

BioPAT® Spectro Ambr® calibration in SIMCA®

A technical description describing how to create and evaluate a multivariate calibration model in SIMCA for use in a Raman BioPAT Spectro Ambr application.

Using spectroscopic techniques like NIR, Mid-IR, Raman, UV, and Fluorescence Spectroscopy, hundreds of signals are recorded per sample or process time point. The use of SIMCA to analyze these multivariate data provides an overview of the signals and translates them into information about analyte concentrations and process quality. The fast and non-invasive spectroscopic sensors make up an important part of the implementation of Process Analytical Technology (PAT) and can in combination with Umetrics multivariate tools provide the basis for a control strategy to be applied in a manufacturing process.

Objective

In this tutorial you will learn how to import Raman spectra and reference concentration measurements combined in a single CSV file exported from the BioPAT Spectro Ambr application. The spectral data and the reference concentration will then be preprocessed and modelled in SIMCA to create a calibration model that can be deployed as a basis for real-time concentration predictions in the BioPAT Spectro Ambr application which the SIMCA-Q prediction engine embedded. The created model can in this context be used for fast routine analyte concentration determination and process control.

The tutorial will describe all steps necessary for applying a selected set of spectral preprocessing algorithms, fitting a multivariate prediction model for a selected analyte and evaluate the quality of that model. The result of the described procedure is a SIMCA file (.usp) containing a calibration model that can be implemented in the BioPAT Spectro Ambr application for analyte predictions based on Raman spectra.

Steps involved are:

- Import and define the BioPAT Spectro Ambr data
- Create a SIMCA project
- Raw data overview
- Spectral preprocessing of Raman data from a bioprocess
- Create the calibration model
- Evaluate the calibration model performance
- Transferring the calibration model to BioPAT Spectro Ambr

There is also a description on how to install and use spectral preprocessing plugins allowing for custom preprocessing to be used within the SIMCA framework. The procedure described in this tutorial can be applied to any case where a predictive multivariate calibration model is desirable. The description should be seen as an example workflow that can be modified to best fit available data and specific situation.

It is important that relevant calibration data is collected from the Ambr system to create reliable analyte prediction models. This document will not describe how to generate good quality calibration data but it is strongly recommended that some type of statistical experimental design procedure is used.

The SIMCA® 16 software

Data import

SIMCA supports the direct import of standard file formats such as spc, JCAMP, xls and csv, as well as a number of instrument vendor formats.

Spectroscopy skin

In addition to the standard (default) layout of ribbons and buttons, SIMCA comes with a special “Spectroscopy skin” which can be useful when working with spectroscopy data. The Spectroscopy skin uses a separate ribbon layout which includes special wizards and changes the default plots.

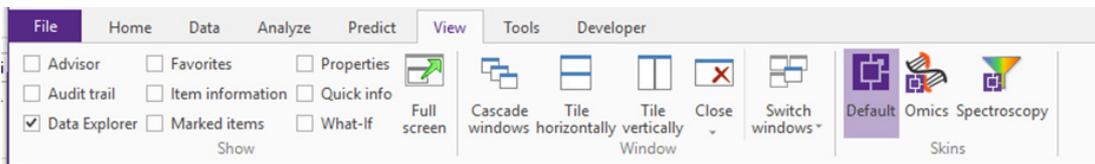
Note: This tutorial requires the use of the Spectroscopy skin for importing the data. Switching to the default view can be done any time after Import.

Some benefits of using the Spectroscopy skin

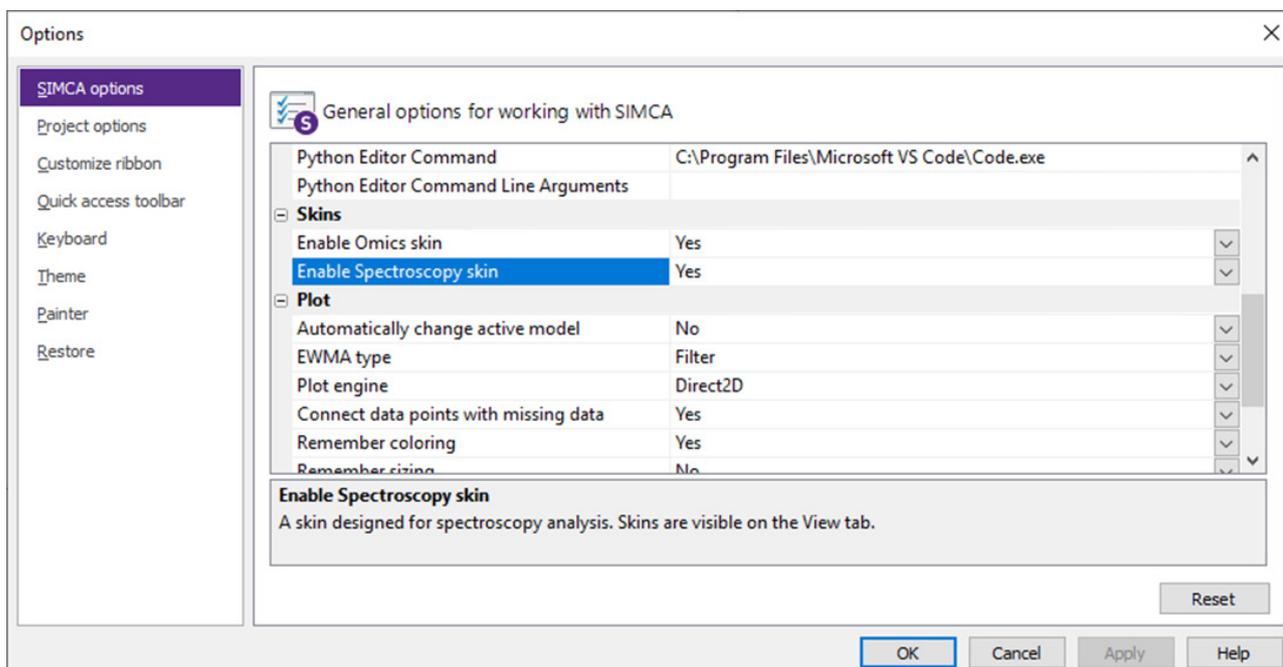
- Specifying X and Y at import will automatically split the data into two datasets
- The default scaling of spectral data is set to centered (ctr)
- Option to define which identifier to use as spectral axis
- Dialog to create multiple spectral plots simultaneously
- Compare filter wizard allowing for parallel creation of alternatively preprocessed models
- Direct import and modelling and prediction of new spectra

Switching between different SIMCA skins

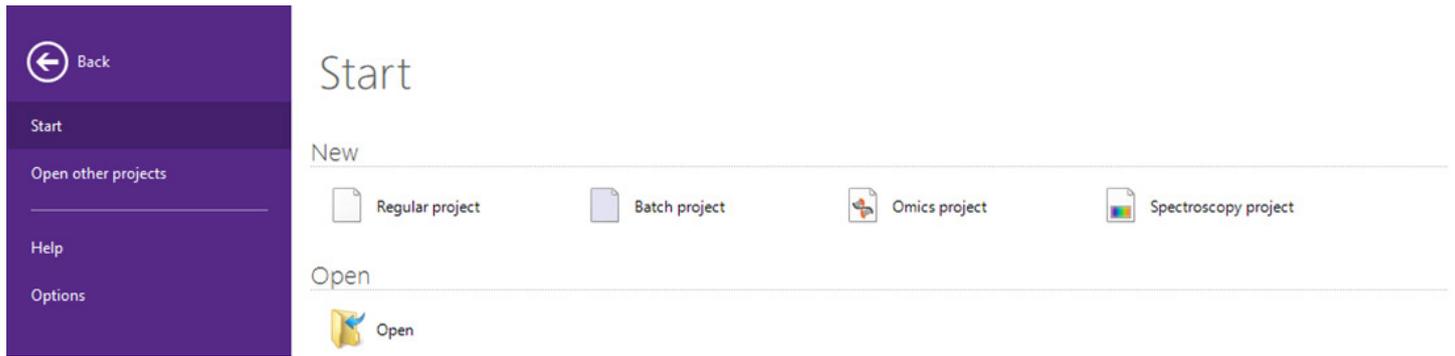
SIMCA 16 is delivered with three different skins: Default, Omics and Spectroscopy. In the view tab, you can switch between these at any point during the analysis.



If you cannot see the three Skin buttons, you need to enable them in SIMCA options. Click on File | Options and select SIMCA options. Then scroll down to the Skins section and enable the ones you want to use.



Once the skins are enabled, you will also have the possibility to choose to create an Omics or Spectroscopy project from the Start page



Using Python plugin for spectral preprocessing

The BioPAT Spectro data used in this tutorial requires the use of spectral preprocessing not included in the SIMCA software. However, SIMCA 16 has the ability to utilize custom spectral preprocessing plugins created in C++ or Python. This tutorial will use a Python plugin to access a baseline correction and a peak area normalization.

To be able to use the plugin you must first acquire the UmPyFilters.py file. Available as download from Sartorius Data Analytics or from Umetrics support (umetrics_support@sartorius.com). This file must be copied into the SIMCA 16 Python directory: "%Userprofile%\AppData\Roaming\Umetrics\SIMCA\16.0\Python" to be available inside SIMCA.

Creating a multivariate calibration model for BioPAT Spectro

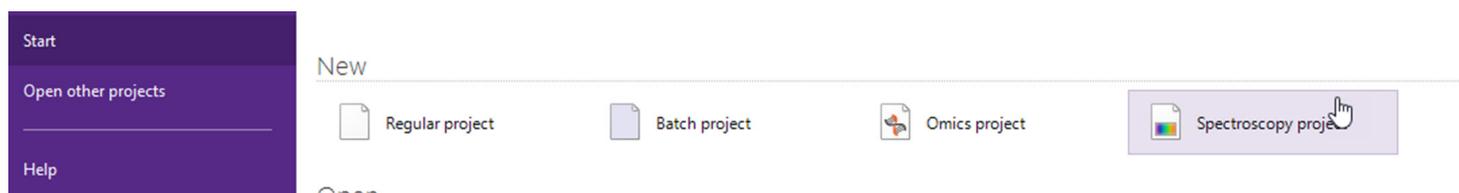
Import and define the BioPAT Spectro Ambr data

Data can come in one or two files from the Ambr application. Here we will show the one file case where spectral data and analyte concentrations are combined in one comma separated text file (.csv). If the calibration data is delivered in two files, one with spectral data and the other with analyte concentrations you will have to import the files separately. Sometimes, the spectrometer delivers spectra in individual files, one spectrum per file, (e.g. in .spc format), which you can import all at once into SIMCA.

Create a SIMCA project

The first step in the data analysis is to create a SIMCA project which will hold all your data and models. The project is stored in a single file with the extension .usp which is the file that you later will transfer back to the Ambr system for model execution.

