clinprotools 2.2

User Manual

Bruker Daltonics

Copyright

Document History
ClinProTools User Manual, Version 2.2 (November 2007)
Part #: 249619
First edition: June 2004
Printed in Germany

Warranty
The information contained in this document is subject to change without notice.

Bruker Daltonik GmbH makes no warranty of any kind with regard to this material, including, but not limited to, the implied warranties of merchantability and fitness for a particular purpose.

Bruker Daltonik GmbH shall not be liable for errors contained herein or for incidental or consequential damages in connection with the furnishing, performance or use of this material.

Bruker Daltonik GmbH assumes no responsibility for the use or reliability of its software on equipment that is not furnished by Bruker Daltonik GmbH.

Bruker Daltonik GmbH

Fahrenheitstrasse 4
28359 Bremen
Germany

Phone: +49 (421) 2205-432
FAX: +49 (421) 2205-106
E-mail: clinprot.support@bdal.de
Internet: http:\\www.bdal.de

Contents

1 PREFACE

The *Bruker Daltonics ClinProTools* 2.2 application (referred to as 'ClinProTools') is an easy-to-use data post-processing software for visualization, data reduction, data mining and building predictive models from protein profiling data using Bruker's Biflex/Reflex, Omniflex, Autoflex or Ultraflex mass spectrometers (MS).

ClinProTools combines intuitive visualization features and multiple mathematical algorithms to generate pattern recognition models for classification and prediction of e.g. disease from mass spectrometry based profiling data. These easy-to-use software features allow customers to rapidly generate and validate biomarker patterns from their protein profiling data.

Key features

The ClinProTools software has the following key features:

- Import of files acquired with Bruker's mass spectrometers, import of ASCII file format possible.
- Display of averaged and single spectra with intuitive visualization features such as virtual gel view and stack view.
- Data processing parameters for baseline subtraction, peak definition, recalibration, normalization etc.
- Statistic analysis of peaks from different spectra.
- Supervised classification model generation and validation using different sophisticated mathematical and bioinformatic algorithms.
- Pattern matching algorithm supporting outlier detection.
- PCA and unsupervised hierarchical clustering.
- Highlighting of the biomarker location. Allows users to visually inspect individual spectrum to verify their results.
- Storage of detailed results for each analysis.

2 INSTALLING AND LICENSING CLINPROTOOLS

Bruker Daltonics ClinProTools 2.2 is supported by Windows2000 and WindowsXP English Version. For details on the required service packs see the read-me file on the installation CD, for system requirements Section 2.1.

Working with ClinProTools 2.2 requires the MATLAB Component Runtime application be installed on your system. Thus, installation first checks if this component is present and if not it prompts you to install the application prior to starting ClinProTools installation.

The ClinProTools software and the MATLAB Component Runtime software are installed from the ClinProTools installation CD delivered. Initial installation of ClinProTools on a computer automatically creates a temporary license valid for 30 days. To work with ClinProTools in future, you have to enter the ClinProTools license key you received. A separate license is needed for usage of the Support Vector Machine algorithm; this license is not part of the 30-day test license.

2.1 System Requirements

CPU: Pentium IV processor equivalent

- Clock: 3 GHz or more for satisfying data handling, double processor machine recommended
- Hard disk: at least 2 GB of free disk space
- Main Memory: 2GB RAM, up to 4 GB are supported
- Operating System: Windows 2000 or Windows XP English Version with the latest Service Packs
- Internet Explorer 7 or 6
- Graphic resolution: 1024x768 pixels, 256 colors or better, optimum 1280x1024 with true colors
- CD–ROM / DVD drive (only for installation)
- Microsoft .NET Platform (will be installed by Setup if not found on computer)

2.2 Installing ClinProTools

The ClinProTools software is installed from the ClinProTools installation CD. If the MATLAB Component Runtime application is not available on your system, you will be prompted installing it prior to the installation of ClinProTools.

Installation notes

- The program should be installed by a user with administrator rights, it is not sufficient to install it as a normal user with "Run as" administrator.
- During setup, the installation of the .Net framework must be affirmed.
- Internet Explorer 6.0 is required and for loading the XML files with style sheets Excel 2002 or higher. Make sure that the Excel security settings (extras/options/securities/macro security) are set to 'low'.
- During the setup, the MATLAB Component Runtime will be installed. Please check "All Users" during MATLAB Component Runtime setup to ensure proper access for all users.
- Sometimes empty tables occur while displaying XML with style sheets in the Internet Explorer. This is due to an out-of-date XML parser registered. Microsoft provides a Replace Mode Tool called Xmlinst.exe, which sets the application reference to a newer XML parser version. The tool can be downloaded from the Microsoft web site (http://www.microsoft.com/downloads/details.aspx?FamilyID=1e6185d7-e4e4-43b1-8056-0e5ecd15a88a&displaylang=en or search for 'Xmlinst.exe' on the web site).
- To ensure that Excel parses the XML files with style sheet properly, make sure that a dot is used as decimal separator by Excel. To enforce this go to the 'Tools/ Option' dialog in Excel. On the 'International' tab at 'Number handling' uncheck 'Use system separators', enter a dot as 'Decimal separator' and a comma as 'Thousands separator'. If this is not set, numbers may be parsed as dates and the like.
- To support more than 2 GB RAM, the /3GB flag has to be set in the boot.ini (Section 2.3).

To install ClinProTools:

1. Start your Windows application.
2. Insert the installation CD into the CD-ROM drive of your computer (e.g. E:). If the 'Autostart' function is activated, the CD browser will start automatically and guide you to start the installation. You can proceed to step 7. Otherwise, if the ‘Autostart’ feature is not turned on, proceed to step 3.
3. Click 

.
4. Click **Run**.
5. In **Open**, type in the command line "E:\setup.exe" (if E: is your CD-ROM drive).

6. Click **OK**. This starts installation and checks if the MATLAB Component Runtime is available on your computer.
7. If the MATLAB Component Runtime is not available, the **InstallShield Wizard** informs you that installing this application is required prior to installing ClinProTools. Click **OK** to start MATLAB Component Runtime installation and follow the **MATLAB Component Runtime – InstallShield Wizard** instructions. When you are asked whether to install the MATLAB Component Runtime for yourself or for anyone who uses your computer it is recommended to choose 'Everyone' if several users work on this computer. Otherwise, there might be the problem that the MATLAB Component Runtime will be available only for the user who installed it. Click **Finish** when you get to the end of the wizard prompts.
8. After installation of the MATLAB Component Runtime, or if it has already been available, installing ClinProTools starts. Follow the **Bruker Daltonics ClinProTools 2.2 – InstallShield Wizard** instructions to set up ClinProTools on your computer. Click **Finish** when you get to the end of the wizard prompts.
9. If you received a license key for ClinProTools, it is recommended to activate the license now (Section 0).

2.3 Supporting More Than 2 GB RAM

More than 2 GB RAM are not supported automatically by Windows XP Professional. To enforce the usage, the /3GB flag has to be set in the boot.ini. In the case of 4 GB RAM, 3 GB are available for the program, while 1 GB is reserved for the operation system. To avoid too little memory left for the operation system we recommend to make use of the /3GB flag only in combination with full 4 GB RAM.

For detailed information, also about other operation systems, please refer to the Microsoft web page at

http://www.microsoft.com/whdc/system/platform/server/PAE/PAEmem.mspx

2.4 Licensing ClinProTools

Installation of ClinProTools on a computer automatically creates a temporary license valid for 30 days. To work with ClinProTools in future enter the license key you received. If you ordered the Support Vector Machine license, also enter the corresponding license key. Permanent as well as temporary licenses can be given. In the latter case, an expiration warning will inform you about the forthcoming expiration firstly 30 days before the license will expire. License keys have to be entered in the Bruker Daltonics LicenseManager, which you can launch from the Windows **Start** menu. The

LicenseManager will list any license currently available for any Bruker Daltonics application. Alternatively, you can open the LicenseManager from ClinProTools using the **LicenseManager** command from the **Compass** menu.

Note: If the license key for the Support Vector Machine is entered when ClinProTools is started, a restart of ClinProTools is necessary to make the Support Vector Machine available.

To license ClinProTools and Support Vector Machine (optionally):

1. Click 

.
2. Click **Programs**.
3. Click **Bruker Daltonics**.
4. Click **Administration**.
5. Click **LicenseManager** to open the **Bruker Daltonics LicenseManager** dialog.

6. Enter the ClinProTools license key in **New license key**.
7. Click **Add**. The button is enabled after entering a correct license key. If the key is valid the license is added to **Existing licenses** (Figure 2-1).
8. Repeat steps 6 and 7 for the Support Vector Machine license key if available.
9. Click **Close**.

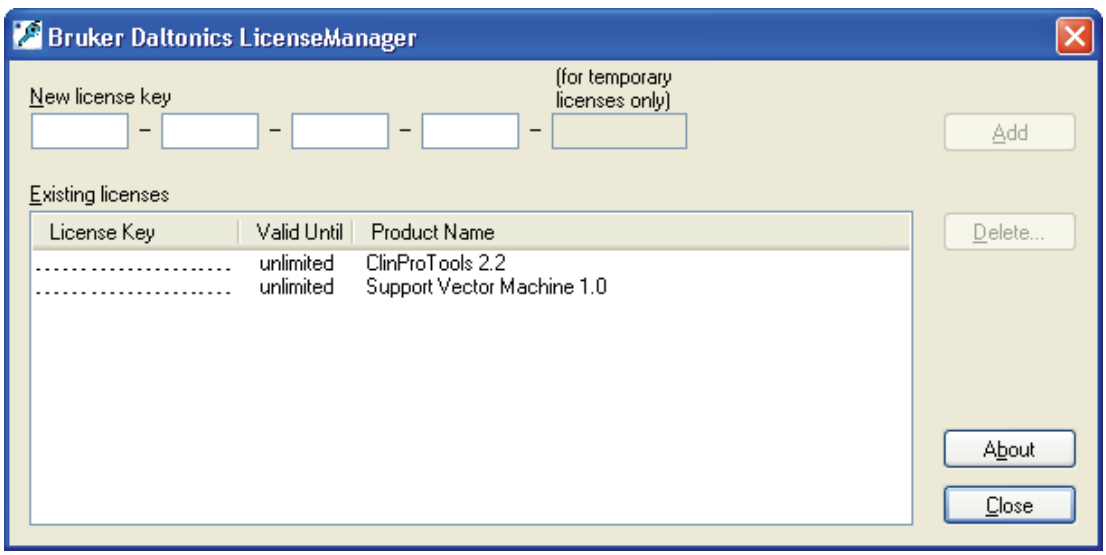


Figure 2-1 Bruker Daltonics LicenseManager dialog with the ClinProTools 2.2 and Support Vector Machine 1.0 licenses present

2.5 Uninstalling ClinProTools

There is no need for uninstalling previous ClinProTools versions when a new ClinProTools upgrade version should be installed on your system. Nevertheless, if you want to uninstall ClinProTools from your system, proceed as follows:

To uninstall ClinProTools:

1. Click 

.
2. Click **Settings**.
3. Click **Control Panel**.
4. Double-click **Add/Remove Program**.
5. Select the ClinProTools 2.2 software from the list of installed programs.
6. Click **Remove**.
7. Confirm the request to remove ClinProTools from your system.

3 DATA ACQUISITION FOR CLINICAL PROTEOMICS

3.1 Introduction

A direct MS analysis of very complex mixtures such as many biological fluids (blood serum, blood plasma etc.) can often end up with unsatisfying spectra quality. Highly concentrated components may suppress minor components, similar mass to charge (*m/z*) ratio peptides and proteins may result in overlapping peaks.

Those features can be avoided when the samples are subjected to prefractionation prior to MS analysis. A selective enrichment of specific peptides, protein fragments and proteins according to their biological, chemical or physical properties can improve spectra quality significantly. Bruker offers an off-line system for enrichment/prefractionation, based on magnetic microbeads with different functionalized surfaces.

The handling of the magnetic beads is simple. They are provided as different kits (ClinProt Kits) and each kit contains a detailed protocol for sample preparation (optimized on blood serum). Additionally, the flexible handling of the beads enables the user to vary the protocols optionally and to adapt them to their special tasks (scaling, concentration variation, multi-step protocols for bead combination...).

3.2 Sample Preparation

For MALDI-TOF MS analysis, it is recommended to prepare the samples on Bruker 384 MTP AnchorChip targets with an optimal anchor diameter of 600 µm (# 209513). In general, target preparation can be performed with a number of different matrices, e.g. according to the mass range of interest, and based on different protocols. For details about MALDI-TOF MS target preparation, please refer to the AnchorChip manual, version 2.2.

In the following a protocol is described which has been optimized for special clinical proteomic approaches to gain profile spectra in the mass range from approx. 1000 to 10000 Da. Please keep in mind that high reproducibility of results is significantly depending on reproducibility of sample preparation in all different steps starting from collecting and storing the samples and ending with target preparation and MALDI-TOF analysis.

Target preparation protocol for profiling samples

Matrix for the mass range 1 – 20 kDa:

Matrix solution: α-cyano-4-hydroxycinnamic acid (HCCA, # 201344, # 201072), 0.3 g/l in ethanol:acetone 2:1 (daily prepared)

Take 1 µl of the sample (purified and directly eluted from the magnetic beads according to the protocol) and mix it thoroughly with 10 µl of the matrix solution. Subsequently, 0.5 to 1 µl of the mixture should be applied onto one anchor position of the target and allowed to dry at room temperature. It is recommended to work continuously as matrix- and several sample-solutions contain very volatile solvents; uncontrolled evaporation may result in decreased preparation quality. The measurement variance is reduced by spotting each sample several times.

Matrix for the mass range 5 – 100 kDa:

Matrix solution: 7.6 mg (50 µmol) 2,5-DHAP are suspended in 375 µl EtOH and 125 µl (10 µmol) of diammonium hydrogen citrate (stock solution: 27 mg in 1.5 ml distilled H_2O) are added. The suspension should be vortexed for at least 1 min followed by sonification for 15 min. The mixture has to be vortexed again (1 min) and the clear matrix solution is suitable for MALDI-TOF MS analysis now.

For target preparation, 2 µl of sample are acidified with 2 µl of 2% TFA. Subsequently, 2 µl of freshly prepared 2,5-DHAP matrix solution have to be added and vigorously mixed. Finally, 1 µl of the mixture should be applied onto one anchor position of the target and allow to dry at room temperature. Parallel spotting on multiple target positions is recommended as well.

3.3 Data Acquisition with flexControl

A Bruker MALDI-TOF mass spectrometer will be delivered with a number of acquisition methods which were specifically adapted to the individual machine during installation and which can be loaded directly into the acquisition software flexControl. Those default methods cover e.g. different mass ranges and it is recommended for inexperienced users to start working with one of them and to adapt the method to their special approaches.

In the following two model methods, created on Autoflex- and Ultraflex-MALDI-TOF MS respectively, for generating clinical proteomics profiles are described (Table 3-1). These parameters are for information only and are to be regarded only as a guide; they should not simply be copied to the user's mass spectrometers as each machine can differ. They represent first easy-to-use starting parameters for the profiling user. Usually, only slight changes are necessary to create an individual method according to each machine. Gray highlighted values in the table represent values that do not signi-

ficantly change between different mass spectrometers and can generally be defined as fixed.

Table 3-1 Recommended method parameters (linear mode) for measuring clinical proteomics profile spectra

Note: Please note that parameter sets have to be optimized for different proteomic samples and for every instrument.

Parameters	**Autoflex - 1 – 10 kDa -**	**Ultraflex - 1 – 10 kDa -**
N_2 pressure	approx. 1700 - 2000 mbar	approx. 1700 - 2000 mbar
Laser	individually adjustable, it is recommended to shoot approx. 15 shots with higher laser power, followed by 30 shots on the same position with approx. half of the initial laser power	individually adjustable, it is recommended to shoot approx. 15 shots with higher laser power, followed by 30 shots on the same position with approx. only half of the laser power
Parameters	**Autoflex - 1 – 10 kDa -**	**Ultraflex - 1 – 10 kDa -**
Shots	30 (the 15 pre-shots should not be added to the sum-spectra) from 15 to 18 different positions on one anchor (sum: 450-550 shots)	30 (the 15 pre-shots should not be added to the sum-spectra) from 15 to 18 different positions on one anchor (sum: 450-550 shots)
Spectrometer		
Ion Source 1	20 kV	25 kV
Ion Source 2	18.4 kV	23.2 kV
Lens	7.5 kV	6 kV
Pulsed Ion Extraction	120 ns	350 ns
Polarity	positive	positive
Matrix Suppression mode	gating	gating
Gating strength	high / maximum	medium / high
Suppress up to	approx. 800 Da	approx. 800 Da

Parameters	**Autoflex - 1 – 10 kDa -**	**Ultraflex - 1 – 10 kDa -**
Detection		
Mass Range	low: 900 – 10,500 Da	low: 900 – 10,500 Da
Detector Gain	1600 - 1800 V	1600 - 1800 V
Sample Rate	1.00	1.00
Electronic Gain	regular, 100 mV	regular, 100 mV
Parameters	**Autoflex - 1 – 10 kDa -**	**Ultraflex - 1 – 10 kDa -**
Real time smooth	high	high
Set up (***Be careful***, *changes are not saved in the method!)*		
Laser Frequency	25 Hz	25 Hz
Laser Attenuator	e.g. 60 / 30	e.g. 60 / 30

4 GETTING STARTED WITH CLINPROTOOLS

4.1 Starting ClinProTools

You can start ClinProTools from Windows **Start** menu. When ClinProTools is installed, a **Bruker Daltonics** folder containing this application (and perhaps other Bruker Daltonics applications) is created in the **Start** menu's **Programs** folder. Alternatively, you can double-click the ClinProTools icon created on your desktop during installation.

If ClinProTools is started without a valid license being present, a message informs you that ClinProTools has not been licensed yet. Confirming this message automatically starts the Bruker Daltonics LicenseManager.

On ClinProTools start-up, the files *SettingsDataPreparation.xml* and *SettingsModelGeneration.xml* are generated containing the corresponding settings. On repeated start, these files are loaded by the application if present; otherwise, a new one with default values is generated. In addition, a file named *SettingsGeneral.xml* is generated in the same way. All files are set up in the ClinProTools folder (Section 4.2).

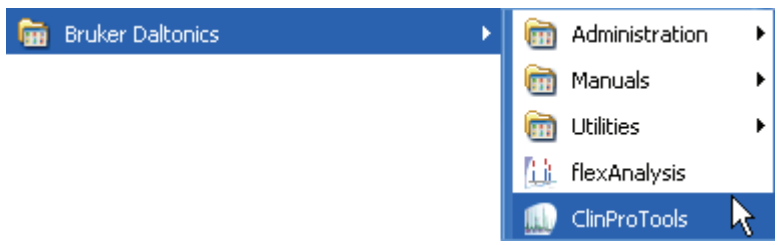
Starting ClinProTools also initializes MATLAB except the option to disable MATLAB is checked in the **General Settings** dialog (Section 9.1.1.12).

For detailed description of the ClinProTools user interface, please refer to Section 5.

To start ClinProTools from Windows Start menu:

1. Click 

.
2. Click **Programs**.
3. Click **Bruker Daltonics**.
4. Click **ClinProTools**. This starts ClinProTools.



5. If ClinProTools has not been licensed yet, confirm the corresponding message and license ClinProTools (Section 0). After licensing, ClinProTools will be started.
6. If the *SettingsDataPreparation.xml* and/or *SettingsModelGeneration.xml* file is/are not available, quit information on starting with default values.

4.2 ClinProTools File Location

All files created by ClinProTools 2.2 will be saved to the ClinProTools folder

"C:\BDAL\ClinProTools_2_2\Files"

The style sheets like ClinProtModel.xls, ClinProtClassification.xls etc. and the default settings files will be installed in this folder, too. On installation, five subfolders are created in the ClinProTools folder, which will be opened by the Load/Save dialogs by default: \ClinProtClassifications, \ClinProtModels, \ClinProtModelSpectraImport, \SettingsDataPreparation and \SettingsModelGeneration.

Note: If you have previously worked with ClinProTools 2.0 and/or ClinProTools 2.1 the corresponding ClinProTools folder "C:\BDAL\ClinProTools_2_0\Files" and/or "C:\BDAL\ClinProTools_2_1\Files" will be kept containing, amongst others, the style sheets for the ClinProTools 2.0 XML and/or ClinProTools 2.1 XML files.

ClinProTools saves temporary *ClinProt*.xml* and *ClinProt*.txt* files in the ClinProTools folder. For example, each time the **Spectra List** or **Peak Statistic** command is performed a new temporary *ClinProtSpectra.xml* resp. *ClinProtStatistic.xml* file is generated getting a running number (e.g. *ClinProtSpectra0001.xml*, *ClinProtSpectra0002.xml*, *ClinProtStatistic0001.xml*, *ClinProtStatistic0002.xml*, etc.).

ClinProTools allows clearing all temporary files in the ClinProTools folder at once using the **General Settings** command from the **File** menu and clicking **Clear Temporary XML Files**. If you do not want to delete all these files, you can select and delete only the desired ones using e.g. the Microsoft Windows Explorer.

4.3 ClinProTools General Settings

ClinProTools allows defining certain general, non-algorithm settings for like file paths, display of *ClinProt*.xml* files etc. If you do not want to work with the defaults you can define own settings. Changed settings can be reset to the defaults.

The ClinProTools general settings are saved in the *SettingsGeneral.xml* file. This file is generated on ClinProTools start-up and updated on each settings change. If this file is not present when ClinProTools is started, a new one with default values will be generated. The *SettingsGeneral.xml* file also collects the file open paths, correlation and statistic settings.

Note: Resetting the general settings is only possible when no spectra are loaded.

To view and change general settings for ClinProTools:

1. From the **File** menu, select **General Settings**.

2. In the **General Settings** dialog, change the default settings if desired.
3. Click **OK**.

To reset general settings including file open paths, statistic and correlation settings to defaults:

1. From the **File** menu, select **General Settings**.
2. In the **General Settings** dialog, click **Reset General Settings**.
3. Confirm the request on resetting to defaults.
4. Click **OK**.

4.4 Three Basic Workflows in ClinProTools

ClinProTools offers three basic workflows, 'Peak Statistic Calculation', 'Model Generation' and 'Classification'. To get familiar with the ClinProTools user interface and basic processing features we recommend that you run these basic workflows with the ClinProTools demo data from your installation CD simply using the ClinProTools' default settings.

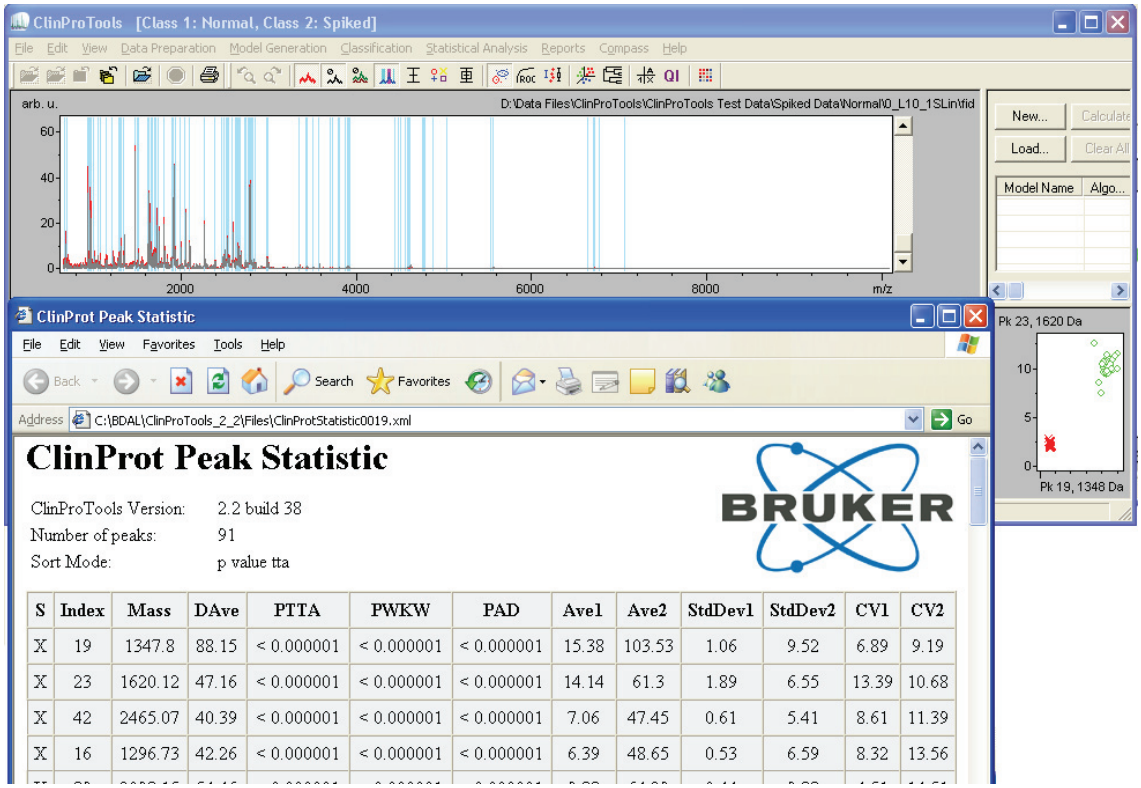
4.4.1 Basic Workflow 'Peak Statistic Calculation'

The basic workflow 'Peak Statistic Calculation' can be used to quickly calculate peak statistics using ClinProTools' default settings. This workflow includes spectra recalibration and average spectra calculation, peak picking and peak calculation as well as peak statistic calculation. The statistic results are automatically shown in the Peak Statistic report and stored as *ClinProtStatistic[number].xml* file.

To run the 'Peak Statistic Calculation' workflow:

1. Load one or more model generation classes (e.g. "Normal" and/or "Spiked" from the "ClinProTools Test Data" folder on the installation CD) using the **Open Model Generation Class** command from the **File** menu or  . One class can be loaded at a time. For this, select the folder with the spectra you want to load as a class. ClinProTools loads all spectra in a folder and its subfolders as one class and prepares them.
2. Start peak statistic calculation using the **Peak Statistic** command from the **Reports** menu or  . This runs the spectra recalibration, spectra averaging and peak calculation processes on the loaded spectra.
3. View the Peak Statistic report (Section 8.1.1.2), which opens automatically and

lists all picked peaks with corresponding statistical data ordered according to the sort mode for peak selection (Figure 4-1).
The Spectra View (Section 5.1.1) shows all picked peaks marked by highlighting their integration regions and the 2D Peak Distribution View (Section 5.1.3.1) displays the distribution of the two first (best separating) peaks of the peak statistic (Figure 4-1).



ClinProt Peak Statistic

BRUKER

ClinProTools Version: 2.2 build 38
Number of peaks: 91
Sort Mode: p value tta

S	Index	Mass	DAve	PTTA	PWKW	PAD	Ave1	Ave2	StdDev1	StdDev2	CV1	CV2
X	19	1347.8	88.15	< 0.000001	< 0.000001	< 0.000001	15.38	103.53	1.06	9.52	6.89	9.19
X	23	1620.12	47.16	< 0.000001	< 0.000001	< 0.000001	14.14	61.3	1.89	6.55	13.39	10.68
X	42	2465.07	40.39	< 0.000001	< 0.000001	< 0.000001	7.06	47.45	0.61	5.41	8.61	11.39
X	16	1296.73	42.26	< 0.000001	< 0.000001	< 0.000001	6.39	48.65	0.53	6.59	8.32	13.56

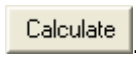
Figure 4-1 Results of running the basic workflow 'Peak Statistic Calculation'

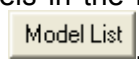
4.4.2 Basic Workflow 'Model Generation'

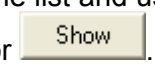
The basic workflow 'Model Generation' can be used to quickly calculate models using ClinProTools' defaults settings. This workflow includes spectra recalibration and average spectra calculation, peak calculation and model generation based on the selected classification algorithm. Data of all models present in the model list or of a selected model can be shown in the Model List report or the Model report using the corresponding command. This stores the respective data as *ClinProtModelList[number].xml* file or *ClinProtModel[number].xml* file, respectively.

To run the 'Model Generation' workflow:

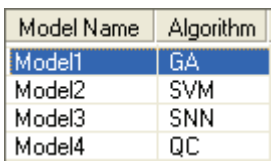
1. Load the two classes "Normal" and "Spiked" from the "ClinProTools Test Data" folder on the installation CD using the **Open Model Generation Class** command from the **File** menu or  . One class can be loaded at a time. For this, select the folder with the spectra you want to load as a class. ClinProTools loads all spectra in a folder and its subfolders as one class and prepares them.
2. Add model parameter sets to the Model List View using the **New Model** command from the **Model Generation** menu or 

. For this, select the classification algorithm (GA, SVM, SNN or QC), click **OK** in the appearing algorithm-specific settings dialog to use the defaults and enter a model name. Repeat this procedure for each of the four algorithms.
3. Start model calculation using the **Calculate** command from the **Model Calculation** menu or 

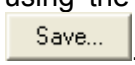
. This runs data preparation (spectra recalibration, spectra averaging and peak calculation) on the loaded spectra and after that generates a model for each added model parameter set. The generated models are entered in the model list.
4. To view the parameters of all models in the model list use the **Model List** command from the **Reports** menu or 

. This opens the Model List report (Section 8.1.1.5) (Figure 4-2).
5. To view a model select it from the list and use the **Show Model** command from the Model List View context menu or 

.



Model Name	Algorithm
Model1	GA
Model2	SVM
Model3	SNN
Model4	QC

This opens the Model report (Section 8.1.1.6), which lists all parameters of the selected model (Figure 4-2).
The Spectra View (Section 5.1.1) shows the peaks that are incorporated in the current model now having red integration regions instead of blue ones (Figure 4-2).
6. Models are not saved automatically. If you want to save a model, select it from the list and save it using the **Save Model As** command from the Model List View context menu or 

.

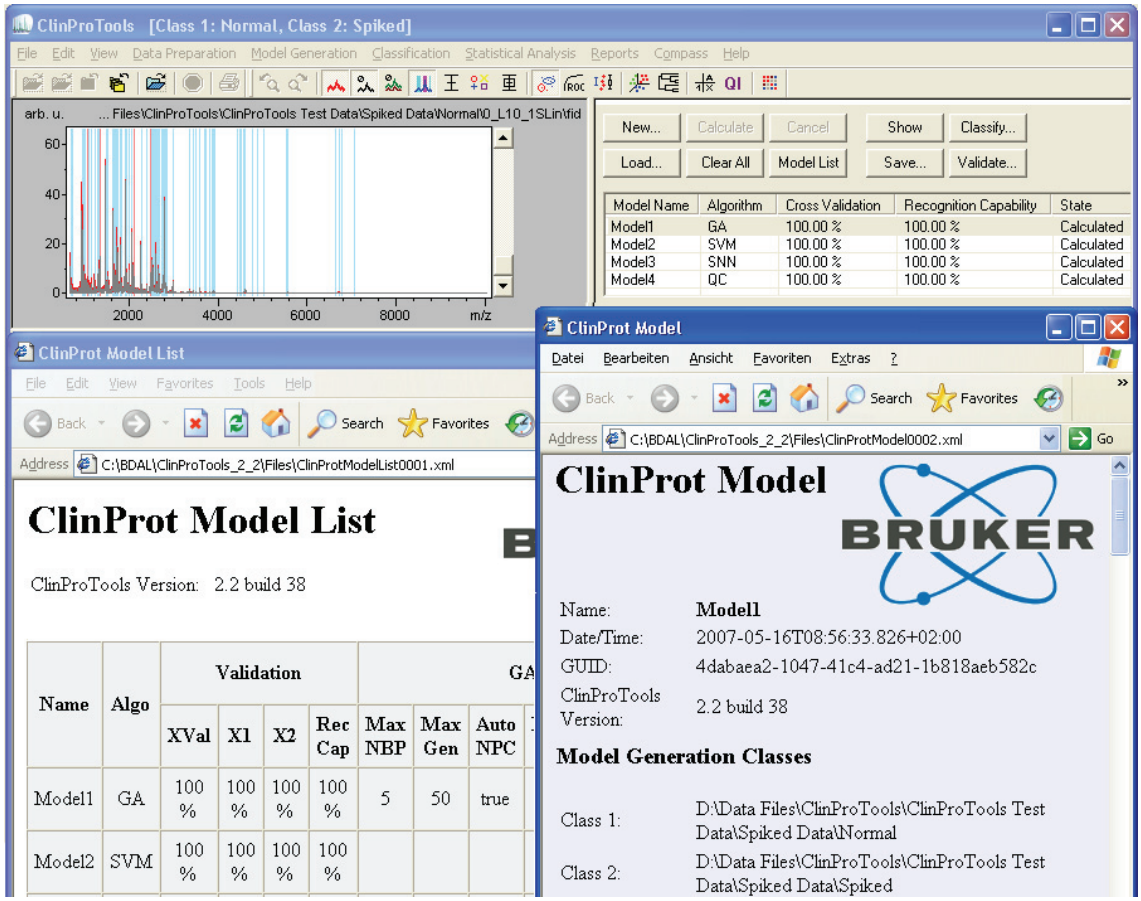


Figure 4-2 Results of running the basic workflow 'Model Generation'

4.4.3 Basic Workflow 'Classification'

The basic workflow 'Classification' can be used to quickly classify test spectra with an existing model. This workflow includes selecting the model to use and the spectra to classify, data preparation of these spectra according to the settings saved in the model and their classification. The classification result is automatically shown in the Classification report and stored as *ClinProtClassification[number].xml* file.

To run the 'Classification' workflow:

1. Generate a model as described in the 'Model Generation' workflow (Section 4.4.2) or load a previously generated and saved model using the **Load Model** command from the **Model Generation** menu or Load....

2. Select the model to use from the model list.

Model Name	Algorithm
Model1	GA
Model2	SVM
Model3	SNN
Model4	QC

3. Load the "To Classify (5 + 5)" spectra collection from the "ClinProTools Test Data" on the CD using the **Classify** command from the **Classification** menu or Classify... . This prepares the spectra according to the parameter settings saved with in the current model and classifies the spectra.

4. View the Classification report (Section 8.1.1.8), which opens automatically and lists the classification results (Figure 4-3).
The Gel View (Section 5.1.2.1) and Spectra View (Section 5.1.1) now also display the spectra of the classified spectra collection (Figure 4-3).

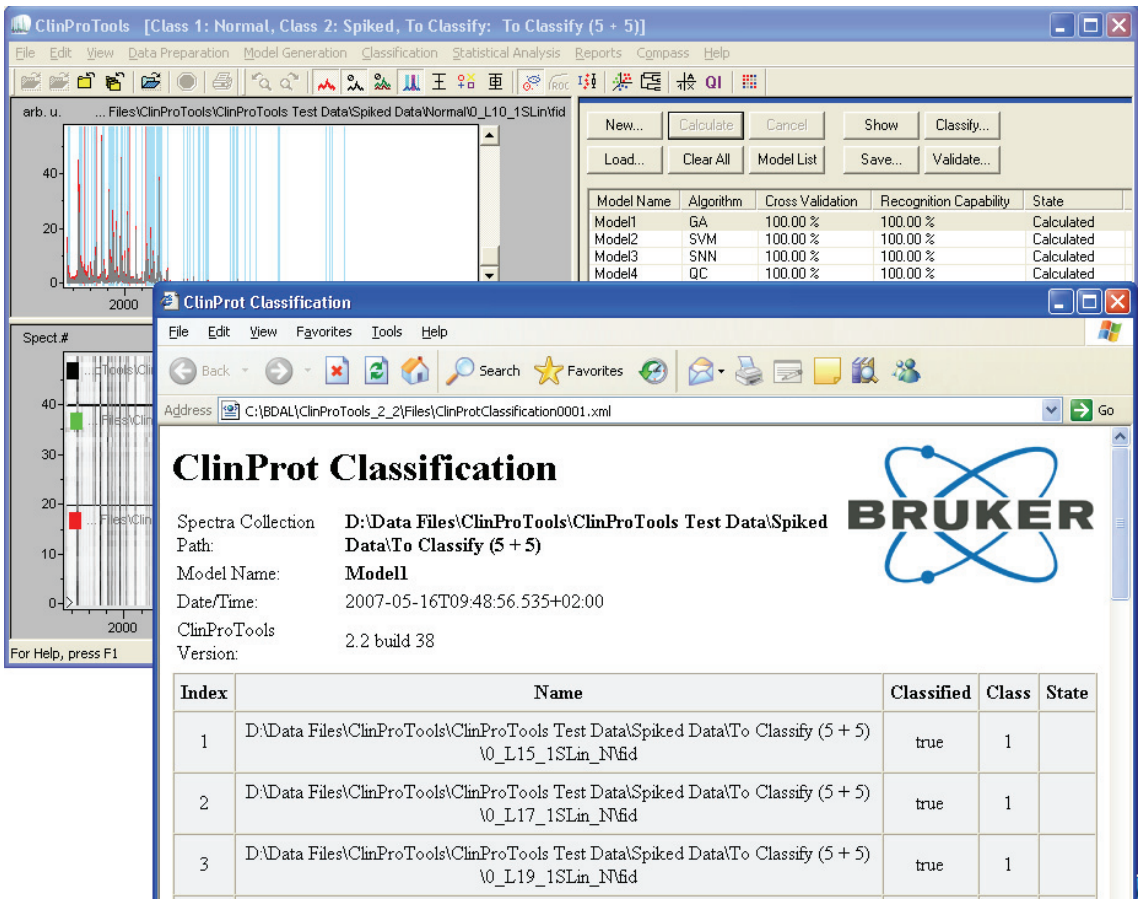


Figure 4-3 Results of running the basic workflow 'Classification'

4.5 Closing ClinProTools

You can close ClinProTools when you have finished your current session. To exit ClinProTools you have to answer a confirmation request.

To close ClinProTools:

1. From the **File** menu, select **Exit** or click the application's 
2. Answer the confirmation request to close ClinProTools.

5 CLINPROTOOLS USER INTERFACE

On starting ClinProTools, the ClinProTools window opens. Herein all processing operations ClinProTools supports are started and most of the results displayed. Only results of a Principal Component Analysis (PCA) or an unsupervised hierarchical clustering display in separate windows, the PCA windows and the Dendrogram window, which originate from the external MATLAB® software tool integrated in ClinProTools.

5.1 ClinProTools Window

The ClinProTools window (Figure 5-1) consists of four views, Spectra View, Gel/Stack View, Peak Statistics View and Model List View, the title bar, menu bar, toolbars and status bar.

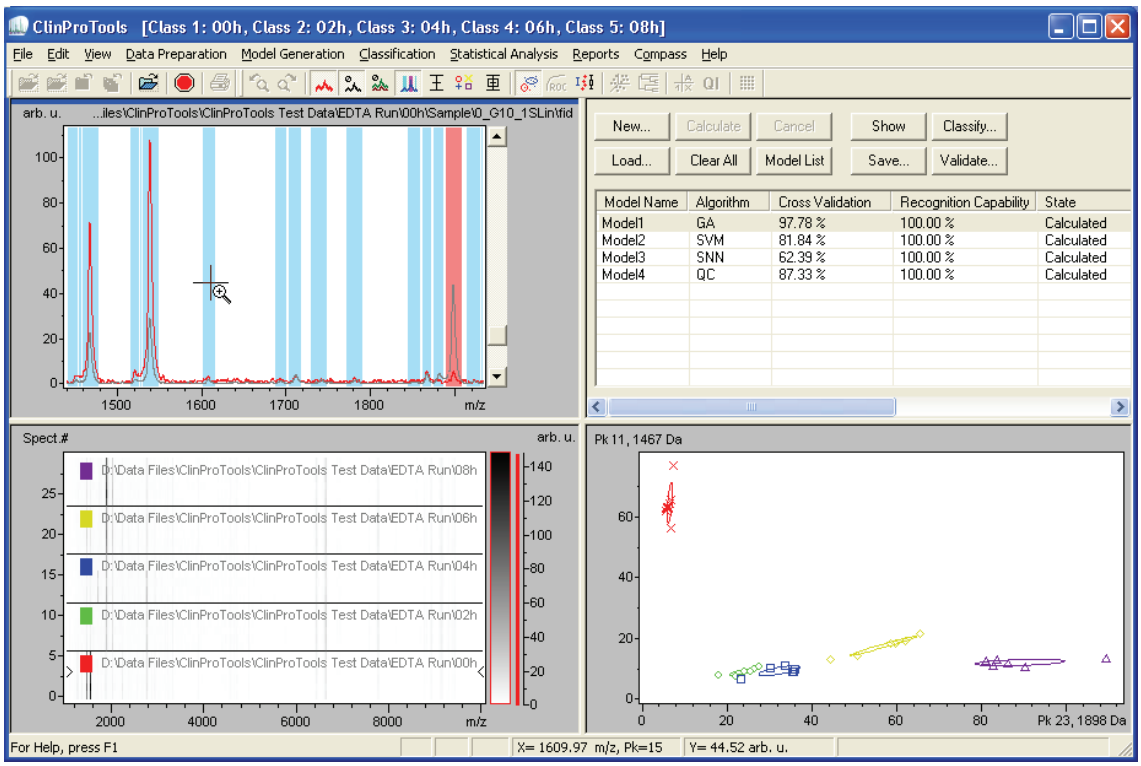


Figure 5-1 ClinProTools window with Spectra View (top left), Model List View (top right), Gel/Stack View (bottom left) and Peak Statistics View (bottom right)

The views are displayed in split view and are opened by default. The active view is marked with a blue selection bar on top. You can resize the views by dragging the horizontal or vertical split borders with the left mouse button held down. All four views can be changed at once by dragging the split cross. Moving borders may result in hiding (a) view(s); a hidden view can be shown again by dragging the corresponding border(s) accordingly. You can alter the data plotting views in various ways (Section 5.1.7). On closing ClinProTools, certain settings for the views are saved and reloaded on next program start.

The colors that are used to display the single spectra of different class membership and the corresponding peak statistic data as well as the calculated average spectrum/ spectra and the noise spectrum are predefined in the system. All single spectra of a certain class get the same color. Class 1 (first loaded class) spectra are displayed in red, class 2 (second loaded class) spectra in green, class 3 (third loaded class) spectra in blue, etc. A maximum of ten different class colors is defined; if you load more than ten classes, class coloring will continue starting again with red. Only in case of using spectra import XML files there is the possibility to define own class colors for displaying the spectra of the referenced classes and corresponding data.

5.1.1 Spectra View

The Spectra View (Figure 5-2) displays the single spectra of the loaded model generation classes, the calculated average and noise spectra, and specific peak statistics. The x-axis records the *m/z* value, the y-axis the peak intensity in arbitrary units. The statistical plots are drawn on a unique scale independent of the peak intensity scale. The kind of spectra and peak statistics displayed depends on the current processing state and the corresponding **View** menu settings (Section 9.1.3.6).

Single spectra are displayed by default with one single spectrum shown at a time (Section 9.1.3.6.1). Alternatively, you can show all single spectra in an overlay spectra plot (Section 9.1.3.6.2). The path and name of the current spectrum are indicated in the top right corner of the view. All single spectra of one class display in the same color that is indicated in the Gel View. To show another single spectrum of the same or another class use the scroll bar or click the spectrum in the Gel View.

After spectra recalibration, the total average spectrum (Section 9.1.3.6.3) is shown by default in gray color. In addition, you can show class average spectra (Section 9.1.3.6.4) displayed with a darker color than the corresponding single spectra (e.g. red > dark red) and/or the noise spectrum in orange (Section 9.1.3.6.5).

After peak calculation and peak selection, all picked peaks are marked by colored integration regions (Section 9.1.3.6.6) by default; included peaks with blue, excluded peaks with gray bars. After model generation, red bars instead of blue ones mark the peaks incorporated in the model selected in the model list. Certain peak statistic data (average with standard deviation, peak distribution, box and whiskers; Sections

9.1.3.6.7 to 9.1.3.6.9) can be displayed for all or a restricted number of peaks. All symbols used are colored like the corresponding class. The peak(s) shown in the Peak Statistics View can be marked with black arrows on the top of the Spectra View (Section 9.1.3.6.11).

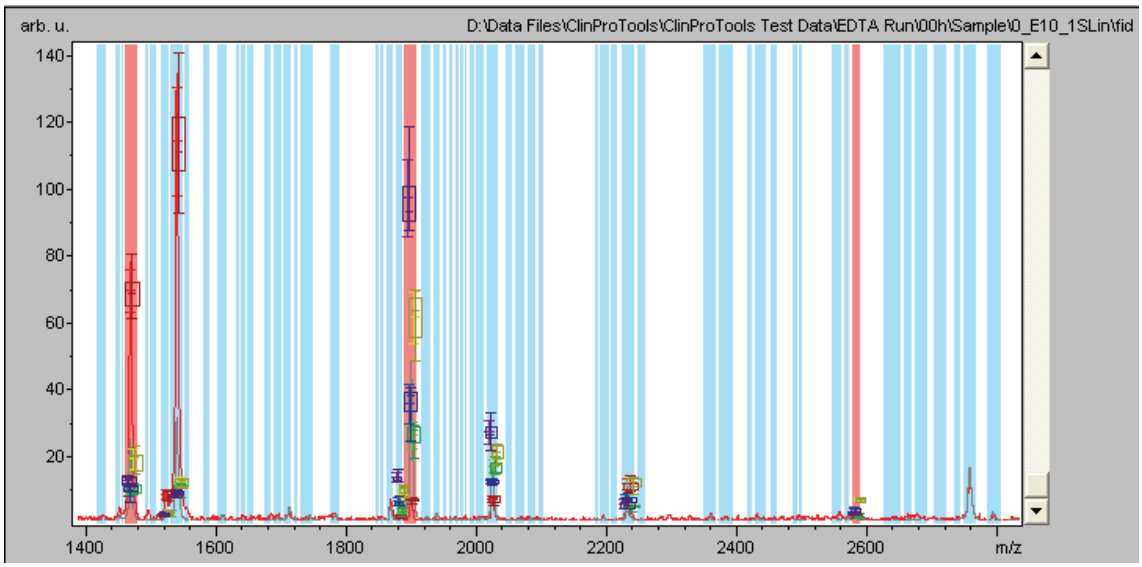


Figure 5-2 **Spectra View after generating a model: A single and the total average spectrum are displayed with marking the picked peaks in blue and the peaks incorporated in the model in red; additionally certain peak statistic data is shown.**

The Spectra View allows manual exclusion/inclusion of unprocessed spectra and picked peaks, manual editing of the average peak list and forcing peaks into a model. Excluded spectra are displayed with a darker color than the respective included ones (e.g. in dark red instead of red). Excluded peaks are marked by gray integration regions instead of blue ones, forced peaks by green ones.

You can switch the Spectra View to distance mode to show *m/z* differences for two selected peaks (Section 9.2.9.10).

5.1.2 Gel/Stack View

The Gel/Stack View consists of two views, Gel View and Stack View. You can toogle between the views using the **Display Type > Gel View** and **> Stack View** commands from the view's context menu. The Gel View is displayed by default.

5.1.2.1 Gel View

The Gel View (Figure 5-3) displays all spectra of the loaded classes arranged in a pseudo-gel like look. The x-axis records the *m/z* value. The left y-axis displays the running spectrum number originating from subsequent spectra loading. The peak intensity is expressed by a color code. The color bar and the right y-axis indicate the relation between the color a peak is displayed with and the peak intensity in arbitrary units. Various color modes are available (Section 9.2.9.6).

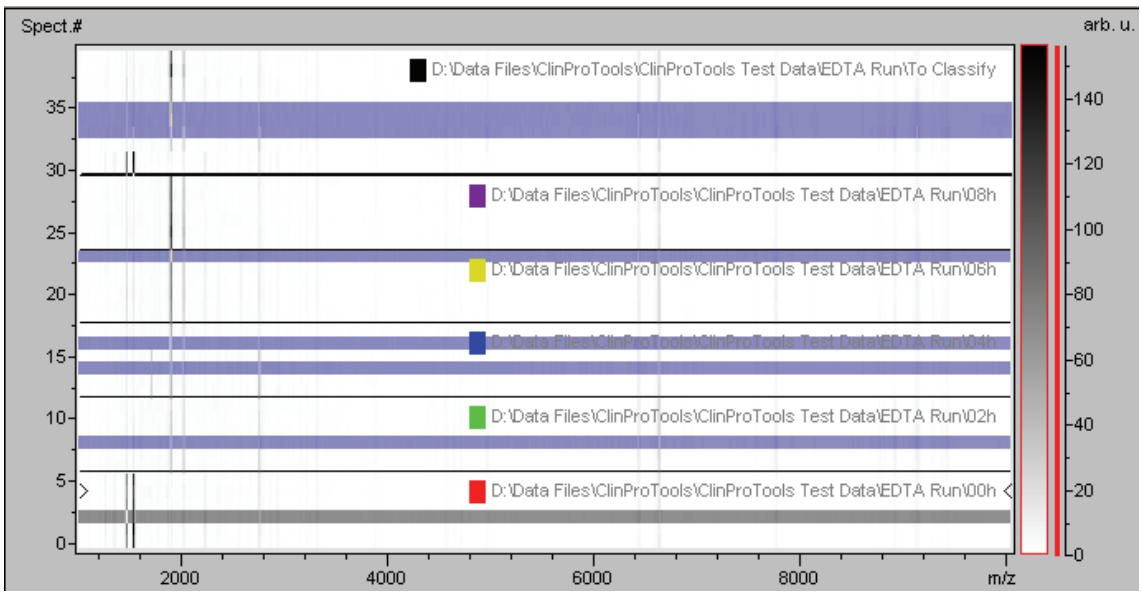


Figure 5-3 **Gel View showing the spectra of five model generation classes (red, green, blue, ocher, violet) and a spectra collection to be classified (black) using a linear gray scale for intensity display**

The spectra of the first loaded class (class 1) are displayed at the bottom of the view, the spectra of the second loaded class (class 2) above the class 1 spectra, the spectra of the third loaded class (class 3) above the class 2 spectra, etc. Each class is separated from the next loaded one by a horizontal line. Classes used in model generation are separated by thin lines, a class used in classification is separated from the last loaded model generation class by a thicker line.

The class names (Section 9.1.3.7.1) consisting of path and folder name of the respective class are shown by default. The color box in front of the class name indicates the color the single spectra contributing this class are displayed with in the Spectra View. 12 class colors are predefined in the system; after 12 classes, repetition of colors will occur.

The current spectrum marker (Section 9.1.3.7.2) marks the spectrum currently shown in the Spectra View. When spectra from multiple measurements are loaded and spectra grouping is enabled, dashed group separators (Section 9.1.3.7.5) are shown by default which separate spectra originating from the same spot.

Like the Spectra View, the Gel View allows manual exclusion/inclusion of unprocessed spectra. Manually as well as automatically excluded spectra are highlighted by default, using colored spectrum states concerning the reason of exclusion (Section 9.1.3.7.3). Excluded spectra can be hidden from the Gel View (Section 9.1.3.7.4).

5.1.2.2 Stack View

The Stack View (Figure 5-4) displays all spectra of the loaded classes in a three dimensional space. The x-axis records the *m/z* value, the y-axis the peak intensity in arbitrary units and the z-axis the loading order. The spectra of the first loaded class are in the foreground, those of the last loaded one in the background. The default orientation of the plot is 30° but you can quickly change it by dragging all axes at once using the mouse (Section 5.1.7.3).

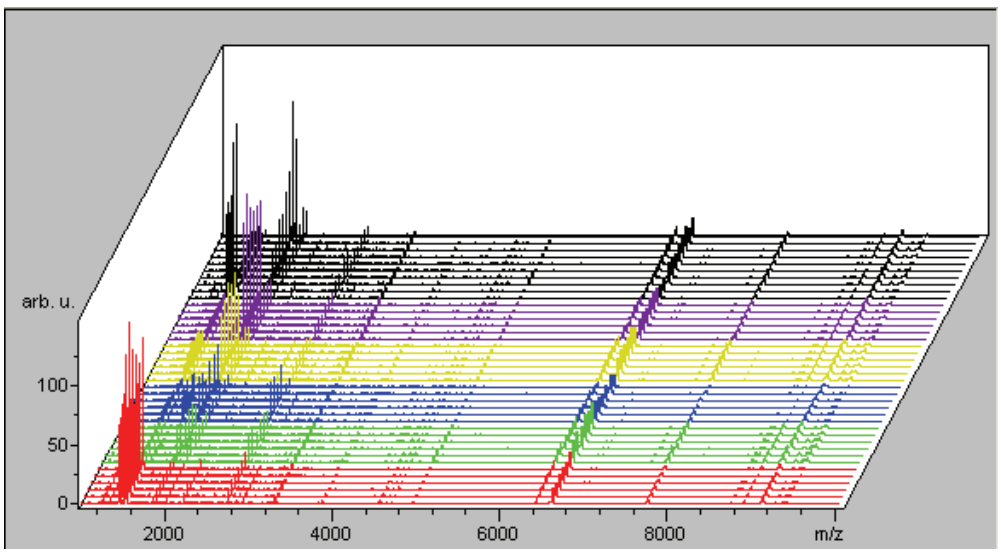


Figure 5-4 **Stack View displaying the spectra of five model generation classes (red, green, blue, ocher, violet) and of a collection to be classified (black)**

The spectra are colored according to their class membership by default. Like in the Spectra View excluded spectra are displayed with a darker color than the corresponding included spectrum (e.g. red > dark red). Excluded spectra can be hidden (Section 9.1.3.7.4). The Stack View can be switched to whitewash mode (Section

9.2.9.26) resulting in a finer structured plot due to resolving overlying structures but hiding the coloring of the class membership of the spectra.

5.1.3 Peak Statistics View

The Peak Statistics View consists of three views, 2D Peak Distribution View, ROC Curve View and Single Peak Variance View. You can toggle between the views using the **Peak Statistics View > 2D Peak Distribution**, **> ROC Curve** and **> Single Peak Variance** commands from the **View** menu. The 2D Peak Distribution View is displayed by default. Switching to ROC Curve or Single Peak Variance View is possible after peak calculation was performed; however, the ROC Curve View can be activated only for the case of two loaded classes.

5.1.3.1 2D Peak Distribution View

The 2D Peak Distribution View (Figure 5-5) displays the distribution of two selected peaks in the non-excluded spectra of the loaded model generation classes. The peak numbers and *m/z* values are indicated on the x- and y-axes. When a classification was performed in standard mode, the view (additionally) displays the respective peak data for the classified spectra.

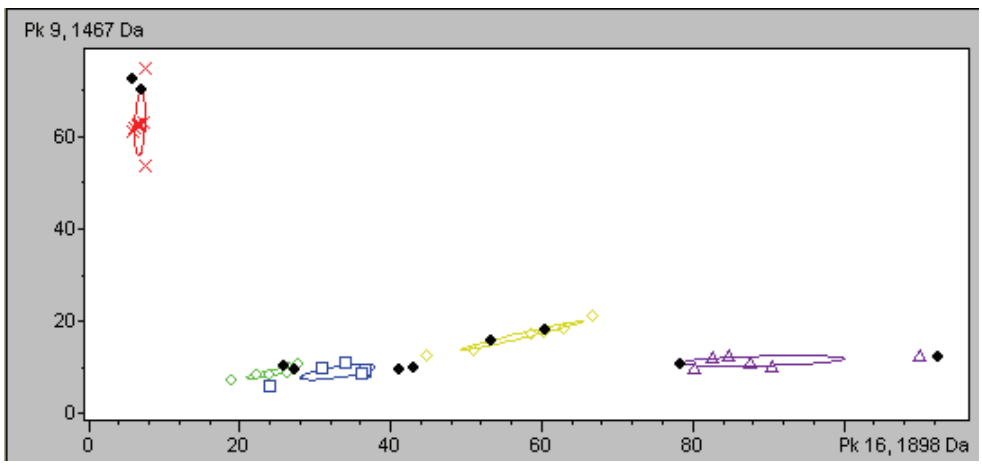


Figure 5-5 **2D Peak Distribution View displaying the distribution of two peaks in the spectra from 5 model generation classes and a classified collection; the ellipses represent the standard deviation of the peak area class average**

The data is shown on a two-dimensional plane. By default, the first two (= best separating) peaks of the current statistic sort order are displayed. Depending on the peak calculation settings, the x-axis shows the peak area/intensity values with respect

to the most important peak in accordance to the sort criterion e.g. its p-value, and the y-axis the peak area/intensity values for the second most important peak, respectively. If the sort criterion is changed, the 2D Peak Distribution View may change too, because there may now be other peaks that are considered as most important. The axis measures are given in arbitrary units which are chosen automatically to fit the plot optimal in the plane. You can change the default peak selection (Section 9.1.3.8.5.1).

