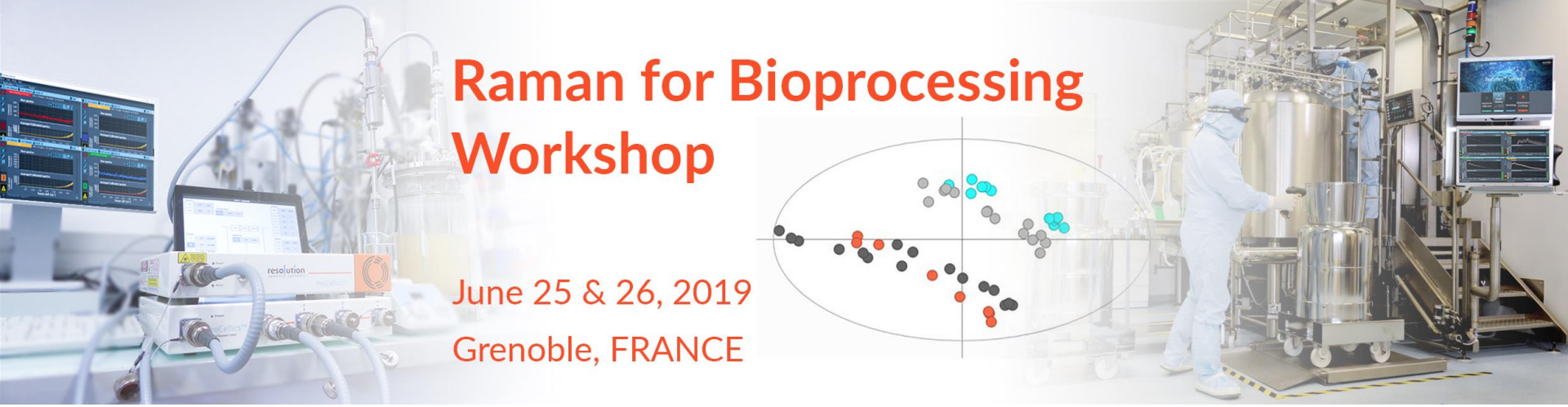


Efficient Raman PLS modeling

Laure PETILLOT – Scientist, Bioprocess Analytics

resolution
spectra systems

Agenda

1. Basics – reminder
   a) Data collection
   b) Model building steps on ProCellics™
   c) Chemometric Analysis on Simca

2. Efficient Raman modeling
   a) Challenge #1: high quantity of calibration data and parameters
   b) Challenge #2: lack of consistency in model-building steps
   c) Challenge #3: achieving real-time monitoring vs external validation
   d) Challenge #4: managing process and setup variations

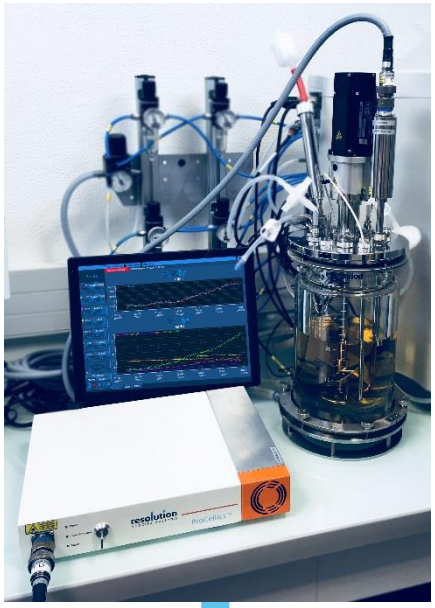
Basics – reminder

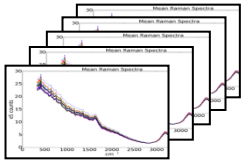
Chemometric analysis

PLS Partial Least Squares regression: quantitative analysis

Multivariate calibration

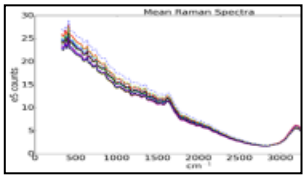
Monitoring





Raman spectra

Raman spectrum TN
Raman spectrum T4
Raman spectrum T3
Raman spectrum T2
Raman spectrum T1



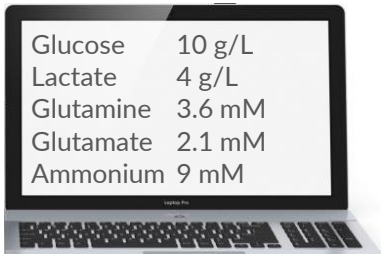
Off-Line analyzer

Partial Least Squares (PLS) regression modeling

Statistical model

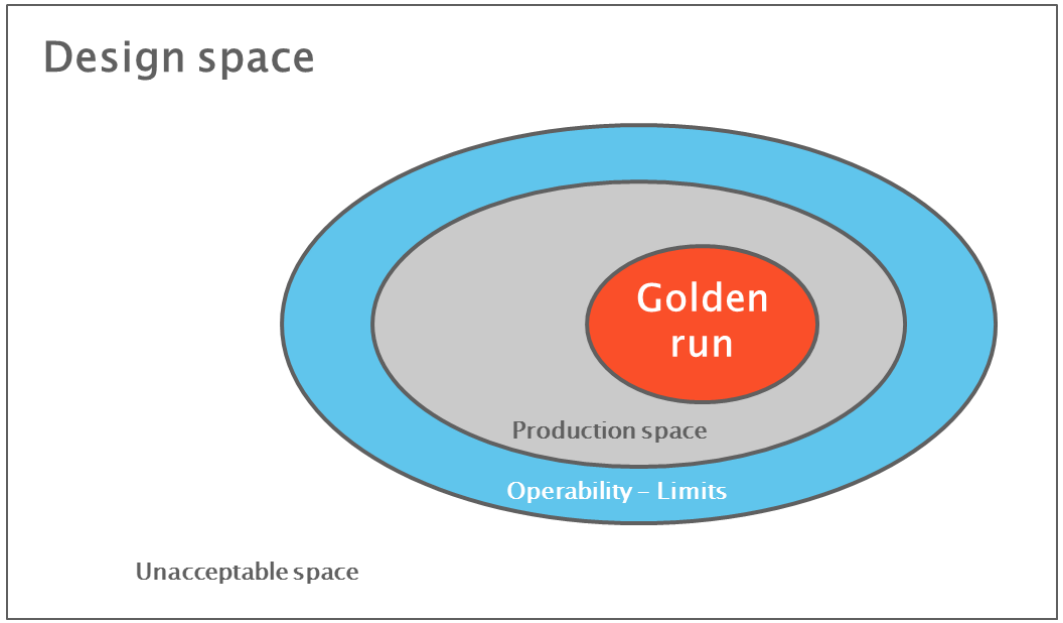
In-line real-time quantitation

Quantitation sample TN
Quantitation sample T4
Quantitation sample T3
Quantitation sample T2
Quantitation sample T1



Data collection

1. Recommendations



Data must be carefully selected to ensure a proper design space

Uniformity	**Uniform distribution** of samples throughout the range of potential variations
Limits	More samples included around the specified limits (in concentrations and in kinetics)
De-correlation	Avoiding correlation between metabolites: achieving **metabolite and cell spikes** (ex: feedings)
Sample number	The number of samples to be included in order to create a **valid calibration model** for quantitative analysis will depend on the **complexity of the process** medium (50 points for basic cell cultures)