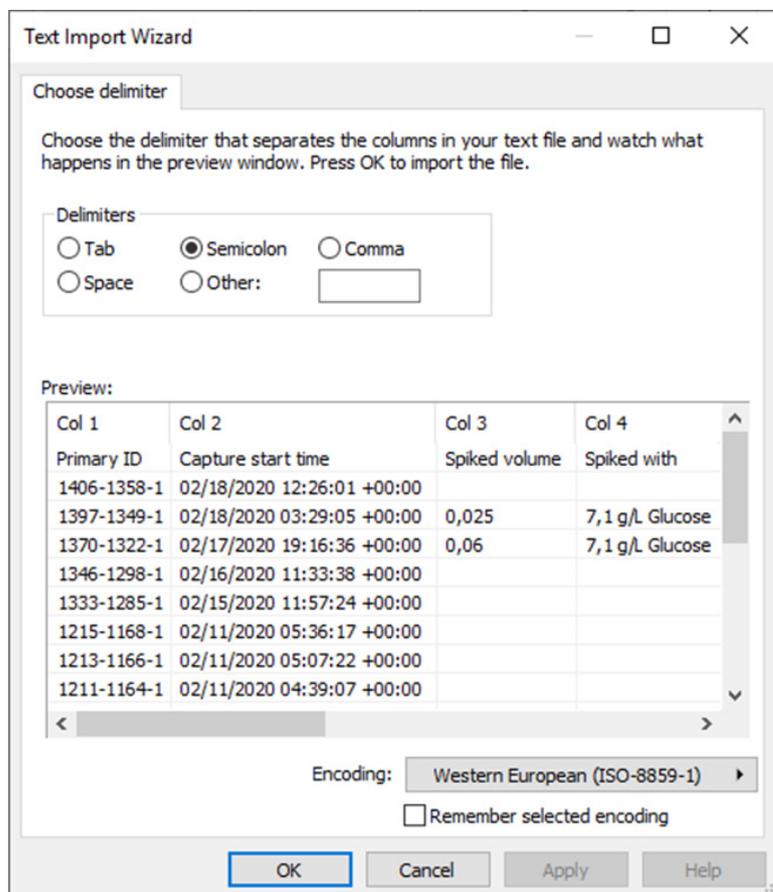
Start SIMCA 16 and select "Spectroscopy project" from the "New" section. If you do not see the "Spectroscopy project" button you need to activate the Spectroscopy skin.



After selection to create a new spectroscopy project, the SIMCA import module will start and you are ready to import your Ambr .CSV file.

If you do not see a File Open dialog, click on the Add data button and select From file.

Locate your Ambr data file and select it. A CSV text import wizard opens and make sure the right delimiter is selected before clicking OK.



The CSV file contains both spectral data and analyte concentrations together with sample information and is opened in a spreadsheet view where you need to define the column content.

During import, SIMCA will guess what the columns are but this must be verified before continuing. It is important the imported dataset has one unique sample, or observation, identifier (Primary ObsID) and one unique variable identifier (Primary VarID). These are typically the first column and row respectively. In the CSV file exported from the BioPAT Spectro data management system, the first column contains unique sample names and should be set as Primary ID. If not already defined as Primary ID, highlight the column and click on the Primary observation ID button. Verify also that the first row is defined as Primary variable ID.

The screenshot shows the SIMCA Import dialog box with the following column definitions:

Column	Variable ID	Observation ID	Data Type	Variable Role
Primary ID	Primary variable ID	Secondary observation ID	Quantitative	Y-variable
Capture start time	Secondary variable ID	Secondary observation ID	Qualitative	Exclude
Spiked volum	Secondary variable ID	Secondary observation ID	Quantitative	Y-variable
Spiked with	Secondary variable ID	Secondary observation ID	Qualitative	Exclude
Glucose	Secondary variable ID	Secondary observation ID	Quantitative	Y-variable
Glucose (befo	Secondary variable ID	Secondary observation ID	Quantitative	Y-variable
Lactate	Secondary variable ID	Secondary observation ID	Quantitative	Y-variable
Lactate (befo	Secondary variable ID	Secondary observation ID	Quantitative	Y-variable
Glutamine	Secondary variable ID	Secondary observation ID	Quantitative	Y-variable
Glutamine (be	Secondary variable ID	Secondary observation ID	Quantitative	Y-variable
Ammonia	Secondary variable ID	Secondary observation ID	Quantitative	Y-variable
Vessel number	Secondary variable ID	Secondary observation ID	Quantitative	Y-variable
Water band n	Secondary variable ID	Secondary observation ID	Quantitative	Y-variable
Glutamate	Secondary variable ID	Secondary observation ID	Quantitative	Y-variable
Glutamate (be	Secondary variable ID	Secondary observation ID	Quantitative	Y-variable
Batch Id	Secondary variable ID	Secondary observation ID	Quantitative	Y-variable

It is also important to define which variables (columns) contain the analyte concentrations (Y-data) and which contain the spectral data (X-data). Select the analyte columns one by one and click on the Y-variable button. The spectra part of the file is to the far right and should already be defined as X-variables (indicated by white cells). All other columns should be defined as Secondary observation ID. Secondary ID:s can be used inside SIMCA for sample search, plot labels, and colors.

We have now defined what we need in the import and can click on the "Finish import" button to proceed into SIMCA after giving the project a name in the dialog that opens.

SIMCA has split the spectral data (X) and the reference concentration data (Y) in two separate datasets for a better overview.

If you import data from separate files, make sure you end up with two datasets as shown below. One dataset with all spectral data (X-data) and one dataset with all the analyte concentrations (Y-data).

The screenshot shows the SIMCA software interface with two datasets:

Dataset - BioPAT Spectro ambr data

Primary ID	Glutamine (before spiking)	Vessel number	Water band normalisation	Glutamate (before spiking)	Batch Id	100	101	102	103	104	105	106	107	108
1	0,82	1	12,6051	0,053	45	385262	371928	357572	342506	327075	311558	296103	280792	
2	0,8	6	14,4933	0,099	50	611468	587288	562690	537022	510071	482359	454928	428763	
3	0,8	4	15,0888	0,177	48	508010	487820	466431	444340	421730	398728	375703	353340	
4	0,55	7	14,6158	0,037	51	433976	418141	400385	381002	360667	340125	319896	300254	
5	0,47	7	15,5371	0,031	51	432926	417098	399928	381596	362644	342961	323450	304242	
6	0,69	1	16,1482	0,403	45	408922	396173	382463	367797	352321	336282	319939	303567	
7	0,67	8	17,55	0,67	7	15,7977	15,6201	15,6315	15,7491	15,7561	15,8912	15,939	16,1482	
8	0,72	7	17,55	0,67	7	15,7977	15,6201	15,6315	15,7491	15,7561	15,8912	15,939	16,1482	
9	0,67	7	17,55	0,67	7	15,7977	15,6201	15,6315	15,7491	15,7561	15,8912	15,939	16,1482	
10	0,67	7	17,55	0,67	7	15,7977	15,6201	15,6315	15,7491	15,7561	15,8912	15,939	16,1482	
11	0,67	7	17,55	0,67	7	15,7977	15,6201	15,6315	15,7491	15,7561	15,8912	15,939	16,1482	
12	0,67	7	17,55	0,67	7	15,7977	15,6201	15,6315	15,7491	15,7561	15,8912	15,939	16,1482	
13	0,67	7	17,55	0,67	7	15,7977	15,6201	15,6315	15,7491	15,7561	15,8912	15,939	16,1482	
14	0,67	7	17,55	0,67	7	15,7977	15,6201	15,6315	15,7491	15,7561	15,8912	15,939	16,1482	
15	0,67	7	17,55	0,67	7	15,7977	15,6201	15,6315	15,7491	15,7561	15,8912	15,939	16,1482	
16	0,67	7	17,55	0,67	7	15,7977	15,6201	15,6315	15,7491	15,7561	15,8912	15,939	16,1482	
17	0,67	7	17,55	0,67	7	15,7977	15,6201	15,6315	15,7491	15,7561	15,8912	15,939	16,1482	
18	0,67	7	17,55	0,67	7	15,7977	15,6201	15,6315	15,7491	15,7561	15,8912	15,939	16,1482	

Dataset - Y-Variables(BioPAT Spectro ambr data)

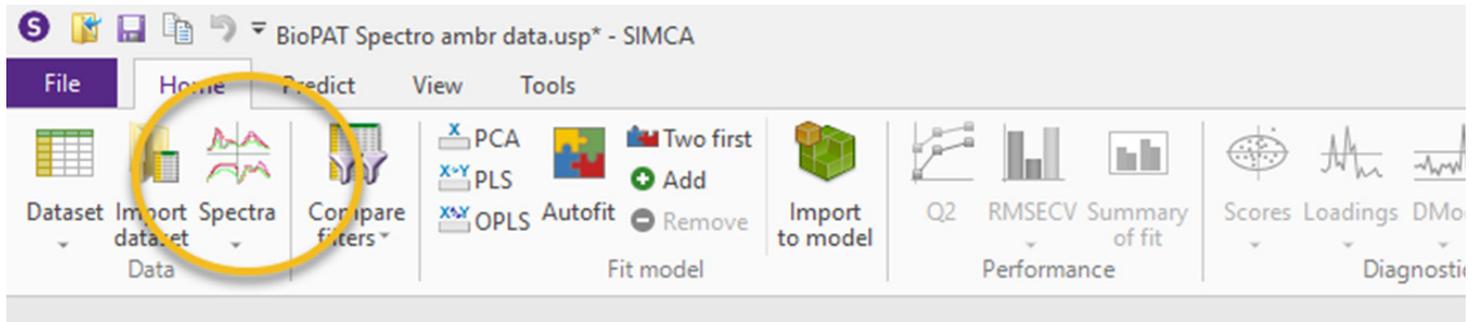
Primary ID	Batch Id	Glucose	Lactate	Glutamine	Ammonia	Glutamate
1	45	2,58	2,807	0,82	8,38	0,053
2	50	4,99576	2,09576	0,678788	6,14303	0,084
3	48	5,959	3,458	0,56	4,326	0,1239
4	51	2,38	4,154	0,55	6,85	0,037
5	51	2,6	3,593	0,47	6,48	0,031
6	45	7,63	5,277	0,69	3,45	0,403
7	45	7,38	6,175	0,67	3,19	0,39
8	44	7,55	6,175	0,72	3,37	0,413
9	51	7,54	6,063	0,67	3,34	0,43
10	51	11,74	4,47745	0,568485	2,97818	0,364848
11	49	8,14	4,94	0,76	3,38	0,409
12	47	7,88	5,389	0,71	3,43	0,428
13	45	5,9	6,961	0,45	3,62	0,362
14	51	6,13	6,512	0,46	3,54	0,36
15	46	6,54	7,747	0,47	3,71	0,408
16	49	6,62	8,42	0,6	3,93	0,396