All data points belonging to the same model generation class resp. to the classified spectra collection are displayed with the same symbol colored according to the class color, e.g. red cross (class 1 spectra), green diamond (class 2 spectra) or black diamond (classified spectra). If multiple measurements are used and spectra grouping is enabled the peaks of all spectra of a group are averaged before they are processed by the algorithms. However, the 2D Peak Distribution View does not show the averaged peaks but the peaks from all spectra.

The ellipses can represent the standard deviation of the class average of the peak areas/intensities or the 95% confidence interval, which is the standard deviation weighted by the reciprocal number of data points (Section 9.1.3.8.5.2). Classified spectra, of course, are shown without such statistic information.

5.1.3.2 ROC Curve View

The ROC Curve View (Figure 5-6) displays the Receiver Operating Characteristic (ROC) curve (Section 6.4.2.2) for the selected peak generated from all included spectra of the loaded two model generation classes. The x-axis records the '1-specificity' in terms of the false positives and the y-axis the 'sensitivity' in terms of the true positives; both axes are given in values between 0 and 1. To set up a ROC curve a decision has to be made whether class 1 or class 2 should be treated as positive (Section 9.1.3.8.2).

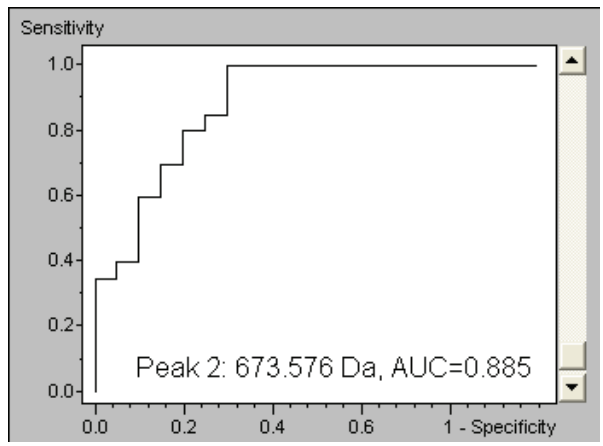


Figure 5-6 ROC Curve View displaying the ROC curve for a peak and corresponding data

The number and *m/z* value of the peak the ROC curve is displayed for is indicated at the bottom of the plot followed by the ROC curve's AUC value (Section 6.4.2.2). If the information is not shown or only partially displayed, broadening the view will help. You can use the view's scroll bar to browse through the different ROC curves over the present set of peaks.

5.1.3.3 Single Peak Variance View

The Single Peak Variance View (Figure 5-7) can display three kinds of statistical data for a selected single peak. The box and whiskers (with/without outliers) (Section 9.1.3.6.9), peak distribution (Section 9.1.3.6.8) or average with standard deviation plots (Section 9.1.3.6.7) calculated from the area/intensity values of the selected peak in the loaded spectra are shown separately for each class. The selected peak is indicated in the top right corner. The view shows the variance for all peaks even if there are only few peaks selected in the statistic settings (Section 9.1.8.5).

Only one kind of statistic can be displayed at a time. The data shown depends on the state of the **View** menu commands **Spectra View > Box & Whiskers**, **Peak Distribution** and **Average & StdDev**. If none of the commands is active (default setting) when switching to the Single Peak Variance View, automatically the box and whiskers plots are set up in both the Spectra and the Single Peak Variance View. If only one command is active, the corresponding statistical data is shown in the view. If more than one command is active, a hierarchical order among the three statistics defines which one is displayed; at this, box and whiskers takes priority over peak distribution and peak distribution again over average with standard deviation.

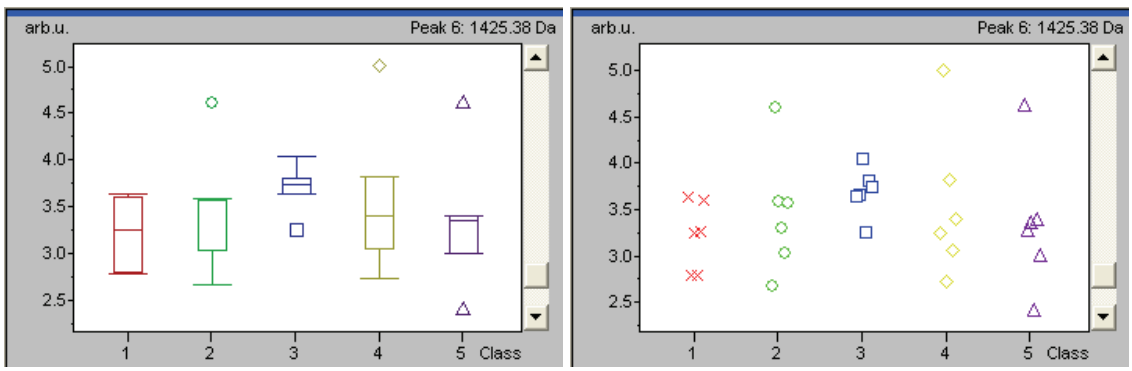


Figure 5-7 Single Peak Variance View displaying the box and whiskers with outliers plots (left) resp. the peak distribution plots (right) for a peak in the spectra from five classes

The plots are drawn on a unique y-scale that is set to auto-scaling by default like in the Spectra View. In horizontal direction, the box and whiskers and the average with standard deviation are spread over the available space and their sizes are automatically adjusted when the window is resized. The size of the peak distribution symbols remains constant when resizing the window. The symbols of one class are slightly horizontally displaced to avoid drawing several symbols in the same place and thus giving a wrong impression of density.

To display data of another peak you can use the view's scroll bar to browse through the peaks or right-click the desired peak it in the Spectra View and use the **Variance for Peak n** command (Section 9.2.9.24).

5.1.4 Model List View

The Model List View (Figure 5-8) lists all models currently available in ClinProTools in a tabular format, the model list. It also offers various buttons to quickly launch model-related commands.

Each model is displayed along with corresponding data, **Model Name**, classifier **Algorithm**, **Cross Validation** and **Recognition Capability** results, current **State** and **Date/Time** of calculation. The data available for a model depends on the model's current **State**. An added but still not calculated model has the state 'Added'. The state 'Calculated' indicates an already calculated model from the current session and the state 'Loaded' a loaded, formerly saved model. The corresponding cross validation and recognition capability results are shown as well as date/time of model calculation. When a model is currently under calculation, the progress of model generation and validation is shown as '% Generating Model' and '% X Val. Calc.'. If an error occurred during model calculation, the state 'ERROR' is displayed; the kind of error can be viewed in the Error report (Section 8.1.1.9). If the cross validation could not be calculated due to not enough spectra (< 20) 'Insufficient Spectra Number' is given under **Cross Validation**.

When a model of the state 'Calculated' or 'Loaded' is selected, the corresponding peak selection that was used to generate this model is displayed in the Spectra View. When no model is selected, the Spectra View shows the current peak selection that will be used in a subsequent model generation process. You can deselect all models by clicking in the background of the Model List View.

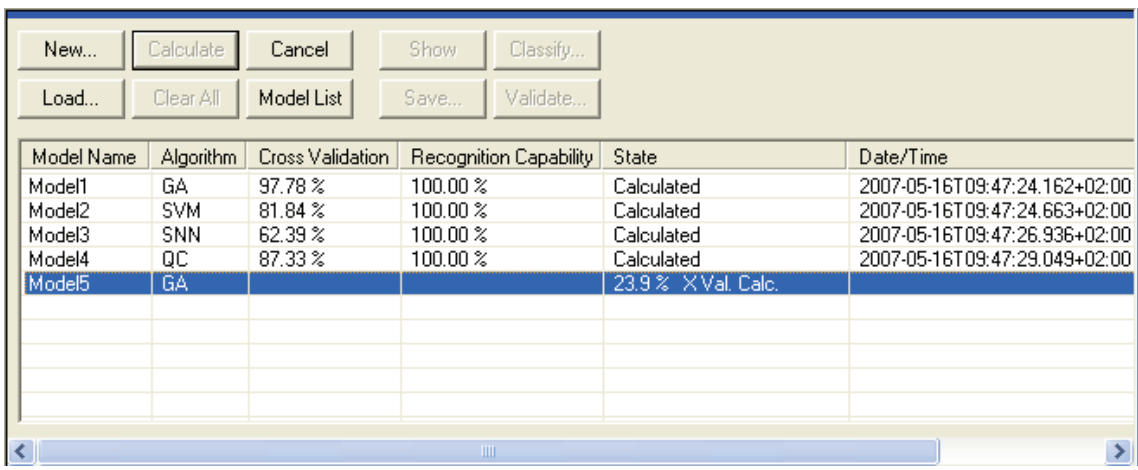


Figure 5-8 Model List View with five models; one is just under calculation

5.1.5 Toolbars

ClinProTools offers the **General** and **View** toolbars (Figure 5-9) with buttons for quick mouse access to certain tools. Each button is supplied with a tool tip showing a short description of its function. Both toolbars are docked below the menu bar by default. You can hide a toolbar and show it again using the **General Toolbar** resp. **View Toolbar** command from the **View** menu. A toolbar can be undocked by double-clicking its slider and docked again by double-clicking its title bar. It can be moved by positioning the mouse cursor in the slider/title bar and dragging the bar with the left mouse button held down to the desired position.

Figure 5-9 General toolbar (left) and View toolbar (right)

5.1.6 Status Bar

The status bar (Figure 5-10) is docked at the bottom of the ClinProTools window. You can hide the status bar and show it again using the **Status Bar** command from the **View** menu.

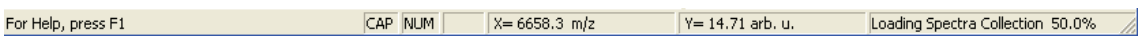


Figure 5-10 Status bar

In the left hand corner, a short help text is displayed. This corresponds to the cursor position and given actions. The next boxes show if the caps lock ("CAP") and the alphanumeric function ("NUM") are activated. In the right hand corner, the progress of a running process is shown.

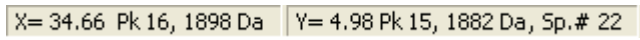
The boxes 'X=' and 'Y=' display the current x- and y-positions of the cursor if the cursor is in a data plotting view and **Coordinates** command for the respective window is active. The data shown depends on the view the cursor is located in, the processing state (before/after peak picking) and whether the cursor is on a peak or between peaks. For the Stack View, no cursor coordinates are displayed.

Spectra View: 

'X' shows the m/z value and 'Y' the intensity value. If the cursor is on a peak, 'Pk' shows the peak number. If the cursor is also on a data point, 'Sp' shows the number of the corresponding spectrum.

Gel View: 

'X' shows the m/z value and 'Y' the spectrum number.

2D Peak Distribution View: 

'X' and 'Y' show the peak area/intensity values of the two peaks selected for display, together with the peak's number and m/z value. If the cursor is on a data point, 'Sp' shows the number of the corresponding spectrum.

ROC Curve View: 

'X' shows the 1-Specificity value and 'Y' the Sensitivity value or vice versa depending on the current definition.

Single Peak Variance View: 

'X' shows the class number and 'Y' the value of the current statistical data. If the cursor is on a data point, 'Sp' shows the number of the corresponding spectrum.

5.1.7 Altering the ClinProTools Data Plotting Views

You can alter the display of the data plotting views (Spectra, Gel, Stack, 2D Peak Distribution, ROC Curve and Single Peak Variance Views) in various ways to adapt the views to your needs.

5.1.7.1 Customizing the Display

The display of a data plotting view can be customized using the commands from the context menu of the view's display region or axes. A changed setting applies to the selected view only.

- To show/hide the cursor coordinates for a view in the status bar activate/deactivate the **Coordinates** command. This does not apply to the Stack View.
- To show/hide the grid in a view, activate/deactivate the **Grid** command. This does not apply to the Stack View.
- To change the background color of the display region of a view, select the **Background Color** command and choose a new color. This does not apply to the Gel View.
- To change the background color of axes of a view, select the **Background Color** command and choose a new color.
- To change the axis font of a view, select the **Axis Font** command and choose a new font.
- To show/hide the scale of the x- or y-axis of a view, select the **Show/Hide X-Axis** or **Show/Hide Y-Axis** command, respectively.

5.1.7.2 Changing the Display Range

The display range of a certain data plotting view can be as follows:

Slave-master behavior of x-axes of Spectra View and Gel/Stack View

The x-axis of the Spectra View and the x-axis of the Gel/Stack View show a slave-master behavior by default. When the scaling of the x-axis in the Gel/Stack View is changed, the x-axis in the Spectra View is always adjusted accordingly. Depending on whether the mouse or a scaling command is used, the adjustment occurs automatically on releasing the mouse button or on the next right-click into the Gel/Stack View. Contrarily, the x-axis of the Gel/Stack View is kept when the x-axis in the Spectra View is changed, but you can force the Gel/Stack View's x-axis to follow the x-axis of the Spectra View by enabling the **Gel/Stack View > Follow Spectra View Mass Range** command from the **View** menu.

Auto-scaling of y-axis of Spectra View or Single Peak Variance View

The y-axis of the Spectra View or the Single Peak Variance View can be set to auto-scaling using the **Auto Scaling** command from the view's context menu. When active, the axis scaling is automatically adjusted to fully display the most intense peak in the current mass range (Spectra View) resp. the maximum statistic value of the current peak in the loaded classes (Single Peak Variance View).

Zooming

To zoom in an area of the Spectra, Gel, 2D Peak Distribution or Single Peak Variance View activate the **Zooming** command in the view's context menu and move the mouse cursor in the view to display the zoom cursor  . To select the desired area position the zoom cursor at the desired start point and drag it with the left mouse button held down to the desired end point. On releasing the mouse button, the enclosed area is zoomed in.

You can also use the mouse wheel to zoom on the axes and in the Spectra View around the position of the mouse cursor (to 65% or 150%, respectively, depending on the direction).

Expanding, contracting and displacing axes

The scaling of the x-axis and y-axis of a view can be changed using the mouse. The scaling cursor is displayed when the mouse cursor is positioned on/below the x-axis () or on/left to the y-axis (). To expand or contract an axis, drag the scaling cursor with the right mouse button held down (right/upwards to expand, left/downwards to contract). To displace an axis, drag the scaling cursor with the left mouse button held down in the desired direction. Alternatively, you can use the various **Scaling** commands available in the view's context menu.

You can also use the mouse wheel to displace an axis when the Shift key or the Ctrl key, respectively, is simultaneously held down. Using the Shift key displaces the axis by 15%, using the Ctrl key shifts the axis by 90% of window extent.

Undoing/Redoing display range changes

ClinProTools stacks the zooming operations for each view separately. You can use the **Undo Zoom** and **Redo Zoom** commands from the **View** menu to undo/redo the last done/undone display range change in the focused view. You can reset the Spectra, Gel, 2D Peak Distribution or Single Peak Variance View to full display of data by double-clicking in the view. If you want to reset only one axis, you can double-click it.

5.1.7.3 Changing the Stack View's Orientation

The orientation of the 3D Stack View is 30° with the base yielding approx. one third of the window height by default. You can quickly change the plot by combined dragging of axes with the mouse. For this, position the mouse in the Stack View to display the 3D cursor  and press the left mouse button. This results in the current plot axes being displayed with bold black lines (Figure 5-11). When dragging the mouse, the orientation of the bold lines changes accordingly indicating the current positioning of three axes. On releasing the mouse button, the Stack View is updated immediately with redrawing all spectra. Moving axes also allows changing the 3D into a 2D plot with displaying all spectra in list view (Figure 5-12).

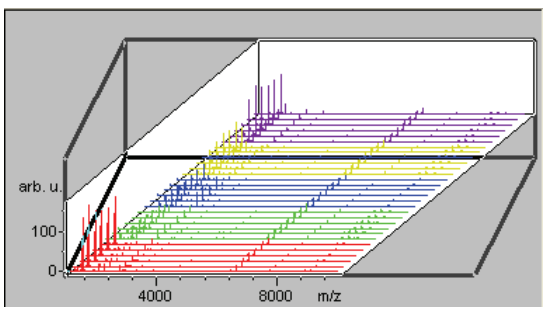


Figure 5-11 Stack View during dragging axes: bold black lines show the current positions of axes

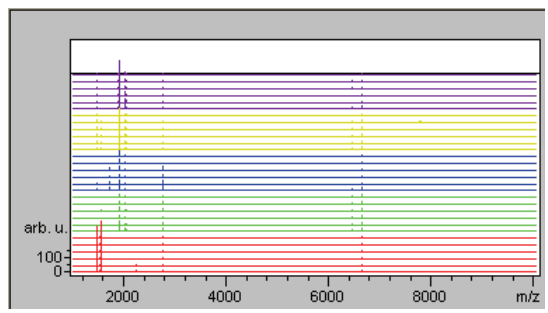


Figure 5-12 List view of spectra after changing the 3D into a 2D plot

5.1.7.4 Resetting the Data Plotting Views

Using the **Reset View Settings** command from the **View** menu allows a combined resetting of certain settings for the data plotting views. Some resets apply to all views (e.g. state of grid and auto-scaling, background color of axes, axis font, data formats for copying graphics), others to only a certain one (e.g. gray scale in Gel View, orientation of Stack View). The complete list of affected settings is given with the description of the command.

5.2 MATLAB Based Windows

Results obtained by the external MATLAB tool integrated in ClinProTools are presented in the MATLAB based PCA windows and Dendrogram window.

5.2.1 PCA Windows

The PCA windows display data of a PCA (Section 6.4.2.3). The windows originate from the external MATLAB software tool integrated in ClinProTools. The PCA main window opens automatically after the PCA is completed. Single Scores or Loadings plot windows, the Influence window and the Variance window can be shown on demand. A once opened PCA window stays open as long as you do not close it or the whole ClinProTools session.

5.2.1.1 PCA Main Window

The PCA main window (Figure 5-13) displays the results of a PCA run performed on the loaded spectra data set(s) (Section 7.5.2.1). It contains eight 3D and 2D plots, four Scores plots (top row) and four Loadings plots (bottom row), showing the data of three selected principle components (PC) (Section 7.5.2.1). The black crosses in the Loadings plots mark the zero axes. Multiple PCA main windows can be open at a time, each representing the results of another PCA run.

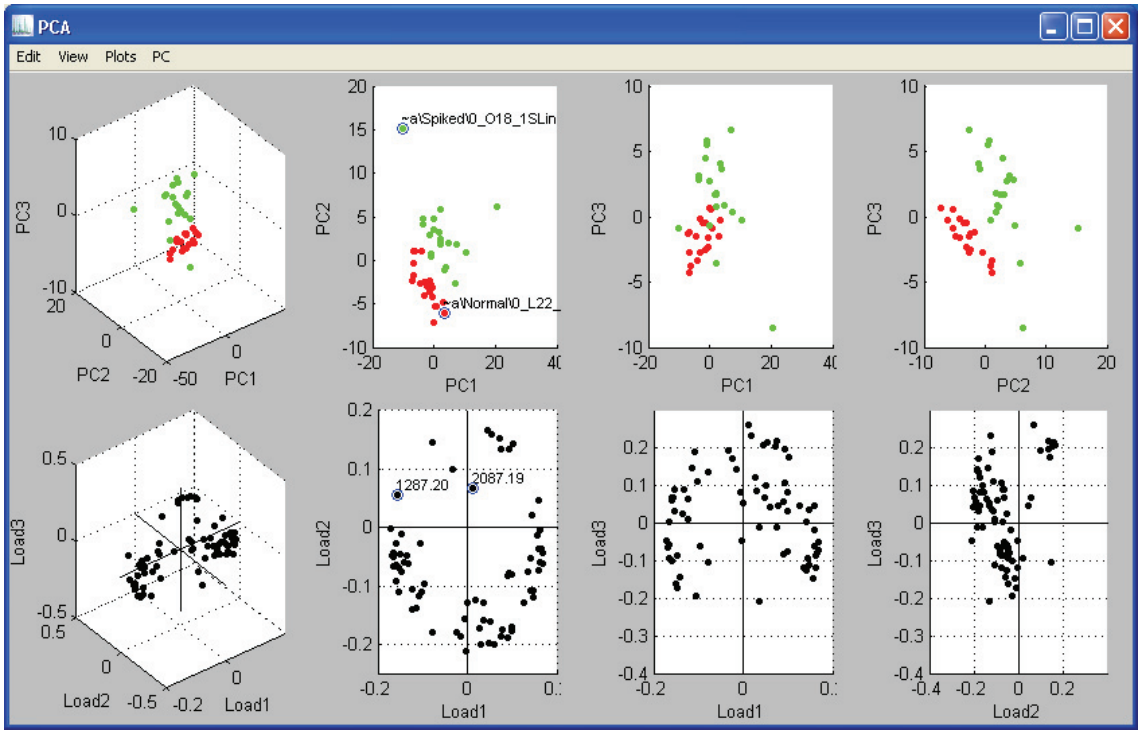


Figure 5-13 PCA main window showing PCA results for two loaded classes in the Scores plots (top row) and Loadings plots (bottom row)

5.2.1.2 Single Scores Plot / Loadings Plot Window

Each Scores plot or Loadings plot of the PCA main window can be displayed in a separate window (Figure 5-14) using the corresponding command in the PCA main window's **Plots** menu.

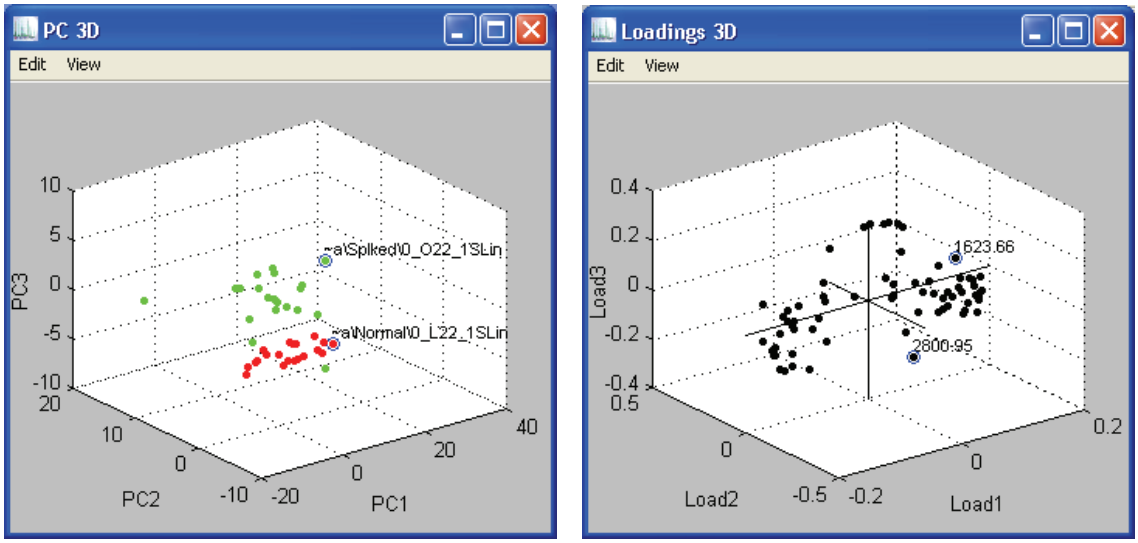


Figure 5-14 3D Scores plot window (left) and 3D Loadings plot window (right)

5.2.1.3 Influence Window

The Influence window (Figure 5-15) displays the Influence plot of the current PCA with respect to the chosen PC number. You can open the window via the **Influence** command from the PCA main window's **Plots** menu.

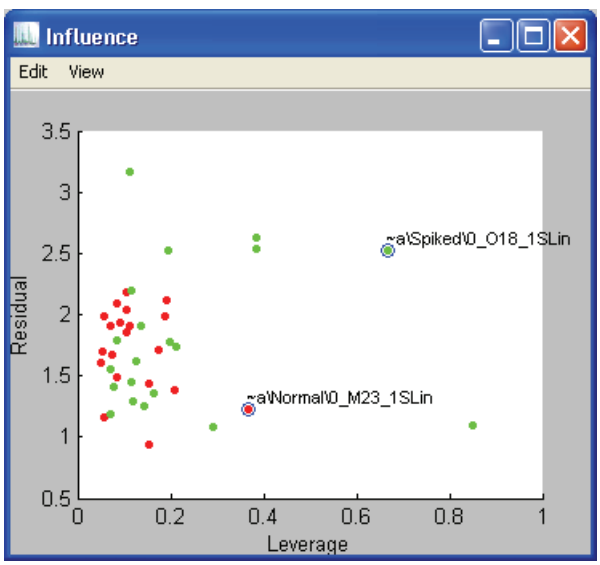


Figure 5-15 Influence window showing an Influence plot

5.2.1.4 Variance Window

The Variance window (Figure 5-16) displays the variance plot of the current PCA (Section 7.5.2.3). You can open the window via the **Variance** command from the PCA main window's **Plots** menu.

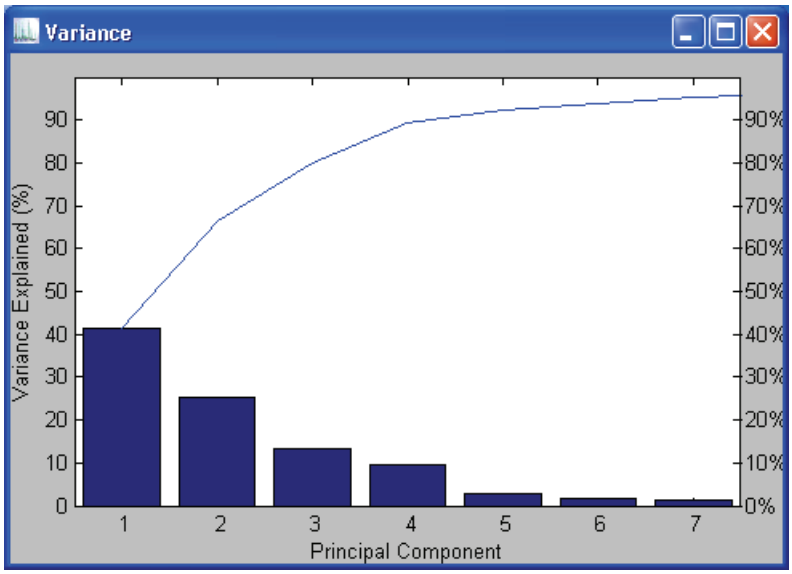


Figure 5-16 Variance window showing a Variance plot

5.2.2 Dendrogram Window

The Dendrogram window (Figure 5-17) displays the result of an unsupervised hierarchical clustering (Section 6.4.2.4) performed on the loaded spectra. It originates from the external MATLAB® software tool integrated in ClinProTools. The Dendrogram window opens automatically after the unsupervised clustering is completed. The display of the dendrogram (e.g. full tree, w/o spectrum paths) depends on the clustering parameters used.

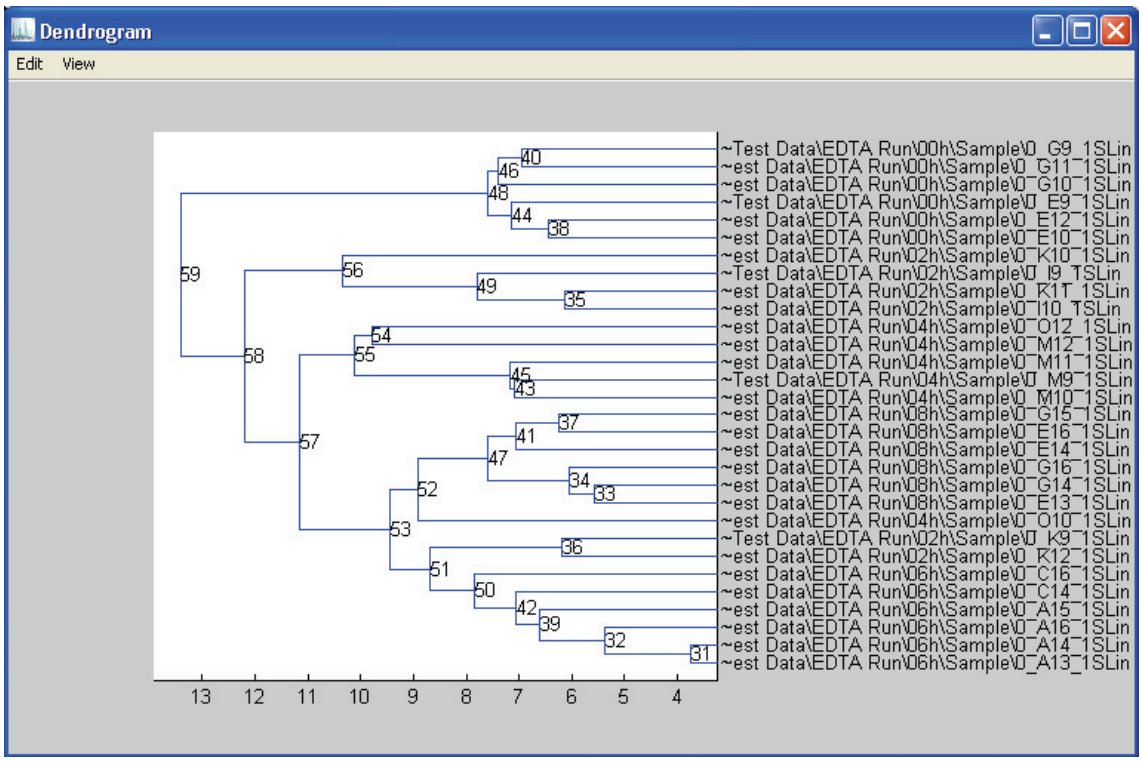


Figure 5-17 Dendrogram window showing a full tree dendrogram for the EDTA Run demo data and corresponding spectra paths

6 BASICS ON DATA PREPARATION, MODEL GENERATION AND SPECTRA CLASSIFICATION WITH CLINPROTOOLS

The following sections provide basic information on data preparation, model generation and validation and spectra classification in ClinProTools. In addition, the used statistical tests and methods as well as certain statistical problems with MS data are described.

6.1 Data Preparation

ClinProTools uses a standard data preparation workflow including spectra pretreatment, peak picking and peak calculation operations. ClinProTools automatically picks the peaks either on the calculated total average spectrum or alternatively on the single spectra and sets up the corresponding peak list. After the automatic picking, it is possible to edit the peaks manually. The result of data preparation is a collection of peak areas resp. maximal intensities for each spectrum. For all spectra, the areas/maximal intensities of the same peaks are calculated so that for each spectrum the same number of peak areas/intensities is obtained.

ClinProTools supports grouping of spectra from multiple measurements. Furthermore, the standard workflow can be supplemented by applying additional filters to modify spectra, reduce data and exclude spectra of lower quality from further processing.

For nearly all data preparation steps there are parameters, which can be chosen to adapt to the kind of spectra used and to control the number of peaks taken into account. The data preparation parameters are set in the **Settings Spectra Preparation** (Section 9.1.4.1) and **Settings Peak Calculation** (Section 9.1.4.2) dialogs.

6.1.1 Standard Data Preparation Workflow

The spectra selected for model generation and classification are treated according to a standard workflow generally including the following steps:

- Baseline subtraction on spectra
- Normalization of spectra
- Recalibration of spectra (optional)
- Average spectra calculation
- Average peak list calculation

- (Peak calculation in the individual spectra)
- Normalization of peak lists for model generation

6.1.1.1 Baseline Subtraction on Spectra

The purpose of the baseline subtraction is to remove the broad structures of a spectrum. If we would not do a baseline correction, the variable level of the baseline, which depends on the preparation, would influence the peak areas quite a lot and would make it difficult to select peaks based on S/N and intensity thresholds. The baseline subtraction is done (1) on the individual spectra to prepare the spectra for the purpose of recalibration and quality checks and (2) again on the average spectrum. The latter baseline correction is done to remove baseline structures, which come up from the averaging of the noise of the individual spectra.

The goal of the baseline algorithms is to remove the broad baseline structures without disturbing the line shape of broad and overlapping peaks too much. For that purpose, ClinProTools offers two algorithms each with a parameter to be able to optimize the baseline correction:

- **Convex Hull Baseline**: This type of baseline algorithm constructs the baseline by fitting multiple parabolas to the spectrum. The baseline is then refined in an iterative way. This is done in a way that the baseline is at least almost always below the spectrum; therefore the name Convex Hull baseline. The **Baseline Flatness** parameter influences baseline construction.
- **Top Hat Baseline**: This type of baseline algorithm constructs the baseline by means of morphology operators. The baseline of the spectrum is obtained in two steps: First, each data point is replaced by the minimum value of the spectrum within *n* data points, which gives the so-called “erosion”. Then within the same number of data points, each value is replaced by the local maximum of the minimal values giving the “opening” of the spectrum, which is the baseline. The number of data points over which the minimum and maximum value is searched for is a function of the mass. The range for the minimum and maximum search can be enlarged with the **% Minimal Baseline Width** parameter. For reference, we refer to J. Serra, "Image Analysis and Mathematical Morphology", Academic Press, New York (1982).

The advantage of the Top Hat baseline is the fact that the tuning parameter is giving a more transparent option to modify the baseline.The advantage of the Convex Hull baseline is that the baseline is a smooth function. Usually, the Convex Hull baseline is preferred for a mass range of up to 10 kDa.

6.1.1.2 Normalization of Spectra

All spectra are normalized to their own TIC (total ion count). Thereby for each spectrum the TIC is determined as the sum of all intensities of the spectrum. Subsequently all intensities of this spectrum are divided by the obtained TIC value. After this procedure all intensities are in the range of [0,1]. For visualization as e.g. in the Spectra View the intensities maybe scaled by some additional factor.

6.1.1.3 Recalibration of Spectra

Usually, a mass spectrum is presented as a plot showing intensity over *m/z* (mass to charge ratio) values. However, the *m/z* values are not obtained directly from a measurement. Instead, these values are computed from the time of flight (TOF), or other raw data, by means of a calibration function. Here, this function is a quadratic mapping between *m/z* and TOF. However, systematic time shifts can be observed for individual measurements, e.g., due to the height profile of the preparation. This finally leads to peak shifts, which can easily be observed in the Gel View. For this purpose, calibration before measurement is necessary. For linear profiling spectra using steel or AnchorChip targets, calibration at one position should be sufficient. The resulting mass deviation over the whole target is mostly < 300 ppm. If Prespotted AnchorChip targets are used, nearest neighbor calibration should be done to prevent higher mass deviation.

In practice, 2000 ppm is an upper estimate for the individual mass error. The task of recalibration is to reduce mass shifts occurred during the measurement. Spectra recalibration is enabled by default but can be turned off.

In order to recalibrate single spectra a list of reference masses is required. Such a list is obtained from the line spectra derived from the original data using peak picking. Only those masses, which occur in at least 30 % of the spectra, are used as reference masses. The recalibration algorithm looks for these reference masses in the peak list of each spectrum. The **ppm Maximal Peak Shift** parameter of the **Settings Spectra Preparation** dialog is used as upper limit of the mass difference between reference mass and peak mass. The calibration function of the spectrum is modified such that the mass error across all assigned pairs is minimized. The number of reference masses depends on the application. For measurements within the mass range 1,000 - 10,000 Da about 40 - 80 reference masses can be expected. Typically 50 % and more of those peaks are used for the recalibration of individual spectra.

As part of the recalibration step, the list of reference masses is generated. After recalibration it is checked by the spectrum quality filter (Section 6.1.3.2) how many of these masses can be found in each spectrum. All spectra with a Spectrum Quality Value < Spectrum Quality Threshold are marked as "Not Recalibratable" and can automatically be excluded.

6.1.1.4 Average Spectra Calculation

From the recalibrated preprocessed individual spectra, a total average spectrum is calculated. The spectra are weighted with the reciprocal size of the classes to get an equal representation of classes with a very different number of spectra. Per-class average spectra are calculated also.

In the literature [J. S. Morris, K. R. Coombes, J. Koomen, K. A. Baggerly, R. Kobayashi, “Feature Extraction and Quantification for Mass Spectrometry in Biomedical Applications Using the Mean Spectrum”, Bioinformatics Advance Access, 2005] it has been shown that using the mean spectrum is in most cases favorable against using peak lists obtained from the individual spectra.

6.1.1.5 Average Peak List Calculation

The ClinProTools data analysis workflows rely on peak information determined for each spectrum. An average peak list representing all important peaks is calculated at first. It contains the start and end positions of these peaks. At those peak locations in all single spectra the area or maximal intensity of the peaks is calculated in the next step. These peak lists are used as features to determine statistical information as well as classification models.

In ClinProTools 2.2, two modes for the generation of the average peak list are available.

- The first one is the 'total average spectrum peak picking approach' for the detection of an average peak list like in ClinProTools 2.1 (standard approach). Thereby the peak picking is applied on the calculated total average spectrum. The identified peak regions by means of start- and end-positions are subsequently mapped to all single spectra.

  Due to the averaging of the spectra the signal to noise for the peak picking procedure is improved and peaks which may be overlooked on a single spectrum, due to noise artefacts, can be easier detected on the total average spectrum. Even small but reproducible peaks will be detectable which would get lost in the noise of single spectra. While this approach is promising and works quite well in case of e.g. two class approaches with similar class sizes, it maybe less appropriate in case of a larger number of classes or in case of classes which are very imbalanced. This again, is due to the averaging property which may not only reduce noise artefacts but may also delete small and rare peaks.
- To overcome this effect an alternative method for the detection of an 'overall average peak list', the 'single spectra peak picking approach', is supported in ClinProTools 2.2. Very briefly, this approach combines multiple peak lists which are obtained by peak pickings on each individual spectrum or, in case of multiple spots, on the averages of them.

6.1.1.5.1 Peak Picking on the Total Average Spectrum

In the total average spectrum peak picking approach (standard approach) an average peak list is generated by picking peaks on the calculated total average spectrum.

The automatic detection of peaks is based on the analysis of a smoothed first derivative. The smoothing is determined by the **Resolution** parameter in the **Settings Spectra Preparation** dialog. Lower resolution values cause stronger smoothing. Then, the zero-crossings of the smoothed derivative are used to identify peaks. Internally an iterative procedure using different resolution values is applied to identify unresolved shoulder-peaks. If the **Resolution** parameter is chosen too large, more and more artificial peaks (spikes) will be found. On the other hand, smaller **Resolution** values will remove more and more unresolved (shoulder) peaks from the peak list. Start and end positions of the peaks are determined by relative slope thresholds of the derivative along the trailing edges of the peak.

In the 'total average spectrum peak picking' approach (standard approach) an average peak list is set up by picking peaks on the calculated total average spectrum. The number of peaks picked on the total average spectrum, and thus the average peak list, can be reduced by applying the **Signal to Noise Threshold** or by giving the **Maximal Peak Number** found in the **Settings Peak Calculation** dialog. Furthermore, peaks which do not exceed a certain percentage of the largest peak (**% Relative Threshold Base Peak**) can be excluded.

The integration regions of the picked peaks can be displayed in the Spectra View (Section 9.1.3.6.6).

6.1.1.5.2 Peak Picking on the Single Spectra

In the single spectra peak picking approach an overall average peak list is calculated over all spectra of all classes by an automatic combination of multiple peak lists. Thereby the peak lists are determined by application of a peak picking procedure for each sample. All single peak lists are merged together such that a large list of peaks is obtained, which in principle includes duplicates that are only distinct by a small error or mass shift.

The procedure starts with the average peak list obtained in the standard approach. This average peak list is used to screen out peaks which are already very common. These peaks immediately become part of the final overall average peak list. On the remaining set of peaks a clustering is applied (Martinez et al, 1993) which is combined with the approach published in DeSieno (1988) to obtain a clustering of the peaks. Thereby the number of clusters is determined in accordance to an overestimate of the number of peaks in the given list of peaks. The clustering forces peaks to be in a cluster which are very close to each other by means of a small mass shift given in ppm. After some postprocessing steps, the clustering is converted into a peak list and combined with the initially obtained average peak list. Thereby, now multiple peaks are

mapped onto a single cluster or peak position. Peaks which are very rare in the set of spectra, say, with a presence of less than 10% can be omitted. In that way an overall peak list is obtained, which contains peaks which show a nearly overall presence but also rare peaks can be detected which may be present in a single class, only.

The obtained overall average peak list is further processed such that overlapping peaks, by means of start/end positions, but not by means of central masses, are made distinct from each other. On the obtained list, peak features such as the area or intensity are calculated. The S/N of a peak is determined as an average of the single S/N values of the peaks which are mapped to this peak location and can be used for subsequent selection procedures.

For details, please refer to:

T.M. Martinez, S.G. Berkovich and K. J. Schulten, "Neural-gas network for vector quantization and its application to time-series prediction", IEEE Transactions on Neural Networks 4: pp 558-569, 1993

D. DeSieno, "Adding a conscience to competitive learning", Proceedings ICNN'88 International Conference on Neural Networks, pp 117-124, 1988

6.1.1.6 Peak Calculation in the Individual Spectra

Calculating peaks in the individual spectra is based on the average peak list picked on the total average spectrum or the overall average peak list picked on the single spectra. Either the peak areas or the maximal peak intensities can be used for peak calculation which is defined in the **Settings Peak Calculation** dialog. Area calculation is applied by default as peak areas have a smaller variation between spectra than intensities of single points have. The area of a peak is obtained by integrating the intensities over the region of the peak according to the selected **Integration Type**. In **Zero Level** integration, the full intensity values are integrated whereas in **End-Point Level** integration only the area above the cutting edge connecting the endpoints is being measured. The two options will yield different areas especially for shoulder peaks. Depending on the classification algorithm used, peak areas may be normalized for being used in model generation (Section 6.1.1.7).

6.1.1.7 Normalization of Peak Lists for Model Generation

With the Genetic Algorithm (Section 6.2.1.1), Support Vector Machine (Section 6.2.1.2) and Supervised Neural Network (Section 6.2.1.3) algorithm the peak lists are normalized before being used in model generation. This is necessary to make different peaks comparable to each other. Otherwise, small peaks would not be treated as equally important.

6.1.2 Spectra Grouping

To increase the measurement quality, the ClinProtRobot supports multiple measurement of the same sample (Section 6.4.3.3). Such a set is called 'spotting' and consists of multiple 'spots'. Spectra belonging to one spotting must be treated different by the software in comparison to independent spectra to avoid errors in the statistical calculation.

To support spectra grouping the **Support Spectra Grouping** option in the **Settings Spectra Preparation** dialog must be enabled. The spectra grouping parser (see below) automatically parses spectra paths and groups the spectra according to their sub folder structure.

Note: This option is suitable only for automatically created spectra by the current ClinProtRobot with the corresponding software. If the option is enabled while using a different folder structure, the parser might by coincidence detect non-existing groups, which will lead to calculation errors.

In each spectra group only one spectrum should be enabled for further processing. This can be done automatically using the **Enable Similarity Selection** option during spectra load, which selects the spectrum most similar to the average of a group. Alternatively, you can manually exclude spectra (Section 7.1.3) leaving one per group. If more than one spectrum per group is processed due to dependent samples the statistic values might be misleading, e.g. in the case of p-values they might be much too low.

Spectra grouping parser

The spectra grouping parser determines the spectra group membership by analyzing the spectra paths names. It works as follows: The list of spectra paths of a spectra collection is processed sequentially. If consecutive paths have the same group folder, they are considered to be belonging to one group. In reverse order, the group folder is the fourth subfolder of the 'fid' file path of a spectrum. (E.g. in the case of ...\F\F1\0_D9\1\1SLin\fid 'F1' is the assumed group folder). The first sub folders name must be one of '1SLin', '1Lin', '1Ref', '1slin', '1lin' or '1ref' ('1SLin' in the example). The second must be a number from 1 to 9 ('1' in the example). The third subfolder must contain an underscore '_' ('0_D9' in the example). If these conditions are not given, the spectrum will not be considered belonging to a group.

The following example will be parsed as two groups F1 and F2 with four spectra each:

...\F\F1\0_D1\1\1SLin\fid
...\F\F1\0_D2\1\1SLin\fid
...\F\F1\0_D3\1\1SLin\fid
...\F\F1\0_D4\1\1SLin\fid
...\F\F2\0_D5\1\1SLin\fid
...\F\F2\0_D6\1\1SLin\fid
...\F\F2\0_D7\1\1SLin\fid
...\F\F2\0_D8\1\1SLin\fid

Note: To make sure that the spectra are parsed correctly verify the spectra grouping performed by the parser. The grouping is displayed in the **Groups** column in the Spectra List report (Section 8.1.1.1). In the Gel View, groups are separated by dashed group separator lines.

6.1.3 Additional Filters

ClinProTools supports additional spectra modifying and selecting filters, which can be applied to modify spectra, reduce data and exclude spectra of lower quality from further processing. Parameters and usage of these filters can be changed in the **Settings Spectra Preparation** dialog. Excluded spectra can be highlighted in the Gel View according to the filter used for exclusion (Section 9.1.3.7.3).

Note: In general, one should be aware that the optional filter process needs additional time during the spectra-loading step but may be very helpful to improve subsequent processing steps.

6.1.3.1 Filters Modifying Spectra

ClinProTools supports various spectra modifying filters, which will be applied during spectra loading if activated, except the recalibration filter that applies to spectra recalibration. The Resolution parameter also applies to peak picking on the average spectrum.

Resolution parameter

The peak detection algorithm needs a hint for the peak width to be able to decide what has to be assumed to be a peak and not just an artifact of a broader peak. Since the peak width in mass units depends on the mass range and on the TOF instruments, resolution is used as a parameter instead (resolution = mass / peak width). Defaults for the resolution are given by choosing the respective mass range from the drop down list in the **Settings Spectra Preparation** dialog. If the Resolution parameter is chosen too large, more and more artificial peaks (spikes) will be found. On the other hand, smaller Resolution values will remove more and more unresolved (shoulder) peaks from the peak list.

Mass range filter

To limit the mass range of the spectra to be analyzed you can specify a minimum and maximum mass. Otherwise, define a mass range that is larger than the experimental mass range, which should be the case if you keep the default values.

Smoothing filter

For data smoothing, the Savitsky Golay algorithm is available. The idea of this algorithm is to calculate polynomials in the neighborhood of each data point to get a smoothing of the data. This can be formulated as

$$\widetilde{y}_i := \sum_{k=-M}^{M} c_k y_{i+k}$$

with coefficients c_k. The parameter *M* in the formula is calculated from the given *m/z* smoothing width, which can be changed within the **Settings Spectra Preparation** dialog. The number of smoothing cycles can also be chosen which gives the option to apply this smoothing filter multiple times. The weights for Savitsky Golay are obtained by considering a least square problem for 2M + 1 nodes.

For details, please refer to M.U.A. Bromba and H. Ziegler, "Analytical Chemistry", 53, pp. 1583-1586 (1981) and to A. Savitzky and M.J.E. Golay, "Analytical Chemistry", 36, pp. 1627-1639 (1964) where also tables of the c_k's can be found.

Data reduction filter

In order to speed up calculations and to reduce the memory consumption, especially for very large data sets, a simple way of data reduction is available. The reduction of data is achieved be replacing every set of "n" consecutive data points by the average of these points. The number "n" is given by the data reduction **Factor** parameter. Typically, the value should be chosen between 1 (no reduction) and 10 (10 fold reducetion). The greater the factor n is chosen the more features will be smoothed out. As a consequence, e.g., shoulder peaks may no longer be resolved. Moreover, lower noise estimates are obtained for reduced data. Best classification results are expected without data reduction.

Note: The data reduction is applied prior to any other data processing and influences all subsequent results.

6.1.3.2 Filters Selecting Spectra

ClinProTools allows the processing of a large number of spectra including multiple measurements of the same sample. Within the set of spectra individual spectra are of different quality regarding noise, chemical artifacts, level of intensity, etc. Therefore and to obtain a faster post processing it is recommended to apply some of the supported quality filters upon the loaded spectra. The filters aim on selecting only 'good' spectra and excluding those of lower quality. Thereby, each filter has its own responsibility e.g. to pass only spectra with a sufficiently small amount of noise.

Note: In general, one should be aware that the optional filter process needs additional time during the spectra-loading step but may be very helpful to improve subsequently processing steps.

The following quality filters will be applied to the spectra with the order listed below if enabled.

Null spectra exclusion filter

In seldom cases it happens, that due to a preparation artifact or some I/O problems, a spectrum contains no data or the intensities are extremely low. In that case, the obtained spectrum cannot be processed in a useful way and should be removed. The null spectra exclusion filter identifies such spectra and removes them from the corresponding class. The spectrum is not further processed and cannot be re-included without deselection of the filter and reloading the class.

Noise spectra exclusion filter

The noise spectra exclusion filter aims on identification of noisy spectra. Due to an inappropriate preparation or measurement distortions spectra with a high amount of noise may be measured. These spectra should be excluded to avoid interfering effects on the further processing. This filter checks a given spectrum in the range of 1 Da – 4 kDa or if the spectrum does not contain this range the check is done on the whole spectrum. It analyses the noise function of the considered spectrum and tries to estimate the objective amount of noise within the spectrum. This estimation is compared to a user-defined **Noise Threshold**. If the estimated noise is too high, the spectrum becomes excluded from the set of spectra. Excluded spectra can manually be re-included (Section 7.1.3), but in general, the decision of the filter should be appropriate by a valid noise threshold setting.

Adduct/Polymer spectra exclusion filter

The samples to measure may contain chemical artifacts like sodium-, potassium-adsorptions or polymer as the most common types. The adduct/polymer spectra exclusion filter identifies and excludes spectra, which show mass shifts corresponding to one or more specified artifacts. Na, Mg, K, PEG and PPG are searched for by default but one can adapt the adduct/polymer settings to the analytical task. ClinProTools offers two exclusion modes, **Strict** and **Less Strict**. The strict exclusion mode aims on exclusion of spectra which show characteristic shifts in the autocorrelation spectrum with respect to the adduct/polymer parameterization. The underlying criterion is strict, which means if such a shift exists and it is not due to randomness the spectrum is excluded. The less strict mode allows spectra with shifts of lower contribution to remain in the spectra set collection. This is determined upon an experimental obtained internal threshold.

Similarity selection filter

With the ClinProt equipment, it is possible to measure a sample more than one time. This is useful since it may happen that a single measurement will fail but within the remaining multiple measurements of the same sample a sufficiently well measurement exists. ClinProTools supports spectra grouping from multiple measurements when the **Support Spectra Grouping** option is set. The user can decide if all valid multiple

measurements should be processed or a selection of one characteristic spectrum per sample should be applied. In the first case, the remaining multiple measurements (after some optional quality filter steps) are averaged to reduce the overall measurement variance. In the second case, the similarity filter will be applied. It aims on the selection of the most characteristic spectrum within a given set of spectra from the same sample.

To obtain a useful selection only those multiple measurements should be processed by the similarity filter, which can be considered as characteristic and similar for the current sample. Therefore, a prefiltering using the noise filter and the adduct filter is recommended. Finally, the similarity filter returns one spectrum per sample using a mathematical similarity measure or the spectrum with the highest intensity if only two spectra remained.

Spectra quality filter

Using the spectra quality filter, both spectra of low quality and spectra that have bad calibration properties, can be detected and excluded. As part of the recalibration step, a list of masses (list of reference masses) which occur very frequently within the entire data set is generated. The number of these masses is called the *Maximum Quality Value*. After recalibration it is checked how many of these masses can be found in each spectrum using the **Maximum Peak Shift** parameter as maximum shift. The number of found masses is called the *Spectrum Quality Value*. The *Spectrum Quality Threshold* is computed as the product

Spectrum Quality Threshold = *Maximum Quality Value* * *% Match to Calibrant Peaks*

All spectra with a *Spectrum Quality Value* < *Spectrum Quality Threshold* are marked as "Not Recalibratable" and excluded by using the **Exclude not Recalibratable Spectra** option. The default value of the **% Match to Calibrant Peaks** parameter is set to 30%.

Some typical values for the *Maximum Quality Value* are:

- studies 1,000 – 10,000 Da: *Maximum Quality Value* = 40..80
- studies 8,000 – 20,000 Da: *Maximum Quality Value* = 10..40
- studies 20,000 – 100,000 Da: *Maximum Quality Value* = 4..10

6.1.4 Manual Peak Editing

ClinProTools supports manual editing. The average peak list can be edited manually by adding or deleting single peaks or changing the integration regions of peaks (Section 7.1.5.2). If the peak number is limited to 0, it is possible to create a list consisting only of manual edited peaks. If manual peak editing is performed after the peak calculation has already been run, a recalculation of the peaks is required.

6.2 Model Generation and Validation

In ClinProTools, 'models' are generated which function as 'classifiers'. The aim of model generation is to describe the spectra of the model generation classes in such a way that new spectra can be classified afterwards. ClinProTools supports four kinds of algorithms for generating classification models.

In general, the classification results can be improved by a meaningful restriction of the number of peaks. This can be done during the data preparation due to a selection based on the signal-to-noise ratio or other criteria (**Limit Peak Number** parameter in the **Settings Peak Calculation** dialog, Section 9.1.4.2). Another possibility is to select the peaks according the **Sort Mode** in the **Settings Peak Selection** dialog (Section 9.1.5.1), which allows using only the peaks with the probably highest class separation capability under a univariate view on the data.

The following sections provide information on ClinProTools' classification algorithms, k-nearest neighbor classification, cross validation and external validation.

6.2.1 Classification Algorithms

ClinProTools supports four kinds of algorithms for generating classification models. All these algorithms are different in their methodology and have advantages and drawbacks. Figure 6-1 illustrates the characteristics of the four classification algorithms.

- **Genetic Algorithm** (GA): This algorithm mimics evolution in nature and is used to select the peak combinations which are most relevant for separation.
- **Support Vector Machine** (SVM): This algorithm is motivated from statistical learning theory and is at first used to determine separation planes between the different data classes. Upon the obtained planes, a peak ranking can be calculated in a second step.
- **Supervised Neural Network** (SNN): This algorithm is a prototype-based classification algorithm. The SNN tries to identify some characteristic spectra for each class, which are named prototypes, and could be somehow considered as prototypical samples of that class.
- **QuickClassifier** (QC): This algorithm is a univariate sorting algorithm. The class averages of the peak areas are stored in the model together with some statistical data like the p-values at certain peak positions. For classification, the peak areas/ intensities are sorted per peak and a weighted average over all peaks is calculated.

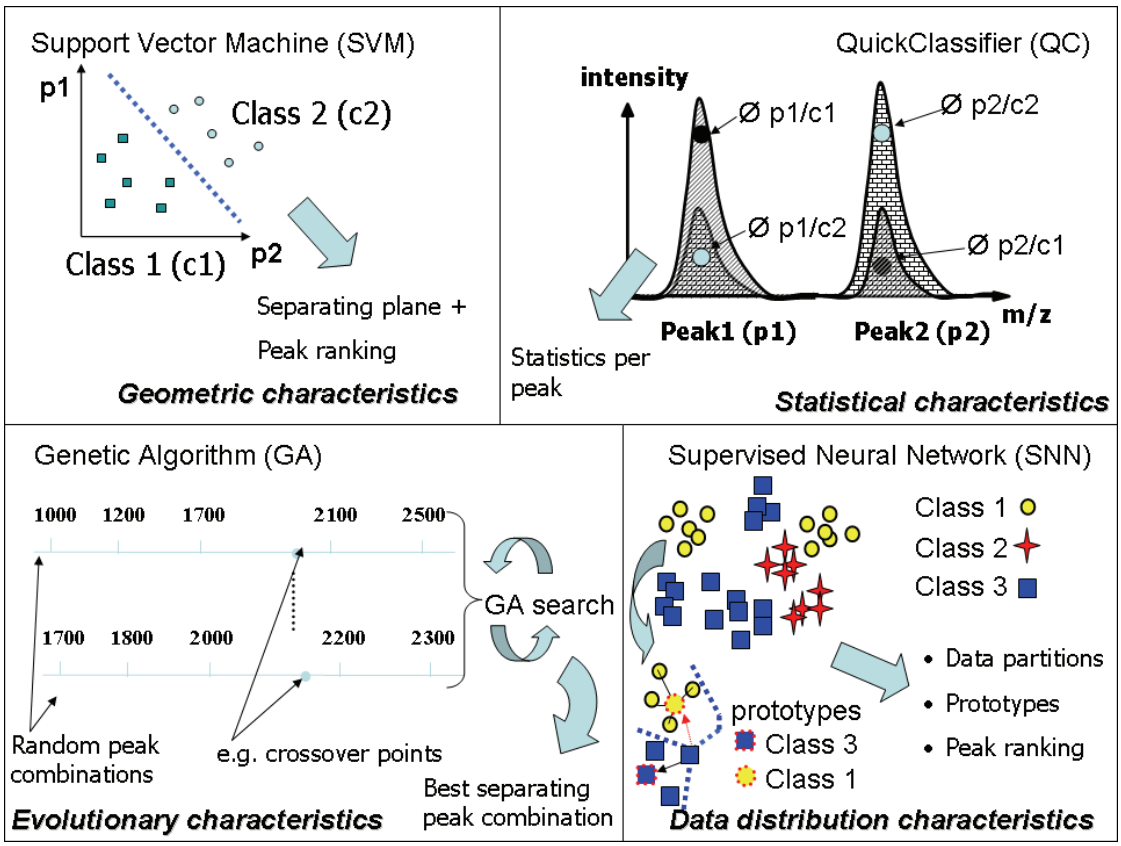


Figure 6-1 Overview of ClinProTools' four classification algorithms

6.2.1.1 Genetic Algorithm

The concept of Genetic Algorithms (GA) was developed by John Holland [J. H. Holland, "Adaptation in Natural and Artificial Systems", University of Michigan Press, Ann Arbor, 1975]. It is based on the idea of evolution in which the fittest individuals have the highest chances of survival. Here, we apply them to select combinations of peaks, which perform best in separating the classes under consideration.

