# BioPAT<sup>®</sup> Spectro Ambr<sup>®</sup> calibration in SIMCA<sup>®</sup>

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.

	File Home Da	ta Analyze	Predict View	v Tools	Developer						
	Advisor Favor	rites 🗌 P	Properties	~~		X		¢		e l	
-	Audit trail Item	information 🗌 🕻	Quick info Full	Cascade	Tile Tile		Switch			pectroscopy	
	✓ Data Explorer	ed items 🗌 V	What-If screen		orizontally vertically	Close	windows*	Default	Unics 5	bectroscopy	
		Show			Window				Skins		

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.

Options			×
SIMCA options	General options for working with	SIMCA	
Customize ribbon	Python Editor Command	C:\Program Files\Microsoft VS Code\Code.exe	^
-	Python Editor Command Line Argument	s	
Quick access toolbar	Skins		
Keyboard	Enable Omics skin	Yes	$\sim$
Theme	Enable Spectroscopy skin	Yes	$\sim$
Daiabas	Plot		
Painter	Automatically change active model	No	$\sim$
Restore	EWMA type	Filter	$\sim$
	Plot engine	Direct2D	$\sim$
	Connect data points with missing data	Yes	$\sim$
	Remember coloring	Yes	$\sim$
	Remember cizing	No	~ ~
	Enable Spectroscopy skin		
	A skin designed for spectroscopy analysis.	Skins are visible on the View tab.	
			Reset
		OK Cancel Ap	ply Help

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

€ Back	Start
Start	New
Open other projects	New       Regular project       Batch project       Omics project   Spectroscopy project
Help	Open
Options	Open

#### 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.

Start	New			
Open other projects	Regular project	Batch project	Omics project	Spectroscopy proj
Help	0000			

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.

Text Import Wiz	ard		- 0	$\times$
Choose delimiter				
	miter that separates the columns preview window. Press OK to imp		and watch what	
Delimiters				
() Tab	Semicolon     Comma			
○ Space	Other:			
Preview: Col 1	Col 2	Col 3	Col 4	^
Primary ID	Capture start time	Spiked volume	Spiked with	
1406-1358-1	02/18/2020 12:26:01 +00:00			
1397-1349-1	02/18/2020 03:29:05 +00:00	0,025	7,1 g/L Glucose	
1370-1322-1	02/17/2020 19:16:36 +00:00	0,06	7,1 g/L Glucose	
1346-1298-1	02/16/2020 11:33:38 +00:00			
1333-1285-1	02/15/2020 11:57:24 +00:00			
	02/11/2020 05:36:17 +00:00			
	02/11/2020 05:07:22 +00:00			
1211-1164-1	02/11/2020 04:39:07 +00:00			~
<			>	
	Encoding:	Western Europea	an (ISO-8859-1)	•
		Remember select	ed encoding	
	OK Can	cel App	Hel	p

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.

• • • •							S	IMCA Import -	[BioPAT Spe	ectro ambr data]									-
File	Home Edit View																		۵
				1 <sup>2</sup> 3	A <sup>B</sup> C		XY			<b>ð ð</b>									
d data	Primary variable ID Variable IDs	Primary Seco observation ID observatio Observatio			Qualitative D	ate/Time X-	variable Y-varia Variable roles	ble Exclude		Clear Apph ormatting formatting Auto-formatting	ng* import								
BioPAT Spe	ectro ambr data 🗙 (	Ð																	
Primary	ID Secondary ID	Secondary ID	Secondary ID	Y	Secondary ID	Y	Secondary ID	Y	Secondary	ID Y	Secondary ID	Secondary ID	Y	Secondary	ID Secondary ID	17	18	19	20 🔨
Primary	ID Capture start time	Spiked volum	Spiked with	Glucose	Glucose (befo	Lactate	Lactate (befor	Glutamine	Glutamine	(be Ammonia	Vessel numbe	Water band no	Glutamate	Glutamate	(be Batch Id		100	101	
1406-135	68-1 02/18/2020 12:26:01	-00:00		2,58	2,58		2,807	0,82		8,38	1	12,6051	0,053	0,053	45		385262	371928	357
1397-134	9-1 02/18/2020 03:29:05	00:00 0,025	7,1 g/L Glucos	4,99576	4,62	2,09576	5 2,47	0,678788	0,8	6,14303	6	14,4933	0,084	0,099	50		611468	587288	562
1370-132	2-1 02/17/2020 19:16:36	-00:00 0,06	7,1 g/L Glucos	5,959	5,47	3,458	4,94	0,56	0,8	4,326	4	15,0888	0,1239	0,177	48		508010	487820	466
1346-129	02/16/2020 11:33:38	-00:00		2,38	2,38	4,154	4,154	0,55	0,55	6,85	7	14,6158	0,037	0,037	51		433976	418141	400
1333-128	85-1 02/15/2020 11:57:24	-00:00		2,6	2,6	3,593	3,593	0,47	0,47	6,48	7	15,5371	0,031	0,031	51		432926	417098	399
1215-116	58-1 02/11/2020 05:36:17	-00:00		7,63	7,63	5,277	5,277	0,69	0,69	3,45	1	16,1482	0,403	0,403	45		408922	396173	382
1213-116	6-1 02/11/2020 05:07:22	-00:00		7,38	7,38	6,175	6,175	0,67	0,67	3,19	8	15,8912	0,39	0,39	44		408525	395144	381
1211-116	4-1 02/11/2020 04:39:07	-00:00		7,55	7,55	6,175	6,175	0,72	0,72	3,37	7	15,7977	0,413	0,413	51		402985	390403	376
1207-116	0-1 02/11/2020 03:34:48	-00:00		7,54	7,54	6,063	6,063	0,67	0,67	3,34	7	15,6201	0,43	0,43	51		400640	387294	372
1204-115	7-1 02/11/2020 02:50:18	-00:00 0,025	33,3 g/L Glucc	11,74	7,89	4,47745	5,277	0,568485	0,67	2,97818	5	15,6315	0,364848	0,43	49		399873	385759	371
1199-115	2-1 02/11/2020 01:37:46	-00:00		8,14	8,14	4,94	4,94	0,76	0,76	3,58	3	15,7491	0,409	0,409	47		403630	389967	375
1195-114	48-1 02/11/2020 00:39:51	00:00		7,88	7,88	5,389	5,389	0,71	0,71	3,43	1	15,7561	0,428	0,428	45		402570	389440	375
1189-114	2-1 02/10/2020 14:39:25	-00:00		5,9	5,9	6,961	6,961	0,45	0,45	3,62	7	15,2682	0,362	0,362	51		391694	378615	364
1179-113	32-1 02/10/2020 12:11:46	-00:00		6,13	6,13	6,512	6,512	0,46	0,46	3,54	2	15,3111	0,36	0,36	46		389956	376523	362
1174-112	27-1 02/10/2020 02:14:51	-00:00		6,54	6,54	7,747	7,747	0,47	0,47	3,71	5	14,9265	0,408	0,408	49		376071	362967	349
		-00:00		6,62			8,42	0,6		3,93		15,0401	0,396		50		384871	370333	354

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).