Raw data overview

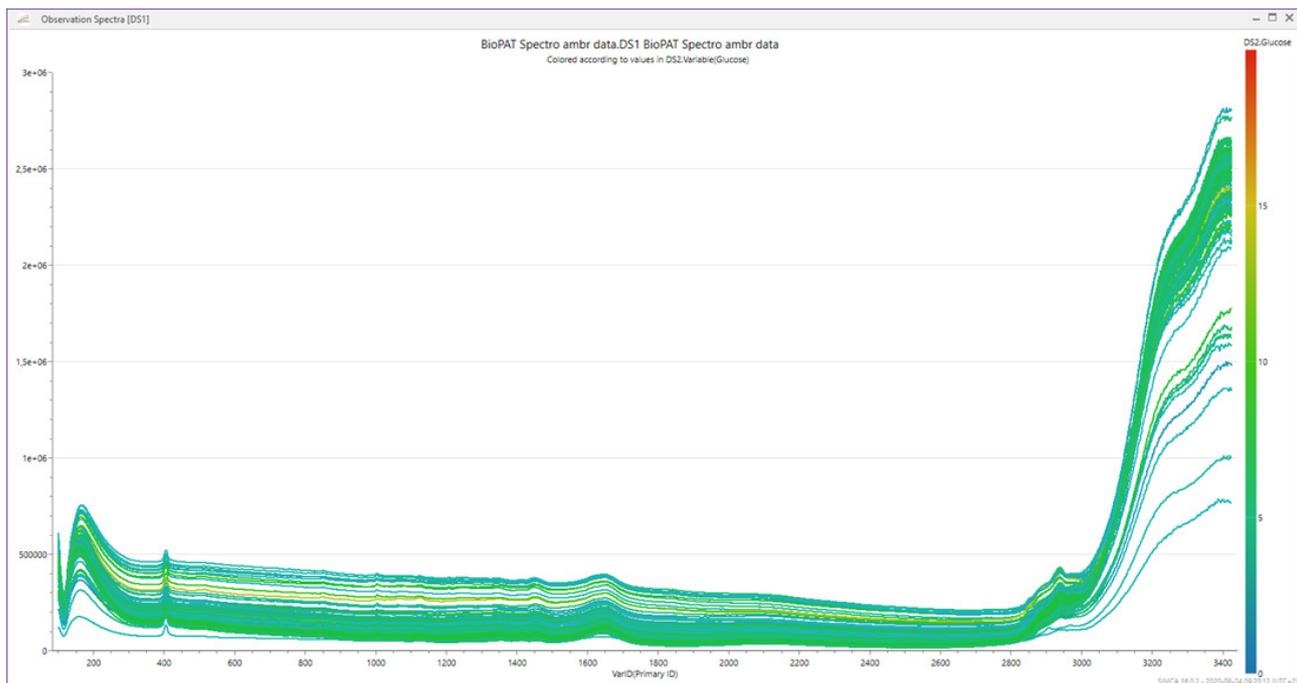
The first step inside SIMCA is to inspect the imported data to verify that it looks as expected and that no outliers or deviating samples are detected in the Raman spectra. This is very important since bad spectral recordings will create bad calibration models. All outliers and deviating samples should be investigated and understood before removing them from the calibration model creation.

To inspect the raw data we start by first looking at the raw spectra then create an overview model of the spectra using Principal Component Analysis (PCA).

Click on the "Spectra" button and a dialog that allows you to select and compare multiple spectral datasets is opened but we only have one at this stage so we just have to click OK.



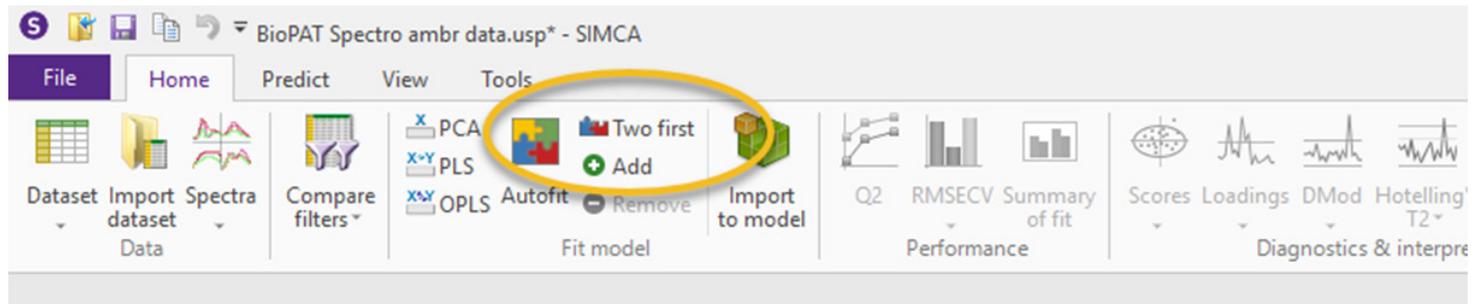
The spectra plot displays all spectra colored by the first Y-variable (Glucose in this example) and in the "Properties" section found in the "Data explorer" pane you can change the coloring vector.



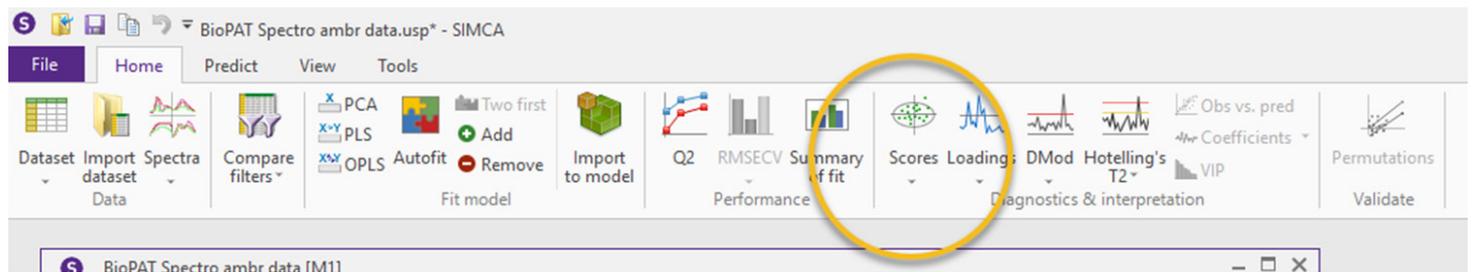
We can see that there are a few spectra that display slightly deviating behavior in the higher shift region (lower right in the plot). To see the identity of a specific spectrum, hover the cursor over the line and a tool tip with information appears. Since this area is related to a water signal, the lower intensities in the deviating spectra could be due to air in the flow cell. Low sample volume in the flow cell results in lower Raman signal and if it is too low it is not useful for building calibration models. There is no strict limit for when a sample should be excluded from the modelling due to a too low signal and has to be considered from case to case.

Outliers in spectral behavior can often also be seen in an overview PCA and this is a way to identify and easily remove the deviating observations from further modelling. A first PCA model of the spectral data (PCA-X) has already been prepared using all samples (observations) and all variables (spectra). This model, highlighted in the project window in the SIMCA workspace provides a first multivariate overview of the available samples to detect unwanted outliers, trends and clusters among the samples.

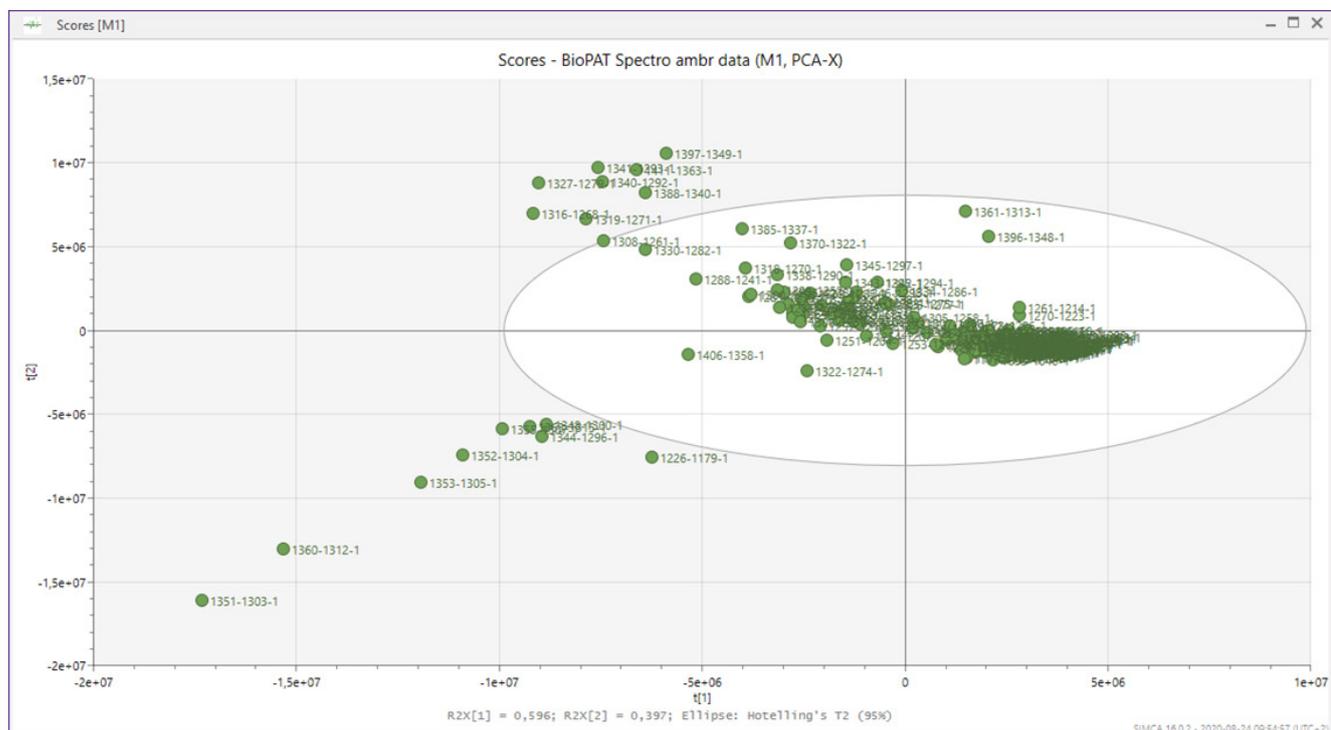
Fit the PCA model on the prepared workset (M1) by clicking on the "Two first" button to calculate the first two principal components for the spectral data.