Pattern determination is used to identify an optimal set of peaks, which gives the best separating model determined upon the model generation spectra used and validated on test spectra or by a cross validation procedure. A brute force approach would not work: A systematic trial of all combinations would take far too long because the number of possible combinations is extremely large. For 1000 given peaks and a desired combination of just 3 peaks, you get 1,000*999*998 = 997,002,000 sets of peaks! Therefore, we need more sophisticated ways to do it.

The advantage of the GA is that it needs much less computational time than the brute force approach while still yielding good results. The drawback is that you obtain only a near optimal solution since you cannot guarantee to find the best combination if you do not test all of them.

How the GA works

The GA works on a population, which consists of a multitude of peak combinations. During selection, the fittest peak combinations are chosen and the less capable are abandoned. This is done by optimizing a cost function, which aims on optimal class separation with variance high between classes. Using the cost function each peak combination is rated by an expense factor, which is used as a measure for the fitness. The crossover combines randomly selected pairs of peak combinations to produce child peak combinations, which replace their parent peak combination. The intention here is to combine two fairly good peak combinations to form even better ones. Of course, crossover of peak combinations can also result in less fit combinations, but these will not survive for a very long time. Finally, a small amount of peak combinations is modified randomly during mutation. This is done to keep genetic diversity and to prevent a premature convergence to a local optimum. The expectation is that the average fitness of all peak combinations rises and the best fitness observed will improve.

Parameterization

The basic and advanced parameters for the GA are defined in the **Settings Genetic Algorithm** dialog (Section 9.1.5.2.1).

The basic parameters define the **Maximal Number of Peaks in Model** and the **Maximal Number of Generations** (iterations for the algorithm to run). When using the default value ('50') for the latter most of the time, this value will not be reached as the stop criteria will halt calculation when no better peak combination is found for a number of iterations. For K-nearest neighbors classification (Section 6.2.2) the **Number of Neighbors** can be set to default odd values.

The **Advanced** parameters define how the initial number of peak combinations within a population is determined, either by an **Automatic Detection** mode or by specifying the **Number of Peak Combinations**. The **Mutation Rate**, which is the likelihood of a mutation, can be set to values ranging from 0.0 (no mutation occurs) to 1.0 (all peak combinations are mutated in each generation). The **Crossover Rate**, which is the likelihood of a crossover between peak combinations, can be set to values ranging from 0.0 (no crossovers) to 1.0 (all peak combinations in each generation are used in crossover and are replaced by their children). Since the GA employs random numbers for selection, crossover and mutation, it is possible and quite likely that different values for most of the parameters (especially for **Crossover Rate** and **Mutation Rate**) may yield different solutions. To make comparisons between peak combinations possible, this randomness can be made the same for all peak combinations to be generated by applying **Use Varying Random Seed**. This seeds the random number generator with a different value each time, so every model is using different random numbers. If this

option is not checked, the GA uses the same number for initializing the random seed in any peak combination to be calculated. This way, randomness is disabled and it is easier to study the effect of algorithm parameters.

Classification result

The result of the GA is the peak combination which is proved to separate best between the different classes.

6.2.1.2 Support Vector Machine Algorithm

The concept of the Support Vector Machines (SVM) was developed by Vladimir Vapnik [V. Vapnik, "Statistical Learning Theory", Wiley and Sons, New York, 1998] and is based on the principle of structural risk minimization (SRM). The aim of SRM is to minimize an upper bound on the expected risk over each of the hypothesis classes of the considered problem. In our case, we have a classification problem with an expected risk of misclassifications. We now are interested on a well-modeled classifier with minimal risk. For the SVM this is formalized in an optimization problem, which can be solved using sophisticated mathematical approaches. In the simplest case, the SVM helps to determine an optimal hyperplane separating two clouds of data (Figure 6-2).

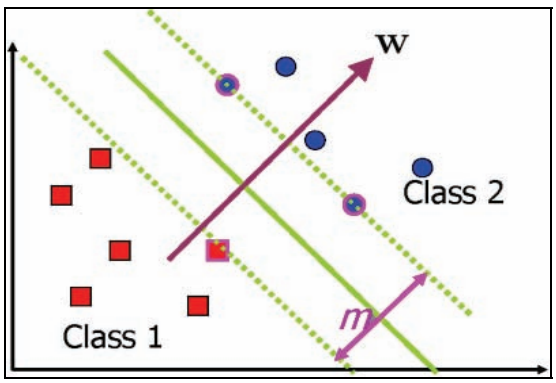


Figure 6-2 Determination of the optimal separating hyperplane; the solid line represents the obtained optimal line (or with more than two peaks hyperplane)

The advantage of the SVM is that it is quite fast in the determination of the peak ranking and a well, formal established pattern recognition tool, yielding good results. The drawback is that it calculates its solution including all peaks as a whole and is by principle a two-class approach, the multi-class solution is obtained by a wrapper method, which may lead only a near optimal solution since you cannot guarantee to find the best combination if you do not test all of them.

If we have classification problems with a large number of peaks, we want to know which peaks separate best. Therefore, the solution obtained from the SVM can be ana-

lyzed in more detail to get a ranking of the contributing peaks. Peaks which have good separation properties are more important for the SVM solution than peaks which do not separate well.

How the SVM works

If we have more than two classes we split the data into a class containing all data points from the current considered class and a rest class which contains the remaining data points. A penalty term 'C' is determined automatically from the data to limit the structural risk of misclassifications and the formal optimization problem is defined. The optimization problem is solved using a quadratic problem solver. A peak ranking is derived from the obtained hyperplane solution. The procedure is iterated until for each class a classifier (class vs rest) is obtained. Upon the obtained SVM model the best number of peaks is determined (if not manual given) by a clustering in the subspace taken from the k best peaks and the (best) solution is stored as the final model.

Parameterization

The parameters for the SVM are defined in the **Settings Support Vector Machine** dialog (Section 9.1.5.2.2). The detection mode for determining the best number of peaks to be integrated in the model has to be defined; you can apply the **Automatic Detection (1-25 Peaks)** mode or specify a **Number of Peaks** (Section 6.2.1.5). For k-nearest neighbors classification (Section 6.2.2) the **Number of Neighbors** can be set to default odd values.

Classification result

The result of the SVM is the peak combination which is proved to separate best between the different classes.

6.2.1.3 Supervised Neural Network Algorithm

The Supervised Neural Network algorithm (SNN) is a prototype-based classification algorithm. If one considers a set of spectra divided into e.g. two classes (cancer, control) the SNN tries to identify some characteristic spectra for each class. These spectra are named prototypes and could be somehow considered as prototypical samples of that class e.g. the prototypical cancer patient from a proteomic point of view. The determination of these prototypes is a complicated task because the classification model is solely based on these prototypes; nevertheless, it should have good generalization abilities for unknown data in an external validation. To fit this needs the SNN allows for metric adaptation which is useful in the search for biomarker candidates, integrates neighborhood cooperation, which typically leads to a better generalization in external validations and is a margin optimizer, which similar to the Support Vector Machine is well founded on mathematical theory.

The SNN, based on the ideas from Kohonen's Learning Vector Quantizers, is a modified version of the Supervised Relevance Neural Gas algorithm. It was developed by

Barbara Hammer, Marc Strickert and Thomas Villmann [B. Hammer, M. Strickert and T. Villmann "Supervised Neural Gas with General Similarity Measure", Neural Processing Letters 21 (1), 21-44 (2005)] and is based on the principle of margin optimization.

As a simple example, a checkerboard data set that consists of two classes, blue and green, with multiple clusters will be considered. These two classes exist in a two-dimensional data space, which is originated by the strange case that the spectra have only two peaks. In Figure 6-3 the first dimension x may be created by peak areas from the first peak and the second dimension y by peak areas from the second peak. Each point in the figure represents a spectrum.

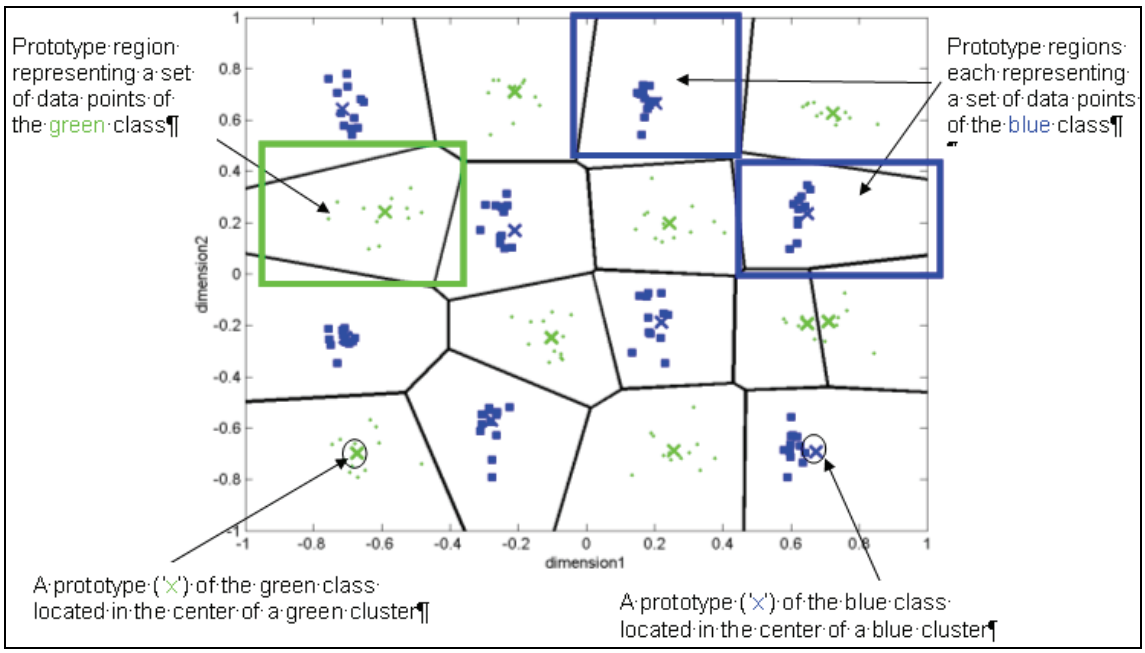


Figure 6-3 Supervised Neural Network algorithm: Determination of class prototypes

The SNN has to learn the characteristic of the two classes in a way that new data points can be classified to one of the two classes. To do this the SNN tries to determine a set of prototypes depicted as 'x' (green class) and 'x' (blue class). Multiple prototypes for one class are typical. Thereby the distance between separating boundaries (the polygons in the figure) of consecutive classes (e.g. green / blue in the figure) should be as large as possible (for the SNN this is formalized in an optimization problem, solved by a gradient descent on a cost function similar as within a neural network approach). The prototypes ('x' and 'x') represent a set of data points from the original data set. This means that all data points which are closer to a prototype I (e.g. from the prototype set of the green class) than to a prototype II belong to prototype I. Thereby, it is claimed that this prototype I is characteristic for these data points. If the prototype I is

of type ‘green’, it is assumed that all of its data points belong to the green class, too. The region, which encloses such a set of data points, is depicted by a polygon.

In that way, the position of the prototypes in the data space of two peaks induces a partitioning giving the spider net in the example. Later classifications are only done upon these prototypes and hence it is important that they are well located in the data space.

The advantage of the SNN is that it determines local classifier models (there can be different regions of prototypes with the same class label) and hence it shows good performance for very multimodal data. In addition, it can determine a peak ranking. It naturally deals with multiple classes. The drawback is that it aims on empirical risk minimization (ERM) which may stick in local minima. This is reduced by neighborhood cooperation.

How the SNN works

In an initial step, a predefined number of prototypes are spread over the data space by use of the Batch-Neural-Gas algorithm, which gives an optimal distribution of the prototypes over the data space in accordance to the data density properties. In a second step, the SNN optimizes the positions of the prototypes with respect to the class information (supervised) minimizing the empirical risk. Thereby the used metric of the data space is adapted such that dimensions, which are relevant for the class separation, are higher weighted than dimensions, which do not contribute to class separation. This procedure of optimizing prototype positions with a combined feature selection is applied iteratively for a predefined upper limit of steps. The algorithm may stop earlier if some convergence criteria are reached.

Parameterization

The parameters for the SNN are defined in the **Settings Supervised Neural Network** dialog (Section 9.1.5.2.3).

The SNN automatically uses the automatic detection mode to determine the best number of peaks to be integrated in the model (Section 6.2.1.5).

The **Advanced** parameters define the prototype determination. The **Upper Limit of Cycles** can be set. This number should be chosen with respect to the complexity of the data and can be evaluated considering the views, the number of picked peaks and the statistics. For complex data sets, the SNN may need longer runtime to get good results. Typically, the value can be taken with defaults. The user setting k = [1-99] is multiplied internally by 100, hence the number of cycles (processing the whole model generation data for one time) is given as k * 100. Typically at least 1000 cycles should be calculated by the algorithm. The **Number of Prototypes** should be chosen with respect to the number of expected sub clusters in the data and the overall data complexity. The default is suitable in general but an increase of prototypes may sometimes improve the model performance, but may also lead to over-fitting if too many prototypes are used. Internally at least one prototype for each class is used.

Classification result

The final model consists of the final prototypes and the learned metric, which can be interpreted as a weighting of input dimensions. The classification takes place using only the prototypes in a nearest neighbor approach. The weighting is subject of change with respect to correlated peaks. This means that if e.g. two peaks have similar importance only one peak will be ranked high and the other peak may vanish. For interpretation, only high weighting values should be analyzed. A low ranking value is – no – indication that the peak is unimportant, but a high ranked peak is probably important.

6.2.1.4 QuickClassifier Algorithm

The QuickClassifier algorithm (QC) is a univariate sorting algorithm. The class averages of the peak areas are stored in the model together with some statistical data like the p-values (Section 6.4.1.6) at certain peak positions. For classification, the peak areas are sorted per peak and a weighted average over all peaks is calculated.

The QC algorithm has a good performance. The univariate approach makes it easy to trace back classification results. The classification allows not only determining the class membership but calculates also a likeliness for each class. If there are only few samples available for the model generation, the validity of the classification seems to be better in comparison to other algorithms in many cases.

How the QC works

At first for each peak position, the class averages of the peak areas are calculated. These averages are stored in the model together with the weights determined from the statistical tests. For classification at each peak position the reciprocal difference of the peak area and the class averages are calculated and normalized. In the next step, over all peak positions from these values weighted averages for the classes are calculated. To determine the class member ship these weighted averages are compared.

Parameterization

The parameters for the QC are defined in the **Settings QuickClassifier** dialog (Section 9.1.5.2.4). The QC automatically uses the automatic detection mode to determine the best number of peaks to be integrated in the model (Section 6.2.1.5). The **Sort Mode** defines the peak ranking as well as the weights used for averaging. In the case of **Difference Average** the difference between the maximal and minimal average area of all classes is used as the weight, in the case of the **P-values** the logarithm of the p-value is used. Several models containing up to the first 25 peaks of the ranking are compared internally to determine the optimal peak number.

Classification result

Apart from the calculated class membership, the classification result contains a likeliness for each class. It is derived from the weighted average and normalized to 1. If the value is 1 for all classes, all classes are equally likely. If the value is below 1, the class

is less likely, if it is higher, it is more likely. The class with the highest likeliness is the predicted one.

6.2.1.5 Detection Modes to Determine the Best Number of Peaks in a Model

Note: The peak rankings in the multivariate algorithms (GA/SVM/SNN) are derived from an analysis in a high-dimensional data space performed on all peaks passed to the algorithm. These rankings may underestimate the individual importance of single peaks as available by use of univariate rankings obtained from statistical tests (t-test, ANOVA ...); in some cases, they might differ significantly. Especially if the number of peaks is limited or fixed, which is an option in GA and SVM, it might happen, that the algorithms choose peaks with a suboptimal univariate classification capability.

If a classification based on a univariate peak ranking is desired, it is advisable to pre-sort the peaks with the Peak Selection available in the Settings Peak Selection dialog in the Model Generation menu. E.g., the peaks used for the algorithms can be reduced to the three best peaks according t-test. In this way, the algorithms are forced to use the best peaks according to a univariate ranking.

ClinProTools offers an automatic and a manual detection mode for determining the best number of peaks to be integrated in a model. The automatic mode (option **Automatic Detection (1-25 Peaks)**) automatically determines the best number of peaks to be integrated in the model with restricting the number of peaks 1 to 25 peaks. For the manual mode, the **Number of Peaks** to be taken has to be specified.

When the automatic detection mode is active, you do not need to manually determine the peak number by iterating the model generation with different settings for the best number of peaks. The algorithm does this internally by an automatic iteration. To have reliable processing times the search for the number of best peaks is restricted to maximal 25 peaks in a model. Therefore, the automatic peak detection will always create models with 1 to 25 peaks. Due to this restriction, it could happen that a manually created model with a larger number of peaks or with all peaks included may give better results than a model obtained by automatic detection. As a second point the automatic detection incorporates no cross validation, hence the best number of peaks is determined on the recognition capability only. Therefore, the obtained model may show over fitting effects.

The detection mode(s) that can be applied depend(s) on the classification algorithm. The SVM supports both the automatic and the manual mode; the QC and the SNN in principle use the automatic mode whereas the GA always works in manual mode.

6.2.2 K-Nearest Neighbor Classification

The *k*-nearest neighbor (k-NN) classifier algorithm is used within the GA and SVM to obtain the final classification. It just uses the distances between points in the n-dimensional space. Remember that each point corresponds to a spectrum. The coordinates of the point are made up of the peak areas of the spectrum. The peak selection is derived from the current GA peak combination or the final SVM peak ranking solution. The idea of k-NN classifiers is to look at the k-nearest neighbors and their spectra class membership.

For details on numerical analysis of the k-NN principle, we refer to T. Hastie, R. Tibshirani and J. Friedman, "The Elements of Statistical Learning", Springer, (2002).

Workflow

The workflow of the k-NN classification is as follows:

1. The distances between all data points (spectra) are calculated.
2. The k-nearest neighbors for each point are determined.
3. Each point is classified according to the class membership of the neighboring points (Figure 6-4).
4. The separation value is calculated which indicates how good the data could be separated and classified with the current parameter of k-NN and the given peak selection.

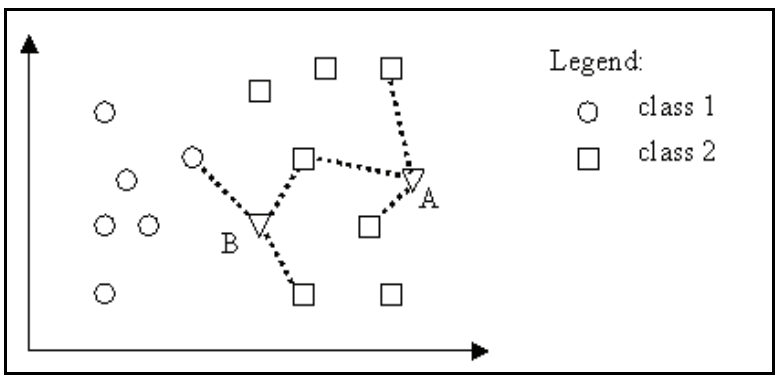


Figure 6-4 Classification of points A and B using three nearest neighbors

Parameterization

The **Number of Neighbors** parameter in the **Settings Genetic Algorithm** / **Support Vector Machine** dialogs (Sections 9.1.5.2.1 and 9.1.5.2.2) determines the ‘k’, i.e. how many neighbors have to be used in comparisons. Per default, ‘k’ can be set only to the odd values '1', '3', '5' and '7' which has been found to perform reasonable well on different data sets. The odd value ensures that in general a classification is obtained using k-NN (unclassified may still happen for e.g. three classes and k = 3, where two

neighbors belong to different classes) and that the solution is sufficiently stable. The case of one neighbor (k = 1) should be used if the number of samples is very small. For a larger number of samples per class k > 1 is recommended.

Classification result

The result of the k-NN classification is a neighborhood matrix in a reduced space of the given peak selection and a classification result for each spectrum which has to be classified.

6.2.3 Cross Validation

Cross validation is a measure for the reliability of a calculated model and can be used to predict how a model will behave in the future. It is a method for evaluating the performance of a classifier for a given data set and under a given parameterization. Different methods for cross validation have been proposed. The generic principle behind cross validation methods is to split (automatically) a given set of data into a model generation set and a test set. The model generation set is used to determine a model by use of the chosen classifier. The test set is than used to evaluate the obtained model and to determine the prediction capability. This procedure is repeated multiple times and the absolute prediction capabilities are accumulated and finally normalized to a relative prediction capability obtained by the cross validation procedure. The kind of splitting into test and model generation set and how the iteration is performed depends on the specific cross validation method.

It should be noted that a cross validation for very small sample sizes (even if only in one class) is not very useful and may give unusual results. Thus, within ClinProTools the cross validation is calculated only if at least 20 not excluded spectra over all groups are available. This must also be kept in mind when working with groups of spectra from multiple measurements (Section 6.1.2); here at least 20 groups must be available. This is since each group is averaged if the similarity selection filter (Section 6.1.3.2) is not on and if the similarity selection filter is on also only one spectrum per group is taken.

For detailed information on cross validation, we refer to M. J. Kearns, Y. Mansur, A.Y. Ng and D. Ron, "An experimental and theoretical comparison of model selection methods", Machine Learning, 27, 7-50, 1997.

Parameterization

ClinProTools supports three different kinds of cross validation that can be set in the **Settings Cross Validation** dialog (Section 9.1.5.7). With activated cross validation, after each model generation a final cross validation is applied. The three modes are illustrated in Figure 6-5 and described in the folowing.

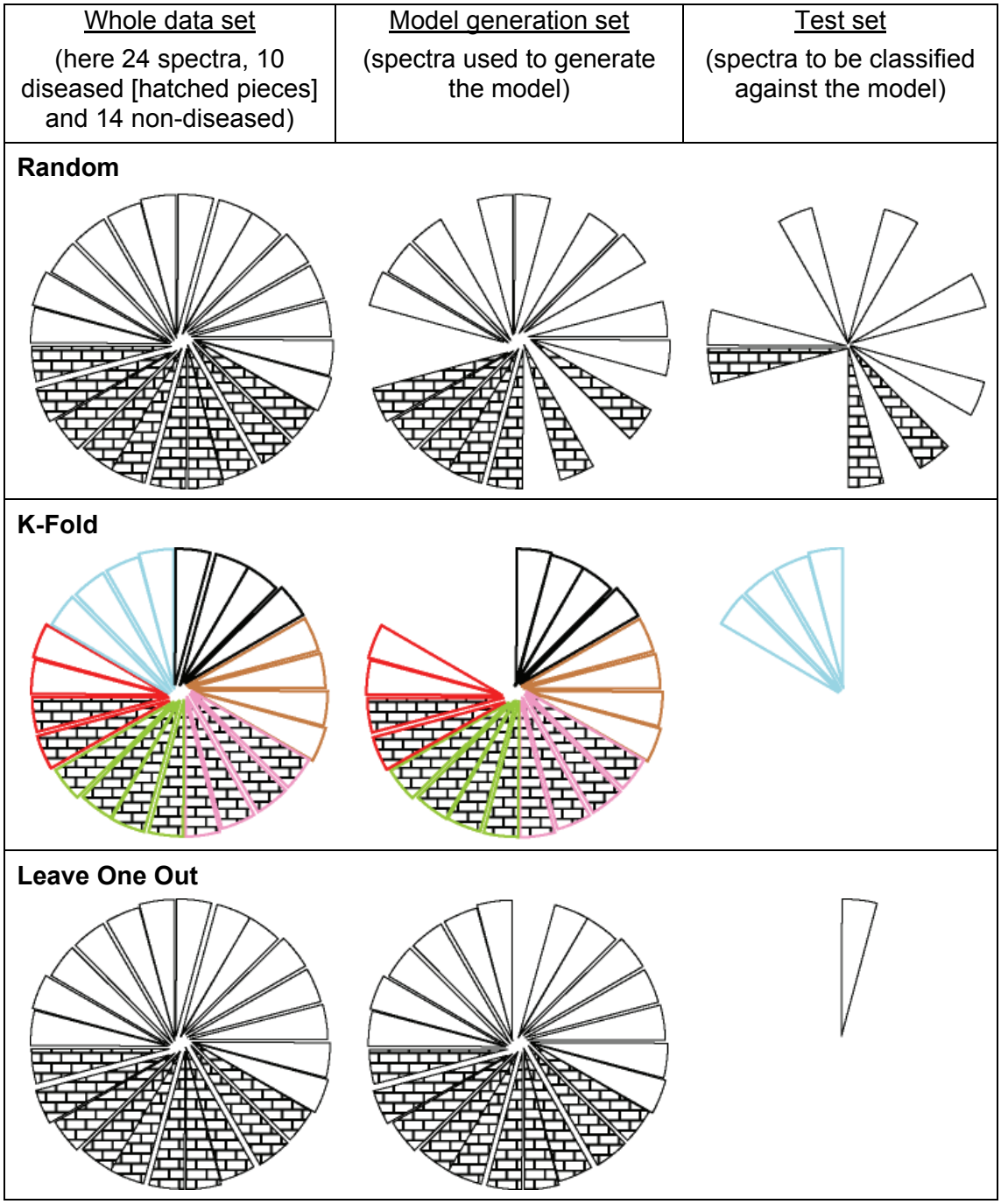


Figure 6-5 Illustration of the cross validation modes used in ClinProTools

- **Random**: A random subset of data points (taken over all classes) is selected and omitted from the model generation procedure. The model is calculated with the remaining data points and the random set of data points is classified against the model. The obtained classification results are stored. This procedure is repeated for a defined number of iterations and finally the averaged classification results give the prediction capability.
- **K-Fold**: The data set is divided into *k* equal parts of data points. Then *k* models are generated where each time a different one of the *k* part is omitted. The omitted part is classified against the model calculated from the remaining *k*-1 parts. The obtained classification results are stored for the *k* models, averaged and returned as the prediction capability.
- **Leave One Out**: As the name suggests exactly one data point is left out. The remaining points are used for model generation. The omitted data point is classified against the model. This procedure is repeated for *n* times, where *n* is the number of data points. The obtained classification results are stored for the *n* models, averaged and returned as the prediction capability.

In general, the choice of the cross validation procedure depends on the number of available data points. For larger data sets, a **K-Fold** or **Random** approach is recommended. If the number of data points is rather small (e.g. less than 30 spectra per class) and it is expected that a high variation within each class exists it is more reliable to use the **Leave One Out** method since in that case more data points remain for the modeling stage.

6.2.4 External Validation

External validation allows, similar to the cross validation measure (Section 6.2.3) obtained during the model generation procedure, predicting the capability of a calculated model.

External validation requires loading new sets of spectra for each class (e.g. control, cancer_stage_1, cancer_stage_2 …). These validation spectra should not have been used in the model generation step and could come e.g. from a fresh measurement (in accordance to the same clinical protocols) of patients. The validation spectra are loaded and prepared in the same way as the spectra used in model generation and then are classified against the model. There is the same selection done as during model generation by the Noise Spectra Exclusion, Adduct/Polymer Spectra Exclusion and Similarity Selection. It is recommended to use only suitable spectra for external validation. In the case of multiple measurements and Similarity Selection switched off, the peaks of a group are not averaged, but the spectra are treated separately. Because of that calculating, the recognition capability with the external validation workflow might differ from the one calculated during model generation, where the peaks are averaged.

The model predicts the probable class membership of the validation spectra. The Validation report (Section 8.1.1.7) includes a so-called confusion matrix with one row and one column for each class. The entries of the matrix indicate how many spectra from one class have been classified to the correct and to other classes. A perfect prediction would give a diagonal matrix and an average of 100% for the 'Correct Classified' values. This view can be used as an indicator for the prediction capability of the model on unknown data and reveals further if some classes are better predicted than others are. In addition, an individual Classification report (Section 8.1.1.8) is set up for each class (when the **Show Single Classifications** option set), which shows for each validation spectrum assigned to this respective class if the model was capable to classify this spectrum and to which class it has been classified. For the QuickClassifier we also obtain a likeliness measure, which indicates some kind of safety regarding the classification of a spectrum to a specific class.

6.3 Spectra Classification

For classification of unknown spectra, a complete classification model is needed. The model contains all information needed to prepare and classify the unknown spectra using the same parameters as were used for model generation. Classification depends on the type of the model-generating algorithm and the underlying classifier. ClinProTools supports two classification modes settable in the ClinProTools general settings:

- **Standard mode**: The standard mode is ClinProTools' normal classification mode. The spectra to be classified are loaded and displayed in ClinProTools. After the classification, the classification result is automatically shown in the Classification report (Section 8.1.1.8) and the 2D Peak Distribution View displays the corresponding peak data for the classified spectra. The classification result can be saved in an XML file on demand. However, since all spectra are kept in the memory the number of spectra that can be classified at a time is limited by the memory size.
- **Batch mode**: The batch mode is an alternative classification mode overcoming the spectra number limitation of the standard mode. Processing a big amount of spectra at a time might for example be necessary in the case ClinProTools is used for classification by the Bruker flexImaging software. In this mode, the spectra to be classified are neither displayed in ClinProTools nor kept in the memory; any number of spectra can be classified at a time. After the classification, a saving dialog pops up automatically to store the classification result in an XML file. The Classification report can be created on demand; however, it is not recommended to display big classifications because the browser used for display might take a long time to process the XML file with style sheet. Large XML files with style sheet should better be opened in Excel.

In both modes, the software holds the classification result as long as the classification is not closed.

6.4 Statistics in ClinProTools

ClinProTools supports various statistical tests and methods, which can be applied to the prepared spectra data. A short introduction to each test/method is given in this section. In addition, some remarks to statistical problems with MS data have to be made.

6.4.1 Statistical Tests

ClinProTools offers various statistical tests. With respect to the constraints for the individual tests, we can differ between tests expecting normal distribution and distribution free tests. The t-test and ANOVA test are tests expecting normal distribution of the underlying data and are for two (t-test) or *k* classes (with $k > 2$, ANOVA test). The supported Wilcoxon test ($k = 2$) or Kruskal-Wallis test ($k > 2$) do not depend on the normal distribution assumption and should be used for a more generic analysis. The Anderson-Darling test in the case of ClinProTools has been adapted to test for normal distributions. Each of these tests calculates the so-called p-value.

For a detailed introduction in statistical tests including mathematical theory, please refer to:

J. M. Chambers and T. J. Hastie, "Statistical Models in S", Wadsworth & Brooks/Cole, 1992.

R.E. Walpole and R.H. Myers, "Probability and Statistics for Engineers and Scientists", 5th ed., Macmillan, 1993.

6.4.1.1 T-Test

A t-test is a statistical hypothesis test in which the test statistic has a Student's t-distribution if the null hypothesis is true. In ClinProTools, we consider the t-test as a statistical test of the null hypothesis that the means of two normally distributed populations are equal. All such tests are usually referred to as Student's t-tests, though strictly speaking that name should only be used if the variances of the two populations are also assumed to be equal; the form of the test used when this assumption is dropped is sometimes called Welch's t-test.

There are different versions of the t-test depending on whether the two samples are

- independent of each other (e.g., individuals randomly assigned into two groups), or
- paired so that each member of one sample has a unique relationship with a particular member of the other sample (e.g., the same people measured before and after an intervention, or IQ test scores of a husband and wife).

For the classification problems considered with ClinProTools, the analyses must be applied on a set of independent individuals where intervention and no intervention are

not mixed. Therefore, we can ignore this variant and consider only independent samples. For the kind of multiple measurements of one sample, the different spectra are averaged to reduce the measurement variance after a possible pre-selection by use of some filter criteria.

If the calculated t-value is greater than the threshold chosen for statistical significance (alpha conventionally equal to 0.05), then the null hypothesis that the two groups do not differ is rejected in favor of the alternative hypothesis, which typically states that the groups do differ.

6.4.1.2 ANOVA Test

In statistics, analysis of variance (ANOVA) is a collection of statistical models and their associated procedures, which compare means by splitting the overall observed variance into different parts. The initial techniques of the analysis of variance were pioneered by the statistician and geneticist Ronald Fisher in the 1920s and 1930s, and are sometimes known as Fisher's ANOVA or Fisher's analysis of variance.

In ClinProTools, we calculate a so-called One-way ANOVA. A One-Way Analysis of Variance is a way to test the equality of three or more means at one time by using variances.

Assumptions:

- The populations from which the samples were obtained must be normally or approximately normally distributed.
- The samples must be independent.
- The variances of the populations must be equal.

The null hypothesis will be that all population means are equal; the alternative hypothesis is that at least one mean is different. If the decision is to reject the null, then at least one of the means is different. However, the ANOVA does not tell you where the difference lies.

6.4.1.3 Wilcoxon Test

The Wilcoxon rank-sum test is a non-parametric alternative to the paired Student's t-test. This test should be used whenever the assumptions that underlie the t-test cannot be satisfied. The test is named for Frank Wilcoxon who proposed this, and the rank-sum test, in 1945.

The null hypothesis tested is that a sample is symmetrically distributed around a specified center. It is often used to test difference scores of data collected before and after an experimental manipulation, in which case the central point would be expected to be zero. Scores exactly equal to the central point are excluded and the absolute values of

the deviations from the central point of the remaining scores are ranked such that the smallest deviation has a rank of 1. Tied scores are assigned for a mean rank. The sums for the ranks of scores with positive and negative deviations from the central point are then calculated separately. A value *S* is defined as the smaller of these two rank sums. *S* is then compared to a table of all possible distributions of ranks to calculate *p*, the statistical probability of attaining *S* from a population of scores that is symmetrically distributed around the central point. As the number of used scores, *n*, increases, the distribution of all possible ranks *S* tends towards the *z*-distribution, so for an *n* of greater than 10 this distribution is used to calculate *p*. This test assumes that the compared sample sets originate at least from a common distribution.

For details, please refer to F. Wilcoxon, "Individual Comparisons by Ranking Methods", Biometrics 1, pp 80-83 (1945).

6.4.1.4 Kruskal-Wallis Test

In statistics, the Kruskal-Wallis one-way analysis of variance by ranks is a non-parametric method. Unlike the analogous one-way analysis of variance, the Kruskal-Wallis test does not assume a normal population. This, like many non-parametric tests, uses the ranks of the data rather than their raw values to calculate the statistic. Since this test does not make a distributional assumption, it is not as powerful as the ANOVA test.

The hypotheses for the comparison of two independent groups are:

- H0 (null hypothesis): The samples come from identical populations
- Ha (alternative hypothesis): The samples come from different populations

Notice that the hypothesis makes no assumptions about the distribution of the populations. These hypotheses are also sometimes written as testing the equality of the central tendency of the populations.

The test statistic for the Kruskal-Wallis test is *H*. This value is compared to a table of critical values for *U* based on the sample size of each group. If *H* exceeds the critical value for *H* at some significance level (usually 0.05) it means that there is evidence to reject the null hypothesis in favor of the alternative hypothesis.

For details, please refer to W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis", Journal of the American Statistical Association 47 (260): pp 583-621 (1952).

6.4.1.5 Anderson-Darling Test

The Anderson-Darling test (AD test; Stephens, 1974) is used to test if a sample of data comes from a population with a specific distribution. It is a modification of the Kolmogorov-Smirnov (KS) test and gives more weight to the tails than does the KS test. The

KS test is distribution free in the sense that the critical values do not depend on the specific distribution being tested. The AD test makes use of the specific distribution in calculating critical values. This has the advantage of allowing a more sensitive test and the disadvantage that critical values must be calculated for each distribution. In the case of ClinProTools, the AD test has been adapted to test for normal distributions. The AD test is an alternative to the chi-square and Kolmogorov-Smirnov goodness-of-fit tests.

The AD test is defined as:

- H0: The data follow a specified distribution (in ClinProTools 2.2 normal distribution).
- Ha: The data do not follow the specified distribution.

The AD test is applicable if at least eight spectra are available. It gives an estimate on the normal distribution assumption. In general, one will consider a large set of peaks with different distribution properties (some peaks maybe normal distributed and some are not). From a formal point of view, the t-test/ANOVA test can only be used if the underlying distribution fits the normality assumption.

Hence, at first one should look at the p-value for the AD test. If it is above e.g. 0.05 one should consider the t-test or ANOVA test; otherwise, the result from Wilcoxon / Kruskal-Wallis (W/KW) test has to be evaluated. The relations are shown in Table 6-1.

Table 6-1: Relation of p-value from AD to p-value from t-test/ANOVA or W/KW

	p-value t-test/ANOVA ≤ 0.05	p-value t-test/ANOVA > 0.05	p-value W/KW ≤ 0.05	p-value W/KW > 0.05
p-value AD > 0.05	interesting peak	uninteresting peak	interesting peak	uninteresting peak
p-value AD ≤ 0.05	not applicable	not applicable	interesting peak	uninteresting peak

Thereby, "interesting peak" means that a peak shows a significant difference between the considered classes in a univariate point of view. The Wilcoxon/Kruskal-Wallis test has less restrictive constraints; hence, the p-value calculation needs a larger set of spectra for a valid p-value estimation. Therefore, if the p-value for AD > 0.05 one should always consider the t-test/ANOVA instead of Wilcoxon/Kruskal-Wallis to derive a decision.

For details, please refer to M. A. Stephens, "EDF Statistics for Goodness of Fit and Some Comparisons", JASA 69 #347, pp 730-737 (1974).

6.4.1.6 P-Value

A p-value is the probability that an observed effect is simply due to chance; it therefore provides a measure of the strength of an association. A p-value does not provide any measure of the size of the effect, and cannot be used in isolation to inform clinical judgment.

P-values are affected both by the magnitude of the effect and by the size of the study from which they are derived, and should therefore be interpreted with caution. In particular, a large p-value does not always indicate that there is no association and, similarly, a small p-value (Section 6.4.3.2) does not necessarily signify an important clinical effect.

Subdividing p-values into 'significant' and 'non-significant' is poor statistical practice and should be avoided. Exact p-values should always be presented, along with estimates of effect and associated confidence intervals. In peak statistics (Section 8.1.1.2), p-values are calculated for each picked peak using the corresponding statistical test. The p-value can be used for selecting peaks for model generation as well as for sorting and showing peak statistic results.

6.4.2 Statistical Methods

ClinProTools offers various statistic methods to calculate and visualize statistical properties of the underlying data, correlation analysis, receiver operating characteristic, principal component analysis and unsupervised clustering. The methods are described in a less formal way as they can be used for mass spectrometric data within ClinProTools:

6.4.2.1 Correlation Analysis

The correlation analysis is used to analyze stochastic relations between random variables upon a given sample set. In our context, the random variable is given by an individual peak and its properties (peak area), and the sample set is the given set of spectra. ClinProTools supports calculating correlation matrices (Section 8.1.1.3) and per-peak correlation lists (Section 8.1.1.4). A correlation matrix is obtained by comparing each peak in the list peak to each other peak whereas a correlation list results from comparing a selected peak to each other peak in the list. In both cases correlation analysis can be calculated over either all classes or only a selected one.

Algorithms for correlation analysis

ClinProTools offers two algorithms for correlation analysis, the standard correlation algorithm and the Kendall's tau-b algorithm:

- **Standard correlation algorithm**: (Default algorithm) The standard algorithm to determine the correlation matrix/list combines ordinary correlation coefficients of pair wise considered peaks in a common matrix.
- **Kendall's tau-b algorithm**: The Kendall's tau-b (KT) algorithm describes a rank correlation coefficient (-1.....+1). It is less frequent than the Spearman rank correlation coefficient (a well-known alternative algorithm that is not supported by ClinProTools because of the below mentioned reasons); however, it is much more powerful. Within the KT approach all value pairs are compared with each other in a common sense and not just two values of a pair, further the error ranks of pairs are evaluated. Hence, the KT algorithm is less sensitive against outliers. KT is more robust and has been recommended if the data do not necessarily come from a bivariate normal distribution. Kendall's tau-b is a nonparametric measure of association based on the number of concordances and discordances in paired observations. Concordance occurs when paired observations vary together, and discordance occurs when paired observations vary differently.

Correlation matrix and per-peak correlation list

The correlation matrix/list is a tool to analyze the relation between the peaks: Do peak areas vary independent of each other or is there a correlation between the variation of the peak areas? The correlation coefficient (cc) can be between -1 and +1. A cc of +1 means that the areas of the two peaks have a perfect positive correlation and go up and down in the same way. If the cc is -1 the two peaks are perfectly negative correlated: If the intensity of the first peak is above the mean level in one spectrum the other peak will be below the mean level and vice versa. Smaller absolute values of the cc indicate that the peak areas are not correlated and vary in an independent way. ClinProTools uses a color code ranging from red (cc = +1) to blue (cc = -1) to highlight different cc ranges in the matrix and thus allows quickly detecting highly correlated peak pairs.

The correlation matrix/list can be used to identify peaks that show concordances or discordances. If two peaks behave similar, i.e. their cc is close to ± 1, and if they are relevant for the classification task, it is quite common that only one of these peaks is incorporated in the model. In this case, 'alternative' peaks, which behave similar to the one available in the model, can be identified using the correlation matrix/list.

In a correlation matrix, it is also possible to identify groups of peaks, which are highly correlated. To group the peaks of the correlation matrix we start with one peak and add all peaks to the first group which have an absolute cc > correlation level with at least one of the peaks which are already part of this group. From the remaining peaks, additional groups are created. Peaks that are not connected to other peaks form a group on their own. Both the peaks within the groups and the groups against each other are sorted according to the sort mode of the statistical settings. Thereby, one could manually deselect highly correlated peaks from the model building stage to simplify the identification of biomarker candidates.

For details, please refer to R. A. Becker, J. M. Chambers, and A. R. Wilks, "The New S Language", Wadsworth & Brooks/Cole, 1988.

6.4.2.2 Receiver Operating Characteristic

ClinProTools calculates a Receiver Operating Characteristic (ROC) curve for each peak within peak calculation. The ROC curve gives a graphical overview about specificity and sensitivity of a test or, within ClinProTools, an evaluation of the discrimination quality of a peak. The sensitivity represents the true positive fraction (TPF) and the specificity the true negative fraction (TNF) (Section 6.4.3.6). The fraction of false negatives (FNF) together with the TPF give a sum of 1 (100%); and the fraction of the false positives (FPF) together with the TNF also give a sum of 1 (100%).

Note: ROC curves can only be generated for the case of two model generation classes because a true/false decision is not possible for more than two classes.

Within Figure 6-6, the ROC curve is explained on an example of two populations – diseased and non-diseased patients. For all patients the same test is performed and numeric results are received for each patient. A plot of these results leads to the upper diagram shown. The vertical (green) line within the diagram indicates an arbitrary chosen threshold: a value above this threshold represents a positive test result and a value below it a negative test result. The position of this cut-off point will determine the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). If the test threshold is moved from left to right, the proportion of the FP decreases, but the TP also decrease simultaneously (lower diagram). The ROC curve (graph right to the lower diagram) is an exploration of what happens to the TPF and the FPF if the position of the arbitrary threshold is varied. The point corresponding to the chosen threshold is shown on the ROC curve as the cross. If the threshold is very high, almost no FP occur on the one hand, but only less TP are identified on the other hand. If the threshold is moved towards a more reasonable, lower value, the number of TP increases (the ROC curve moves steeply up). Finally, a region will be reached where there is a remarkable increase in FP and the ROC curve slopes off as the test threshold is moved down to ridiculously low values.

The best possible prediction method would yield a graph that was a point in the upper left corner of the ROC space, i.e. 100% sensitivity (all TP are found) and 100% specificity (no FP are found). A completely random predictor would give a 45-degree diagonal, the so-called no-discrimination line. Thus, the closer the ROC curve follows the left-hand border and then the top border of the ROC space, the more accurate it is and the closer the curve comes to the diagonal, the less useful is the test at discriminating between two populations. A more precise way of characterizing this "closeness to the diagonal" is to look at the area under the ROC curve (AUC). The area measures discrimination, which is the ability of the test to correctly classify those with and without

disease. The closer the area is to 0.5, the less useful is the test, and the closer it is to 1.0, the better is the test.

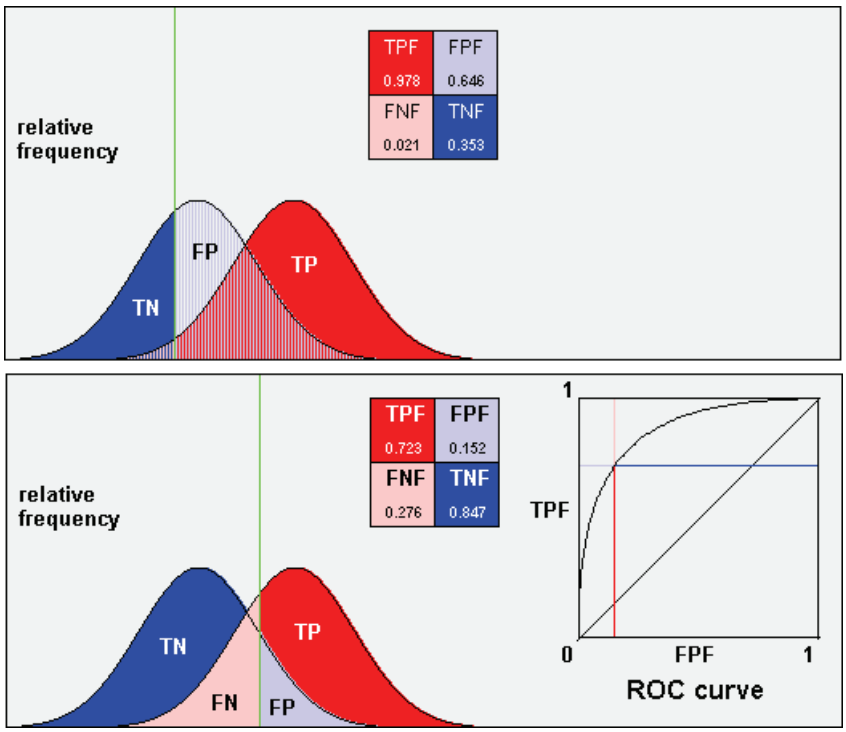


Figure 6-6 Graphical explanation of the ROC curve

In ClinProTools, the ROC curve is generated similar as explained above whereby a peak is considered as the random variable, which can be interpreted as a test separating two populations. The peak area or the intensity of the peak represents the threshold that is used to reach the separation into the two groups. ROC curves for all calculated peaks can be viewed in the ROC Curve View as shown in Figure 6-7. On the x-axis the '1-specificity' in terms of the false positives is given and on the y-axis, the sensitivity in terms of the true positives is recorded; for this it is assumed that the first loaded class is the diseased one and the second loaded class is the non-diseased one. Both axes are given in values between 0 and 1. At the bottom of the plot, the peak number, peak position and AUC value are given. If the data is separable by a univariate approach considering only one peak as a test criterion the ROC Curve View may already indicate this peak by a high AUC value close to 1.0.

Note: ROC curves and their AUC values are only estimations and become more confident with an increasing number of samples.

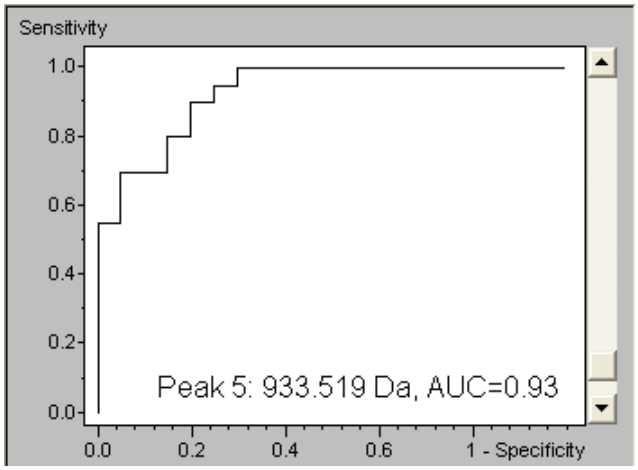


Figure 6-7 ROC curve for a good separating peak with high AUC value

6.4.2.3 Principal Component Analysis

ClinProTools offers a statistical data analysis in terms of principal component analysis (PCA). The PCA is managed by an external MATLAB software tool, which is integrated in ClinProTools.

PCA is a broadly used mathematical technique designed to extract, display and rank the variance within a data set. The overall goal of PCA is to reduce the dimensionality of a data set while simultaneously retaining the information present in the data. In data sets with many groups of variables, variables often show similar behavior and contain redundant information. In the case of mass spectra, the variables are represented by the intensity at defined masses. According to the resolution, the number of these variables can be very high. The PCA reduces the number of dependent variables contained within the spectra set via replacing groups of variables by a single new variable. By this, a set of new variables, so called principal components will be generated. Each principal component (PC) is a linear combination of the original variables. All principal components are orthogonal to each other, so there is no redundant information. In many cases (depending on the complexity of the data set), only few PCs (compared to the large number of original variables) contain most of the variance. The full set of PCs is as large as the original set of variables, nevertheless only the first PCs are of interest mostly; higher PCs contain very detailed spectra information and the highest PCs contain spectra noise.

Figure 6-8 describes the transformation of a data set to PCs in a simplified graphic. Actually, each sample (spectrum) can be plotted in an m-dimensional space of variables. Diagram A shows a plot of the spectra (represented by grey points) in a three-dimensional space of variables as simplification. The PCA ranks the variables according to their influence on the data set. Upon PCA calculation, the original coordinates of

the diagram are transformed to new coordinates ranked by the variance each coordinate explains. The new axes are called PCs (B). PC1 describes the largest variance within the data set; PC2 describes the second largest variance, and is orthogonal to PC1, etc. This is indicated by the strength and the orientation of the arrows in diagram B. The variance explained by a PC is calculated as sum of the individual variance (C).

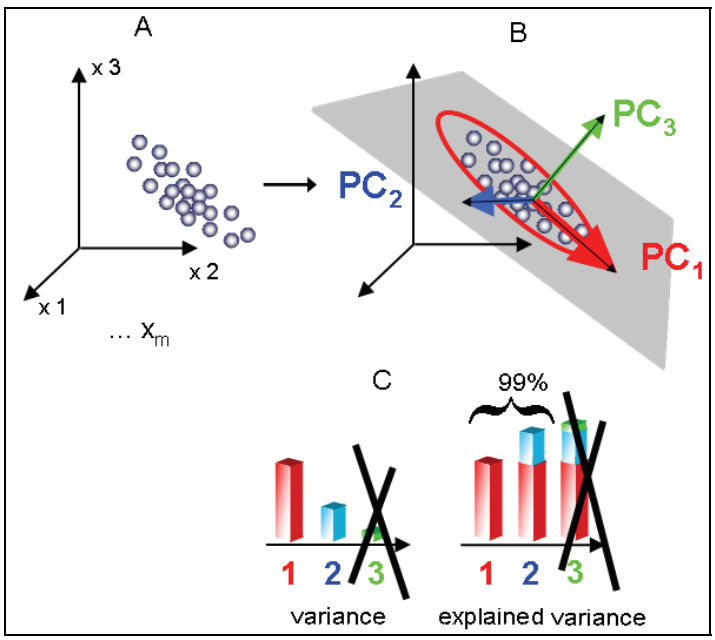


Figure 6-8 Simplified representation of the generation of PCs from a data set (explanations are given in the text)

From the PCA results so called scores and loadings can be derived and displayed in various plots (Section 7.5.2.1):

- **Scores**

The score output represents the original data mapped into the new coordinate system which is defined by the PCs. Within the Scores plot, outliers from a group or from several groups can be discovered and visualized. Outliers are samples which are extreme or do not fit the PCA model. Independent from the PC coordinates all Scores plots contain the same sample number as the original data set.

- **Loadings**

During the calculation of PCs, the variables (peaks) obtain different loadings in dependence on their contribution to the explained variance of a PC. The values for the loadings are between -1 and 1. A negative value indicates a negative loading of the respective variable, a positive value reports a positive loading of the variable and a value of ‘0’ shows, that the respective variable has not influence on the variability of the

PC. In the case of mass spectra, the loadings give information about the contribution of single peaks to the variance covered by the respective PC.

For details, please refer to I. T. Jolliffe, “Principal Component Analysis” (2002), Springer, 2nd edition.

6.4.2.4 Unsupervised Clustering

A clustering workflow has been added using a hierarchical clustering algorithm. The calculation can be done on PCA-transformed data or on the untransformed peak lists. If the PCA is used, limiting the PCs to those necessary for explaining 95% of the variance serves as a good data reduction.

After performing the calculation, the class membership of the data sets is stored as *ClinProtClustering.xml* in the CPT folder.

With hierarchical clustering in the data space the distance of the data points is calculated based on a metric. For a given number of classes the data are grouped accordingly. A dendrogram presenting the hierarchy is displayed. The complete tree C*linProtClustering.tree.xml* is exported to the CPT folder. If the full tree option is set, the spectra paths can be displayed at the end of the branches. Node numbers corresponding to the nodes in *ClinProtClustering.tree.xml* are displayed at the dendrogram nodes. In the additional output *ClinProtClustering.tree2.xml*, the XML structure is identical to the tree structure; each node is represented by a XML element with two sub node elements. The number of sub nodes is available as an attribute at each node. If a limited number of classes has been chosen, the class number at the end of the branches corresponds to the class number in *ClinProtClustering.xml*.

The hierarchical clustering uses the MATLAB algorithms. For a further documentation of the parameters, one can refer to the corresponding MATLAB documentation, available at http://www.mathworks.com (search for *pdist* for the Distance Method parameters and for *linkage* for the Linkage Method parameters).

For an easy-to-understand introduction in hierarchical clustering in general, see http://home.dei.polimi.it//matteucc/Clustering/tutorial_html/hierarchical.html.

6.4.2.5 Pattern Matching for Outlier Detection

If peak picking on single spectra is chosen (Section 6.1.1.5.2), the overall averaged peak list as well as per-class averaged peak lists are calculated. If a statistical model is generated, these data will be serialized too. When spectra are classified against this model, a peak list of the spectrum will be generated using the same peak picking parameters used during model generation. This peak list will be matched against the overall and per-class peak lists stored in the model. The resultant "scores" are

displayed in the Classification report (Section 8.1.1.8). The range of the scores is 0..1, with 0 for no correspondence between spectrum and the spectra used during model generation. The higher the value, the better is the correspondence.

This value can be used to support outlier detection. Spectra with a low overall score respectively with low scores for all classes have a different peak pattern and are therefore candidates for outliers.

6.4.3 Remarks on Statistical Problems with MS Data

There are certain statistical problems with MS data, which are described in the following sections.

6.4.3.1 Common Statistical Pitfalls - Generic Remarks

Statistical constraints

Most statistical tests set some constraints on their applicability. Some tests e.g. expect a normal distribution or the variables needs to be independent. These constraints must be kept in mind when using the statistical methods within ClinProTools. Some tests are robust if the constraints are not completely valid; so it is in general safe to additionally process an ANOVA test if the normal distribution is not perfectly given. Other methods may be more sensitive and the results may become very inaccurate and invalid.

Multiple measurements

If during the measurements, samples are spotted multiple times and are processed with ClinProTools we are focused with the multiple measurement (mm) problem (Section 6.4.3.3). If mm are used, they must be processed as mm. This has to be specified by the user in the spectra preparation settings; ClinProTools then automatically processes the mm in the correct form sample by sample. If mm are processed without the corresponding option set, each spectrum is considered as an individual sample. This would lead to invalid cross validation statistic results.

Unequal class sizes

Unequal class sizes are very common in clinical research especially by considering cancer and control classes. This has some effects, which should be kept in mind. For the data preparation, the peaks are picked on an average spectrum. In the next step of the statistic calculation, very small sample sizes for one class may give a bias in the test procedure. In addition, the classification procedures are more or less affected by different sample sizes. For the GA the optimization could be dominated by e.g. one large class. The QC relies on some statistical measurements; hence, the QC may be affected by unequal sample sizes similar as the statistic calculation. The SVM looks for extreme boundary points to determine the hyperplanes; hence, it is less effected by unequal sample sizes as long as the boundaries have a clear definition. However, in

the cross validation it may happen that exactly this boundary points are removed and the SVM performs poor.