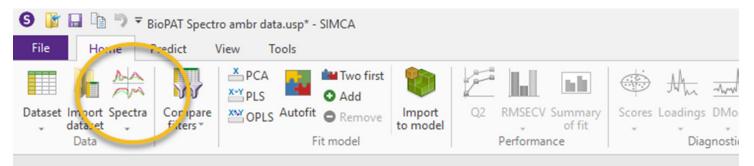
	Home Pr	edict View Tools															
			Add		ΈĿ	<b>b</b> h		the work	M. Alk	Cobs vs. prec Coefficients	14	2					
t Imp	ort Spectra	Compare XX ODIC Autof		Import	Q2 RMSE	CV Summary	Scores Load	dings DMod		w coencients	Permuta						
data		filters *	-	to model	*	of fit			T2*	IN VIP							
Da	ta		Fit model		Perfor	mance		Diagnostics	& interpretati	on	Valid	ate					
<b>3</b> B		ambr data [M1]								_ 0	×						
umber	Model			N R2X(cu	-) D2V(-			Date Title		Hierarchical	_						
umper	Model M1	Type PCA-X	A	N R2X(cu 172	n) R2Y(c	um) Q2(cu	m) [ 2020-0			Hierarchical							
	MI	PCA-X		172			2020-0	18-24									
	Dataset - B	oPAT Spectro ambr data															
																- 0	×
	1	7	8	9		10	C.	11	12	13	14	15	16	17	18	19	×
1	1	7			ormalisation				12		14 102			17 105	18		^
1	1	7 Glutamine (before spiking)														19	^
2	1 Primary ID	7 Glutamine (before spiking) 0,82		Water band n		Glutamate (be		Batch Id	100	101	102	103	104	105	106	19 107 10	^
2	1 Primary ID 1406-1358-1	7 Glutamine (before spiking) 0,82 0,8	Vessel number 1	Water band n 12,6051		Glutamate (be 0,053		Batch Id 45	100 385262	101 371928	102 357572	103 342506	104 327075	105 311558	106 296103	19 107 10 280792	^
2 3 4 5	1 Primary ID 1406-1358-1 1397-1349-1 1370-1322-1 1346-1298-1	7 Glutamine (before spiking) 0,82 0,8 0,8 0,55	Vessel number 1 6	Water band n 12,6051 14,4933 15,0888 14,6158		Glutamate (bet 0,053 0,099		Batch Id 45 50	100 385262 611468	101 371928 587288	102 357572 562680	103 342506 537022	104 327075 510071	105 311558 482359	106 296103 454928	19 107 10 280792 428763	^
2 3 4 5 6	1 Primary ID 1406-1358-1 1397-1349-1 1370-1322-1 1346-1298-1 1333-1285-1	7 Glutamine (before spiking) 0,82 0,8 0,8 0,55 0,47	Vessel number 1 6	Water band n 12,6051 14,4933 15,0888 14,6158 15,5371		Glutamate (bet 0,053 0,099 0,177 0,037 0,031		Batch Id 45 50 48 51 51	100 385262 611468 508010	101 371928 587288 487820 418141 417098	102 357572 562680 466431	103 342506 537022 444340 381002 381596	104 327075 510071 421730	105 311558 482359 398728 340125 342961	106 296103 454928 375703 319896 323450	19         10           280792         428763           353340         300254           304242         304242	^
2 3 4 5 6	1 Primary ID 1406-1358-1 1397-1349-1 1370-1322-1 1346-1298-1	7 Glutamine (before spiking) 0,82 0,8 0,8 0,55 0,47	Vessel number 1 6 4 7	Water band n 12,6051 14,4933 15,0888 14,6158		Glutamate (bel 0,053 0,099 0,177 0,037		Batch Id 45 50 48 51	100 385262 611468 508010 433976	101 371928 587288 487820 418141	102 357572 562680 466431 400385	103 342506 537022 444340 381002	104 327075 510071 421730 360667	105 311558 482359 398728 340125	106 296103 454928 375703 319896	19         10           107         10           280792         428763           353340         300254	^
2 3 4 5 6 7 8	1 Primary ID 1406-1358-1 1397-1349-1 1370-1322-1 1346-1298-1 1333-1285-1 1215-1168-1 1213-1166-1	7 Glutamine (before spiking) 0,82 0,8 0,55 0,47 0,69 0,67	Vessel number 1 6 4 7 7 1 8	Water band n 12,6051 14,4933 15,0888 14,6158 15,5371 16,1482		Glutamate (bet 0,053 0,099 0,177 0,037 0,031 0,403	fore spiking)	Batch Id 45 50 48 51 51	100 385262 611468 508010 433976 432926	101 371928 587288 487820 418141 417098	102 357572 562680 466431 400385 399928	103 342506 537022 444340 381002 381596 367797	104 327075 510071 421730 360667 362464 352321	105 311558 482359 398728 340125 342961	106 296103 454928 375703 319896 323450	19         10           280792         428763           353340         300254           304242         304242	^
2 3 4 5 6 7 8 9	1 Primary ID 1406-1358-1 1397-1349-1 1370-1322-1 1346-1298-1 1333-1285-1 1215-1168-1 1213-1166-1 1211-1164-1	7 Glutamine (before spiking) 0,82 0,8 0,8 0,55 0,47 0,69 0,67 0,72	Vessel number 1 6 4 7 7 1 8 7	Water band n 12,6051 14,4933 15,0888 14,6158 15,5371 16,1482	'-Variables(Bir	Glutamate (bet 0,053 0,099 0,177 0,037 0,031 0,403 0,403	fore spiking) mbr data)	Batch Id 45 50 48 51 51 45	100 385262 611468 508010 433976 432926 408922	101 371928 587288 487820 418141 417098 396173	102 357572 562680 466431 400385 399928	103 342506 537022 444340 381002 381596 367797	104 327075 510071 421730 360667 362464 352321	105 311558 482359 398728 340125 342961 336282	106 296103 454928 375703 319896 323450 319939	19         10           280792         428763           353340         300254           3004242         303567           300943         298671	^