Data collection: samples

2. Example of a standard batch data collection

Standard batch with normal sample strategy and feeds with different elements

Batch	Day	Feed	Samples before spike	Samples after spike	"Free samples"
"Standard" culture batch example	D1	-	0	0	1
	D2	Medium	1	1	0
	D3	-	0	0	1
	D4	Medium / Glucose	1	1	0
	D5	-	0	0	1
	D6	Medium/ Glutamate	1	1	0
	D7	-	0	0	1
	D8	Medium / Glucose	1	1	0
	D9	-	0	0	1
	D10	Medium/ Glutamate	1	1	0
	D11	-	0	0	1
	D12	Medium	1	1	0

Data collection: samples

3. Example of a special batch data collection

Special batch with adapted sample strategy and feeds with only one element per feed

Batch	Day	Feed	Samples before spike	Samples after spike	"Free samples"
"Special" culture batch example	D1	-	0	0	1
	D2	Medium	1	1	0
	D3	Glucose	1	1	0
	D4	Medium	0	0	1
	D5	Glutamate	1	1	0
	D6	Medium	0	0	1
	D7	Glucose	1	1	0
	D8	Medium	0	0	1
	D9	Glutamate	1	1	0
	D10	Medium	0	0	1
	D11	-	0	0	1
	D12	Medium	1	1	0

Data collection: spectra

4. Spectra acquisition

Acquisition interface: spectra, notes, samples

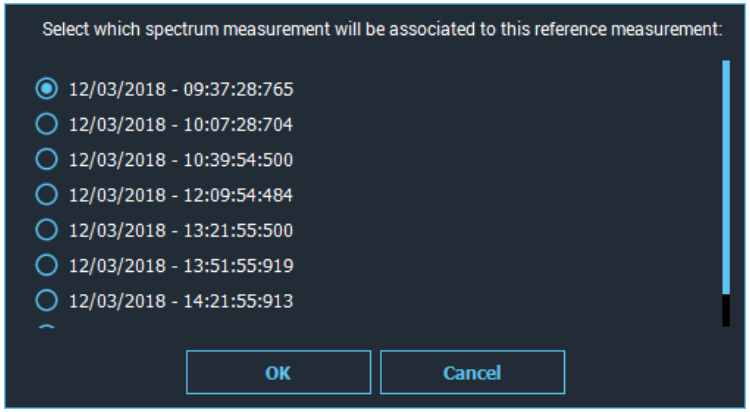
Model building steps on ProCellics™

1. Reference data association

- By clicking on the  button during your batch acquisition, a line is created in a reference data association table
- Each line corresponding to one spectral acquisition corresponds to one of your off-line measurement

	Name	Date - hour	TCD (Cells/mL)	VCD (Cells/mL)	Ammonia (mM)	Lactate (g/L)
1	New sample 1	12/03/2018 - 09:37:28				
2	New sample 2	12/03/2018 - 10:07:28				
3	New sample 3	12/03/2018 - 12:09:54				
4	New sample 4	12/03/2018 - 13:51:55				
5	New sample 5	12/03/2018 - 14:21:55				
6	New sample 6	12/03/2018 - 14:51:55				

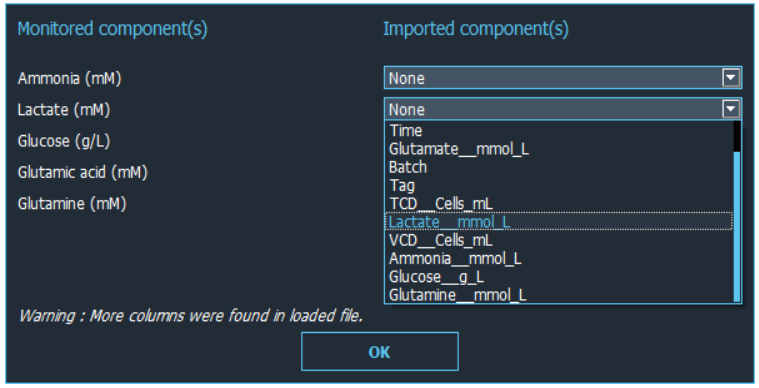
- A reference can also be created by clicking the “Add reference”  button



Model building steps on ProCellics™

1. Reference data association

- **Automatic**: download your .csv file (example: Nova Biomedical, Cedex Roche, etc.) and associate the correct column of your file with the components of the sub-project



- **Copy and paste** from your off-line measurement file

- **Manually** enter the values

	Name	Date - hour	TCD (Cells/mL)	VCD (Cells/mL)	Glucose (g/L)	Glutamine (mM)	Ammonium (mM)	Glutamic acid (mM)	Lactate (g/L)
1	Sample 1	20/02/2019 - 17:19:26	420667	412333	6.84667	9.78	1.21333	1.79667	0.01
2	Sample 2	20/02/2019 - 17:40:45	409333	404333	10.3433	8.73	1.14	2.15667	0.01
3	Sample 3	21/02/2019 - 09:17:32	536333	527000	10.1567	7.88333	1.95	2.01667	0.316667
4	Sample 4	21/02/2019 - 16:35:13	737000	725000	9.94667	7.04667	2.54	2.4	0.493333
5	Sample 5	22/02/2019 - 09:13:20	1.39267e+6	1.38533e+6	9.14667	5.37667	4.17	2.57	1.00667
6	Sample 6	22/02/2019 - 17:39:07	2.09033e+6	2.076e+6	8.51	4.14	5.20333	2.82333	1.31333
7	Sample 7	23/02/2019 - 09:58:00	4.062e+6	4.04067e+6	7.1	1.81667	7.29333	2.99333	1.84
8	Sample 8	24/02/2019 - 10:08:35	8.04267e+6	7.97233e+6	3.68333	1.62667	11.13	3.41667	2.75
9	Sample 9	24/02/2019 - 10:52:45	7.964e+6	7.90467e+6	5.94333	6.55667	10.2567	3.46667	2.49667
10	Sample 10	25/02/2019 - 10:37:37	1.09707e+7	1.0893e+7	3.54667	0.38	12.2367	3.66	2.35
11	Sample 11	25/02/2019 - 11:52:55	1.15823e+7	1.15323e+7	7.58667	5.3	12.1267	3.16	2.23667
12	Sample 12	25/02/2019 - 16:55:05	1.23427e+7	1.2256e+7	7.23333	3.08667	12.8867	3.05333	2.17

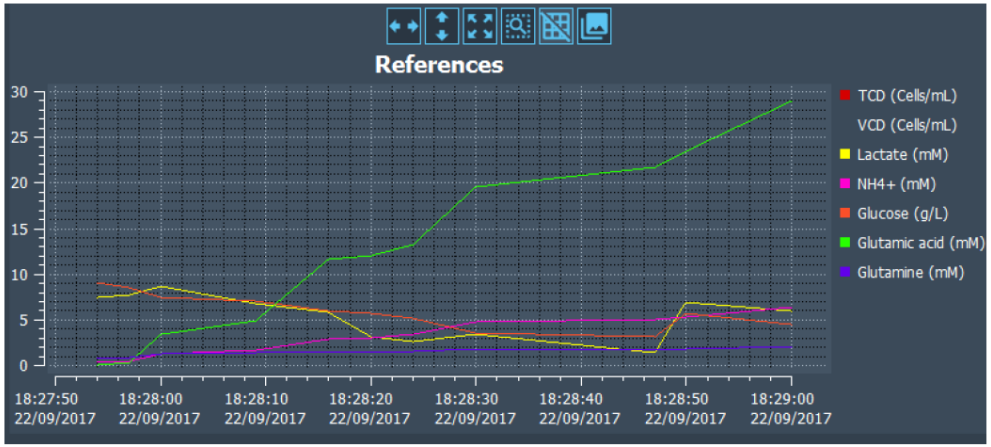
Model building steps on ProCellics™

1. Reference data association

➔ **A spectra association is created**: each sample in the table is associated to the nearest/previous/next spectrum

