



Mestrelab Research

BioHOS 3.1

MANUAL



Document Number

P/N 242 R2

COPYRIGHT

©2023 MESTRELAB RESEARCH S.L.

All rights reserved. No parts of this work may be reproduced in any form or by any means - graphic, electronic, or mechanical, including photocopying, recording, taping, or information storage and retrieval systems - without the written permission of the publisher.

Products that are referred to in this document may be either trademarks and/or registered trademarks of the respective owners. The publisher and the author make no claim to these trademarks.

While every precaution has been taken in the preparation of this document, the publisher and the author assume no responsibility for errors or omissions, or for damages resulting from the use of information contained in this document or from the use of programs and source code that may accompany it. In no event shall the publisher and the author be liable for any loss of profit, or any other commercial damage caused or alleged to have been caused directly or indirectly by this document.

CONTENTS

1. INSTALLATION AND LICENSING	4
2. NEW FEATURES SINCE VERSION 3.0.....	5
3. THE RIBBON INTERFACE	5
4. GENERAL	5
5. WORKSPACE RESULTS DISPLAY	6
6. DATA IMPORT AND PROCESSING	6
6.1. DIRECT IMPORT OF PROCESSED DATA.....	8
6.2. STACKED ITEMS.....	8
6.3. CLASSES: IDENTIFICATION AND COLORING.....	10
6.4. PHASING AND BASELINE CORRECTION	11
6.5. BLIND REGIONS AND CUTS	11
6.6. SOLVENT SIGNAL REMOVAL	13
6.7. DENOISE BY VOI COMPRESSION	14
6.8. REFERENCE ALIGNMENT.....	15
7. DATA ANALYSIS	15
7.1. 2D PEAK PICKING	15
8. ECHOS ANALYSIS.....	17
8.1. POINTS FITTING AND CORRELATION COEFFICIENT (R)	17
8.2. VIEWING THE RESIDUALS.....	18
8.3. EXPORT DATA	19
9. CCSD COMPARISON	19
9.1. REFERENCE	20
9.2. CCSD RESULTS	20
9.3. MANUAL REMOVAL OF POINTS.....	22
10. 1D PROFILE	23
10.1 MATRIX ANALYSIS	27
11. MULTIVARIATE STATISTICS (PCA, SIMCA, AND PLS)	28
11.1. SELECTING THE SPECTRA FOR THE ANALYSIS.....	28
11.2. DATA PREPARATION.....	29

11.2.1	<i>Binning</i>	29
11.2.2	<i>Data Processing</i>	30
11.2.3	<i>PCA Options</i>	32
11.2.4	<i>SIMCA Options</i>	34
11.2.5	<i>PLS Options</i>	35
11.3.	PCA RESULTS.....	36
11.4.	SIMCA RESULTS.....	49
11.5.	PLS RESULTS.....	52

The evaluation of protein higher-order structure (HOS) has been recognized as a key analytical method in the manufacture of monoclonal antibody (mAb) therapeutics and other recombinant proteins. There is a considerable wealth of literature to support the use of NMR for the assessment¹, a reference standard is available², and a round-robin evaluation was successfully performed³. An up-to-date list of key literature is available in Bruker's Pharma Journal Club⁴.

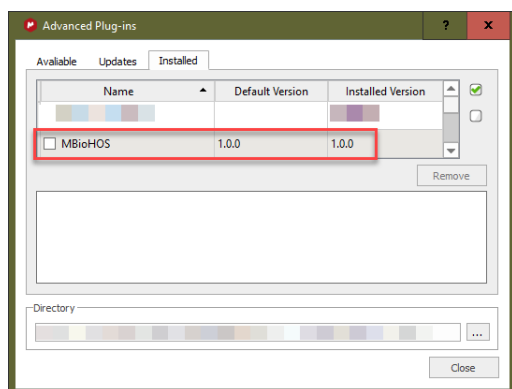
Published methods to assess mAb HOS using NMR generally use one of these approaches:

- “Fingerprint” spectral comparison
- Peak movements
- Chemometric analysis

In all cases, representative “reference” spectra (or spectrum) are required against which a test spectrum is evaluated.

1. Installation and licensing

MestReNova (Mnova) versions ≥ 14.3 have the “BioHOS” capability included with the installation. It must be installed from the “File > Advanced Plug-ins” panel by selecting the check box in the “Available” tab and restarting Mnova.



You will, in addition, require a valid license. A 45-day trial can be made available upon request, and a quote supplied for purchase.

For sales and support information, please contact:

¹ Kiss, R.; Fizil, Á.; Szántay, C. *J. Pharm. Biomed. Anal.* **2018**, *147*, 367–377.

² <https://www.nist.gov/programs-projects/nist-monoclonal-antibody-reference-material-8671>

³ Brinson, R. G.; Marino, J. P.; Delaglio, F.; Arbogast, L. W.; Evans, R. M.; Kearsley, A.; Gingras, G.; Ghasriani, H.; Aubin, Y.; Pierens, G. K.; Jia, X.; Mobli, M.; Grant, H. G.; Keizer, D. W.; Schweimer, K.; Stahl, J.; Widmalm, G.; Zartler, E. R.; Lawrence, C. W.; Reardon, P. N.; Cort, J. R.; Xu, P.; Ni, F.; Saeko, Y.; Kato, K.; Parnham, S. R.; Tsao, D.; Blomgren, A.; Rundolf, T.; Trieloff, N.; Schmieder, P.; Ross, A.; Skidmore, K.; Chen, K.; Keire, D.; Freedberg, D. I.; Suter-Stahel, T.; Wider, G.; Ilc, G.; Plavec, J.; Bradley, S. A.; Baldisseri, D. M.; Sforca, M. L.; Zeri, A.; Wei, J. Y.; Szabo, C. M.; Amezcuca, C. A.; Jordan, J. B.; Wikstrom, M. *MAbs* **2018**, *11* (1), 94–105.

⁴ <https://www.bruker.com/en/products-and-solutions/mr/nmr-pharma-solutions/journal-club.html>

2. New features since version 3.0

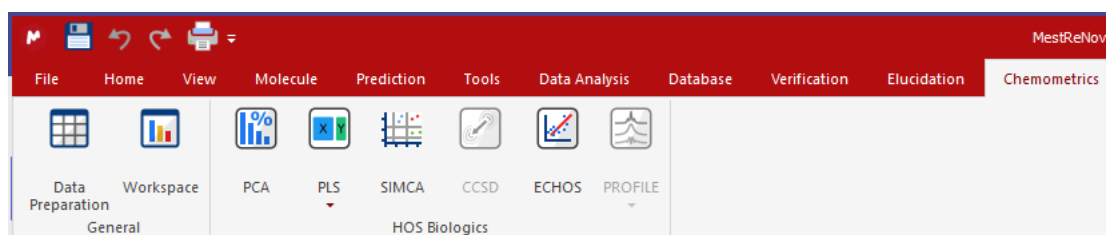
In addition to the features implemented in Mnova BioHOS 2.0, such as Matrix evaluation, 1D Profile, PCA model, and Distance calculation, we have also implemented new features since Mnova BioHOS 3.0:

SIMCA (Soft Independent Modelling of Class Analogies). Model spectra of different classes can be used to generate a “SIMCA model”, and tools are available for model validation. New samples are then classified by the model.

PLS (Partial Least Squares). Sample spectra and their relationship to response variables, e.g. concentrations are modeled. There are tools available for model validation. The PLS model predicts the response variables of new samples.

3. The ribbon interface

Once installed and licensed, a new ribbon tab called “Chemometrics” will become available in the Mnova window. Analysis methods that are unavailable for the loaded dataset will have their buttons greyed out.



4. General

Four analysis methods are currently available, and the best for each user’s case must be determined empirically.

Generally, you should apply the same sample preparation, acquisition, and processing steps for all spectra. The latter is facilitated by using a Processing template.

In all cases, the analysis input will be 2 or more NMR spectra that have been *stacked*, so that they appear as a single item on a single page in a document.

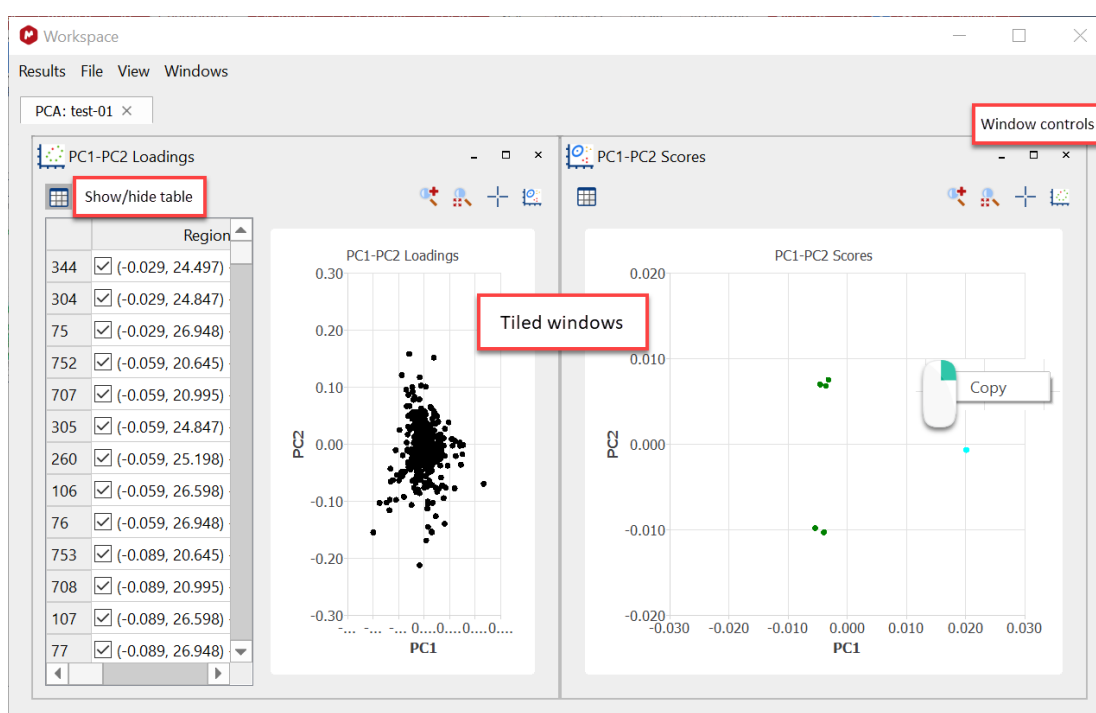
5. Workspace results display

All results are displayed as plots and tables using the *Workspace* window. The display options are contextual to the analysis, and each is described in detail in the analysis sections.

The *Workspace* window has tabs that allow switching between analyses: the name of each tab is set when the analysis is performed.

Each plot has, alongside, the underlying data table.

One or more windows can be viewed in the Workspace, and these have the usual, general properties, such as maximise, arrange, etc.



6. Data import and processing

2D data processing will be applied by Mnova when the spectra are imported. These are described fully in the full Mnova manual, but a few topics will be highlighted.

Recent development acquisition methods may require an individual calculated apodization function⁵. In this case, a direct import of the processed spectrum into Mnova is needed.

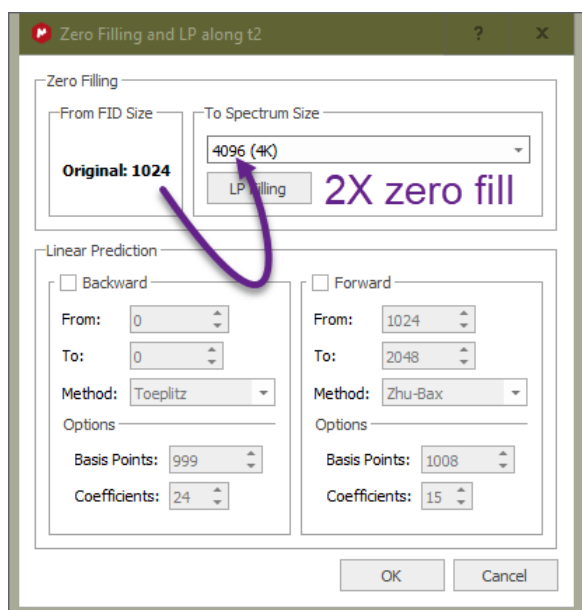
⁵ P. Rössler, D. Mathieu & A. D.Gossert, Angew. Chem. Int. Ed. 2008. <https://doi.org/10.1002/anie.202007715>

1D Profile uses a special processing routine for these kinds of spectra. There is baseline only on the left and right.

Several methods can be used to import the data:⁶

- Data browser [preferred]
- File > Open... ("ser", raw data files)
- Drag and drop a data file or folder into the Mnova window

Mnova will read the processing parameters from the source file, and either apply equivalents to these, or a standard, “advised” processing. If NUS was used for data acquisition, then it will be automatically detected and processed correctly. As a general rule, 2-fold zero filling should be applied to both dimensions before FFT.



Every effort should be made to ensure the best data quality before any analysis. This starts with optimised sample preparation (concentration, excipients, etc.) and continues to spectrum acquisition (temperature, general spectrometer performance and optimisation, pulse sequence, etc.)

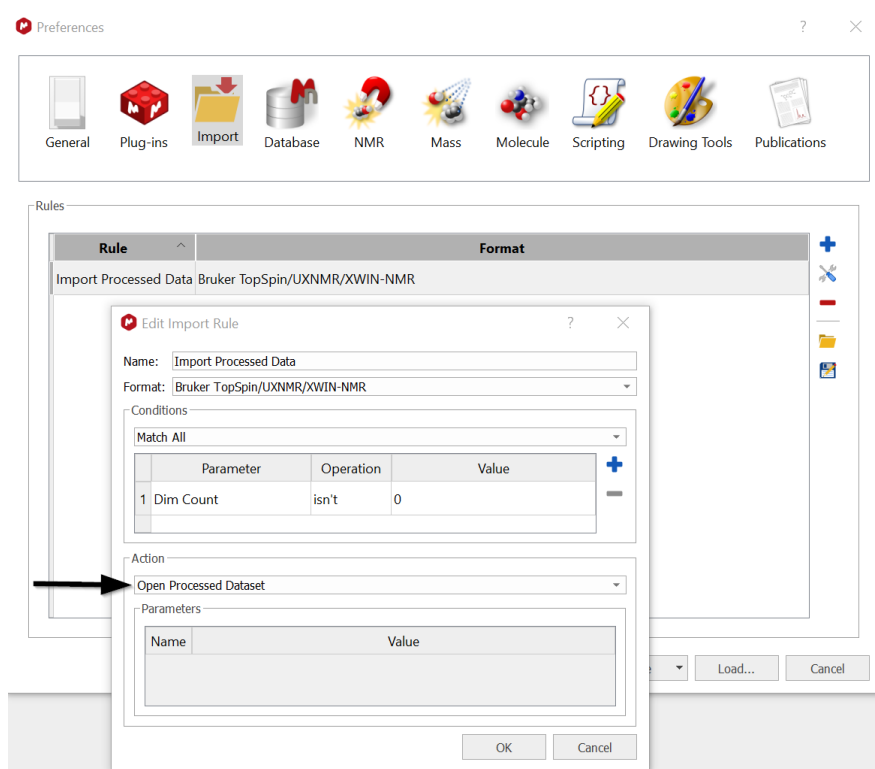
Basic data processing (FFT) is followed by processing stages:

- Phasing (automatic)
- Baseline correction (automatic)
- t1 ridge elimination
- Definition of a region-of-interest using:
 - Blind regions
 - Cuts
- PCA: Data compression (VOI)
- If required: Reference alignment

⁶ For this manual, data from a Bruker spectrometer will be shown, but this analysis will work with data from JEOL or Agilent/Varian instruments, too.

6.1. Direct import of processed data

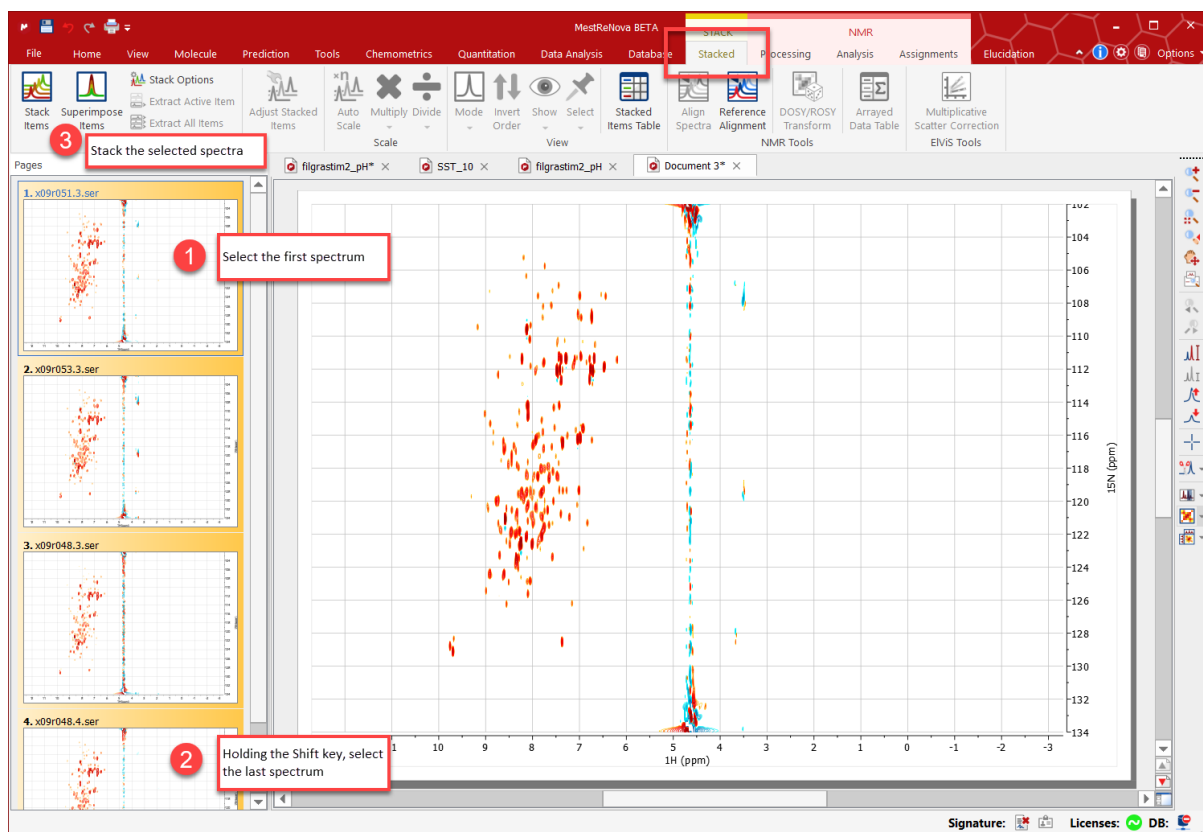
Per default, Mnova reads the raw data and does the processing. Advances in acquisition may require a special processing, so that Mnova is not able to process these spectra using the new features. To ensure that the processed data is read directly, it needs to be configured in the *preferences*. To do so, a new input rule needs to be defined. The action needs to be Open Processed Dataset.



6.2. Stacked items

Mnova has significant capability to stack closely related spectra, making their further processing and analysis greatly simplified. Please see Part X of the Mnova manual for fuller details.

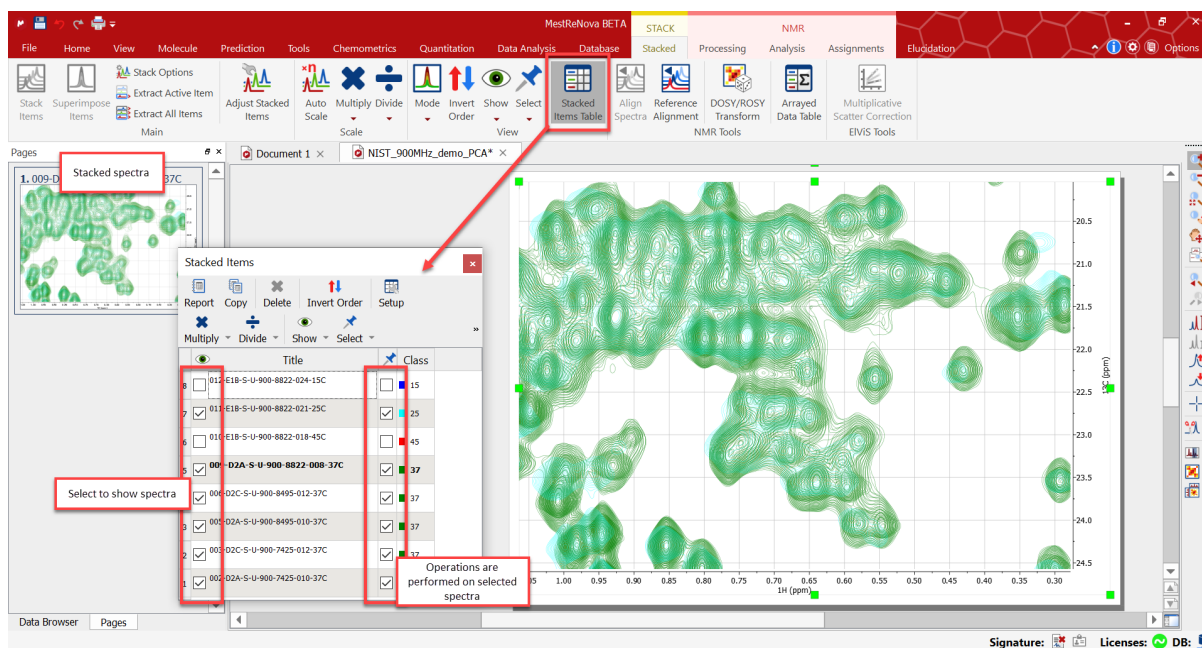
Spectra can, for example, be imported individually and joined to form a stack of spectral vectors. The spectral items in the stack maintain all their data content, and the single spectra in the document can be deleted. Any spectrum in a stack can be (re)converted to a single spectral item.



The Stacked Items Table is shown. Here, the spectra are shown superimposed, but other viewing options can be used (Stacked > Mode).