2 3 4 5 6 7 8 9 9	1 Primary ID 1406-1358-1 1397-1349-1 1370-1322-1 1346-1298-1 1333-1285-1 1215-1168-1 1213-1166-1 1211-1164-1 1207-1160-1	7 Glutarnine (before spiking) 0,82 0,8 0,55 0,47 0,69 0,67 0,72 0,67	Vessel number 1 6 4 7 7 1 8 7 7 7	Water band n 12,6051 14,4933 15,0888 14,6158 15,5371 16,1482 Dataset - Y 1	'-Variables(Bi	Glutamate (bet 0,053 0,099 0,177 0,037 0,031 0,403 0PAT Spectro ar 12	fore spiking) mbr data) 13	Batch Id 45 50 48 51 51 45  14	100 385262 611468 508010 433976 432926 408922 15	101 371928 587288 487820 418141 417098 396173 	102 357572 562680 466431 400385 399928	103 342506 537022 444340 381002 381596 367797	104 327075 510071 421730 360667 362464 352321 350551 346978 340380	105 311558 482359 398728 340125 342961 336282 334255 331305 324260	106 296103 454928 375703 319896 323450 319939 317569 315158 308408	19         10           280792         428763           353340         300254           300254         303567           303567         300943           298671         292832	^
2 3 4 5 6 7 8 9 10 11	1 Primary ID 1406-1358-1 1397-1349-1 1370-1322-1 1346-1298-1 1333-1285-1 1215-1168-1 1211-1164-1 1207-1160-1 1204-1157-1	7 Glutamine (before spiking) 0,82 0,8 0,8 0,55 0,47 0,69 0,67 0,67 0,67 0,67	Vessel number 1 6 4 7 7 1 8 7 7 5 1 1	Water band n 12,6051 14,4933 15,0888 14,6158 15,5371 16,1482 Dataset - V 1 Primary ID	'-Variables(Bio 11 Batch Id	Glutamate (bel 0,053 0,099 0,177 0,037 0,031 0,403 DPAT Spectro ar 12 Glucose	fore spiking) mbr data) 13 Lactate	Batch Id 45 50 48 51 51 51 45  14 Glutamine	100 385262 611468 508010 433976 432926 408922 15 Ammonia	101 371928 587288 487820 418141 417098 396173 16 Glutamate	102 357572 562680 466431 400385 399928	103 342506 537022 444340 381002 381596 367797 - □ ×	104 327075 510071 421730 360667 362464 352321 350551 346978 340380 341154	105 311558 482359 398728 340125 342961 336282 334255 331305 324260 325718	106 296103 454928 375703 319896 323450 31939 317569 315158 308408 310062	19         10           280792         428763           428763         33340           300254         304242           303567         300943           296671         292832           294375	^
2 3 4 5 6 7 8 9 10 11 11 12	1 Primary ID 1406-1358-1 1397-1349-1 1370-1322-1 1346-1298-1 1333-1285-1 1215-1168-1 1215-1160-1 1207-1160-1 1204-1157-1 1199-1152-1	7 Glutamine (before spiking) 0,82 0,82 0,55 0,47 0,69 0,67 0,72 0,67 0,67 0,67 0,67	Vessel number 1 6 4 7 7 1 8 7 7 1 8 7 1 3 2	Water band n 12,6051 14,4933 15,0888 14,6158 15,5371 16,1482 Dataset - V 1 Primary ID 1406-1358-1	-Variables(Bin 11 Batch Id 45	Glutamate (bel 0,053 0,099 0,177 0,037 0,037 0,031 0,403 0 PAT Spectro ar 12 Glucose 2,58	fore spiking) mbr data) 13 Lactate 2,807	Batch Id 45 50 48 51 51 45 45 45 45 45 45 45 45 45 45 45 45 45	100 385262 611468 508010 433976 432926 408922 15 Ammonia 8,38	101 371928 587288 487820 418141 417098 396173 16 Glutamate 0,053	102 357572 562680 466431 400385 399928	103 342506 537022 444340 381002 381596 367797 - □ ×	104 327075 510071 421730 360667 362464 352321 350551 346978 340380 341154 344391	105 311558 482359 398728 340125 342961 336282 331305 324260 325718 328219	106 296103 454928 375703 319896 323450 31939 317569 315158 308408 310062 312078	19         107         10           280792         428763         353340           300254         304242         303567           300943         298671         292832           294375         296177         296177	^
2 3 4 5 6 7 8 9 10 11 12 13	1 Primary ID 1406-1358-1 1397-1349-1 1370-1322-1 1336-1298-1 1333-1285-1 1215-1168-1 1213-1166-1 1211-1164-1 1207-1160-1 1204-1157-1 1195-1148-1	7 Glutamine (before spiking) 0,82 0,8 0,8 0,55 0,47 0,69 0,67 0,72 0,67 0,67 0,67 0,67 0,67 0,67 0,67 0,67	Vessel number 1 6 4 7 7 1 8 7 7 5 1 3 2 1 3 2 1 3	Water band n 12,6051 14,4933 15,0888 14,6158 15,5371 16,1482 Dataset - 1 1406-1358-1 1397-1349-1	-Variables(Bin 11 Batch Id 45 50	Glutamate (bel 0,053 0,099 0,177 0,037 0,031 0,403 0 PAT Spectro ar 12 Glucose 2,58 4,99576	fore spiking) mbr data) 13 Lactate 2,807 2,09576	Batch Id 45 50 48 51 51 45 45 45 45 45 45 0, 82 0, 678788	100 385262 611468 508010 433976 432926 408922 15 Ammonia 8, 38 6, 14303	101 371928 587288 487820 418141 417098 396173 16 Glutamate 0,053 0,084	102 357572 562680 466431 400385 399928	103 342506 537022 444340 381002 381596 367797 - □ ×	104 327075 510071 421730 360667 362464 352321 346978 346978 341154 344391 343834	105 311558 482359 398728 340125 342961 336282 334255 331305 324260 325718 328219 327528	106 296103 454928 375703 319896 323450 319939 317569 315158 308408 310062 312078 311168	19         107         10           280792         428763         353340           303254         304242         303567           300943         298671         292832           294375         296177         295006	^