Different classes with pre and post treatments

Currently, ClinProTools considers each loaded class as independent with respect to the other classes. In addition, it is expected that the samples are independent to each other, in the sense that e.g. two loaded samples do not come from the same clinical specimen. Therefore, ClinProTools does not completely support a scenario where one class contains samples (e.g. cancer) before clinical treatment and the second class contains samples post a clinical treatment. To handle such cases a modeling of the semantic of the clinical specimens and the experimental design is necessary which will be part of a future version of ClinProTools.

6.4.3.2 Small P-Value Phenomenon

Within ClinProTools, different kinds of statistical tests are offered to the user. They can be used to identify peaks, which show a significant difference between the considered classes. Within clinical proteomics, we are in general confronted with a small set of samples and a large number of identified peaks. From the identified peaks, ClinProTools derives some characteristics such as the peak area/intensity, which are considered as features and used for the further processing. These peak areas/intensities are the values that are analyzed by the statistical tests as well as by the classification algorithms.

The statistical tests give a p-value for each peak. This p-value is a probability measure for the strength of an association between the different classes. The exact value and the reliability of the p-value depend on some aspects, which are explained in the following.

The statistical tests differ in their performance and their requirements with respect to the underlying data (in our case the peak areas). As a result not each test can (from a theoretical / formal point of view) be applied to each set of data. In addition, the performance (power) of the test depends on a specific combination of constraints. In general, each of the supported tests makes at least the following constraints:

- C1: number of classes (2 classes, ≥ 2 classes)
- C2: kind of distribution of the underlying data (normal distribution, arbitrary distribution) but the same for each class
- C3: number of necessary disjunctive samples (small, large)
- C4: number of features (small, large)
- C5: properties of the measurement

To constraint C1:

The first constraint is a strict constraint in the sense, that some tests are not applicable if the number of classes (e.g. control, cancer1, cancer2) is larger than 2. This is the

case for the t-test and the Wilcoxon test, which are only applicable for two-class scenarios. In general, one can say that as larger the number of classes as more complicated the test scenario.

To constraint C2:

If the data follow a specific distribution, we already know a lot about the data and their behavior. Therefore, tests have been developed which give very nice powerful test procedures as long as the data are nearly e.g. normally distributed. The t-test and the ANOVA test are tests of this kind and their results are only valid (from a formal point of view) if the underlying data fits the normal distribution constraint. Applying these tests on non-normally distributed data will distort the test results.

To constraint C3:

This constraint is very important and common for all tests. In general, a large number of disjunctive (no multiple measurements or duplicates) samples improves the power of the test and make thus the results more reliable. A small sample size on the other hand increases the probability for a wrong test decision.

To constraint C4:

In clinical proteomics, we search for potentially interesting peaks or masses, which are capable to separate different classes (e.g. cancer, control). In contrast to the common assumption the more the better, a large number of features does not necessarily improve the performance of the search for such candidates. This is mainly because many of the features are not important for the underlying classification question and hence are in some sense - noise. To overcome this it is important to have a good parameterization for the peak detection (e.g. S/N threshold) to pick only peaks that have a good S/N-ratio and are likely to be important. The classification and statistical testing problem becomes much easier if the number of potentially important peaks/ features is small and the features are not just noise.

To constraint C5:

It is important to be careful about dependent samples (Section 6.4.3.4), e.g. samples, which are measured from the same clinical person, cannot be considered to be independent. The same applies to multiple measurements of the same sample (multiple spotting) (Section 6.4.3.3). This has effects on the determined features (peak areas). The first case has to be controlled by the user. If dependent samples are used in the same class, the subsequent results are more or less inaccurate. The second point can be controlled by the ClinProTools multiple measurement handling.

If multiple measurements (mm) or dependent samples are considered as independent, the test results can become distorted. For mm which are considered to be independent we could get p-values which are extremely small whereas if the mm are considered valid the p-values are in correct ranges.

Another problem in the determined features is a large number of zero's. If for a larger number of spectra at a picked peak the peak area is close to zero, we are confronted

with this problem. Some tests are very sensitive to a large number of (close to) zero values and the p-values may again be unrealistic small. If such effects are observed one should take a closer look on the values for this feature (mass) by analyzing the exported peak list. The peak list can be exported to XML or CART format (Section 8.4).

From the above explanations, it becomes obvious that the more flexible the constraints (≥ 2 classes, distribution free, small sample size, large number of features) the more complicated the test scenario. For very relaxed constraints, most tests are not powerful enough and the obtained results are in fact invalid. To overcome this one has to except some constraints to make the problem more suitable. For example, one could increase the number of disjunctive samples by 10 or 100 (if possible) - this will improve correctness of the underlying estimations and improve the performance of the tests. It is also very common to reduce the number of features to a smaller subset (e.g. 100 features) by omitting features which are probably unimportant, e.g. because of some pre-knowledge.

If one ignores these problems or the data just does not fit to these aspects the obtained p-values from the statistical tests are in fact poorly estimated and may be inappropriate. If e.g. the number of samples is small and the normal distribution assumption is not true, p-values from the t-test or the ANOVA test may be very (unrealistic) small. To take now just the alternative distribution-free test is also no solution hence this test requires a much larger number of samples to obtain the same power as the distribution-free test, because it does not any longer depends on a specific distribution. This has to be kept in mind - it may still be true that more relevant peaks have smaller p-values than unimportant peaks but the exact p-value is not any longer valid.

6.4.3.3 Multiple Measurements of the Same Sample

Multiple measurements (mm) occur if the same sample is measured multiple times. This can automatically be done by the ClinProt measurement system by multiple spots of the same sample. The obtained spectra (generally 4 for each sample) are stored in a common directory named by the sample_id. These measurements should (in general) be very similar.

ClinProTools has to handle samples measured with mm in a special manner for e.g. formal reasons regarding statistics and model building. If mm are available for a sample the **Support Spectra Grouping** option in the **Settings Spectra Preparation** dialog has to be activated before opening any files. This enables ClinProTools to search for specific directory structures for mm as created by the measurement system. By default, the measurement system automatically manages the directory structure and mm are supported valid.

Note: Therefore, it is strongly recommended not to modify the directory structure under a sample set top folder. Otherwise invalid groupings may occur.

If mm are present an additional processing option is available in ClinProTools. The similarity selection filter (Section 6.1.3.2) can be used to select a characteristic spectrum from the mm, which will end in one spectrum per sample and finally the ordinary processing queue as if there would be only single measurements. If the similarity selection filter is not used multiple measurements are averaged before further processing. All remaining processing steps e.g. model generation, statistics calculation, etc. are done upon these averaged spectra (one averaged spectrum per sample).

It is very important to handle mm in this special way. If mm are considered as independent measurements (spectra grouping not active) the statistics will become very inaccurate, because the number of spectra is artificially increased e.g. by a factor of 4 when 4 mm are available per sample and the underlying statistics (e.g. variance etc.) are in fact invalid. In addition, the model generation is effected and the cross validation may be inaccurate.

6.4.3.4 Dependent Measurements of Different Samples from the Same Clinical Person

Another scenario occurs if dependent measurements of different samples from the same clinical person are used. In fact, ClinProTools is currently not designed for this purpose and takes no care about dependent samples in a set of spectra. The results should be seen under this strong constraint.

6.4.3.5 Multiple Hypothesis Testing - Analyzing a Large Number of Peaks at the Same Time

The identification of biomarker candidates within ClinProTools focuses on the detected peaks over a given set of spectra. If the number of detected peaks is large, we also have a large number of features (peak areas). Applying statistical tests on each feature at the same time forms the case of the so-called multiple hypothesis testing.

The application of a statistical test on a single feature aims on single hypothesis testing. This is the case when we want to know if the given feature (derived from a peak) shows a significant difference between the considered classes. This is a single hypothesis. In general, however, the number of features is large and for each feature, we create a hypothesis, which is tested by some statistical test. If we do so, we are considering multiple hypotheses.

Testing each of a large number of hypotheses at the same alpha-level as for a single hypothesis normally leads to a large number of false positives, i.e. features are called significant although there is no expression change in reality. To overcome this problem different p-value adjustment procedures have been proposed in the statistical literature.

ClinProTools automatically applies the so-called Benjamini & Hochberg p-value adjustment procedure to adjust the p-values to observe the multiple hypothesis problem.

For details, we refer to: S. Dudoit and J. Popper Shaffer and J. C. Boldrick, "Multiple Hypothesis Testing in Microarray Experiments", Statistical Science, Vol. 18(1), pp 71-103, 2003.

6.4.3.6 How to Determine Sensitivity and Specificity from External Validation

The sensitivity of a binary (two-class) classification algorithm, such as a blood test to determine whether a person has a certain disease, is a parameter that expresses something about the test's performance. The same applies to the specificity. The semantic of these two characteristics depends on the setting of the positive and the negative class. In a binary scenario, the corresponding values can be derived from the ClinProTools result output as follows:

Workflow:

- Create a binary classification model.
- Click Validate... or select the **External Validation** command from the **Classification** menu or the Model List View context menu.
- Load external data for both classes and start external validation.
- Three new XML files are shown (Validation, Classification Validation class 1 and Classification Validation class 2 reports)
- Read from the confusion matrix in the Validation report (Section 8.1.1.7) the number of true positives, false negatives, true negatives and false positives and calculate sensitivity/specificity (see the following example).

Determination of sensitivity, specificity and positive, negative prediction values

If we assume the positive class (e.g. diseased) is class 1 and the negative class (e.g. control) is class 2, sensitivity and specificity can be derived by considering the confusion matrix in the Validation report as follows (Figure 6-9):

ClinProt Validation

Class	Name	Correct Classified Part of Valid Spectra	1	2	0	Inv.
1	Diseased	88.9 %	24(1)	3(3)	0	0
2	Control	73.1 %	7(2)	19(4)	0	0

Figure 6-9 Validation report (used in our example)

From the confusion matrix, which in our example includes the columns labeled with '**1**' (classifications to class 1), '**2**' (classifications to class 2), '**0**' (unclassified spectra) and '**Inv.**' (number of invalid spectra), count for:

Class 1 (diseased)

(1) the number of correct positive classified spectra as 'TP' (= true positives)

(2) the number of wrong positive classified spectra as 'FP' (= false positives)

Class 2 (control)

(3) the number of wrong negative classified spectra as 'FN' (= false negatives)

(4) the number of correct negative classified spectra as 'TN' (= true negatives)

Thereby, the rows can be seen as the true classifications taken from the sample set and the columns indicate the prediction of the machine including not classifiable samples.

Considering the example of the above Validation report for the validation of two classes (class 1: diseased, class 2: control) one obtains:

Sensitivity

The sensitivity of such a test is the probability that the test has a positive outcome when the tested person is truly diseased.

sensitivity = true positives / (true positives + false negatives)

sensitivity = (TP) / (TP + FN)

In our example validation:

sensitivity = 24 / (24 + 3)

sensitivity = 88.9 %

Specificity

The specificity of such a test is the probability that the test has a negative outcome when the tested person is truly not diseased.

specificity = true negatives / (true negatives + false positives)

specificity = (TN) / (TN +FP)

In our example validation:

specificity = 19 / (19 + 7)

specificity = 73,1 %

Sensitivity alone does not tell us all about the test, because a 100%-sensitivity can trivially be achieved by labeling all test cases positive and a 100%-specificity can trivially be achieved by labeling all test cases negative. However, in the first case, the

specificity would be zero and in the second case, the sensitivity would be zero, respectively.

A test with a high sensitivity has fewer Type II errors, a test with a high specificity has fewer Type I errors. For explanation of Type I and Type II errors please refer to the Glossary (Appendix A.2).

Note

- Sensitivity = (true positives) / (true positives + false negatives)
- Specificity = (true negatives) / (true negatives + false positives)
- Positive prediction = (true positives) / (true positives + false positives)
- Negative prediction = (true negatives) / (true negatives + false negatives)

References

http://gim.unmc.edu/dxtests/Default.htm

7 WORKFLOWS IN DETAIL

In the following the ClinProTools workflows will be described in detail; here not just the default settings are used like described for the basic workflows in Section 4.4. This section includes details on loading and preparing data for model generation, generating and validating classification models, classifying spectra, calculating peak statistic and correlation analysis and performing PCA and unsupervised clustering.

7.1 Spectra Loading and Data Preparation

All spectra that are loaded for peak statistic calculation, model generation or classification have to be prepared either automatically by applying specified parameters or manually by user action. During loading, the spectra are prepared by applying various spectra modifying and selecting filters. Further data preparation concerning spectra recalibration and averaging, average peak list calculations as well as peak calculation and selection is performed when launching the respective ClinProTools workflows.

7.1.1 Defining the Data Preparation Settings

The data preparation settings specify how spectra preparation and recalibration, average spectra calculation, average peak list calculation and peak calculation is performed. They are specified in the **Settings Spectra Preparation** and **Settings Peak Calculation** dialogs. The settings are automatically stored in the *SettingsDataPreparation.xml* file, which is loaded when ClinProTools is started and is updated on each settings change. To keep special settings you can save them in an XML file.

Although only used in model generation the **Settings Peak Selection** parameters can also be specified in the context of defining the data preparation settings since initial peak selection is a part of the peak calculation workflow. The settings are automatically stored in the *SettingsModelGeneration.xml* file.

7.1.1.1 Setting the Spectra Preparation Parameters

The **Settings Spectra Preparation** dialog (Section 9.1.4.1) defines the settings for preparing spectra. Most of the parameters apply to filters that modify or select spectra during spectra loading. Some parameters also apply to peak picking for spectra recalibration as well as to spectra recalibration itself and averaging. You can use the default parameters or specify own settings suitable for your data. Alternatively, you can

load a data preparation settings file or reset the current settings to the default values (Section 7.1.1.4).

Note: If these parameters are changed after spectra loading or even processing the views may become cleared to prevent the spectra from further processing and a message will inform you on how to proceed.

To set the spectra preparation parameters:

1. From the **Data Preparation** menu, select **Settings Spectra Preparation**.
2. In the **Settings Spectra Preparation** dialog, specify the parameters as desired and click **OK**.
3. If the views become cleared, quit the message and do the required action.

7.1.1.2 Setting the Peak Calculation Parameters

The **Settings Peak Calculation** dialog (Section 9.1.4.2) defines the settings for peak picking on either the total average spectrum or the single spectra and peak calculation in the individual spectra. You can use the default parameters or specify own settings suitable for your data. Alternatively, you can load a data preparation settings file or reset the current settings to the default values (Section 7.1.1.4).

Note: If the parameters are changed after spectra processing, the views may become cleared to prevent the spectra from further processing and a message will inform you on how to proceed.

To set the peak calculation parameters:

1. From the **Data Preparation** menu, select **Settings Peak Calculation**.
2. In the **Settings Peak Calculation** dialog, specify the parameters as desired and click **OK**.
3. If the views become cleared, quit the message and do the required action.

7.1.1.3 Setting the Peak Selection Parameters

Although the peak selection becomes effective only in model generation, it is part of the peak calculation workflow and thus its settings are specified in the context of defining the data preparation settings. However, you can change the current selection afterwards (Section 7.2.1.2).

The **Settings Peak Selection** dialog (Section 9.1.5.1) defines the settings for selecting peaks for model generation. By default, ClinProTools uses all picked peaks in model generation but you can specify that only selected best peaks with respect to the chosen sort mode should be included. You can use the default parameters or specify own settings suitable for your data. Alternatively, you can load a stored model generation

settings file or reset the current settings to the default values (Section 7.2.1.1.3).

Note: The peak selection settings may strongly influence the quality of the chosen classification algorithm. In many cases, a reasonable reduction of peaks improves the classification performed by the algorithms.

Note: If the parameters are changed after running the peak calculation workflow, the current peak selection is changed according to the new settings.

To set the peak selection parameters:

1. From the **Model Generation** menu, select **Settings Peak Selection**.
2. In the **Settings Peak Selection** dialog, specify the parameters as desired and click **OK**. If the peak calculation workflow has already been run, the current peak selection is changed according to the new parameter settings.

7.1.1.4 Saving, Loading and Resetting the Data Preparation Settings

The data preparation settings are automatically stored in the *SettingsDataPreparation.xml* file which is updated on each settings change. To keep the data preparation settings you have adapted to special analytical tasks you can save them in an XML file with a specified name. This allows loading the settings again. Changed settings can be reset to the defaults. Loading a data preparation settings file or resetting the current settings to their defaults is always possible; however, if spectra have already been loaded you might have to close the spectra and load them again or repeat the previously processing depending on which data preparation settings have been changed.

To save the current data preparation settings:

1. From the **Data Preparation** menu, select **Save Settings Data Preparation**. This opens the **Save Data Preparation Settings File** dialog with the SettingsDataPreparation folder as the default storage location.
2. Specify the file name and target folder and click **Save**.
3. If you have selected an existing file name, answer the confirmation request to overwrite the file.

To load a data preparation settings file:

1. From the **Data Preparation** menu, select **Load Settings Data Preparation**. This opens the **Load Settings Data Preparation File** dialog with the SettingsDataPreparation folder opened by default.
2. Navigate to the file you want to load. Double-click it or select it and click **Open**. This overwrites the current data preparation settings with the loaded ones.
3. If spectra are currently loaded, follow the instructions in the appearing message on how to proceed.

To reset the current data preparation settings to the defaults:

1. From the **Data Preparation** menu, select **Reset Settings Data Preparation**.
2. Confirm the appearing request to reset the current settings to the defaults.
3. If spectra are currently loaded, follow the instructions in the appearing message on how to proceed.

7.1.2 Loading Spectra in ClinProTools

To generate a model or calculate statistics one or several classes have to be loaded. ClinProTools loads all spectra in a folder and its subfolders recursively as one class.

ClinProTools supports loading spectra of the X-Mass, BAF und ASCII file formats. For loading ASCII files (Appendix A.4), the Null Spectra Exclusion filter (Section 6.1.3.2) in the **Settings Spectra Preparation** dialog has to be disabled.

For model generation two classes must be loaded at least. Single classes can be loaded for peak statistic operations, performing PCA or unsupervised clustering. You have to load all classes that should be included in model generation or statistic operations before you start any data processing (e.g. recalibration, peak calculation). A later loading of additional classes is not possible without starting complete data preparation and thus spectra loading again.

There are two ways to load spectra in ClinProTools:

- Via the **Open Model Generation Class** command from the **File** menu, you can select a folder and load all spectra contained as one model generation class. This operation has to be repeated until all model generation classes you want to open were loaded.
- Alternatively, a spectra import XML file (Appendix A.4) can be opened via the **Open Spectra Import XML** command. This automatically loads all spectra in the referenced folders as different model generation classes with respect to folder definition.

Before loading a class, the available memory is checked against the memory needed to load the selected spectra provided the **Check Memory on Load** option in the **General Settings** dialog (Section 9.1.1.12) is set. If the memory is insufficient, a warning message will appear which asks you whether to continue.

Upon opening a class all spectra in the selected/referenced folder are loaded and prepared according to the current spectra preparation settings. This includes baseline subtraction and normalization of spectra as well as various additional filtering processes. The loaded spectra are displayed in the Spectra View and Gel/Stack View. The first loaded collection is referred to as 'class 1:' in the ClinProTools title bar, the second as 'class 2:', etc.

A running spectra loading can be canceled by clicking  or 

. This cancels the current class loading and also closes and unloads all classes opened before.

A list of all loaded spectra can be viewed in the Spectra List report (Section 8.1.1.1) using the **Spectra List** command from the **Report** menu. The report also informs about the spectra's current include/exclude state and certain data acquisition parameters.

7.1.2.1 Opening a Model Generation Class

To open a model generation class you have to select the folder which contains the spectra you want to load as one class. The loading procedure has to be repeated for each model generation class of interest.

To open a model generation class:

1. From the **File** menu, select **Open Model Generation Class** or click  .
2. In the **Browse for Folder** dialog, navigate to the folder that contains the spectra you want to load and click **OK**. This loads all spectra in this folder and perhaps subfolders as one model generation class.
3. If a message about a too low memory size appears, decide how to continue.
4. Repeat steps 1 to 3 for each model generation class you want to load.

7.1.2.2 Opening a Spectra Import XML File

Opening a file of the *ClinProtSpectraImport.xml* format (Appendix A.4) allows loading a list of referenced spectra of different model generation classes at once. The import file can contain a path list of spectra or only class paths.

To open a spectra import XML file:

1. From the **File** menu, select **Open Spectra Import XML** or click  .
2. In the **Open Spectra Import XML** dialog select the desired XML import file and click **Open**. This loads all referenced spectra according to their class membership.

7.1.3 Manually Excluding/Including a Spectrum

Spectra can automatically be excluded by applying specific selecting filters (Section 6.1.3.2) during spectra loading. In addition, you can manually exclude spectra you do not want to use in further processing or re-include spectra that have (automatically) been excluded before. Spectra can only be excluded or included before any further processing is started.

Excluded spectra are displayed in the Spectra View and Stack View with a darker color than the included spectra of the same class (e.g. in dark red instead of red). In the Gel View, excluded spectra can be marked by a default color code (Section 9.1.3.7.3)

which indicates the reason of exclusion. Moreover, excluded spectra can be hidden from being displayed in the Gel/Stack View (Section 9.1.3.7.4).

To exclude/include a spectrum manually:

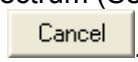
1. In the Spectra View or Gel View, select the spectrum you want to exclude/include.
2. From the **Edit** menu, select **Exclude Spectrum** resp. **Include Spectrum**. Alternatively, you can select the command from the Spectra View context menu.

7.1.4 Recalibrating Spectra and Calculating Average Spectra

The recalibration workflow performs spectra recalibration if enabled (default setting) as well as total average spectrum and class average spectra calculation from all non-excluded spectra. The recalibration of spectra is based on the corresponding settings in the **Settings Spectra Preparation** dialog (Section 9.1.4.1).

The recalibration workflow can be started manually using the **Recalibration** command from the **Data Preparation** menu. The workflow will be run automatically if a workflow that requires spectra being recalibrated is launched without the recalibration step has been performed yet.

The spectra that are found by the spectra quality filter (Section 6.1.3.2) to be 'not recalibratable' are marked as such and become excluded if the corresponding option is set. In the Gel View, not recalibratable spectra can be marked with a special color code according to their state (Section 9.1.3.7.3); different colors are used for 'not recalibratable but included' and 'not recalibrated, excluded' spectra. The total average spectrum is calculated and shown in the Spectra View by default (Section 9.1.3.6.3). The class average spectra are also calculated and can be shown on demand (Section 9.1.3.6.4). The same applies to the calculated noise spectrum (Section 9.1.3.6.5).

A running workflow can be canceled by clicking  or 

. This clears all views. To continue start recalibration again or close all spectra and load new classes.

To recalibrate spectra and calculate average spectra:

1. Specify the recalibration and average spectrum calculation parameters in the **Settings Spectra Preparation** dialog as desired. If spectra are loaded and you change parameters that affect spectra preparation during loading, the views become cleared. In this case, close all spectra and open the classes again.
2. From the **Data Preparation** menu, select **Recalibration**.

7.1.5 Setting up the Average Peak List

The average peak list determines the peaks to be calculated in the individual spectra. It collects all peaks that were picked on either the total average spectrum or the single spectra as well as manually edited peaks. Each peak gets an index number and is defined by its *m/z* value and integration region.

7.1.5.1 Calculating the Average Peak List

The average peak list can be calculated by automatically picking peaks on either the total average spectrum or the single spectra. Automatic calculation is based on the peak picking parameters defined in the **Settings Peak Calculation** dialog (Section 9.1.4.2). The picked peaks are indicated by gray integration regions in the Spectra View. A later manual editing of the found peaks is possible (Section 7.1.5.2).

The average peak list calculation workflow can be started manually using the **Average Peak List Calculation** command from the **Data Preparation** menu. The workflow will be run automatically if a workflow that requires an average peak list being calculated is launched without the average peak list calculation step has been performed yet.

To calculate the average peak list:

1. Specify the peak picking parameters in the **Settings Peak Calculation** dialog as desired.
2. From the **Data Preparation** menu, select **Average Peak List Calculation**.

7.1.5.2 Manually Editing the Average Peak List

The average peak list can be edited manually. You can add new peaks to the list, change existing peaks with respect to their integration region or remove peaks from the list. Moreover, a pure manual peak editing is possible as an alternative to automatic peak picking.

Manual peak editing requires the average peak list calculation workflow to be run first. This is needed even if a pure manual peak list editing should be performed. For pure manual peak editing, average peak list calculation must be run with the **Limit Peak Number** option in the **Settings Peak Calculation** dialog (Section 9.1.4.2) activated and the **Maximal Peak Number** set to '0'.

Editing peaks is also possible after peak (statistic) calculation or model generation. However, this resets the current peak calculation, indicated by the integration regions of all picked peaks change to gray color, and thus requires recalculation of peaks.

To add a new peak:

1. In the Spectra View, zoom in the peak you want to add.
2. Right click the peak and select **Add Peak** from the view's context menu. This displays the distance cursor.
3. Move the cursor lines to the positions where the peak should start and end, and click the right mouse button.
4. Confirm the dialog on adding a peak with the given integration region.
5. If the selected integration region overlaps with that of an already existing peak, confirm the message on peak adding is refused and repeat steps 2 to 4.

To change the integration region of a peak:

1. In the Spectra View, zoom in the peak you want to change.
2. Right click the peak and select **Edit Peak n** from the view's context menu. This displays the distance cursor, which marks the current integration limits.
3. Move the cursor lines to the new positions where the peak should start and end and click the right mouse button.
4. Confirm the dialog on changing the peak's integration region as stated.

To remove a peak:

1. In the Spectra View, right click the peak you want to remove and select **Remove Peak n** from the view's context menu.

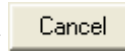
7.1.6 Calculating Peaks and Optionally Selecting Peaks for Model Generation

The peak calculation workflow calculates the peaks stored in the average peak list in the single spectra and corresponding peak statistic data as well as selects the peaks to be included in model generation. In case of two loaded model classes, also the ROC curves per peak (Section 6.4.2.2) are generated.

Peak calculation is based on the peak calculation settings in the **Settings Peak Calculation** dialog (Section 9.1.4.2). Either the peak areas, which for the GA, SVM and SNN are normalized, or the maximal peak intensities can be used. Peak selection is performed according to the settings in the **Settings Peak Selection** dialog (Section 9.1.5.1). All picked peaks are taken by default but you can define that only the best peaks according to the chosen sort mode should be selected.

The peak calculation workflow can be started manually using the **Peak Calculation** command from the **Data Preparation** menu. The workflow will be run automatically if a workflow that requires peaks being calculated is launched without the peak calculation step has been performed yet.

After peak calculation, the integration regions of the peaks selected for model generation are marked blue in the Spectra View. The 2D Peak Distribution View plots the data of the first two peaks of the peak list by default. The results of peak calculation can be viewed by setting up the Peak Statistic report (Section 8.1.1.2). The calculated statistical data (average and standard deviation, 1D peak distribution, box and whiskers) can be shown in the Spectra View on demand (Sections 9.1.3.6.7 to 9.1.3.6.9).

A running workflow can be canceled by clicking  or 

. This clears all views. To continue start peak calculation again or close all spectra and load new classes.

To calculate and optionally select peaks:

1. Specify the peak calculation and optionally the peak selection parameters in the **Settings Peak Calculation** and/or **Settings Peak Selection** dialogs as desired.
2. From the **Data Preparation** menu, select **Peak Calculation**.
3. To show the corresponding Peak Statistic report select the **Peak Statistic** command from the **Reports** menu.

7.1.7 Manually Excluding/Including a Peak

After running the peak calculation workflow all picked peaks are indicated by colored integration regions in the Spectra View. Included peaks, i.e. peaks that will be used in model generation, are indicated by blue integration regions and excluded peaks by gray ones (Section 9.1.3.6.6). You can exclude currently included peaks as well as include currently excluded peaks. The latter applies to both manually excluded peaks and peaks excluded by automatic peak selection.

Note: Exclusion/inclusion of peaks is only possible after the peak calculation workflow was run and if currently, no model generation workflow is running.

To exclude/include a peak manually:

1. In the Spectra View, right-click in the integration region of the peak you want to exclude/include and select **Exclude Peak n** resp. **Include Peak n**.

7.2 Model Generation and Validation

A classification model can be generated by applying one of the four classification algorithms supported by ClinProTools to all included peaks in the non-excluded spectra of the loaded model generation classes. The resulting model can automatically be validated internally by cross validation within the model generation workflow. An external validation can also be performed after model generation.

7.2.1 Generating a Model

You can generate a new model when at least two model generation classes are loaded and the data preparation has already been done or you have defined the settings you want to use for it. To generate a new model, you have to add a new model parameter set to the model list that defines the classification algorithm to be used and the algorithm-related model parameters. The peaks that should be included in model generation have to be determined as well as the settings for cross validation before model calculation is started. There is no fixed order in which these individual steps have to be done; the following description will start with adding a new model parameter set.

7.2.1.1 Defining the Model Generation Settings

The model generation settings define how model generation, validation and spectra classification is performed. They are specified in the **Settings Peak Selection** (Section 9.1.5.1), **Settings Genetic Algorithm** (Section 9.1.5.2.1), **Settings Support Vector Machine** (Section 9.1.5.2.2), **Settings Supervised Neural Network** (Section 9.1.5.2.3), **Settings QuickClassifier** (Section 9.1.5.2.4) and **Settings Cross Validation** (Section 9.1.5.7) dialogs. You can use the default parameters or specify own settings suitable for your data. The settings are automatically stored in the *SettingsModelGeneration.xml* file, which is loaded when ClinProTools is started and is updated on each settings change. To keep special settings you can save them in an XML file with a specified name.

7.2.1.1.1 Adding a Model Parameter Set to the Model List

To calculate a new model you have to add a new model parameter set to the model list. This includes selecting the classification algorithm, setting the algorithm-related model parameters and specifying the model name. Entering a model name is optionally but can be forced by checking the **Force Entering Model Name** option in the **General Settings** dialog (Section 9.1.1.12). The model name can still be edited after the parameter set was entered in the model list via the **Edit Model Name** command from the Model List View context menu but only as long as model calculation is not started.

To add a model parameter set:

1. From the **Model Generation** menu, select **New Model** or click 

.
2. In the **Choose Algorithm** dialog, select the classification algorithm to be used. Clicking **OK** opens the corresponding algorithm-specific **Settings [Algorithm]** dialog.
3. Depending on the chosen algorithm, define the model parameters for the:
   - GA in the **Settings Genetic Algorithm** dialog (Section 9.1.5.2.1)

- SVM in the **Settings Support Vector Machine** dialog (Section 9.1.5.2.2)
- SNN in the **Settings Supervise Neural Network** dialog (Section 9.1.5.2.3)
- QC in the **Settings QuickClassifier** dialog (Section 9.1.5.2.4)

Then click **OK**.

4. In the **Model Name** dialog (Section 9.1.5.2.5), enter the name for the new model if desired. Click **OK** to enter the new model parameters set with the specified name in the model list getting the state 'Added'. If the **Force Entering Model Name** option is active and you have not entered a model name, a message informs you that a model name is needed. Quit the message, enter a name and click **OK**.

7.2.1.1.2 Setting the Cross Validation Parameters

ClinProTools supports three kinds of cross validation (Section 6.2.3) that can be chosen in the **Settings Cross Validation** dialog (Section 9.1.5.7). It is strongly recommended to apply one kind of cross validation to verify that the obtained models give valid results on unseen data. With activated cross validation, after each model generation a final cross validation is applied. You must keep in mind that cross validation in ClinProTools requires at least 20 non-excluded spectra over all classes being available. This also applies to working with groups of spectra from multiple measurements; here at least 20 groups must be available.

You can use the default parameters or specify own settings suitable for your data. Alternatively, you can load a model generation settings file or reset the current settings to the default values.

Note: If you change the cross validation settings when models of the state 'Calculated' are present in the Models List, these models are automatically reset to the state 'Added'. In this case, model calculation has to be performed again. This ensures that all models in the list are based on the same cross validation settings.

To set the cross validation parameters:

1. From the **Model Generation** menu, select **Settings Cross Validation**.
2. In the **Settings Cross Validation** dialog, specify the parameters as desired and click **OK**. If the model list contains models of the state 'Calculated', this resets the models to the state 'Added'.

7.2.1.1.3 Saving, Loading and Resetting the Model Generation Settings

The model generation settings are automatically stored in the *SettingsModelGeneration.xml* file which is updated on each settings change. To keep the model generation settings you have adapted to special analytical tasks you can save them in an XML file with a specified name. This allows loading these settings again. Changed settings can also be reset to the defaults.

To save the current model generation settings:

1. From the **Model Generation** menu, select **Save Settings Model Generation**. This opens the **Save Model Generation Settings File** dialog with the SettingsModel-Generation folder as the default storage location.
2. Specify the file name and target folder and click **Save**.
3. If you have selected an existing file name, answer the confirmation request to overwrite the file.

To load the model generation settings:

1. From the **Model Generation** menu, select **Load Settings Model Generation**. This opens the **Load Settings Model Generation File** dialog with the SettingsModel-Generation folder being opened by default.
2. Navigate to the file you want to load. Double-click it or select it and click **Open**. This overwrites the current data preparation settings with the loaded ones.

To reset the current model generation settings to the defaults:

1. From the **Model Generation** menu, select **Reset Settings Model Generation**.
2. Confirm the appearing request to reset the current settings to the defaults.

7.2.1.2 Checking and Optionally Changing the Current Peak Selection

Before running the model generation workflow, you should check the current peak selection that was set up within the peak calculation workflow and change it if desired. All peaks with a blue integration region will be included in model generation whereas peaks with a gray integration region will be excluded.

Note: The peak selection settings may strongly influence the quality of the chosen classification algorithm. In many cases, a reasonable reduction of peaks improves the classification performed by the algorithms.

You can change the current selection by modifying the parameters in the **Settings Peak Selection** dialog (Section 9.1.5.1). For example, if you do not want to include all peaks in model generation (default setting), you can restrict the peaks to be taken; this selects only the best peaks according to the chosen sort mode. Additionally or alternatively, you can change the current selection by manually excluding/including peaks (Section 7.1.7). Moreover, you can force (a) certain peak(s) into the model to be generated (Section 7.2.1.3).

7.2.1.3 Forcing a Peak into a Model

After running the peak calculation workflow, you can force a peak into a model, which means the respective peak must be incorporated in the generated model. A forced peak is marked by a green integration region before as well as after model generation. Forcing a peak into a model can be canceled by selecting the command for the respective peak again.

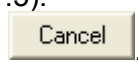
To force a peak into a model:

1. In the Spectra View, right-click in the integration region of the peak you want to force and select **Force Peak n into Model**.

7.2.1.4 Calculating a Model

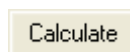
After entering a model parameter set in the model list, you can calculate a corresponding model using the **Calculate** command from the **Model Generation** menu. This command runs model calculation automatically on all models of the state 'Added' currently present in the model list. If the loaded spectra are not fully prepared when launching model calculation, first the required workflows (recalibration, average peak list calculation and/or peak calculation) are run according to the current settings.

In model generation, all or only the selected peaks of the prepared, non-excluded spectra of the loaded model generation classes are used. The peaks that separate best between the loaded classes are searched for using the chosen classification algorithm and the algorithm-related model parameters. Cross validation is performed on the model and the recognition capability is calculated if both options are not deactivated. The progress of model generation is shown in the Model List View's **State** column. The results of cross validation and recognition capability calculation are entered in the model list after model generation has finished. The model's state is changed to 'Calculated'. You can show the XML file of the calculated model (Section 7.2.1.5).

The calculation of new models can be canceled by clicking  or 

. The model's state is set back to ‘Added’.

To calculate a model:

1. From the **Model Generation** menu, select **Calculate** or click 

. This successively calculates all models of the state 'Added' contained in the model list.

7.2.1.5 Showing a Single Model

You can show a calculated model in the Model report (Section 8.1.1.6). This contains the model generation classes and all data preparation and model generation parameters that were used for setting up that model as well as the results of cross valida-

tion and recognition capability if calculated.

To show a single model:

1. In the Model List View, right-click the model you want to show and select **Show Model** or click Show.

7.2.1.6 Showing All Models in the Model List

You can show all models currently in the model list in the Model List report (Section 8.1.1.5). This includes all models of any state with corresponding parameters. The data shown depends on the models' current state.

To show all models in the model list:

1. From the **Reports** menu, select **Model List** or click Model List.

7.2.1.7 Saving a Model

If you want to keep a calculated model you can save it in an XML file with a specified name. This allows loading the model again e.g. for external validation or classification of test spectra.

To save a model:

1. In the Model List View, right-click the model you want to save and select **Save Model As** or click Save....
2. In the **Save Model** dialog, specify the file name and target folder. The ClinProt-Models folder is the default storage location.
3. Click **Save**.
4. If you have selected an existing name, answer the confirmation request to overwrite the file.

7.2.1.8 Removing a Single or All Models from the Model List

You can remove a model from the model list if you do not want to have it there any longer or clear the complete model list at once. This unloads the selected resp. all models present.

Note: Please remember that models are not saved automatically in ClinProTools. Thus, if you have calculated a new model you should first consider whether to save it (Section 7.2.1.7) before removing it.

To remove a single model from the model list:

1. In the Model List View, right-click the model you want to remove and select **Remove Model**.

To clear the model list:

1. From the **Model Generation** menu, select **Clear All** or click Clear All.

7.2.1.9 Loading a Model

A model that was saved in a XML file can be loaded in the model list again e.g. to perform external validation on it or classify unknown test spectra in it. A loaded model gets the state 'Loaded'.

To load a model:

1. From the **Model Generation** menu, select **Load Model** or click Load... This opens the **Load Model** dialog with the ClinProtModels folder opened by default.
2. Navigate to the model you want to load. Double-click it or select it and click **Open**. This enters the model in the Model List View.

7.2.2 Validating a Model Externally

You can perform an external validation on a calculated model using spectra of known class membership that have not been used in generating the respective model. For each class in the model corresponding spectra must be loaded. The validation test spectra are loaded and prepared like the model generation spectra used to generate the model and then are classified in the model. All settings for the respective workflows are stored in the model. From the confusion matrix in the Validation report (Section 8.1.1.7) you can obtain how well your current model classifies the known test spectra.

To perform an external validation on a model:

1. In the model list, select the model you want to validate. If the desired model is currently not in the model list, load it.
2. From the **Classification** menu or the Model List View context menu, select **External Validation** or click Validate....
3. In the **External Validation** dialog:
   - Select for each class present an appropriate spectra collection to be classified.

     Enter the path and name of the folder containing the collection or click Browse and navigate to the respective folder.

- Specify whether single classification reports (Section 8.1.1.8) per class should be shown.

4. Click **OK** to start the external validation workflow. This prepares and classifies the validation spectra and then shows the Validation report and the single Classification reports if created.

7.3 Spectra Classification

Within ClinProTools, unknown spectra can be classified in a classification model set up for the respective analytical task. The spectra to be classified are loaded and prepared as the spectra that were used for model generation and are classified in the model according to the data preparation and model generation parameters stored in the model. ClinProTools supports two modes for spectra classification launching different workflows.

7.3.1 Changing the Classification Mode

Spectra classification can be run in standard or in batch mode (Section 6.3) with the standard mode being active by default. If you want to run your next spectra classification in another mode than that currently active, you have to change the mode before starting the classification. Changing the mode is not possible when a classification is running or loaded.

To change the current classification mode:

1. From the **File** menu, select **General Settings**.
2. In the **Settings General** dialog, set **Classify in Batch Mode** as needed: Check the option to work in batch mode or if the mode has been previously switched to batch mode uncheck the option to work in standard mode again. Then click **OK**.

7.3.2 Selecting a Model for Spectra Classification

To classify spectra you have to select the model to be used in the model list. The model should be suitable for the data you want to analyze.

To select the model to use:

1. In the model list, select the model you want to use. If the desired model is currently not in the model list, load it.

7.3.3 Selecting the Spectra to be Classified and Running Classification

After selecting a classification model, you have to select the spectra collection to be classified and run classification. The spectra are prepared according to the data preparation parameters stored in the model and then classified based on the respective model generation parameters.

The classification workflow depends on the active classification mode:

- In the standard mode, all spectra to be classified are loaded in ClinProTools and displayed in the Spectra, Gel and Stack views; the class color is 'black'. The classifycation result is automatically shown in the Classification report (Section 8.1.1.8) and stored as *ClinProtClassification[number].xml* file. The 2D Peak Distribution View displays corresponding peak data. You can save the result with a specified name (Section 7.3.4).
- In the batch mode, no spectra are displayed in the ClinProTools GUI. After classification is finished, the **Save Classification** dialog opens to save the classification result in an XML file with a specified name. The corresponding Classification report can be shown on demand (Section 7.3.5); however, in the case of big XML files it is not recommendded to create the report.

Independent of the mode the software holds the classification as long as you do not close the classification (Section 7.3.6).

To select a spectra collection and run classification:

1. From the **Classification** menu, select **Classify**.
2. In the **Browse For Folder** dialog, navigate to the folder that contains the spectra collection to be classified and click **OK**. This runs the classification workflow corresponding to the active mode.

7.3.4 Saving the Classification Result

The classification result for the selected spectra collection can be saved in an XML file with a specified name. In the standard mode, you have to call up the saving dialog whereas in the batch mode the workflow opens that dialog automatically. Saving the result is possible as long as you do not close the classification (Section 7.3.6).

To save the classification result:

1. From the **Classification** menu, select **Save Classification** to open the **Save Classification** dialog if it is not shown automatically. This dialog opens with the ClinProtClassification folder as the default storage location.

2. Enter the file name or select one from the folder list and click **Save**. If you have selected an existing file name, answer the confirmation request to overwrite it.

7.3.5 Showing the Classification Result

The classification result can be shown in the Classification report (Section 8.1.1.8) and stored as *ClinProtClassification[number].xml* file. In the standard mode, the workflow automatically creates and shows the Classification report; you may show the result again if you closed the report. In the batch mode, the workflow does not set up the Classification report. You can create the report on demand; however, displaying big classifications is not recommended because the browser used for display might take a very long time to process the XML file with style sheet. Showing the classification result is possible as long as you do not close the current classification (Section 7.3.6).

To show the classification result:

1. From the **Classification** menu, select **Show Classification**.

7.3.6 Closing the Classification

Closing a classification removes the current classification result from the memory and in the standard mode it also unloads the classified spectra and removes them from the views.

To close the classification:

1. From the **Classification** menu, select **Close Classification** or click  .

7.4 Peak Statistic and Correlation Analysis Calculation

7.4.1 Calculating Peak Statistic

The peak statistic calculation workflow is in principle the same like the peak calculation workflow but additionally creates and shows the Peak Statistic report (Section 8.1.1.2). That report lists all peaks picked in the spectra of the loaded model generation classes with corresponding state, peak area/intensity and statistical data. By default, the peak statistic workflow uses the current peak selection settings (Section 9.1.5.1) defined for

model generation but if desired you can define different settings for reporting. For example, you can sort peaks by *m/z* value, an option that is not available with the peak selection parameters.

To calculate peak statistic with/without changing peak statistic settings:

1. If you want to change the current peak statistic settings, select **Settings Statistic** from the **Reports** menu. Otherwise processed to step 3.
2. In the **Settings Statistic** dialog uncheck **Use Selection/Sort Mode From 'Settings Peak Selection' Dialog**. Then select the desired sort mode.
   If you want to display peak statistic data in the Spectra View only for a restricted number of peaks enter the respective peak number.
   Click **Peak Statistic** to immediately show the Peak Statistic report or click **OK** to close the dialog with changing the current settings.
3. From the **Reports** menu, select **Peak Statistic** or click  . This shows the corresponding Peak Statistic report.

7.4.2 Calculating Correlation Analysis

A correlation analysis (Section 6.4.2.1) can be calculated either for all peaks resulting in setting up a correlation matrix or for a selected peak which creates a per-peak correlation list. In both cases, the correlation analysis can be calculated over either all classes or only a specified one using one of the two correlation algorithms available. The result of a correlation matrix calculation is automatically displayed in the Correlation Matrix report (Section 8.1.1.3) and stored as *ClinProtCorrelationMatrix[number].xml* file. For a per-peak correlation list calculation the Correlation List report (Section 8.1.1.4) is shown automatically after the calculation is finished and stored as *ClinProtCorrelationList[number].xml* file.

The correlation matrix workflow automatically runs the spectra recalibration, average peak list calculation and/or peak calculation workflows if these have not been performed when launching correlation matrix calculation. In contrast, per-peak correlation list calculation can be done only after peak calculation was performed; otherwise, the respective command is disabled.

To calculate a correlation matrix:

1. From the **Reports** menu, select **Correlation Matrix**. If peak calculation has not been performed yet, the required workflows are run prior to opening the **Correlation Matrix** dialog.
2. In the **Correlation Matrix** dialog, define the parameters for correlation matrix calculation and click **OK**. This calculates correlation analysis on all peaks and shows the results in the Correlation Matrix report.

To calculate a per-peak correlation list:

1. In the Spectra View, right-click the peak for which you want to calculate correlation analysis and select **Correlation List for Peak n**.
2. In the **Correlation List** dialog, define the parameters for correlation list calculation and click **OK**. This calculates correlation analysis on the selected peak and shows the results in the Correlation List report.

7.5 Performing PCA

To get more information about the variability within model generation classes and thus the homogeneity/heterogeneity of a spectra set, a PCA can be carried out within ClinProTools. In the context of PCA the separation into classes is ignored, i.e. all data is treated as one group. It is also possible to apply PCA to a single loaded class only, e.g. to detect subgroups or outliers within the model generation class. PCA can be performed on grouped spectra, too. The PCA is carried out by an external MATLAB software tool, which is started automatically within ClinProTools.

7.5.1 Calculating a PCA

A PCA is calculated on all non-excluded spectra in the loaded spectra set(s) and requires two valid spectra with three peaks being available at least. The PCA workflow automatically runs the spectra recalibration, average peak list calculation and/or peak calculation workflows if these have still not been performed when launching PCA calculation. After the PCA is completed, the PCA main window opens displaying the results.

To calculate a PCA:

1. Open the spectra set(s) on which you want to calculate a PCA.
2. If certain spectra should not be included in PCA exclude them.
3. From the **Statistical Analysis** menu, select **PCA** or click  .
4. In the **PCA** dialog, check whether normalized data should be used. Click **OK** to start PCA. If required the spectra recalibration, average peak list calculation and/or peak calculation workflows will be run prior to starting PCA.
5. View the results of the PCA.

7.5.2 Viewing PCA Results

The results of a PCA can be viewed in the Scores and Loadings plots, the Influence plot and the Variance plot. All generated data of a PCA is stored as *ClinProtPCA.xml* file in the ClinProTools folder and can be viewed by launching this file. This file will be overwritten when running a new PCA.

7.5.2.1 Scores Plots and Loadings Plots

The PCA main window displays the results of a PCA in eight different 3D and 2D Scores plots and Loadings plots (Figure 5-13). By default, the scores and loadings concern the first three PCs, PC1, PC2 and PC3, which usually explain most of the variance within in the data set. The number of calculated PCs complies with the total number of peaks in the average peak list.

Scores plots

The top row shows four Scores plots with variable axis definitions. The top left plot is a 3D plot; the following Scores plots are 2D plots which show the three selected PCs in all possible combinations. The axes of the Scores plots record arbitrary units. Within the Scores plots one point represents one spectrum and each plot contains as many points as non-excluded spectra are in the used data set(s). The points are shown in the same color like the spectra in the Spectra View; in the example shown in the figure above two model generation classes were used for PCA.

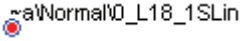
The Scores plots display for each spectrum the corresponding scores of the selected PCs. They show how the spectra are distributed in the corresponding sub-space determined by the selected PCs and visualize the relationship between different spectra, e.g. whether different groups are separated from each other or which spectra may be outliers.

Loadings plots

The bottom row shows four Loadings plots with variable axis definitions. As for the Scores plots, one 3D plot and three 2D plots are given. The Loadings plots are coupled to the Scores plots, i.e. the plots below each other belong together and refer to the same PCs. The axes of the Loadings plots record the loadings between -1 and 1. Load1 shows the loadings for PC1, Load2 the loadings for PC2, etc. Within the Loadings plots each point represents one peak and each plot contains as many points as non-excluded peaks are in the average peak list of the used data set(s). For better visualization, black crosses mark the zero axes.

The Loadings plots display for each peak the loadings the selected PCs. They show how principal components are related to the original peaks. Peaks that are far away from the central cloud are responsible for the variance within the data set.

The following operations may help you to better view single data, get more details and document results:

- Each Scores plot or Loadings plot of the PCA main window can be displayed in a separate window via the corresponding command of the **Plots** menu.
- To view the data in a Scores plot or Loadings plot in detail you can change the display of the plot by zooming, panning and rotating operations via the **Zoom**, **Pan** and **Rotate** commands of the **View** menu.
- If you want to know which spectrum corresponds to a data point in a Scores plot or which *m/z* value corresponds to a data point in a Loadings plot, just click the desired point with the left mouse button. This marks the selected data point with file () resp. *m/z* () description. Clicking a marked data point again removes the information from that point. To (un)mark data points the **Mark Data Points** command of the **View** menu must be active (default setting).
- The scores and loadings of PC1, PC2 and PC3 are plotted against each other by default. If you want to view the data of another PC set, you can change the PC selection via the **PCs** command of the **PC** menu. Each PC change updates all plots in the PCA main window accordingly but previously set up single plot windows remain unchanged.
- A graphic of the content of the PCA main window or a single Scores plot or Loadings plot window can be copied to the clipboard via the **Copy** command of the **Edit** menu. This allows pasting the graphic into an appropriate application.
- All PCA data generated during the current PCA are stored as *ClinProtPCA.xml* file in the ClinProTools folder. This includes the calculated variances, scores and coefficients. Double-clicking that file displays the respective data. The file is overwritten on each PCA run.

7.5.2.2 Influence Plot

The Influence plot (Figure 5-15) shows which influence a spectrum has on the current PCA model. It is based on a certain number of most variant PCs you have to specify to set up the plot via the **Influence** command in the **Plots** menu. Each data point in this plot represents a spectrum included in the PCA model.

The Influence plot is a diagnostic tool for the identification of outliers. The vertical axis is a measure how far away a spectrum is from the model space (distance to model). The horizontal axis is a measure how far away a spectrum is from the model center after being projected into the model space and thus of the leverage of a spectrum to the model. Strong outliers have high leverage on the model, i.e. strong "power" to pull the PCA model toward themselves, and may "consume" one PC just because of their existence. (The term *leverage* derives from the Archimedean principle that anything can be lifted out of balance as long as the lifter has a long enough lever).

The display of the Influence plot can be changed by zooming and panning operations, data points can be marked with spectrum information and the content of the plot can be copied to the clipboard like described for the Scores and Loadings plots (Section 7.5.2.1). If you want to set up the Influence plot for a different number of PCs, again select the **Influence** command and enter the desired number of PCs.

7.5.2.3 Variance Plot

The Variance plot (Figure 5-16) can be displayed via the **Variance** command in the **Plots** menu. It displays the explained variance in percent contributed by the single given PCs. The blue curve starting from the first bar demonstrates the accumulated variance from PC to PC. The number of PCs concerned in the plot depends on in the investigated data. Basically, the plot displays as many PCs as are needed to explain at least 95% of the variance within the data set but it is limited to displaying ten PCs at most. This may result in the sum-of-explained-variance curve not reaching the 95% mark in each case.

7.6 Performing Unsupervised Clustering

To get more information about the variability within model generation classes and thus the homogeneity/heterogeneity of a spectra set, an unsupervised hierarchical clustering can be carried out within ClinProTools. In the context of unsupervised clustering the separation into classes is ignored, i.e. all data is treated as one group. It is also possible to apply unsupervised clustering to a single loaded class only, e.g. to detect subgroups or outliers within the model generation class. Unsupervised clustering can be performed on grouped spectra, too. The unsupervised clustering is carried out by an external MATLAB software tool, which is started automatically within ClinProTools.

7.6.1 Calculating an Unsupervised Clustering

An unsupervised clustering is calculated on all non-excluded spectra in the loaded spectra set(s) and requires three valid spectra with three peaks being available at least. The unsupervised clustering automatically runs the spectra recalibration, average peak list calculation and/or peak calculation workflows if these have still not been performed when launching unsupervised clustering calculation. After the unsupervised clustering is completed, the Dendrogram window opens displaying the created dendrogram.

To calculate an unsupervised clustering:

1. Open the spectra set(s) you want to cluster.

2. If certain spectra should not be included in unsupervised clustering exclude them.
3. From the **Statistical Analysis** menu, select **Unsupervised Clustering** or click  .
4. In the **Unsupervised Clustering** dialog, specify the parameters as desired and click **OK** to start unsupervised clustering. If required the spectra recalibration, average peak list calculation and/or peak calculation workflows will be run prior to starting unsupervised clustering.
5. View the resulting dendrogram.

7.6.2 Viewing the Unsupervised Clustering Result

The result of an unsupervised hierarchical clustering of spectra can be viewed in the Dendrogram window (Figure 5-17). The created dendrogram shows the clusters calculated from the spectra (= classes) and the distances among the single clusters. Depending on the clustering parameter settings used the created dendrogram shows either the full tree of spectra with/without spectra paths or is limited to a specified number of clusters the spectra were assigned to.

The corresponding data is stored in the follwing files in the ClinProTools folder:

- *ClinProtClustering.xml* contains a list of classes with all the spectra paths of the spectra belonging to the classes (only if number of classes is limited). This file will display automatically after the clustering is completed if the corresponding option is set in the **Unsupervised Clustering** dialog and the **Create Full Tree** option is unchecked.
- *ClinProtClustering.tree.xml* exports all the nodes with distances of the hierarchical clustering in the form of a linkage list.
- *ClinProtClustering.tree2.xml* exports the XML tree, each node is represented by a node element, which contains two sub node elements.

8 REPORTING DATA

ClinProTools offers various types of reports to report specific spectra, peak statistic, model, validation or classification data as well as error information. Open reports can be saved and printed. Graphics from views/plots can be printed or copied to the clipboard. Peak list data can be exported to XML or CART format.

8.1 Creating ClinProtTools Reports

ClinProTools offers various types of reports for showing specific data (Section 8.1.1). All ClinProTools reports are created as XML files except the Error report which is of the *.txt format. They can be opened with either the Microsoft Internet Explorer or Excel; the respective application can be chosen in the **General Settings** dialog (Section 9.1.1.12). Multiple reports can be open at a time.

All XML files contain style sheet references, which transform them into HTML when opened with a web browser (Microsoft Internet Explorer 6.0 is strongly recommended). The referenced style sheet must be in the same folder as the XML file. To ensure that Excel parses the XML files with style sheet properly, make sure that a dot is used as decimal separator in Excel. To enforce this go to the 'Tools/Option' dialog in Excel. On the 'International' tab at 'Number handling' uncheck 'Use system separators', enter a dot as 'Decimal separator' and a comma as 'Thousands separator'. If this is not set, numbers may be parsed as dates and the like.

The XML files are stored with a consecutively numbered default name (e.g. *'ClinProt-Statistic0001.xml'*, *'ClinProtValidation0001.xml'*) in the ClinProTools folder. The corresponding style sheets (suffix *.xsl) have been installed there by the setup. These XML files will stay in this folder as long as you do not delete them either by automatically removing all temporary XML files (Section 4.3) or manually removing selected ones. All files can also be saved via the browsers **Save As** command (Section 8.1.2). If you like to store the files at another location it is advisable to store a copy of the style sheet there, too.

A previously created report can be shown again by double-clicking its file name in the Windows Explorer. This applies to both the reports automatically saved by the system and the reports manually saved with a specified name.

Note: If the XML file you want to show is stored at another location than in the ClinProTools folder, make sure that a copy of the corresponding style sheet (suffix *.xsl) is stored in this folder, too.

Concerning problems with empty XML tables, please see the installation notes (Section 2.2). To avoid the BRUKER logo to be displayed in Excel (which is wrong positioned

due to an error in Excel), check 'Hide all' in 'Objects' in the 'View' tab in the 'Tools/ Options' dialog.

Details of a report will be shown by hovering with the mouse over the table items (Figure 8-1). To fit a page to the window in Internet Explorer the text size can be changed via the **Text Size** command from the Explorer's **View** menu or with the mouse wheel while holding down the Ctrl button.