Active spectrum

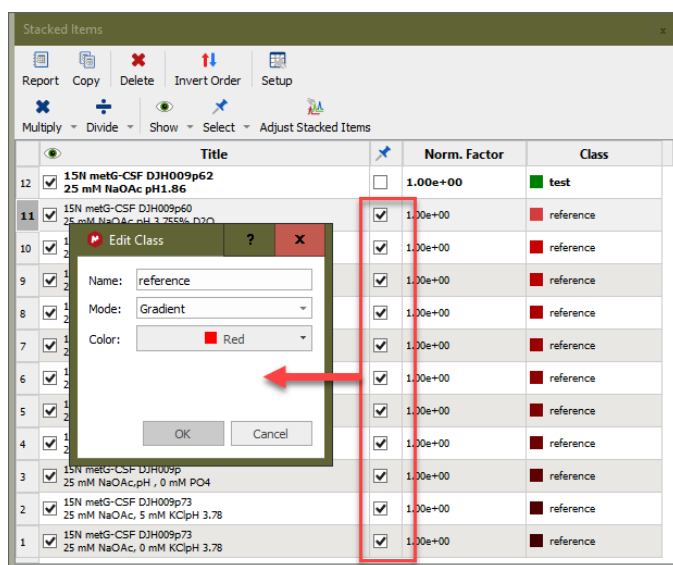
Note in the table that one spectrum in the table is shown in bold type, and the plot lines are thicker. This, the active spectrum, has significance for processing. The active spectrum can be changed by double clicking on a row in the table or using the mouse scroll wheel with the Shift key held down.



6.3. Classes: identification and coloring

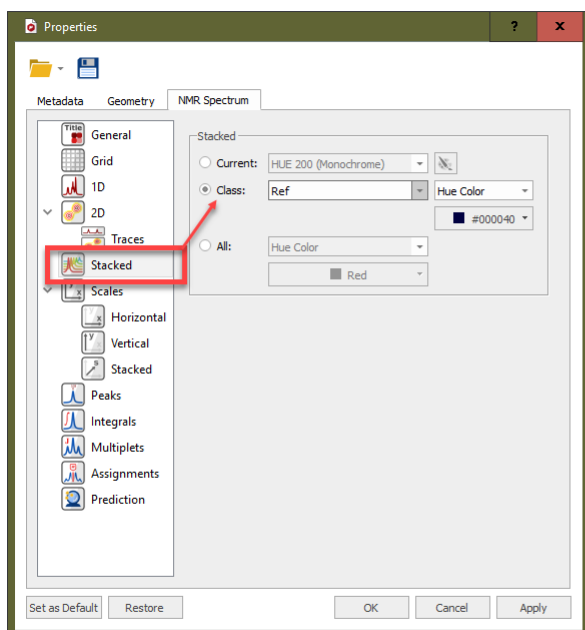
It is often convenient to assign the different spectra in the array different class names, for example “Reference” and “Test”. Once defined, this perpetuates through Mnova and the HOS analyses. Classes can also have an associated color, which further simplifies results display.

Here we show how to change these attributes from the *Stacked Items* display, although the same can apply in different tables. Color attributes can be changed in several places, including Table > Properties > Stacked.

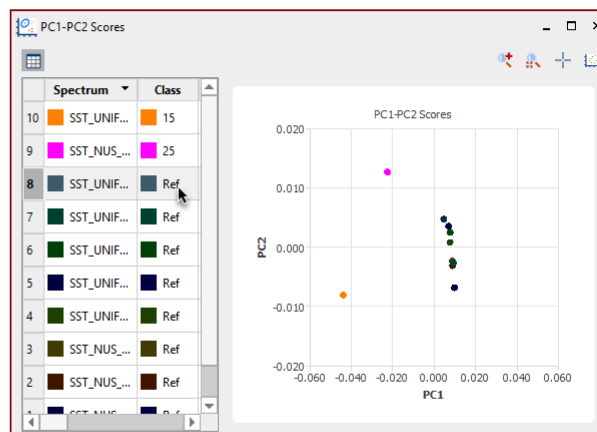


Select (drawing pin check box) rows [1...11], which belong to the same class. Double-click in the Class field to set the class name and color attributes using the Edit Class window.

Class names and colors are propagated throughout Mnova. They can be seen (and changed) in the spectrum Properties, and the BiologicsHOS “Workspace” tables.



Properties



PCA Scores plot

6.4. Phasing and baseline correction

These corrections are described in detail in the full manual. Generally, the imported values will work well, and further optimisation is not required.

Automatic phasing will usually provide a good result: choose “Regions” autophase for f2 and f1.

Example processing selections:

Apodisation: sine-bell shifted 60°.

Zero-filling: to 2K (f2) and twice for f1 (e.g., 64 → 256 points)

Auto-phase correction: Regions method (both dimensions)

Baseline correction: Polynomial (both dimensions)

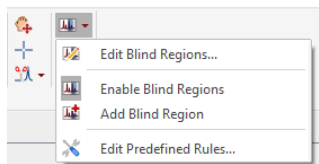
t1 noise reduction (optional): Processing > More Processing > Reduce t1 noise (see below)

6.5. Blind regions and cuts

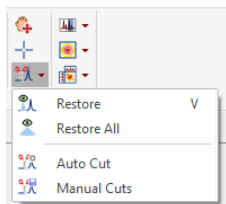
Blind regions and cuts are used to define a “region of interest” (ROI) – a part of the spectrum that has useful peaks. Other regions should be removed from the analysis in this way, as they can contribute noise that will

adversely affect the analysis. Regions of high peak density and, therefore, spectral overlap can be excluded for a better analysis.

These options are available from the “View” tab.



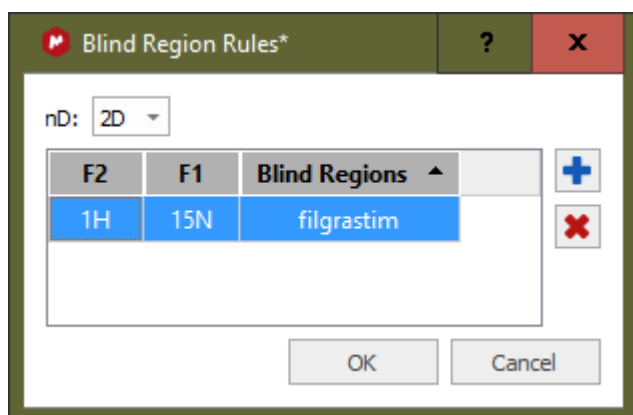
Blind regions menu



Cuts menu

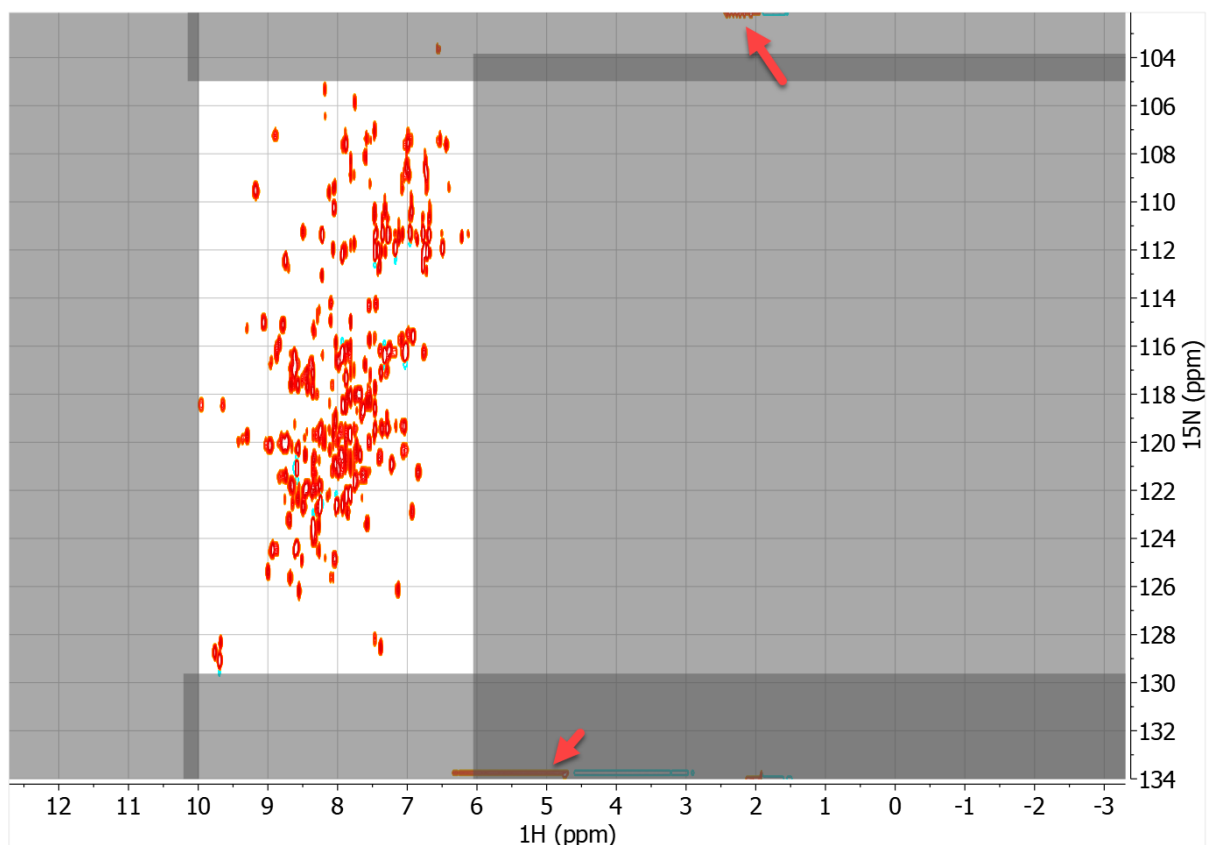
Cuts must be applied to each new spectrum – they cannot be applied from a file. Cut regions can be copied from 1 spectrum and pasted to others. (Ctrl + C followed by Home > Paste > NMR Zoom and Cuts)

Blind regions can be applied automatically when a spectrum is imported. For this, an existing group of blind regions are first saved as a file, and this is then used as a Blind Region Rule. In this example, the set of blind regions were saved to a file called “filgrastim”, and the dialogue shows the rule being created:



Once the rule has been created, new data imports will have these blind regions applied.

For CCSD analysis, blind regions used for the reference spectrum are applied automatically to the test spectrum.

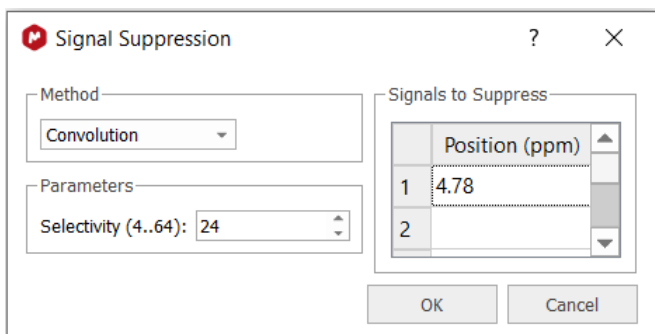


In this spectrum, 4 blind regions were used to define a ROI that excludes axial peaks and other noise. Arrows show intensity that should be excluded from the analysis.

Tip: check that the blind regions you specify go all the way to the edge of the spectrum, if this is what you want. Use the table in View > Edit blind regions.

6.6. Solvent signal removal

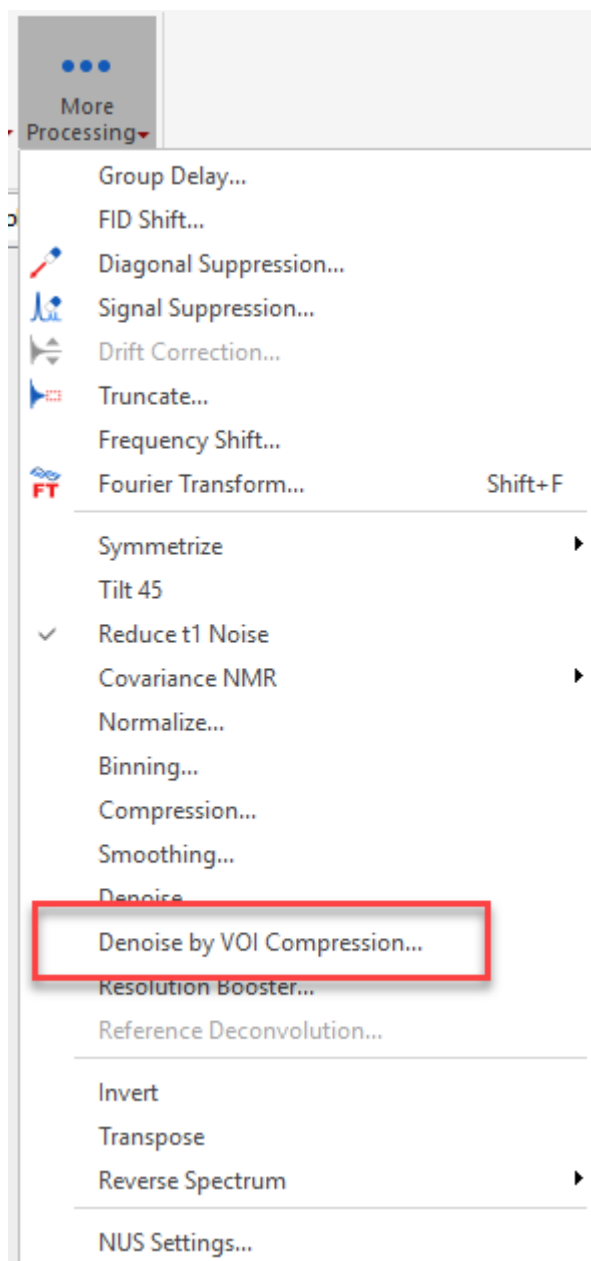
Processing > More processing > Signal Suppression



6.7. Denoise by VOI compression

Noise regions in the spectrum are identified, and the point values set to zero. This functionality is most often used for PCA analysis to reduce the number of bins.

This is available from the “More Processing” menu.



6.8. Reference alignment

This general capability in Mnova will be useful for ECHOS analysis. CCSD has this capability built in, and binning for PCA will usually remove the effect of small chemical shift changes.

Stacked > NMR Tools > Reference Alignment

The functionality will be applied automatically: no user input or specifications are required.

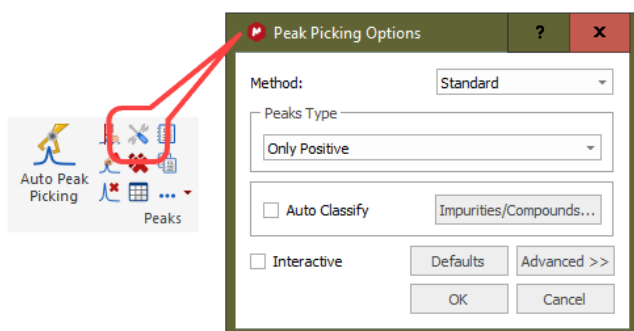
7. Data analysis

Aside from peak picking, which is required for CCSD analysis, no general analysis in Mnova is required for HOS analysis.

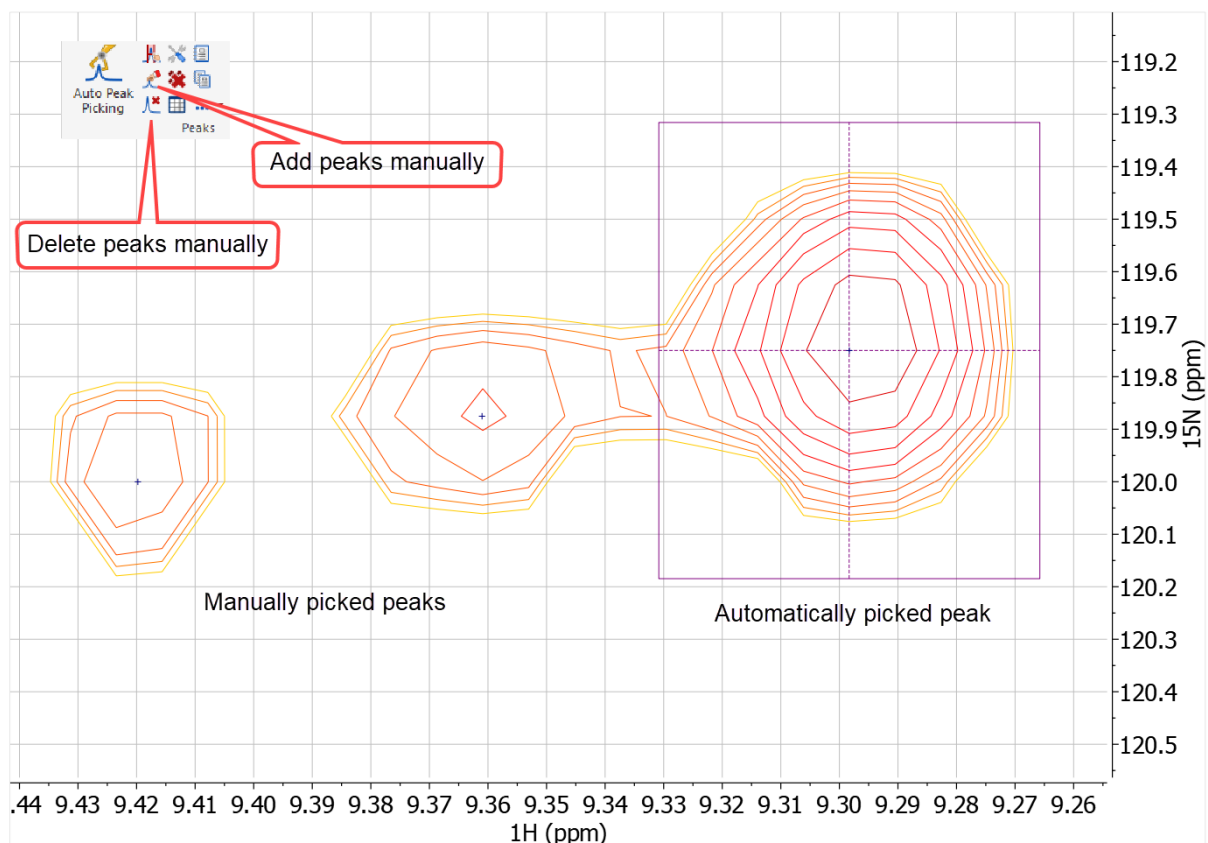
7.1. 2D peak picking

For CCSD, the reference spectrum should be peak-picked. It is possible to import existing peak lists in Sparky or TopSpin format.

Start by displaying the spectrum and choosing a vertical intensity expansion (mouse scroll wheel) that shows the peaks of interest, but no noise is called a “Visual threshold”. Use these peak picking options from the Analysis > Auto Peak Picking > Options button:



You should inspect the peak picks and refine them *before* CCSD analysis: delete any small peaks that may have been picked and manually pick ones that may have been missed.



Note that there are further automatic 2D peak picking options available but not described here.

After peak picking you can choose to *Annotate* peaks using the Peaks table. Annotations can be shown on the spectrum, which is a useful way to indicate the reference peak.

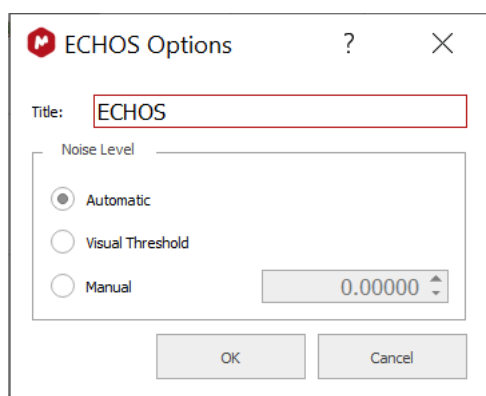
	δ (ppm)	f1 (Hz)	f1 (ppm)	f2 (Hz)	Intensity	Width f2	Width f1	Flags	Annotation
1	2.05	4123.0	18.22	1841.4	17.7	23.42	89.66	None	Reference
2	1.79	3837.5	16.95	1607.0	36.9	21.76	86.33	None	
3	1.75	4673.9	20.65	1576.1	9.7	35.28	105.75	None	
4	1.74	3903.9	17.25	1566.9	18.9	25.49	90.10	None	
5	1.73	3698.1	16.34	1560.7	7.2	22.47	97.42	None	

8. ECHOS analysis

ECHOS⁷ was initially described as a simple method for spectral comparison: a scatter plot is created by pairwise intensity comparisons of points with the exact coordinates in the reference- and test spectra. The plot axes are the point intensities, and their absolute values are not relevant.

An ECHOS analysis is performed on two spectra, and they are not distinguished. The spectra should be acquired under identical conditions. However, if required, interpolations are applied to make the analysis valid. The calculation does not align the spectra. Aligning the spectra needs to be done prior to the analysis. Normalization of amplitudes is not required.

Points are selected for comparison if in at least one spectrum the intensity is above the noise level. There are three ways to define the noise level: *Automatic*, *Visual Threshold*, and *Manual*.

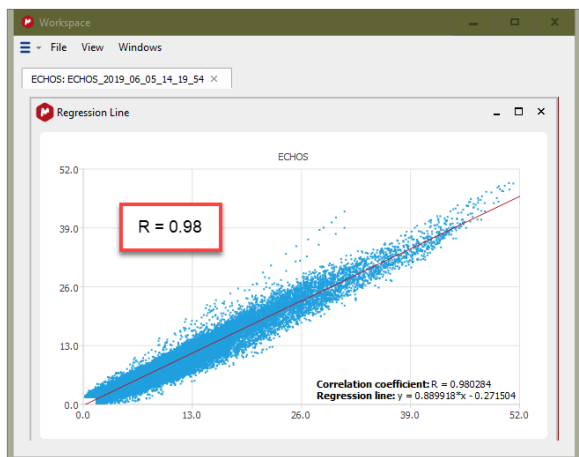


In case *Automatic* is chosen, the spectrum display is adjusted to the calculated noise level.

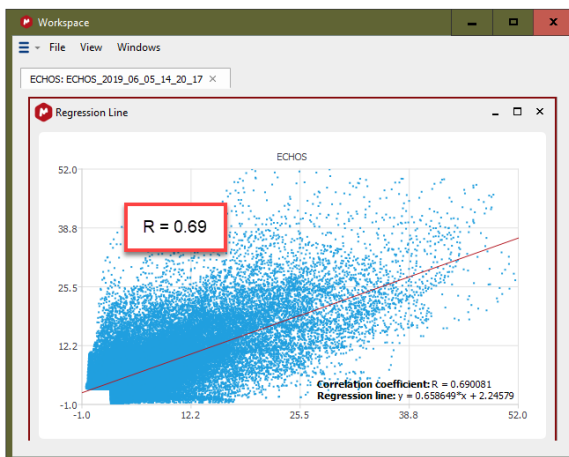
8.1. Points fitting and correlation coefficient (R)

A straight-line fit is applied to the scatter plot of points: when the spectra are most similar, the fit's correlation coefficient is closest to 1.0.