2 3 4 5 6 7 8 9 10 11 12 13 14	1 Primary ID 1406-1358-1 1397-1349-1 1370-1322-1 1346-1298-1 1333-1285-1 1213-1166-1 1211-1164-1 1207-1160-1 1204-1157-1 1199-1152-1 1195-1148-1 1189-1142-1	7 Glutamine (before spiking) 0,82 0,8 0,8 0,55 0,47 0,66 0,67 0,72 0,67 0,67 0,67 0,76 0,67 0,76 0,71 0,45	Vessel number 1 6 4 7 7 1 8 7 7 5 1 3 2 2 1 3 7 4 3 7 4	Water band n 12,6051 14,4933 15,0888 14,6158 15,5371 16,1482 Dataset - V 1 <b>IPrimary ID</b> 1406-1358-1 1397-1349-1 1370-1322-1	-Variables(Bir 11 Batch Id 45 50 48	Glutamate (be) 0,053 0,099 0,177 0,037 0,031 0,403 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	fore spiking) mbr data) 13 Lactate 2, 807 2, 09576 3, 458	Batch Id 45 50 48 51 51 45  6 <b>Glutamine</b> 0, 82 0, 678788 0, 56	100 385262 611468 508010 433976 432926 408922 15 Ammonia 8,38 6,14303 4,326	101 371928 587288 487820 418141 396173 396173 16 Glutamate 0,053 0,084 0,1239	102 357572 562680 466431 400385 399928	103 342506 537022 444340 381002 381596 367797 - □ ×	104 327075 510071 421730 360667 362464 352321 350551 346978 340380 341154 343834 343834 335010	105 311558 482359 398728 340125 342961 336282 334255 331305 324260 325718 328219 327528 319461	106 296103 454928 375703 319896 323450 319939 315158 308408 310062 312078 311168 303579	19         10           280792         10           428763         353340           300254         303254           300943         296671           29294375         296177           295006         287573	^
2 3 4 5 6 7 8 9 10 11 12 13 14 15	1 Primary ID 1406-1358-1 1397-1349-1 1370-1322-1 1346-1298-1 1215-1168-1 1215-1168-1 1215-1168-1 1207-1160-1 1207-1160-1 1204-1157-1 1199-1152-1 1189-1142-1 1189-1142-1 1179-1132-1	7 Glutarmine (before spiking) 0,82 0,8 0,8 0,55 0,47 0,69 0,67 0,72 0,67 0,67 0,67 0,72 0,67 0,72 0,67 0,72 0,67 0,72 0,67 0,72 0,67 0,72 0,65 0,69 0,62 0,62 0,62 0,62 0,62 0,62 0,62 0,62	Vessel number 1 6 4 7 7 7 7 7 7 7 7 7 1 3 2 1 3 2 1 3 2 1 3 2 5 1 3 2 5 1 3 2 5 5 1 3 2 5 5 5 5 5 5 5 5 5 5 5 5 5	Water band n 12,6051 14,4933 15,0888 14,6158 15,5371 16,1482 1 Primary ID 1406-1358-1 1370-1322-1 1370-1322-1 1346-1298-1	-Variables(Bin 11 Batch Id 45 50 48 51	Glutamate (bet 0,053 0,099 0,177 0,037 0,031 0,403 0,403 0,403 0,403 0,403 0,403 0,403 0,403 0,403 0,403 0,403 0,403 0,403 0,403 0,403 0,403 0,403 0,403 0,403 0,50 0,50 0,50 0,50 0,50 0,50 0,50 0,	fore spiking) mbr data) 13 Lactate 2,807 2,09576 3,458 4,154	Batch Id 45 50 48 51 51 45  14 <b>Glutamine</b> 0,82 0,678788 0,56 0,55	100 385262 611468 508010 433976 432926 408922 15 Ammonia 8,38 6,14303 4,326 6,85	101 371928 587288 487820 418141 417098 396173 	102 357572 562680 466431 400385 399928	103 342506 537022 444340 381002 381596 367797 - □ ×	104 327075 510071 421730 360667 362464 352321 346978 34050 341154 344391 344391 343834 335010 331773	105 311558 482359 398728 340125 342961 336282 334265 324260 325718 328219 327528 319461 316082	106 296103 454928 375703 319896 323450 319939 317569 315158 308408 310062 312078 311168 303579 300197	19         107         10           280792         428763         353340           303254         303567         303567           303943         298671         292832           294375         296177         295006           297573         28425         32424	^
2 3 4 5 6 7 8 9 10 11 12 13 14 15 16	1 Primary ID 1406-1358-1 1397-1349-1 1370-1322-1 1346-1298-1 1333-1285-1 1213-1166-1 1211-1164-1 1207-1160-1 1204-1157-1 1199-1152-1 1195-1148-1 1189-1142-1	7 Glutamine (before spiking) 0,82 0,8 0,8 0,55 0,47 0,69 0,67 0,67 0,67 0,67 0,67 0,67 0,76 0,67 0,76 0,71 0,45 0,46 0,47	Vessel number 1 6 4 7 7 1 8 7 7 5 1 3 2 2 1 3 7 4 3 7 4	Water band n 12,6051 14,4933 15,0888 14,6158 15,5371 16,1482 16,1482 10 1406-1358-1 1397-1349-1 1370-1322-1 1346-1298-1	-Variables(Bio 11 Batch Id 45 50 48 51 51	Glutamate (be) 0,053 0,099 0,177 0,037 0,031 0,403 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	fore spiking) mbr data) 13 Lactate 2, 807 2, 09576 3, 458	Batch Id 45 50 48 51 51 45  6 <b>Glutamine</b> 0, 82 0, 678788 0, 56	100 385262 611468 508010 433976 432926 408922 15 Ammonia 8,38 6,14303 4,326	101 371928 587288 487820 418141 396173 396173 16 Glutamate 0,053 0,084 0,1239	102 357572 562680 466431 400385 399928	103 342506 537022 444340 381002 381596 367797 - □ ×	104 327075 510071 421730 360667 362464 352321 350551 346978 340380 341154 343834 343834 335010	105 311558 482359 398728 340125 342961 336282 334255 331305 324260 325718 328219 327528 319461	106 296103 454928 375703 319896 323450 319939 315158 308408 310062 312078 311168 303579	19         10           280792         10           428763         353340           300254         303254           300943         296671           29294375         296177           295006         287573	^