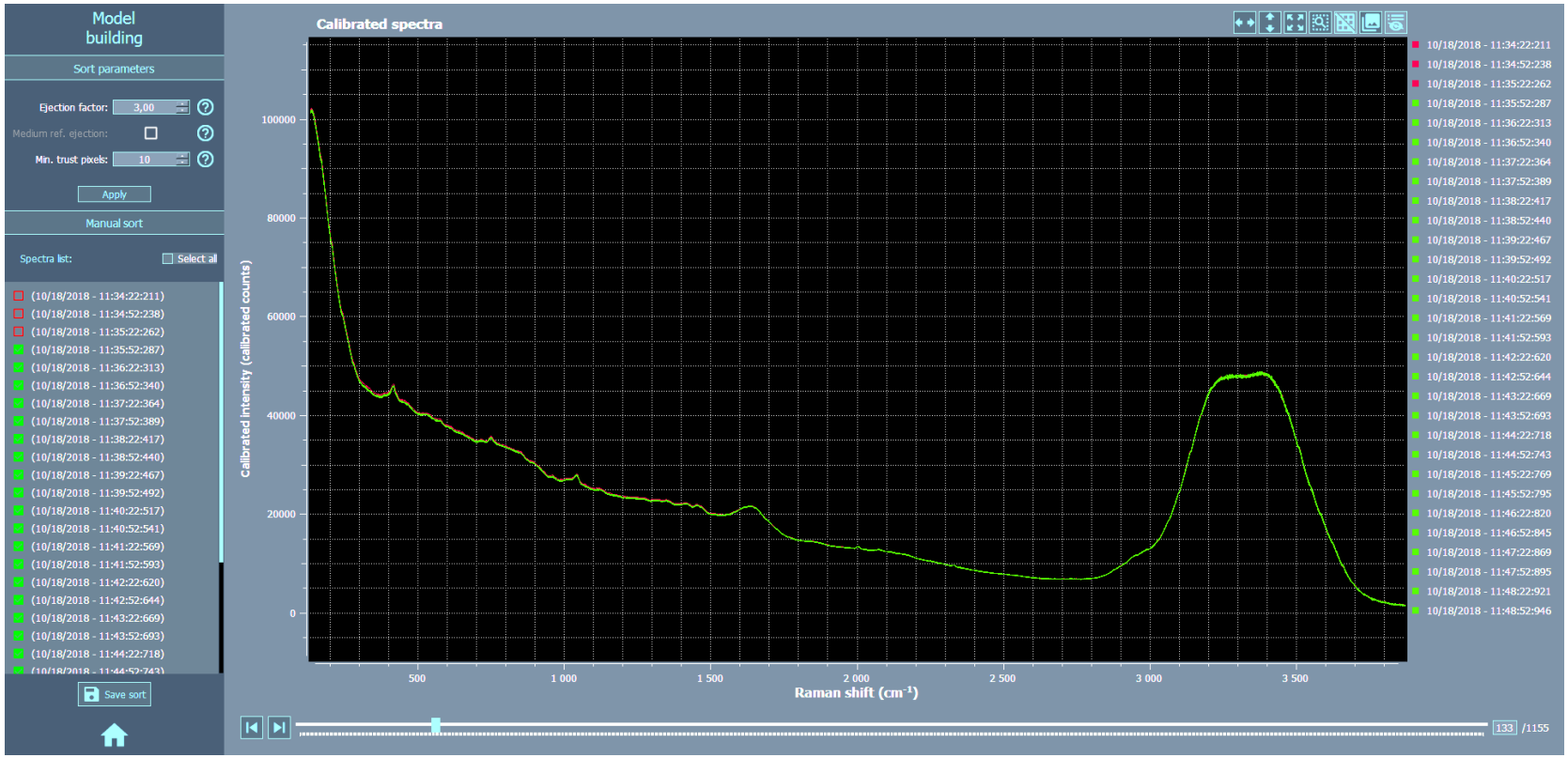
- Visualization of notes and graphs: the user can visualize the “Notes” saved during the batch - with the button  - and the “References” graph corresponding to the reference data kinetic during the batch

	date - hour	
1	06/09/2017 - 11:34	Medium reference acquisition.
2	06/09/2017 - 11:34	Start batch acquisition.
3	06/09/2017 - 11:35	Reference added.
4	06/09/2017 - 11:37	Note added : Glucose Feeding 1g/L
5	06/09/2017 - 11:38	Pause batch.
6	06/09/2017 - 11:39	Resume batch.
7	06/09/2017 - 11:41	Stop batch acquisition.



Model building steps on ProCellics™

2. Elimination of aberrant spectra

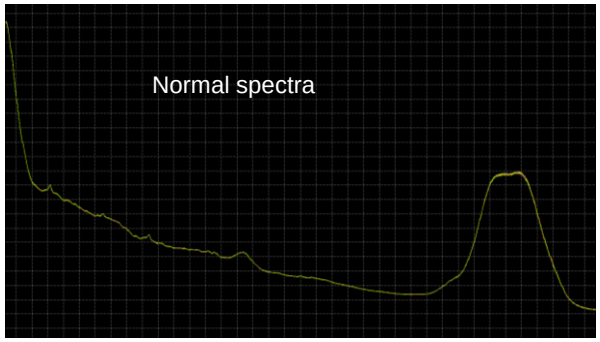


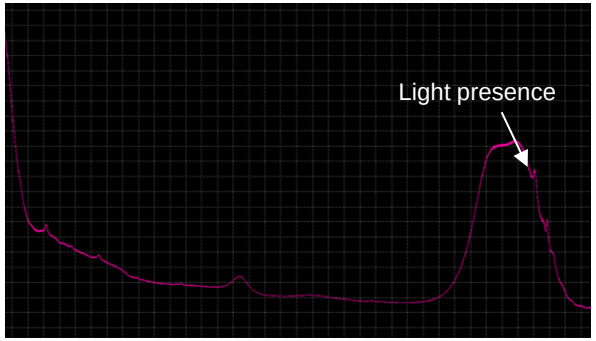
- **Automatic**: restriction to 3*σ (intra and with reference medium spectra (available in a next version))
  This treatment will allow to automatically unselect the aberrant spectra in one acquisition