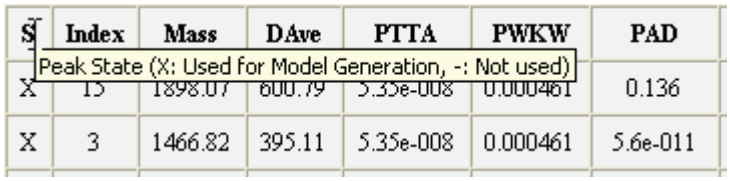


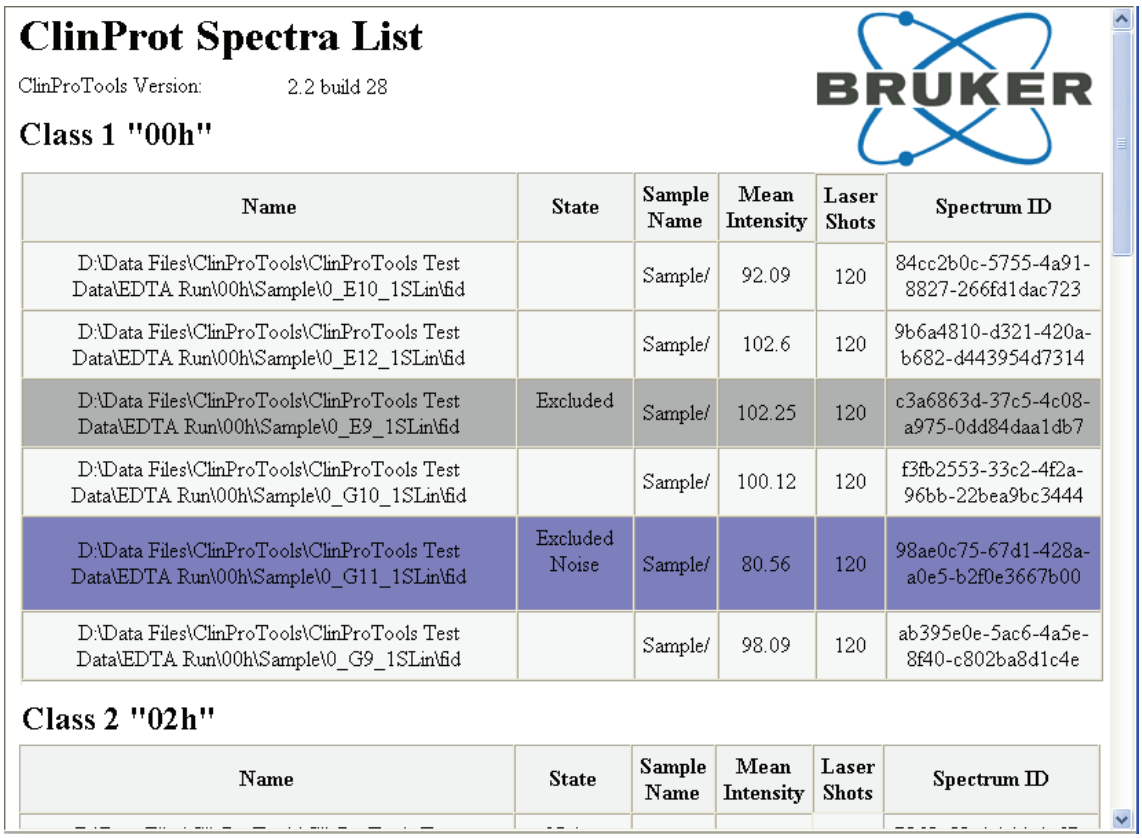
Figure 8-1 Displaying details of a report (here for the item 'S')

8.1.1 ClinProTools Report Types

8.1.1.1 Spectra List Report

The Spectra List report *ClinProtSpectra.xml* (Figure 8-2) is created and shown using the **Spectra List** command from the **Reports** menu. This report lists all loaded spectra with corresponding data grouped according to their class membership. The following data is displayed for each spectrum:

Column	Description
Name	Path and name of the spectrum.
State	Inclusion/exclusion state of the spectrum. Excluded spectra are marked by an 'Excluded' entry. In case of exclusion by filters, the filter is given (e.g. 'Excluded Noise'). The rows of excluded spectra are colored according to the reason of exclusion (the color code is the same as used in the Gel View, Section 9.1.3.7.3).
Sample Name	Sample name.
Mean Intensity	Average intensity before TIC normalization.
Laser Shots	Number of shots.
Spectrum ID	ID of the spectrum.
Groups	Grouping of spectra; available when the **Support Spectra Grouping** option in the **Settings Spectra Preparation** dialog is enabled.



ClinProt Spectra List

ClinProTools Version: 2.2 build 28

Class 1 "00h"

Name	State	Sample Name	Mean Intensity	Laser Shots	Spectrum ID
D:\Data Files\ClinProTools\ClinProTools Test Data\EDTA Run\00h\Sample\0_E10_1SLin\fid		Sample/	92.09	120	84cc2b0c-5755-4a91-8827-266fd1dac723
D:\Data Files\ClinProTools\ClinProTools Test Data\EDTA Run\00h\Sample\0_E12_1SLin\fid		Sample/	102.6	120	9b6a4810-d321-420a-b682-d443954d7314
D:\Data Files\ClinProTools\ClinProTools Test Data\EDTA Run\00h\Sample\0_E9_1SLin\fid	Excluded	Sample/	102.25	120	c3a6863d-37c5-4c08-a975-0dd84daa1db7
D:\Data Files\ClinProTools\ClinProTools Test Data\EDTA Run\00h\Sample\0_G10_1SLin\fid		Sample/	100.12	120	f3fb2553-33c2-4f2a-96bb-22bea9bc3444
D:\Data Files\ClinProTools\ClinProTools Test Data\EDTA Run\00h\Sample\0_G11_1SLin\fid	Excluded Noise	Sample/	80.56	120	98ae0c75-67d1-428a-a0e5-b2f0e3667b00
D:\Data Files\ClinProTools\ClinProTools Test Data\EDTA Run\00h\Sample\0_G9_1SLin\fid		Sample/	98.09	120	ab395e0e-5ac6-4a5e-8f40-c802ba8d1c4e

Class 2 "02h"

Name	State	Sample Name	Mean Intensity	Laser Shots	Spectrum ID

Figure 8-2 Spectra List report (section)

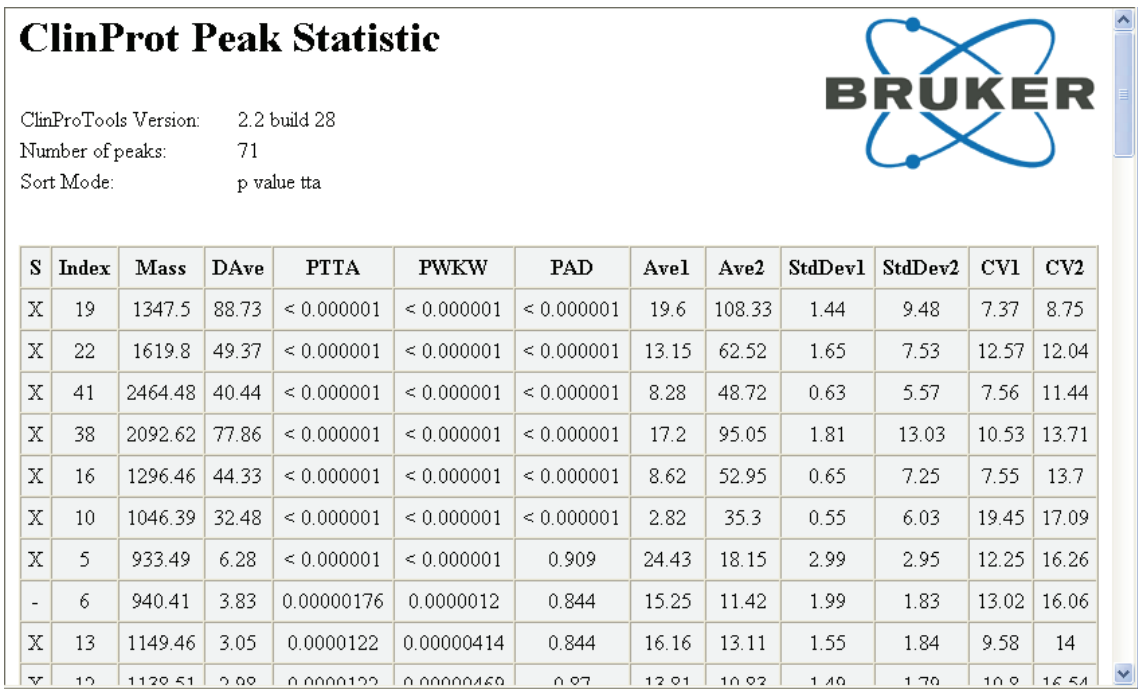
8.1.1.2 Peak Statistic Report

The Peak Statistic report *ClinProtStatistic.xml* (Figure 8-3) is created and shown using the **Peak Statistic** command from the **Reports** menu or by clicking  . This report shows a table with all peaks picked in peak calculation along with several values. The total number of peaks and the used sort mode are shown above the table.

The following data is displayed for each peak:

Column	Description
S	Inclusion/exclusion state of the peak: 'X' used for model generation (= included); '-' not used for model generation (= excluded).
Index	Peak index.
Mass	*m/z* value.
DAve	Difference between the maximal and the minimal average peak area/ intensity of all classes.

Column	Description
PTTA	P-value of t-test (2 classes; Section 6.4.1.1) or ANOVA test (> 2 classes; Section 6.4.1.2), range 0..1; 0: good, 1: bad. Preferable for normal distributed data.
PWKW	P-value of Wilcoxon test (2 classes; Section 6.4.1.3) or Kruskal-Wallis test (> 2 classes; Section 6.4.1.4), range 0..1; 0: good, 1: bad. Preferable for not normal distributed data.
PAD	P-value of Anderson-Darling test (Section 6.4.1.5); gives information about normal distribution; range 0..1; 0: not normal distributed, 1: normal distributed.
Ave*N*	Peak area/intensity average of class *N.*
StdDev*N*	Standard deviation of the peak area/intensity average of class *N.*
CV*N*	Coefficient of variation in % of class *N*.



ClinProt Peak Statistic

ClinProTools Version: 2.2 build 28
Number of peaks: 71
Sort Mode: p value tta

S	Index	Mass	DAve	PTTA	PWKW	PAD	Ave1	Ave2	StdDev1	StdDev2	CV1	CV2
X	19	1347.5	88.73	< 0.000001	< 0.000001	< 0.000001	19.6	108.33	1.44	9.48	7.37	8.75
X	22	1619.8	49.37	< 0.000001	< 0.000001	< 0.000001	13.15	62.52	1.65	7.53	12.57	12.04
X	41	2464.48	40.44	< 0.000001	< 0.000001	< 0.000001	8.28	48.72	0.63	5.57	7.56	11.44
X	38	2092.62	77.86	< 0.000001	< 0.000001	< 0.000001	17.2	95.05	1.81	13.03	10.53	13.71
X	16	1296.46	44.33	< 0.000001	< 0.000001	< 0.000001	8.62	52.95	0.65	7.25	7.55	13.7
X	10	1046.39	32.48	< 0.000001	< 0.000001	< 0.000001	2.82	35.3	0.55	6.03	19.45	17.09
X	5	933.49	6.28	< 0.000001	< 0.000001	0.909	24.43	18.15	2.99	2.95	12.25	16.26
-	6	940.41	3.83	0.00000176	0.0000012	0.844	15.25	11.42	1.99	1.83	13.02	16.06
X	13	1149.46	3.05	0.0000122	0.00000414	0.844	16.16	13.11	1.55	1.84	9.58	14
X	12	1138.51	2.98	0.0000122	0.00000469	0.87	13.91	10.83	1.49	1.79	10.8	16.54

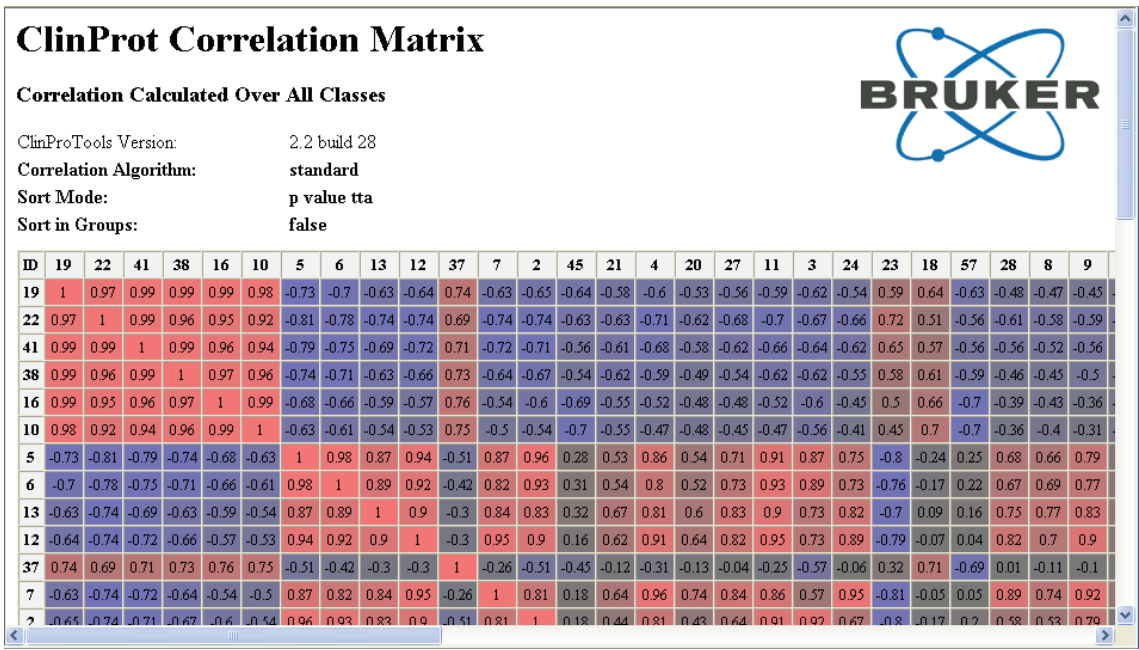
Figure 8-3 Peak Statistic report (section)

8.1.1.3 Correlation Matrix Report

The Correlation Matrix report *ClinProtCorrelationMatrix.xml* (Figure 8-4) is created and shown using the **Correlation Matrix** command from the **Reports** menu. The correlation analysis (Section 6.4.2.1) is based on the settings defined in the **Correlation Matrix** dialog.

The report shows the correlation matrix that lists all peak pairs with their calculated correlation coefficient (cc) ranging from +1 to -1. In addition, a color code ranging from red (cc = +1) to blue (cc = -1) is used to highlight different cc ranges in the matrix. This allows quickly detecting highly correlated peak pairs. Whether correlation was calculated over all classes or a specified one is reported above the table as well as the used correlation algorithm (standard or Kendall's tau), the sort mode and whether sorting in groups was performed.

Note: To get the indices and masses that correspond to a table cell hover with the mouse over the table cell of interest.



ClinProt Correlation Matrix

Correlation Calculated Over All Classes

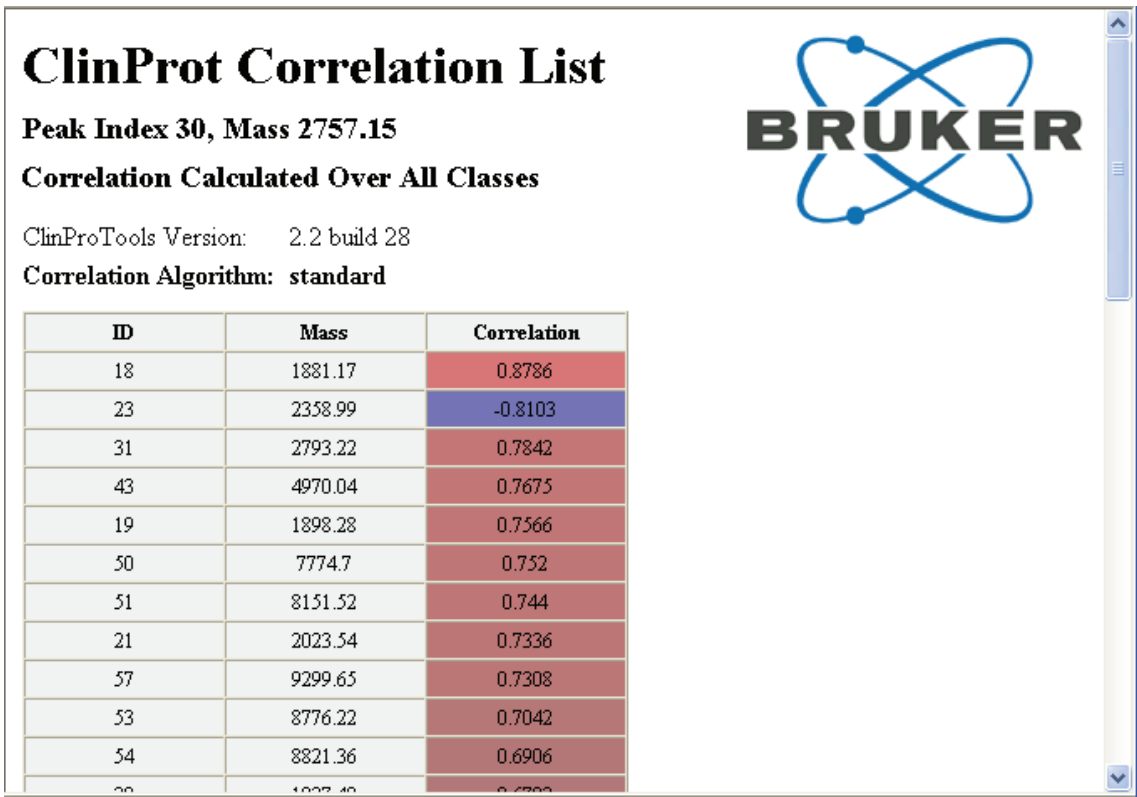
ClinProTools Version:	2.2 build 28
Correlation Algorithm:	**standard**
Sort Mode:	**p value tta**
Sort in Groups:	**false**

ID	19	22	41	38	16	10	5	6	13	12	37	7	2	45	21	4	20	27	11	3	24	23	18	57	28	8	9
19	1	0.97	0.99	0.99	0.99	0.98	-0.73	-0.7	-0.63	-0.64	0.74	-0.63	-0.65	-0.64	-0.58	-0.6	-0.53	-0.56	-0.59	-0.62	-0.54	0.59	0.64	-0.63	-0.48	-0.47	-0.45
22	0.97	1	0.99	0.96	0.95	0.92	-0.81	-0.78	-0.74	-0.74	0.69	-0.74	-0.74	-0.63	-0.63	-0.71	-0.62	-0.68	-0.7	-0.67	-0.66	0.72	0.51	-0.56	-0.61	-0.58	-0.59
41	0.99	0.99	1	0.99	0.96	0.94	-0.79	-0.75	-0.69	-0.72	0.71	-0.72	-0.71	-0.56	-0.61	-0.68	-0.58	-0.62	-0.66	-0.64	-0.62	0.65	0.57	-0.56	-0.56	-0.52	-0.56
38	0.99	0.96	0.99	1	0.97	0.96	-0.74	-0.71	-0.63	-0.66	0.73	-0.64	-0.67	-0.54	-0.62	-0.59	-0.49	-0.54	-0.62	-0.62	-0.55	0.58	0.61	-0.59	-0.46	-0.45	-0.5
16	0.99	0.95	0.96	0.97	1	0.99	-0.68	-0.66	-0.59	-0.57	0.76	-0.54	-0.6	-0.69	-0.55	-0.52	-0.48	-0.48	-0.52	-0.6	-0.45	0.5	0.66	-0.7	-0.39	-0.43	-0.36
10	0.98	0.92	0.94	0.96	0.99	1	-0.63	-0.61	-0.54	-0.53	0.75	-0.5	-0.54	-0.7	-0.55	-0.47	-0.48	-0.45	-0.47	-0.56	-0.41	0.45	0.7	-0.7	-0.36	-0.4	-0.31
5	-0.73	-0.81	-0.79	-0.74	-0.68	-0.63	1	0.98	0.87	0.94	-0.51	0.87	0.96	0.28	0.53	0.86	0.54	0.71	0.91	0.87	0.75	-0.8	-0.24	0.25	0.68	0.66	0.79
6	-0.7	-0.78	-0.75	-0.71	-0.66	-0.61	0.98	1	0.89	0.92	-0.42	0.82	0.93	0.31	0.54	0.8	0.52	0.73	0.93	0.89	0.73	-0.76	-0.17	0.22	0.67	0.69	0.77
13	-0.63	-0.74	-0.69	-0.63	-0.59	-0.54	0.87	0.89	1	0.9	-0.3	0.84	0.83	0.32	0.67	0.81	0.6	0.83	0.9	0.73	0.82	-0.7	0.09	0.16	0.75	0.77	0.83
12	-0.64	-0.74	-0.72	-0.66	-0.57	-0.53	0.94	0.92	0.9	1	-0.3	0.95	0.9	0.16	0.62	0.91	0.64	0.82	0.95	0.73	0.89	-0.79	-0.07	0.04	0.82	0.7	0.9
37	0.74	0.69	0.71	0.73	0.76	0.75	-0.51	-0.42	-0.3	-0.3	1	-0.26	-0.51	-0.45	-0.12	-0.31	-0.13	-0.04	-0.25	-0.57	-0.06	0.32	0.71	-0.69	0.01	-0.11	-0.1
7	-0.63	-0.74	-0.72	-0.64	-0.54	-0.5	0.87	0.82	0.84	0.95	-0.26	1	0.81	0.18	0.64	0.96	0.74	0.84	0.86	0.57	0.95	-0.81	-0.05	0.05	0.89	0.74	0.92
2	-0.65	-0.74	-0.71	-0.67	-0.6	-0.54	0.96	0.93	0.83	0.9	-0.51	0.81	1	0.18	0.44	0.81	0.43	0.64	0.91	0.92	0.67	-0.8	-0.17	0.2	0.58	0.53	0.79

Figure 8-4 Correlation Matrix report (section)

8.1.1.4 Correlation List Report

The Correlation List report *ClinProtCorrelationList.xml* (Figure 8-5) is created and shown using the **Correlation List for Peak n** command from the Spectra View context menu. The correlation analysis (Section 6.4.2.1) is based on the settings defined in the **Correlation List** dialog; it can be calculated over either all classes or only a specified one. The report lists the correlation coefficients (cc) that were calculated by comparing the selected peak (given by its index and *m/z* value above the list) to each other peak in the peak list. Whether correlation was calculated over all classes or a specified one is reported above the table as well as the used correlation algorithm (standard or Kendall's tau). The single compared peaks are ordered with decreasing absolute correlation value. Like in the Correlation Matrix report (Section 8.1.1.3) a color code ranging from red (cc = +1) to blue (cc = -1) is used to highlight different cc ranges in the list. This allows quickly detecting highly correlated peak pairs.



ClinProt Correlation List

Peak Index 30, Mass 2757.15

Correlation Calculated Over All Classes

ClinProTools Version: 2.2 build 28

Correlation Algorithm: standard

ID	Mass	Correlation
18	1881.17	0.8786
23	2358.99	-0.8103
31	2793.22	0.7842
43	4970.04	0.7675
19	1898.28	0.7566
50	7774.7	0.752
51	8151.52	0.744
21	2023.54	0.7336
57	9299.65	0.7308
53	8776.22	0.7042
54	8821.36	0.6906

Figure 8-5 Correlation List report (section)

8.1.1.5 Model List Report

The Model List report *ClinProtModelList.xml* (Figure 8-6) is created and shown using the **Model List** command from the **Reports** menu or by clicking Model List. It lists the parameters of all loaded models in a table.

The following data is displayed for each model:

Column	Description
Name	Model name.
Algo	Classification algorithm used.
Validation	Results from cross validation (overall and for each class) and recognition capability calculation. If cross validation could not be calculated due to not enough spectra this is indicated by 'Insuf.' under **XVal**.
GA Param	GA-specific parameter settings (filled for GA models only).
Best Peaks	Peak detection mode (filled for SVM, SNN and QC models).
SNN Param	SNN-specific parameter settings (filled for SNN models only).
QC Param	QC-specific parameter settings (filled for QC models only).
KNN Param	Number of neighbors in k-NN classification (filled for GA, SVM and SNN models).
Column	Description
X Val Param	Cross validation parameter settings.
Date/Time	Date and time of model calculation.
GUID	Globally unique identifier of the model.

ClinProt Model List

ClinProTools Version: 2.2 build 28

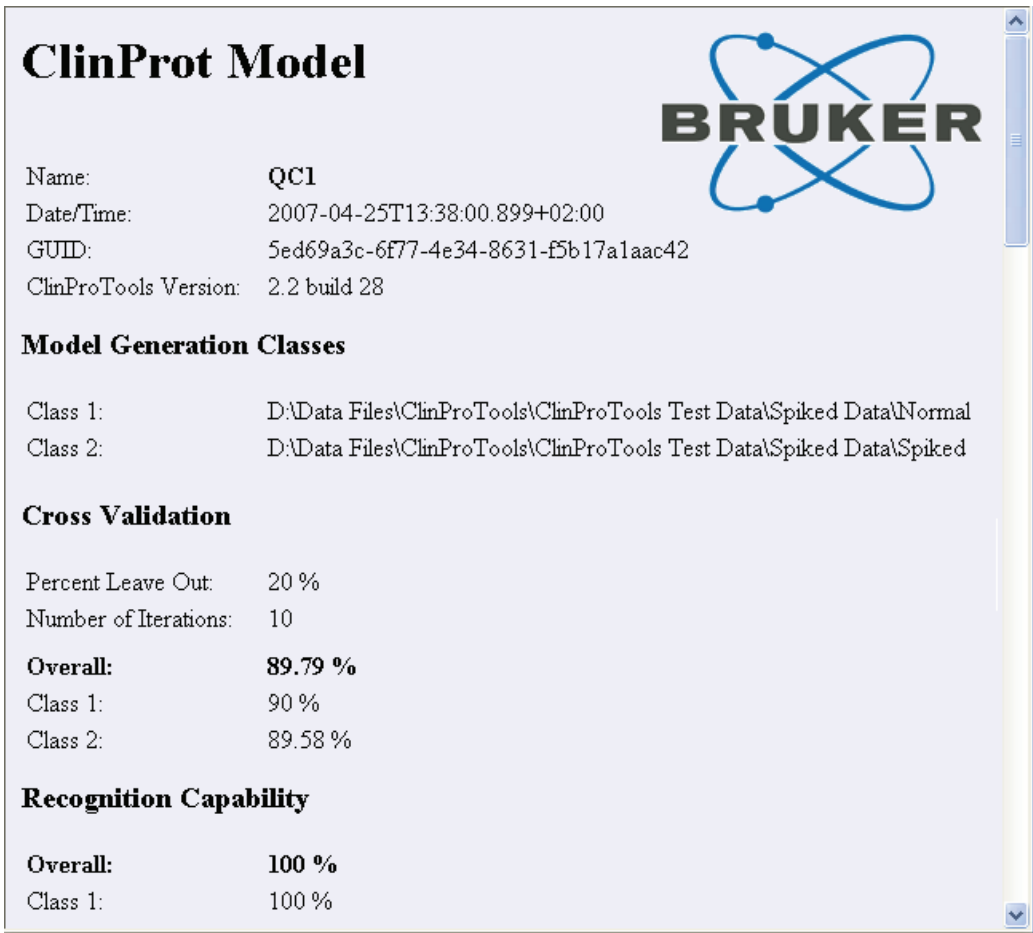
BRUKER

Name	Algo	Validation				GA Param							Best Peaks		SNN Param			QC Param	KNN Param	X
		XVal	X1	X2	Rec Cap	Max NBP	Max Gen	Auto NPC	Num PCs	Mut Rate	Cross Rate	Var RS	Auto BPD	Num BstP	UL Cycl	Auto NPTr	Num PrT	Sort Mode	Num OfNN	Mode
GA1	GA	95.5 %	100 %	90.9 %	100 %	5	50	true		0.2	0.5	false							3	random
SVM1	SVM	95.5 %	100 %	90.9 %	100 %								true						3	random
SNN1	SNN	100 %	100 %	100 %	100 %								true		1000	true				random
QC1	QC	89.8 %	90 %	89.6 %	100 %								true					p value tta		random

Figure 8-6 Model List report (section)

8.1.1.6 Model Report

The Model report *ClinProtModel.xml* (Figure 8-7) is created and shown for the selected model using the **Show Model** command from the Model List View context menu or by clicking Show. It lists all model generation classes, parameters and results of the model. If the single spectra peak picking approach was used the recognition scores of all the spectra matched against the overall average peak list are also stored in the model and listed in the report. The same applies to the recognition scores of the spectra of the different classes matched against the corresponding per-class average peak list.



ClinProt Model

BRUKER

Name:	QC1
Date/Time:	2007-04-25T13:38:00.899+02:00
GUID:	5ed69a3c-6f77-4e34-8631-f5b17a1aac42
ClinProTools Version:	2.2 build 28

Model Generation Classes

Class 1:	D:\Data Files\ClinProTools\ClinProTools Test Data\Spiked Data\Normal
Class 2:	D:\Data Files\ClinProTools\ClinProTools Test Data\Spiked Data\Spiked

Cross Validation

Percent Leave Out:	20 %
Number of Iterations:	10
Overall:	**89.79 %**
Class 1:	90 %
Class 2:	89.58 %

Recognition Capability

Overall:	**100 %**
Class 1:	100 %

Figure 8-7 Model report (section)

8.1.1.7 Validation Report

The Validation report *ClinProtValidation.xml* (Figure 8-8) is created and shown after performing a validation using the **External Validation** command from the **Classification** menu or by clicking Validate... . It contains the external validation results. The last part of the table contains the confusion matrix. This gives an overview how validation data have been classified. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. A perfect classification would have entries only on the diagonal, which means that all validation data have been classified to their own class.

The following data is displayed for each class:

Column	Description
Class	Classes in the model (numbers 1 ..*N*).
Name	Path and name of the validation spectrum/spectra collection assigned to the respective class.
Correct Classified Part of Valid Spectra	Percentage of correctly classified part of valid spectra per class.
N	Number of spectra classified in predicted class.
0	Number of unclassified spectra.
Inv.	Number of invalid spectra (at present only filled by 'not recalibratable' spectra).

ClinProt Validation

Date/Time: 2007-04-26T08:42:36.075+02:00
ClinProTools Version: 2.2 build 28

Class	Name	Correct Classified Part of Valid Spectra	1	2	0	Inv.
1	D:\Data Files\ClinProTools\ClinProTools Test Data\Spiked Data\External Validation\Normal	100 %	3	0	0	0
2	D:\Data Files\ClinProTools\ClinProTools Test Data\Spiked Data\External Validation\Spiked	100 %	0	3	0	0

Figure 8-8 Validation report

8.1.1.8 Classification Report

The Classification report *ClinProtClassification.xml* (Figure 8-9) is created and shown after performing a classification using the **Classify** command from the **Classification** menu or by clicking 

. It contains the classification per spectrum in a table. If the single spectra peak picking approach was used additionally the recognition scores of the classified spectra matched against the overall average peak list as well as against the per-class average peak lists are given. Classification reports are also created and shown when the **Show Single Classifications** option is checked for external validation (Section 9.1.6.2). For each class a separate report is set up.

ClinProt Classification

BRUKER

Spectra Collection Path: **D:\Data Files\ClinProTools\ClinProTools Test Data\Spiked Data\To Classify (5 + 5)**
Model Name: **QC1**
Date/Time: 2007-09-14T11:35:57.437+02:00
ClinProTools Version: 2.2 build 65

Index	Name	Classified	Class	Class1	Class2	State	Score	Score1	Score2
1	D:\Data Files\ClinProTools\ClinProTools Test Data\Spiked Data\To Classify (5 + 5) \0_L15_1SLin_N\fid	true	1	1.97	0.03		0.202	0.167	0.215
2	D:\Data Files\ClinProTools\ClinProTools Test Data\Spiked Data\To Classify (5 + 5) \0_L17_1SLin_N\fid	true	1	1.99	0.01		0.215	0.186	0.185
3	D:\Data Files\ClinProTools\ClinProTools Test Data\Spiked Data\To Classify (5 + 5) \0_L19_1SLin_N\fid	true	1	1.97	0.03		0.197	0.152	0.184
4	D:\Data Files\ClinProTools\ClinProTools Test Data\Spiked Data\To Classify (5 + 5) \0_M19_1SLin_N\fid	true	1	1.96	0.04		0.236	0.218	0.195
5	D:\Data Files\ClinProTools\ClinProTools Test Data\Spiked Data\To Classify (5 + 5) \0_M20_1SLin_N\fid	true	1	1.99	0.01		0.223	0.194	0.194
6	D:\Data Files\ClinProTools\ClinProTools Test Data\Spiked Data\To Classify (5 + 5) \0_N14_1SLin_S\fid	true	2	0.16	1.84		0.196	0.125	0.215
7	D:\Data Files\ClinProTools\ClinProTools Test Data\Spiked Data\To Classify (5 + 5) \0_N17_1SLin_S\fid	true	2	0.05	1.95		0.215	0.128	0.233

Figure 8-9 Classification report for spectra when a QC model was used which is based on single spectra peak calculation and without information from the ClinProtRobot (section)

The following data is displayed for each spectrum:

Column	Description
Index	Order of spectra loading.
Name	Path and name of the spectrum.
Classified	Whether or not the algorithm was able to classify the spectrum.
Class	Estimated class.
Class *N*	To which extent the spectra belongs to class *N*. > 1: good, = 1: average, < 1: bad (for QC only)
State	Spectrum state information.
Score	Recogniton score of the classified spectrum matched against the overall average peak list (for single spectra peak picking only).
Score N	Recognition score of the classified spectrum matched against the per-class average peak list (for single spectra peak picking only).

The following items are only generated for spectra with information from ClinProtRobot:

Column	Description
Sample Name	Sample name.
Sample ID	One-to-one ID.
Sample Group	Group membership like *disease* or *normal.*
Sample Type	Kind of the sample.
Patient	Patient name.
Comment	Arbitrary comment.
Source	Source plate name and position.
Processing Step	Pipetting step.
Preparation Method	Pipetting method.

8.1.1.9 Error Report

The ClinProt Error report *ClinProtError.txt* (Figure 8-10) is created and shown using the **Show Error** command from the Model List View context menu. This command is enabled if the selected model has the state 'ERROR'. A message informs you about where and why the error has occurred and what you can do.

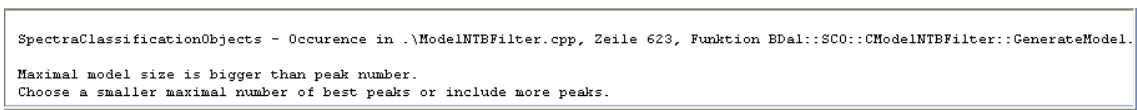


Figure 8-10 Error report: Example of an error message

8.1.2 Saving a Report

Each report set up in ClinProTools is automatically saved in an XML file with a consecutively numbered default name and stored in the ClinProTools folder. If desired you can save a shown report via the browser's **Save As** command either in the ClinProTools folder or at another location.

Note: If you like to store the files at another location it is advisable to store a copy of the corresponding style sheet there, too.

8.1.3 Printing a Report

You can print a shown report using the browser's **Print** command.

8.2 Printing a Graphic of a Data Plotting View

You can print a graphic of the current content of the selected data plotting view w/o previewing it.

Note: If the lines in a printout from the Spectra, 2D Peak Distribution or ROC Curve View are too thin, you can use the **Display Mode > 2 Pixel** and **3 Pixel** commands from the view's context menu to display and thus print thicker lines.

Note: It may be advisable to limit the printer's resolution to 300 dpi. For example, a resolution of 600 dpi produces four times the number of data as a resolution of 300 dpi does and a resolution of 1200 dpi even produces sixteen times the number. Thus, printing will take (much) longer or even may be stopped when using a higher resolution.

Note: Printing a graphic of the Gel or Stack View may take much time. Alternatively, you can copy the graphic to the clipboard with only the **Bitmap to Clipboard** command being active, paste it into e.g. Microsoft Paint or PowerPoint and then print it from there.

To print a graphic of a data plotting view:

1. Select the view of which you want to print a graphic.
2. Depending on whether or not you want preview the graphic proceed as follows:
   - To preview the graphic: Select **Print Preview** from the **File** menu. This sets up the graphic in the preview window as it would be printed. If you want to print the graphic now, click **Print** and proceed to step 3. Otherwise, close the preview window.

- To directly print the graphic: Select **Print** from the **File** menu or click  or press the keys Ctrl+P.

3. In the **Print** dialog, select the printer and print options and click **OK**.

8.3 Copying a Graphic of a Data Plotting View, a PCA Plot or a Dendrogram

You can copy a graphic of the focused data plotting view, PCA plot or dendrogram to the clipboard in order to paste the graphic into an appropriate application.

ClinProTools copies graphics of data plotting views as a bitmap with a resolution of 800*600 pixels by default. Alternatively, ClinProTools can copy graphics as a metafile with a resolution of 8000*6000 pixels. You can define whether a bitmap and/or a metafile are/is copied via the **Bitmap to Clipboard** and **Metafile to Clipboard** commands from the **Edit** menu. If both types are activated, the program that pastes the clipboard's contents into its document determines which of these formats it uses (Microsoft Paint uses bitmap by default; Microsoft Word, Excel and PowerPoint prefer metafile). Whereas the high resolution of the metafile format offers superior graphics quality, some programs (e.g. Microsoft Word) can get extremely sluggish due to the amount of data when a Gel View is copied as metafile. By selecting 'Tools' > 'Options' > 'View' > 'Show picture place holders' from Microsoft Word's menu, you can avoid the redisplay of the graphics on every move.

Graphics of PCA plots (entire PCA main window, single Sores plot, single Loadings plot, Influence plot) and dendrograms are copied as metafiles by the MATLAB tool.

To copy a graphic of a data plotting view:

1. Select the view of which you want to copy a graphic to the clipboard.
2. Depending on which type(s) of graphic should be used activate the **Bitmap to Clipboard** and/or **Metafile to Clipboard** command(s) from the **Edit** menu.
3. From the **Edit** menu of the ClinProTools window, select **Copy**. Or click  or press the keys Ctrl+C. This copies the corresponding graphic(s) to the clipboard.
4. Paste the graphic into the desired application. If two graphics are on the clipboard, the pasting application will take the appropriate one.

To copy a graphic of a PCA plot or dendrogram:

1. From the **Edit** menu of the PCA window or Dendrogram window, select **Copy** to copy a graphic to the clipboard.
2. Paste the graphic into the desired application.

8.4 Exporting the Peak List to XML or CART Format

After running peak calculation, specific data of the current peak list can be exported to XML or CART format to use the data in downstream applications.

For XML export, three different XML formats are supported (Appendix A.4). The XML2 Files format differs from the XML Files format in that it additionally provides the class and spectra paths as attributes. The XML3 Files format is similar to the XML2 Files format but a style sheet reference for ClinProtPeakList.xsl is added to facilitate working with peak lists in Excel.

As an alternative, the peak area resp. the intensity data of the peak list can be exported to the CART (ASCII) format (*.dat; by Salford Systems, San Diego, CA, USA) (Appendix A.4). For each spectrum, the class membership and the areas resp. intensities of the picked peaks given by their *m/z* value are exported.

To export the peak list to XML or CART format:

1. From the **File** menu, select **Peak List Export**. This opens the **Peak List Export** dialog.
2. Navigate to the folder where you want to save the exported file.
3. Enter a name for the exported file or select one from the folder list.
4. In **Files of Type**, select the desired format, **XML Files**, **XML2 Files**, **XML3 Files** or **CART Files**.
5. Click **Save**. If you have chosen an existing file name, confirm the appearing message to overwrite the file.

9 REFERENCE PART

The following sections describe the ClinProTools menus (Section 9.1) and context menus (Section 9.2) as well as the MATLAB based menus (Section 9.3).

9.1 ClinProTools Menus

9.1.1 File Menu

The **File** menu offers the following commands (Figure 9-1):

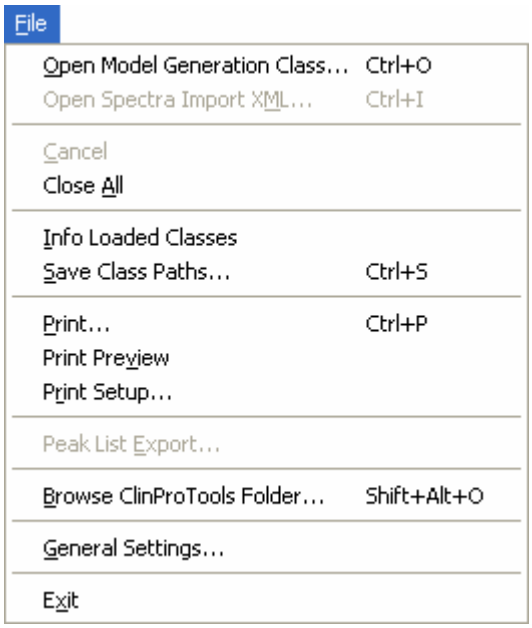


Figure 9-1 File menu

Command	Used to …
Open Model Generation Class	Open the selected spectra collection for model generation.
Open Spectra Import XML	Open the selected spectra import XML file and load the referenced spectra.
Cancel	Cancel any current loading/calculation/model generation/classification process.

Command	Used to …
Close All	Close and unloads all spectra and models.
Info Loaded Classes	Show path information about the loaded spectra collections.
Save Class Paths	Save the paths of the loaded model generation classes as spectra import XML file.
Print	Print a graphic of the active data plotting view.
Print Preview	Preview the graphic to be printed for the active data plotting view.
Print Setup	Set up the printer and printing options.
Peak List Export	Export the peak list in XML or CART format.
Browse ClinProTools Folder	Browse the ClinProTools folder.
General Settings	Define general ClinProTools settings.
Exit	Close ClinProTools.

9.1.1.1 Open Model Generation Classes Command

The **Open Model Generation Class** command is used to open a model generation class. ClinProTools loads all spectra in a folder and its subfolders recursively as one model generation class. ClinProTools supports loading spectra of the X-Mass, BAF und ASCII (Appendix A.4) file formats. For loading ASCII file formats, the null spectra exclusion filter (Section 6.1.3.2) has to be disabled. For model generation you have to load two classes at least; single classes can be loaded for peak statistic operations including PCA. One class can be opened at a time.

The command opens the **Browse For Folder** dialog (Figure 9-2). Navigate to the folder of the class to be opened, select it and click **OK**. This loads and prepares the selected spectra and displays them in the Spectra View and Gel/Stack View. If the **Check Memory on Load** option (Section 9.1.1.12) is set, first the available memory is checked against the memory needed to load the spectra. If it is insufficient, a warning message appears which asks you whether to continue. You have to repeat the loading procedure for each class of interest. The first loaded collection is referred to as 'class 1:' in the ClinProTools title bar, the second as 'class 2:', etc.

Shortcuts

Button: 

Keys: Ctrl+O

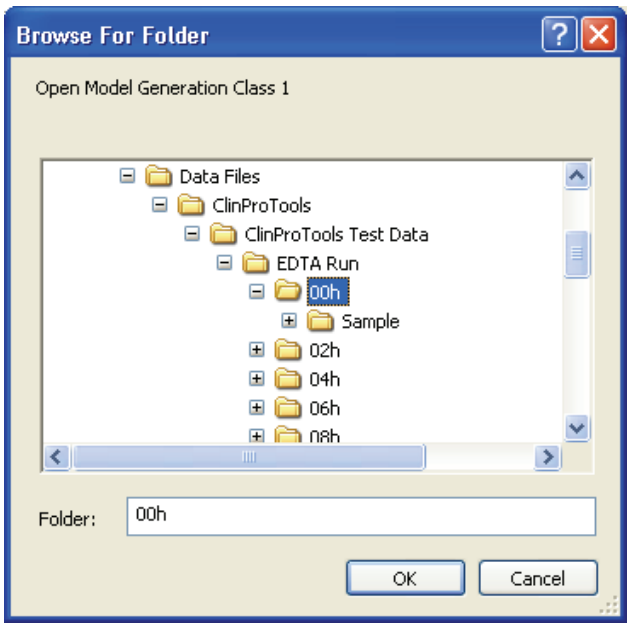


Figure 9-2 Browse For Folder dialog for opening a model generation class

9.1.1.2 Open Spectra Import XML Command

The **Open Spectra Import XML** command is used to open a *ClinProtSpectraImport.xml* and load the referenced spectra accordingly. The spectra import XML format can hold either a path list of spectra or only class paths (Appendix A.4). In ClinProTools, spectra import XML files can be saved via the **Save Class Paths** command.

The command opens the **Open Spectra Import XML** dialog with the ClinProtSpectraImport folder opened by default. Navigate to the file you want to load and click **Open**. Loading and preparation of spectra is performed as described with the **Open Model Generation Class** command.

After opening a spectra import XML file no additional spectra import XML file can be loaded. However, you can add further spectra collections via the **Open Model Generation Class** command.

Shortcuts

Button: 

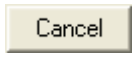
Keys: Ctrl+I

9.1.1.3 Cancel Command

The **Cancel** command cancels any currently running spectra loading, recalibration, peak calculation, model generation or classification process. The effect of canceling depends on the running process. When canceling a

- Spectra loading process the model generation class currently being loaded as well as all previously loaded classes are unloaded. You have to start loading classes again.
- Data preparation process (spectra recalibration, peak calculation) the data plotting views become temporarily cleared. You have to run the canceled process again to redraw the data in the views. Alternatively, if you do not want continue processing, you can select the **Close All** command from the **File** menu.
- Model generation, validation or classification process the data plotting views remain unchanged. You can start the respective process again.

Shortcut

Button:  or 

9.1.1.4 Close All Command

The **Close All** command closes and unloads all spectra and models in order to load new model generation classes. Confirm the request to close all spectra.

Shortcut

Button: 

9.1.1.5 Info Loaded Classes Command

The **Info Loaded Classes** command shows path information about the loaded spectra collections. All classes are listed in an automatically opened information box with their corresponding paths numbered with respect to their loading order (Figure 9-3).

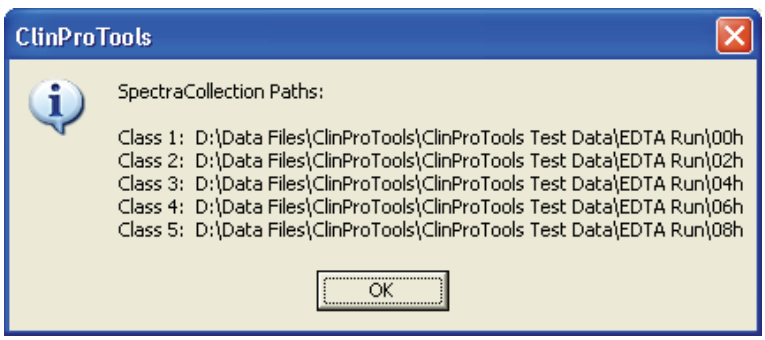


Figure 9-3 Path information about loaded spectra classes

9.1.1.6 Save Class Paths Command

The **Save Class Paths** command is used to save the paths of the currently loaded model generation classes as *ClinProtSpectraImport.xml* (Appendix A.4). This allows loading the referenced spectra via the **Open Import Spectra XML** command.

The command opens the **Save Class Paths as Spectra Import XML** dialog with the ClinProtSpectraImport folder as the default storage location. Enter the file name or select one from the folder list and click **Save**. If you have selected an existing file name, answer the confirmation request to overwrite the file.

Shortcut

Keys: Ctrl+S

9.1.1.7 Print Command

The **Print** command is used to print a graphic of the content of the active data plotting view. The command opens the **Print** dialog to specify the printer and printing options and start printing.

Note: It may be advisable to limit the printer's resolution to 300 dpi. For example, a resolution of 600 dpi produces four times the number of data as a resolution of 300 dpi does and a resolution of 1200 dpi even produces sixteen times the number. Thus, printing will take (much) longer or even may be stopped when using a higher resolution. Where in the **Print** dialog the resolution can be set depends on the respective printer.

Note: Printing a graphic of the Gel or Stack View may take much time. Alternatively, you can copy the graphic with only the **Bitmap to Clipboard** command (Section 9.1.2.3) being active to the clipboard, paste it into e.g. Microsoft Paint or PowerPoint and then print it from there.

Shortcuts

Button: 

Keys: Ctrl+P

9.1.1.8 Print Preview Command

The **Print Preview** command previews the graphic for the active data plotting view as it would be printed. The graphic is set up in the ClinProTools preview window. You can click the preview's **Print** button to print the graphic now; otherwise, close the preview.

9.1.1.9 Print Setup Command

The **Print Setup** command is used to set up the printer and printing options. The command opens the **Print Setup** dialog.

9.1.1.10 Peak List Export Command

The **Peak List Export** command is used to export specific peak list data to the CART or XML format. For the XML export, three different XML formats are available (Appendix A.4).

The command opens the **Peak List Export** dialog to select the export format (**XML Files**, **XML2 Files**, **XML3 Files** or **CART Files**) and specify the file name and target folder. Clicking **Save** exports the peak list data in the selected format.

9.1.1.11 Browse ClinProTools Folder Command

The **Browse ClinProTools Folder** command browses the ClinProTools folder C:\BDAL\ClinProTools_2_2\Files.

Shortcuts

Button: 

Keys: Shift+Alt+O

9.1.1.12 General Settings Command

The **General Settings** command is used to define general, non-algorithm settings for ClinProTools. The general settings are saved to the *SettingsGeneral.xml* file, which also stores file open paths, statistic and correlation settings. The command also allows resetting the settings saved in the *SettingsGeneral.xml* file to the defaults as well as removing all temporary XML files from the ClinProTools folder. The command opens the **Settings General** dialog (Figure 9-4).

Show Tables With…

Choose whether you want to view most of the *ClinProt*.xml* files with the browser specified in **Browser** or with Excel.

Browser. Uses the browser selected in **Browser**.

Excel (2002 or Higher). Uses Excel.

Note: You need Excel 2002 or newer, because the older versions do not support XML with style sheets. When Excel starts, check the option "Open the file

with the following style sheet applied ...".
The Excel security settings (extras/options/securities/macro security) must be set to 'low'.
To avoid the BRUKER logo to be displayed in Excel (which is wrong positioned due to an error in Excel), check 'Hide all' in 'Objects' in the 'View' tab in the 'Tools/Options' dialog.

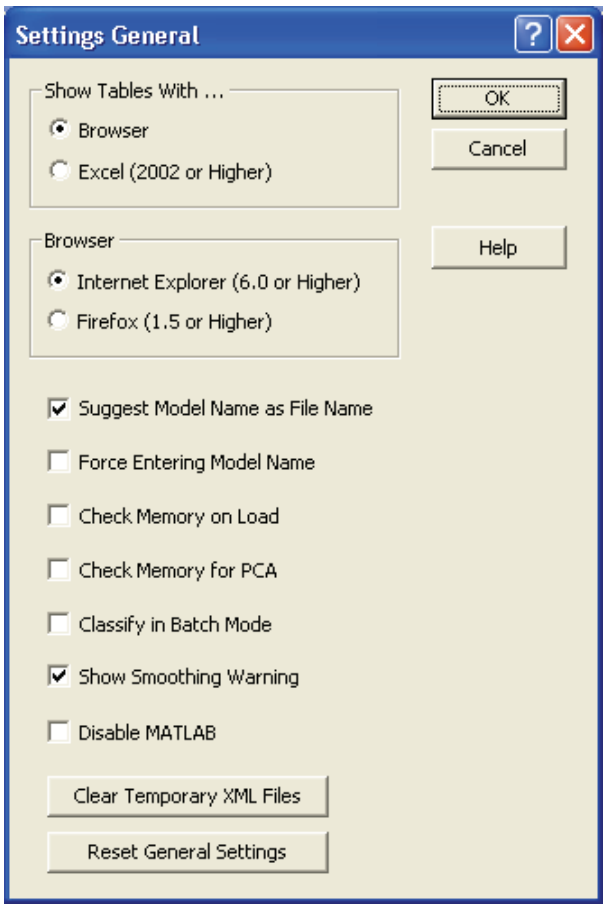


Figure 9-4 Settings General dialog (default setting)

Browser

If **Browser** is selected in **Show Tables With** select the browser you want to use:

Internet Explorer (6.0 or Higher). Uses Internet Explorer. Requires version 6.0 or higher installed.

Firefox (1.5 or Higher). Uses Firefox. Requires version 1.5 or higher installed.

Suggest Model Name as File Name

Check this option if the name entered during adding a new model to the model list should be suggested as *'ModelName.xml'* in the **Save Model As** dialog (Section 9.2.9.19).

Force Entering Model Name

Check this option if a new model should be added to the model list only if a model name has been entered.

Check Memory on Load

Check this option if it should be checked on spectra loading whether the available memory is sufficient to load the selected spectra. If the needed memory size exceeds the available memory, the machine might slow down or come to a standstill. The memory size is rated as insufficient when "needed memory x 2 > available memory". In this case, a warning message is launched which asks you whether to continue. You can set an option to skip this message in future.

Check Memory for PCA

Check this option if it should be checked on PCA start whether the available memory is sufficient to perform PCA on the loaded data set(s). If the needed memory size exceeds the available memory, a warning message is launched.

Classify in Batch Mode

Check this option to activate the batch mode and uncheck it to activate the standard mode for classification (Section 6.3).

Clear Temporary XML Files

Removes all temporary *ClinProt*.xml* and *ClinProt*.txt* files from the ClinProTools folder after you have confirmed the corresponding request.

Reset General Settings

Resets the current general settings including file open paths, statistic and correlation settings to defaults after you have confirmed the corresponding request.

Show Smoothing Warning

Check this option if a smoothing warning should appear when you select peak picking on Single Spectra (**Settings Peak Calculation** dialog) but smoothing is currently not enabled (**Settings Spectra Preparation** dialog).

Disable MATLAB

Check this option if MATLAB should be disabled. Checking this option is not recommend since without MATLAB it is impossible to run single spectra peak picking, PCA and unsupervised clustering.

9.1.1.13 Exit Command

The **Exit** command is used to close ClinProTools. Confirm the confirmation request to quit ClinProTools.

Shortcut

Button: application's 

9.1.2 Edit Menu

The **Edit** menu (Figure 9-5) offers the following commands:

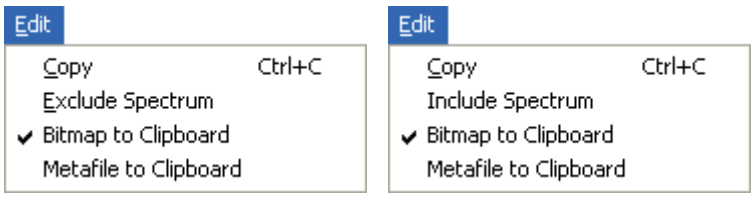


Figure 9-5 Edit menu when the current spectrum is included (left) or excluded (right)

Command	Used to …
Copy	Copy a bitmap and/or a metafile graphic of the selected data plotting view to the clipboard according to the states of the graphic format commands.
Exclude/Include Spectrum	Exclude/Include the selected spectrum manually.
Bitmap to Clipboard	Activate/Deactivate the bitmap format for copying graphics to the clipboard.
Metafile to Clipboard	Activate/Deactivate the metafile format for copying graphics to the clipboard.

9.1.2.1 Copy Command

The **Copy** command copies a graphic of the selected data plotting view to the clipboard. This allows pasting that graphic into another application. By default, ClinProTools copies graphics as a bitmap with a resolution of 800*600 pixels. Alternatively, ClinProTools can set up a metafile with a resolution of 8000*6000 pixels. Whether a bitmap and/or a metafile are/is created depends on the settings of the **Bitmap to Clipboard** and **Metafile to Clipboard** commands from the **Edit** menu. If both types are activated, the program that pastes the clipboard's contents into its document determines which of these formats it uses.

Shortcuts

Button: 
Keys: Ctrl+C

9.1.2.2 Exclude/Include Spectrum Command

The **Exclude Spectrum** command excludes the selected spectrum from further processing. The **Include Spectrum** command includes the selected, manually or automatically, spectrum. The command available depends on the spectrum's current state. Spectra can be excluded or included only before any spectra processing (e.g. recalibration, peak calculation) is performed.

In the Spectra and Stack views, all excluded spectra are displayed darker colored than the included spectra of the same class, e.g. in dark red instead of light red (Figure 9-6 top). In the Gel View and the Spectra List report (Section 8.1.1.1) manually excluded spectra are highlighted by dark gray bars (Figure 9-6 bottom) when the **Gel/Stack View > Colored Spectrum State** command from the **View** menu is active; the automatically excluded spectra are then colored according to the reason of exclusion.