### Raw data overview

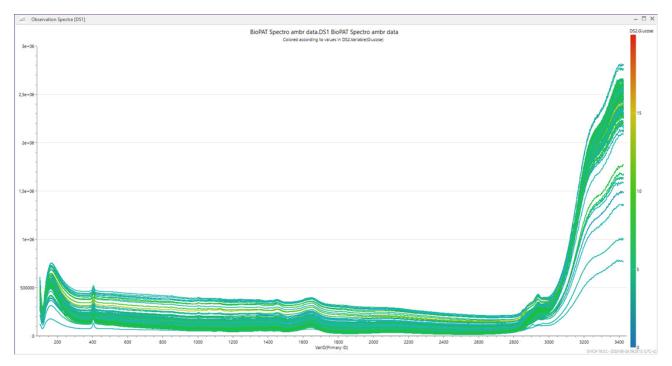
The first step inside SIMCA is to inspect the imported data to verify that it looks as expected and that no outliers or deviating samples are detected in the Raman spectra. This is very important since bad spectral recordings will create bad calibration models. All outliers and deviating samples should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



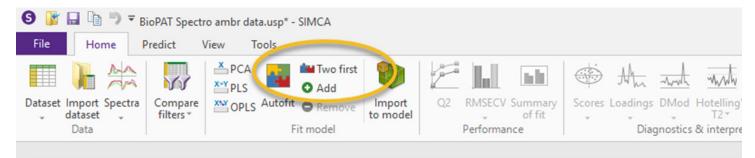
The spectra plot displays all spectra colored by the first Y-variable (Glucose in this example) and in the "Properties" section found in the "Data explorer" pane you can change the coloring vector.



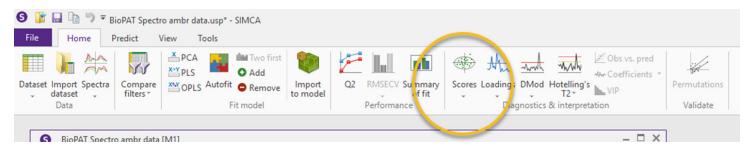
We can see that there are a few spectra that display slightly deviating behavior in the higher shift region (lower right in the plot). To see the identity of a specific spectrum, hover the cursor over the line and a tool tip with information appears. Since this area is related to a water signal, the 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.

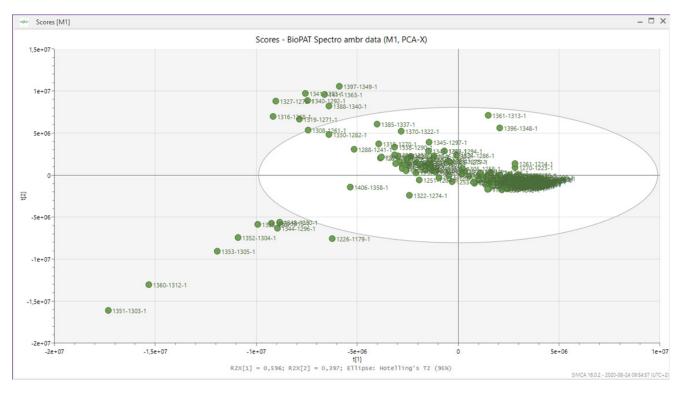
Fit the PCA model on the prepared workset (M1) by clicking on the "Two first" button to calculate the first two principal components for the spectral data.



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.

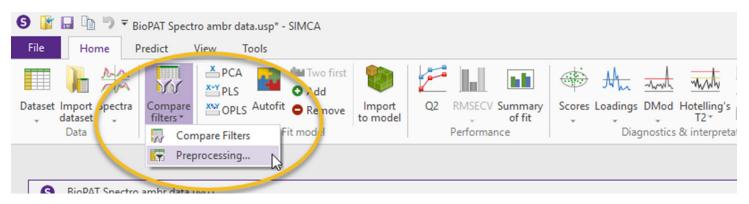
A common challenge with creating multivariate models based on spectroscopic data is that spectra often contains unwanted variation introduced through imperfections in instrument setup or external disturbances. To remove the effect of such irrelevant variation from the spectra and standardize the spectroscopic 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 1<sup>st</sup> 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.

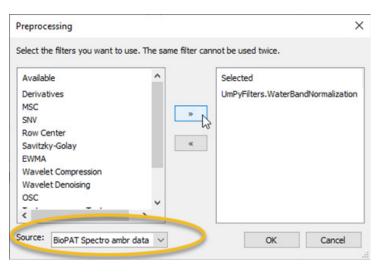
To perform the desired waterband normalization in SIMCA 16 we need to use a Python preprocessing plugin. We will start by applying this normalization and later we will add the additional two preprocessing algorithms (SNV and 1st derivative).

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.

	andNorn		observations to ti		
Variables: 3326, inc	luded: 3326, selected	:0	Observations: 172, ind Observation ID	State	:ted: 0
✓ 100	Incl.		✓ 1406-1358-1	Ind.	
V 101	Ind.		✓ 1397-1349-1	Ind.	
V 102	Ind.		√ 1370-1322-1	Ind.	
✓ 103	Ind.		√ 1346-1298-1	Ind.	
✓ 104	Ind.		✓ 1333-1285-1	Ind.	
✓ 105	Incl.		✓ 1215-1168-1	Ind.	
✓ 106	Ind.		✓ 1213-1166-1	Ind.	
✓ 107	Ind.		√ 1211-1164-1	Ind.	
✓ 108	Ind.		✓ 1207-1160-1	Ind.	
✓ 109	Ind.		✓ 1204-1157-1	Ind.	
✓ 110	Ind.	~	✓ 1199-1152-1	Ind.	~
Find ID	• ۵	✓ ×	Find ID	ام	• v ×

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.