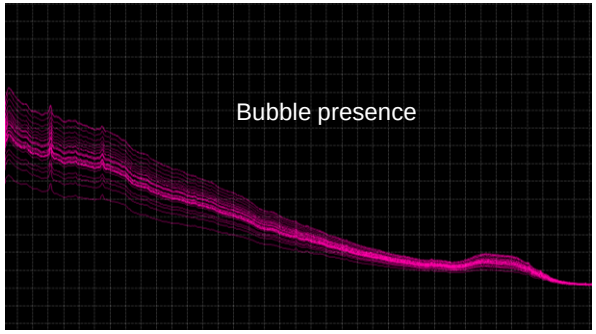
Model building steps on ProCellics™

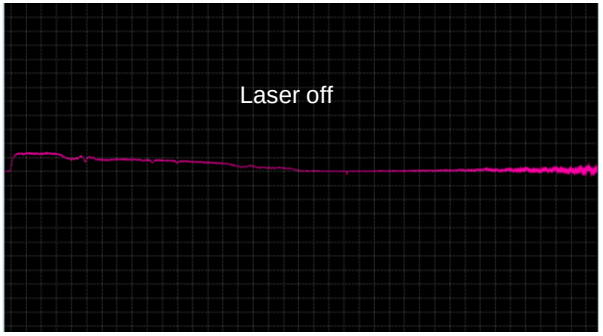
2. Elimination of aberrant spectra

- **Manual**:
  - Adjustable ejection factor and trusted pixel number
  - Manual (de)selection in the list ➔ if you have a bubble / light / laser- off, etc. you may have to **unselect all the acquisition**





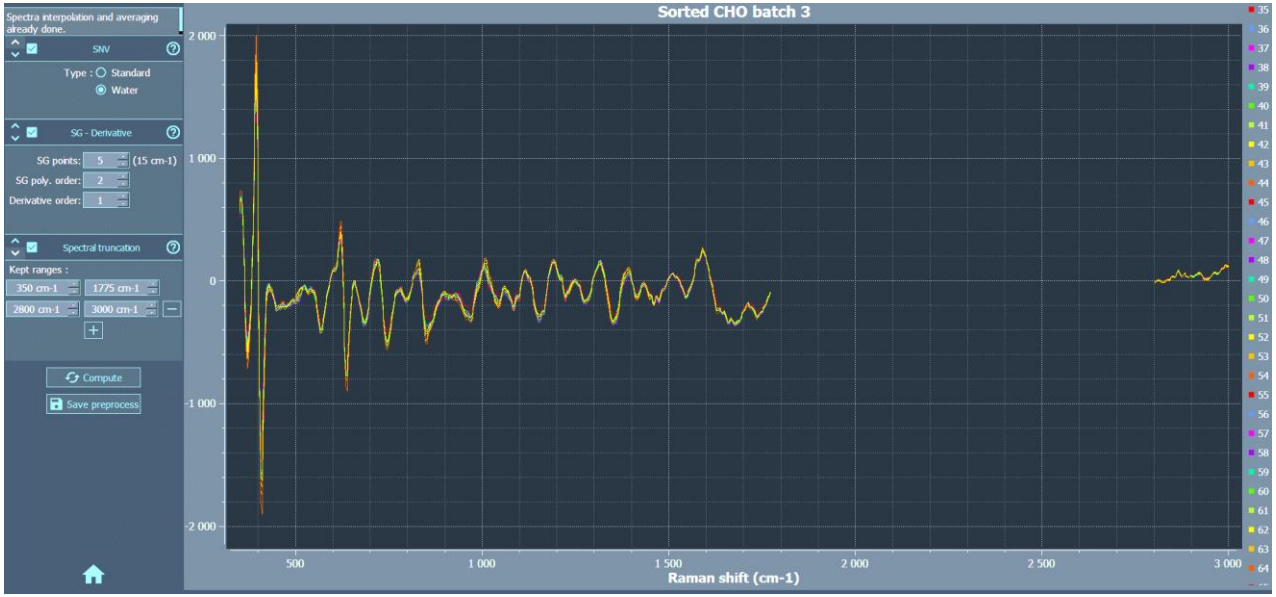




Model building steps on ProCellics™

3. Preprocessing step

Demonstration of spectral differences by pre-processing Raman spectra

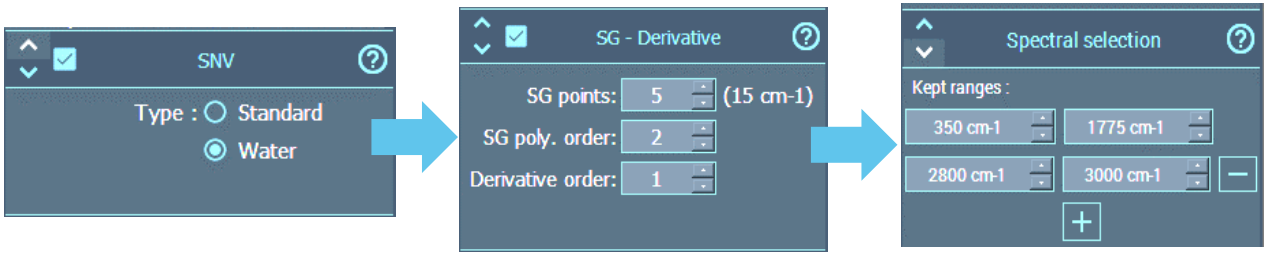


3 pre-processing tools:

- **Normalization**
- **SG - Derivatives**
- **Spectral truncation**

The order of combined pre-treatments makes a difference!

Example of pre-process:

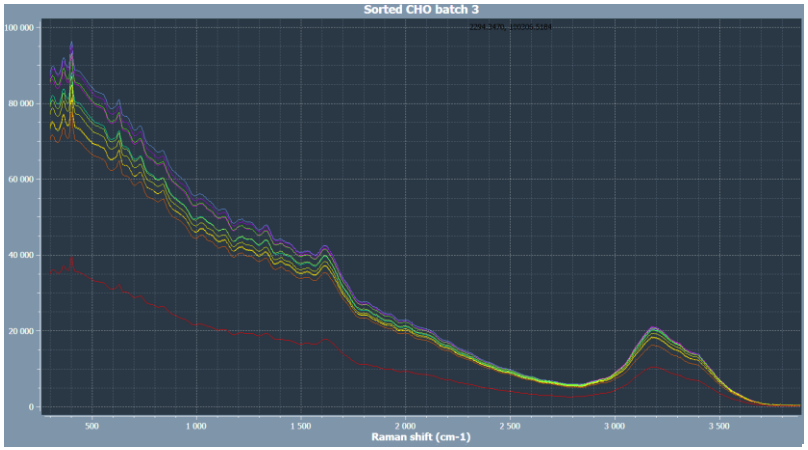


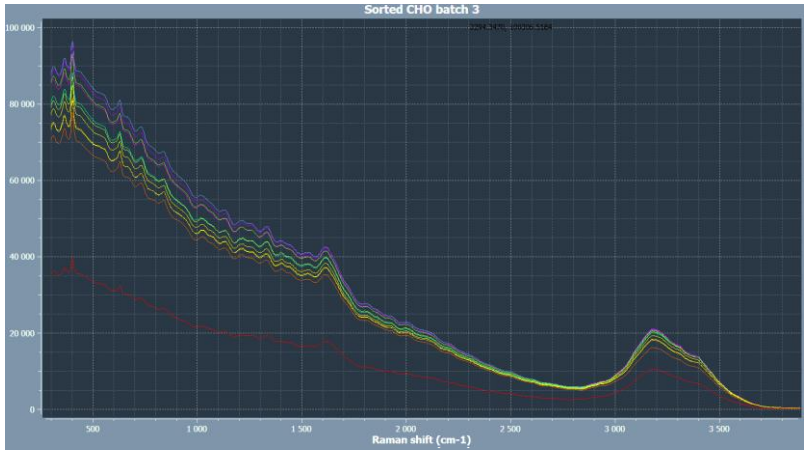
Model building steps on ProCellics™

3. Preprocessing step

Normalization: when applying the normalization filter, each spectrum is "normalized" by substracting the mean (by Raman shift) and dividing by the standard deviation. Two different algorithms are proposed:

- **Water normalization:** normalization is applied to the water spectrum (3,100-3,700 cm^{-1} band)
- **Standard normalization:** normalization is applied to the entire spectrum

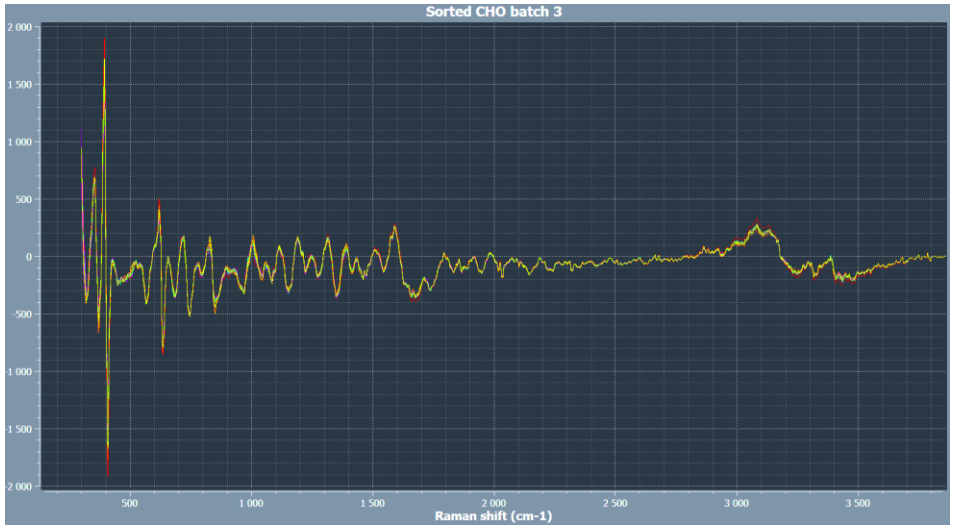




Model building steps on ProCellics™

3. Preprocessing step

SG-Derivatives: Savitzky-Golay-Derivatives are based on fitting a low degree polynomial function (usually of quadratic or cubic degree) on the data, followed by applying the first or second derivative from the resulting polynomial at points of interest

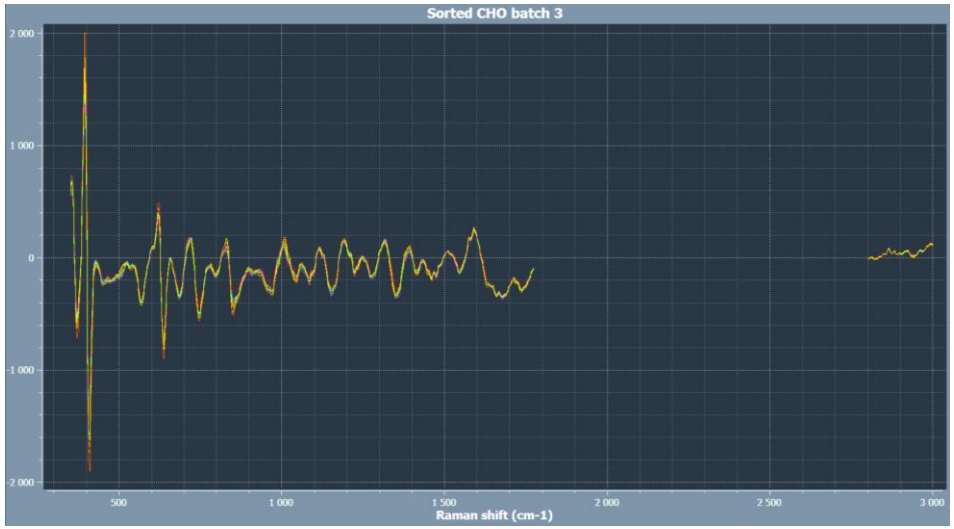


First derivative removes additive baseline

Model building steps on ProCellics™

3. Preprocessing step

Spectral Restriction to Regions Of Interest (ROI): if the region of interest is limited in so-called “fingerprint” bands, the user can precisely select the Raman-shifts ranges to analyze making a spectral truncation

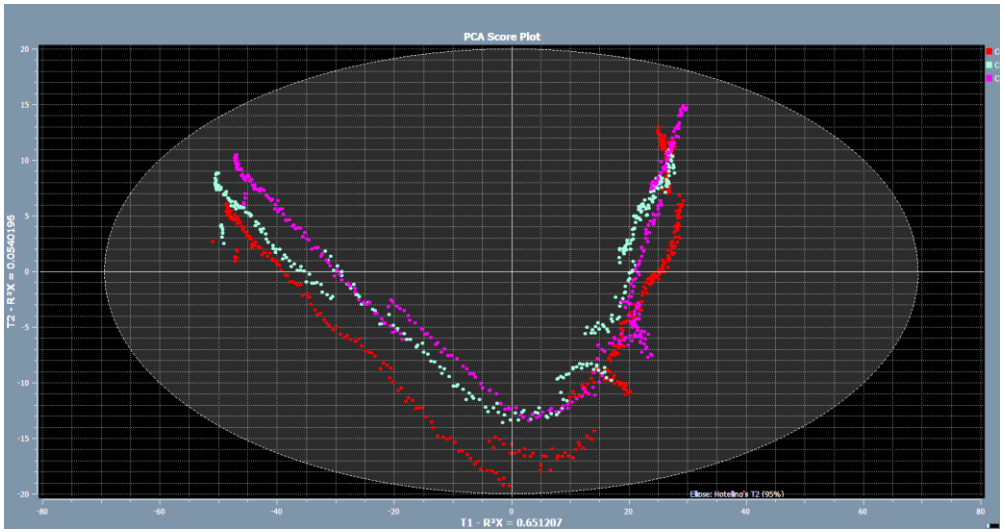


Selection of “finger-print” signals

Model building steps on ProCellics™

4. PCA Score Plot and Hotelling's T^2 Plot – All spectra

PCA Score Plot



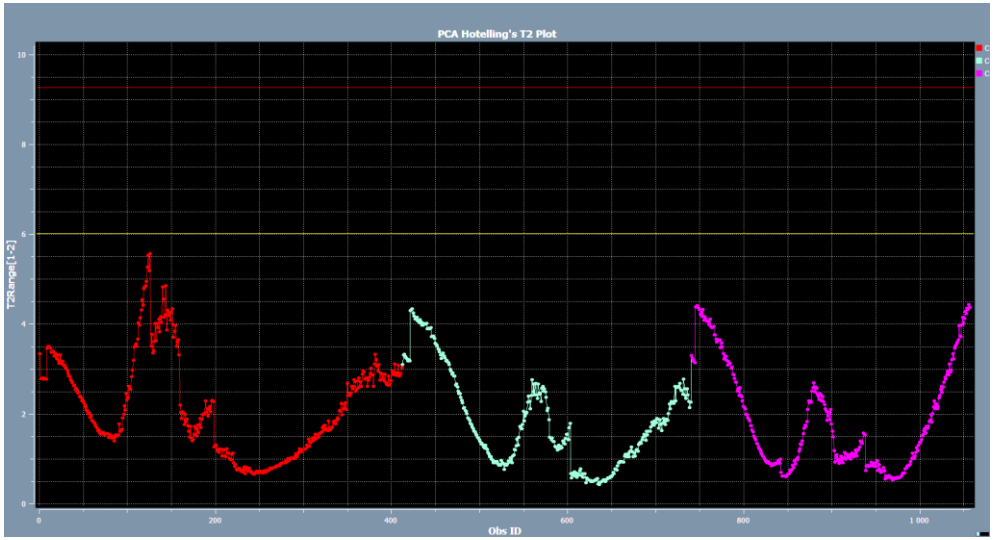
Principle

- "Simplification" of variables: seek to extract relevant information contained in spectra = synthesize spectra
- Graphical representations and distribution of each spectrum