When the model is calculated, click on the Scores button to create a scatter plot of the first model.



In the scores plot, each sample spectra is shown as a point. Look for samples (observation in SIMCA) that are different than the majority of samples. In the displayed score scatter plot, it is clear that a group of observations, in the lower left corner are deviating from the majority of points. These must be investigated and understood before they are removed from the modelling. After investigation (not shown) it is confirmed that the group in the lower left are the same samples as was seen having a low water signal in the raw spectra.



Spectral preprocessing of Raman data from a bioprocess

After inspection of the raw Raman spectra and decision on which samples to use in the calibration model it is time to investigate what type of spectral filtering, or preprocessing, is best suited for the data at hand. This may be different for each new system and analyte but here we show an example of the procedure.

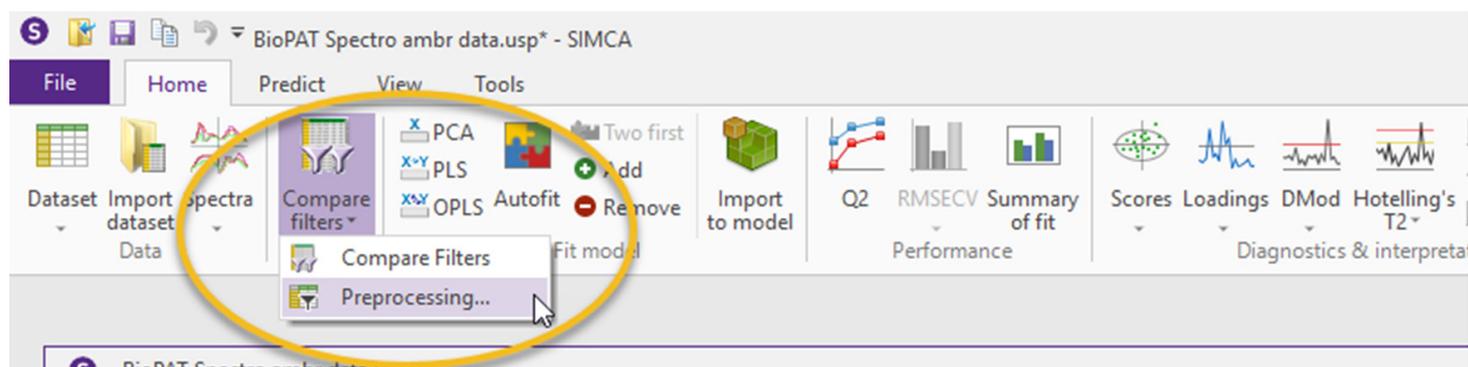
A common challenge with creating multivariate models based on spectroscopic data is that spectra often contains unwanted variation introduced through imperfections in instrument setup or external disturbances. To remove the effect of such irrelevant variation from the spectra and standardize the spectroscopic signal, SIMCA offers several spectral filters including Multiplicative Signal Correction (MSC), Standard Normal Variate (SNV) and 1st, 2nd or 3rd derivatives. Through the use of SIMCA's ability to handle plugins, one can also create custom filters and apply them in the same fashion as the pre-installed methods. In this tutorial we will use a preprocessing plugin created using Python.

For the calibration models using Raman spectra measured in a bioreactor (e.g. ambr or STR) it is good practice to use the size (area) of a water band peak to normalize each spectra before creating the multivariate models. The normalization corrects for different sensitivities in the equipment setup and similar instrument variations. In this tutorial where we model non-spiked Glucose we will use a peak area normalization in combination with Standard normal variate (SNV) and 1st derivative transformation in that order. Other methods and algorithms may be required for other analytes and other data. Some analytes may not require any normalization or preprocessing at all.

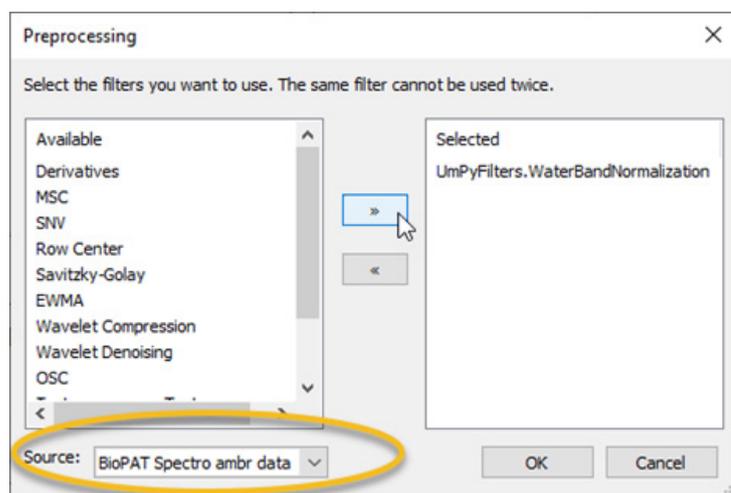
To perform the desired waterband normalization in SIMCA 16 we need to use a Python preprocessing plugin. We will start by applying this normalization and later we will add the additional two preprocessing algorithms (SNV and 1st derivative).

Note! The Peak area normalization algorithm used in the Python plugin may take some time to complete. If you have a few hundred spectra you can expect the preprocessing to take several minutes.

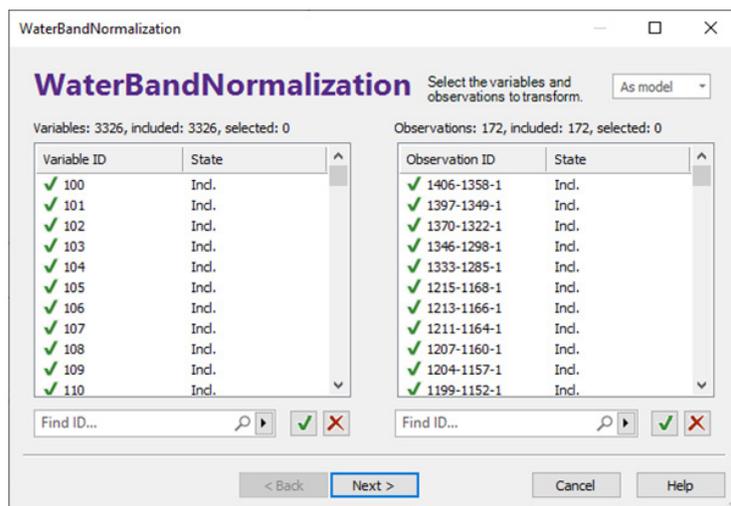
Since we know exactly what preprocessing to apply, we are not interested in comparing filters but only performing a specific preprocessing. To access this dialog, click on the little down arrow on the Compare filters button and select Preprocessing from the drop down menu.



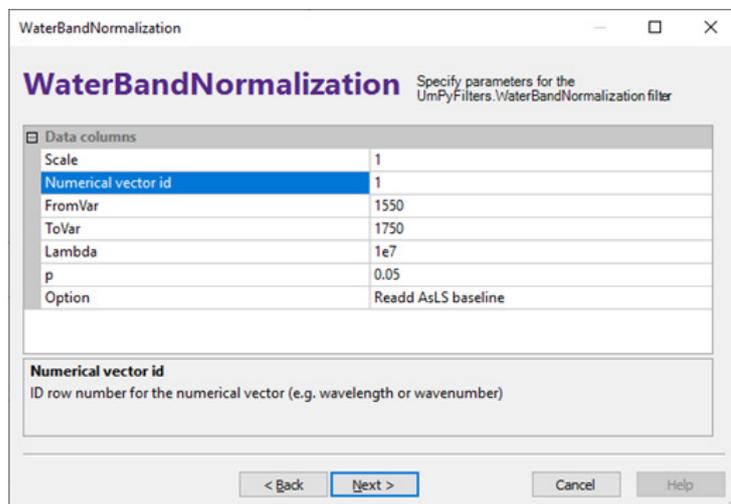
In the dialog that appears, check that the source dataset is your spectral dataset (here called BioPAT Spectro ambr data) and then select and move the UmPyFilters.WaterBandNormalization filter from the left list to the right list of selected filters. Click OK when finished.



Next step is to decide on which variables and observations to include in the preprocessed dataset. We keep all variables and all observations at this stage so we click Next.



The waterband normalization is using the Asymmetric Least Squares (AsLS) algorithm to subtract a baseline before calculating the area over the selected peak. There are several settings available for this filter and they are briefly described here.



Scale: A scaling factor for the resulting spectra. May be useful when comparing filters from different instrument setups. (Here we use 1)

Numerical vector id: The row number in your dataset that contains the numerical variable ID:s (e.g. wavenumbers). The Raman shifts are found in row 1 in this dataset.

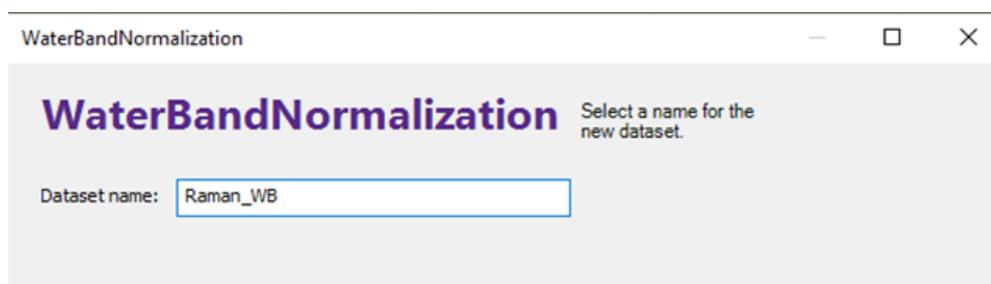
FromVar and ToVar: The Wavenumbers or Raman shifts that frames the waterband peak. By default this is using the range 1550-1750 cm⁻¹.

Lambda: The smoothing factor used in the AsLS algorithm. The default is 1e7

p: The asymmetry factor used in the AsLS algorithm. The default is 0.05

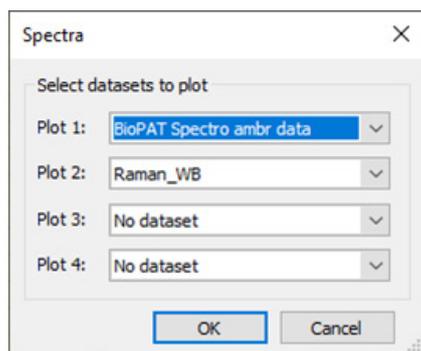
Option: Three options are available: Readd AsLS baseline, With AsLS baseline correction, and Without baseline correction. We recommend that you use the default Readd AsLS baseline.

Last step of the preprocessing is to name the dataset that will be created. A good praxis is to indicate the source and the preprocessing steps in the dataset name.