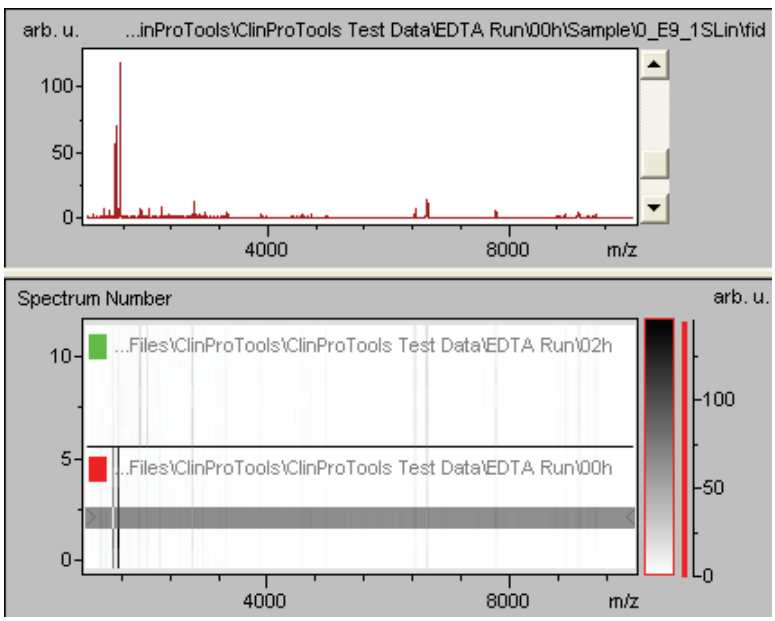


Figure 9-6 Display of a manually excluded spectrum: in the Spectra View (top) the spectrum is indicated by a dark red instead of a light red color, in the Gel View (bottom) it is highlighted with a dark gray bar

9.1.2.3 Bitmap to Clipboard Command

The **Bitmap to Clipboard** command defines that a bitmap graphic of the selected data plotting view should be copied to the clipboard when using the **Copy** command. A bitmap graphic is copied with a resolution of 800*600 pixels. This format is active by default.

9.1.2.4 Metafile to Clipboard Command

The **Metafile to Clipboard** command defines that a metafile graphic of the selected data plotting view should be copied to the clipboard when using the **Copy** command. A metafile graphic is copied with resolution of 8000*6000 pixels. Whereas the high resolution of the metafile format offers superior graphics quality, some programs (e.g. Microsoft Word) can get extremely sluggish due to the amount of data when a Gel View graphic is copied as metafile. By selecting 'Tools' > 'Options' > 'View' > 'Show picture place holders' from Word's menu, you can avoid the redisplay of the graphics on every move.

9.1.3 View Menu

The **View** menu (Figure 9-7) offers the following commands:

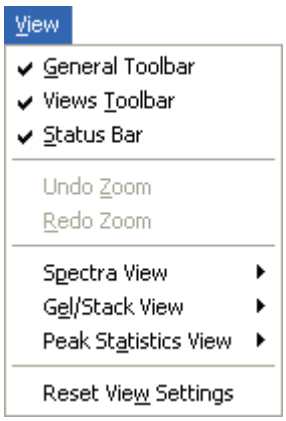


Figure 9-7 View menu

Command	Used to …
General Toolbar	Show/Hide the General toolbar.
View Toolbar	Show/Hide the View toolbar.
Status Bar	Show/Hide the status bar.

Command	Used to …
Undo Zoom	Undo last zooming operation.
Redo Zoom	Redo previously undone zooming operation.
Spectra View	Pop up commands for showing data in the Spectra View.
Gel/Stack View	Pop up commands for showing data the Gel/Stack View.
Peak Statistics View	Pop up commands for showing data in the Peak Statistics View.
Reset View Settings	Reset certain settings of the data plotting views to the defaults.

9.1.3.1 General Toolbar Command

The **General Toolbar** command shows/hides the General toolbar. The General toolbar is shown by default.

9.1.3.2 View Toolbar Command

The **View Toolbar** command shows/hides the View toolbar. The View toolbar is shown by default.

9.1.3.3 Status Bar Command

The **Status Bar** command shows/hides the status bar. The status bar is shown by default.

9.1.3.4 Undo Zoom Command

ClinProTools stacks the zooming operations you perform in the Spectra, Gel, 2D Peak Distribution or Single Peak Variance View for each view separately. The **Undo Zoom** command undoes the last change in zoom range performed in the currently focused view.

Shortcut

Button: 

9.1.3.5 Redo Zoom Command

The **Redo Zoom** command restores the previously undone zoom range in the currently focused view.

Shortcut

Button: 

9.1.3.6 Spectra View Popup Command

Pointing to **Spectra View** offers the following commands (Figure 9-8):

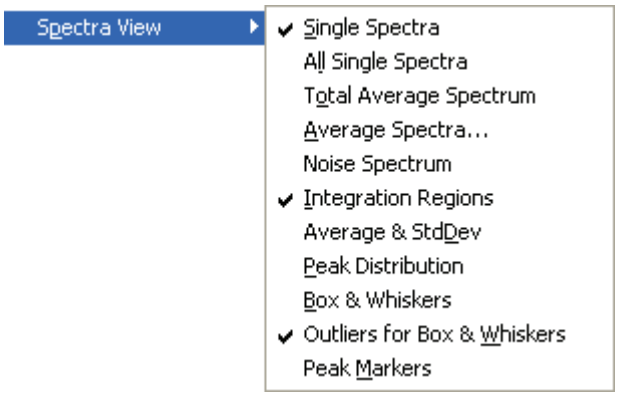


Figure 9-8 Spectra View submenu

Command	Used to …
Single Spectra	Show/Hide the single spectra with one single spectrum displayed at a time.
All Single Spectra	Show/Hide the simultaneous overlaid display of all single spectra.
Total Average Spectrum	Show/Hide the total average spectrum.
Average Spectra	Show/Hide the average spectra for (a) specified class(es).
Noise Spectrum	Show/Hide the noise spectrum.
Integration Regions	Show/Hide the integration regions of picked peaks.
Average & StdDev	Show/Hide the class averages of peak area/intensity with standard deviation.
Peak Distribution	Show/Hide the peak distribution plot with respect to peak area/intensity.
Box & Whiskers	Show/Hide the box & whiskers plots for the peak area/intensity per class.

Command	Used to …
Outliers for Box & Whiskers	Show/Hide the outliers for box & whiskers plots for the peak area/intensity per class.
Peak Markers	Show/Hide the markers for the peak(s) selected in the 2D Peak Distribution, ROC Curve or Single Peak Variance View.

9.1.3.6.1 Spectra View > Single Spectra Command

The **Single Spectra** command shows/hides single spectra of the loaded classes in the Spectra View with displaying only one single spectrum at a time. All single spectra within a class are displayed in the same predefined class color; excluded spectra are shown with a darker color than the included ones. As an alternative, you can use the **Spectra View > All Single Spectra** command to display all single spectra simultaneously and overlaid. The separately displayed single spectra are shown by default.

Shortcut

Button: 

9.1.3.6.2 Spectra View > All Single Spectra Command

The **All Single Spectra** command shows/hides single spectra of the loaded classes in the Spectra View with displaying all single spectra simultaneously and overlaid. All single spectra within a class are displayed in the same predefined class color; excluded spectra are shown with a darker color than the included ones. As an alternative, you can use the **Spectra View > Single Spectra** command to display only one single spectrum at a time.

9.1.3.6.3 Spectra View > Total Average Spectrum Command

The **Total Average Spectrum** command shows/hides the total average spectrum in the Spectra View. The total average spectrum (Section 6.1.1.4) is calculated from all non-excluded spectra within the spectra recalibration process. It is displayed in gray color and is shown by default.

Shortcut

Button: 

9.1.3.6.4 Spectra View > Average Spectra Command

The **Average Spectra** command is used to show/hide class average spectra in the Spectra View. Class average spectra are calculated within the spectra recalibration workflow. They are displayed in a darker color than the corresponding single spectra (e.g. in dark red instead of red). The class average spectra are hidden by default.

The command opens the **Display of Averages** dialog (Figure 9-9) to choose the class(es) for which the corresponding average spectrum should be shown.

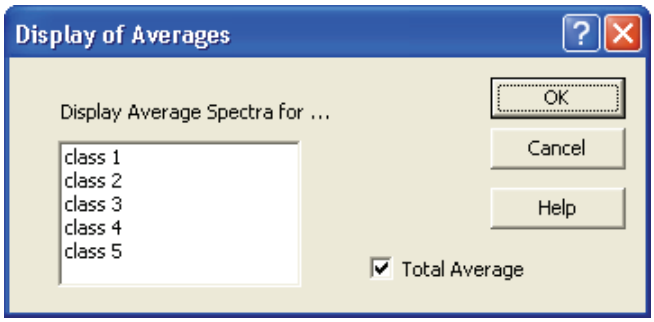


Figure 9-9 Display of Averages dialog

Display Average Spectra for …

Lists all loaded classes. To show class average spectra select the respective class(es) from this list. If a currently shown class average spectrum should be hidden again, deselect the respective class in this list.

Total Average

Check this option to display the total average spectrum (same as **Spectra View > Total Average Spectrum** command from **View** menu).

OK

Shows the average spectrum/spectra for the selected class(es) in the Spectra View. Previously shown class average spectra are hidden if not longer contained in the class selection.

Shortcut

Button: 

9.1.3.6.5 Spectra View > Noise Spectrum Command

The **Noise Spectrum** command shows/hides the calculated noise spectrum used in average spectrum calculation in the Spectra View. The noise spectrum is displayed in orange color and is hidden by default.

9.1.3.6.6 Spectra View > Integration Regions Command

The **Integration Regions** command shows/hides the integration regions of the picked peaks in the Spectra View (Figure 9-10). The integration regions are highlighted with different colors concerning the current state of the peak. Non-excluded peaks are indicated in blue and excluded ones in gray. Peaks incorporated in the selected calculated model are marked red. Peaks forced into a model are green highlighted before and after model generation as well. The integration regions are shown by default.

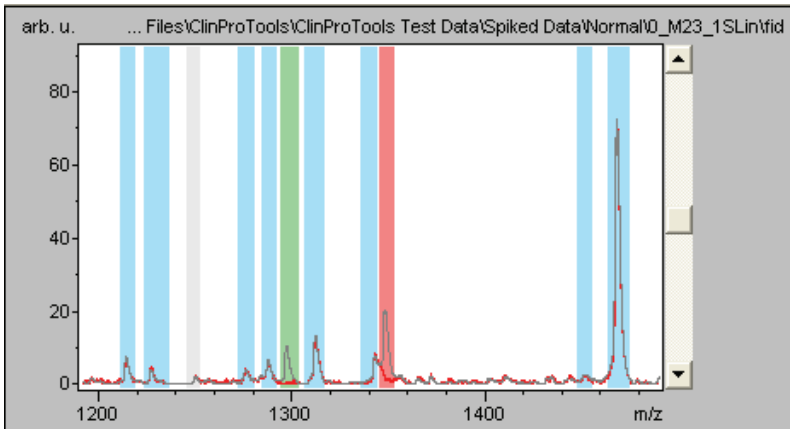


Figure 9-10 Coloring of integration regions of peaks that are included (blue), excluded (gray), forced into the model (green) and incorporated in the model (red)

Shortcut

Button: 

9.1.3.6.7 Spectra View > Average & StdDev Command

The **Average & StdDev** command shows/hides the per-class average with standard deviation plots in the Spectra View (Figure 9-11). These represent the calculated average of the peaks areas/intensities in the single spectra of a class with the corresponding standard deviation on both sides. These bars are colored like the corresponding class. The plot is drawn on a unique scale independent of the peak intensity scale. These bars are hidden by default.

Shortcut

Button: 

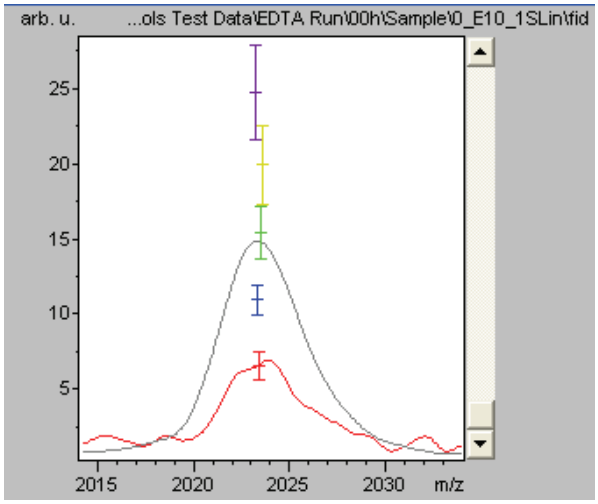


Figure 9-11 Average with standard deviation bars marking the peak area/intensity averages of five classes

9.1.3.6.8 Spectra View > Peak Distribution Command

The **Peak Distribution** command shows/hides the 1D distribution of the peak areas/ intensities in the Spectra View (Figure 9-12). The 1D peak distribution plots the areas/ intensities of the respective peak in the single spectra of the loaded classes as separate values. Like in the 2D peak distribution, values of peaks from different classes are displayed with different predefined symbols (e.g. cross, circle) that are colored according to the respective class color. The plot is drawn on a unique scale independent of the peak intensity scale. The 1D peak distribution is hidden by default.

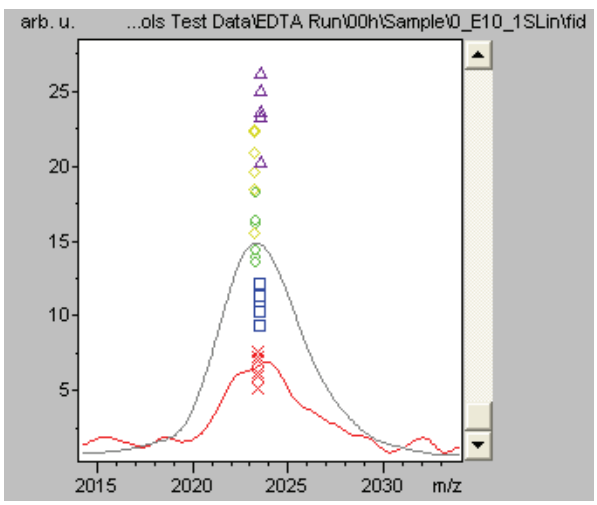


Figure 9-12 1D peak distribution plotting the single areas/intensities for the respective peak in the spectra of five classes

Shortcut

Button: 

9.1.3.6.9 Spectra View > Box & Whiskers Command

The **Box & Whiskers** command shows/hides the per-class box & whiskers plots for the peak area/intensity in the Spectra View (Figure 9-13). In this standard box plot, the top and bottom end marks of the plot, the so-called whiskers, indicate the maximum and minimum peak area/intensity within a given class. The box indicates the 25%-quartile (bottom) and the 75%-quartile (top) and the horizontal intersection denotes the median. 50% of the values fall into this interquartile range and the whiskers give you an impression of how much the remaining 50% of the values spread.

Outliers are not indicated in the standard box plot. You can display modified box & whiskers plots showing outliers by also activating the **Spectra View > Outliers for Box & Whiskers** command.

The box & whiskers plots give a graphic representation of homogeneity of the areas of a certain peak in the spectra of one class. They allow assessment of the quality of the peaks in a model. A peak where the box & whiskers of the individual classes are well separated with only minimal overlap of the whiskers is better suited for classification than a peak with overlapping box & whiskers. The plot is drawn on a unique scale independent of the peak intensity scale. The box & whiskers plots are hidden by default.

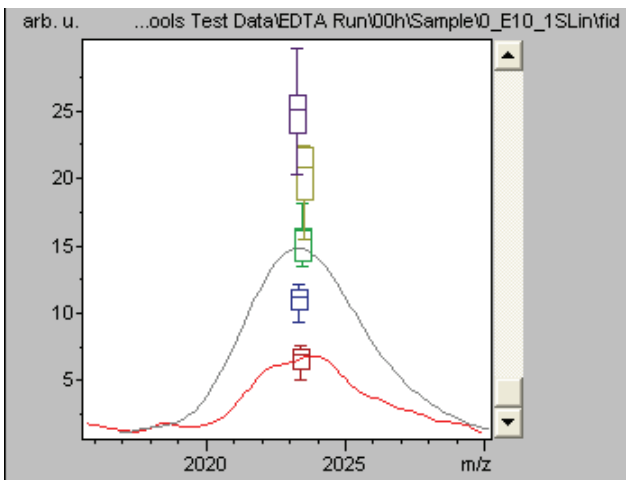


Figure 9-13 Standard box & whiskers plot calculated from the areas/intensities of the respective peak in the spectra of five classes

Shortcut

Button: 

9.1.3.6.10 Spectra View > Outliers for Box & Whiskers Command

The **Outliers for Box & Whiskers** command shows/hides the outliers for the per-class box & whiskers plots in the Spectra View (and also in the Single Peak Variance View) when the **Spectra View > Box & Whiskers** command is active. This toggles the box & whiskers plots between the standard box plot (command not active) and the modified box plot (command active) (Figure 9-14).

The modified box plot differs from the standard box plot with respect to the meaning of the displayed whiskers and the additional display of outliers which are measured values that do not fall inside the whiskers. In the modified box plot, the end mark of the top whisker denotes the value that is calculated by adding 1.5 times the interquartile range (range between 25%- and 75% quartile) to the largest measured value which is smaller than or equals the 75%-quartile. The end mark of the bottom whisker denotes the value that is calculated by subtracting 1.5 times the interquartile range from the smallest measured value which is larger than or equals the 25%-quartile. Thus, ca. 95% of all measured values are inside the whiskers if the whisker length is 1.5 times the interquartile range. All measured values that are larger or smaller, respectively, than the end marks of the whiskers are indicated as outliers. Outliers are denoted in the modified box plot by symbols which correspond to the symbols used for the peaks of the respective class in the 1D peak distribution (Section 9.1.3.6.8).

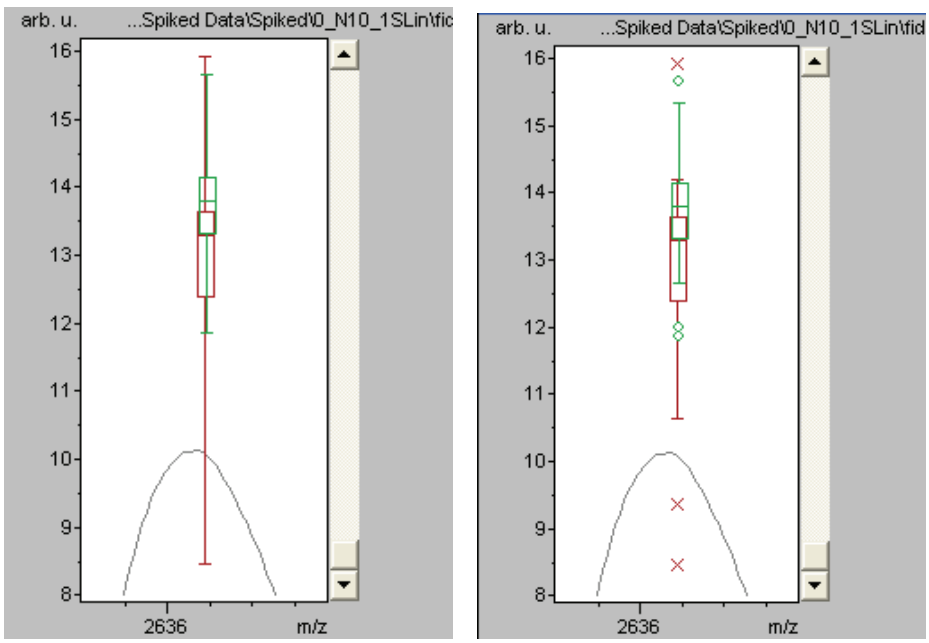


Figure 9-14 **Standard box plot (left) and modified box plot (right) indicating the outliers (three by red crosses and three green circles) not belonging to the 95% of values inside the whiskers**

9.1.3.6.11 Spectra View > Peak Markers Command

The **Peak Markers** command shows/hides the peak markers in the Spectra View (Figure 9-15). A black arrow marker (v) at the top of the view indicates the peak(s) for which corresponding data is shown in the 2D Peak Distribution, ROC Curve or Single Peak Variance View. The peak markers are hidden by default.

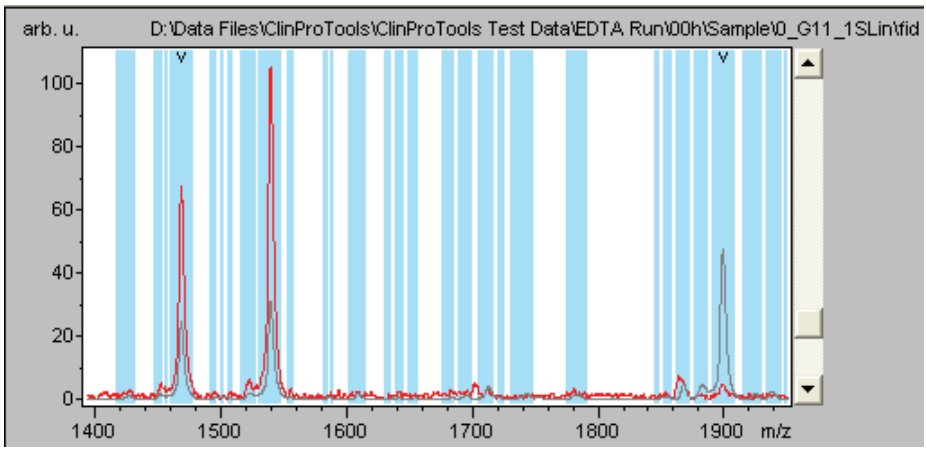


Figure 9-15 Peak markers indicating in the Spectra View the two peaks selected in the 2D Peak Distribution View

9.1.3.7 Gel/Stack View Popup Command

Pointing to **Gel/Stack View** offers the following commands (Figure 9-16):

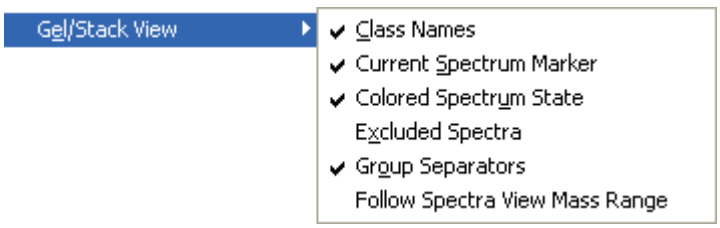


Figure 9-16 Gel/Stack View submenu

Command	Used to …
Class Names	Show/Hide the class names in the Gel View.
Current Spectrum Marker	Show/Hide the current spectrum marker in the Gel View.
Colored Spectrum State	Mark/Do not mark the spectrum state with modified colors in the Gel View.

Command	Used to …
Excluded Spectra	Hide/Show excluded spectra in the Gel View and Stack View.
Group Separators	Show/Hide group separators in Gel View.
Follow Spectra View Mass Range	Force/Do not force the Gel View to follow the mass range of the Spectra View.

9.1.3.7.1 Gel/Stack View > Class Names Command

The **Class Names** command shows/hides the names and color-coding of the loaded classes in the Gel View. The class names are shown by default.

9.1.3.7.2 Gel/Stack View > Current Spectrum Marker Command

The **Current Spectrum Marker** command shows/hides the current spectrum marker in the Gel View. Two arrow markers (> <) at the left and the right border of the Gel View (Figure 5-3) indicate the single spectrum that is selected in the Spectra View. The current spectrum marker is shown by default.

9.1.3.7.3 Gel/Stack View > Colored Spectrum State Command

The **Colored Spectrum State** command shows/hides the coloring of spectra according to their state in the Gel View as well as in the Spectra List report (Section 8.1.1.1). Different spectrum states concerning automatic exclusion by specific filters or manual exclusion are indicated by predefined colors. In addition, not recalibratable but not excluded spectra are marked in pink.

Table 9-1 lists the colors used and explains their meaning. Colored spectrum states are shown by default.

Table 9-1 Explanation of coloring of spectra in the Gel View and Spectra List

Color		Description
	light gray	Null spectrum excluded by null spectra exclusion filter.
	yellow	Spectrum excluded by exception (see comment in Spectra List)
	lilac	Spectrum excluded by noise spectra exclusion filter (Figure 9-17).
	green	Spectrum excluded by adduct/polymer spectra exclusion filter.
	turquoise	Spectrum excluded by similarity selection filter (Figure 9-18).
	red	Not recalibratable; excluded by spectra quality filter.
	pink	Not recalibratable; not excluded.
	dark gray	Spectrum manually excluded (Figure 9-17).

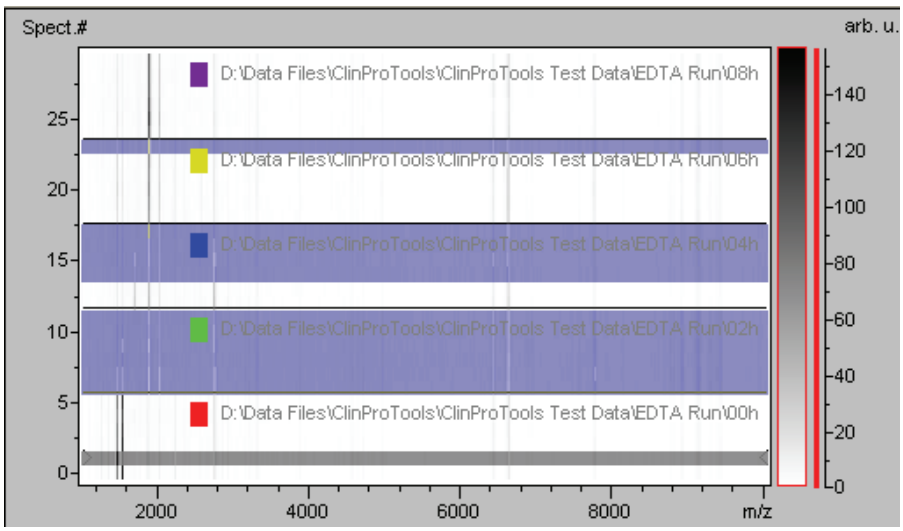


Figure 9-17 **Coloring of spectra excluded by the noise spectra exclusion filter during spectra loading (lilac) and by manual exclusion, respectively (dark gray)**

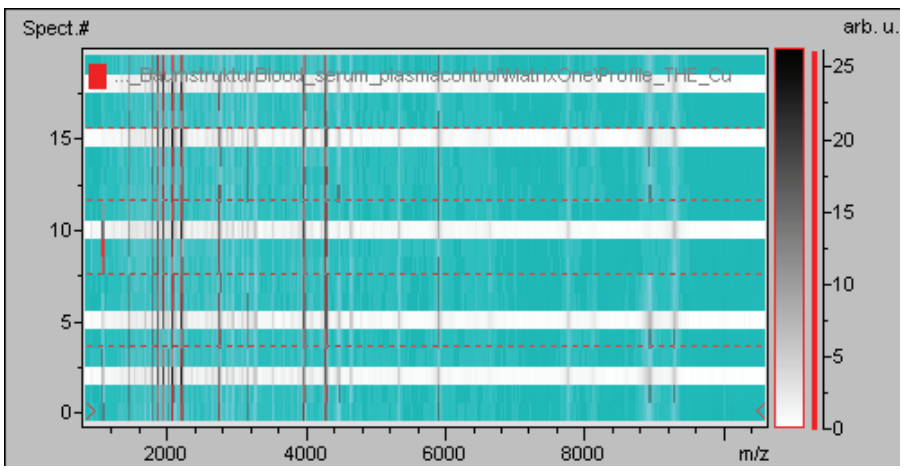


Figure 9-18 **Coloring of spectra excluded by the similarity selection filter (turquoise); the most characteristic spectrum of each group remains included and thus is not specially colored**

9.1.3.7.4 Gel/Stack View > Excluded Spectra Command

The **Excluded Spectra** command shows/hides excluded spectra in the Gel and Stack views. Excluded spectra are shown by default.

9.1.3.7.5 Gel/Stack View > Group Separators Command

The **Group Separators** command shows/hides group separators in the Gel View (Figure 9-19). Group separators are horizontal dashed lines, which separate spectra groups resulting from multiple measurements of samples within one class. These markers are only displayed if the **Support Spectra Grouping** option in the **Settings Spectra Preparation** dialog is checked on spectra loading. Group separators are shown by default.

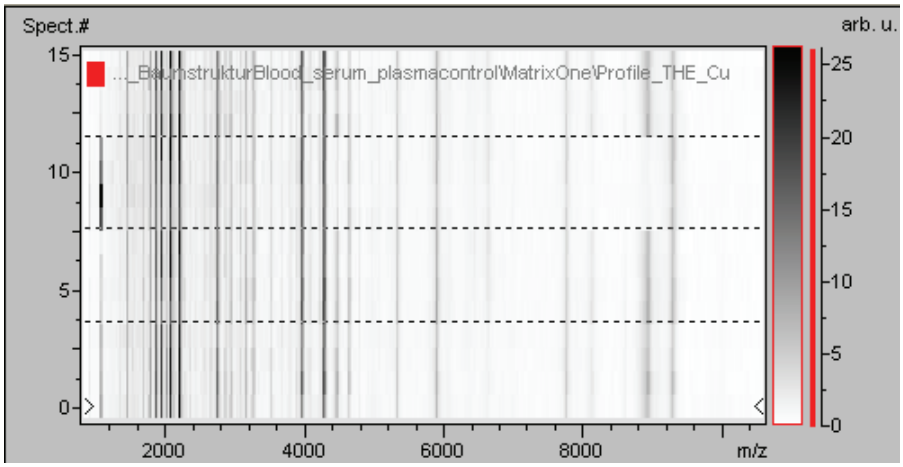


Figure 9-19 Separation of groups of spectra from multiple measurements by horizontal group separators, here each group consists of four spectra

9.1.3.7.6 Gel/Stack View > Follow Spectra View Mass Range Command

The **Follow Spectra View Mass Range** command force or do not forces the Gel/Stack View to follow the mass range of the Spectra View. By default, the x-axis of the Spectra View and the x-axis of the Gel/Stack View show a slave-master behavior. The x-axis of the Spectra View always follows the x-axis of the Gel/Stack View when the latter is changed but contrarily the x-axis of the Gel/Stack View is kept when the x-axis in the Spectra View is changed. Enabling the command switches the x-axis of the Gel/Stack View to dependence on the x-axis of the Spectra View and disabling it resets to the default behavior.

9.1.3.8 Peak Statistics View Popup Command

Pointing to **Peak Statistics View** offers the following commands (Figure 9-20):

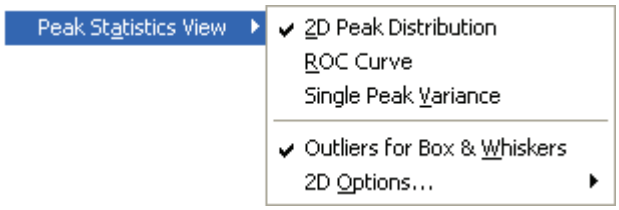


Figure 9-20 Peak Statistics View submenu

Command	Used to …
2D Peak Distribution	Switch to 2D Peak Distribution View and display the 2D peak distribution for two selected peaks.
ROC Curve	Switch to ROC Curve View and display the ROC curve for the selected peak.
Single Peak Variance	Switch to Single Peak Variance View and display the current peak statistics for the selected peak.
Outliers for Box & Whiskers	Show/Hide the outliers for box & whiskers plots for the peak area/intensity per class in the Single Peak Variance View.
2D Options	Pop up command for displaying data in the 2D Peak Distribution View.

9.1.3.8.1 Peak Statistics View > 2D Peak Distribution Command

The **2D Peak Distribution** command switches the Peak Statistics View to 2D Peak Distribution View to display the 2D peak distribution for two selected peaks.

Shortcut

Button: 

9.1.3.8.2 Peak Statistics View > ROC Curve Command

The **ROC Curve** command is used to switch the Peak Statistics View to ROC Curve View to display the ROC curve for the current peak of two loaded model generation classes. This requires first a decision to be made whether class 1 or class 2 should be treated as positive. The decision remains valid as long as the ROC Curve View is not closed but can be changed by selecting the command again and making a new decision. The current decision also applies to the **ROC Curve for Peak n** command from the Spectra View context menu.

The command opens the **ROC** dialog (Figure 9-21) to specify the class to be treated as positive. Clicking the corresponding button opens the ROC Curve View and displays the ROC curve for the current peak.

Shortcut

Button: 

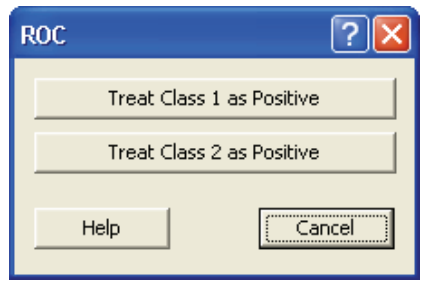


Figure 9-21 ROC dialog

9.1.3.8.3 Peak Statistics View > Single Peak Variance Command

The **Single Peak Variance** command switches the Peak Statistics View to Single Peak Variance View to display statistic data (box and whiskers, peak distribution or average with standard deviation) for the current peak. The data shown depends on the state of the **View** menu commands **Spectra View > Box & Whiskers**, **Peak Distribution** and **Average & StdDev** (Section 5.1.3.3). The Single Peak Variance View can also be launched via the **Variance for Peak n** command from the Spectra View context menu.

Shortcut

Button: 

9.1.3.8.4 Peak Statistics View > Outliers for Box & Whiskers Command

The **Outliers for Box & Whiskers** command shows/hides the outliers for the per-class box & whiskers plots in the Single Peak Variance View (and also in the Spectra View) when the **Spectra View > Box & Whiskers** command is active. This toggles the box & whiskers plots between the standard box plot (command not active) and the modified box plot (command active) (Figure 9-22). For description of standard and modified box plot, please refer to Section 9.1.3.6.10).

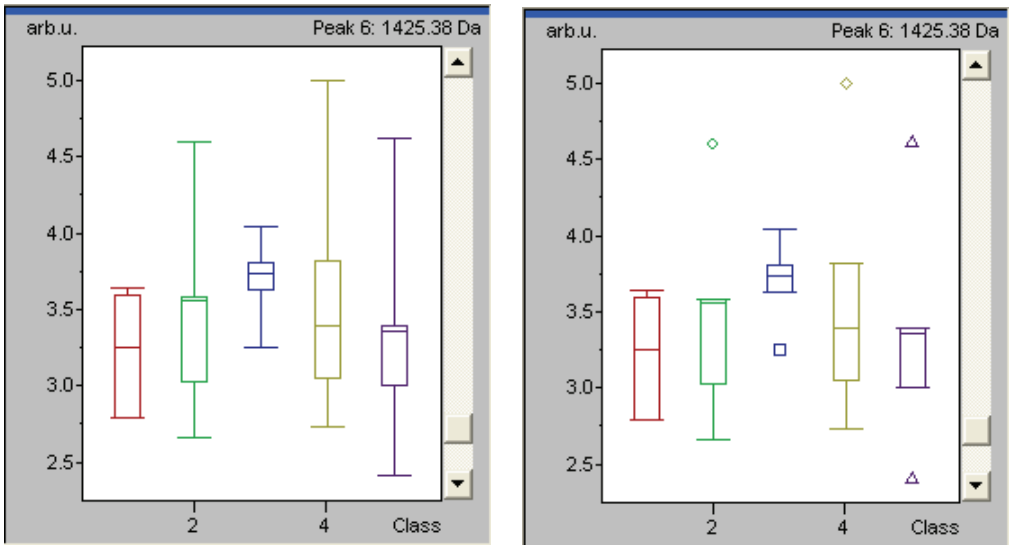


Figure 9-22 Standard box plot (left) and modified box plot (right) indicating the outliers (diamonds, box, triangles) not belonging to the 95% of values inside the whiskers

9.1.3.8.5 Peak Statistics View > 2D Options Popup Command

Pointing to **2D Options** offers the following commands (Figure 9-23):

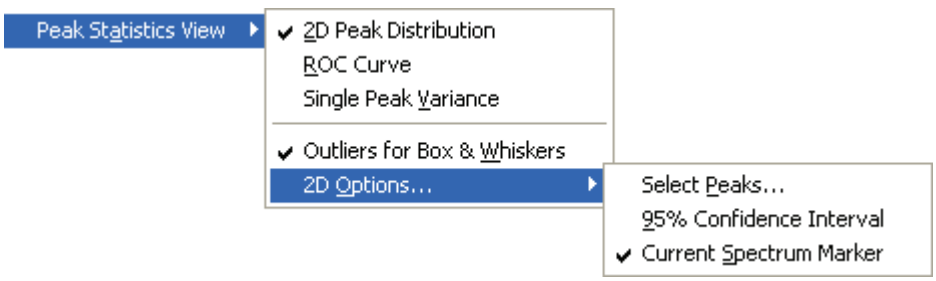


Figure 9-23 2D Options submenu

Command	Used to …
Select Peaks	Select two peaks to be displayed in the 2D Peak Distribution View.
95% Confidence Interval	Display the 95% confidence interval or standard deviation in the 2D Peak Distribution View.
Current Spectrum Marker	Mark/Do not mark in the 2D Peak Distribution View the data point corresponding to the current spectrum.

9.1.3.8.5.1 Peak Statistics View > 2D Options > Select Peaks Command

The **Select Peaks** command is used to change the current peak selection in the 2D Peak Distribution View. By default, the first two peaks of the current statistic sort order set via the **Settings Statistic** command (Section 9.1.8.5) are displayed. The horizontal axis plots the first, the vertical axis the second peak. The command opens the **Peak Distribution** dialog (Figure 9-24) to select two peak indices for displaying the corresponding peak distribution.

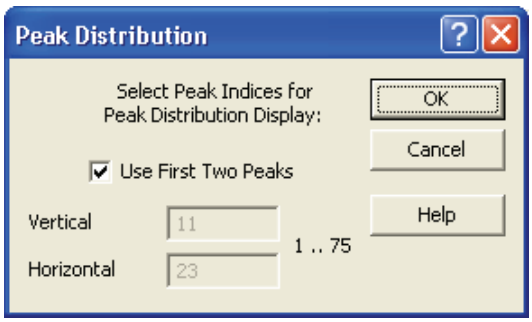


Figure 9-24 Peak Distribution dialog

Use First Two Peaks

Check this option if the first to peaks of the current statistic sort order should be displayed. Uncheck this option if you want to display the peak distribution for another pair of peaks, which you have to specify in **Vertical** and **Horizontal**.

Vertical / Horizontal

Enter the index of the peak to be displayed on the vertical axis.

Horizontal

Enter the index of the peak to be displayed on the horizontal axis.

OK

Updates the 2D Peak Distribution View for the now selected peaks.

9.1.3.8.5.2 Peak Statistics View > 2D Options > 95% Confidence Interval Command

The **95% Confidence Interval** command toggles the 2D Peak Distribution View between displaying the calculated 95% confidence interval (command is active) (Figure 9-25) or standard deviation (command is inactive) (Figure 9-26) for each class as an ellipse. The 95% confidence interval is the standard deviation weighted by the reciprocal number of data points. The ellipses are displayed according to the classes' color-coding. For information about the confidence interval, we refer to J. M. Chambers and T. J. Hastie, "Statistical Models in S", Wadsworth & Brooks/Cole (1992).

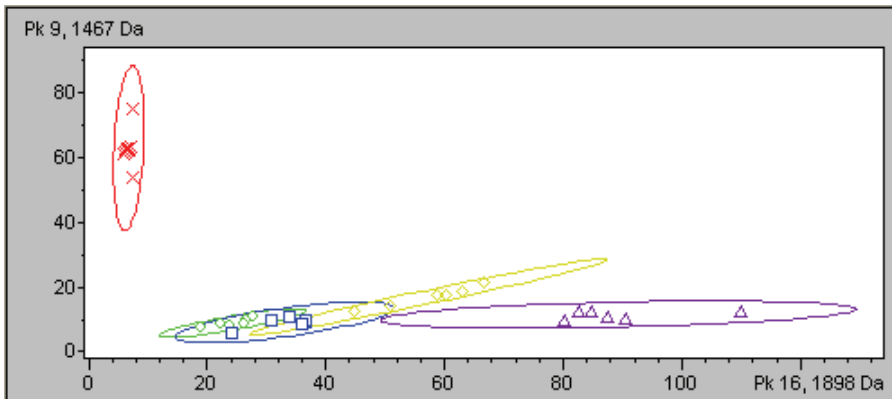


Figure 9-25 Display of 95% confidence interval

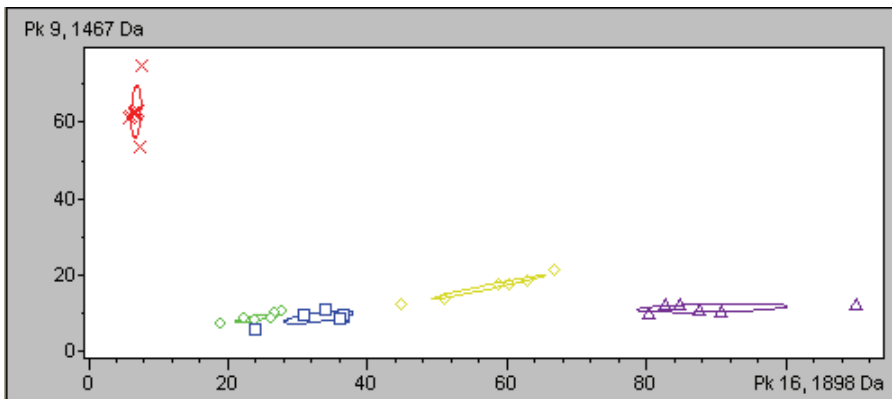


Figure 9-26 Display of standard deviation

9.1.3.8.5.3 Peak Statistics View > 2D Options > Current Spectrum Marker Command

The **Current Spectrum Marker** command shows/hides in the 2D Peak Distribution View the marking of that data point that corresponds to the current spectrum in the Spectra View (Figure 9-27). The respective data point is marked by bold display.

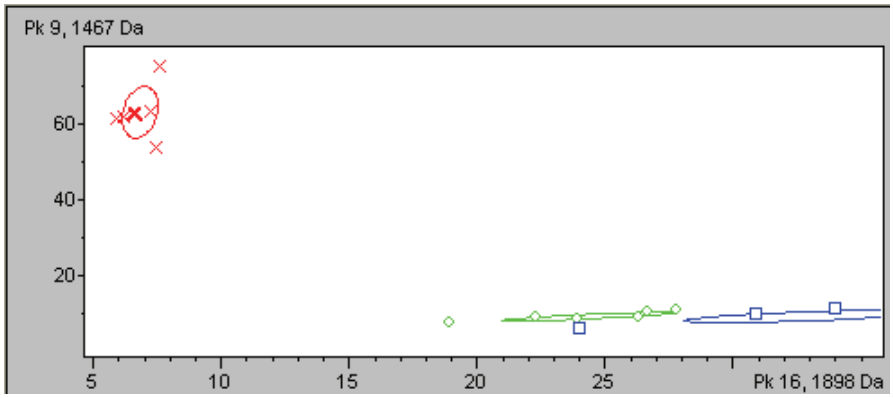


Figure 9-27 Marking the data point that corresponds to the current spectrum by bold display (here: bold red cross in top left corner)

9.1.3.9 Reset View Settings Command

The **Reset View Settings** command resets certain settings of the data plotting views to the default settings. Selecting this command opens a confirmation request to reset the view settings. Click **Yes**, to reset them, click **No** to keep the current settings.

The following defaults are restored:

View	Resets to
All views:	- default split view partition - no grid - no auto-scaling - zooming enabled - full display of data (zoom reset) - background of display region: white (does not apply to Gel View) - background of axes: gray (RGB(192,192,192)) - axis font: Arial, standard size - hidden scales shown again - only BMP format to clipboard
Spectra View:	- spectra display: line, no markers - line width: 1 pixel
Gel View:	- displayed - color scheme: gray scale - color intensity: quadratic
Stack View:	- orientation 30° with basis approx. one third of view height - colored spectra (no whitewash)
Peak Statistic views:	- point width: 1 pixel

9.1.4 Data Preparation Menu

The **Data Preparation** menu offers the following commands (Figure 9-28):

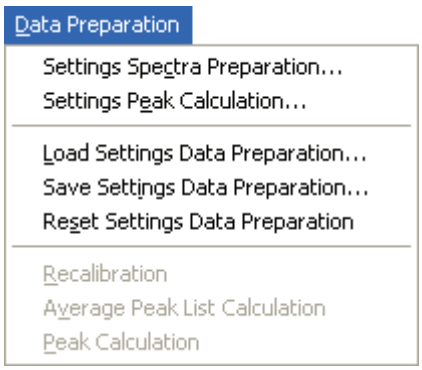


Figure 9-28 Data Preparation menu

Command	Used to …
Settings Spectra Preparation	Define the spectra preparation and recalibration settings.
Settings Peak Calculation	Define the peak picking and peak calculation settings.
Load Settings Data Preparation	Load the selected data preparation settings XML file.
Save Settings Data Preparation	Save the current data preparation settings in an XML file with a specified name.
Reset Settings Data Preparation	Reset the current data preparation settings to their defaults.
Recalibration	Recalibrate spectra and calculates average spectra.
Average Peak List Calculation	Calculate the average peak list on the total average spectrum.
Peak Calculation	Calculate peaks and peak statistic in the single spectra.

9.1.4.1 Settings Spectra Preparation Command

The **Settings Spectra Preparation** command is used to set the parameters for preparing spectra (by modification and selection) and picking recalibration masses during spectra loading as well as for spectra recalibration. The settings are stored with the peak calculation settings in the *SettingsDataPreparation.xml* file. The command opens the **Settings Spectra Preparation** dialog (Figure 9-29).

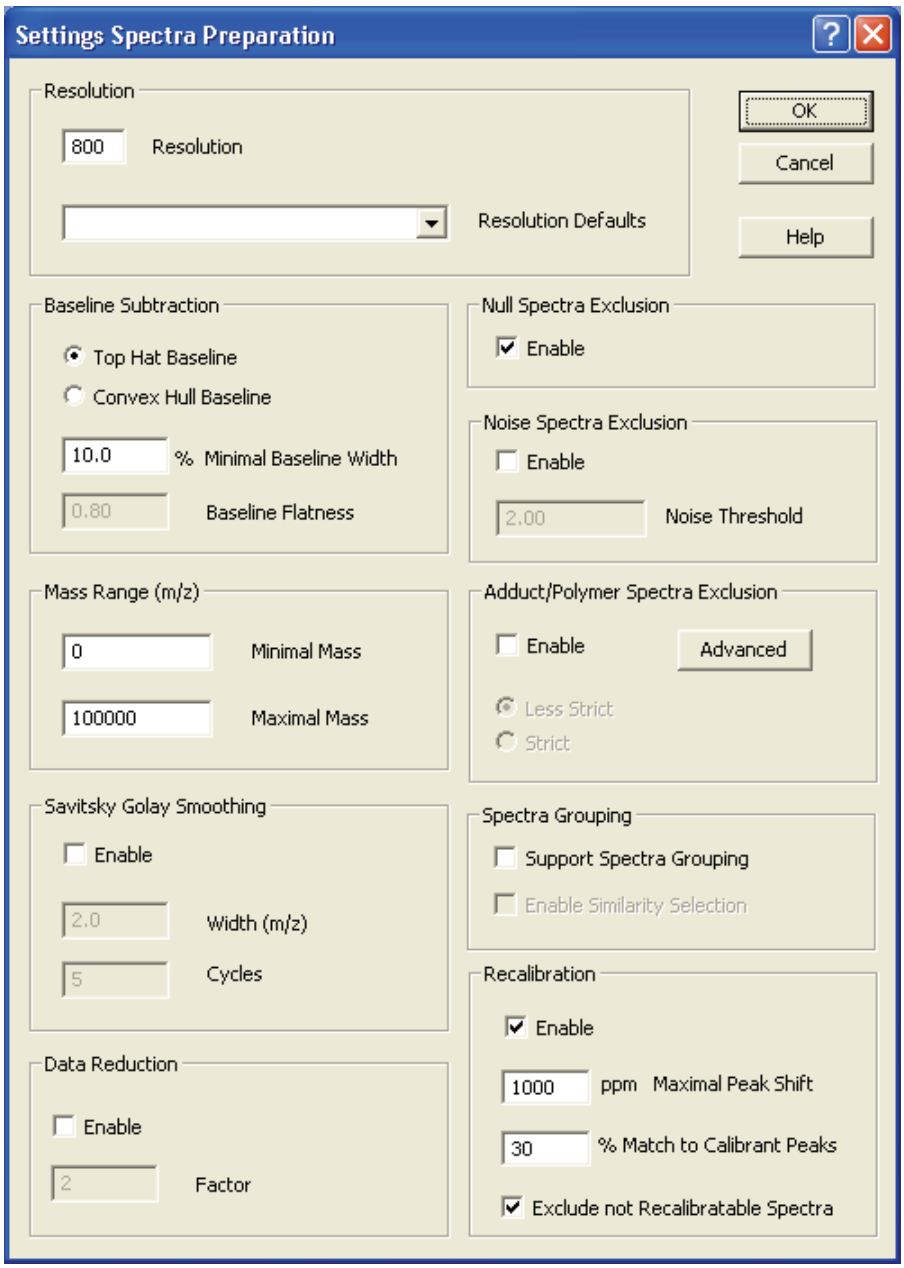


Figure 9-29 Settings Spectra Preparation dialog (default setting)

In **Resolution**, define the resolution (Section 6.1.3.1) to be applied to the peak detection algorithm as a hint for the peak width.

Resolution

Enter the resolution to be applied to detecting peaks for individual spectra (used for recalibration) and peaks for the total average spectrum (used for model building). If this parameter is chosen too large, more and more artificial peaks (spikes) will be found. On the other hand, smaller resolution values will remove more and more unresolved (shoulder) peaks from the peak list. Alternatively, you can select the respective mass range from the drop down list below to enter the corresponding default resolution here.

Resolution Defaults

Lists specific mass ranges for which default resolution values can be loaded into **Resolution**. This requires the *ResolutionDefaults.xml* file be present in the ClinProTools folder. If this file is not present, the box is disabled and shows "--- No Defaults Available ---". To enter a default resolution, select the mass range from the list.

In **Baseline Subtraction**, define the parameters for the baseline subtraction filter:

Top Hat Baseline

Select this option to perform Top Hat baseline subtraction (Section 6.1.1.1) which constructs the baseline by means of morphology operators. The range for the minimum and maximum search can be enlarged with the **% Minimal Baseline Width** parameter.

Convex Hull Baseline

Select this option to perform Convex Hull baseline subtraction (Section 6.1.1.1) which constructs the baseline by fitting multiple parabolas to the spectrum. The **Baseline Flatness** parameter influences baseline construction.

% Minimal Baseline Width

For **Top Hat Baseline** selected, enter the % minimal baseline width. This parameter influences the level of details to which the baseline approaches the spectrum. If this value is larger than 0.0 it tells the algorithm that the range in mass units should be at least the given fraction of the mass of the actual data point for which the baseline has to be calculated. If this value is increased, groups of overlapping peaks will be less affected but the flatness of the baseline is also reduced. This is especially of interest for mass ranges > 20 kDa where the baseline correction might otherwise remove too much from broad overlapping peaks.

Baseline Flatness

For **Convex Hull Baseline** selected, enter the flatness of the baseline. This parameter influences the number of parabola used to explain the baseline and the flatness of the resulting spectrum, which is obtained by subtracting the baseline from the spectrum. The larger the flatness value the finer the baseline will approach the spectrum.

In **Mass Range**, define the parameters for the mass range filter (Section 6.1.3.1). If you want to limit the mass range, specify a minimal and a maximal mass. Otherwise, define a mass range that is larger than the experimental mass range, which should be the case if you keep the default values.

Minimal Mass

Enter the minimal mass of the mass range. All masses below this value will be cut before loading the spectra.

Maximal Mass

Enter the maximal mass of the mass range. All masses above this value will be cut before loading the spectra. **Maximal Mass** must be at least two times larger than **Minimal Mass**.

In **Savitsky Golay Smoothing**, define the parameters for the smoothing filter (Section 6.1.3.1):

Enable

Check this option if the Savitsky Golay smoothing filter should be enabled.

Width (Da)

Enter the smoothing width in Dalton.

Cycles

Enter the number of smoothing iterations.

In **Data Reduction**, define the parameters for the data reduction filter (Section 6.1.3.1):

Enable

Check this option if the data reduction filter should be enabled. Setting this option allows speeding up calculations and reducing the memory consumption, especially for very large data sets. However, best classification results are expected without data reduction.

Note: The data reduction is applied prior to any other data processing and influences all subsequent results.

Factor

Enter the data reduction factor. This sets the number of consecutive data points in a set that are to be replaced by the average of these points. Typically, the value should be chosen between 2 (double reduction) and 10 (10-fold reducetion). The greater the factor is chosen the more features will be smoothed out. As a consequence, e.g., shoulder peaks may no longer be resolved.

In **Null Spectra Exclusion**, define the parameters for the null spectra exclusion filter (Section 6.1.3.2):

Enable

Check this option if the null spectra exclusion filter should be enabled to sort out spectra with a very low intensity. In general, this option should be disabled only in the case of third party spectra.

In **Noise Spectra Exclusion**, define the parameters for the noise spectra exclusion filter (Section 6.1.3.2):

Enable

Check this option if the noise spectra exclusion filter should be enabled. Detected spectra will be excluded.

Noise Threshold

Enter the noise threshold. It can be considered as a required mean signal-to-noise value of the spectra in the considered range. Values ≥ 1 can be applied; the higher the value the stricter the detection will be.

In **Adduct/Polymer Spectra Exclusion**, define the parameters for the adduct/polymer spectra exclusion filter (Section 6.1.3.2):

Enable

Check this option if the adduct/polymer spectra exclusion filter should be enabled.

(exclusion mode)

Select at which adduct/polymer level the spectra should be excluded:

- **Less Strict**. Excludes only spectra with adducts/polymers of a higher level. This mode allows spectra with shifts of lower contribution to remain in the spectra set collection. This is determined upon an experimental obtained internal threshold.
- **Strict**. Excludes spectra with adducts/polymers of any level. This mode aims on exclusion of spectra which show characteristic shifts in the autocorrelation spectrum with respect to the adduct/polymer parameterization. The underlying criterion is strict, which means if such a shift exists and it is not due to randomness the spectrum is excluded.

Advanced

Allows defining advanced adduct/polymer spectra exclusion parameters.

The button opens the **Settings Adduct/Polymer Spectra Exclusion** dialog (Figure 9-30) to customize settings for that filter. The dialog can only be edited if **Enable** is checked.

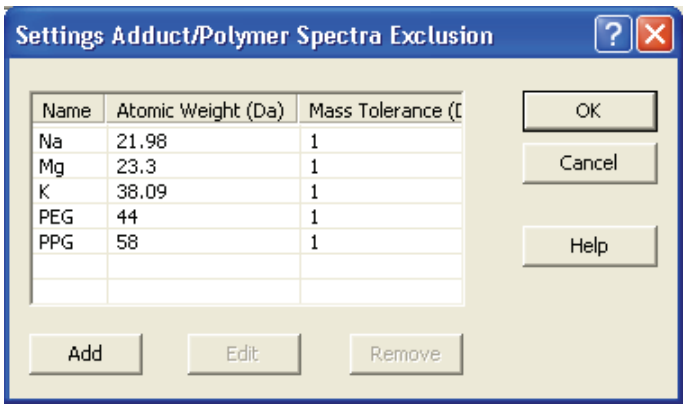


Figure 9-30 Settings Adduct/Polymer Spectra Exclusion dialog (default setting)

Name lists the adducts/polymers the filter should search the spectra for. Sodium (**Na**), magnesium (**Mg**), potassium (**K)**, polyethylenglycol (**PEG**) and polypropyleneglycol (**PPG**) are contained by default. You can add new adducts/polymers to this list and change or remove existing adducts/polymers.

Atomic Weight (Da) lists the corresponding atomic weights of the adducts/polymers.

Mass Tolerance (Da) lists the corresponding mass tolerance allowed for detecting the adduct/polymer peak.

Add / Edit allows adding a new resp. editing the selected adduct/polymer. The button opens the **Adduct/ Polymer Property** dialog (Figure 9-31) to add a new adduct/polymer to be searched for by the filter or to edit the selected one with respect to atomic weight and/or mass tolerance. Do the desired entries/changes and click **OK** to add the new adduct/polymer to the **Settings Adduct/Polymer Spectra Exclusion** dialog resp. change the selected one.

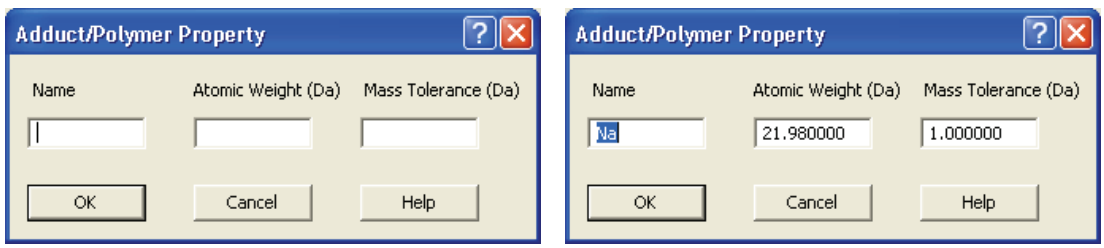


Figure 9-31 Adduct/Polymer Property dialog for adding a new (left) or editing an existing adduct/polymer (right)

Remove removes the selected adduct/polymer.