PCA can be used for many different purposes:

- Identification of data groups and trends
- Faster detection of outliers

Hotelling's T^2 Plot



Model building steps on ProCellics™

5. Export dataset to SIMCA

Export

	Preprocess result name	Preprocessing	Export to SIMCA	Model association
1	PreProcess CHO	Done	Done	Do

Creating a .csv file to import the data into SIMCA

Host name: localhost
Port: 3306
Database name: bioProcess
Login : simca_user
Password : test
View name : PreProcAndRefExport_view_1
Creation date : 2018-04-27 17:09:02
Creator : test

.csv file export

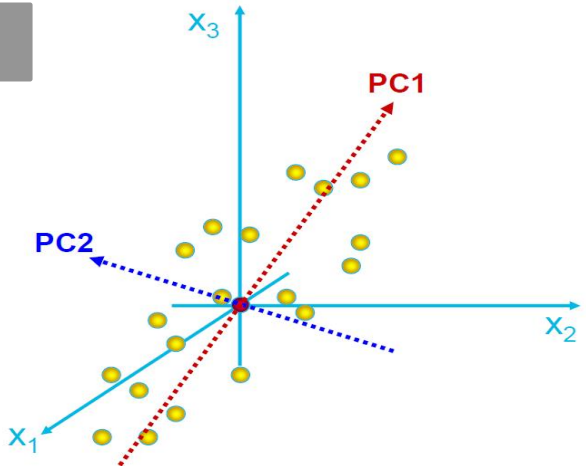
Select which components you want to export :
TCD (Cells/mL)
VCD (Cells/mL)
NH4+ (mM)
Lactate (mM)
Glucose (g/L)
Glutamine (g/L)
All
None
Select time scale: ms seconds min hours days
File path: C:/Users/f.thomas/Desktop/export_monitoring.csv
OK
Cancel

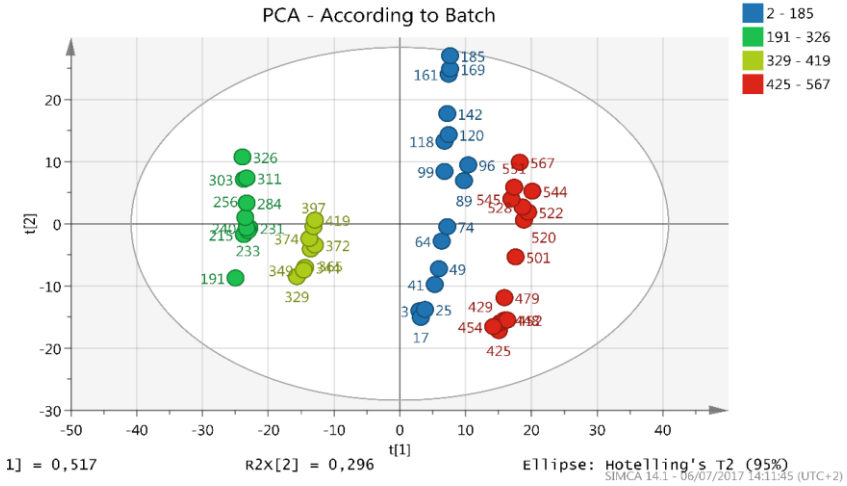
SIMCA

Chemometric analysis on Simca

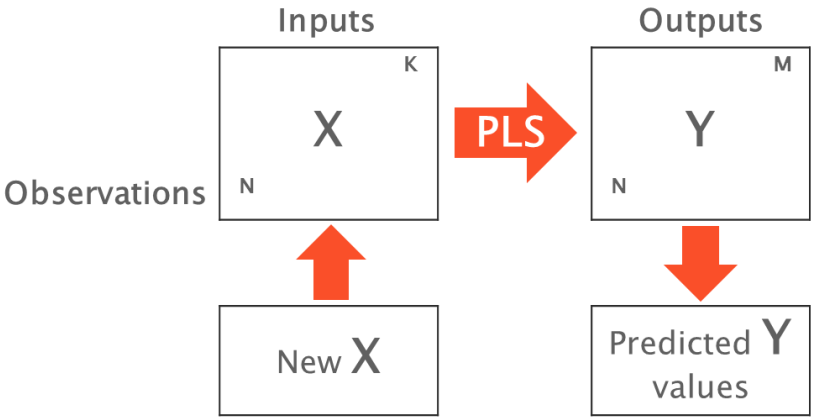
1. Model Building

PCA





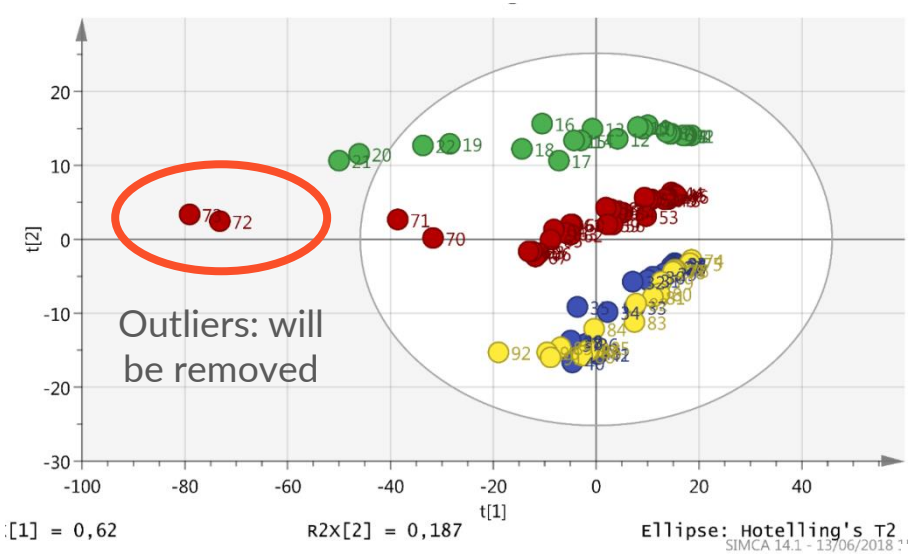
PLS



PLS models = relationship between blocks
X (spectra) and Y (off-line data)