⁷ Easy Comparability of Higher Order Structure



Similar spectra



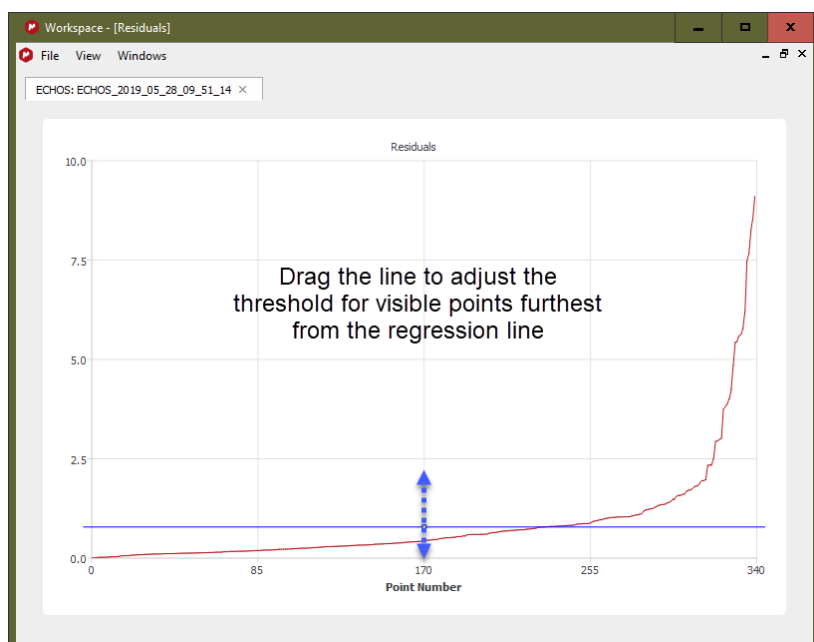
Dissimilar spectra

On the LHS we see the ECHOS result for 2, similar spectra (NIST samples at 50° C) where R=0.98. On the RHS we compared less similar spectra (NIST samples at 50° C and 25° C) where, now, R=0.69.

8.2. Viewing the residuals

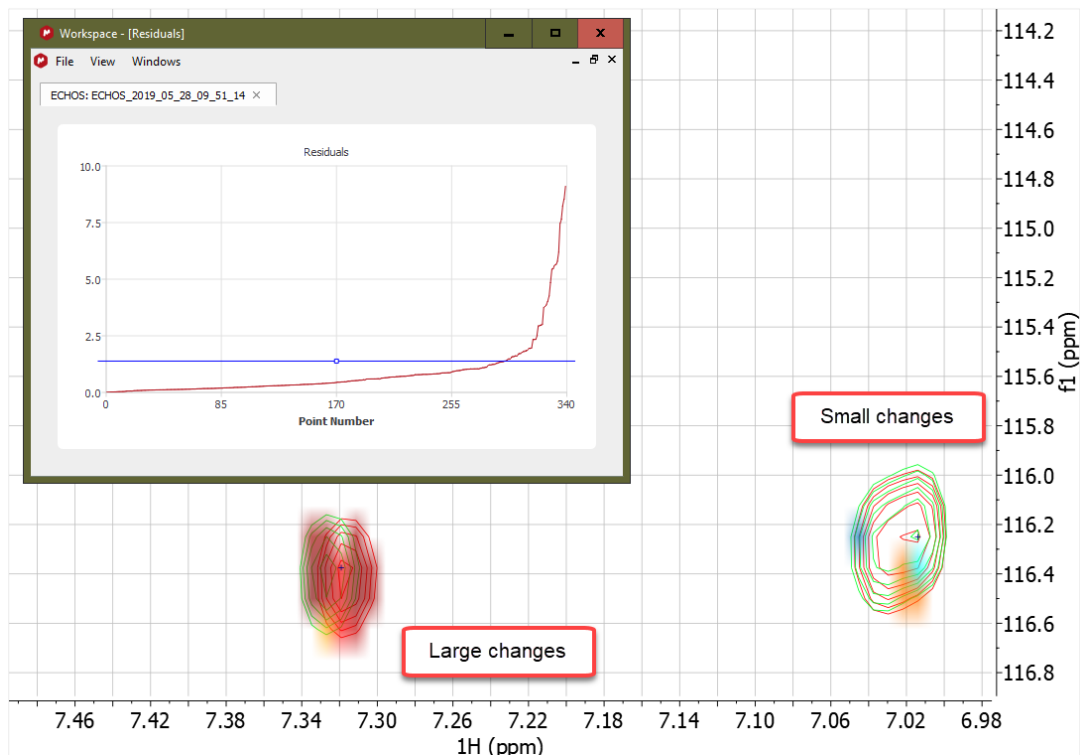
It is useful to visualize the spectra's parts that differ the most, as shown by their deviation from the fitted line in the regression plot.

It is possible in the Mnova window, where the points' deviation from the fitted line is plotted as a *heat map* together with the contour plot of the actual spectra. The larger the deviation, the hotter or colder the color code. The threshold for viewing the points can be set from the height of the horizontal blue line, using the mouse:



As the threshold is lowered, points with smaller deviations will be shown together with the spectral contour plots.

Now you can see where the spectra differ. The spectra are shown as the blue and green contours, and the ECHOS-derived deviations are co-plotted "hot" and "cold" shades.



8.3. Export data

The data points used in the Regression line plot can be exported as a CSV file (Workspace > File > Export to CSV...) The point intensities from each spectrum is followed by the data point coordinate.

9. CCSD comparison

Peaks are picked for the reference spectrum (see above). When CCSD is chosen, the matching peaks in the test spectrum/spectra are automatically determined, their shift changes computed, and the CCSD for each peak pair calculated. An average CCSD for the sample is computed and this may be a good, overall measure of the sample quality. The larger the average CCSD, the more peaks moved – and the less similarity exists between samples.

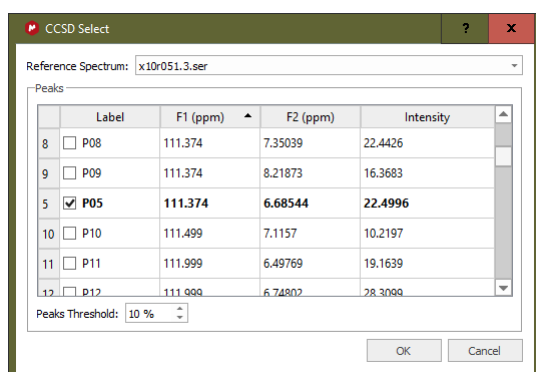
Briefly, these steps should be followed:

1. Overlay (stack) the reference and test spectrum/spectra. Assign classes (optional).
2. Pick peaks in the reference spectrum [Future: import a list of peaks.]
3. Apply Chemometrics > CCSD.
4. Select the reference spectrum from the drop-down box.

- Choose the reference peak. This will be used to both *align* and *scale* spectra. In the example above, this is peak **P05**.
Hint: If an "Annotation" is applied to a peak in the peak list table then this is shown in the CCSD table: this makes it easier to identify.
- Click on the OK button for the analysis to be performed.

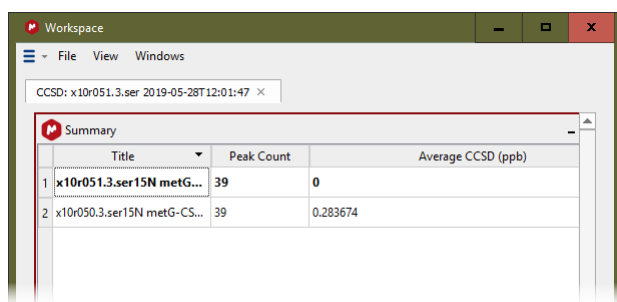
9.1. Reference

The analysis starts with the specification of the reference spectrum. The peak in the 2D spectrum should be taken as a reference: its position and amplitude should not change between all spectra. If the peak was annotated, then this will be shown as the peak label in the table.



9.2. CCSD results

Average CCSD



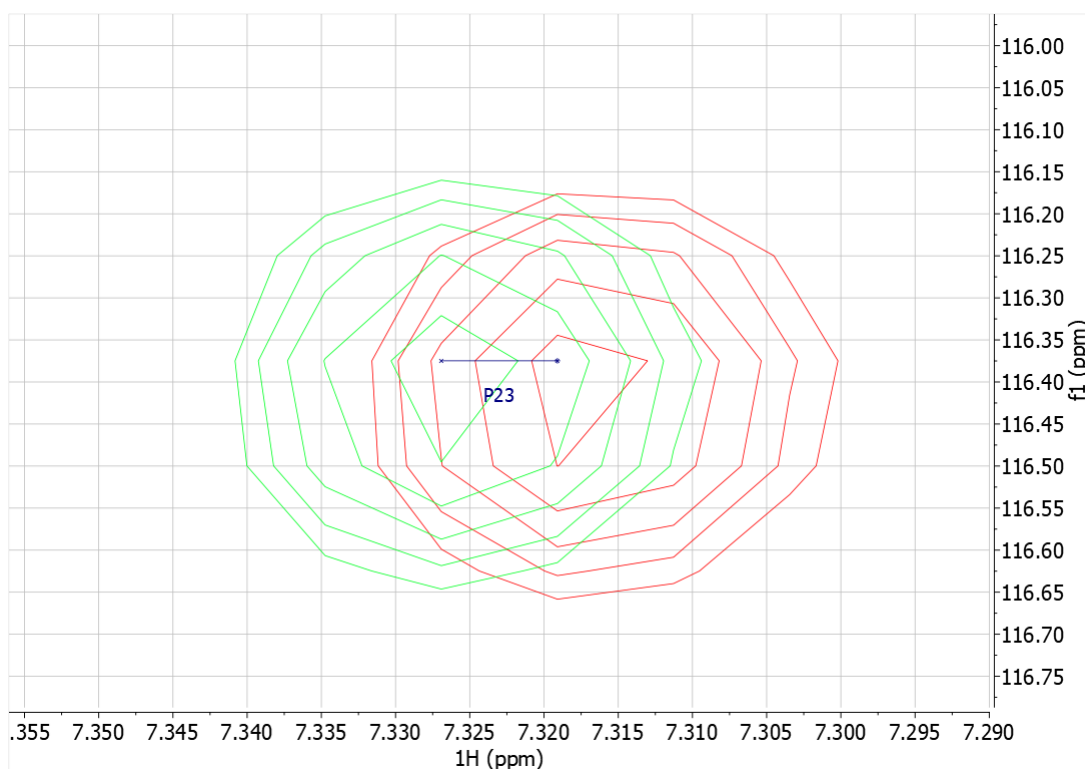
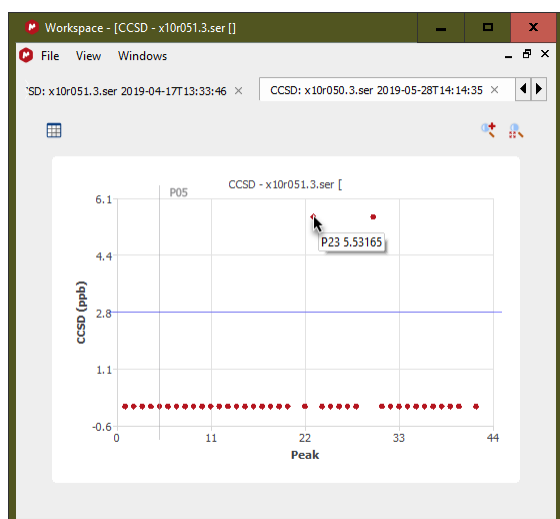
The *average CCSD* (ppb) will be shown for each test spectrum. The smaller the value, the more similar the spectra.

"Peak count" indicates the number of matching peaks that were automatically associated between the spectra. A smaller number, shown in red, indicates that some peaks in the Reference spectrum could not be matched to a peak in the test spectrum.

CCSD

The individual CCSD values for each peak can be seen. Peaks with relatively large shifts are quickly identified.

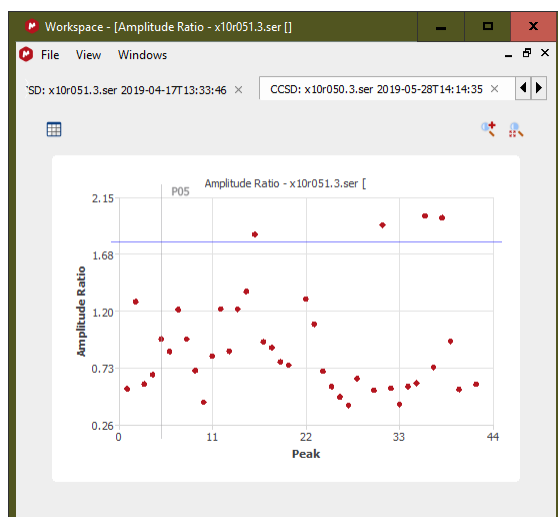
Double-click with the mouse in a point to zoom in on the region of the spectrum where the peak occurs.



We show the peak movement with a blue line, here for Peak #23.

Amplitude ratio

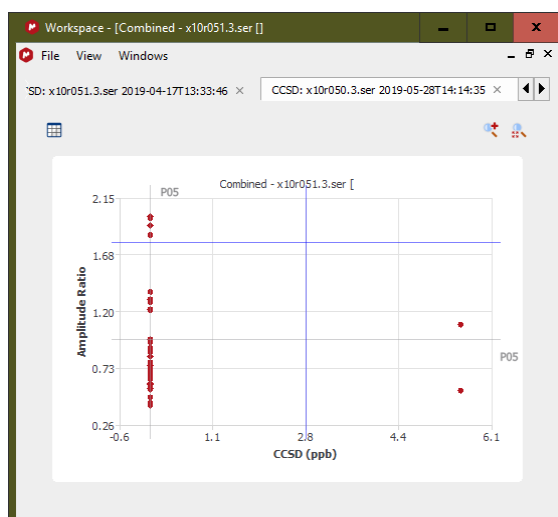
The ratio of peak heights for each peak pair is plotted in a similar way, to now show broadening effects.



Again, double-clicking on a point on the plot will cause the relevant spectral region to be shown.

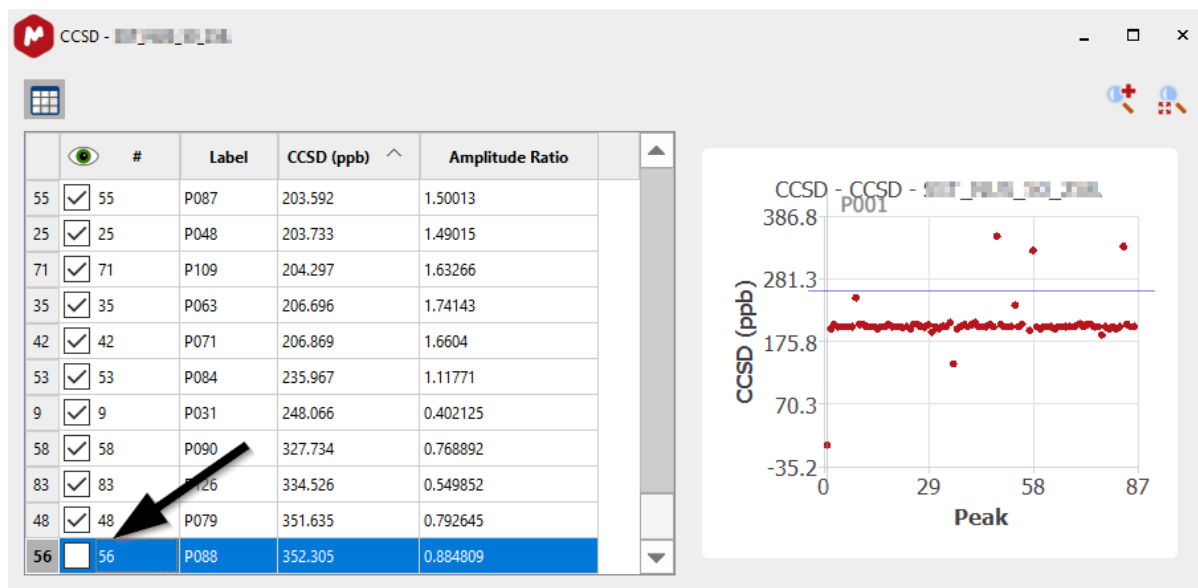
Combined plot

Finally, the CCSD and peak amplitude ratios are plotted against each other. Peaks can be identified by hovering the mouse on the point. Now, one can determine if the same peak both shifts and broadens.



9.3. Manual removal of points

The user may decide that a peak comparison should not be used: the peaks may be very small, or the algorithm failed to find a good partner for the peak. Removing bad points will lower the average CCSD, presumably to display a more accurate representative analysis. Individual activation/deactivation is available in the table view



10. 1D Profile

1D Profile^{8,9,10} analysis uses 1D NMR spectra. Small spectral changes are detected, and statistical analysis is performed on evaluations. Repeated measurements are required to cover variations from sample preparation, measurement, and the results' statistical analysis. There should be a minimum of 3 measurements per batch. Inter- and intra class correlations are made and statistically represented as a box and whisker plot.

The algorithm works in the following way

S: the NMR spectrum of the pure protein with optimization of Ph0/BC

C: the "convoluted" spectrum obtained by applying a broadening factor to **S**

F: the "fingerprint" spectrum, the difference of **S-C**

A correlation coefficient (=matching factor) is determined using **F**(fingerprint) and **C**(ref) spectra from 2 different samples. The matching factor is logarithmic.

The analysis is fully automated in Mnova, consisting of these steps:

⁸ Poppe, L.; Jordan, J. B.; Lawson, K.; Jerums, M.; Apostol, I.; Schnier, P. D. Profiling Formulated Monoclonal Antibodies by 1H NMR Spectroscopy. *Anal. Chem.* 2013, 85 (20), 9623–9629. <https://doi.org/10.1021/ac401867f>

⁹ Poppe, L.; Jordan, J. B.; Rogers, G.; Schnier, P. D. On the Analytical Superiority of 1D NMR for Fingerprinting the Higher Order Structure of Protein Therapeutics Compared to Multidimensional NMR Methods. *Anal. Chem.* 2015, 87 (11), 5539–5545. <https://doi.org/10.1021/acs.analchem.5b00950>

¹⁰ Poppe, L.; Knutson, N.; Cao, S.; Wikström, M. In Situ Quantification of Polysorbate in Pharmaceutical Samples of Therapeutic Proteins by Hydrodynamic Profiling by NMR Spectroscopy. *Anal. Chem.* 2019, 91 (12), 7807–7811. <https://doi.org/10.1021/acs.analchem.9b01442>

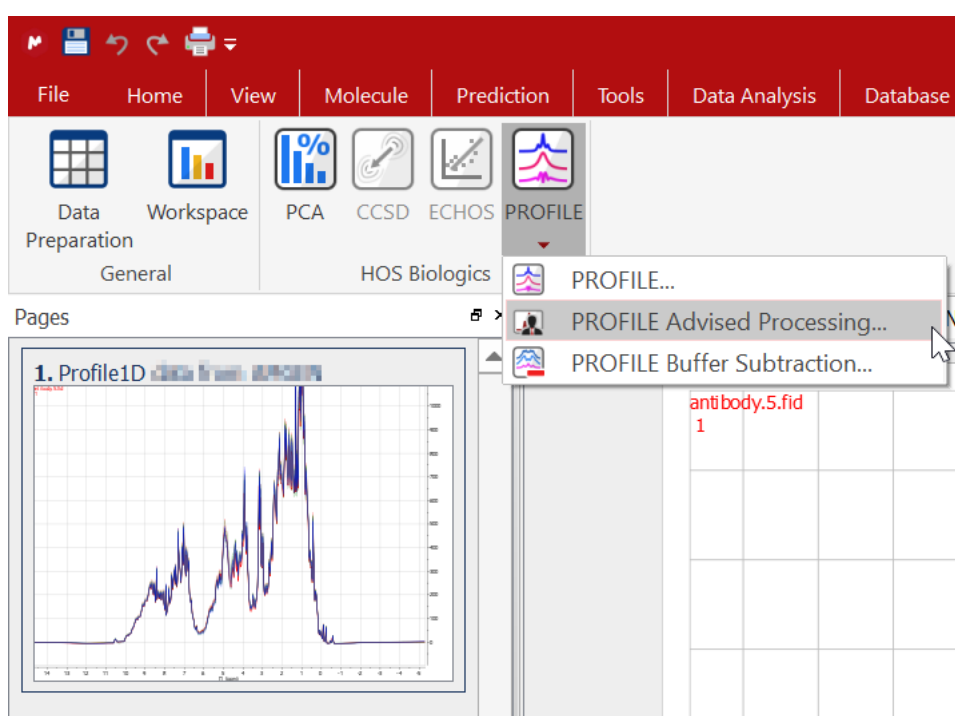
1. Advised processing¹¹ including phase and baseline optimization
2. Buffer/excipient signals subtraction