In **Spectra Grouping**, define the parameters for spectra grouping (Section 6.1.2) and the similarity selection filter (Section 6.1.3.2):

Support Spectra Grouping

Check this option if multiple spots of one sample should be treated as a group (spectra grouping).

Note: This option is suitable only for automatically created spectra by the current ClinProtRobot with the corresponding software. If the option is enabled while using a different folder structure, the parser might by coincidence detect not existing groups, which will lead to calculation errors.
If switched on, manually copied spectra may also be parsed as a group if it has the same folder structure as generated by the ClinProtRobot.

Enable Similarity Selection

Check this option if the similarity selection filter should be enabled to detect the most suitable spectrum of a spectra group; not suitable spectra will be excluded.

In **Recalibration**, define the parameters for recalibration (Section 6.1.1.3) and the spectra quality filter (Section 6.1.3.2):

Enable

Check this option if spectra should be recalibrated.

ppm Maximal Peak Shift

Enter the maximal mass shift allowed for a peak in recalibration in ppm. Values from 1 to 2000 ppm can be set.

% Match to Calibrant Peaks

Enter the percentage match to calibrant peaks value which is multiplied with the Maximum Quality Value (= number of reference masses) to determine the Spectra Quality Threshold of the filter. The spectrum's Spectrum Quality Value must reach this threshold so that the spectrum is not marked as 'not recalibratable'. 0% means no exclusion; the highest reasonable value probably is 80 %.

Exclude Not Recalibratable Spectra

Check this option if spectra that are marked as 'not recalibratable' should be excluded in further processing.

OK

Changes the spectra preparation settings. Depending on the current processing state, the views may become cleared to prevent the loaded spectra from further processing; then a message will inform you on how to proceed.

9.1.4.2 Settings Peak Calculation Command

The **Settings Peak Calculation** command is used to set the parameters for picking peaks and calculating peak areas/intensities and statistical data. The settings are stored with the spectra preparation settings in the *SettingsDataPreparation.xml.* The command opens the **Settings Peak Calculation** dialog (Figure 9-32).

Note: The peak calculation settings, especially the **Signal to Noise Threshold**, may strongly influence the quality of the chosen classification algorithm.

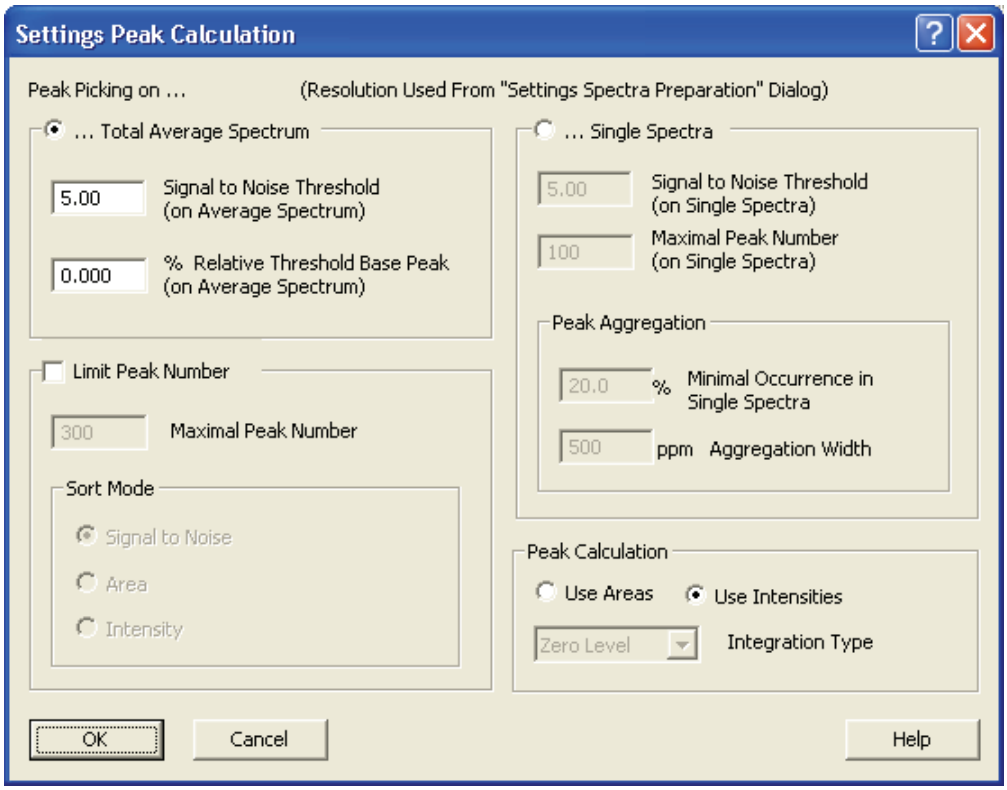


Figure 9-32 Settings Peak Calculation dialog (default setting)

In **Peak Picking**, define how to pick peaks:

…Total Average Spectrum

Check this option if peaks should be picked on the total average spectrum and the overall average peak list should be calculated from this spectrum like in ClinProTools 2.1 (Section 6.1.1.5.1).

Signal to Noise Threshold (on Average Spectrum)

Enter the minimum signal-to-noise ratio a peak must have in order to be detected. The higher this value, the less peaks are detected but the higher is the quality of the detected peaks. Reasonable values are 2.0 and above.

% Relative Threshold Base Peak (on Average Spectrum)

Enter the minimum relative intensity a peak must have with respect to the base peak in order to be detected. Peaks with a lower relative intensity are excluded.

… Single Spectra

Check this option if peaks should be picked on the single spectra and averaged peak lists over all classes and the single classes should be calculated (Section 6.1.1.5.2). In addition, a peak statistic for further use in pattern matching algorithms is stored. The averaged peak list over all classes is used instead of the average peak list obtained from the total average spectrum in CPT 2.1. Because overlapping peak ranges from different classes will be cut into separate non-overlapping ranges, it is preferable to use peak intensities instead of areas, which might better represent the different peaks in this case.

If smoothing (in **Settings Spectra Preparation** dialog) is currently not enabled when selecting this option, a warning will appear which recommends enabling smoothing. You can skip this warning in future by checking the corresponding option in this message and turn it on again by enabling the **Show Smoothing Warning** option in the the **Settings General** dialog.

Signal to Noise Threshold (on Single Spectra)

Enter the minimum signal-to-noise ratio a peak must have in order to be detected. The higher this value, the less peaks are detected but the higher is the quality of the detected peaks. Reasonable values are 2.0 and above.

Maximal Peak Number (on Single Spectra)

Enter the maximal number of peaks to pick on a single spectrum.

Minimal Occurrence in Single Spectra

Enter in which part of the spectra a peak must occur at least [in %] to be added to the average peak list. In the case of few classes and little differences between the classes a quite high value can be chosen.

Note: If there are more classes and the classes differ much, a lower value has to be chosen, e.g. in the case of 8 classes with total different peak patterns a value lower than 1/8 = 12.5 % has to be chosen.

Aggregation Width

Enter the range of clusters [in ppm] for peak aggregation during average peak list generation as well as during the classification in the case of the statistical test based algorithm.

Limit Peak Number

Check this option if you want to limit the number of peaks to pick to a **Maximal Peak Number** according to a selected **Sort Mode**.

Maximal Peak Number

Enter the maximal number of peaks to pick. If more peaks have been found, the first N best peaks according to the selected **Sort Mode** are kept. You can set the number to '0' to allow pure manual peak editing (Section 7.1.5.2).

Sort Mode

Select the sort mode according to which the N best peaks are selected:

Signal to Noise. Sorts by signal-to-noise ratio.
Area. Sorts by area.
Intensity. Sorts by intensity.

In **Peak Calculation**, define how to calculate the peaks in the individual spectra:

Use Areas

Check this option if the peak areas should be used for peak calculation; these are determined based on the selected **Integration Type**.

Use Intensities

Check this option if the maximum peak intensities based on zero level should be used for peak calculation.

Integration Type

For **Use Areas** selected, choose the integration type for calculating peak areas; these two options will yield different areas especially for shoulder peaks:

End-Point Level. Integrates only the area above the cutting edge connecting the start and end points of the peak.
Zero Level: Integrates the full intensity values.

OK

Changes the current peak calculation settings. Depending on the current processing state, the views may become cleared to prevent the loaded spectra from further processing; then a message will inform you on how to proceed.

9.1.4.3 Load Settings Data Preparation Command

The **Load Settings Data Preparation** command is used to load a stored data preparation settings XML file. Loading a data preparation settings file is always possible; however, if spectra have already been loaded you might have to close the spectra and load them again or repeat the previous processing depending on which data preparation settings have been changed. The command opens the **Load Settings Data Prepara-**

tion File dialog with the SettingsDataPreparation folder opened by default. Navigate to the file you want to load and click **Open**. This overwrites the current data preparation settings with the loaded ones. If spectra are currently loaded, a message informs you on how to proceed.

9.1.4.4 Save Settings Data Preparation Command

The **Save Settings Data Preparation** command is used to save the current spectra preparation and peak calculation settings in an XML file with a specified name. The command opens the **Save Data Preparation Settings File** dialog with the Settings-DataPreparation folder as the default storage location. Enter a file name or select one from the folder list and click **Save**. If you have selected an existing file name, answer the confirmation request to overwrite the file.

9.1.4.5 Reset Settings Data Preparation Command

The **Reset Settings Data Preparation** command is used to reset the current spectra preparation and peak calculation settings to their defaults. Resetting the data preparation settings is always possible; however, if spectra have already been loaded you might have to close the spectra and load them again or repeat the previously processing depending on which data preparation settings have been changed. The command displays a confirmation request to reset to defaults. Click **Yes** to reset the current settings, click **No** to retain them. If spectra are currently loaded, a message informs you on how to proceed.

9.1.4.6 Recalibration Command

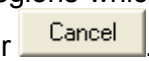
The **Recalibration** command is used to run the recalibration workflow. This includes recalibration of the spectra of the loaded classes and calculation of the total average spectrum and the class average spectra from all not excluded spectra. Recalibration is performed based on the recalibration masses already picked during spectra loading and the current recalibration settings (Section 9.1.4.1).

The command runs spectra recalibration and average spectra calculation. The spectra quality filter (Section 6.1.3.2) marks spectra, which are not recalibratable. These become excluded if the corresponding option is set. The calculated total average spectrum is shown in the Spectra View by default. The additionally calculated class average spectra can be shown on demand (Section 9.1.3.6.4). You can cancel the running process by clicking  or 

.

9.1.4.7 Average Peak List Calculation Command

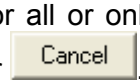
The **Average Peak List Calculation** command is used to run the average peak list calculation workflow. This automatically picks peaks on either the total average spectrum or the single spectra based on the current peak picking settings (Section 9.1.4.2) and determines the integration regions. The calculated average peak list can be edited manually (Section 7.1.5.2).

The command runs the peak picking and average peak list calculation. If the recalibration workflow has not been performed yet when selecting this command, it will be automatically run before the average peak list calculation workflow starts. The picked peaks are indicated in the Spectra View by gray marked integration regions which are shown by default.You can cancel the running process by clicking  or 

.

9.1.4.8 Peak Calculation Command

The **Peak Calculation** command is used to run the peak calculation workflow. This calculates the peaks stored in the average peak list in the single spectra and specific peak statistic data. In case of two loaded model classes, also the ROC curves per peak (Section 6.4.2.2) are generated. Optionally, it also performs peak selection for model generation. Peak calculation is based on the current peak calulation settings (Section 9.1.4.2). Either the peak areas or the maximal peak intensities can be used. Peak areas are normalized for model generation when using the GA, SVM or SNN. Peak selection is performed according to the current peak selection settings (Section 9.1.5.1).

The command runs peak and peak statistic calculation as well as peak selection if specified. If the recalibration or the average peak list calculation workflow has not been performed when selecting this command, the respecttive workflow(s) iwill be automatically run before the peak calculation workflow starts. The result of peak calculation can be viewed in the Peak Statistic report (Section 8.1.1.2). Depending on the peak selection settings, either all peaks or only the selected best ones are marked as included in model generation (shown by blue instead of gray integration regions). Depending on the current the statistic settings (Section 9.1.8.5) and Spectra View settings (Section 9.1.3.6) certain peak statistic data are shown for all or only the selected peaks. You can cancel the running process by clicking  or 

.

9.1.5 Model Generation Menu

The **Model Generation** menu offers the following commands (Figure 9-33):

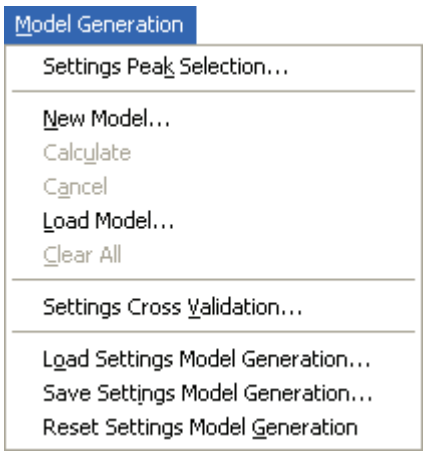


Figure 9-33 Model Generation menu

Command	Used to …
Settings Peak Selection	Define the peak selection settings.
New Model	Add a new model parameter set to the model list; this launches selecting algorithm, setting algorithm-specific model parameters and specifying model name.
Calculate	Start model generation.
Cancel	Cancel the current loading/calculation/generation/-classification process.
Load Model	Load the selected model.
Clear All	Clear the Model List View.
Settings Cross Validation	Define the cross validation settings.
Load Settings Model Generation	Load the selected model generation settings XML file.
Save Settings Model Generation	Save the current model generation settings to an XML file with specified name.
Reset Settings Model Generation	Reset the current model generation settings to their defaults.

9.1.5.1 Settings Peak Selection Command

The **Settings Peak Selection** command is used to define which peaks should be used in model generation. By default, all picked peaks are used but you can define that only a restricted number of best peaks with respect to the selected sort mode are taken. The peak selection settings are applied to the spectra within the peak calculation workflow; however, the resulting selection will become effective only in model generation. The settings are stored with the cross validation, GA, SVM, SNN and QC settings in the *SettingsModelGeneration.xml* file, which is updated on each settings change. The command opens the **Settings Peak Selection** dialog (Figure 9-34).

Note: The peak selection settings may strongly influence the quality of the chosen classification algorithm. In many cases, a reasonable reduction of peaks improves the classification performed by the algorithms.

Note: Apart from the peak selection settings to use in model generation, you can define differing settings for calculating peak statistic (Section 9.1.8.5).

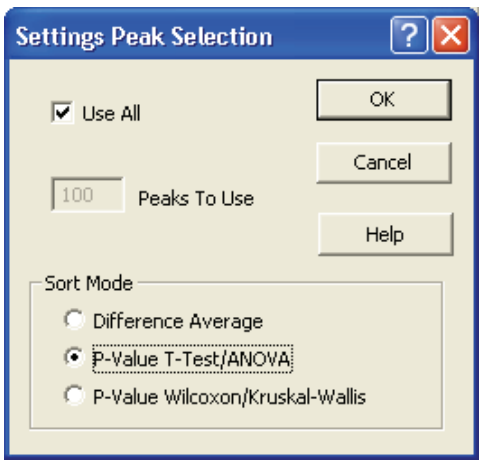


Figure 9-34 Settings Peak Selection dialog (default setting)

Use All

Check this option if you want to use all picked peaks in model generation. Uncheck it if you want to restrict the number of peaks to a maximal number of best ones.

Peaks To Use

If **Use All** is unchecked, enter the maximal number of best peaks to use. The peaks are selected with respect to the **Sort Mode**.

Sort Mode

Select how to sort peaks if only the best ones should be used:

Difference Average. Sorts the peaks by the difference between the maximal and the

minimal average peak area of all classes.

P-Value T-Test/ANOVA. Sorts the peaks by the p-value from t-test (Section 6.4.1.1) / ANOVA test (Section 6.4.1.2).

P-Value Wilcoxon/Kruskal-Wallis. Sorts the peaks by the p-value from Wilcoxon test (Section 6.4.1.3) / Kruskal-Wallis test (Section 6.4.1.4).

OK

Changes the current peak selection settings. If peak selection has already been performed, the current selection is changed according to the new settings.

9.1.5.2 New Model Command

The **New Model** command is used to add a new model parameter set to the model list. This launches selecting the classification algorithm to use, specifying the algorithm-specific parameters and entering a model name. The command opens the **Choose Algorithm** dialog (Figure 9-35) to select the classifycation algorithm.

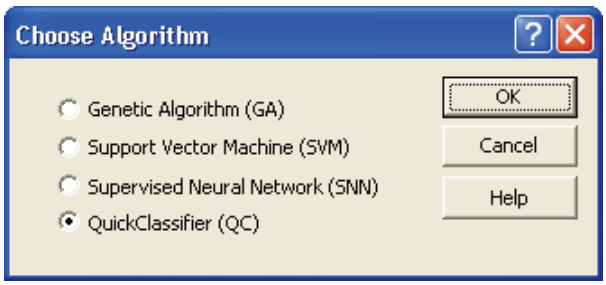


Figure 9-35 Choose Algorithm dialog (default setting)

Select the classification algorithm to be used to generate a new model:

Genetic Algorithm (GA). Uses the Genetic Algorithm.

Support Vector Machine (SVM). Uses the Support Vector Machine.

Note: For usage of the Support Vector Machine, a separate license is needed. This option is disabled when the license is not present.

Supervised Neural Network (SNN). Uses the Supervised Neural Network.

QuickClassifier (QC). Uses the QuickClassifier.

OK

Depending on the selected algorithm opens the corresponding dialog for setting algorithm-specific parameters: **Settings Genetic Algorithm** (Section 9.1.5.2.1), **Settings Support Vector Machine** (Section 9.1.5.2.2), **Settings Supervised Neural Network** (Section 9.1.5.2.3) or **Settings QuickClassifier** (Section 9.1.5.2.4) dialog.

Shortcut

Button: New...

9.1.5.2.1 Settings Genetic Algorithm Dialog

The **Settings Genetic Algorithm** dialog (Figure 9-36) defines the basic and advanced parameters for the GA. The settings are stored with the peak selection, cross validation, SVM, SNN and QC settings in the *SettingsModelGeneration.xml* file, which is updated on each settings change.

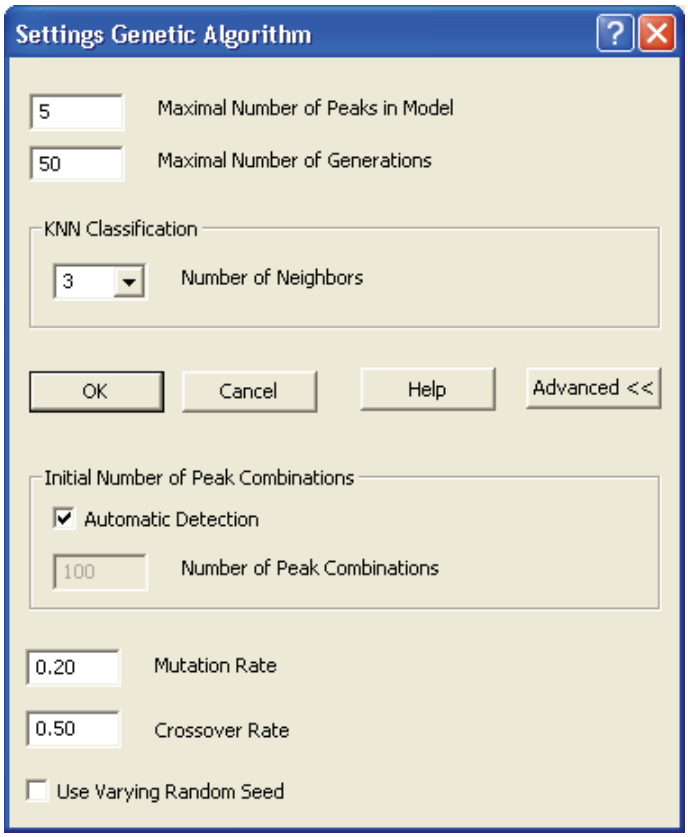


Figure 9-36 Settings Genetic Algorithm dialog (default setting)

Maximal number of Peaks in Model

Enter the maximal number of peaks included in the model.

Maximal Number of Generations

Enter the maximal number of generations (iterations) for the algorithm to run. Most of the time, this number will not be reached as the stop criteria will halt calculation when no better peak combination is found for a number of iterations.

In **KNN Classification**, define how to perform k-nearest neighbor classification (Section 6.2.2):

Number of Neighbors

Enter the number of neighbors ('k') to be used. Per default, 'k' can be set only to the odd values '1', '3', '5' and '7' which has been found to perform reasonable well on different data sets. The odd value ensures that in general a classification is obtained using k-NN (unclassified may still happen for e.g. three classes and k = 3, where two neighbors belong to different classes) and that the solution is sufficiently stable. The case of one neighbor (k = 1) should be used if the number of samples is very small. For a larger number of samples per class k > 1 is recommended.

Advanced >> / Advanced <<

Expands/Contracts the dialog to display/hide the advanced GA parameters.

In **Initial Number of Peak Combinations**, define how to determine the initial number of peak combinations within the population.

Automatic Detection

Check this option if the initial number of peak combinations should be determined automatically. To automatically determine the number of peak combinations (npc) the following heuristic formula is used:

NPC = 100 + (NumberOfPickedPeaks x 20) / (MaximalNumberOfPeaksInModel + 1)

Number of Peak Combinations

Enter the initial number of peak combinations if **Automatic Detection** is not set.

Mutation Rate

Enter the mutation rate, which is the likelihood of a mutation. In ClinProTools, a mutation is the random exchange of a peak within a peak combination by a randomly selected new one. The values can range from 0.0 (no mutation occurs) to 1.0 (all peak combinations are mutated in each generation).

Crossover Rate

Enter the crossover rate, which is the likelihood of a crossover between peak combinations. The values can range from 0.0 (no crossovers) to 1.0 (all peak combinations in each generation are used in crossover and are replaced by their children).

Use Varying Random Seed

Since the GA employs random numbers for selection, crossover and mutation, it is possible and quite likely that different values for most of the parameters (especially for **Crossover Rate** and **Mutation Rate**) may yield different solutions. To make comparisons between peak combinations possible, this randomness can be made the same for all peak combinations to be generated.

Check this option to seed the random number generator with a different value each time, so every model is using different random numbers. Uncheck this option if the GA should use the same number for initializing the random seed in any peak combination

to be calculated. This way, randomness is disabled and it is easier to study the effect of algorithm parameters.

OK

Opens the **Model Name** dialog (Section 9.1.5.2.5) to specify a name for the model.

9.1.5.2.2 Settings Support Vector Machine Dialog

The **Settings Support Vector Machine** dialog (Figure 9-37) defines the parameters for the SVM. The settings are stored as described for the GA settings (Section 9.1.5.2.1).

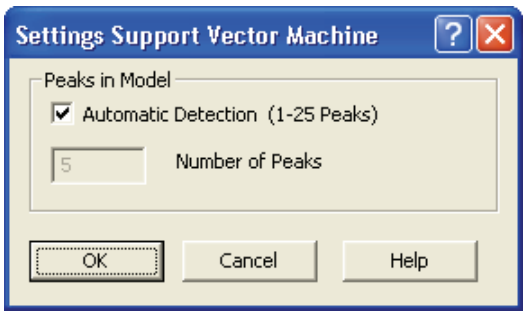


Figure 9-37 Settings Support Vector Machine dialog (default setting)

In **Peaks in Model**, define how the number of best peaks necessary for model generation should be determined:

Automatic Detection (1-25 Peaks)

Check this option if automatic peak detection (Section 6.2.1.5) should by performed which automatically determines the best number of peaks to be integrated in the model by an internal iteration. The search for the number of best peaks is restricted to maximal 25 peaks in a model.

Number of Peaks

If **Automatic Detection (1-25 Peaks)** is not set enter the number of peaks that must be integrated in the model.

OK

Opens the **Model Name** dialog (Section 9.1.5.2.5) to specify a name for the model.

9.1.5.2.3 Settings Supervised Neural Network Dialog

The **Settings Supervised Neural Network** dialog (Figure 9-38) defines the basic and advanced parameters for the SNN. The SNN automatically uses automatic peak detection (Section 6.2.1.5) to determine the best number of peaks to be integrated in

the model (maximum is 25 peaks). The settings are stored as described for the GA settings (Section 9.1.5.2.1).

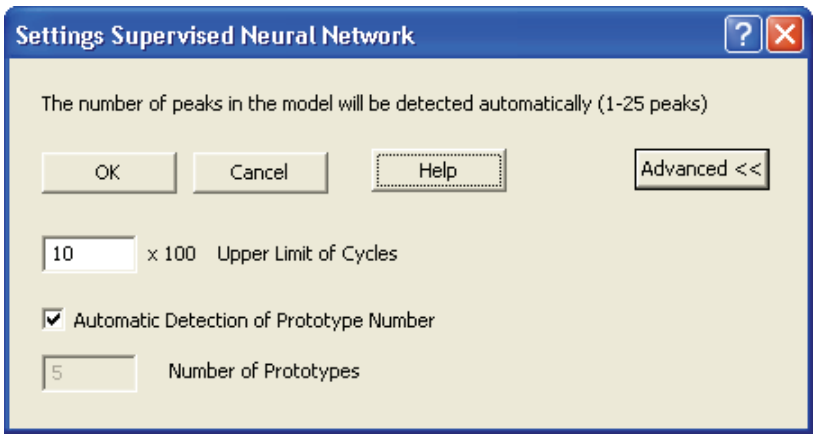


Figure 9-38 Settings Supervised Neural Network dialog (default setting)

Advanced >> / Advanced <<

Expands/Contracts the dialog to display/hide the advanced SNN parameters.

Upper Limit of Cycles

Enter a value (multiplied by 100) for the upper limit of cycles to run for optimizing the prototype positions. This number should be chosen with respect to the complexity of the data and can be evaluated considering the views, the number of picked peaks and the statistics.

Automatic Detection of Prototype Number

Check this option if automatic detection of prototype number should be applied. Uncheck it if a fixed number of prototypes should be used and specify the number in **Number of Prototypes**.

Number of Prototypes

Enter the number of prototypes to use if automatic detection should not be applied. The number of prototypes should be chosen with respect to the number of expected sub clusters in the data set and the overall data complexity.

9.1.5.2.4 Settings QuickClassifier Dialog

The **Settings QuickClassifier** dialog (Figure 9-39) defines the sort/weight mode for the QC. The QC automatically uses automatic peak detection (Section 6.2.1.5) to determine the best number of peaks to be integrated in the model (maximum is 25 peaks). This setting is stored as described for the GA settings (Section 9.1.5.2.1).

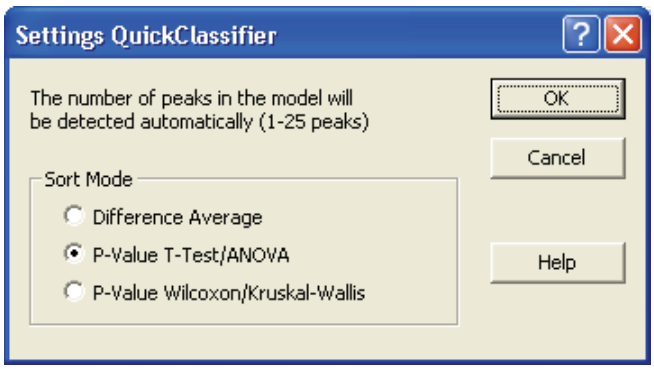


Figure 9-39 Settings QuickClassifier dialog (default setting)

Sort Mode

Select the sort mode used for peak ranking and as weight.

Difference Average. Sorts and weights the peaks by the difference between the maximal and the minimal average peak area of all classes.

P-Value T-Test/ANOVA. Sorts the peaks by the p-value (Section 6.4.1.6) from t-test (Section 6.4.1.1) / ANOVA test (Section 6.4.1.2).

P-Value Wilcoxon/Kruskal-Wallis. Sorts the peaks by the p-value from Wilcoxon test (Section 6.4.1.3) / Kruskal-Wallis test (Section 6.4.1.4).

OK

Opens the **Model Name** dialog (Section 9.1.5.2.5) to specify a name for the model.

9.1.5.2.5 Model Name Dialog

The **Model Name** dialog (Figure 9-40) is used to specify a name for the new model to be entered in the model list.

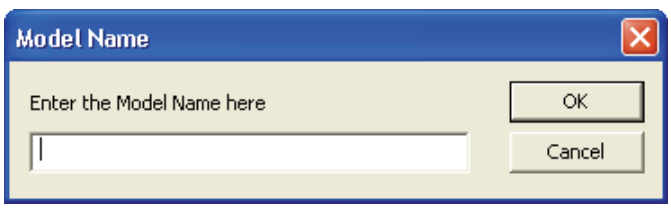


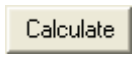
Figure 9-40 Model Name dialog

Entering a model name is optionally when the **Force Entering Model Name** option in the **General Settings** dialog (Section 9.1.1.12) is not set. This dialog is also opened when selecting the **Edit Model Name** command (Section 9.2.9.11) for changing the name of a parameter set in the model list that has still not been calculated.

9.1.5.3 Calculate Command

The **Calculate** command is used to calculate a model of the state 'Added' present in the Model List View. Model generation uses the peak calculation results of all included peaks of the single, non-excluded spectra and is based on the settings for the selected classification algorithm and the cross validation settings. The command calculates all 'added'-state models present in the list at once. If the recalibration, average peak list calculation or peak calculation workflow has not been performed when selecting this command the respective workflow(s) will be automatically run before model calculation starts. After model generation is completed, the corresponding model data is entered in the model list with changing the state of the model(s) into 'Calculated'.

Shortcut

Button: 

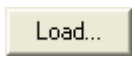
9.1.5.4 Cancel Command

The **Cancel** command is used to cancel any currently running spectra loading, recalibration, peak calculation, model generation or classification process. Same as **Cancel** command from **File** menu.

9.1.5.5 Load Model Command

The **Load Model** command is used to load the XML file of a model that has previously been saved with a specified name (Section 9.2.9.19). This allows performing classification or external validation. The command opens the **Load Model** dialog with the ClinProtModels folder opened by default. Navigate to the model you want to load and click **Open**. This enters the model in the model list with the state 'Loaded'.

Shortcut

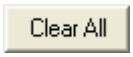
Button: 

9.1.5.6 Clear All Command

The **Clear All** command is used to clear the Model List View. This removes all items currently present.

Note: ClinProTools does not save models automatically. Thus, before selecting this command you should check which model(s) you want to keep and save it/them (Section 9.1.1.3).

Shortcut

Button: 

9.1.5.7 Settings Cross Validation Command

The **Settings Cross Validation** command is used to set the parameters for cross validation. Cross validation in ClinProTools requires that at least 20 not excluded spectra over all groups are available. This also applies to working with groups of spectra from multiple measurements (Section 6.1.3.2); here at least 20 groups must be available. The settings are stored as described for the GA settings (Section 9.1.5.2.1). The command opens the **Settings Cross Validation** dialog (Figure 9-41).

Note: If you change the cross validation settings when models of the state 'Calculated' are present in the Models List, the respective models are reset to the state 'Added' and have to be calculated again. This ensures that all models in the list are calculated based on the same cross validation settings.

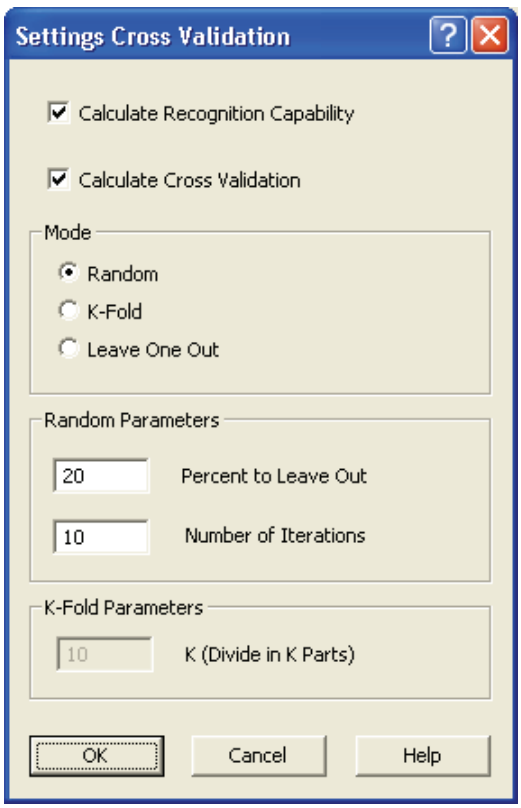


Figure 9-41 Settings Cross Validation dialog (default setting)

Calculate Recognition Capability

Check this option if the recognition capability of the generated model should be calculated. The recognition capability is one measure to describe the performance of a classification algorithm. It is calculated for a determined model as the relative number of correct classified data points by the classifier for the given model under the constraint that all tested data is previously used for the determination of the model or training of the classifier.

Calculate Cross Validation

Check this option if cross validation should be performed on the generated model using the cross validation procedure selected in **Mode**.

Note: It is strongly recommended to apply one kind of cross validation to verify that the obtained models give valid results on unseen data.

Mode

Select the mode for calculating cross validation:

Random. Selects a random subset of data points (taken over all classes) and omits it from the model generation procedure (Section 6.2.3). The parameters for this mode are specified under **Random Parameters**.

K-Fold. Divides the set of data points into *k* equal parts and generates *k* models where each time a different one of the *k* parts is omitted (Section 6.2.3). The parameter for this mode is specified under **K-Fold Parameters**.

Leave One Out. Leaves exactly one data point out and uses the remaining points for model generation (Section 6.2.3).

Note: In general, the choice of the cross validation mode depends on the number of available data points. For larger data sets, a **K-Fold** or **Random** approach is recommended. If the number of data points is rather small (e.g. less than 30 spectra per class) and it is expected that a high variation within each class exists it is more reliable to use the **Leave One Out** method since in that case more data points remain for the modeling stage.

In **Random Parameters**, define the **Random** cross validation mode if set:

Note: It is safe to keep these parameters with defaults.

Percent to Leave Out

Enter the percentage of data points to leave out per iteration.

Number of Iterations

Enter the number of iterations to perform.

In **K-Fold Parameters**, define the **K-Fold** cross validation mode if set:

K (Divide in K Parts)

Enter the number of parts to divide the set of data points.

OK
Changes the current cross validation settings. If the model list contains models of the state 'Calculated' these models will be reset to the state 'Added'.

9.1.5.8 Load Settings Model Generation Command

The **Load Settings Model Generation** command is used to load a stored model generation settings XML file which includes the peak selection, GA, SVM, SNN, QC and cross validation settings. The command opens the **Load Settings Model Generation File** dialog with the SettingsModelGeneration folder opened by default. Navigate to the file you want to load and click **Open**. This overwrites the current model generation settings with the loaded ones.

9.1.5.9 Save Settings Model Generation Command

The **Save Settings Model Generation** command is used to save the current model generation settings in an XML file with a specified name. The command opens the **Save Model Generation Settings File** dialog with the SettingsModelGeneration folder as the default storage location. Enter a file name or select one from the folder list and click **Save**. If you have selected an existing file name, answer the confirmation request to overwrite the file.

9.1.5.10 Reset Settings Model Generation Command

The **Reset Settings Model Generation** command is used to reset the current peak selection, GA, SVM, SNN, QC and cross validation settings to their defaults. The command opens a confirmation request to reset to defaults. Click **Yes** to reset current settings, click **No** to retain them.

9.1.6 Classification Menu

The **Classification** menu offers the following commands (Figure 9-42):

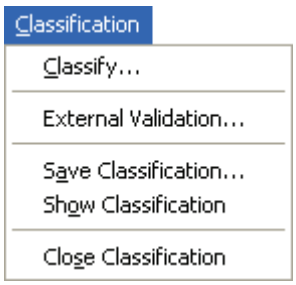


Figure 9-42 Classification menu

Command	Used to …
Classify	Classify the spectra in the selected collection with the chosen model.
External Validation	Validate the selected model externally using test spectra for each class.
Save Classification	Save the current classification result in an XML file with a specified name.
Show Classification	Show the classification result for the currently classified spectra in the Classification report.
Close Classification	Close the current classification and in non-batch classification mode unloads the classified spectra, too.

9.1.6.1 Classify Command

The **Classify** command is used to classify a selected spectra collection with the chosen model. All spectra are prepared and processed according to the parameter settings stored in the respective model. The classification workflow is run according to the active classification mode, standard or batch mode (Section 6.3).

The command opens the **Browse For Folder** dialog to select the spectra to be classified. Navigate to the folder of the respective spectra collection, select it and click **OK** to start classification. How the classification workflow proceeds depends on the active classification mode:

- In standard mode, the selected spectra are loaded in ClinProTools and display in the Spectra, Gel and Stack views with a black class color. After the classification, the Classification report (Section 8.1.1.8) shows the classification result; the report is stored as *ClinProtClassification[number].xml* file. The 2D Peak Distribution View displays the corresponding peak data for the classified spectra.

- In batch mode, the spectra to be classified are not displayed in the ClinProTools GUI. After classification is finished, the **Save Classification** dialog opens to save the classification result in an XML file with a specified name.

In both modes, the classification result is still held by the software as long as the classification is not closed.

Shortcut

Button: 

9.1.6.2 External Validation Command

The **External Validation** command is used to validate the selected model externally. For external validation (Section 6.2.4), you should use spectra of which you know the class membership but which were not used to generate the model. The command opens the **External Validation** dialog (Figure 9-43). For each class in the current model you have to select validation spectra. The data of the validation spectra is prepared as stored in the model.

The classification result for the validation spectra is shown in the Validation report (Section 8.1.1.7) and stored as *ClinProtValidation[number].xml* file. For a perfect classification, the confusion matrix would have entries only on the diagonal, which means that all validation data have been classified to their own class. In addition, Classification reports (Section 8.1.1.8) can be shown for each class in the model separately; the corresponding data is stored as *ClinProtClassification[number].xml* files.

Note: In the validation workflow, each spectrum of the collection will be used; there is no selection done by the noise spectra exclusion and adduct/polymer spectra exclusion or similarity selection filters (Section 6.1.3.2) - if there is something detected, the spectra are only marked but not excluded. Therefore it is recommended to use only suitable spectra for external validation.

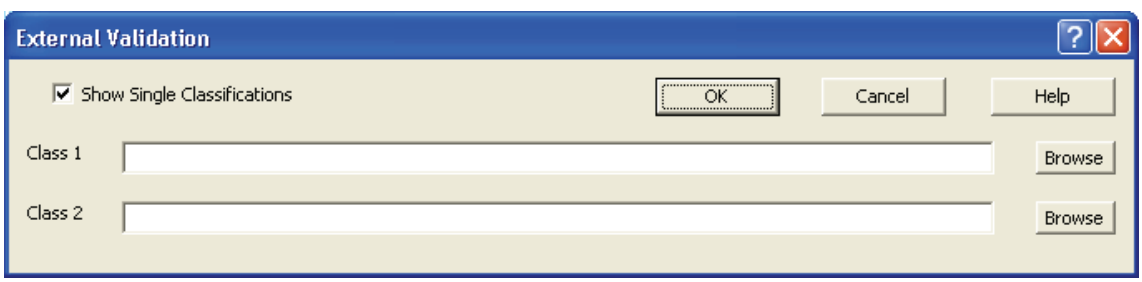


Figure 9-43 External Validation dialog for a two-class model (default setting)

Show Single Classifications

Check this option if a Classification report should be shown for each class in the model.

Class 1, **Class 2** … **Class n**

Enter name and path of the validation spectra for class 1, class 2, … class n. Alternatively, you can select the respective spectra via a browser dialog. For this, click Browse of the respective class entry box, navigate to the spectra and click **OK**.

OK

Classifies the validation spectra and shows the results in the Validation report. If defined, additionally a Classification Validation report is shown for each class in the model separately.

9.1.6.3 Save Classification Command

The **Save Classification** command is used to save the classification result for the loaded spectra collection in an XML file with a specified name When in standard mode, you have to select the command to open the saving dialog whereas in batch mode, the saving dialog opens automatically within the workflow after classification is finished. Saving the classification result is possible as long as you do not close the current classification. The command opens the **Save Classification** dialog with the ClinProt-Classification folder as the default storage location. Enter the file name or select one from the folder list and click **Save**. If you have selected an existing file name, answer the confirmation request to overwrite the file.

9.1.6.4 Show Classification Command

The **Show Classification** command is used to show the classification result for the classified spectra collection in the Classification report (Section 8.1.1.8). When in standard mode, you can use the command to show the result again if you have already closed the automatically created report. When in batch mode, you can use the command to create the Classification report as it is not automatically created within the workflow; however, it is not recommended to display big classifications because the browser used for display might take a long time to process the XML file with style sheet. Large XML files with style sheet should better be opened in Excel. Showing the classification result is possible as long as you do not close the current classification.

9.1.6.5 Close Classification Command

The **Close Classification** command is used to close the current classification. This removes the current classification result from the memory. In standard mode, it also unloads the classified spectra and removes them from the ClinProTools GUI.

Shortcut

Button: 

9.1.7 Statistical Analysis Menu

The **Statistical Analysis** menu offers the following commands (Figure 9-44):

Figure 9-44 Statistical Analysis menu

Command	Used to …
PCA	Perform a PCA on the loaded spectra.
Unsupervised Clustering	Perform an unsupervised clustering on the loaded spectra.

9.1.7.1 PCA Command

The **PCA** command is used to run the PCA workflow performing a PCA (Section 6.4.2.3) on the non-excluded spectra of the loaded spectra data set(s). This calculates a PCA and shows the PCA results in the PCA window. All generated PCA data is stored as *ClinProtPCA.xml* file in the ClinProTools folder. A PCA requires two valid spectra with three peaks being available at least. The command can be applied to several classes or to only a single class. However, in the context of PCA the separation into classes will be ignored, i.e. all data will be treated as one group.

The command opens the **PCA** dialog (Figure 9-45) to define whether the data should be normalized before running the PCA.

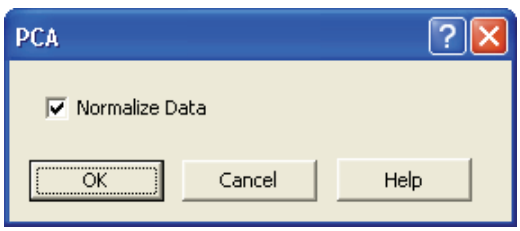


Figure 9-45 PCA dialog (default setting)

Clicking **OK** runs a PCA on the non-excluded spectra in the data set(s). If the spectra recalibration, average peak list calculation and/or peak calculation workflows have not been performed yet, the respective workflow(s) will be automatically run before starting PCA calculation. If the **Check Memory for PCA** option in the **General Settings** dialog is set, first the available memory is checked if it is sufficient for PCA on the loaded data set(s). After the PCA is completed, the PCA main window opens displaying the results of the PCA in the Scores and Loadings plots (Section 7.5.2).

Shortcut

Button: 

9.1.7.2 Unsupervised Clustering Command

The **Unsupervised Clustering** command is used to run the unsupervised clustering workflow performing an unsupervised hierarchical clustering (Section 6.4.2.4) on the non-excluded spectra of the loaded spectra data set(s). This calculates clusters from the spectra (= classes) and creates a dendrogram showing the distances among the single clusters. The corresponding data is stored in the files *ClinProtClustering.xml*, *ClinProtClusteringTree.xml* and *ClinProtClusteringTree2.xml* in the ClinProTools folder.

An unsupervised clustering needs three valid spectra with tree peaks being available at least. The command can be applied to several classes or to only a single class. However, in the context of unsupervised clustering the separation into classes will be ignored, i.e. all data will be treated as one group.

The command opens the **Unsupervised Clustering** dialog (Figure 9-46) to define the settings for unsupervised clustering and start clustering.

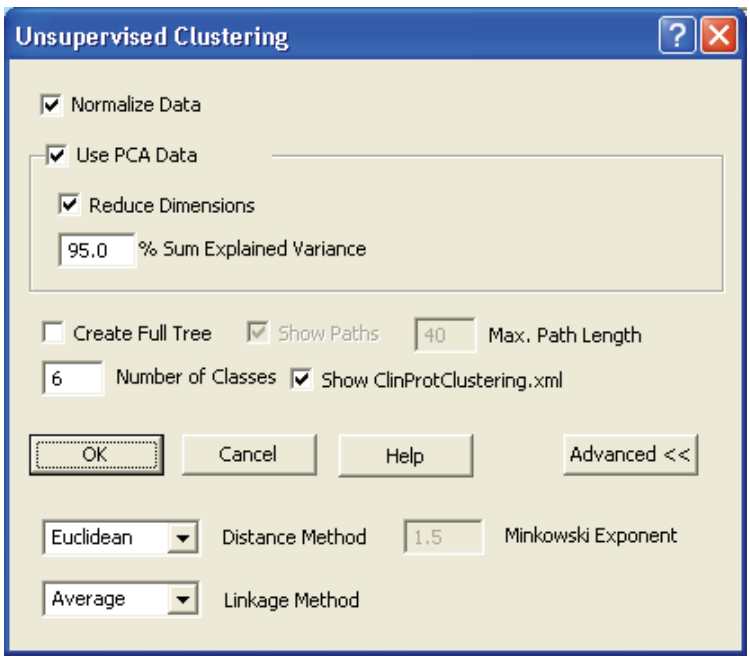


Figure 9-46 Unsupervised Clustering dialog (default setting)

Normalize Data

Check this option if the data should be normalized before running an unsupervised clustering.

Use PCA Data

Check this option if PCA transformed data should be used.

Reduce Dimensions

Check this option if only the first PCs necessary to explain a part of the variance should be regarded. The respective part of variance has to be specified under **% Sum Explained Variance**.

% Sum Explained Variance

Enter the minimum percentage of the sum of the variances the first PCs must have.

Create Full Tree

Check this option if the hierarchical clustering should create the complete dendrogram instead of a given maximum number of classes.

Show Paths

For **Create Full Tree** checked, check this option if the spectra paths should be shown at the end branches (singleton nodes).

Max. Path Length

Enter the upper limit of spectra path length if **Create Full Tree** and **Show Paths** are checked.

Number of Classes

For **Create Full Tree** unchecked, enter up to how many classes the clustering hierarchy should be calculated.

Show ClinProtClustering.xml

Check this option if the *ClinProtClustering.xml* should be launched in the browser if **Create Full Tree** not checked.

Advanced >> / <<

Shows/Hides the advanced parameters.

Distance Method

Select the metric to be used for distance calculation.

Euclidian. Uses Euclidian metric.
Minkowski. Uses Minkowski metric.
Cosine. Uses Cosine metric
Correlation. Uses correlation metric.
Spearman. Uses Spearman metric
Chebychev. Uses Chebychev metric.

Minkowski Exponent

For **Minkowski** metric chosen, enter the Minkowski exponent.

Linkage Method

Select the linkage method to be used for distance calculation.

Average. Uses average linkage.
Ward. Uses ward linkage.

Clicking **OK** then runs an unsupervised hierarchical clustering on the non-excluded spectra in the data set(s). If the spectra recalibration, average peak list calculation and/or peak calculation workflows have not been performed yet, the respective workflow(s) will be automatically run before the unsupervised clustering workflow starts. After clustering is completed, the Dendrogram window opens displaying the clustering result (Section 7.6.2).

Shortcut

Button: 

9.1.8 Reports Menu

The **Reports** menu offers the following commands (Figure 9-47):

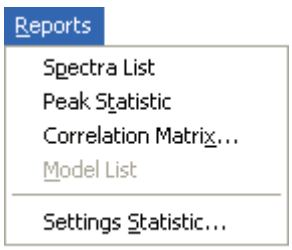


Figure 9-47 Reports menu

Command	Used to …
Spectra List	Show the Spectra List report.
Peak Statistic	Show the Peak Statistic report.
Correlation Matrix	Define correlation parameters and show the Correlation Matrix report.
Model List	Show the Model List report.
Settings Statistic	Define settings for calculating peak statistic and showing certain statistical data in the Spectra View.

9.1.8.1 Spectra List Command

The **Spectra List** command creates and shows the Spectra List report (Section 8.1.1.1) and stores the data as *ClinProtSpectraList[number].xml* file. The Spectra List report lists all loaded spectra with corresponding data.

9.1.8.2 Peak Statistic Command

The **Peak Statistic** command creates and shows the Peak Statistic report (Section 8.1.1.2) and stores the data as *ClinProtStatistic[number].xml* file. The calculation is based on the current statistic settings (Section 9.1.8.5). The Peak Statistic report lists all picked peaks with corresponding data. The 2D Peak Distribution View displays the first two peaks of the selected statistical sort order by default if currently active. If the spectra recalibration, average peak list calculation and/or peak calculation workflow(s) have not been performed yet, the respective workflow(s) will be automatically run before calculating the peak statistic.

Note: By default, the peak statistic is calculated using the same settings as the peak selection. However, you can use differing settings if desired.

Shortcut

Button: 

9.1.8.3 Correlation Matrix Command

The **Correlation Matrix** command is used to calculate a correlation analysis (Section 6.4.2.1) which compares each peak in the peak list to each other peak and to create and show the corresponding Correlation Matrix report (Section 8.1.1.3). The correlation analysis can be calculated over either all classes or only a specified one. The result is stored as *ClinProtCorrelationMatrix[number].xml* file. The settings for correlation matrix setup are stored in the *SettingsGeneral.xml* file, which is updated each time you change the correlation settings or any other settings saved to this file.

Note: Resetting the general settings also resets the current correlation settings.

The command opens the **Correlation Matrix** dialog (Figure 9-48) to specify how to set up the correlation matrix and to start correlation analysis. If the spectra recalibration, average peak list calculation and/or peak calculation workflows have not been performed yet, the respective workflow(s) will be automatically run before opening the dialog.

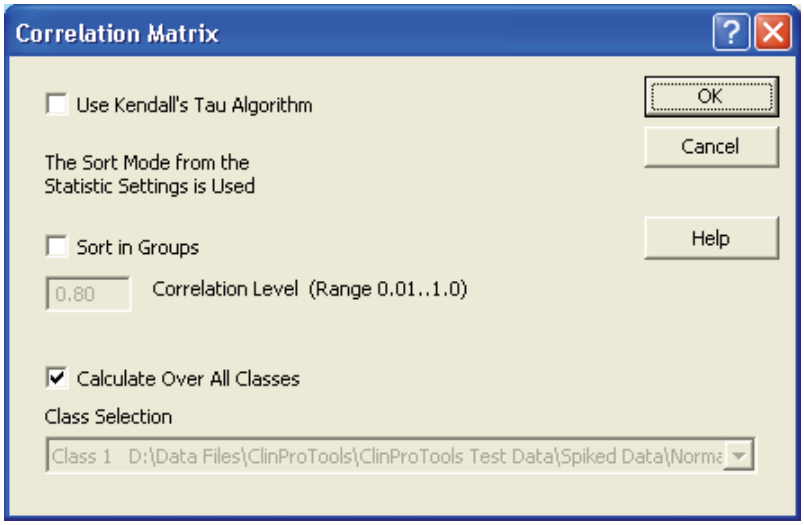


Figure 9-48 Correlation Matrix dialog (default setting)

Use Kendall's Tau Algorithm

Check this option if the Kendall's tau-b algorithm (Section 6.4.2.1) should be used for correlation analysis. Uncheck it to use the standard correlation algorithm.

Sort in Groups

Check this option if the peaks should be sorted in correlation groups.

Correlation Level (Range 0.01..1.0)

Enter a correlation level as absolute correlation value for building correlation groups. Recommended values are 0.7..0.95.

Calculate Over All Classes

Check this option if correlation should be calculated over all model generation classes. Uncheck it if you want to calculate correlation over only the class selected in **Class Selection**.

Class Selection

If **Calculate Over All Classes** is not checked, select from this list the model generation class over which you want to calculate correlation.

OK

Calculates the correlation analysis over all or the selected class and shows the results in the Correlation Matrix report.

9.1.8.4 Model List Command

The **Model List** command creates and shows the Model List report (Section 8.1.1.6) and stores the data as *ClinProtModelList[number].xml* file. The Model List report lists all models currently contained in the Model List View.

Shortcut

Button: Model List

9.1.8.5 Settings Statistic Command

The **Settings Statistic** command is used to define the settings to be used to calculate peak statistic. By default, the same settings are used as are defined for the peak selection (Section 9.1.5.1) but you can define differing settings, e.g. to set up a Peak Statistic report sorted by mass value. Furthermore, you can limit the number of peaks for which statistical data (average with standard deviation, peak distribution, and box and whiskers) is displayed in the Spectra and Single Peak Variance views when corresponding **View** menu commands are active. The statistic settings are stored in the *SettingsGeneral.xml* file, which is updated each time you change the statistic or any other settings saved to this file. The command opens the **Settings Statistic** dialog (Figure 9-49).

Note: Resetting the general settings also resets the current statistic settings.

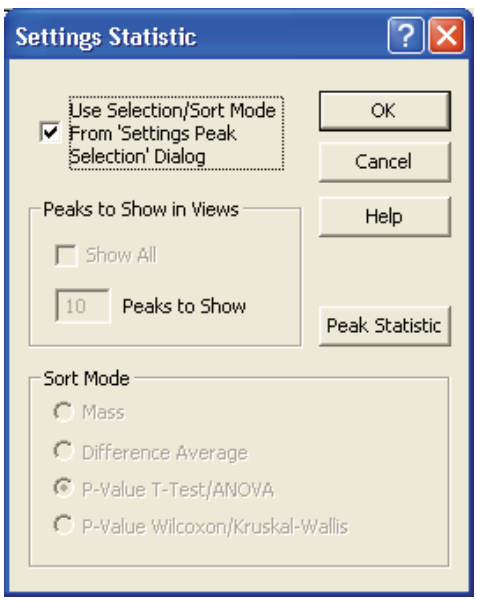


Figure 9-49 Settings Statistics dialog (default setting)

Use Selection/Sort Mode from the 'Settings Peak Selection' Dialog

Check this option if the selection/sort mode defined in the **Settings Peak Selection** dialog (Section 9.1.5.1) should also be used in peak statistic. Otherwise, uncheck this option and specify parameters as desired.

In **Peaks to Show in Views**, define whether statistical data in the Spectra and Single Peak Variance views should be displayed for all peaks or only a specified number of best peaks with respect to the selected sort mode:

Show All

Check this option if statistical data should be shown for all peaks. Uncheck it if you want to limit the number of peaks.

Peaks to Show

If **Show All** is not checked, enter the maximal number of peaks for which statistical data should be shown. The selected **Sort Mode** determines the order of peaks.

In **Sort Mode**, define how to sort the peaks in the Peak Statistic report if the current sort mode from peak selection should not be used. In addition, this setting defines the selection of peaks for which statistical data should be shown in the Spectra View if the peak number is limited.

Mass. Sorts by the *m/z* value.

Difference Average. Sorts the peaks by the difference between the maximal and the minimal average peak area of all classes.

P-Value T-Test/ANOVA. Sorts the peaks by the p-value from t-test (Section 6.4.1.1) / ANOVA test (Section 6.4.1.2).

P-Value Wilcoxon/Kruskal-Wallis. Sorts the peaks by the p-value from Wilcoxon test (Section 6.4.1.3) / Kruskal-Wallis test (Section 6.4.1.4).

Peak Statistic

Creates and shows the Peak Statistic report (Section 8.1.1.2) (same function as the **Peak Statistic** command from **Reports** menu).

OK

Changes the current statistic settings. If there are changes that concern the number of peaks for which statistical data is shown the Spectra View is updated accordingly.

9.1.9 Compass Menu

The **Compass** menu offers the following command (Figure 9-50):

Figure 9-50 Compass menu

Command	Used to …
LicenseManager	Launch the Bruker Daltonics LicenseManager.

9.1.9.1 LicenseManager Command

The **LicenseManager** command is used to view, add and delete licenses for Bruker Daltonics applications. It opens the **Bruker Daltonics LicenseManager** dialog (Figure 2-1) showing all licenses currently present for Bruker Daltonics applications. To add a new license, enter the license key you received in **New license key** and click **Add**. A new line is added to **Existing licenses** with the key you have entered, the product name and the date until the license will be valid. To delete an existing license select it in **Existing licenses**, click **Delete** and confirm the corresponding request.

Note: If the license key for the Support Vector Machine is entered when ClinProTools is started, a restart of ClinProTools is necessary to make the Support Vector Machine available.

9.1.10 Help Menu

The **Help** menu offers the following commands (Figure 9-51):

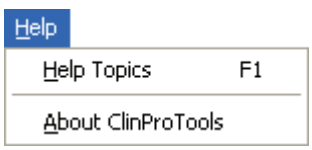


Figure 9-51 Help menu

Command	Used to …
Help Topics	Launch ClinProTools Help.
About ClinProTools	Show copyright and license information for your ClinProTools installation.