Chemometric analysis on Simca

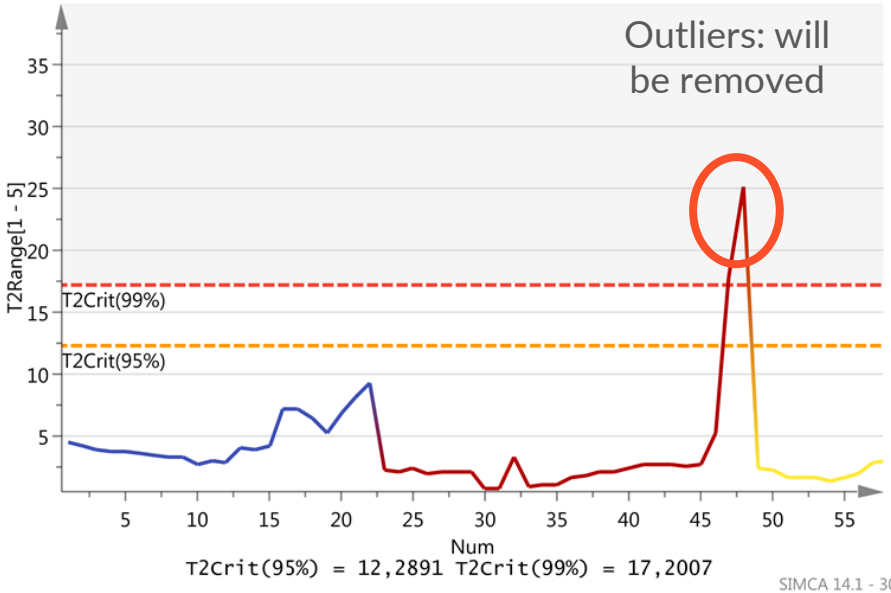
2. Managing outliers



Outliers in data can distort predictions and affect the accuracy if they are not detected an handled appropriately, especially in regression models

In order to detect the presence of outliers, two tools are available:

- the score plot
- the Hotelling's T^2 plot



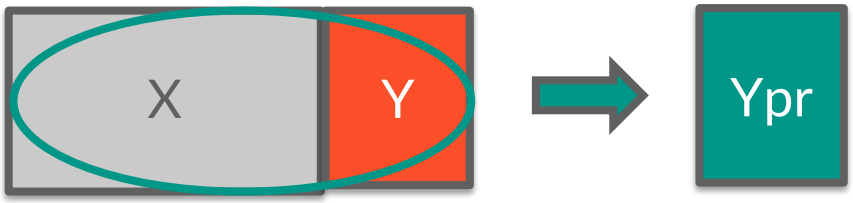
Chemometric analysis on Simca

3. Measuring the performance of a model

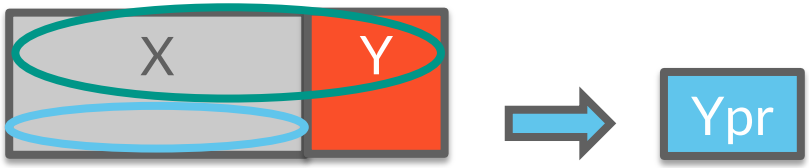
RMS (Root Mean Square) definition: $\sqrt{\left(\frac{(Y_{obs} - Y_{pred})^2}{n - 1}\right)}$

Frequently used to measure the model performance:

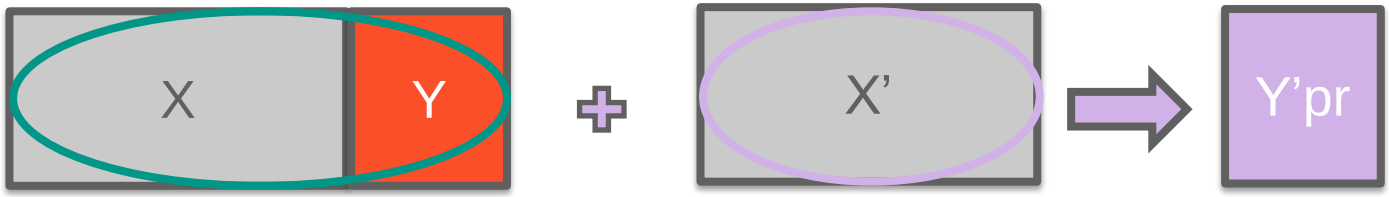
- Average prediction error
- Expressed in the same units as off-line data
- **RMSEE (Estimation Error):** Ypred is calculated from the same data as the ones used to make the model



- **RMSECV (Cross-Validation error):** some data are excluded (division in small datasets) and predicted. The successive models are realized until the entire rotation is obtained.



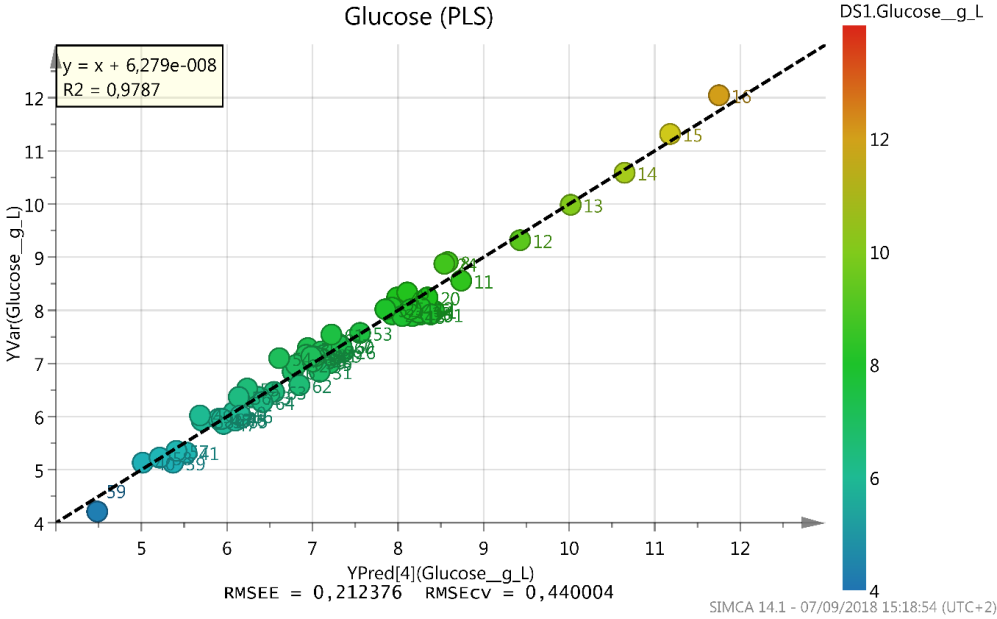
- **RMSEP (Prediction Error):** two separate sets of data. The error is calculated for the prediction set only.