Alternatively, blind regions to exclude excipient signals

3. Analysis
4. Results viewing (spectra)

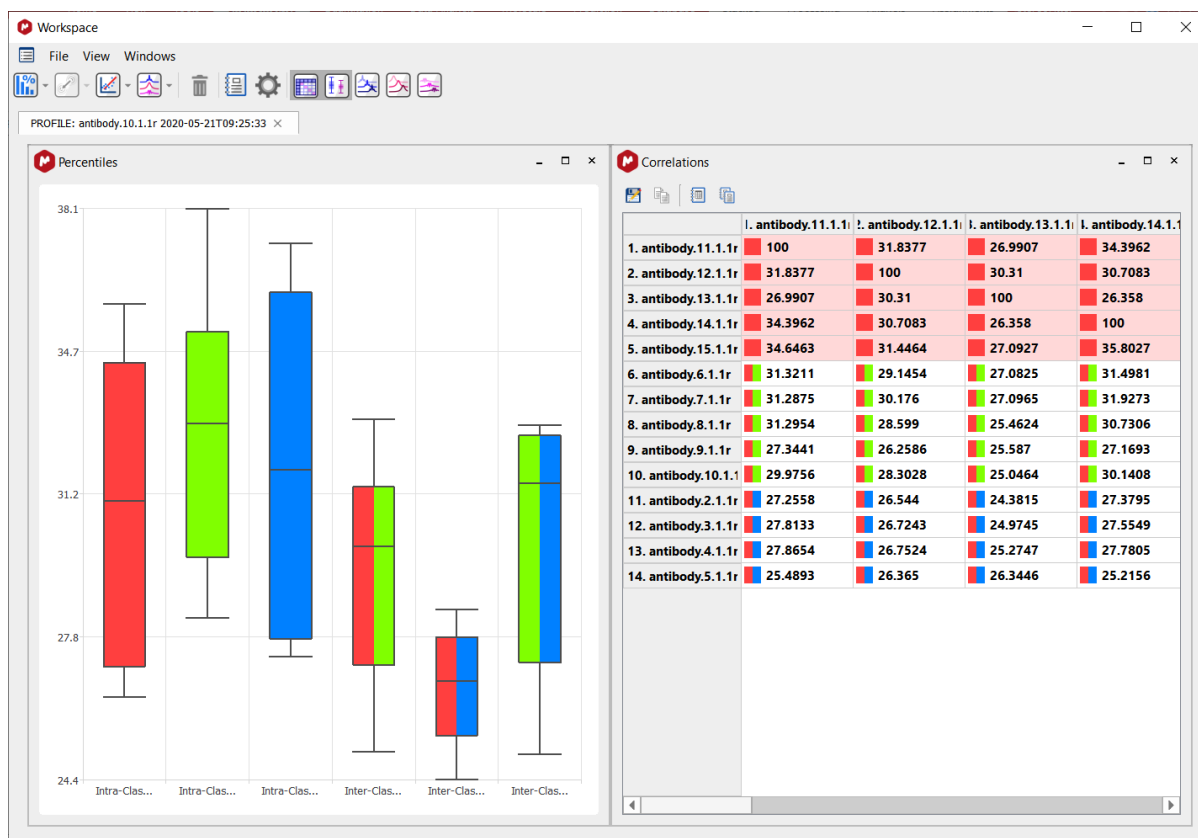
Shows intermediate steps in the analysis, and spectral regions that contribute most to differences in "fingerprint" plot

5. Cross-correlation statistical analysis of classes (box plots)
6. Table of results



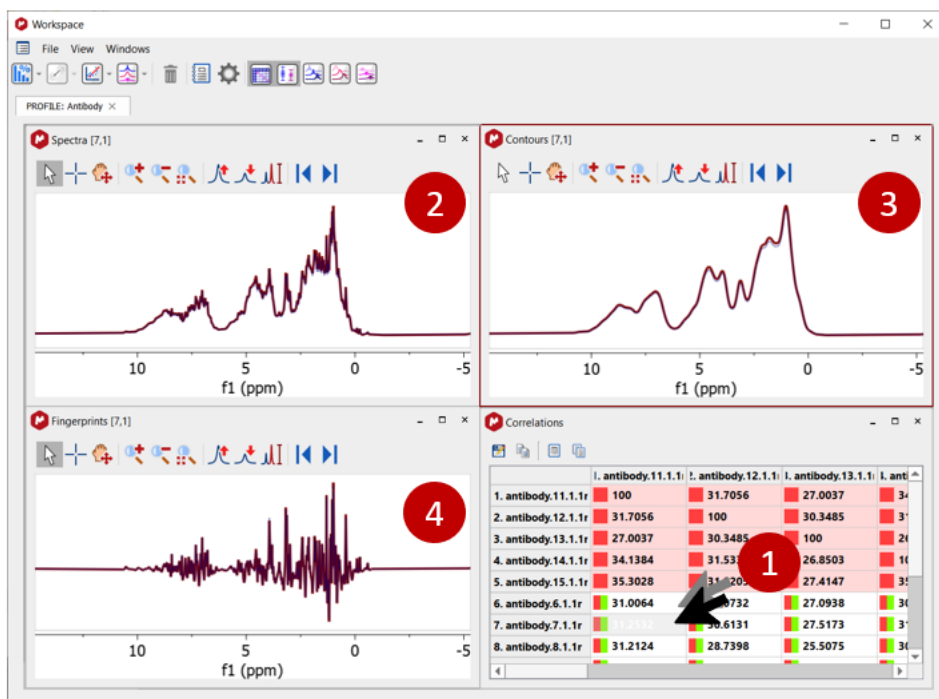
When the calculation is finished, a table of the comparisons is available. Individual access to the comparisons is available.

¹¹ Advised processing: a processing template is applied if an unprocessed *fid* file is opened. Processed files (1rr/2rr) that are imported are not reprocessed by Mnova.



In this case we have 3 groups: blue=reference; red = failed batch; green= working batch. The reference/fail (red-blue box) matching factors are lower than the intra-group matches (red and blue). It is regarded as *failure*. In contrast, the green-blue box shows a distribution with the distributions of the individual class of spectra. It is regarded as a *pass*.

By clicking on individual comparisons in the Table, detailed information is available.



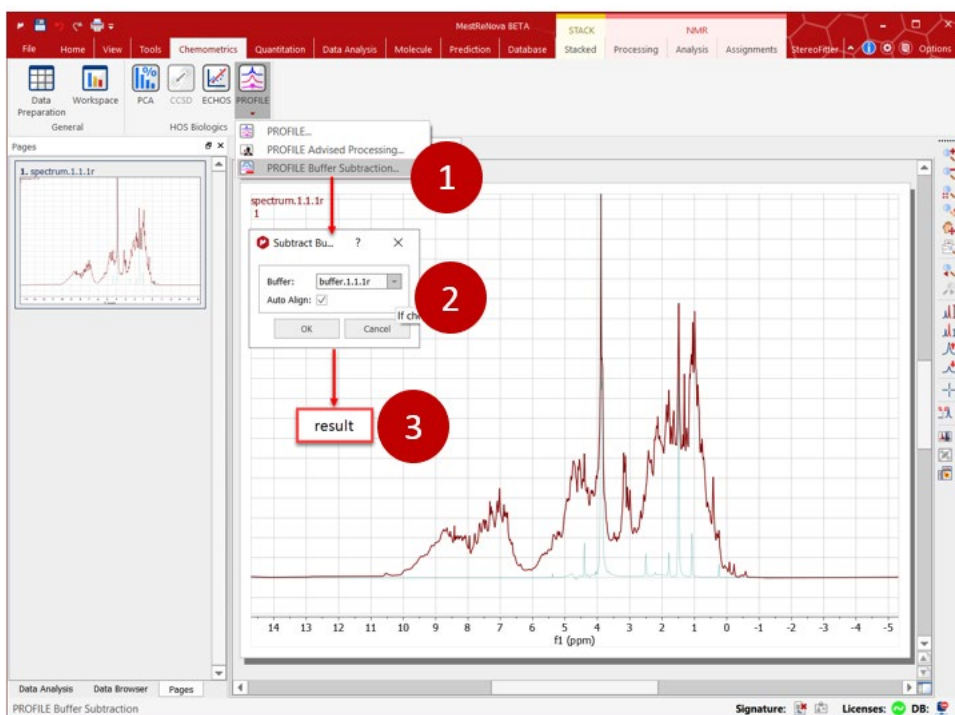
1. Data table (See chapter Matrix Analysis for details)
2. Input *spectra B*
3. Broadened spectra **C**
4. *Fingerprints (F, subtraction spectrum, 2-3)*

Buffer Subtraction

If your spectral stack has (sharper) buffer spectra/spectrum, this will be detected automatically, and the option provided to subtract it (with optimization) from all the protein spectra.

The buffer spectrum is shown with the protein ones until the subtraction is performed.

Only one buffer spectrum is used for subtraction.

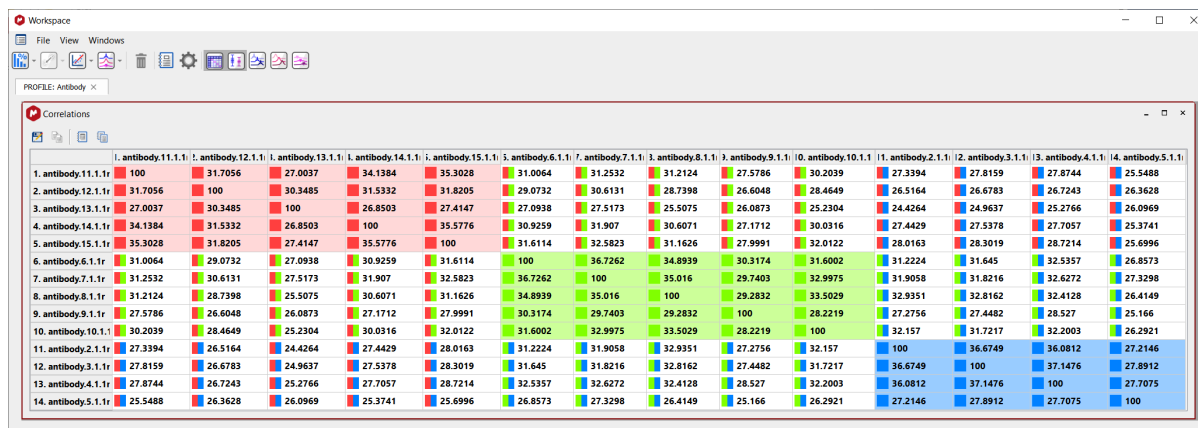


Because the method looks for small changes to the protein signals, unimportant regions must be excluded using blind regions. These may be at/near where the water saturation was done, and buffer signals removed, and the noise regions on each side of the main spectrum.

PCA can be used to make sure that blind regions have been set properly.

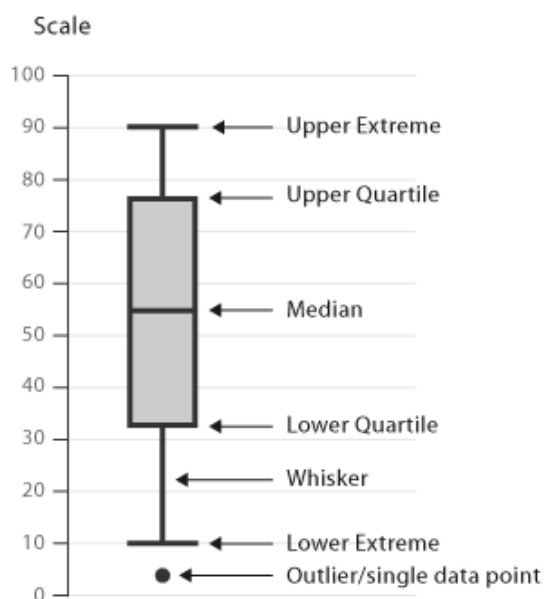
10.1 Matrix Analysis

ECHOS, CCSD, and Profile1D calculate similarity between two spectra. All samples are automatically compared with each other, and a table of results with color-coding makes it easy to compare *inter*- and *intra*class samples. Each cell has a colored square symbol that matches the class(es).



Box-Whisker plots

The distribution of the similarity values can be shown in Box-Whisker plots. Therefore, a boxplot gives graphical information on the location, the dispersion, and the *skewness* of a data set. The box ranges from the first quartile to the third quartile. A vertical line goes through the box at the median. The whiskers go from each quartile to the minimum or maximum. The maximum range is 1.5 times the interquartile range. Samples outside this border are regarded as an outlier. Each section contains 25% of the data.



It is available at the moment for Profile1D only.

11. Multivariate Statistics (PCA, SIMCA, and PLS)

Multivariate statistics is a powerful and essential tool in modern data analysis and modeling, providing a way to simultaneously analyze complex relationships and patterns among multiple variables. The use of unsupervised chemometrics for the analysis of groups of NMR spectra has a long history.¹² It is often central to metabolomics studies, and many applications, refinements, and specialized software packages exist to perform the analysis. Widely used techniques in multivariate statistics are Principal Component Analysis (PCA), Soft Independent Modelling of Class Analogies (SIMCA), and Partial Least Squares (PLS).

PCA is a method for reducing the dimensionality of a dataset while retaining as much of the original variation as possible. It works by transforming the original data (spectra) into a new set of variables, called principal components, which are linear combinations of the original variables. These components are arranged in order of decreasing variance and can be used to visualize the structure of the data and identify patterns and outliers.

SIMCA is a statistical method for supervised classification of data. The method requires a training data set (spectra) and their class membership. The term “soft” indicates that the classification of new samples is not strictly in one class. The method is based on PCA, only significant components are used.

PLS, on the other hand, is a method for modeling the relationship between two sets of variables, such as spectra and response variables, e.g. concentrations. It works by finding a set of latent variables, called components, that capture the maximum covariance between the two sets of variables. PLS allows not only to predict response variables in new spectra, but can also be used to force separation between groups.

The distinctive advantage of having PCA and PLS available in Mnova is having a link between the statistical variables and the underpinning NMR data. It is very useful, for example, in identifying from loadings plots the spectral regions that result in spectra being classified as dissimilar. Before 1D Profile analysis, a PCA analysis can help identify the remaining parts from the formulation buffer. Measurements of distances between classes and distances of test spectrum to a model, represented by a cluster of reference spectra, are available.

The multivariate analysis starts with three or more stacked, 1D or 2D NMR spectra. Typically for the analysis of HOS, these will be methyl ¹H-¹³C HSQC data. Having a set of reference spectra is useful – at least 6. It is useful for the spectra to have assigned class names and for these to be similarly colored.

11.1. Selecting the spectra for the analysis

It is possible to create a stacked array of spectra, and only perform PCA analysis on a subset of the data. The data sets to be used are set from the Stacked items table: this is useful to eliminate outliers that may affect the analysis. (See Influence Plot below.)

¹² Emwas, A. H.; Saccenti, E.; Gao, X.; McKay, R. T.; dos Santos, V. A. P. M.; Roy, R.; Wishart, D. S. *Metabolomics* **2018**, *14* (3), 1–23.

Stacked Items ✕

Report Copy Delete Invert Order Setup

Multiply Divide Show Select Adjust Stacked Items

	<input type="checkbox"/>		<input type="checkbox"/>	Class
10	<input checked="" type="checkbox"/>	SST_U	<input type="checkbox"/>	15
9	<input checked="" type="checkbox"/>	SST_NUS_25_174.1.1.2rr	<input checked="" type="checkbox"/>	25
8	<input checked="" type="checkbox"/>	SST_UNIFORM_50_112.1.1.2rr	<input checked="" type="checkbox"/>	Ref
7	<input checked="" type="checkbox"/>	SST_UNIFORM_50_103.1.1.2rr	<input checked="" type="checkbox"/>	Ref
6	<input checked="" type="checkbox"/>	SST_UNIFORM_50_102.1.1.2rr	<input checked="" type="checkbox"/>	Ref
5	<input checked="" type="checkbox"/>	SST_UNIFORM_50_064.1.1.2rr	<input checked="" type="checkbox"/>	Ref
4	<input checked="" type="checkbox"/>	SST_UNIFORM_50_063.1.1.2rr	<input checked="" type="checkbox"/>	Ref
3	<input checked="" type="checkbox"/>	SST_NUS_50_301.1.1.2rr	<input type="checkbox"/>	Ref
2	<input checked="" type="checkbox"/>	SST_NUS_50_300.1.1.2rr	<input checked="" type="checkbox"/>	Ref
1	<input checked="" type="checkbox"/>	SST_NUS_50_299.1.1.2rr	<input checked="" type="checkbox"/>	Ref

11.2. Data preparation

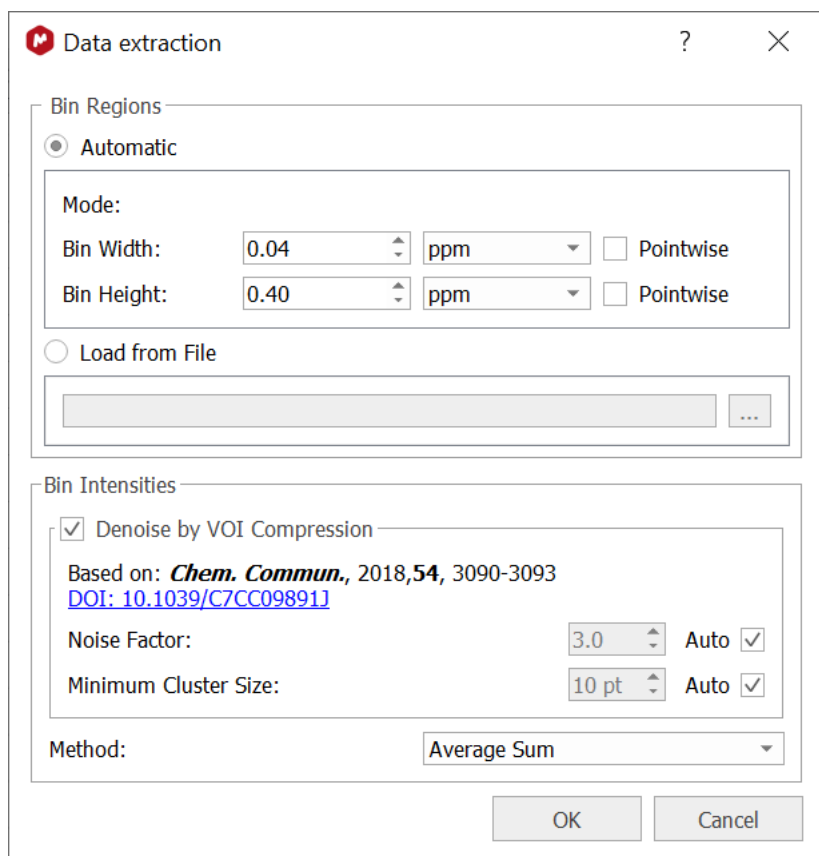
Data preparation for multivariate statistics is similar for PCA and PLS. In the dialog, there is a common and a method-specific part.

The choices made in data preparation can significantly affect the analysis result. We suggest you start with a set of reference spectra and iteratively find the options that result in the tightest clustering in the scores plot (see below).

Many of the options are described mathematically, and their effectiveness is considered in the review by Wishart and colleagues (see reference on the last page).

11.2.1 Binning

The purpose of binning is to simplify the spectrum and reduce the effect of small spectral frequency changes. PCA relies heavily on matrix operations, and the number of values heavily influences computational times. So, binning also helps to reduce computation times. Typical binning options are shown below. The bin sizes were chosen as about the size of an average peak.



Pointwise "binning" should be chosen carefully, as it will result in very large data matrices that can make Mnova unstable.

Smaller bin sizes will increase the data matrix, making computational times longer and computer memory requirements larger.

Load from File allows the user to either import a TXT file of integration regions or a CSV file containing previously computed bins. The latter can be saved from the Data Preparation tool. This capability is, effectively, a way to define regions of interest.

VOI denoising is a good choice because it sets the points in noise regions to zero, and bins are not required there. Thus, the matrix size is effectively reduced.

11.2.2 Data Processing

Data Integrity check

It allows the bins to be "cleaned", removing those having negative values, etc.

Filter

Filtering methods are used to remove bins that are null and do not display any changes among spectra series. By default, if a variable (bin) shows zeros among all rows (spectra) it is discarded.

Five options are possible. With Standard Deviation (**SD**), Median Absolute Deviation (**MAD**) and Interquartile Range (**IQR**) a fixed fraction (default 10%) of the bins is discarded (e.g. if the matrix is composed by 100 bins

then 10 bins are discarded, and the selection is based on the Filter method chosen). In practice SD, MAD and IQR are calculated for all bins. Furthermore, a percentage value of the total bins (with the lowest SD, or MAD or IQR values) are discarded. In the case of Mean Value or Median Value, the user is asked to input a value for the Mean or the Median. Only bins that display a lower value than the inputted one are discarded.

Normalization

It is an operation that is performed on the rows of the matrix. Four strategies are available:

1. Sum: every element on a row is divided by the sum of all elements of the same row
2. Median: every element on a row is reduced by the median value of all the bins that constitute the same row
3. Reference Spectrum (or reference series of spectra): upon normalization by the sum, every element of a row is divided by the corresponding element of the row of the selected reference spectrum (e.g. when you have a reference spectrum and you wish to compare all the other spectra relative to it). If you select a bundle of spectra (like all spectra belonging to the same class), normalization is performed on the calculated average spectrum
4. Reference Bin: user inputs chemical shift value (in ppm) of a reference peak of interest and, automatically, Mnova identifies the bin that comprises the selected peak. Furthermore, every component of a row is divided by the reference component of the corresponding row

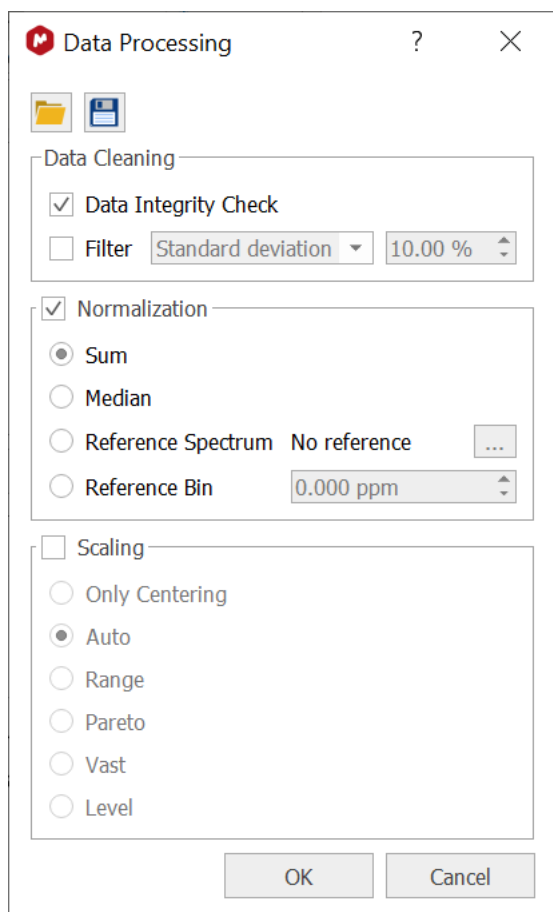
Scaling

The purpose of scaling is to make the numerical values of each variable equally significant (on average) - which can be undesirable. Each variable is divided by its scaling factor. An operation is performed on the columns of the matrix. Autoscaling, Range, Pareto, Vast and Level scaling are available. The mathematical expressions are well known.