Data columns	
Scale	1
Numerical vector id	1
FromVar	1550
ToVar	1750
Lambda	1e7
p	0.05
Option	Readd AsLS baseline
umerical vector id row number for the numerical	rector (e.g. wavelength or wavenumber)

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

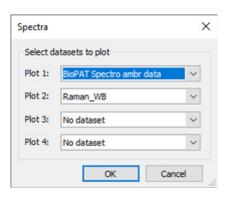
Option: Three options are available: Readd AsLS baseline, With AsLS baseline correction, and Without baseline correction. We recommend that you use the default Readd AsLS baseline.

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.

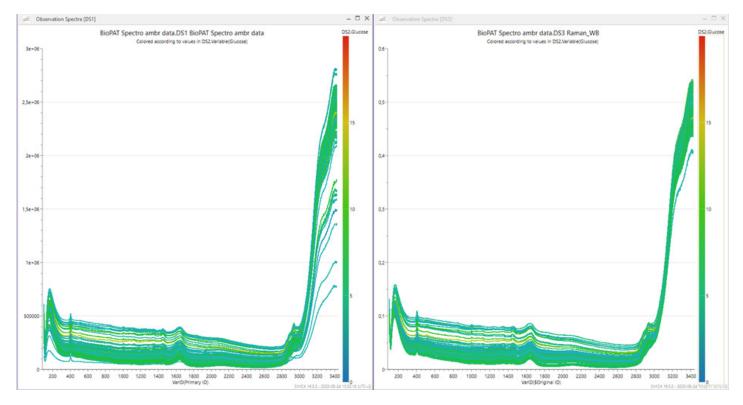
WaterBandNorma	alization		_	$\times$
Waterl	BandNormalization	Select a name for the new dataset.		
Dataset name:	Raman_WB			

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 1<sup>st</sup> 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.

elect the filters you want to u	use. The	ame filter ca	nnot be used twice.	
Available	,	1	Selected	
Derivatives			SNV	
MSC			Derivatives	
SNV		*		
Row Center				
Savitzky-Golay		*		
EWMA				
Wavelet Compression				
Wavelet Denoising				
OSC				
	-			

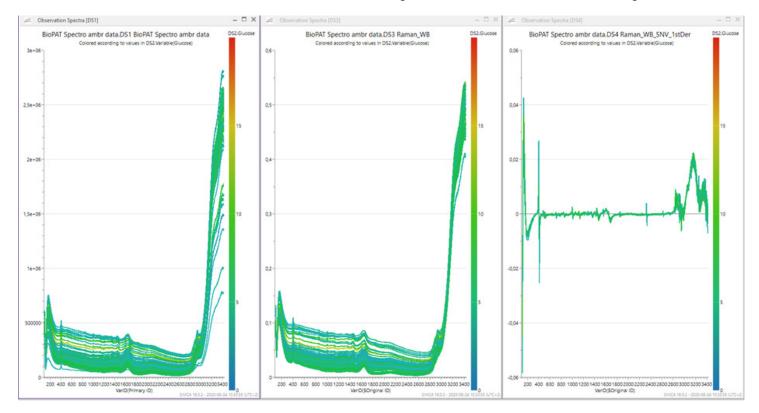
After leaving the variables and observations page, the settings page for Derivatives appear (SNV has no settings). Make sure 1<sup>st</sup> derivative is selected. Leave all the other settings as they are unless you are experienced in how to fine tune the settings.

Chained-Filters		$\Box$ ×
Chained Specify the parameters for the derivation. Select an ob- and derivated data in the plots. The number of points in least 5.	ervation in the list to show each sub-model must be	the original odd and at
Original data:	Observations:	
4,5 ]	Observation ID	^
3,5	1406-1358-1	
2,5	1397-1349-1	
1,5-	1370-1322-1	
0,5	1346-1298-1	
0 400 800 1200 1600 2000 2400 2800 3200	1333-1285-1	
0 400 000 1200 1000 2000 2000 5200	1215-1168-1	
Derivated data:	1213-1166-1	
	1211-1164-1	
0.02	1207-1160-1	
	1204-1157-1	
off the second s	1199-1152-1	
-0,02 -	1195-1148-1 1189-1142-1	
-0.04	1179-1132-1	
0 400 800 1200 1600 2000 2400 2800 3200	1179-1132-1	~
	11/4-112/-1	¥
Derivation specification     Polynomial order     Points in each sub-model       Ist derivative     Ist derivative     Ist derivative       3rd derivative     Cubic     Ist derivative	del: Distance between e	ach point:
< Back Next >	Cancel	Help

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

Chained-Filters	—	$\times$
Chained Select a name for the new dataset.		
Dataset name: Raman_WB_SNV_1stDer		

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.



### Create the calibration model

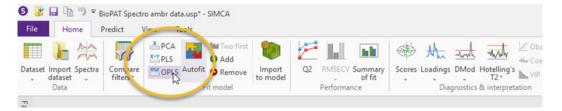
Once a preprocessing chain has been decided, it is time to build the calibration model.

Besides the preprocessing, we also need to decide on the spectral range to use in the model. If no prior knowledge exist, use the full range available but in this case we know from earlier investigations that Glucose has a Raman signal between 450 and 1800 cm-1. The best range is of course different for each specific analyte and need to be decided from prior knowledge. Prior knowledge can be based on literature results or generated in-hose through an analysis of a well-designed (DOE) set of analyte mixtures. Loadings from multivariate models can be used as a tool to find the best variable ranges. The range used in this investigation was found relevant for Glucose during laboratory tests.

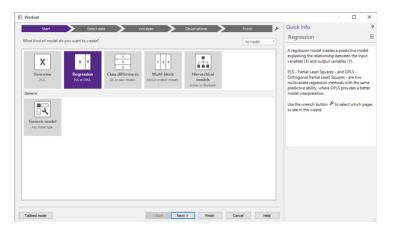
The next step is to select a representative set of samples (observations) that will have analyte concentrations spanning the interesting concentration ranges and at the same time spanning the normal spectral variation. A common way to address this is to use spiked samples where samples are spiked (added) with known amounts of the interesting analytes. This is especially important for analytes with low concentrations and low variability. For many analytes in a bioprocess setting, spiking is required but how to perform a relevant spiking strategy is not part of this tutorial.

To begin the modelling we start by defining the workset. In SIMCA, the workset is the collection of samples and variables that a model is to be built on. This is sometimes called training set or calibration set. In the setup of the workset, we also need to set some important model parameters.