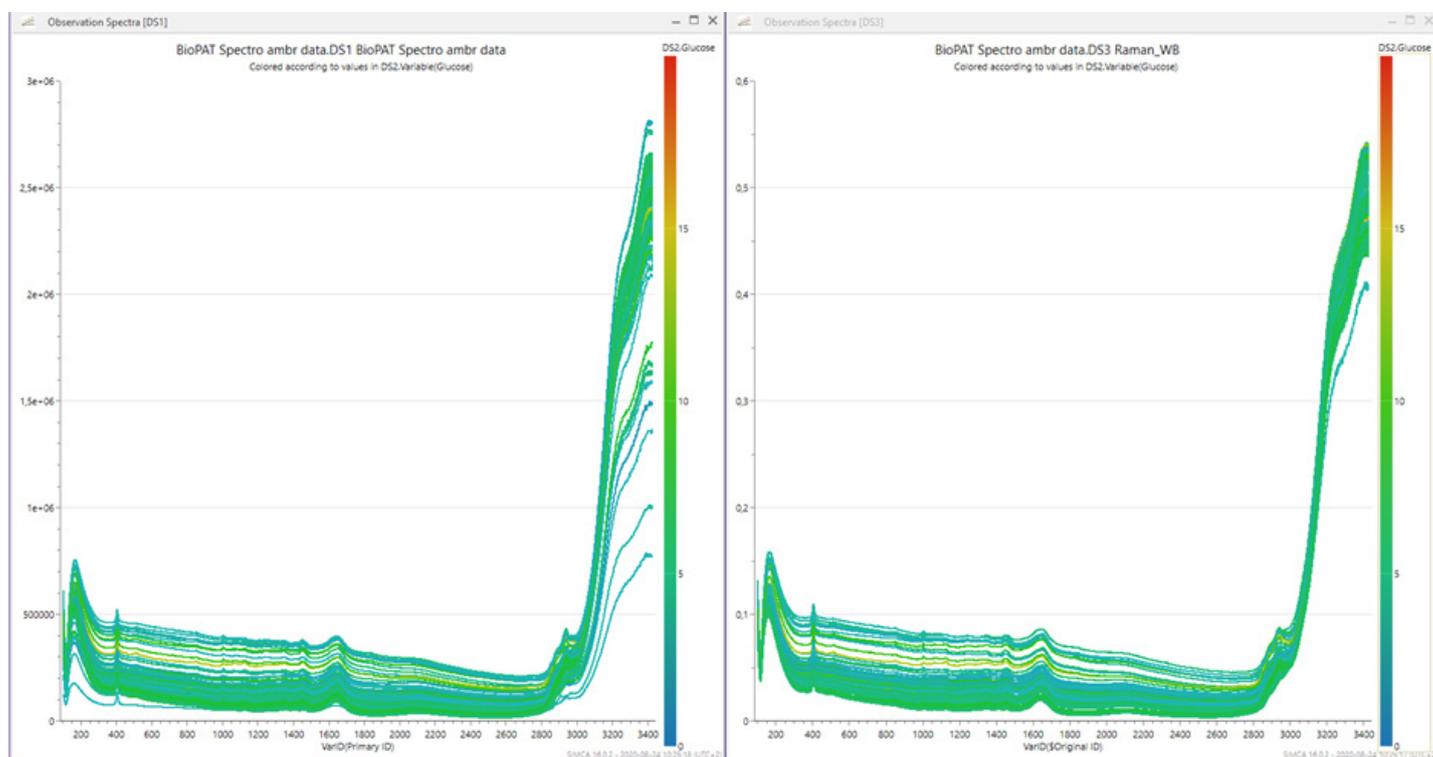


After naming the filtered dataset, you click Finish to complete the normalization

After the AsLS calculations are finished, your new dataset will appear and you can see the result of the preprocessing. Click on the Spectra button in the Home tab and select both the Raw Raman dataset and the filtered dataset (Raman_WB) and click OK.

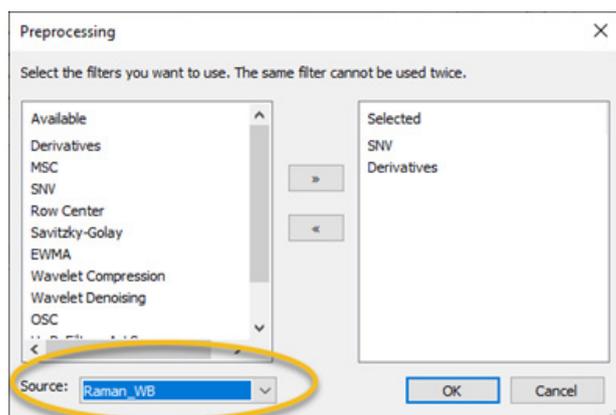


You now get two spectra plots that show the raw spectra to the left and the normalized spectra to the right. We can see that the area normalization adjusts the spectra with low signal in the right hand part of the spectra which makes them fit better into the shape of all samples.

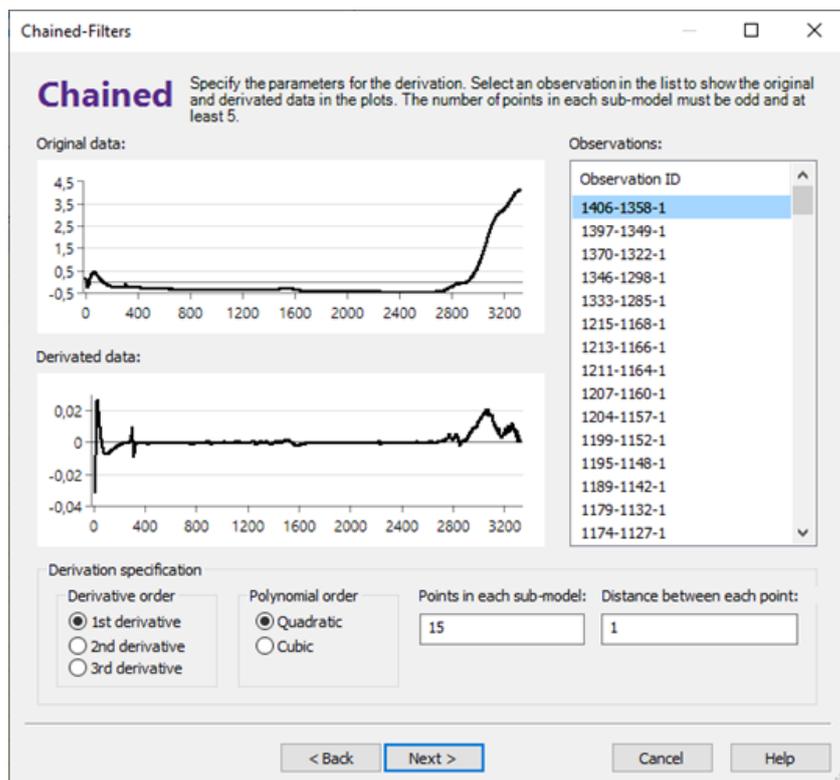


But we can see that there is still a clear shift in the baseline between samples that we want to remove and therefore it was decided to apply two additional filters, SNV and 1st derivative.

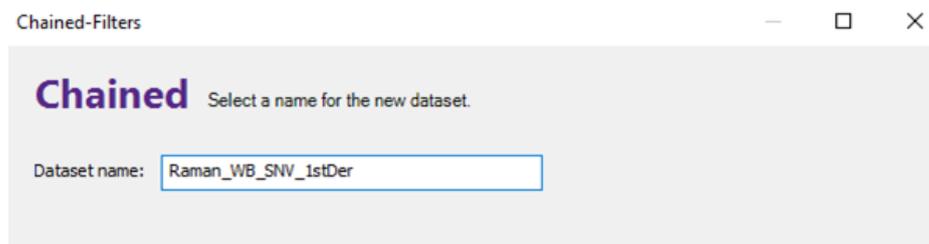
Start the preprocessing wizard in the same way as earlier. Change the source dataset to the waterband normalized dataset and select SNV followed by derivatives. The selected order is the order in which they will be applied. Click OK and on the next page, keep all variables and observations in the dataset.



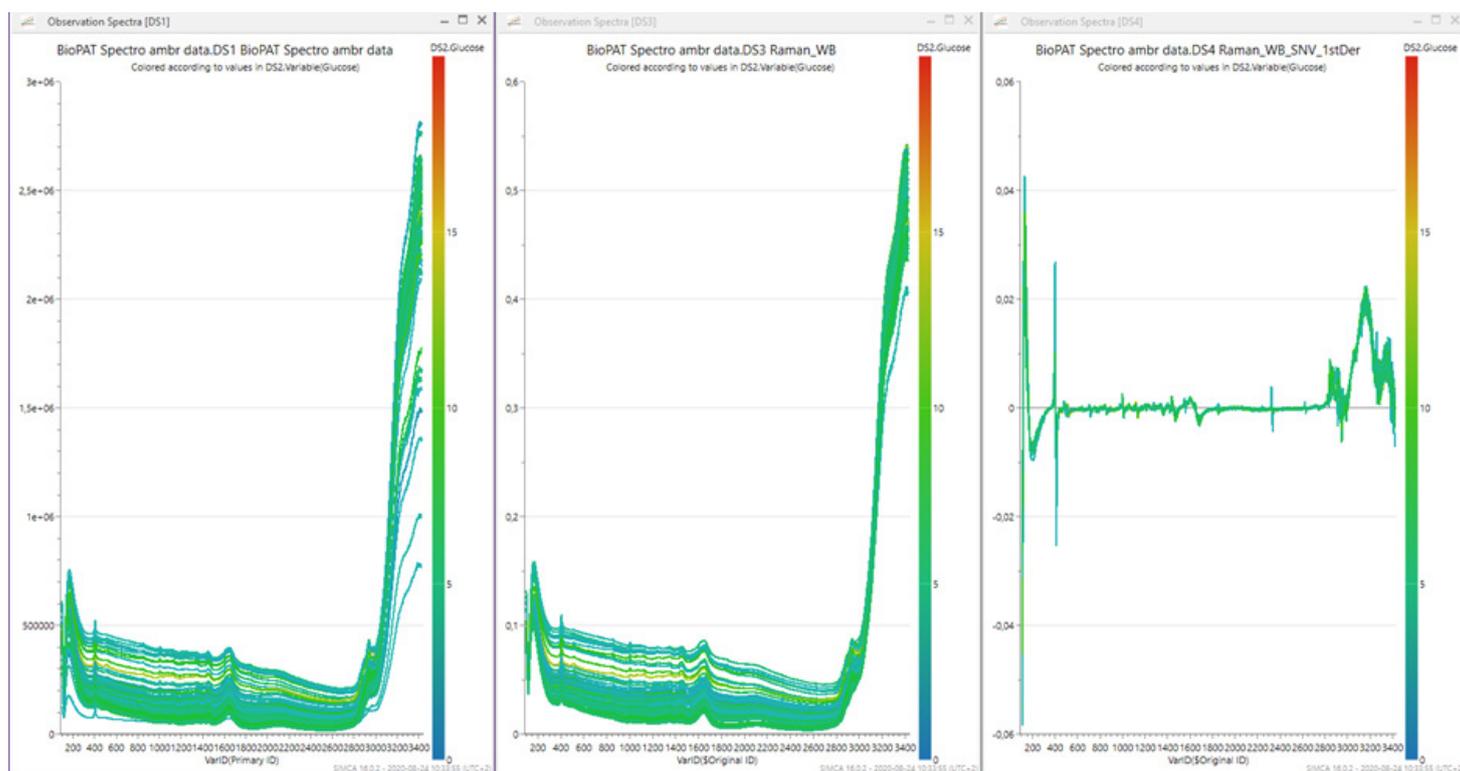
After leaving the variables and observations page, the settings page for Derivatives appear (SNV has no settings). Make sure 1st derivative is selected. Leave all the other settings as they are unless you are experienced in how to fine tune the settings.