Mean centering is always applied before any other scaling. For PCA analysis this is done even if the Scaling option is deselected to ensure that intense and weak signals are equally represented.¹³

In this example, Sum normalization, but no scaling was performed.

¹³ <https://bmcbgenomics.biomedcentral.com/articles/10.1186/1471-2164-7-142>



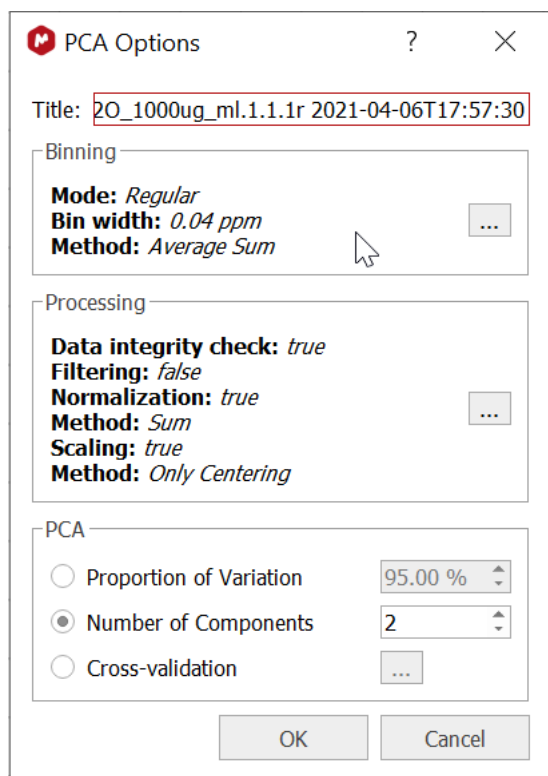
11.2.3 PCA Options

The main panel is used to change or refine the data preparation parameters (see below). The last used values will be automatically saved and presented, facilitating the workflow after they have been optimized.

A title for the analysis can be specified and will later be shown on the tab in the Workspace windows. The number of PCA components can be determined either by

- The proportion of variation value
- Direct enter the number of PCA components
- Cross Validation

Click on the ellipsis (...) to change the Binning or Processing options.



The image shows a 'PCA Options' dialog box with the following settings:

- Title: 2O_1000ug_ml.1.1.1r 2021-04-06T17:57:30
- Binning:
 - Mode: Regular
 - Bin width: 0.04 ppm
 - Method: Average Sum
- Processing:
 - Data integrity check: true
 - Filtering: false
 - Normalization: true
 - Method: Sum
 - Scaling: true
 - Method: Only Centering
- PCA:
 - Proportion of Variation: 95.00 %
 - Number of Components: 2
 - Cross-validation: ...

Buttons: OK, Cancel

Cross validation

When you create a PCA model, you need to ensure that the model is fit for its purpose. Cross-validation helps to determine the optimal number of PCs and estimates the quality of the model.

The basic idea behind cross-validation is to leave out spectra (test data) from all the spectra (train data) and quantify the model's effectiveness. Q^2 is the error of the cross-validated model, R^2 the error of the model without cross-validation. These errors are plotted against the number of PCs so that a valid number of principal components is chosen. If Q^2 decreases with the number of principal components, it is regarded as overfitting. The number of PCs should be chosen, when Q^2 is at the maximum. For a valid model, Q^2 needs to be higher than 0.5.

Workflow:

1. Load and process the training data
2. PCA, select "Cross Validation"
3. Set the criteria, and run
4. View CV results
5. Assess model, and set the number of required components

PCA Options

Title: 2O_1000ug_ml.1.1.1r 2021-04-07T07:44:59

Binning

Mode: *Regular*
 Bin width: *0.04 ppm*
 Method: *Average Sum*

Processing

Data integrity check: *true*
 Filtering: *false*
 Normalization: *true*
 Method: *Sum*
 Scaling: *true*
 Method: *Only Centering*

PCA

Proportion of Variation 95.00 %
 Number of Components 2
 Cross-validation

OK Cancel

Cross-validation Options

Min Components: 2
 Max Components: 10
 Split method: K-Fold
 Fold Number: 10

Use variable prediction

OK Cancel

11.2.4 SIMCA Options

A title for the analysis can be specified and will later be shown on the tab in the Workspace windows. Click on the ellipsis (...) to change the Binning or Processing options.

SIMCA Options

Title: My SIMCA Model

Binning

Mode: *Regular*
 Bin width: *0.04 ppm*
 Method: *Sum*

Processing

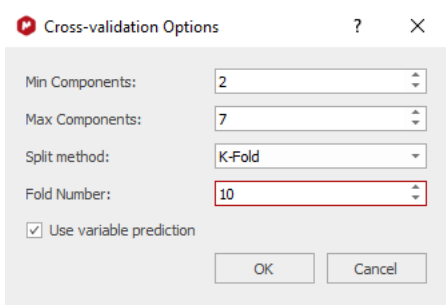
Data integrity check: *true*
 Filtering: *false*
 Normalization: *true*
 Method: *Sum*
 Scaling: *true*
 Method: *Pareto*

Cross-validation

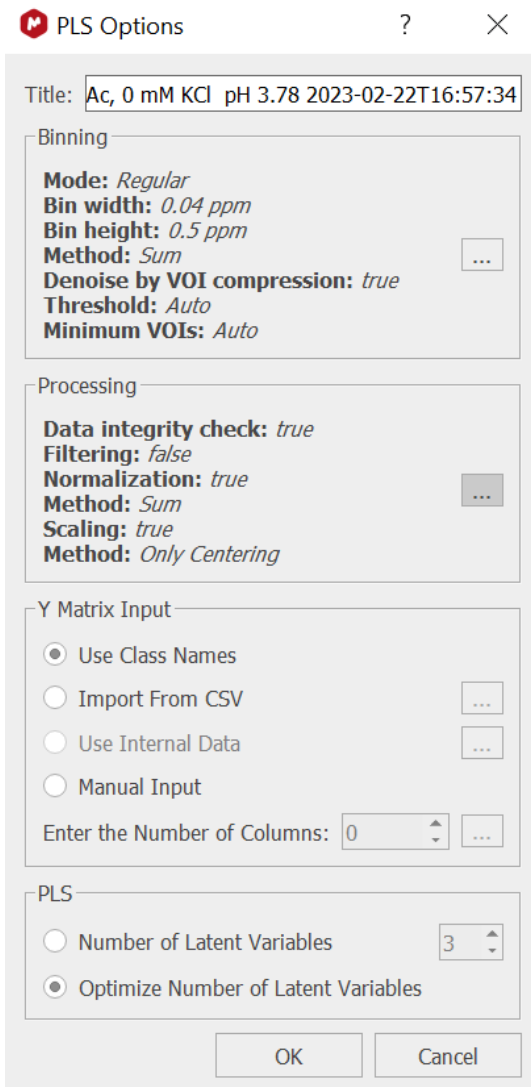
Min Components: 2
 Max Components: 2
 Split Method: *K-Fold*
 Fold Number: 10

OK Cancel

The “Cross-validation” is identical to the calculation used for PCA, see above.



11.2.5 PLS Options



Y Matrix Input

In PLS (Partial Least Squares), the Y table refers to the response or target variable table, which contains the variable(s) we want to model and predict based on the predictor variables in the X table. The Y table is stored in the Mnova file.

The Y table typically consists of one or more continuous or categorical variables, also known as the dependent variables, response variables, or outcome variables. The goal of PLS is to identify the relationship between the predictor variables in the X table and the response variables in the Y table, and to use this relationship to predict new values of the response variables based on new values of the predictor variables.

In cases of more than one response variable, it is possible to use all of them in one model or create a model for each variable individually. Models using one response variable only, are known as “PLS-1”. PLS-1 is a widely used multivariate regression technique that is particularly useful in cases where the number of predictor variables is large, and collinearity or multicollinearity exists among the predictors. It is also used in cases where the relationship between the predictor variables and the response variable is nonlinear, and where the number of observations is relatively small compared to the number of predictor variables.

There are several methods to add the Y matrix:

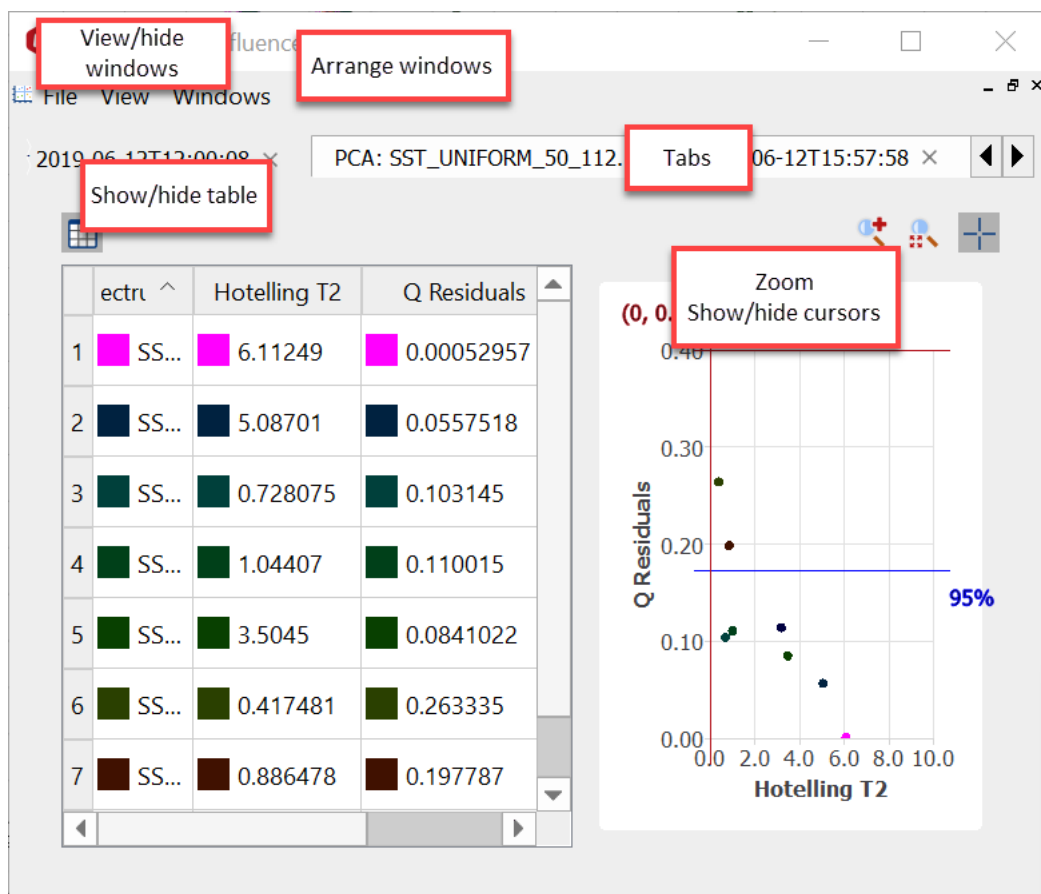
- Directly use class names, the class names are translated internally into numbers,
- Import a CSV file
- Previously entered data can be edited.
- Manual Input opens a table view, data can be copied in, e.g. from Microsoft Excel Sheets.

PLS

In this section, the number of latent variables can be entered manually or optimized automatically.

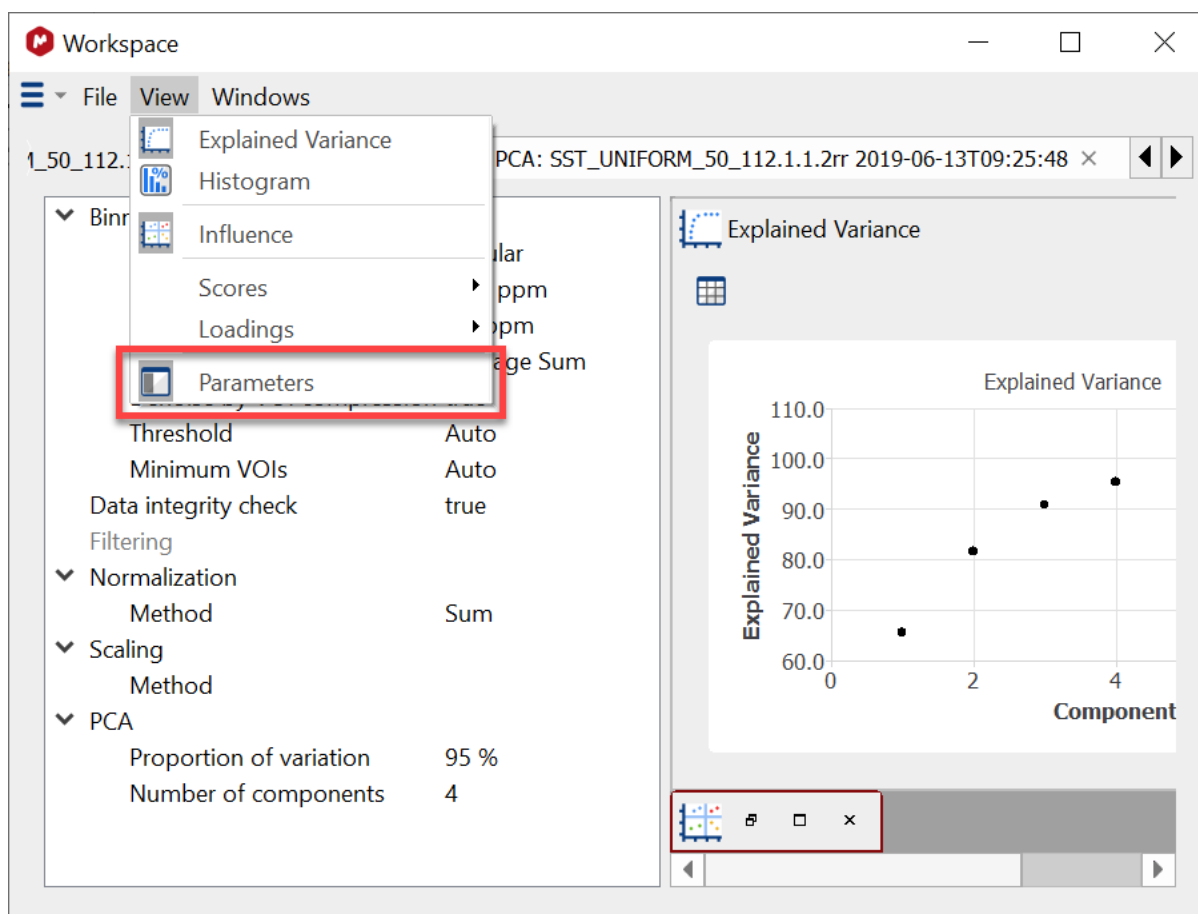
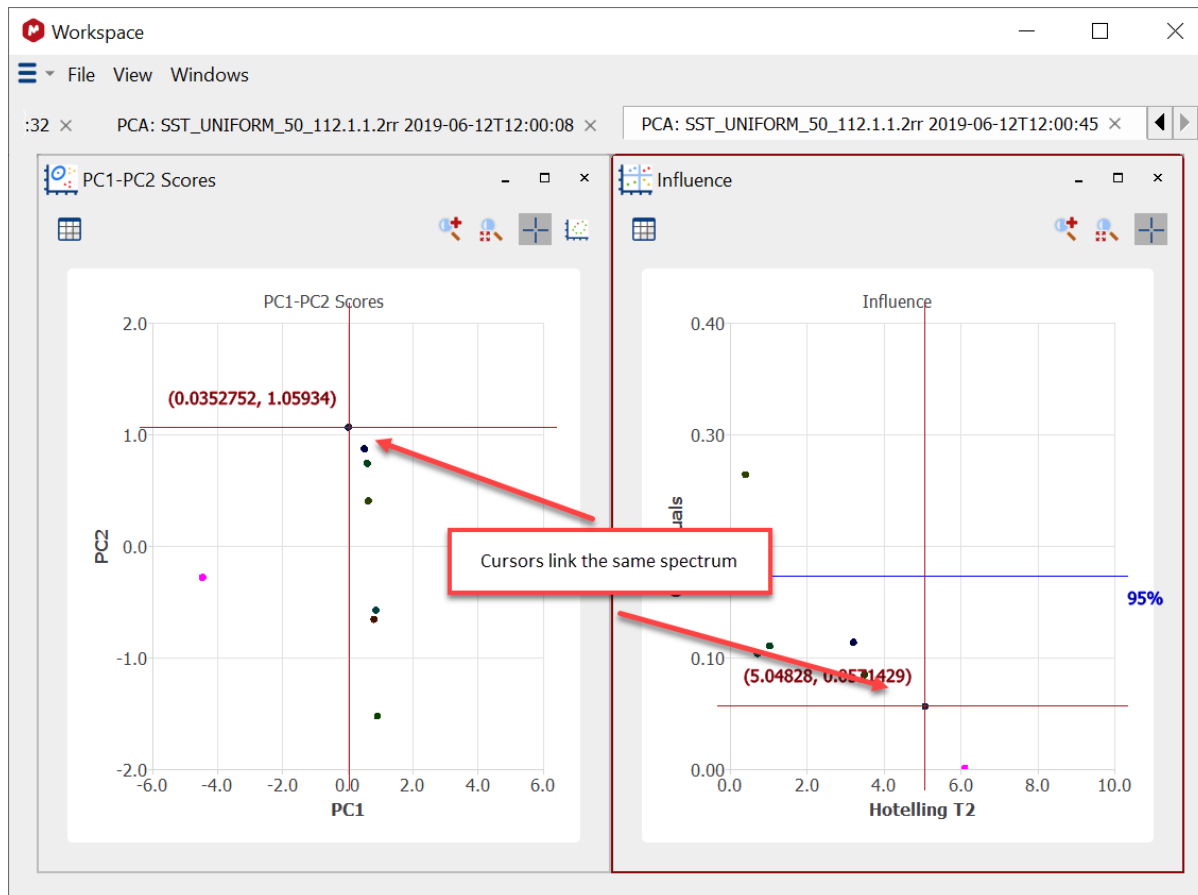
11.3. PCA Results

All results are shown as windows in the Workspace window, which can be minimized, tiled, etc. Previous analysis results may be accessed from tabs at the window's top. A window (generally) has a table behind the graphical data, and this can be shown and used alternately with the plots. It is generally possible to zoom into a region.



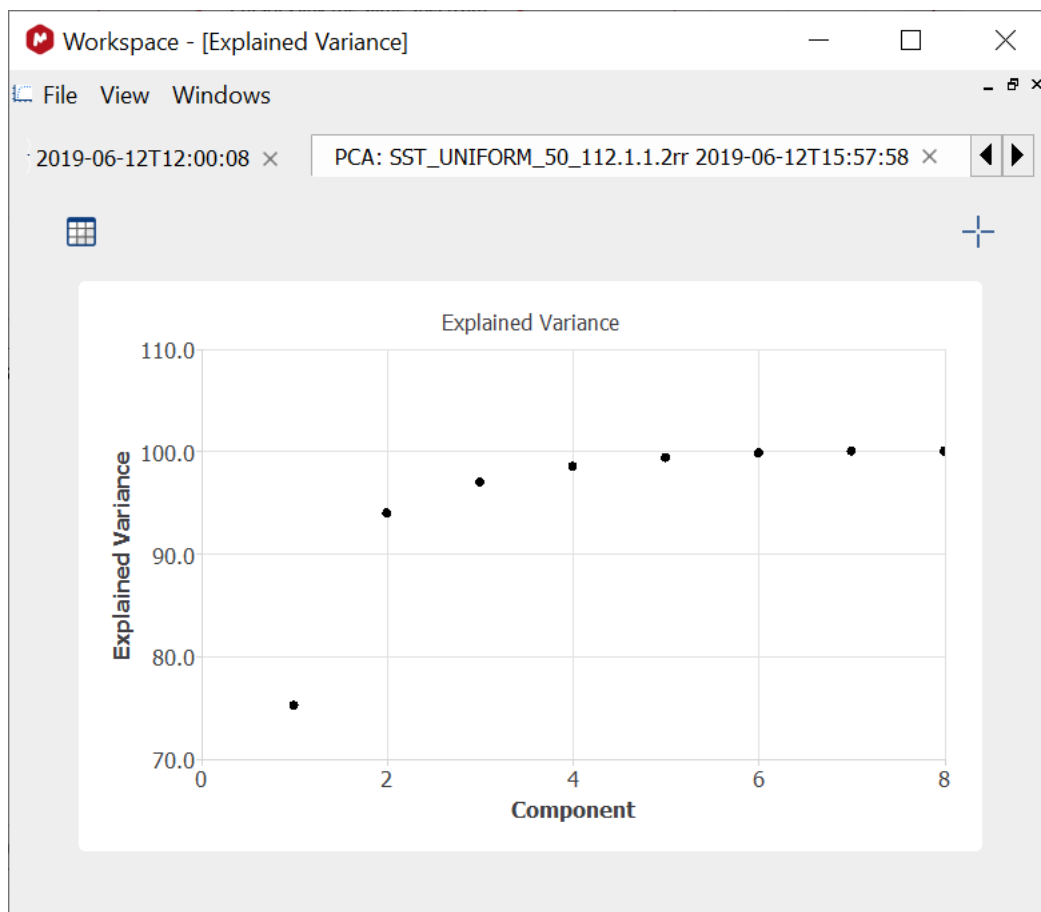
Where relevant, cursors can be coordinated between windows, e.g.:

The data preparation parameters can be shown from the View > Parameters menu item.