9.1.10.1 Help Topics Command

The **Help Topics** command launches ClinProTools Help which is used like other help applications running under Windows.

Shortcut

Key: F1

9.1.10.2 About ClinProTools Command

The **About ClinProTools** command shows copyright and license information for your ClinProTools installation (Figure 9-52).

Figure 9-52 About Bruker Daltonics ClinProTools dialog

9.2 ClinProTools Context Menus

9.2.1 Spectra View Context Menu

The Spectra View context menu offers the following commands:

Command	Used to …
Coordinates	Show/Hide the display of cursor coordinates in the status bar.
Grid	Show/Hide the grid in the view.
Scaling	Pop up scaling commands for the view.
Auto Scaling	Activate/Deactivate auto-scaling in the view.
Zooming	Activate/Deactivate the zoom in mode in the view.
Undo Zoom	Same as **Undo Zoom** command from **View** menu.
Redo Zoom	Same as **Redo Zoom** command from **View** menu.
Distance	Switch the view to distance measurement mode.
Display Mode	Pop up display modes for the view.
Background Color	Define the background color of the display region of views.
View Spectrum Info	Show the spectrum info for the selected spectrum.
Exclude / Include Spectrum	Same as **Exclude/Include Spectrum** command from **Edit** menu.
Exclude / Include Peak	Exclude/Include the selected peak in model generation.
Force Peak into Model	Force the selected peak into the next generated model.
Show Spectrum	If the peak distribution or the box and whisker plot with outliers is activated, show in the Spectra View the spectrum that corresponds to the right-clicked data point in peak distribution.
Add Peak	Add a new peak to the peak list.
Remove Peak	Remove the selected peak from the peak list.
Edit Peak	Change the integration region of the selected peak.
ROC Curve for Peak / Variance for Peak	Display in the ROC Curve view the ROC curve for the selected peak resp. in the Single Peak Variance View the variance for this peak.
Correlation List for Peak	Calculate per-peak correlation analysis for selected peak.

9.2.2 Gel View Context Menu

The Gel View context menu offers the following commands:

Command	Used to …
Coordinates	Show/Hide the display of cursor coordinates in the status bar.
Grid	Show/Hide the grid in the view.
Scaling	Pop up scaling commands for the view.
Zooming	Activate/Deactivate the zoom in mode in the view.
Undo Zoom	Same as **Undo Zoom** command from **View** menu.
Redo Zoom	Same as **Redo Zoom** command from **View** menu.
Distance	Switch the view to distance measurement mode.
Display Type	Pop up commands for toggling between Gel and Stack views.
Display Mode	Pop up display modes for the view.
Exclude/Include Spectrum	Same as **Exclude/Include Spectrum** command from **Edit** menu.

Right clicking the Gel View's color bar opens a context menu containing the same commands as offered by the **Display Mode** command.

9.2.3 Stack View Context Menu

The Stack View context menu offers the following commands:

Command	Used to …
Scaling	Pop up scaling commands for the view.
Display Type	Pop up commands for toggling between Gel and Stack views.
Whitewash	Switch the view to whitewash mode.
Background Color	Define the background color of the display region of views.

9.2.4 2D Peak Distribution View Context Menu

The 2D Peak Distribution View context menu offers the following commands:

Command	Used to …
Coordinates	Show/Hide the display of cursor coordinates in the status bar.
Grid	Show/Hide the grid in the view.
Scaling	Pop up scaling commands for the view.
Zooming	Activate/Deactivate the zoom in mode in the view.
Undo Zoom	Same as **Undo Zoom** command from **View** menu.

Command	Used to …
Redo Zoom	Same as **Redo Zoom** command from **View** menu.
Display Mode	Pop up display modes for the view.
Background Color	Define the background color of the display region of views.
Select Peaks	Same as **Peak Statistics View > 2D Options > Select Peaks** command from **View** menu.
Show Spectrum	Show in the Spectra View the spectrum that corresponds to the right-clicked data point.

9.2.5 ROC Curve View Context Menu

The ROC Curve View context menu offers the following commands:

Command	Used to …
Coordinates	Show/Hide the display of cursor coordinates in the status bar.
Grid	Show/Hide the grid in the view.
Display Mode	Pops up display modes for the view.
Background Color	Define the background color of the display region of views.

9.2.6 Single Peak Variance View Context Menu

The Single Peak Variance View context menu offers the following commands:

Command	Used to …
Coordinates	Show/Hide the display of cursor coordinates in the status bar.
Grid	Show/Hide the grid in the view.
Scaling	Pop up scaling commands for the view.
Auto Scaling	Activate/Deactivate auto-scaling in the view.
Zooming	Activate/Deactivate the zoom in mode in the view.
Undo Zoom	Same as **Undo Zoom** command from **View** menu.
Redo Zoom	Same as **Redo Zoom** command from **View** menu.
Display Mode	Pop up display modes for the view.
Background Color	Define the background color of the display region of views.
Show Spectrum	Show in the Spectra View the spectrum that corresponds to the right-clicked data point.

9.2.7 X/Y-Axes Context Menus

Right clicking on the x-axis or y-axis of a view opens a context menu offering the following commands:

Command	Used to …
Hide/Show X-Axis	Show/Hide the x-scale of the selected view.
Hide/Show Y-Axis	Show/Hide the y-scale of the selected view.
Axis Font	Define the axis font for all views.
Background Color	Define the background color of the axes.

9.2.8 Model List View Context Menu

The Model List View context menu offers the following commands:

Command	Used to …
Show Model	Show the selected model in the Model report.
Save Model As	Savesthe selected model as XML file with a specified name.
Remove Model	Remove the selected model from the view.
Edit Model Name	Edit the model name for a parameter set of an added model parameter list.
Classify	Same as **Classify** command from **Classification** menu.
External Validation	Same as **External Validation** command from **Classification** menu.
Show Error	Show the Error report for a model's 'ERROR' entry.

9.2.9 Commands Available from Context Menus Only

The following section describes commands available from context menus only (in alphabetical order).

9.2.9.1 Add Peak Command

The **Add Peak** command is used to manually add a new peak to the average peak list. Adding peaks is possible after average peak list calculation as well as after peak (statistic) calculation or model generation. In the latter cases, however, the current peak calculation will be reset, which is indicated in that the integration regions of all peaks change to gray color. This requires running the peak calculation resp. model generation workflow again.

The command displays the distance cursor (Figure 9-53). Move the vertical cursor lines, as described with the **Distance** command (Section 9.2.9.10), to the positions where the peak should start and end, and click the right mouse button. Confirm the appearing request to add a peak with the stated integration region (Figure 9-54). This displays the peak with a gray colored integration region. If the selected integration region overlaps with the integration region of an already existing peak, adding of the new peak is refused.

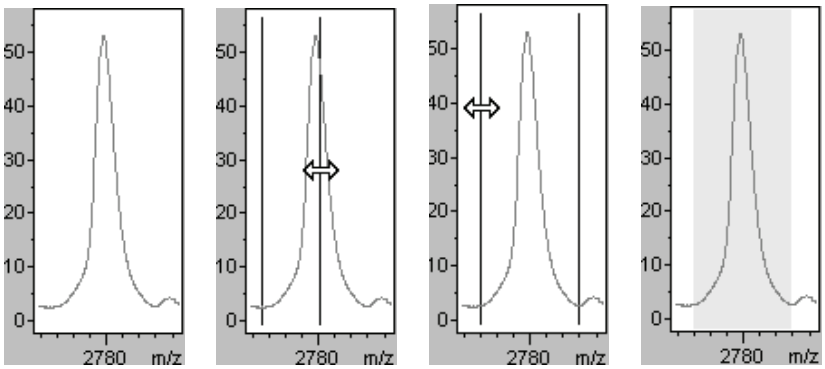


Figure 9-53 Adding a new peak manually using the Distance cursor

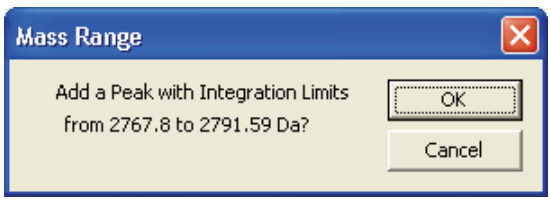


Figure 9-54 Mass Range dialog to confirm adding the stated peak

9.2.9.2 Auto Scaling Command

The **Auto Scaling** command activates/deactivates the auto-scaling mode in the Spectra View or Single Peak Variance View. When auto-scaling is active, the y-axis scaling is automatically adjusted to fully display the most intense peak in the current mass range (Spectra View) resp. the maximum statistic value of the current peak in the loaded classes (Single Peak Variance View).

9.2.9.3 Background Color Command

The **Background Color** command is used to change the background color of the display regions views. Selecting this command from the context menu of any of the axes allows changing the background color of the axes of all views. The command

opens the standard **Color** dialog to select the desired color from a list of **Basic colors** or defined **Custom colors**.

9.2.9.4 Coordinates Command

The **Coordinates** command shows/hides the cursor coordinates in the status bar (Section 5.1.6). When the coordinates mode is active and the cursor is positioned in a data plotting view the corresponding x- and y-data is displayed in the status bar. The data shown depends on the focused view, the processing state and the cursor position with respect to peak position.

9.2.9.5 Correlation List for Peak N Command

The **Correlation List for Peak n** command is used to calculate a correlation analysis (Section 6.4.2.1) which compares the selected peak to each other peak in the peak list and to create and show the corresponding Correlation List report (Section 8.1.1.4). The per-peak correlation can be calculated over either all classes or only a specified one. The result is stored as *ClinProtCorrelationList[number].xml* file. The settings for correlation list setup are stored in the *SettingsGeneral.xml* file, which is updated each time you change the correlation settings or any other settings saved to this file.

Note: Resetting the general settings also resets the current correlation settings.

The command opens the **Correlation List** dialog to specify how to set up the correlation List and to start correlation analysis. For descritption of the parameters, please refer to the **Correlation Matrix** command (Section 9.1.8.3). Clicking **OK** calculates the correlation analysis over all or the selected class and shows the results in the Correlation List report.

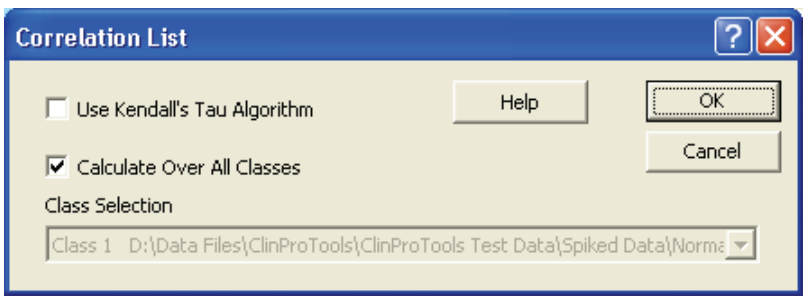


Figure 9-55 Correlation List dialog (default setting)

9.2.9.6 Display Mode Command (Gel View)

The Gel View's **Display Mode** popup offers commands to define the color scheme and intensity mode of the view.

9.2.9.7 Display Mode Command (2D Peak Distribution, ROC Curve, Single Peak Variance Views)

The Peak Statistic View's **Display Mode** popup offers commands to define the line width of points in the views. They can be useful for printing if lines are too thin.

9.2.9.8 Display Mode Command (Spectra View)

The Spectra View's **Display Mode** popup offers commands to define how data points are displayed and connected in the view.

9.2.9.9 Display Type Command

The **Display Type** popup offers commands to toggle between Gel and Stack View.

9.2.9.10 Distance Command

The **Distance** command switches the Spectra View resp. the Gel View to distance mode, which allows determining *m/z* differences between two selected points in a spectrum. The distance cursor displays in that view where distance measurement was launched. Distance measurement behaves similar in both views, but the shape of the distance cursor differs; the procedure described below concerns the Spectra View. When you have finished distance measurement, you can deactivate the distance mode by clicking the right mouse button.

Distance measurement in Spectra View

When activating this mode the distance cursor appears in the center of the Spectra View (Figure 9-56). This cursor consists of two vertical lines and a two-headed arrow. One line is fixed whereas the other is moveable and follows the mouse. The *m/z* position of the fixed line and the measured *m/z* difference from fixed to movable line are displayed as 'X' and 'dX' values in the status bar. You can switch from fixed to moveable line by clicking the left mouse button.

To measure a distance, position the moveable line on the point of the spectrum where you want to start measurement and click the left mouse button. This fixes the moveable line at the selected position and switches the previously fixed line to movable (Figure

9-57). Now place the second line on the point of the spectrum where you want to end your distance measurement (Figure 9-58). While moving the line the displayed *m/z* difference value is continuously updated. The *m/z* difference is given as positive or negative value depending on the current direction in which you move the moveable line with respect to the fixed line. The absolute difference value between both points is always the same. You can change the current sign by a left click (Figure 9-59).

Illustration of *m/z* distance measurement in the Spectra View:

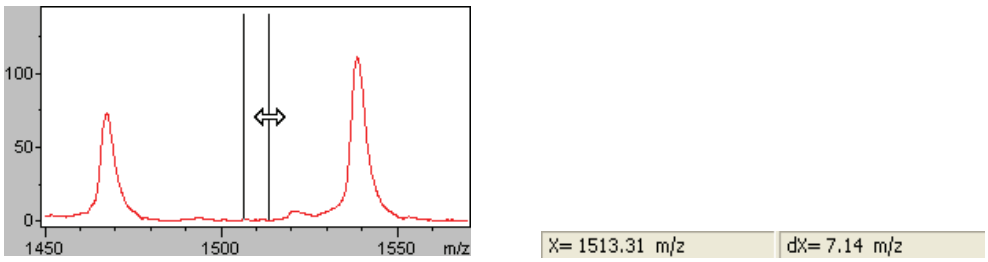


Figure 9-56 Display of distance cursor in the Spectra View after selecting the Distance command

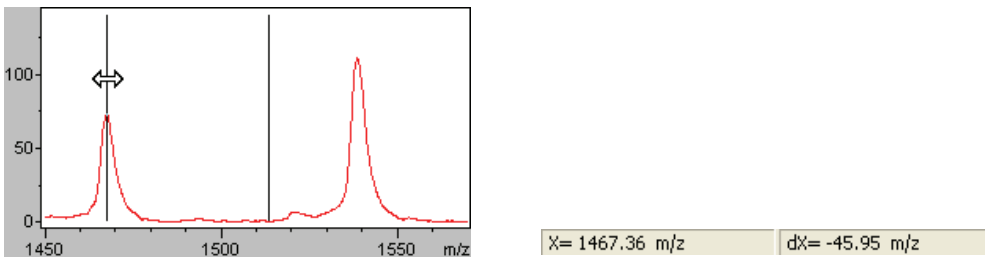


Figure 9-57 Positioning the moveable line on the point where distance measurement should start

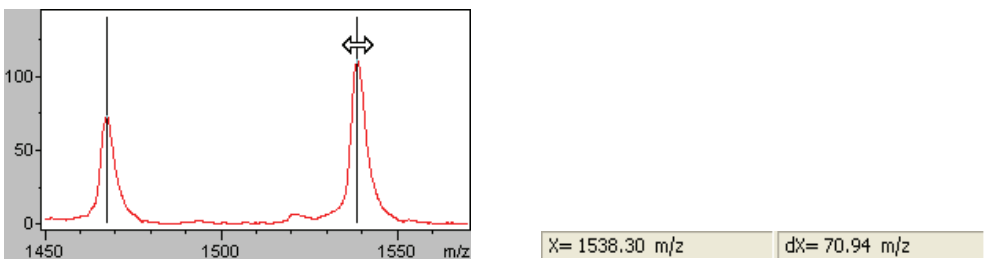


Figure 9-58 Switching fixed and moveable lines and then positioning the now moveable line on the point where distance measurement should end

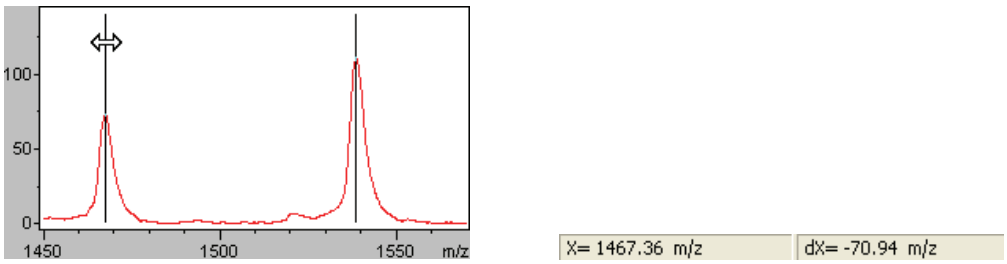


Figure 9-59 Left-clicking again changes the sign of the difference value accordingly

9.2.9.11 Edit Model Name Command

The **Edit Model Name** command is used to edit the name of a model parameter set in the model list. This allows entering a name if no name was specified when adding the parameter set or changing the current name. Editing the model name is possible as long as model calculation is not started. The command opens the **Model Name** dialog (Figure 9-40) to specify a model name. Clicking **OK** enters the (new) model name in the model list.

9.2.9.12 Edit Peak N Command

The **Edit Peak n** command is used to change the current integration region of the selected peak. Editing peaks is possible after average peak list calculation as well as after peak (statistic) calculation or model generation. In the latter cases, however, the current peak calculation will be reset, which is indicated in that the integration regions of all peaks change to gray color. This requires running the peak calculation resp. model generation workflow again.

The command displays the distance cursor (Figure 9-60). Its two vertical lines mark the current limits of the integration region of the selected peak. Move the cursor lines, as described with the **Distance** command (Section 9.2.9.10), to the new positions where the peak should start and end, and click the right mouse button. Confirm the appearing request to change the integration region of the selected peaks as stated (Figure 9-61). This changes the integration region of the peak accordingly.

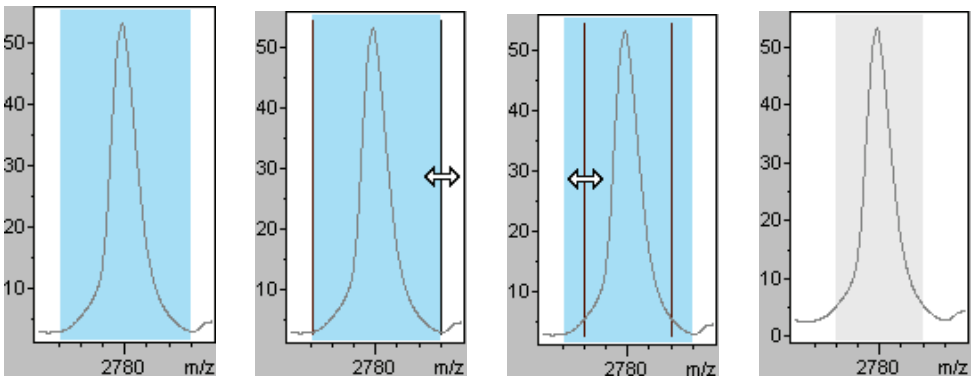


Figure 9-60 Changing the integration region of a peak using the Distance cursor

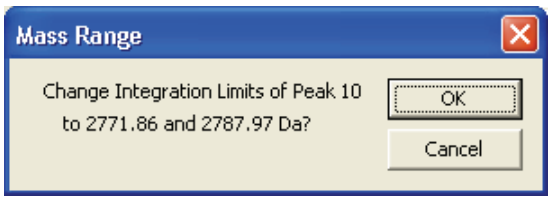


Figure 9-61 Mass Range dialog to confirm changing of integration limits

9.2.9.13 Exclude / Include Peak N Command

The **Exclude/Include Peak n** command excludes or includes, respectively, the selected peak in model generation. Peaks can be excluded/included after running the peak calculation workflow. All included peaks have a blue integration region in the Spectra View, all excluded peaks a gray one (Figure 9-62). Only included peaks are passed on to the algorithms. In the Peak Statistic report (Section 8.1.1.2) excluded peaks are indicated by a '-' entry in the **S** (= state) column (first column).

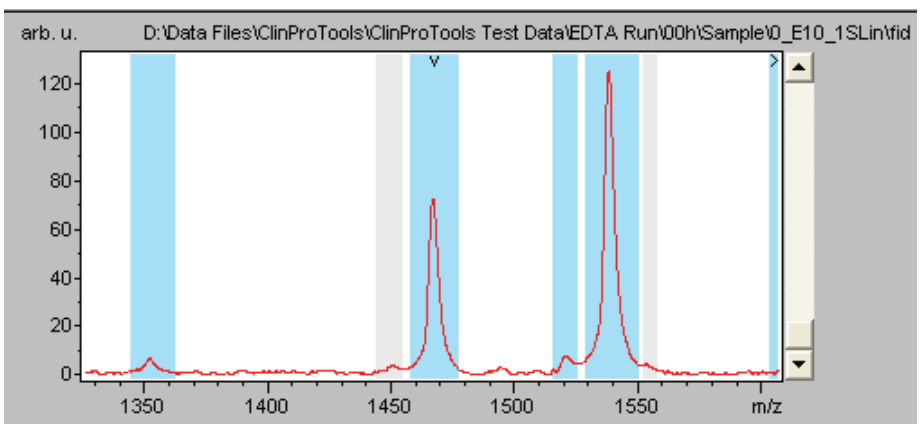


Figure 9-62 Display of included (blue) and excluded (gray) peaks in the Spectra View

9.2.9.14 Force Peak N into Model Command

The **Force Peak n into Model** command forces the selected peak into the model to be generated. Peaks can be forced after running the peak calculation workflow. A forced peak is marked by a green integration region before as well as after model generation. Forcing a peak into a model can be canceled by selecting the command for the respective peak again.

9.2.9.15 Grid Command Command

The **Grid** command shows/hides the grid in a data plotting view. This command applies to the selected view only. The grid properties cannot be changed.

9.2.9.16 Remove Model Command

The **Remove Model** command removes the selected model from the Model List View.

Note: Please remember that models are not automatically saved in ClinProTools. Thus, if you have calculated a new model you should first consider if you want to save it (Section 9.2.9.19) before removing it.

9.2.9.17 Remove Peak N Command

The **Remove Peak n** command removes the selected peak from the average peak list. Removing peaks is possible after average peak list calculation as well as after peak (statistic) calculation or model generation. In the latter two cases, however, the current peak calculation will be reset, which is indicated in that the integration regions of all peaks change to gray color. This requires running the peak calculation resp. model generation workflow again.

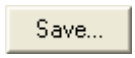
9.2.9.18 ROC Curve for Peak N Command

The **ROC Curve for Peak n** command displays the ROC curve (Section 6.4.2.2) for the selected peak in the ROC Curve View. The command is only enabled if this view is active. Whether class 1 or class 2 is currently treated as positive depends on the decision made when switching to ROC Curve View via the **Peak Statistics View > ROC Curve** command from the **View** menu.

9.2.9.19 Save Model As Command

The **Save Model As** command is used to save the selected model in an XML file with a specified name. The command opens the **Save Model** dialog with the ClinProtModels folder as the default storage location. Enter a new or select an existing model name and click **Save**. If you have selected an existing name, answer the confirmation request to overwrite the file.

Shortcut

Button: 

9.2.9.20 Scaling Command

The **Scaling** popup offers commands for changing the display range of a data plotting view:

Command	Used to …
Expand manually	Change axes scaling in the data plotting views based on values entered in the **Manual Scaling** dialog (see below).
Times 2	Decrease the y-range by 2.
Divide by 2	Increase the y- range by 2.
Offset plus	Shift the y-range up.
Offset minus	Shift the y-range down.
Expand	Expand the x- range
Contract	Contract the x-range.
Move left	Move the x-range to the left.
Move right	Move the x-range to the right.
Reset	Reset x- and y-range to full display of data.

Manual Scaling dialog

Use the **Manual Scaling** dialog (Figure 9-63) to manually change the scaling of x- and/or y-axis in a view. The dialog differs depending on the view whre it was launched. For the Gel View, you can also change the scale of intensity axis and for the Stack View the scale of the z-axis (= spectrum number axis). You can enter new values or reset the current values to full display of data in the respective view.

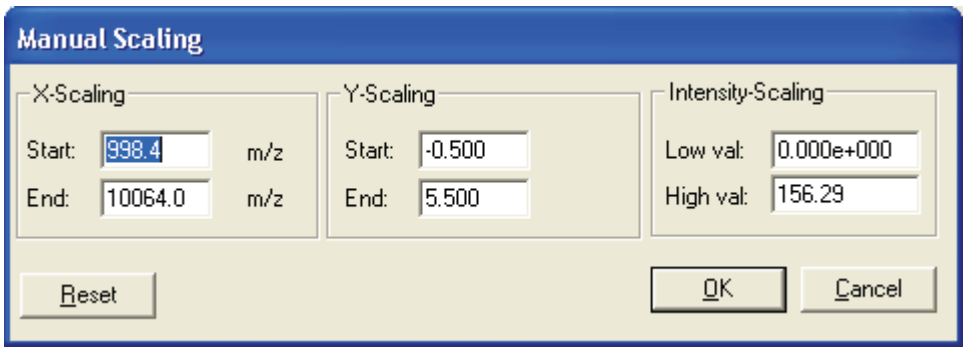


Figure 9-63 Manual Scaling dialog (here for the Gel View)

9.2.9.21 Show Error Command

The **Show Error** command shows the Error report (Section 8.1.1.9) if the state of the selected model has been set to 'ERROR' during model generation.

9.2.9.22 Show Model Command

The **Show Model** command creates and shows the Model report (Section 8.1.1.6) for the selected model and store the data as *ClinProtModel[number].xml* file.

Shortcut

Button: Show

9.2.9.23 Show Spectrum Command

The **Show Spectrum** command shows in the Spectra View that spectrum that corresponds to the data point you selected in the Spectra View, 2D Peak Distribution View or Single Peak Variance View by right-clicking on or close to the data point. The command is only available if a data point was right-clicked.

9.2.9.24 Variance for Peak N Command

The **Variance for Peak n** command displays statistic data (average with standard deviation, peak distribution and/or box and whiskers for the area/intensity of the peak) for the selected peak. This command is available when the Single Peak Variance View is active.

9.2.9.25 View Spectrum Info Command

The **View Spectrum Info** command is used to show specific information about the selected single spectrum. The command opens the **Spectrum Information** dialog which displays the information stored for the current spectrum (Figure 9-64). To view information about another spectrum you can keep the dialog open and just select another spectrum in the Spectra View.

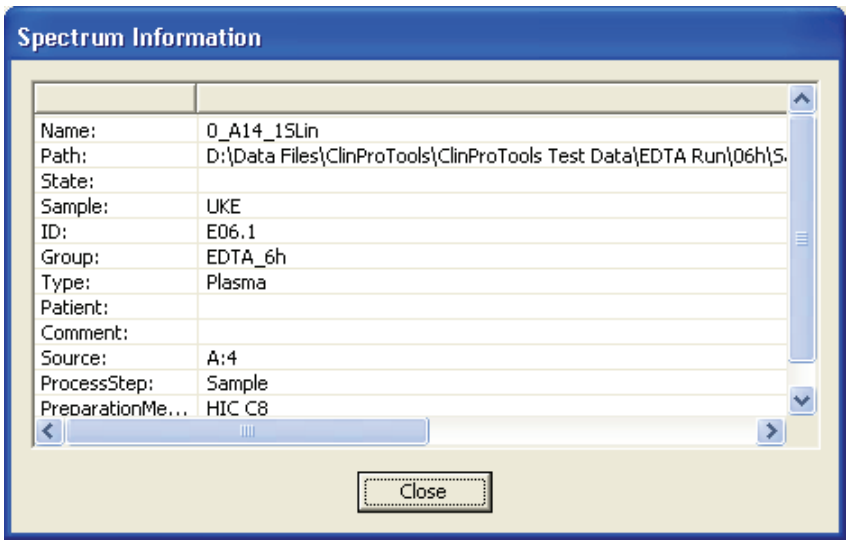


Figure 9-64 Spectrum Information dialog

9.2.9.26 Whitewash Command

The **Whitewash** command switches the Stack View to whitewash mode. In this mode, the plot is structured finer due to resolving overlying structures. All spectra are drawn in black color; thus, their class membership is not shown (Figure 9-65).

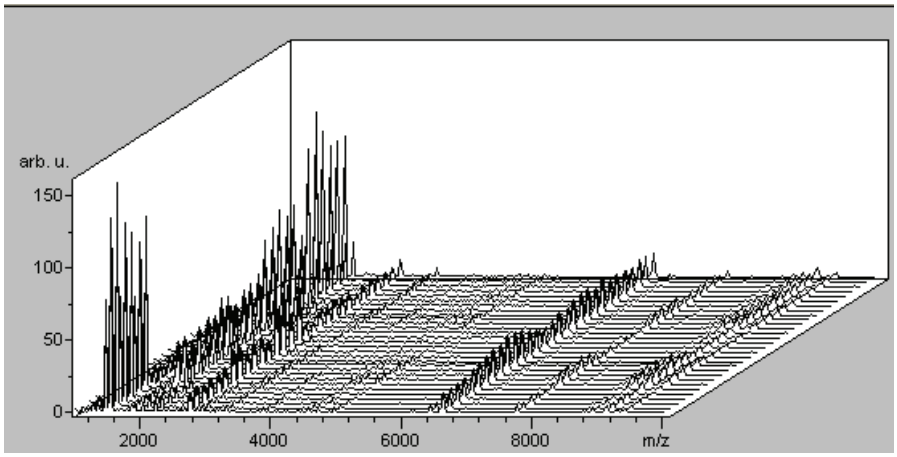


Figure 9-65 Stack View in whitewash mode

9.2.9.27 Zooming Command

The **Zooming** command activates/deactivates the Zoom for the Spectra, Gel, 2D Peak Distribution or Single Peak Variance View to zoom in the selected range (Section 5.1.7.2). When this command is active for a view, the Zoom in cursor is displayed when you move the mouse in the respective view. Otherwise, the mouse cursor is displayed.

9.3 MATLAB Based Menus

The PCA windows and the Dendrogram window provide own menus originating from the external MATLAB tool integrated in ClinProTools. The menus and commands available depend on the particular window.

9.3.1 Edit Menu

The **Edit** menu of a PCA window or the Dendrogram window offers the following command:

Command	Used to …
Copy	Copy a graphic of the focused MATLAB based window to the clipboard.

9.3.1.1 Copy Command

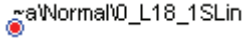
The **Copy** command copies a metafile graphic of the focused MATLAB based window to the clipboard. This allows pasting that graphic into an appropriate application.

9.3.2 View Menu

The **View** menu of a PCA window or the Dendrogram window can offer the following commands:

Command	Used to …
Mark Data Points	Switch to marking data points mode (applicable to PCA plots only).
Zoom	Switch to zoom mode (applicable to 2D plots only).
Pan	Switch to pan mode (applicable to 2D plots only).
Rotate 3D	Switch to rotation mode (applicable to 3D PCA plots only).

9.3.2.1 Mark Data Points Command

The **Mark Data Points** command switches to the marking data points mode. This mode allows marking data points in the Scores plots with corresponding file description (e.g. 

) and in the Loadings plots with peak (*m/z*) description (e.g.

). When this mode is active, the first left mouse button click on a data point marks the respective point and a second click on that point removes the attached information again.

9.3.2.2 Zoom Command

The **Zoom** command switches to zoom mode which allows zooming operations (zoom in/out, reset zooming) in the 2D plots. To zoom in a 2D plot, move the mouse cursor into the desired plot. Position the zoom in cursor at the desired start point and draw it holding the left mouse button pressed to the desired end point. On releasing the mouse button, the enclosed area is zoomed in. The zooming-in steps you perform in a plot are stacked for each plot separately which allows stepwise zooming out of the plots. Double clicking with the left mouse button restores the plot's original view.

When the command is active, right clicking a 2D plot pops up a context menu offering the following commands:

Command	Used to …
Zoom Out	Stepwise zoom out the selected plot if it was zoomed in before.
Reset to Original View	Restore the plot's original view.
Zoom Options	
Unconstrained Zoom	Allow unconstrained zooming.
Horizontal Zoom	Allow zooming in horizontal direction only.
Vertical Zoom	Allow zooming in vertical direction only.

9.3.2.3 Pan Command

The **Pan** command switches to the pan mode. This mode allows moving all data points within a 2D plot as an entity in any direction. This e.g. enables longer descriptions attached to data points to be read. To pan a plot, move the mouse cursor into the desired plot so that it changes into the pan cursor . Click the plot with the left mouse button and while holding the mouse button pressed, shift the data points in the desired direction. Double clicking with the left mouse button restores the plot's original view.

When the command is active, right clicking a 2D plot pops up a context menu offering the following command:

Command	Used to …
Reset to Original View	Restore the plot's original view.

9.3.2.4 Rotate 3D Command

The **Rotate** command switches to the rotation mode that allows rotating the 3D plots. To rotate a 3D plot, move the mouse cursor into the desired plot so that it changes into the rotation cursor . Click the plot with the left mouse button and while holding the mouse button pressed rotate the plot in the desired direction. Double clicking with the left mouse button restores the plot's original view.

When the command is active, right clicking a 3D plot pops up a context menu offering the following command:

Command	Used to …
Reset to Original View	Restore the plot's original view.

9.3.3 Plots Menu

The **Plots** menu of the PCA main window offers the following commands:

Command	Used to …
Variance	Display the Variance plot in the Variance window.
Influence	Display the Influence plot for the chosen PC number in the Influence window.
PC 3D	Display the 3D Scores plot in the PC 3D window.
PC A	Display the left 2D Scores plot in the PC A window.
PC B	Display the middle 2D Scores plot in the PC B window.
PC C	Display the right 2D Scores plot in the PC C window.
Loadings 3D	Display the 3D Loadings plot in the Loadings 3D window.
Loadings A	Display the left 2D Loadings plot in the Loadings A window.
Loadings B	Display the middle 2D Loadings plot in the Loadings B window.
Loadings C	Display the right 2D Loadings plot in the Loadings C window.

9.3.3.1 Variance Command

The **Variance** command displays the Variance plot of the PCA in the Variance window.

9.3.3.2 Influence Command

The **Influence** command is used to display the Influence plot for a specified number of PCs. The command opens the **Influence** dialog (Figure 9-66) to set the number of PCs

to be concerned. The number of PCs needed to explain 95% of the variance in the spectra set is suggested by default. Clicking **OK** creates the corresponding Influence plot and shows it in the Influence window.

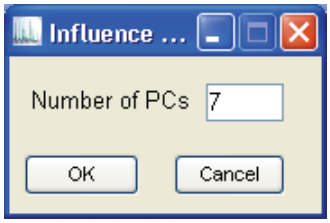


Figure 9-66 Influence dialog

9.3.4 PC Menu

The **PC** menu of the PCA main window offers the following command:

Command	Used to …
PCs	Select the PCs for which corresponding data should be displayed in Scores plots and Loadings plots.

9.3.4.1 PCs Command

The **PCs** command is used to select the PCs for which you want to view data in the Scores plots and Loadings plots. It opens the **PCs** dialog (Figure 9-67) to enter the desired PC numbers. Clicking **OK** updates the data in the plots of the PCA main window accordingly; plots previously set up in a separate window remain unchanged.

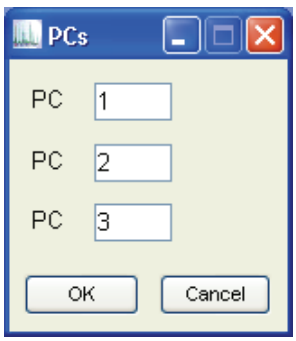


Figure 9-67 PCs dialog

10 ERROR TREATMENT

Error reports

In the case that an error occurs, please send an error report to:

clinprot.support@bdal.de

The error report should contain the following:

- The dump file *.dmp from C:\BDALSystemData\Dumpfiles if a message box has appeared with 'a dump file has been generated'.
- If a message box titled 'ACO' or 'SCO' pops up with an error, the log file C:\BDALSystemData\BFADSLOG.TXT; an additional screen shot is also good
- The ClinProTools build number (from the **About** box)
- A short description of the workflow
- Is the error reproducible, if yes, how?
- The data resp. which kind of data have been used
- The settings files *SettingsDataPreparation.xml*, *SettingsModelGeneration.xml* und *SettingsGeneral.xml* from the ClinProTools folder
- In the case of an 'ERROR' entry in the Model List View, select the **Show Error** command from the view's context menu: the Internet Explorer pops up and the file can be saved with the Explorer's **Save As** command.

Reset of ClinProTools in case of error message

After encountering an error message, the ClinProTools software could be in an intermediate state and it might be necessary to reset the software using the **Close All** command from the **File** menu.

External shutdown of ClinProTools in case of need

The running ClinProTools application can be shut down externally with the help of the *Windows Task-Manager*. It is available via the shortcut Strg+Alt+Del. Mark the ClinProTools process in 'Applications' or 'Processes' and click 'End Task' resp. 'End Process'.

If the application crashes, the process might still be running. The process has to be shut down externally; otherwise, it is not possible to re-launch the application. This might also be the case, if you find it impossible to start the application.

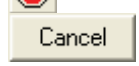
In the case of a dead lock (processing does not end and cannot be cancelled) during data preparation or model generation, the application has also to be shut down externally.

If you try to load a very large amount of data it might happen, that the computer runs out of memory. In this case, processing slows down extremely or comes to a standstill. Canceling does now longer work then and the process has to be shut down externally.

A APPENDIX

A.1 Quick Reference on Menus, Commands, Tool Buttons and Shortcuts in ClinProTools 2.2

The following table lists the menus available in ClinProTools 2.2 and their commands. The corresponding toolbar button and shortcut, if available and the meaning are also included.

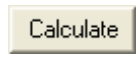
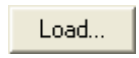
Menu commands	Used to …	Button	Shortcut
File menu			
Open Model Generation Class	Open the selected model generation class.		Ctrl+O
Open Spectra Import XML	Open the selected spectra import XML file and load the referenced spectra.		Ctrl+I
Cancel	Cancel any current loading/calculation/ model generation/ classification process.	 	
Close All	Close and unload all spectra and models.		
Info Loaded Classes	Show paths information about the loaded spectra collections.		
Save Class Paths	Save the paths of the loaded model generation classes as spectra import XML file.		Ctrl+S
Print	Print a graphic of the active data plotting view.		Ctrl+P
Print Preview	Preview the graphic of the active data plotting view.		
Print Setup	Set up the printer and printing options.		
Peak List Export	Export the peak list to XML or CART format.		
Browse ClinProTools Folder	Browse the ClinProTools folder.		Shift+Alt +O
General Settings	Define general ClinProTools settings.		
Exit	Close ClinProTools.		

Menu commands	Used to …	Button	Shortcut
Edit menu			
Copy	Copy a bitmap and/or a metafile graphic of the selected data plotting view to the clipboard.		Ctrl+C
Exclude/Include Spectrum	Exclude/Include the selected spectrum.		
Bitmap to Clipboard	Activate/Deactivate the bitmap format for copying graphics to the clipboard.		
Metafile to Clipboard	Activate/Deactivate the metafile format for copying graphics to the clipboard.		
View menu			
General Toolbar	Show/Hide the General toolbar.		
View Toolbar	Show/Hide the View toolbar.		
Status Bar	Show/Hide the status bar.		
Undo Zoom	Undo the last zoom range change in the selected view.		
Redo Zoom	Redo the last undone zoom range change in the selected view.		
Spectra View >	Pop up the following commands for the Spectra View:		
> Single Spectra	Show/Hide the single spectra.		
> All Single Spectra	Show/Hide the overlaid display of all single spectra.		
> Total Average Spectrum	Show/Hide the total average spectrum.		
> Average Spectra	Show/Hide the class average spectra.		
> Noise Spectrum	Show/Hide the noise spectrum.		
> Integration Regions	Show/Hide the integration regions.		
> Average & StdDev.	Show/Hide the average with standard deviation.		
> Peak Distribution	Show/Hide the 1D peak distribution.		
> Box & Whiskers	Show/Hide the box and whiskers.		
> Outliers for Box & Whiskers	Show/Hide the outliers for box & whiskers plots.		
> Peak Markers	Show/Hide the peak markers.		
Gel/Stack View >	Pop up the following commands for the Gel/Stack View:		
> Class Names	Show/Hide the class names in Gel View.		

Menu commands	Used to …	Button	Shortcut
> Current Spectrum Marker	Show/Hide the current spectrum marker in Gel View.		
> Colored Spectrum State	Mark/Do not mark the spectrum state with modified colors in Gel View.		
> Excluded Spectra	Show/Hide the excluded spectra in Gel/Stack View.		
> Group Separators	Show/Hide the group separators in Gel View.		
> Follow Spectra View Mass Range	Force/Do not force the x-axis of Gel/Stack View to follow the Spectra View mass range.		
Peak Statistics View >	Pop up the following commands:		
> 2D Peak Distribution	Switch to 2D Peak Distribution View and display the 2D peak distribution for two selected peaks.		
> ROC Curve	Switch to ROC Curve View and display the ROC curve for the selected peak.		
> Single Peak Variance	Switch to Single Peak Variance View and display peak statistic for the selected peak.		
> Outliers for Box & Whiskers	Show/Hide the outliers for box & whiskers plots in the Single Peak Variance View.		
> Options			
> Select Peaks	Select two peaks to display in the 2D Peak Distribution View.		
> 95% Confidence Interval	Display the 95% confidence interval or the standard deviation.		
> Current Spectrum Marker	Mark/Do not mark the data point that corresponds to the current spectrum.		
Reset View Settings	Reset certain current view settings to the defaults.		

Data Preparation menu

Settings Spectra Preparation	Define the spectra preparation and recalibration settings.		
Settings Peak Calculation	Define the average spectra and peak calculation settings.		
Load Settings Data Preparation	Load the selected data preparation settings XML file.		

Menu commands	Used to …	Button	Shortcut
Save Settings Data Preparation	Save the current data preparation settings as an XML file with a specified name.		
Reset Settings Data Preparation	Reset the current data preparation settings to their defaults.		
Recalibration	Recalibrate spectra and calculates average spectra.		
Average Peak List Calculation	Calculate average peak list.		
Peak Calculation	Pick peaks, calculate peak areas and peak statistic.		
Model Generation menu			
Settings Peak Selection	Define the peak selection settings.		
New Model	Add a new model parameter set to the model list.		
Calculate	Start model generation.		
Cancel	Cancel any current loading/calculation/ model generation/ classification process.		
Load Model	Load the selected model.		
Clear All	Clear the model list.		
Settings Cross Validation	Define the cross validation settings.		
Load Settings Model Generation	Load the selected model generation settings XML file.		
Save Settings Model Generation	Save the current model generation settings as an XML file with a specified name.		
Reset Settings Model Generation	Reset current model generation settings to their defaults.		
Classification menu			
Classify	Classify the spectra in the selected collection with chosen model.		
External Validation	Validate the selected model externally using test spectra for each class.		
Save Classification	Save the current classification result in an XML file with a specified name.		

Menu commands	Used to …	Button	Shortcut
Show Classification	Show the classification result for the classified spectra in the Classification report.		
Close Classification	Close the current classification.		
Statistical Analysis menu			
PCA	Perform PCA on the spectra data set(s).		
Unsupervised Clustering	Perform hierarchical clustering on the spectra data set(s).		
Reports menu			
Spectra List	Create and show the Spectra List report.		
Peak Statistic	Create and show the Peak Statistic report.		
Correlation Matrix	Create and show the Correlation Matrix report.		
Model List	Create and show the Model List report		
Settings Statistic	Define settings for calculating peak statistic and showing certain statistical data in the Spectra View.		
Compass menu			
LicenseManager	Launch Bruker Daltonics LicenseManager.		
Help menu			
Help Topics	Launch ClinProTools online help.		F1
About ClinProTools	Display copyright and license information about the present ClinProTools installation.		

A.2 Glossary

Class

A class is a set of spectra originating from samples e.g. of the same disease state. The model generation classes have to be sorted e.g. according to the state of disease and are used to generate a model which will be applied to explain the class membership.
ClinProTools loads all spectra in a folder and its subfolders recursively as one class. If the ClinProtRobot is used and multiple spotting takes place, it is important to switch on the **Support Spectra Grouping** option.

Classification

Classification means the determination of the class membership of a given spectrum.

Classification model

A classification model is the result of generating a model. It contains data preparation characteristics as well as classifier characteristics. It can be used for classification of spectra of unknown status. It may be saved (as XML file) to be reloaded later.

Classifier algorithm

A classifier means an algorithm used in generating ***Classification models***. ClinProTools offers four classifier algorithms, the ***Genetic Algorithm***, the ***Support Vector Machine***, the ***Supervised Neural Network*** and the ***QuickClassifier***.

Correlation analysis

The correlation analysis is used to analyze stochastic relations between random variables upon a given sample set. In our context, the random variable is given by an individual peak and its properties (peak area), and the sample set is the given set of spectra. A correlation matrix resp. correlation list is set up as the result of correlation analysis.

Crossover

Crossover means the combination of two randomly selected individuals to produce two new individuals by interchanging parts during GA.

Cross validation

This type of ***Validation*** should be used for automatic validation during model generation. During cross validation a small part of all spectra is left out in model generation and cluster analysis. These spectra are then classified, and the number of correct and wrong class predictions is determined. This procedure is repeated several times, and the correct and wrong class predictions are accumulated for each class.

External validation

This type of ***Validation*** uses a separate test set which has not been used for generating the ***Classification model***. For all spectra of the test set the true class mem-

bership has to be known. During validation, all spectra of the test set are classified. The predicted class memberships are compared to the true class memberships and ***Sensitivity*** and ***Specificity*** can be calculated.

Generation

During one generation of a ***Genetic Algorithm*** a new ***Population*** is created.

Genetic Algorithm (GA)

The Genetic Algorithm is a stochastic search algorithm, which mimics evolution in nature. It is used for the optimization of an objective function (called fitness function) for a large number of solutions, which we call peak combinations. It considers many possible peak combinations simultaneously.

Individual

An individual is the entity that is artificially evolved by the ***Genetic Algorithm***. An individual consists of a set of peaks, which is used for ***k-NN Classification*** to determine the selective power of this set. Individuals can be ***mutated*** or subjected to ***Crossover***. The individual that is deemed to be best at differentiating the model generation spectra is returned from the Genetic Algorithm as the classification model.

K-nearest neighbor (k-NN) classification

The k-nearest neighbor (k-NN) classifier algorithm is used within the ***Genetic Algorithm*** and ***Support Vector Machine*** to obtain the final classification. It just uses the distances between points in the n-dimensional space. The peak selection is derived from the current GA peak combination or the final SVM peak ranking solution. The idea of k-nearest neighbor classifiers is to look at the k-nearest neighbors and their spectra class membership.

Model generation data

Model generation data are classes of spectra, which have been used for the generation of certain models.

Multiple measurement

A multiple measurement is a measurement of the same sample by applying multiple spotting on the target. The obtained spectra, e.g. four spectra of the same sample, are in general quite similar and must be considered in a common sense. The ClinProt measurement software automatically collects multiple measurements in a common folder named by the sample_id. In general, multiple measurements are used to reduce the risk of measure.

Mutation

The random modification of an ***Individual*** during the ***Genetic Algorithm***. In ClinProTools, one peak in the individual is replaced by another randomly selected peak.

Normalization

The scales for all the features of the spectra are rescaled to one standard.

Over fitting

Over fitting means that an obtained classification model performs much better on the model generation classes than on the test data. In general, this is an indicator that some parameters during model generation are too strong adapted to the specifics of the model generation data.

Population

A population means a large collection of ***Individuals*** within the ***Genetic Algorithm***. This may include hundreds to several thousands of individuals.

Principle Component Analysis

Principle Component Analysis (PCA) is a broadly used mathematical technique designed to extract, display and rank the variance within a data set. The overall goal of PCA is to reduce the dimensionality of a data set while simultaneously retaining the information present in the data. In ClinProTools, the PCA reduces the number of dependent variables contained within the spectra set via replacing groups of variables by a single new variable. By this, a set of new variables, so-called principal components (PCs) is generated. In many cases (depending on the complexity of the data set), only few PCs (compared to the large number of original variables) contain most of the variance.

P-value

A p-value is the probability that an observed effect is simply due to chance; it therefore provides a measure of the strength of an association. A p-value does not provide any measure of the size of the effect, and cannot be used in isolation to inform clinical judgment. P-values are affected both by the magnitude of the effect and by the size of the study from which they are derived, and should therefore be interpreted with caution. In particular, a large p-value does not always indicate that there is no association and, similarly, a small p-value does not necessarily signify an important clinical effect.

QuickClassifier

The QuickClassifier (QC) is a univariate sorting algorithm based upon classical test statistic. It determines characteristics for each peak upon its statistical properties. These characteristics are used to set up a model for later classification.

Recalibration

Recalibration means the alignment of a number of spectra using the most prominent peaks.

Receiver Operating Characteristic

The Receiver Operating Characteristic (ROC) curve gives a graphical overview about ***Specificity*** and ***Sensitivity*** of a test or, within ClinProTools, an evaluation of the discrimination quality of a peak. The ROC curve is an exploration of what happens to the true positives and the false positives if the position of an arbitrary threshold is varied. This arbitrary cut-off point splits the values into a fraction representing a positive test result (values above the point) and a fraction representing a negative test result (values below the point). In ClinProTools, ROC curves can only be generated for the case of two model generation classes because a true/false decision is not possible for more than two classes.

Recognition capability

The recognition capability is one measure to describe the performance of a ***Classifier***. It is calculated for a determined model as the relative number of correct classified data points by the classifier for the given model under the constraint that all tested data is previously used for the determination of the model or training of the classifier. In other words, the recognition capability indicates how good a determined model is able to classify the data, which is used for model generation.

If the recognition capability is low, the classifier was not able to learn the underlying data characteristic. This may happen if it is not possible to determine a relation between the properties of the data and the given labeling. A high recognition capability however does not necessarily mean that the data are separable or the model is good. If e.g. all data are learned by heart it is easy to predict the class label for this data if they are represented but for unknown data the model may fail to predict the labeling.

Selection

The fittest ***Individuals*** are selected and the less capable ones are abandoned during ***Genetic Algorithm*** processing. This is done by optimizing a cost function, which aims on optimal class separation with high variance between classes. Using the cost function each peak combination is rated by an expense factor, which is used as a measure for the fitness.

Sensitivity

The sensitivity is the percentage of correctly classified positives. If your aim is to identify diseased people, sensitivity is the ability to correctly identify those who have the disease (the proportion of people with a disease who have a positive test result). This measurement can be derived from the model if a two-class scenario is analyzed.

Specificity

The specificity is the percentage of correctly classified negatives. If your aim is to identify diseased people, specificity is the ability to correctly identify those who do not have the disease (the proportion of people without disease who have a negative test result). This measurement can be derived from the model if a two-class scenario is analyzed.

Supervised Neural Network

The Supervised Neural Network (SNN) is a prototype-based classification algorithm. The SNN tries to identify some characteristic spectra for each class which are named prototypes and which could be somehow considered as prototypical samples of that class.

Support Vector Machine

The Support Vector Machine (SVM) is an algorithm for the determination of optimal separating planes between different data classes. It uses formal approaches from optimization theory to separate the given data sets. Upon the obtained planes, a peak ranking can be calculated in a second step.

Test data

Test data are spectra to be classified by the software, using a model containing peak patterns generated by the model generation process.

Type I and Type II errors

Type I error and Type II error are common measurements in statistical testing and described in the following table:

		Expected decision (Reality)	
		Positive Class	Negative Class
Machine / Test decision	Positive Class	True positive (correct decision)	False negative (type I error / false decision type I)
	Negative Class	False positive (type II error / false decision type II)	True negative (correct decision)

Unsupervised clustering

The hierarchical clustering approach is used as an unsupervised clustering method in ClinProTools. A hierarchy of clusters is generated, which is represented in the form of a dendrogram (a tree).

Validation

After generation of a model, this needs to be validated. Validation, in case of a two-class scenario, yields estimates of ***Sensitivity*** and ***Specificity***, which describe the quality of the model. Two types of validation are used by the software, ***Cross validation*** and ***External validation***.

A.3 Abbreviations

ANOVA	analysis of variance
AUC	area under the ROC curve
cc	correlation coefficient
cv	cross validation
FN	false negative
FNF	false negatives fraction
FP	false positive
FPF	false positives fraction
GA	Genetic Algorithm
GB	Gigabyte
GUID	globally unique identifier
I/O	input/output
k-NN	k-nearest neighbor
KT	Kendall's tau-b algorithm
mm	multiple measurement
npc	number of peak combinations
PCA	principle component analysis
p-value	probability value
QC	QuickClassifier algorithm
RAM	random access memory
rc	recognition capability
ROC	Receiver Operating Characteristic (also Receiver Operating Curve)
SNN	Supervised Neural Network algorithm
SRM	structural risk minimization
SVM	Support Vector Machine algorithm
TIC	total ion count
TOF	time of flight
TN	true negative
TNF	true negatives fraction
TP	true positive
TPF	true positives fraction
XML	extensible markup language

A.4 Data Exchange Formats

ASCII Import

We support Ciphergen ASCII format. The format will be detected automatically. The ASCII files in the spectra collection folder must all have the suffix ".csv", the first line must contain "*M/Z*,Intensity" while the following lines must contain the *m/z*-intensity pairs separated by a comma:

M/Z,Intensity
92.665657,16.709906
92.853842,16.697838
93.042196,17.022505
93.230717,16.194437
93.419405,14.812662
93.608261,13.874493

XML Spectra Import

Details of the *ClinProtSpectraImport.xml* import format with path list of spectra to be loaded for statistic calculation and model generation are as follows:

```
<ClinProtSpectraImport Version="0.0">
<Class Name="Class A">
  <Element Path="C:\A\0_M16_1SLin\fid"/>
  <Element Path="C:\A\0_M17_1SLin\fid"/>
  <Element Path="C:\A\0_M18_1SLin\fid"/>
</Class>
<Class Name="Class B">
  <Element Path="C:\B\0_M19_1SLin\fid"/>
  <Element Path="C:\B\0_M20_1SLin\fid"/>
</Class>
</ClinProtSpectraImport>
```

It is also possible to provide only class paths (Name attribute is ignored in this case):

```
<ClinProtSpectraImport Version="0.0">
<Class Path="C:\A"/>
<Class Path="C:\B"/>
</ClinProtSpectraImport>
```

Mixed formats containing both class and spectra paths are not allowed.

In both cases, an optional RGB attribute can be set changing the displayed class color:

```
<ClinProtSpectraImport Version="0.0">
<Class Path="C:\A" RGB="255 128 255"/>
```

```
<Class Path="C:\B" RGB="128 128 0"/>
</ClinProtSpectraImport>
```

CART (ASCII) Peak List Export

Details of the CART (ASCII) format (*.dat; by Salford Systems, San Diego, CA, USA) are as follows: The first line contains the column header. The column headed 'No' contains an ongoing index while 'Class' contains the class number the spectrum belongs to. The following columns are all prefixed with 'A_' for peak area resp. 'I_' for peak intensity. The mass of the peak is coded into the column title after the prefix, with the decimal dot exchanged by an underscore.

This is an example of the content of a CART export file (four peaks, two classes, three spectra in class 1, two spectra in class 2) where calculated peaks areas were exported:

"No","Class","A_1467_95","A_1623_92","A_1911_27","A_2779_25"

0,1,190.1123,147.5010,224.8864,296.0563
1,1,198.4078,135.5611,223.3063,290.2748
2,1,195.7452,130.1778,217.7962,299.5370
3,2,207.2164,142.4266,228.4644,289.9513
4,2,183.8811,125.1170,211.8117,282.1209

XML Peak List Export

Peak list export to XML supports three XML formats, <ClinProToolsPeakLists>, <ClinProToolsPeakLists2/> and <ClinProToolsPeakLists3/>.

Details of the **XML Files** format are as follows:

```
<ClinProToolsPeakLists>
    <Masses/>      List of peak masses
    <ClassPaths/>Paths of spectra folder
    <SpectraPaths>
        <Class/>       List of spectra paths per class
    </SpectraPaths>
    <Areas>
        <Class/>       List of peak lists (areas) per class
    </Areas>
    <Intensities>
        <Class/>       List of peak lists (intensities) per class
    </Intensities>
</ClinProToolsPeakLists>
```

The **XML2 Files** format generates an alternative XML format <ClinProToolsPeak-Lists2/> with the class and spectra paths provided as attributes.

The **XML3 Files** format <ClinProToolsPeakLists3/> is similar to **XML2 Files** but a style sheet reference for ClinProtPeakList.xsl is added to facilitate working with peak lists in Excel.

A.5 Part Numbers

249614 Software-Package ClinProTools 2.2
249620 License ClinProTools 2.2
245575 License Support Vector Machine 1.0
249619 ClinProTools User Manual

I INDEX

D

E

F

G

H

I

K

L

M

Q

R

S

T

U

V

W

X

Y

Z