Give the new dataset a name indicating source and sequence of filters.



Now you can compare the three different datasets using the spectra plot button: Raw, Waterband normalized, and the dataset where the three filters have been chained. The dataset to the right is the one we will use in the modelling of Glucose.



On the variables page, we start by selecting the spectral range we are going to use. To make it easier to find the correct range, the original variable names can be added to the list through a right click menu option.

Variables included: 3317 (N=3312, Y=5), excluded: 14, selected: 1

Variable ID	\$Original ID	Info	Dataset
Y Glucose			Y-Variables(BioPAT Spectro...
Y Lactate			Y-Variables(BioPAT Spectro...
Y Glutamine			Y-Variables(BioPAT Spectro...
Y Ammonia			Y-Variables(BioPAT Spectro...
Y Glutamate			Y-Variables(BioPAT Spectro...
--- \$1stDer:SNV:100_1	100		Raman_WB_SNV_1stDer
--- \$1stDer:SNV:101_1	101		Raman_WB_SNV_1stDer
--- \$1stDer:SNV:102_1	102		Raman_WB_SNV_1stDer
--- \$1stDer:SNV:103_1	103		Raman_WB_SNV_1stDer
--- \$1stDer:SNV:104_1	104		Raman_WB_SNV_1stDer
--- \$1stDer:SNV:105_1	105		Raman_WB_SNV_1stDer
--- \$1stDer:SNV:106_1	106		Raman_WB_SNV_1stDer
--- \$1stDer:SNV:107_1	107		Raman_WB_SNV_1stDer
X \$1stDer:SNV:108_1	108		Raman_WB_SNV_1stDer
X \$1stDer:SNV:109_1	109		Raman_WB_SNV_1stDer
X \$1stDer:SNV:110_1	110		Raman_WB_SNV_1stDer
X \$1stDer:SNV:111_1	111		Raman_WB_SNV_1stDer
X \$1stDer:SNV:112_1	112		Raman_WB_SNV_1stDer
X \$1stDer:SNV:113_1	113		Raman_WB_SNV_1stDer
X \$1stDer:SNV:114_1	114		Raman_WB_SNV_1stDer
X \$1stDer:SNV:115_1	115		Raman_WB_SNV_1stDer
X \$1stDer:SNV:116_1	116		Raman_WB_SNV_1stDer
X \$1stDer:SNV:117_1	117		Raman_WB_SNV_1stDer
X \$1stDer:SNV:118_1	118		Raman_WB_SNV_1stDer

Quick Info: Variable | \$1stDer:SNV:108_1

Statistics: N 172, Missing values 0 (0%)

Find and highlight the interesting spectral range, in this Glucose calibration example we use the range 450-1800 cm^{-1} . Then click on the arrow next to the Find box below the list and Invert the selection before clicking on the Exclude button.

Variables included: 3317 (N=3312, Y=5), excluded: 14, selected: 1351

Variable ID	\$Original ID	Info	Dataset
X \$1stDer:SNV:1788_1	1788		Raman_WB_SNV_1stDer
X \$1stDer:SNV:1789_1	1789		Raman_WB_SNV_1stDer
X \$1stDer:SNV:1790_1	1790		Raman_WB_SNV_1stDer
X \$1stDer:SNV:1791_1	1791		Raman_WB_SNV_1stDer
X \$1stDer:SNV:1792_1	1792		Raman_WB_SNV_1stDer
X \$1stDer:SNV:1793_1	1793		Raman_WB_SNV_1stDer
X \$1stDer:SNV:1794_1	1794		Raman_WB_SNV_1stDer
X \$1stDer:SNV:1795_1	1795		Raman_WB_SNV_1stDer
X \$1stDer:SNV:1796_1	1796		Raman_WB_SNV_1stDer
X \$1stDer:SNV:1797_1	1797		Raman_WB_SNV_1stDer
X \$1stDer:SNV:1798_1	1798		Raman_WB_SNV_1stDer
X \$1stDer:SNV:1799_1	1799		Raman_WB_SNV_1stDer
X \$1stDer:SNV:1800_1	1800		Raman_WB_SNV_1stDer
X \$1stDer:SNV:1801_1	1801		Raman_WB_SNV_1stDer
X \$1stDer:SNV:1802_1	1802		Raman_WB_SNV_1stDer
X \$1stDer:SNV:1803_1	1803		Raman_WB_SNV_1stDer
X \$1stDer:SNV:1804_1	1804		Raman_WB_SNV_1stDer
X \$1stDer:SNV:1805_1	1805		Raman_WB_SNV_1stDer
X \$1stDer:SNV:1806_1	1806		Raman_WB_SNV_1stDer
X \$1stDer:SNV:1807_1	1807		Raman_WB_SNV_1stDer
X \$1stDer:SNV:1808_1	1808		Raman_WB_SNV_1stDer
X \$1stDer:SNV:1809_1	1809		Raman_WB_SNV_1stDer
X \$1stDer:SNV:1810_1	1810		Raman_WB_SNV_1stDer
X \$1stDer:SNV:1811_1	1811		Raman_WB_SNV_1stDer

Quick Info: Variable | \$1stDer:SNV:450_1, \$1stDer:SNV

Statistics: N 232372, Missing values 0 (0%)

Final step on the variables page is to define the Glucose variable as your target analyte, your Y-variable. Select Glucose in the list and click on the Y button before proceeding to the next page.

Variables included: 1352 (N=1351, Y=1), excluded: 1979, selected: 1

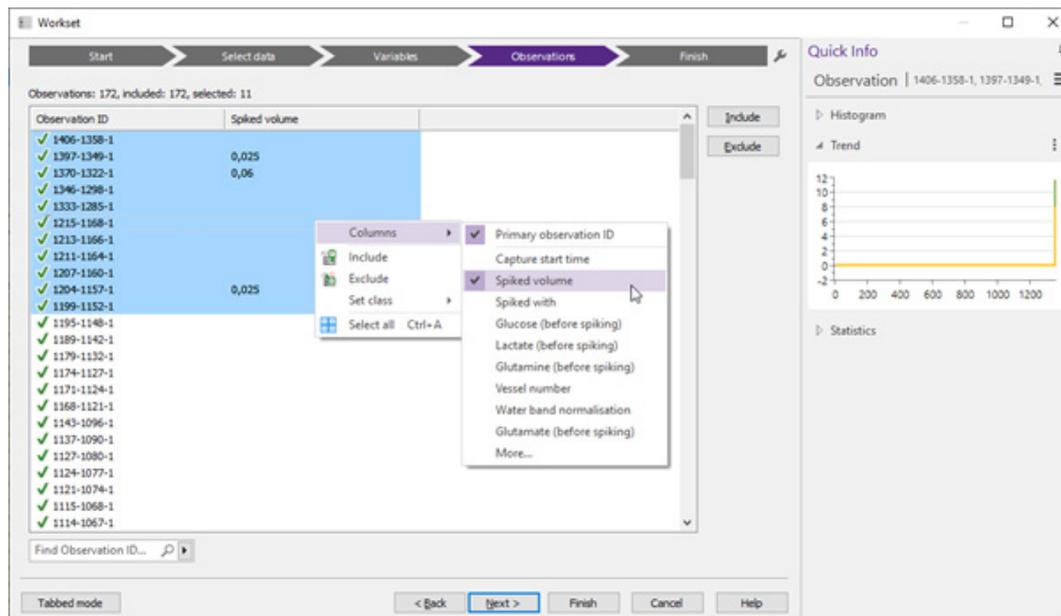
Variable ID	\$Original ID	Info	Dataset
Y Glucose			Y-Variables(BioPAT Spectro...
--- lactate			Y-Variables(BioPAT Spectro...
--- Glutamine			Y-Variables(BioPAT Spectro...
--- Ammonia			Y-Variables(BioPAT Spectro...
--- Glutamate			Y-Variables(BioPAT Spectro...
--- \$1stDer:SNV:100_1	100		Raman_WB_SNV_1stDer
--- \$1stDer:SNV:101_1	101		Raman_WB_SNV_1stDer
--- \$1stDer:SNV:102_1	102		Raman_WB_SNV_1stDer
--- \$1stDer:SNV:103_1	103		Raman_WB_SNV_1stDer
--- \$1stDer:SNV:104_1	104		Raman_WB_SNV_1stDer
--- \$1stDer:SNV:105_1	105		Raman_WB_SNV_1stDer
--- \$1stDer:SNV:106_1	106		Raman_WB_SNV_1stDer
--- \$1stDer:SNV:107_1	107		Raman_WB_SNV_1stDer
--- \$1stDer:SNV:108_1	108		Raman_WB_SNV_1stDer
--- \$1stDer:SNV:109_1	109		Raman_WB_SNV_1stDer
--- \$1stDer:SNV:110_1	110		Raman_WB_SNV_1stDer
--- \$1stDer:SNV:111_1	111		Raman_WB_SNV_1stDer
--- \$1stDer:SNV:112_1	112		Raman_WB_SNV_1stDer
--- \$1stDer:SNV:113_1	113		Raman_WB_SNV_1stDer
--- \$1stDer:SNV:114_1	114		Raman_WB_SNV_1stDer
--- \$1stDer:SNV:115_1	115		Raman_WB_SNV_1stDer
--- \$1stDer:SNV:116_1	116		Raman_WB_SNV_1stDer
--- \$1stDer:SNV:117_1	117		Raman_WB_SNV_1stDer
--- \$1stDer:SNV:118_1	118		Raman_WB_SNV_1stDer