Explained variance

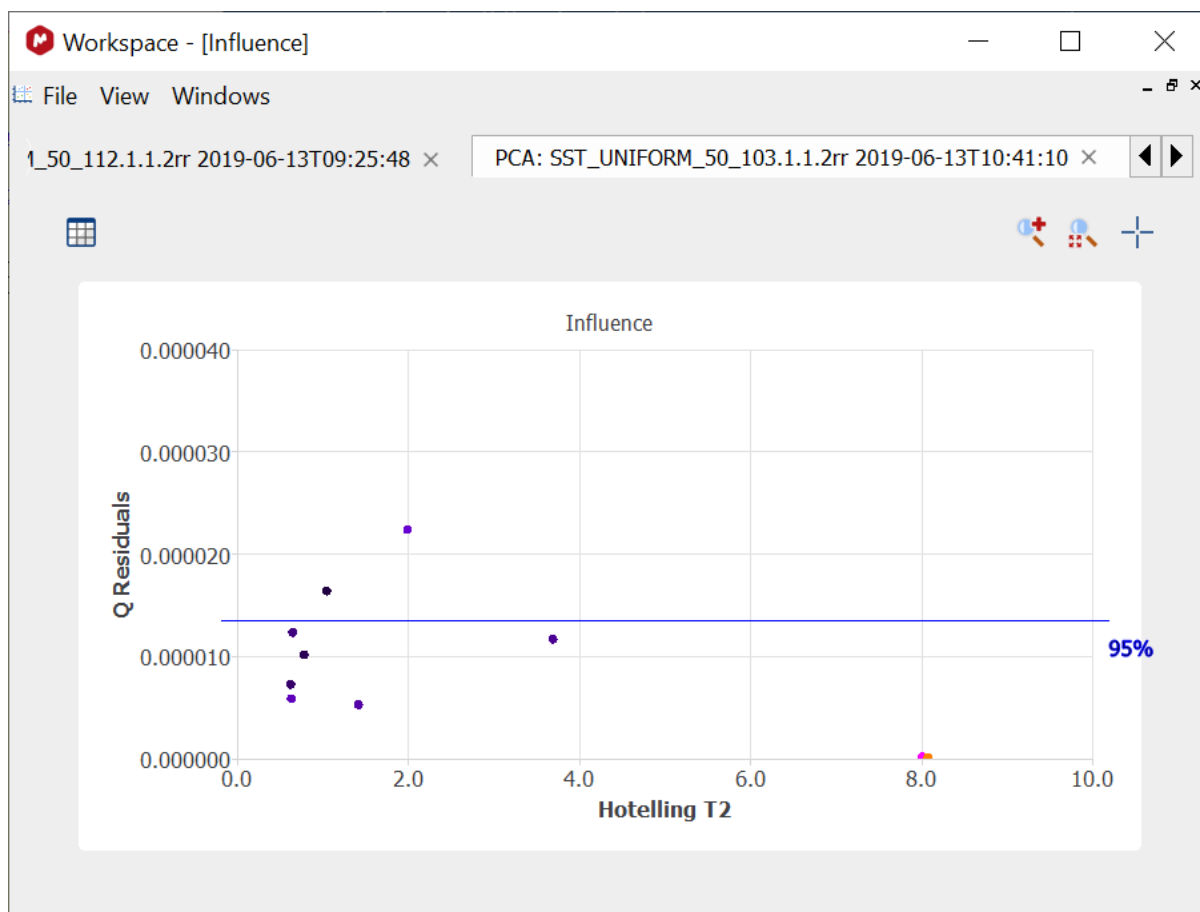
The number of available components is influenced by the "Proportion of variation" % chosen during data preparation and the result of the explained variance plot. In this case, 3 components explain ca 95% of the variance; therefore, only these need to be considered.



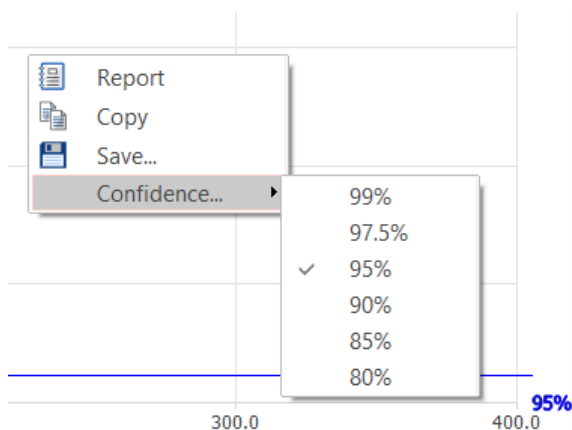
Influence plot

It is a plot of Hotelling's T^2 values (x-axis) vs. Q residuals (y-axis). Summary statistics help explain how well a model describes a given sample and why that sample has observed scores in a given model.

A description of this plot's use is beyond this manual's scope, but it provides a compact view of both residual and score outliers (and inliers). The Hotelling's T^2 values indicate how far a component is from the center of the model: as expected, the test samples are much more distant than the reference.



You can select the confidence limit that is plotted with the blue line by clicking with the right mouse button on the graph:



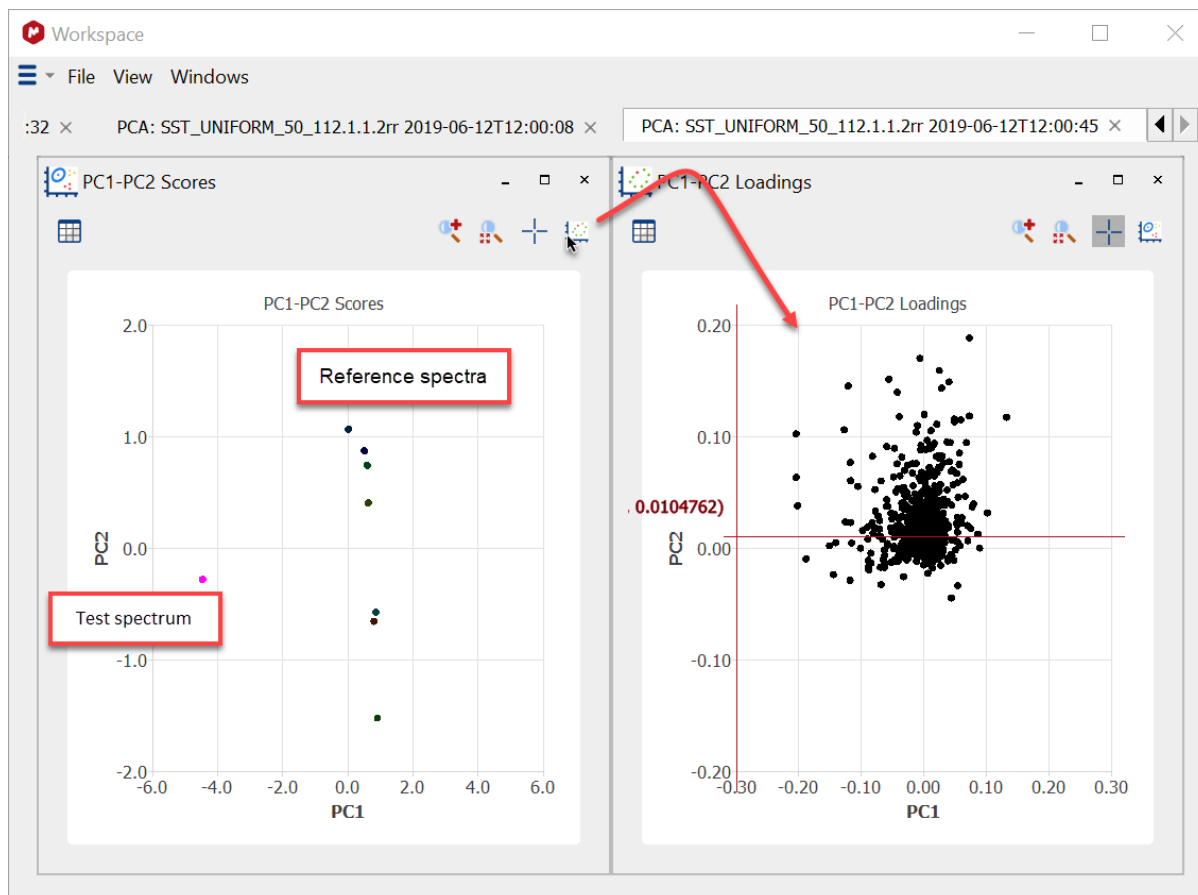
Scores plots

To select the components that best separate the data sets it can be useful to see all the possibilities, here PC1-PC2, PC1-PC3, and PC2-PC3. Only 2 components can be viewed at a time.

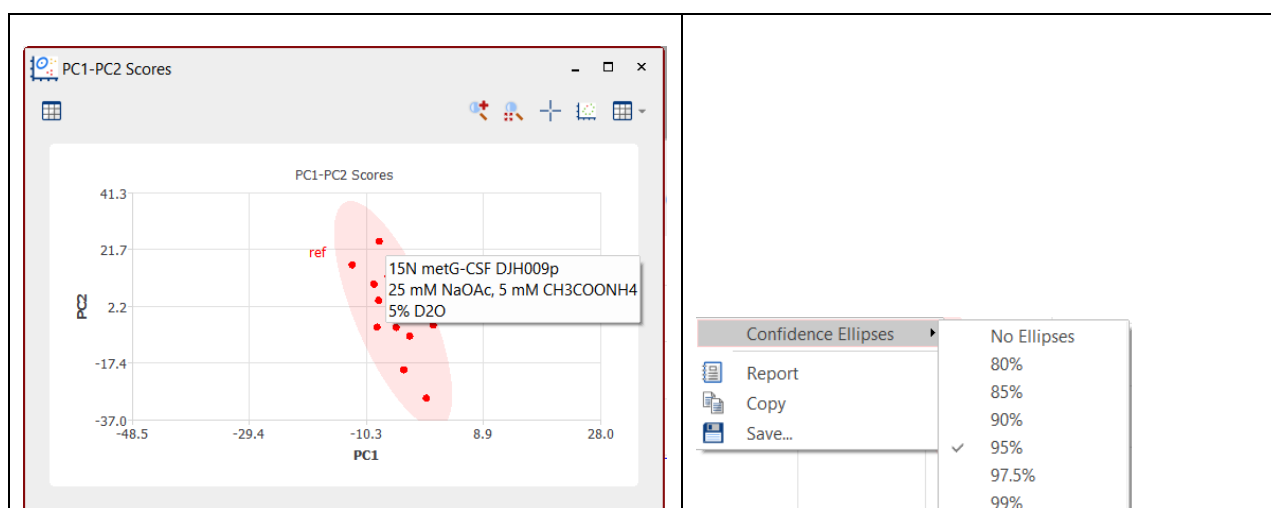


Double-click on any plot in the window to show that single scores plot or select one from the Workspace > View > Scores menu.

We focus on the PC1-PC2 scores plot to see the distance between the test spectrum and the reference spectra cluster.



Hotelling's T^2 ellipses for classes in the score plot can be selected with a right mouse click in the plot areas. Here, the confidence ellipses Limit can be set.



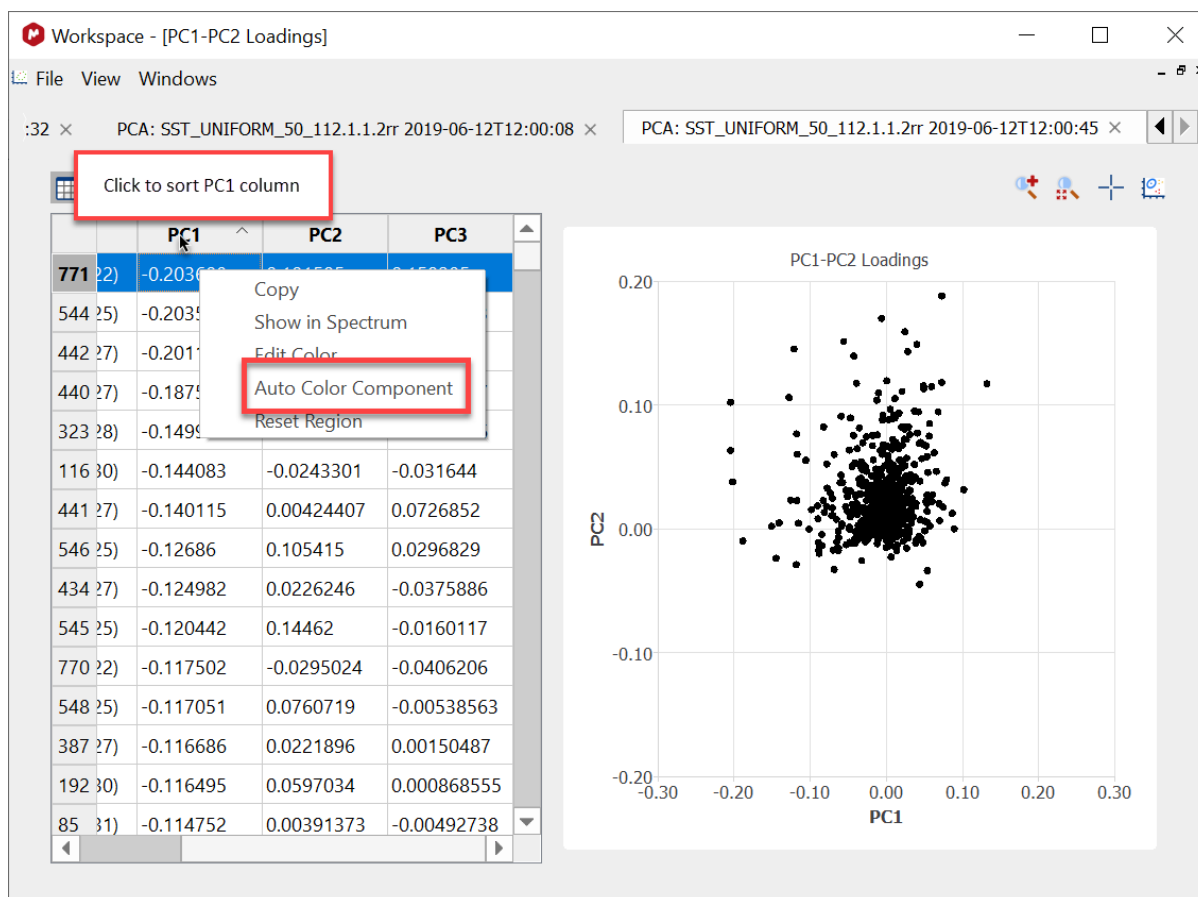
PCA Loadings

The correct loading plot for a specific component pair can be viewed with a button click (see figure, above) or from the *View* menu.

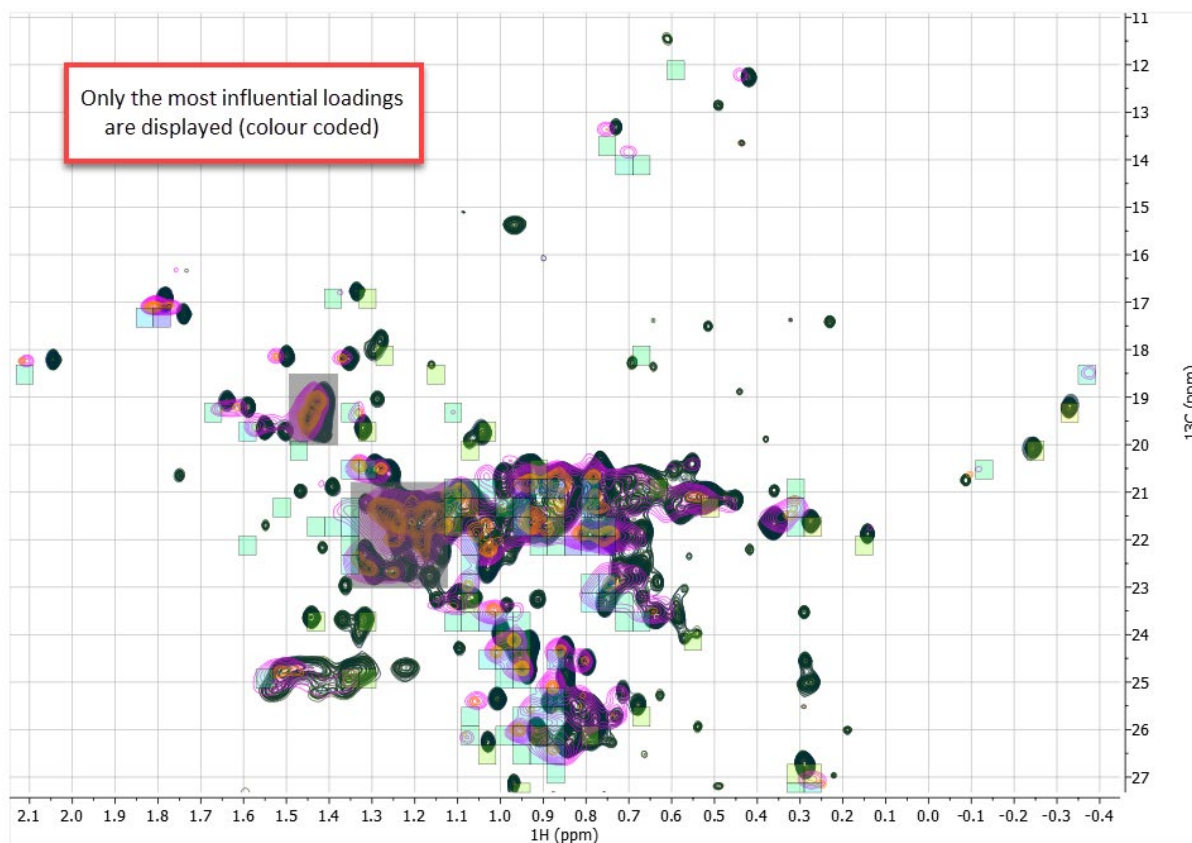
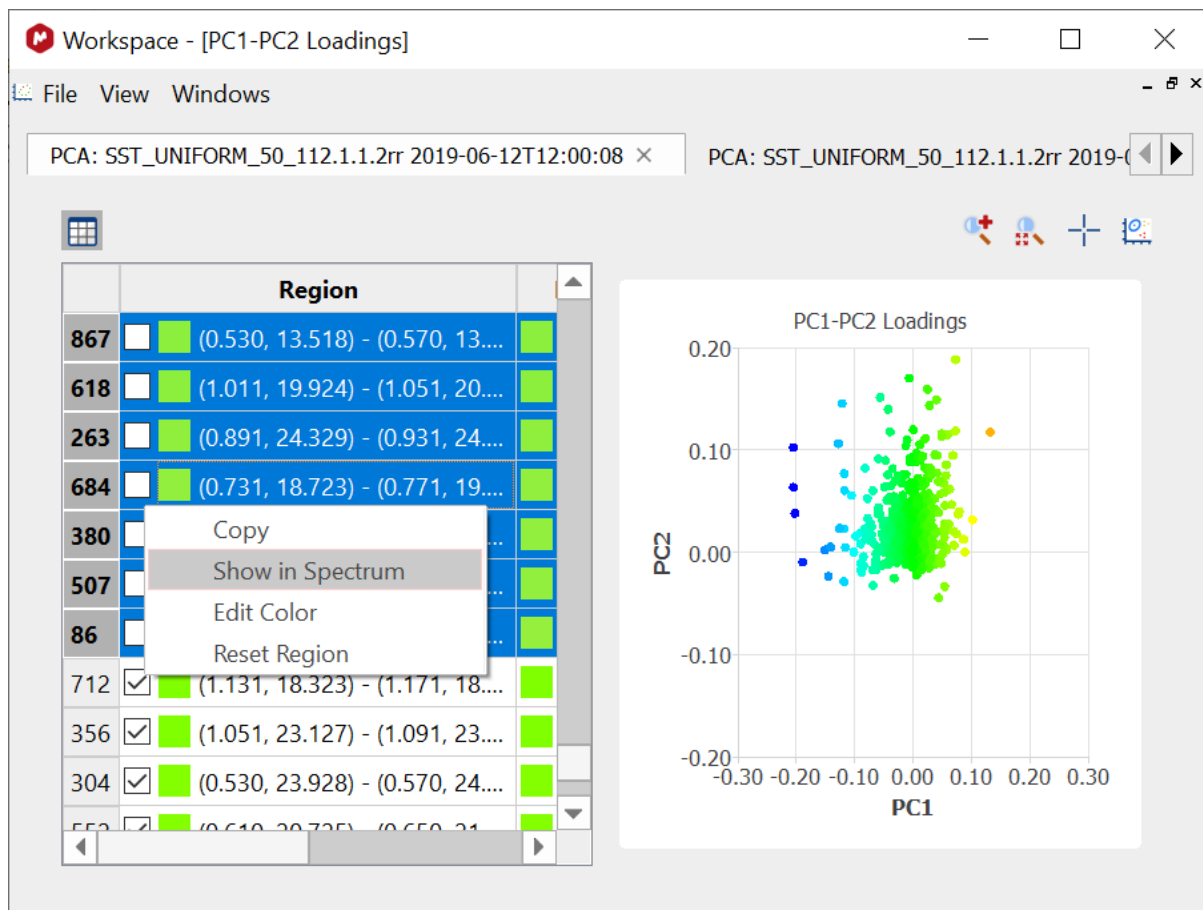
Each point on the loadings plot relates to a specific bucket on the NMR spectrum, and loadings furthest from the origin have the most influence. We, therefore, need ways to connect the loadings to the NMR spectrum.

It is useful to (a) color the loadings as a heat map, and (b) show only the most important buckets – the coldest and hottest colors.

1. Order the PC1 values by clicking on the header in the table
2. With a right mouse button click, choose to Auto color components



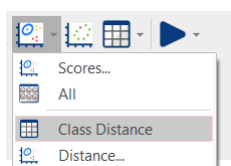
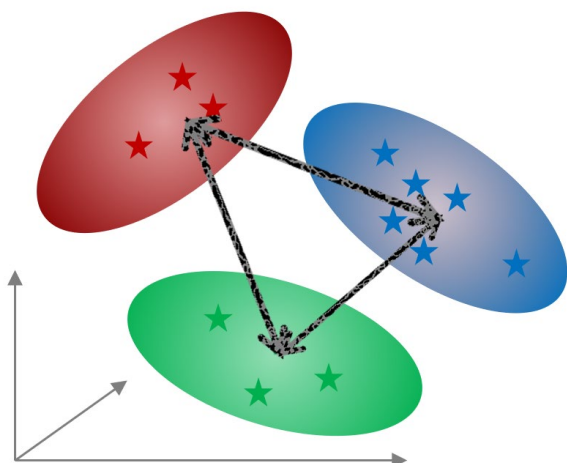
Now we can select the most influential bins to be shown on the NMR spectrum: select a row somewhere near the top of the table. Holding down the Shift key, select another one near the bottom of the table. These bins will usually be colored green. The selected rows show as having a blue background. Now, if you click with the right mouse button you can deselect these rows from being shown.



Distance calculations

The Mahalanobis distance describes how many standard deviations a point is from the mean of a distribution. Distances are calculated between classes or between classes and an individual spectrum. Chen et. al^{14,15} showed the application of distances to compare Biologics.

Global Class distances are calculated from multi-dimensional ellipsoids using either the selected number of principal components or two selected PCs.