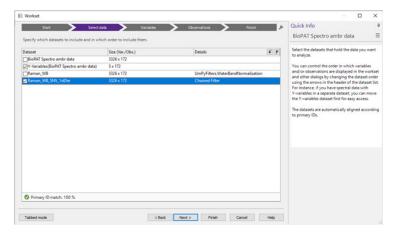
Click on the type of model you want to build. We use Orthogonol Partial Least Squares (OPLS) for this tutorial.



In the workset wizard, select the Regression model mode to guide you through the necessary steps. Click Next to start setting up the calibration model.



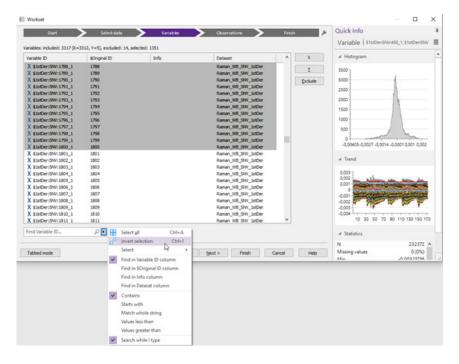
On the Select data tab, you select the datasets that hold your (preprocessed) spectral data and your Y-data (the Glucose concentrations). Click Next.



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.

ariable ID	\$Original ID	Info		Dataset	^	×	.⊿ Histogram	
Y Glucose Y Lactate Y Gutamine 9 Janbers 5997: 500-15 9 Jastbers 5997: 500-15 X Stabers 5997: 500-15 X Stabers 5997: 500-15 X Stabers 5997: 500-15 X Stabers 5997: 100-15 X Stabers 5997: 110-15 X Stabers 5997: 111-15 X Stabers 59971: 111-15 X Stabers 5997: 111-15 X Stabers 5997; 111-15 X Staber	100 101 102 103 105 105 105 105 100 110 111 112 113 114	Columns X Y	<b>v</b> :	Y-Variables(BioPAT Spectro Y-Variables(BioPAT Spectro Y-Variables(BioPAT Spectro Y-Variables(BioPAT Spectro Y-Variables(BioPAT Spectro Raman_WB_SW1stDer Raman_WB_SW1stDer Raman_WB_SW_1stDer Raman_WB_SW_1stDer Raman_WB_SW_1stDer Raman_WB_SW_1stDer Raman_WB_SW_1stDer Raman_WB_SW_1stDer Raman_WB_SW_1stDer Raman_WB_SW_1stDer Raman_WB_SW_1stDer Raman_WB_SW_1stDer Raman_WB_SW_1stDer Raman_WB_SW_1stDer Raman_WB_SW_1stDer		Y Exclude	70 60 40 30 40 30 40 40 40 40 40 40 40 40 40 4	025-0,
X \$1stDer:SNV:115_1 X \$1stDer:SNV:116_1 X \$1stDer:SNV:117_1 X \$1stDer:SNV:118_1	115 116 117 118	Switch X & Y Select all C	trl+A	Raman_WB_SNV_1stDer Raman_WB_SNV_1stDer Raman_WB_SNV_1stDer Raman_WB_SNV_1stDer			-0,05	1

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.



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.

ariable ID	SOriginal ID	Info	Dataset	•	x	.⊿ Histogram	
Glucose	songna ID	one	Y-Variables(BioPAT Spectro		_		
Gucose			Y-Variables(BioPAT Spectro		Y,	60 1	
- Glutamine			Y-Variables(BioPAT Spectro		5		
Amnonia			Y-Variables(BioPAT Spectro		Exclude	50-	
Glutamate			Y-Variables(BioPAT Spectro			40-	
\$1stDer:SNV:100_1	100		Raman_WB_SNV_1stDer			~~	
\$1stDer:5W:100_1	101		Raman_WB_SWV_1stDer			30-	
\$1stDer:SNV:102_1	102		Raman_WB_SNV_1stDer				
\$1stDer:SW:103_1	103		Raman_W8_SWV_1stDer			20-	
\$1stDer:SNV:104_1	104		Raman_W8_SWV_1stDer			10	
\$151Der:SNV:105_1	105		Raman_WB_SNV_1stDer			10-	
\$1stDer:SNV:106_1	106		Raman_W8_SWV_1stDer			0	
\$1stDer:SNV:107_1	107		Raman_WB_SWV_1stDer			1 2 3 4 5 6 7 8 9 10 11	12 13 14 15
\$1stDer:SNV:100_1	108		Raman_WB_SNV_1stDer				
\$1stDer:SNV:109_1	109		Raman_WB_SNV_1stDer			# Trend	
\$1stDer-SW-110-1	110		Raman_WB_SNV_1stDer				
\$1stDer:SNV:111-1	111		Raman_WB_SNV_1stDer			16 3	
\$1stDer-SNV:112-1	112		Raman_WB_SNV_1stDer			14 114	
\$1stDer:SW:113-1	113		Raman_WB_SNV_1stDer			12-10-1-1-1	
\$1stDer-SW-114 1	114		Raman_WB_SNV_1stDer				th la
\$1stDer:SNV:115-1	115		Raman_WB_SNV_1stDer			W 4/14	NmN
\$1stDer: SW:110 1	116		Raman WB_SNV_1stDer			1 UN V V	1 D.
\$1stDer:SNV:117_1	117		Raman_WB_SNV_1stDer			å1	
\$1stDer:SW:110-1	118		Raman WB SNV 1stDer	¥		10 30 50 70 90 110	130 150 1
nd Variable ID							

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.

	72, selected: 11								Observation   1406-1358-1, 1397-1349-1,
bservation ID	Spiked volum	e					^	Include	D Histogram
1406-1358-1								Exclude	4 Trend
1397-1349-1	0,025							Proven a	
1370-1322-1	0,06								12 1
1346-1298-1									10-
1333-1285-1									8
1215-1168-1			Columns		V	Primary observation ID			6
1213-1166-1									4-
1211-1164-1		12	Include			Capture start time			0
1207-1160-1	10000	35	Exclude		V	Spiked volume			.21
1204-1157-1	0,025		Set class			Spiked with	1.11		0 200 400 600 800 1000 1200
1199-1152-1 1195-1148-1		-							
1195-1148-1		=	Select all	Ctrl+A		Glucose (before spiking)			Statistics
1179-1132-1					1.	Lactate (before spiking)			
1174-1127-1						Glutamine (before spiking)			
1171-1124-1						Vessel number			
1168-1121-1									
1143-1096-1						Water band normalisation			
1137-1090-1						Glutamate (before spiking)			
1127-1080-1						More			
1124-1077-1					-		-		
1121-1074-1									
1115-1068-1									
							~		

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 "O" 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.