Quick Info: Variable | Glucose

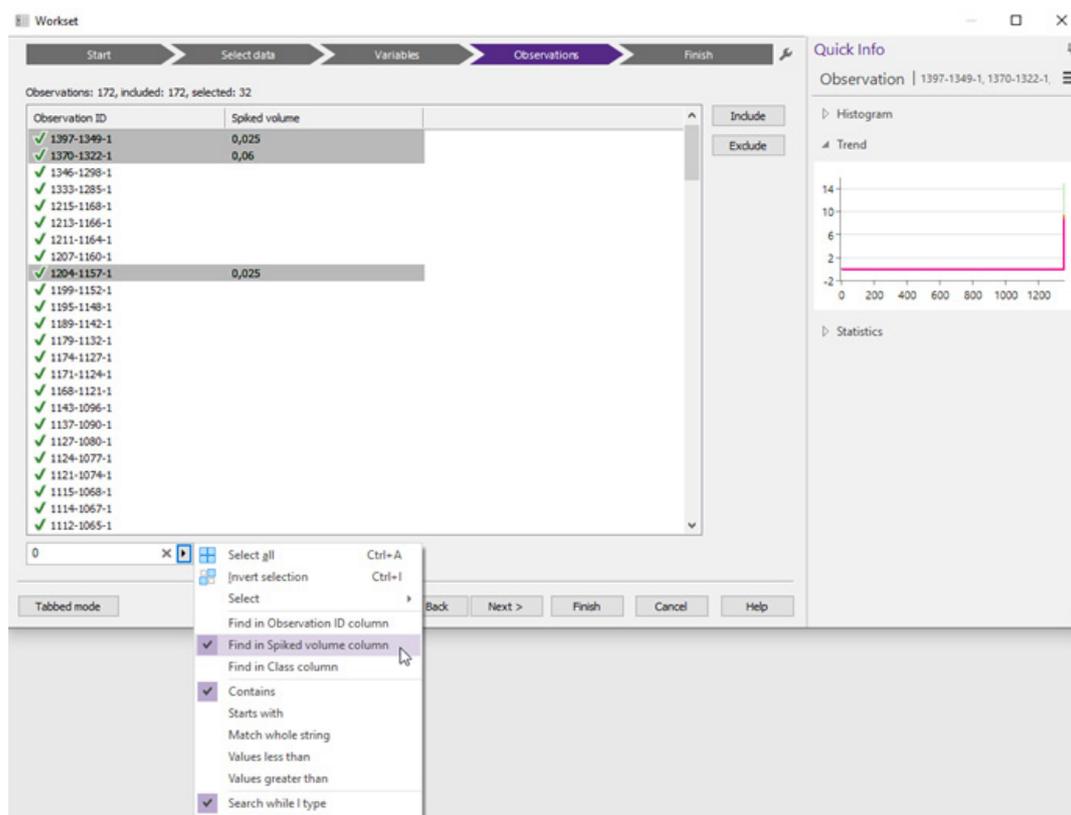
Statistics: N 172, Missing values 0 (0%)

On the Observations page, we should select the samples we want to include in our calibration model. Here we show how to use all samples that has been spiked with Glucose but you select the ones that are relevant for your calibration model.

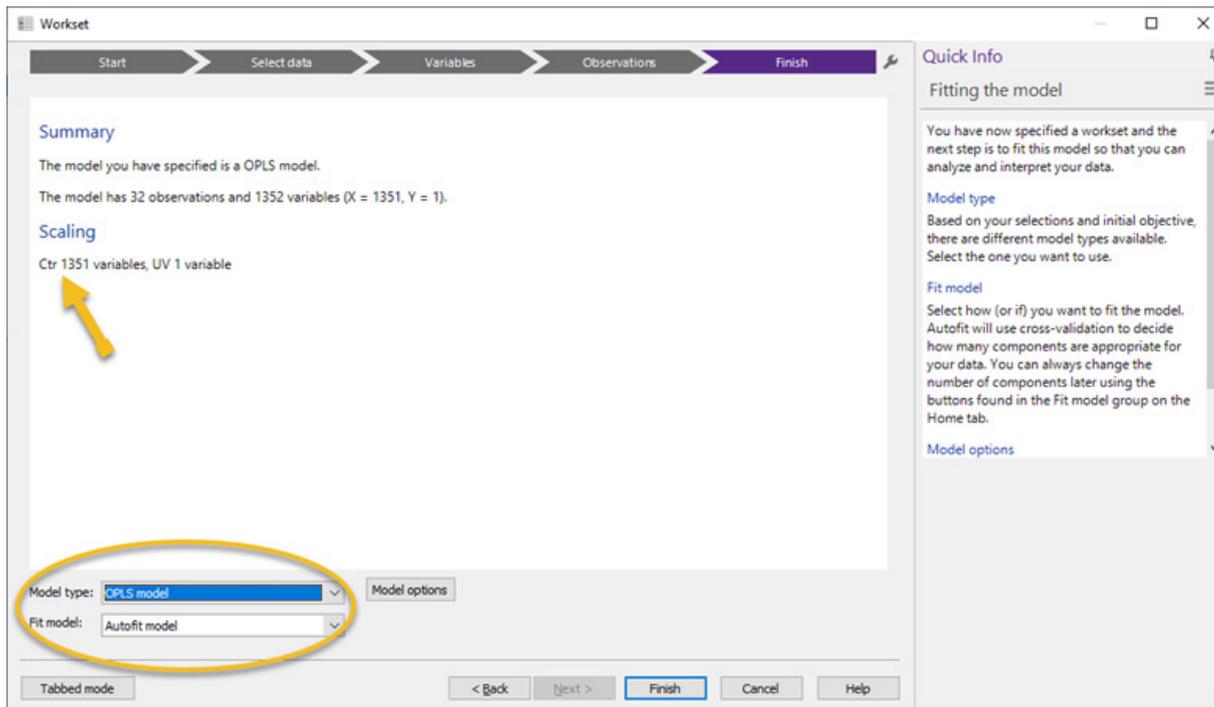
Start by adding the Spiked volume column through the right click menu.



Then use the Find Observations box to locate all observations that have an added volume (contains "0" in the Spiked volume column) and then Invert the selection (from the same menu). Exclude this selection using the Exclude button which leaves all the spiked samples to be used as training set for the model.

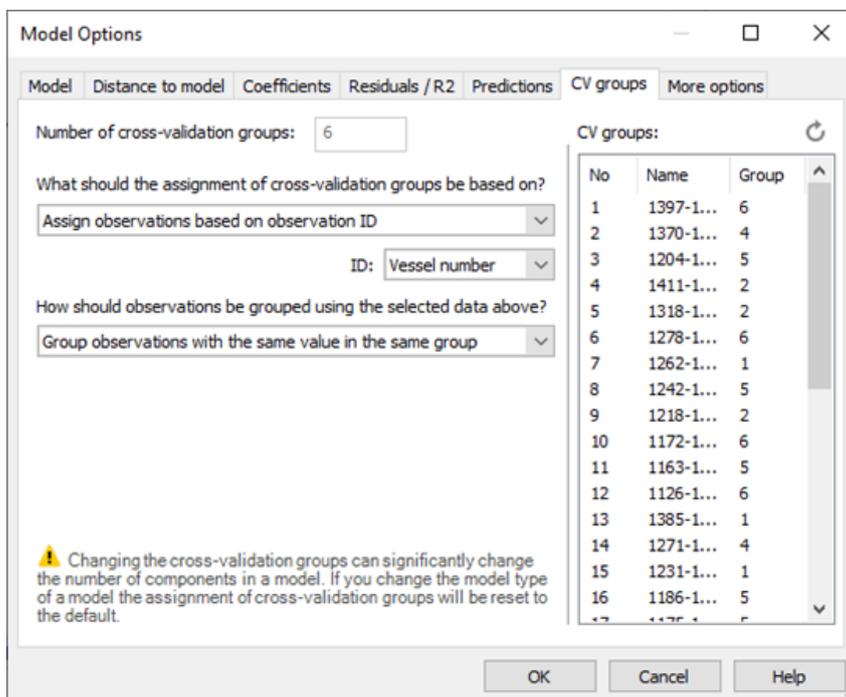


Click Next and move to the last page in the wizard. Here you confirm the model selection (PLS or OPLS) and you decide on the complexity of the model. We select OPLS and use Autofit to decide on the model complexity using cross validation. Note that for spectral data, only centering (Ctr) of the data is used . but UV (Unit variance) scaling is used for the Y-variable. If this is not stated under the scaling section on the Finish summary page, click on the little wrench tool next to the Finish tab up top, open the Scale tab and change the scaling (not shown).



Usually, you do not need to make changes in the Model options dialog but for the samples collected on multiple bioreactors (e.g. an ambr experiment), we want to make sure that the cross validation part of the algorithm (internal validation of the model complexity) takes into account the different vessels. We want all samples collected from a specific vessel to be in the same cross validation group.

Click on the Model options button and proceed to the CV groups tab

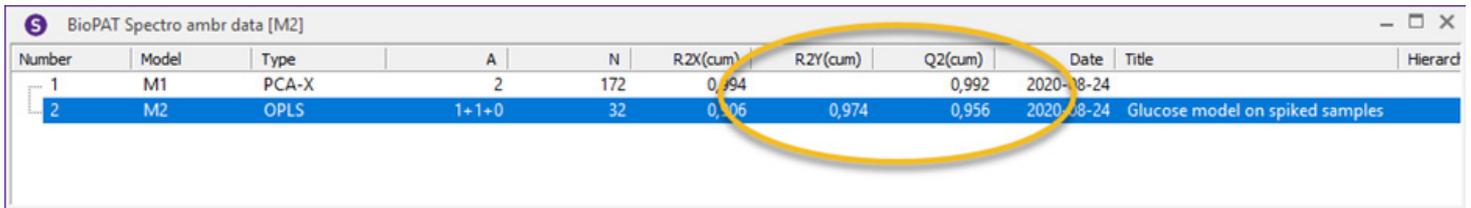


In the CV groups tab, select to Assign observations based on observation ID. Select the Vessel number ID and make sure that you Group observations with the same value in the same group. Click on the refresh symbol over the CV groups list to update the list.

Then click OK and exit the workset wizard by clicking Finish.
A model is now fitted according to cross validation rules decided in SIMCA.