Workspace - [Class Distance]

File View Test Windows

PCA: blend_demo_data/0p_ns64_323_900_5 2020-08-13T08:25:57

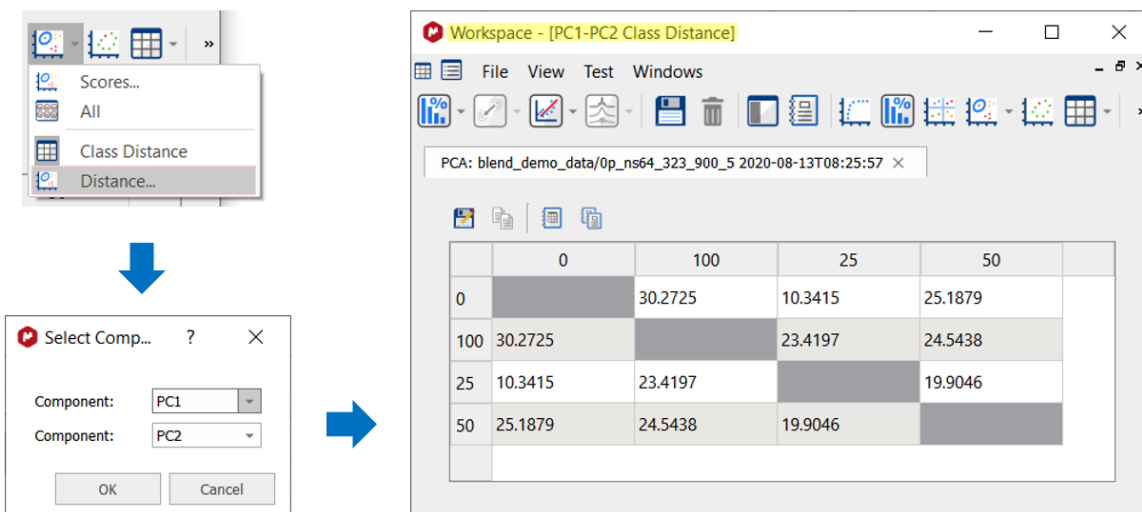
	0	100	25	50
0		48.3516	20.1598	25.3145
100	48.3516		23.4624	32.3866
25	20.1598	23.4624		34.5023
50	25.3145	32.3866	34.5023	

Using 3 components

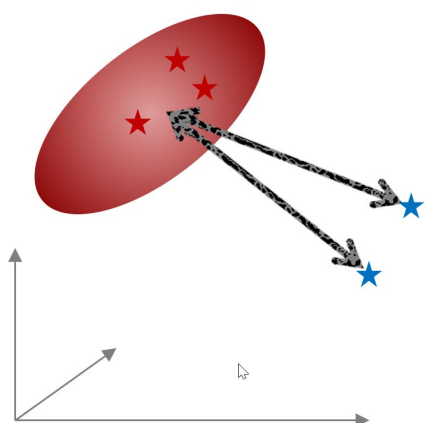
Distances, taking only two PCs into account, are also available.

¹⁴ Chen, K., Park, J., Li, F. et al. Chemometric Methods to Quantify 1D and 2D NMR Spectral Differences Among Similar Protein Therapeutics. *AAPS PharmSciTech* **19**, 1011–1019 (2018). <https://doi.org/10.1208/s12249-017-0911-1>

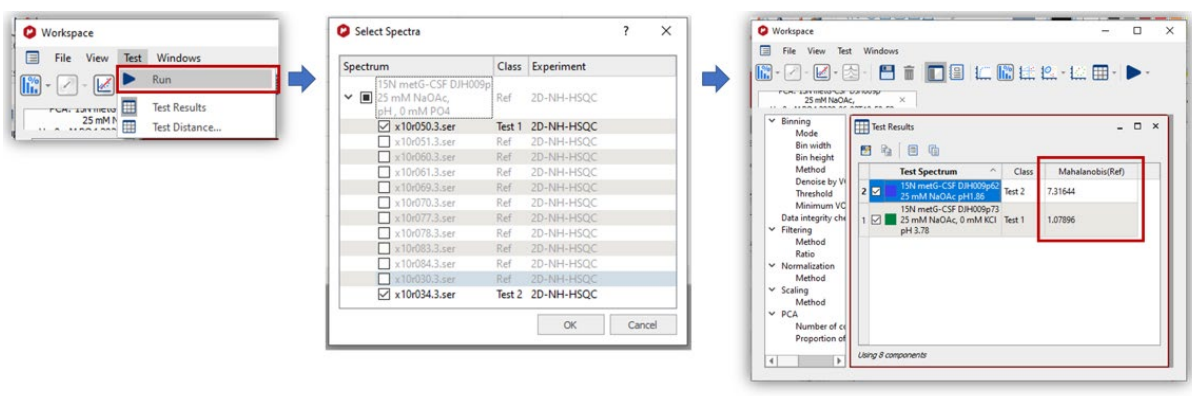
¹⁵ Wang, D.; Park, J.; Patil, S. M.; Smith, C. J.; Leazer, J. L.; Keire, D. A.; Chen, K. An NMR-Based Similarity Metric for Higher Order Structure Quality Assessment Among U.S. Marketed Insulin Therapeutics. *J. Pharm. Sci.* **2020**, *109* (4), 1519–1528. <https://doi.org/10.1016/j.xphs.2020.01.002>



It is possible to measure the distance between classes and individual spectra. The spectra can be taken from any loaded document. It allows detecting to which class a spectrum belongs.



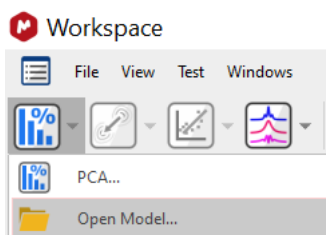
This calculation is available via the *Run* command. Individual spectra can be selected.



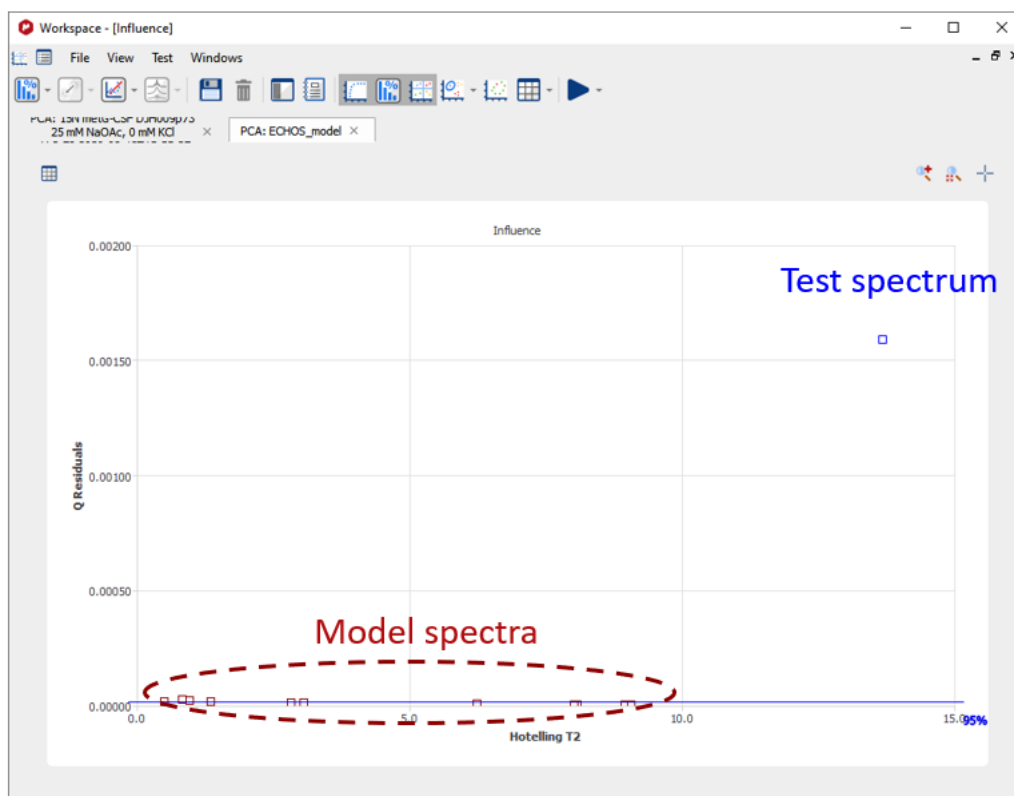
PCA testing using a model

Test spectra (aka "training dataset") of "target" material (e.g., QC pass) are used to create a model. Before use, the model should be cross-validated (Q2, R2).

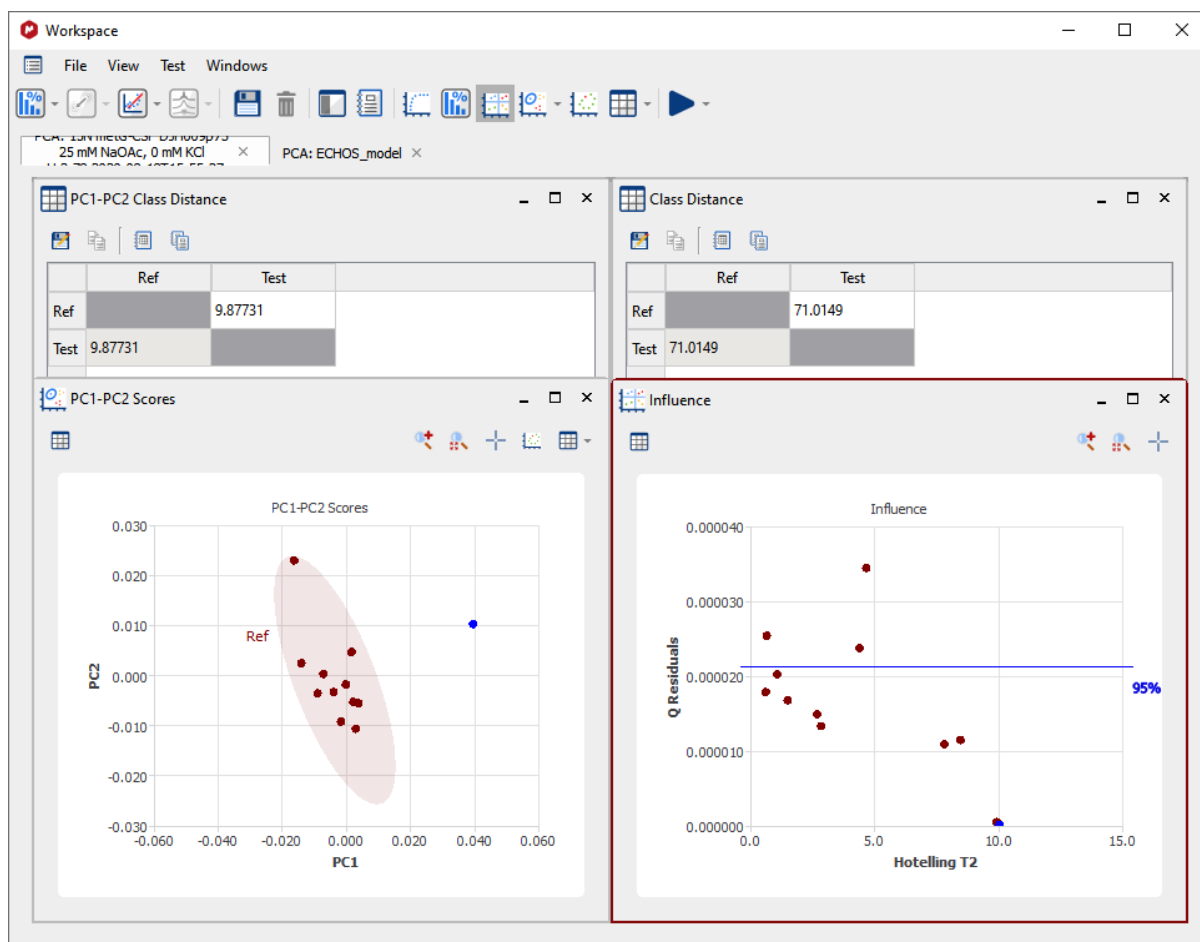
The model can be saved as a file, which can then be used to test new samples without opening the model's spectra. The current PCA analysis can be stored via *File > Save Model*. Loading of the model is possible from the drop-down menu of the PCA icon:



Select the spectra to test and call *Run* to perform the test. Use the Influence plot to see if the model explains the test spectra. It is also possible to use a set of spectra to create a model "on the fly" and then compare test spectra against this. No model file is required.

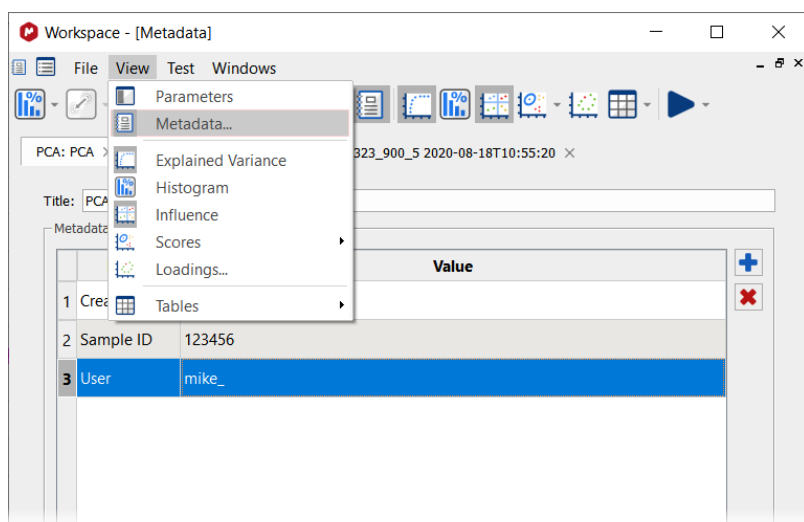


Distance to Model values and scores plots are useful for inspection. The spectroscopic reason for outlier can be detected in combination with the loadings plot.



Metadata

It is possible to create metadata to associate with a sample.

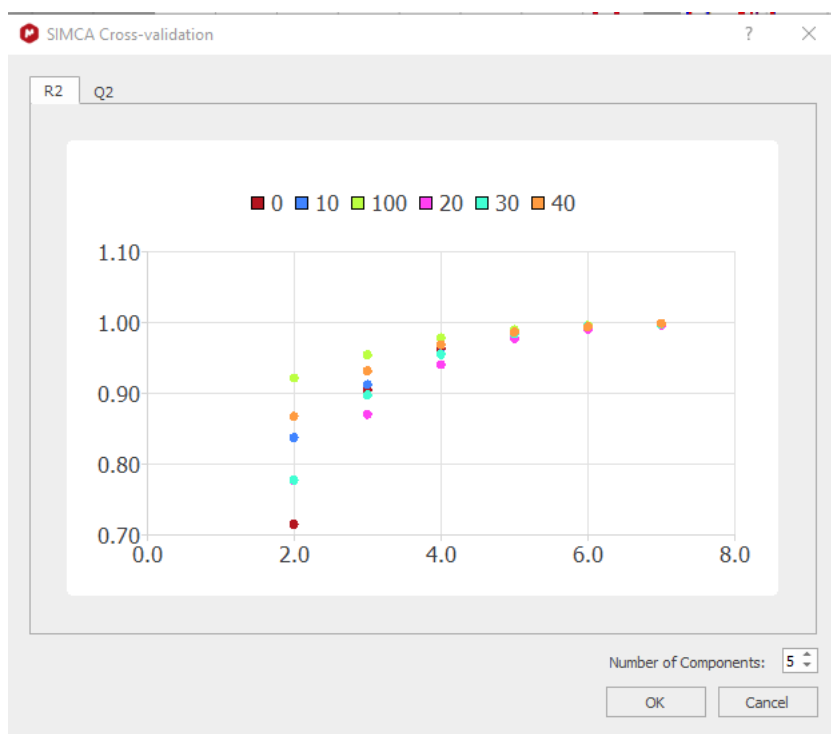


Mixing Mnova with other statistics software

Sometimes, it is necessary to use Mnova to extract the data and partially prepare it for further analysis using their favorite statistical software. We offer a wide range of table output/export options to support this.

11.4. SIMCA Results

After calculation, a window opens showing the R^2 and Q^2 results together with the determined number of principal components.



After clicking on "OK", all results are shown as windows in the Workspace window. Previous analysis results may be accessed from tabs at the windows top, and it is generally possible to zoom into a region.

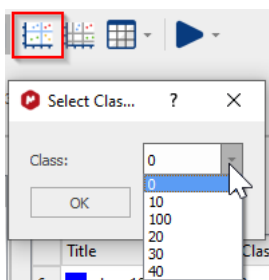
Model Table

The Model Table shows all model spectra and their respective classification. "Green" color means that the sample is classified in this group.

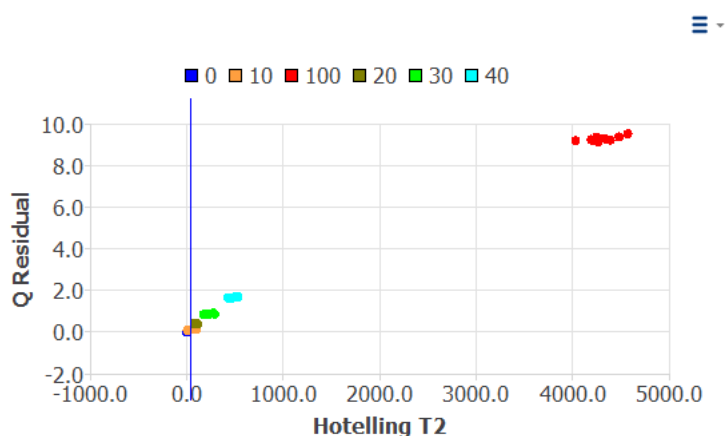
Title	Class	0	10	100	20	30	40
6 depe100As.60.1.1r	0	0.793	273.55	20759.378	1657.997	2025.272	5134.609
7 depe100As.70.1.1r	0	0.74	268.16	20741.664	1639.473	2026.189	5119.117
8 depe100As.80.1.1r	0	0.278	266.149	20689.101	1636.28	2023.898	5110.582
9 depe100As.90.1.1r	0	0.276	273.982	20576.508	1619.122	2032.562	5115.689
21 depe100Bs.10.1.1r	100	31654.166	20863.849	0.345	27384.867	12872.181	10745.969
30 depe100Bs.100.1.1r	100	31065.347	20493.691	0.168	26798.197	12642.683	10519.392
22 depe100Bs.20.1.1r	100	31367.127	20734.458	0.803	27202.676	12794.793	10655.18
23 depe100Bs.30.1.1r	100	32158.048	21265.094	0.287	27968.091	13198.181	10882.899

Influence Plot

The Influence plot can be selected per class.

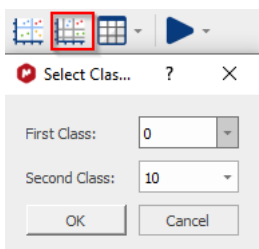


It shows the classification threshold. A general description of the Influence Plot can be found in the PCA section.

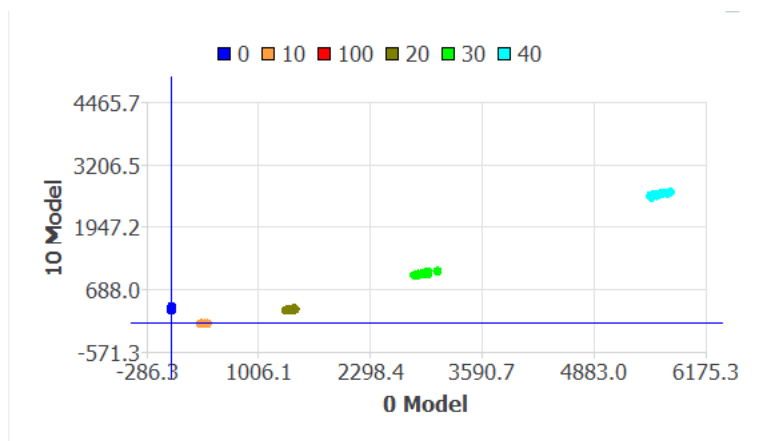


Coomans Plot

The Coomans plot is used to evaluate the classification performance of the model. It shows the distances between samples and class centroids in the PCA space, allowing for the identification of potential misclassifications or outliers. It is used to understand the discriminatory ability of variables or the classification performance of a model in multivariate data analysis. This plot helps to visualize the relationships between variables, classes, and samples in a reduced-dimensional space.

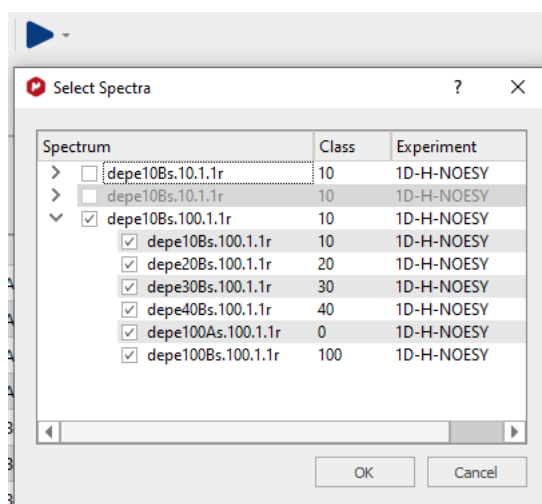


The plot shows the classification threshold



SIMCA Classification

The SIMCA model can be stored and loaded for future use. Classification is done by selecting the spectra to test.