Start	Select data	Variables	Observations	Finish	۶	Quick Info
Observations: 172, included:	172, selected: 32					Observation   1397-1349-1, 1370-1322-1,
Observation ID	Spiked volume			^	Include	Histogram
√ 1397-1349-1 √ 1370-1322-1 √ 1370-1322-1 √ 1330-1328-1 √ 1330-1385-1 √ 1330-1385-1 √ 1213-1166-1 √ 1211-1166-1 √ 1211-1166-1 √ 1211-1167-1 √ 1399-1152-1 √ 1399-1152-1 √ 1399-1152-1 √ 1399-1152-1	0,025 0,06 0,025				Exclude	4 Trend 14 10 6 2 -2 0 200 400 600 800 1000 1200
✓ 1179-1132-1 ✓ 1179-1132-1 ✓ 1179-1124-1 ✓ 1179-1124-1 ✓ 1179-1124-1 ✓ 1179-11990-1 ✓ 1137-1090-1 ✓ 1137-1090-1 ✓ 1127-1090-1 ✓ 1122-1077-1 ✓ 1122-1077-1 ✓ 1122-1077-1 ✓ 1122-1076-1 ✓ 1112-1065-1	💽 📻 Select all	Ctri+A		۷		D Statistics
Tabbed mode	Invert selection Select Find in Observation	Ctrl+I Back	Next > Fin	sh Cancel	Help	
	Find in Spiked volu     Find in Class colum	me column				
	Contains Starts with Match whole string Values less than					
	Values greater than					

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).

1 Workset		×
Start Select data Variables Observations Finish 🖌	Quick Info	ņ
	Fitting the model	≡
Summary The model you have specified is a OPLS model. The model has 32 observations and 1352 variables (X = 1351, Y = 1). Scaling Ctr 1351 variables, UV 1 variable	You have now specified a workset and the next step is to fit this model so that you can analyze and interpret your data. Model type Based on your selections and initial objective, there are different model types available. Select the one you want to use. Fit model Select how (or if) you want to fit the model. Autofit will use cross-validation to decide how many components are appropriate for your data. You can always change the number of components later using the buttons found in the Fit model group on the Home tab.	
Model type:     CircLS model       Fit model:     Autofit model       Tabbed mode     < Back	Model options	~

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

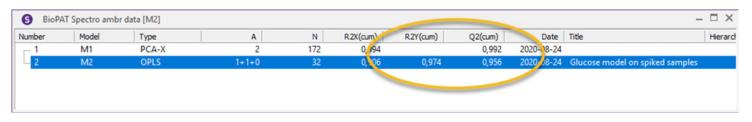
Nodel Options		-		×
Model Distance to model Coefficients Residuals / R2 Predictions	CV groups	More op	tions	
Number of cross-validation groups: 6	CV groups	:		Ç
What should the assignment of cross-validation groups be based on?	No N	lame	Group	^
Assign observations based on observation ID	1 1	397-1	6	
Assign observations based on observation ID 🛛 🗸	2 :	370-1	4	
ID: Vessel number V	3 :	204-1	5	
	4 :	411-1	2	
How should observations be grouped using the selected data above?	5 1	318-1	2	
Group observations with the same value in the same group	6 1	278-1	6	
	7 :	262-1	1	
	8 1	242-1	5	
	9 1	218-1	2	
	10 1	172-1	6	
	11 1	163-1	5	
	12 1	126-1	6	
	13 1	385-1	1	
	14 1	271-1	4	
Changing the cross-validation groups can significantly change the number of components in a model. If you change the model type	15 1	231-1	1	
of a model the assignment of cross-validation groups will be reset to	16 1	186-1	5	
the default.		175.1	-	~
OK	C	ancel	He	p

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.

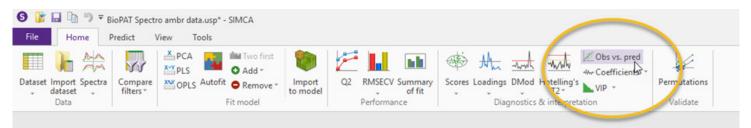
### Evaluate the calibration model performance

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

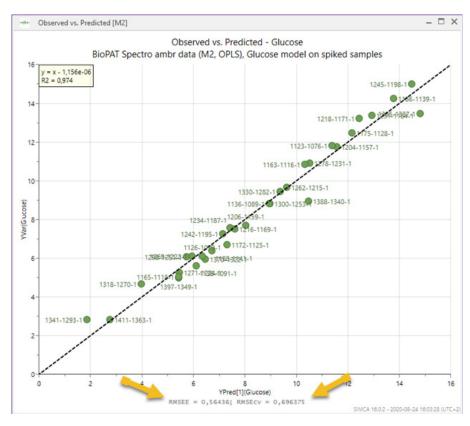


It is also important at this point to check if there are any outliers in the regression model. This is done by investigating the sample spectral residuals (DModX) and the Hotelling's T2 which shows how far away from model center each observation is. In this example, no extreme outliers were found (not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



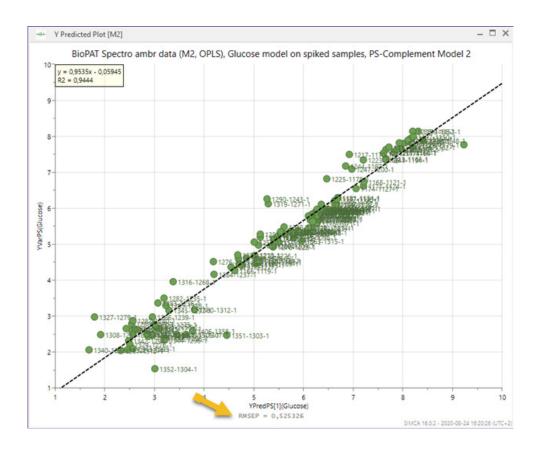
If this is not good enough, the model may have to be refined by e.g. changing the variable range, preprocessing chain or settings, or outlier removal. Apply the changes and fit a new model and compare the results to see if predictive ability improves.

### 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<sup>®</sup> 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