Evaluate the calibration model performance

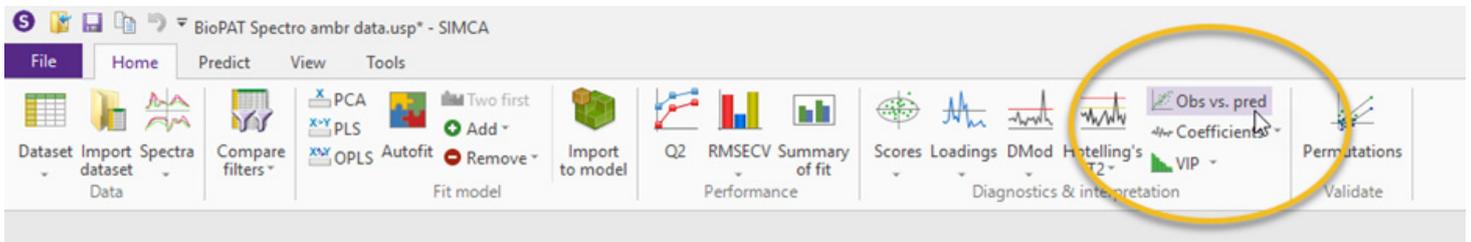
Once a model has been fitted, it needs to be evaluated to see how well it performs. First step is to look at the overall model statistics seen in the Project window. R2Y shows how well the Glucose concentration can be modelled and Q2Y shows how well the Glucose concentrations can be predicted in the cross validation algorithm. R2Y and Q2Y should be as close to 1 as possible. There is no strict cut off but for multivariate calibration models a Q2 higher than 0.95 is expected and desired. In this example we see a 1+1+0 component OPLS model with R2Y of 0.974 and a Q2Y of 0.956 which is OK



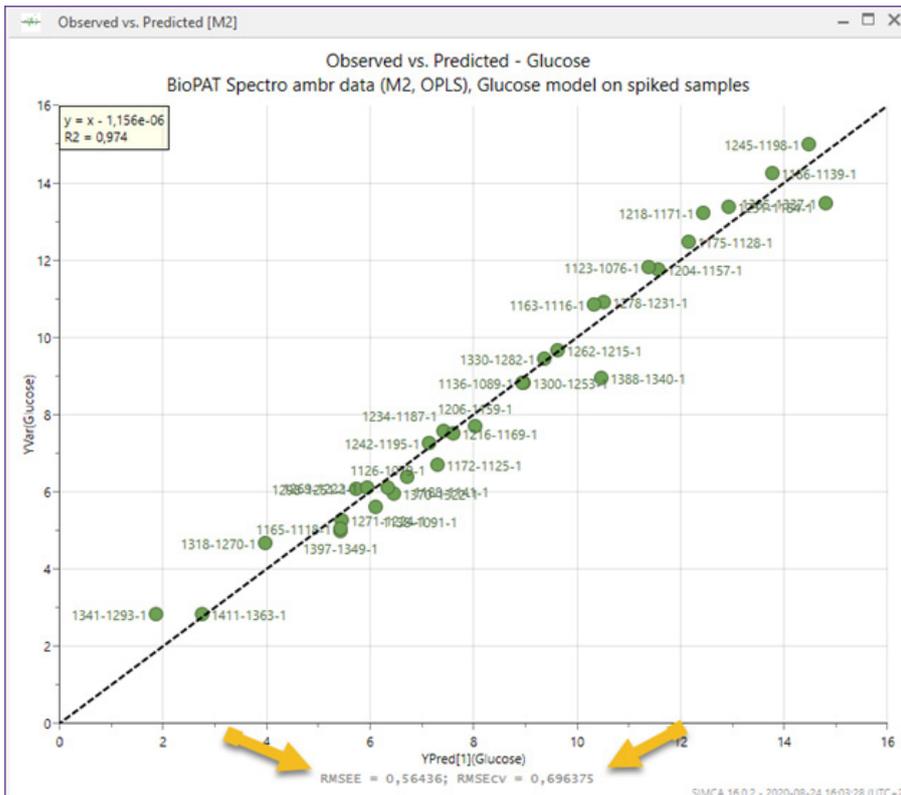
Number	Model	Type	A	N	R2X(cum)	R2Y(cum)	Q2(cum)	Date	Title	Hierard
1	M1	PCA-X	2	172	0.994		0.992	2020-08-24		
2	M2	OPLS	1+1+0	32	0.906	0.974	0.956	2020-08-24	Glucose model on spiked samples	

It is also important at this point to check if there are any outliers in the regression model. This is done by investigating the sample spectral residuals (DModX) and the Hotelling's T2 which shows how far away from model center each observation is. In this example, no extreme outliers were found (not shown).

To get a better understanding on how well the model performs in predicting Glucose concentration, the Observed vs. Predicted plot for Glucose is useful.



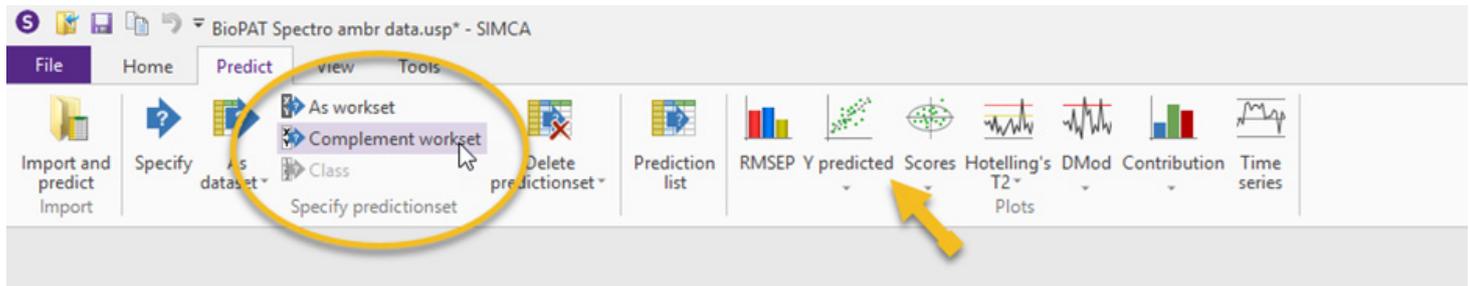
The Obs/pred plot that is displayed shows the predicted and observed values for all the training set observations. At the bottom there are two important model quality numbers: RMSEE and RMSECV (Root Mean Squared Error for Estimates and Cross Validation respectively). RMSECV is the more important one and is an estimate of the average error in concentration units for the predictions during the cross validation algorithm. Here it shows that the average error of the predictions are 0.70 g/l.



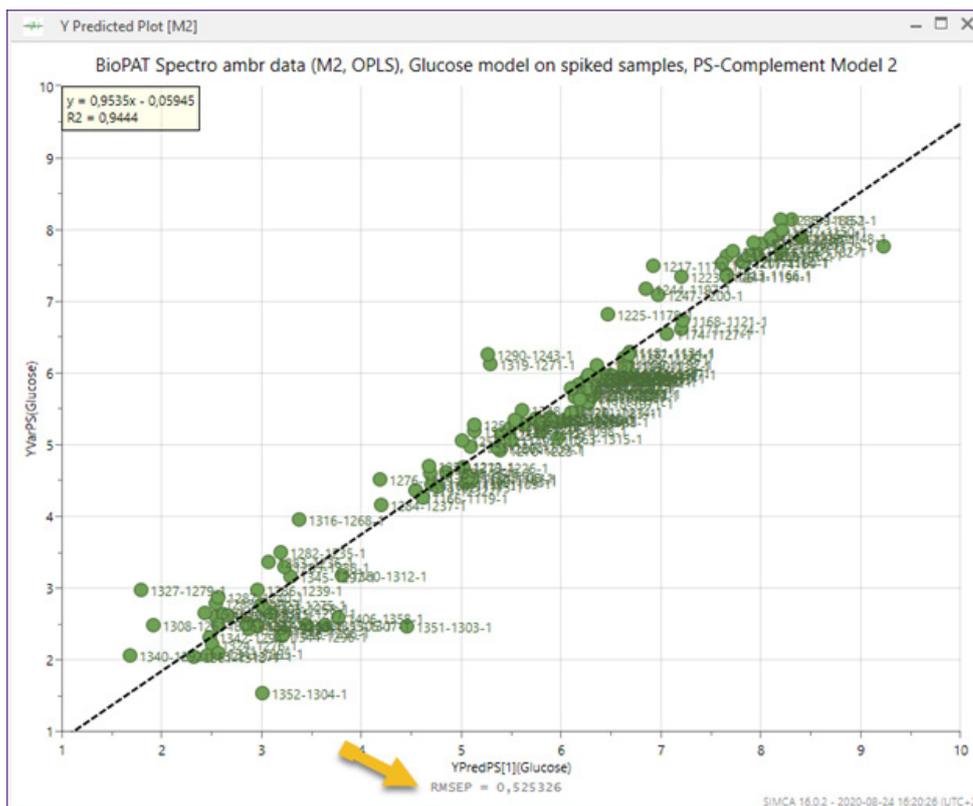
If this is not good enough, the model may have to be refined by e.g. changing the variable range, preprocessing chain or settings, or outlier removal. Apply the changes and fit a new model and compare the results to see if predictive ability improves.

Testing the model

A final measure on the predictive ability of the model is to apply the model to samples that have not been part of the model generation. To test the created model on the excluded samples (i.e. all the non-spiked samples) go to the predict model and click on the Complement workset button and then the Y predicted button to create the Obs/Pred plot for the prediction (or validation) set of observations.



The same preprocessing as was used for the training set observations is now applied to the predictions set (which for AsLS filtering may take some time). At the bottom of the Prediction Obs/Pred plot the average error for the true predictions, RMSEP is displayed. It shows that the average prediction error for the non-spiked data in this dataset is 0.53 g/l for Glucose. The true performance of the prediction model will only be available after deploying the model in the BioPAT Spectro and following the prediction error over a period of time.



This concludes the walk through of how a Glucose calibration model can be created in SIMCA based on data collected using the BioPAT Spectro ambr setup.

Transferring the calibration model to BioPAT Spectro ambr

Once a satisfactory calibration model have been established it needs to be transferred over to the BioPAT Spectro ambr application for execution.

The model, and all the necessary preprocessing, is stored in the SIMCA file (.usp) which can be copied over to the BioPAT Spectro ambr system and the desired calibration model selected. The system is then ready to use the calibration model for prediction of Glucose concentration on new samples.

Sartorius Data Analytics – Change a little. Grow a lot.

We help organizations grow. The Umetrics® Suite of Data Analytics Solutions helps you harness the wealth of data within your organization. Our expertise in data analytics can help you identify vital elements to improve the results of your research, development

and manufacturing processes. With improved process understanding and more consistent product quality, you'll be able to reduce risk, get to market faster, and grow your

business. Our complete solution encompasses software, training, support and project

management. And as part of Sartorius, a global company with more than 9,000 employees, we give you the backing of an international presence.

Experience the benefits for your business today. Find out how our solutions can help your business to grow, whatever industry you are in.

Visit www.sartorius.com/umetrics for details or to download a free 30-day trial.

Sartorius Data Analytics

Phone: +46 40 664 25 80

E-mail: umetrics@sartorius.com

www.sartorius.com/umetrics

Simplifying Progress