The result is shown in the Test Table, with Green/Red color coding:

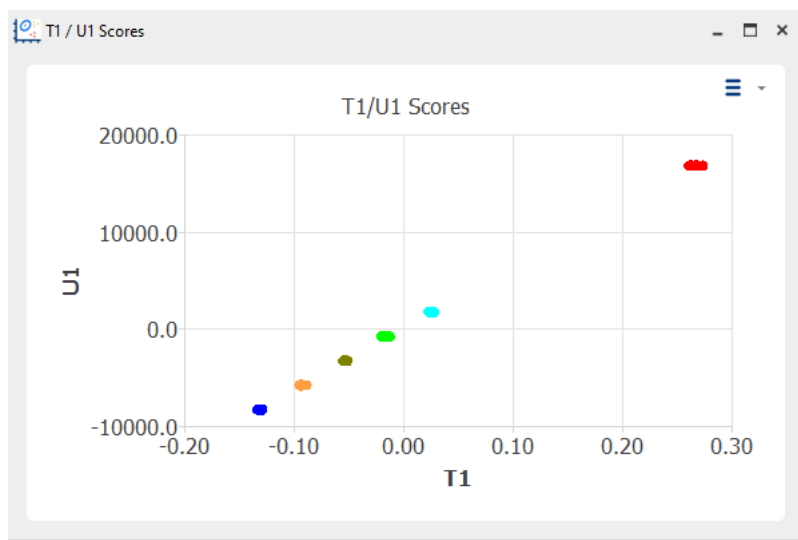
Spectrum	0	10	100	20	30	40
1 depe100As.100.1.1r	0.013	319.003	20439.715	1791.978	2118.463	5212.278
3 depe100Bs.100.1.1r	31065.347	20493.691	0.168	26798.197	12642.683	10519.392
2 depe10Bs.100.1.1r	346.393	0.25	17480.538	430.189	957.344	3204.33
4 depe20Bs.100.1.1r	1419.462	301.488	14653.507	0.181	230.998	1751.019
5 depe30Bs.100.1.1r	2807.651	973.998	11641.852	375.467	0.807	662.305
6 depe40Bs.100.1.1r	5578.607	2613.052	8348.968	2270.053	572.661	0.092

11.5. PLS Results

All results are shown as windows in the Workspace window, which can be minimized, tiled, etc. Previous analysis results may be accessed from tabs at the windows top, and it is generally possible to zoom into a region.

T1/U1 Scores plot

The T/U scores plot is a graphical representation of the scores of the samples in the data matrix X and the scores of the target/response variables Y in the PLS model.



The T/U scores plot can provide useful information about the structure and relationships between the samples in the data set, and the response variables in the PLS model. Specifically, the plot can be used to:

1. Assess model performance I: PLS determines a linear correlation between the samples and response variables. The individual sample location should be close to the diagonal.
2. Assess model performance II: The separation between different groups of samples in the plot can indicate how well the PLS model is able to capture the underlying structure of the data.
3. Identify potential outliers: Samples that are located far away from the diagonal or other samples in the plot may be potential outliers and should be further examined.
4. Identify sample clusters: Samples that are located close to each other in the plot may belong to the same cluster or group and may share similar characteristics.

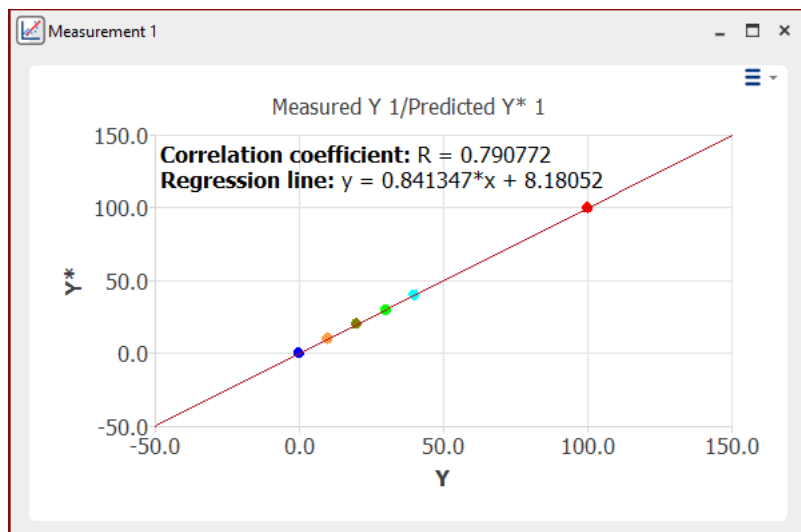
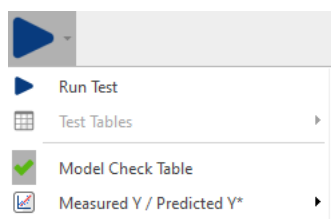
Model Check Table

The Model Check Table gives a numeric overview of the model performance, especially on the performance of the prediction. It shows the class, expected (Y) and predicted (Y*) value. The error, RMSECCV (Root Mean Squared Error of Calibration after Cross Validation), and outlier indicators give additional insight if a sample is an outlier.

	Spectrum ^	Class	Y 1	Y* 1	Error 1	Outlier DmodX	Outlier Leverages	Outlier Spectra Residuals
7	depe10Bs.70.1.1r	10	10.00	10.05	± 0.25			
8	depe10Bs.80.1.1r	10	10.00	9.82	± 0.25			
9	depe10Bs.90.1.1r	10	10.00	10.01	± 0.26			
10	depe10Bs.100.1.1r	10	10.00	10.02	± 0.27			
11	depe20Bs.10.1.1r	20	20.00	20.10	± 0.21			

Measured Y vs. Predicted Y* Plot

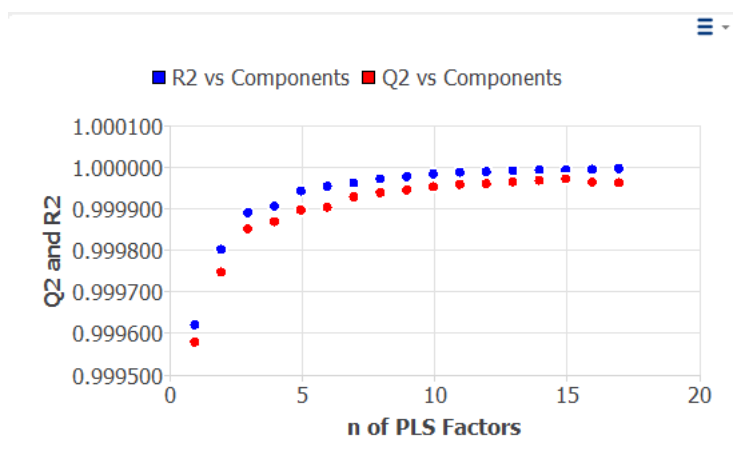
This plot shows the values in the Model Check Table as a graphic. This plot is available from the “Run Test” Icon



The correlation coefficient R measures the degree to which the data points in the Y and Y* scatter plot fall along a straight line. A correlation coefficient close to 1 indicates a strong linear relationship between Y and Y*, while a value close to 0 indicates no correlation.

R² / Q²

Q² and R² are both measures of the goodness of fit of a regression model, but they are calculated and interpreted differently.



R², or the coefficient of determination, is a measure of the proportion of the variance in the response variable that is explained by the predictor variables in the model. Specifically, R² is calculated as the ratio of the explained variance to the total variance in the response variable. R² ranges from 0 to 1, where 0 indicates that the predictor variables have no explanatory power and 1 indicates that the predictor variables explain all the variance in the response variable.

Q², or the cross-validated R², is a measure of the predictive power of the regression model, calculated using cross-validation. Cross-validation is a technique that involves splitting the data set into training and validation sets, fitting the regression model to the training set, and then predicting the response variable in the validation set using the fitted model. Q² is calculated as the ratio of the explained variance in the validation set to the total variance in the validation set. Q² ranges from -1 to 1, where 1 indicates a perfect predictive power of the model, 0 indicates that the model is no better than random guessing. Negative values indicate that the model performs worse than random guessing.

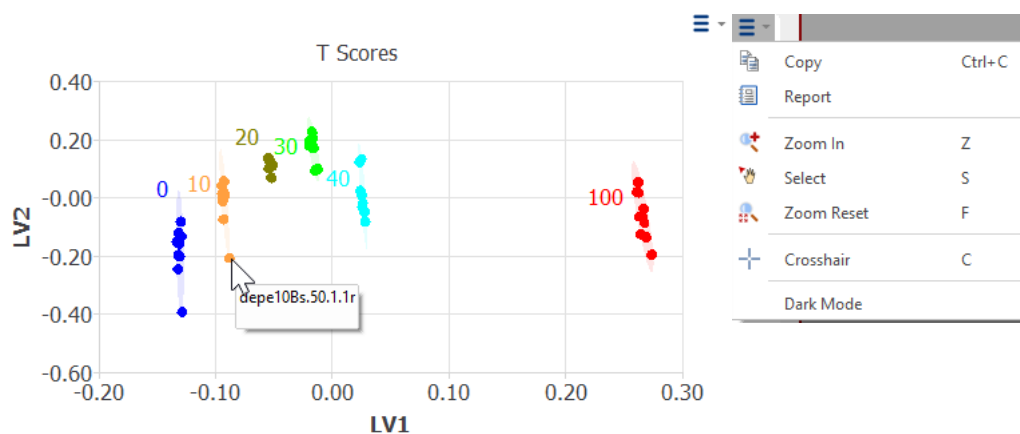
A good choice of a regression model depends on the specific goals of the analysis and the nature of the data set. In general, a good model should have a high R² and a high Q², indicating that the predictor variables explain a large proportion of the variance in the response variable and that the model has good predictive power. However, a high R² does not necessarily imply a high Q², and a model with a high R² but a low Q² may suffer from overfitting, meaning that it captures noise in the data set rather than the underlying patterns. Therefore, it is important to evaluate both R² and Q² to ensure that the model has a good balance between explanatory power and predictive power, and to use appropriate validation techniques to avoid overfitting.

PLS T / U scores

The T-scores provide a reduced-dimensional representation of the original X data that captures the most important variation in the data and is most strongly related to the response variable Y. Each point in the T-scores represents a spectrum. U-scores are based on the original Y values.

The scores can be used for various purposes, such as:

1. To visualize the relationships between the samples in the data set and the latent variables in the PLS model.
2. To identify potential outliers or clusters of samples in the data.
3. To assess the performance of the PLS model in predicting the response variable Y.
4. To make predictions of the response variable for new samples based on their scores on the PLS components.



The scores plot shows the name of the samples and confidence ellipses. Additional commands are available via the upper right corner icon or a right mouse click.

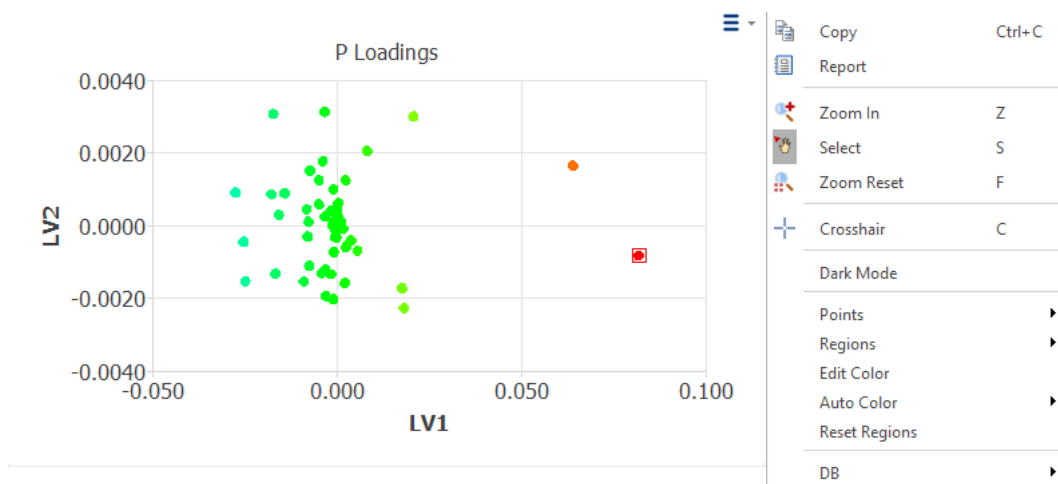
PLS P/Q loadings

In Partial Least Squares (PLS), P-loadings are the estimated values of the linear combinations of the original variables in the data matrix X that are used to construct the latent variables or components in the PLS model. Q-loadings are based on the Y data matrix.

The loadings are estimated during the PLS analysis to maximize the covariance between the X matrix and the response variable Y. The loadings can be used for various purposes, such as:

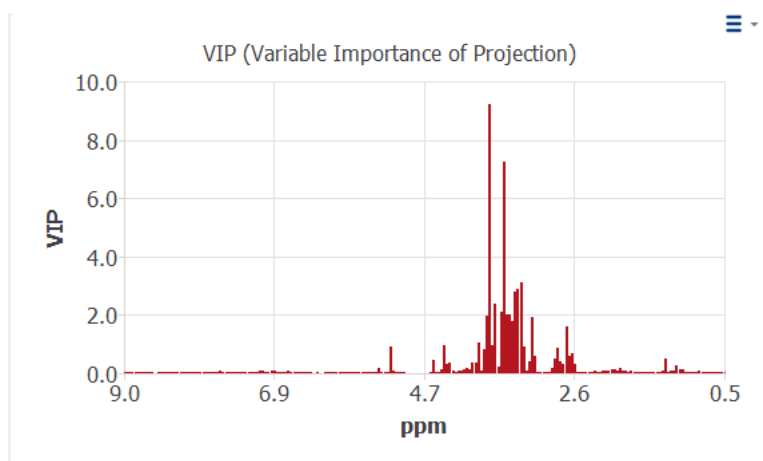
1. To identify the variables most strongly related to each component and assess each variable's importance in the PLS model.
2. To visualize the relationships between the variables in the data set and the latent variables in the PLS model.
3. To identify potential collinearity issues or redundancies among the variables in the data set.
4. To interpret the PLS model and to gain insights into the underlying structure and relationships among the variables in the data set.

An advantage of the BioHOS software is that we can correlate the loadings with the spectra.



Variable Importance in the Projection (VIP)

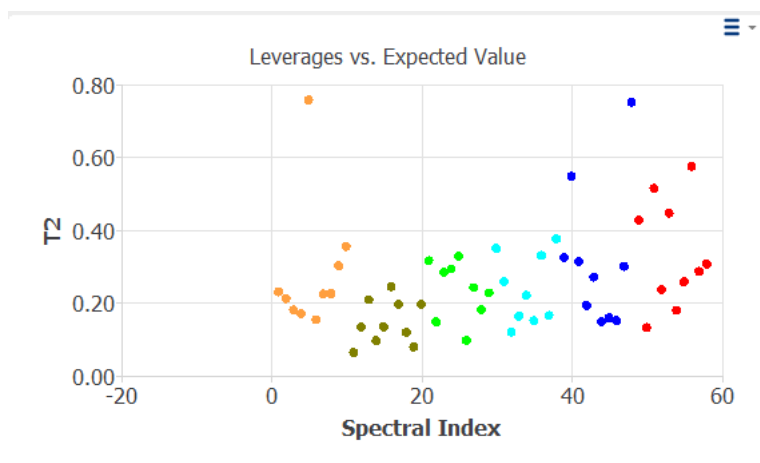
The Variable Importance in the Projection (VIP) plot is a diagnostic tool to assess the importance of each variable (bin) in the PLS model. The VIP plot shows the VIP score for each variable, which measures the variable's importance in explaining the variance in the response variable Y.



The VIP score indicates the relative importance of each variable in the PLS model, taking into account the contribution of all components to the explained variance in Y. A larger VIP score indicates that the corresponding variable is more important in explaining the variance in Y and has a larger impact on the model. Conversely, a smaller VIP score indicates that the corresponding variable is less important in explaining the variance in Y and has a smaller impact on the model. Typically, variables with VIP scores greater than 1 are considered important in the model, while variables with VIP scores less than 0.8 or 0.5 may be considered less important.

Leverage

Leverage is a diagnostic tool to assess the influence of each observation on the regression model. In simple terms, leverage measures how far an observation is from the center of the predictor variable space, relative to the other observations in the data set.

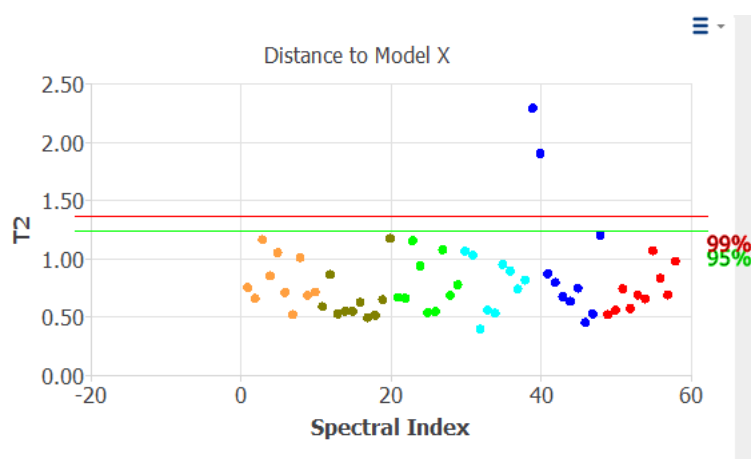


Leverage can be used to identify influential observations in the data set that may greatly impact the regression model. Observations with high leverage are located far from the center of the predictor variable space and may have a large influence on the estimation of the regression coefficients. These observations are sometimes called "leverage points" or "high-leverage points".

However, high leverage alone does not necessarily mean that an observation is influential. The influence of an observation on the regression model also depends on the value of the response variable and the other predictor variables. In addition, the presence of outliers or collinearity in the data set may affect leverage interpretation.

Distance to model (DMod)

Distance to model (DMod) is a diagnostic tool to assess the model's goodness of fit and to identify potential outliers or influential observations in the data set. DMod is a measure of the distance between each observation in the data set and the PLS model in terms of the predicted values and the estimated uncertainties of the model.



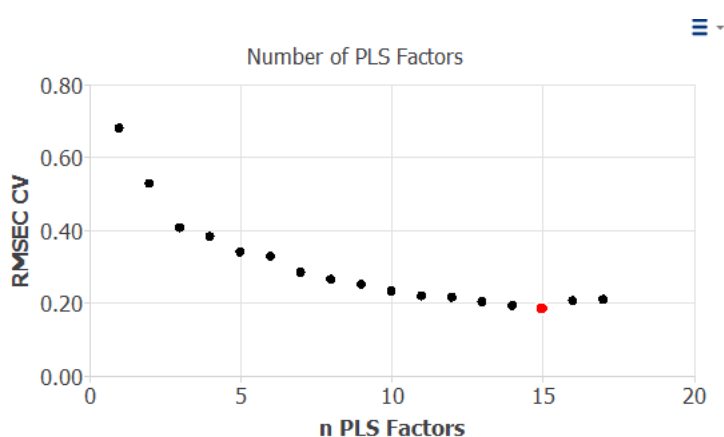
DMod is calculated as the difference between the actual response values for each observation and the corresponding predicted values from the PLS model, normalized by the estimated residual standard deviation of the model.

The DMod values can be interpreted as a measure of the relative distance of each observation from the PLS model, in units of the estimated residual standard deviation of the model. A larger DMod value indicates that the corresponding observation is farther away from the model and maybe a potential outlier or influential observation. Conversely, a smaller DMod value indicates that the corresponding observation is closer to the model and has a smaller influence on the model.

Common thresholds for identifying potential outliers or influential observations are 95% or 99% confidence intervals. However, the threshold may depend on the specific data set and the purpose of the analysis.

Number of PLS factors

One possibility to determine the number of PLS factors is to find an optimal RMSECCV. RMSECCV stands for Root Mean Squared Error of Cross-Validation. It is a measure of the prediction error of a regression model, calculated using cross-validation. Cross-validation is a technique that involves splitting the data set into training and validation sets, fitting the regression model to the training set, and then predicting the response variable in the validation set using the fitted model. RMSECCV is calculated as the square root of the mean squared error across all validation sets.



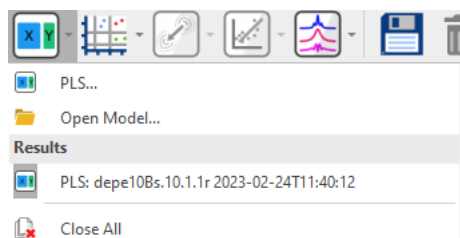
RMSECCV is a measure of the average prediction error of the model in units of the response variable. A lower value of RMSECCV indicates a better predictive performance of the model, meaning that the model can accurately predict the response variable for new observations. However, the interpretation of RMSECCV depends on the specific data set and the range of the response variable. For example, a RMSECCV value of 10 may be considered high if the range of the response variable is small, but may be considered low if the range of the response variable is large.

RMSECCV can be used to compare different regression models' performance or select the optimal number of components. However, it should be noted that RMSECCV is sensitive to outliers and the distribution of the response variable, and may not capture all aspects of the model's predictive performance. Therefore, it is important to evaluate other measures such as Q^2 and R^2 , and to use appropriate validation techniques such as external validation to ensure the generalizability of the model.

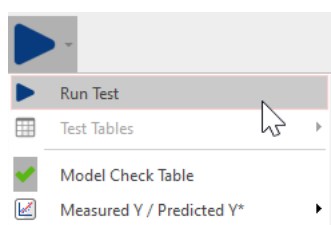
PLS Prediction

PLS prediction uses the identified components to predict the value of Y for new spectra. It is achieved by binning the spectra in the same way as it was used for model building. The regression coefficients of the model are applied to the new binning table calculate the predicted value of Y.

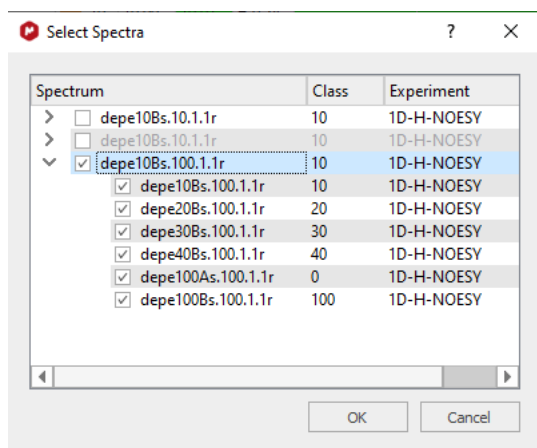
The PLS Model can be stored and loaded for PLS prediction.



The PLS prediction is available via the “Play” icon.



In the spectra selection dialog, the spectra used to create the model are grey but can be selected.



The results are presented in a table similar to the Model Check Table.

	Spectrum	Class	Y 1	Y* 1	Error 1	Outlier DmodX	Outlier Leverages	Outlier Spectra Residuals
1	depe10Bs.100.1.1r	10		10.02	± 0.27			
2	depe20Bs.100.1.1r	20		20.13	± 0.24			
3	depe30Bs.100.1.1r	30		29.91	± 0.25			
4	depe40Bs.100.1.1r	40		40.01	± 0.27			
5	depe100As.100.1.1r	0		0.04	± 0.32			
6	depe100Bs.100.1.1r	100		100.11	± 0.26			

In the column Y* the predicted values are presented, PLS checks if the model can explain the spectra by checking the Distance to Model, Leverage and Spectral Residuals. The error value is based on RMSECCV and how well the model can explain the spectrum.