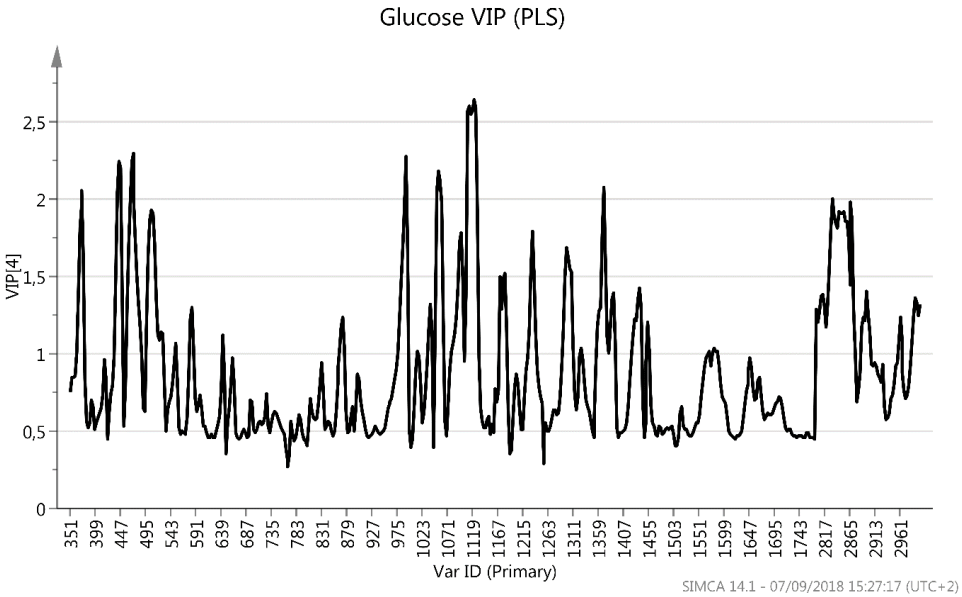


Chemometric analysis on Simca

4. Check the model linearity and contribution

Reference vs predicted: check the linearity of your model and have the confirmation of your RMSEcv





VIP tool: check that the Raman bands of your metabolite contribute well to your model

Chemometric analysis on Simca

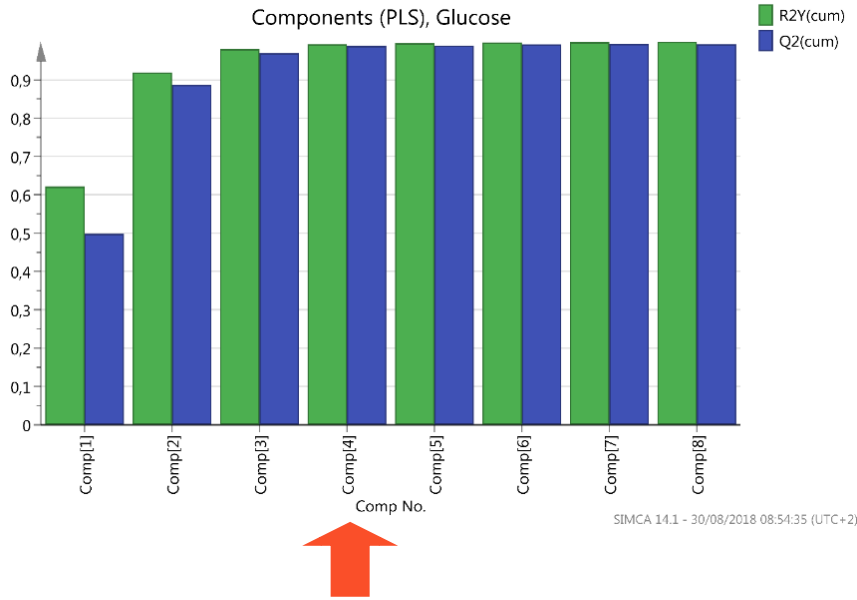
5. Choosing the number of components

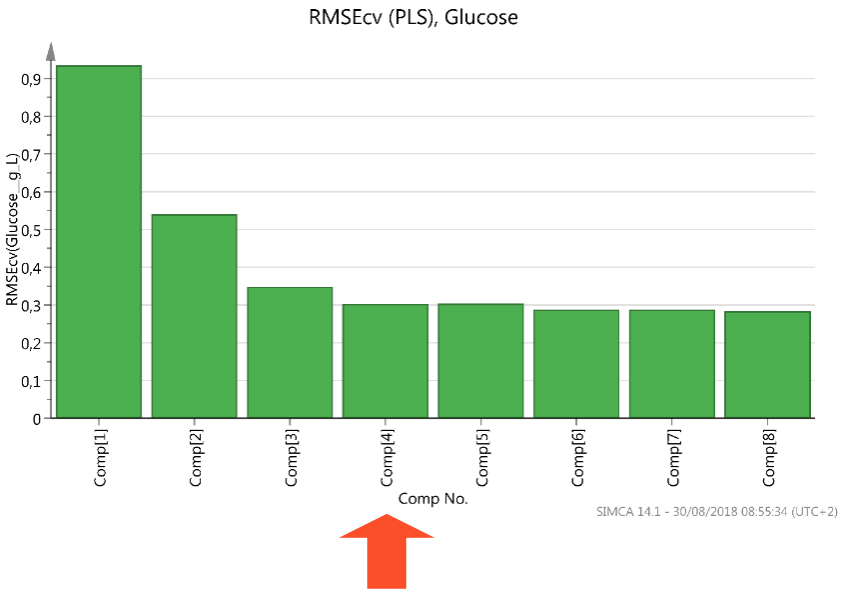
Components : new uncorrelated variables that successively maximize variance
To minimize the expected error, the number of component must be carefully chosen

- fitting the current data too poorly ➔ bad consistency in prediction
- fitting the current data too well ➔ model that does not generalize well to other data (with biological variability for example)

Cross validation

- reliable method for choosing the number of components
- avoids overfitting data by not reusing the same data to both fit a model and to estimate prediction error





Efficient Raman modeling

Challenge #1: high quantity of calibration data and parameters

Typical calibration data set and processing

Reference data	3 batches of 12 days 15 samples per batch	45 off-line quantitations of 3 to 15 parameters	130 to 700 values
Raman data	1 spectrum every 30 seconds	35,000 spectra of 1k pixels	35M unique values
Acquisition parameters	3: laser power, integration time, averaging		
Processing parameters	6 per model: number of components, SNV (normalization), Savitzky-Golay (3), ROI		

Manual sorting and processing with multiple interfaces
> 2- to 5-day effort per modeling session

➡ Incompatible with routine method

➡ How the calibration data set and parameters could be included in the data integrity requirements?

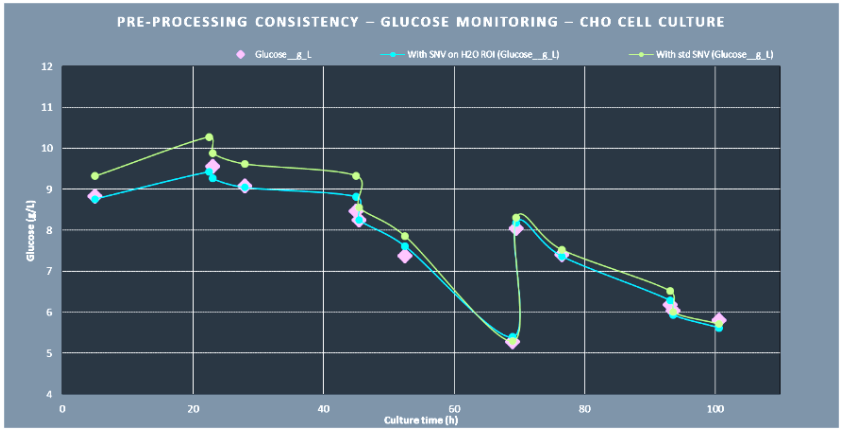
Our answer: data management to ease model building

Bioprocess monitoring is easily achieved thanks to a data pipeline and management

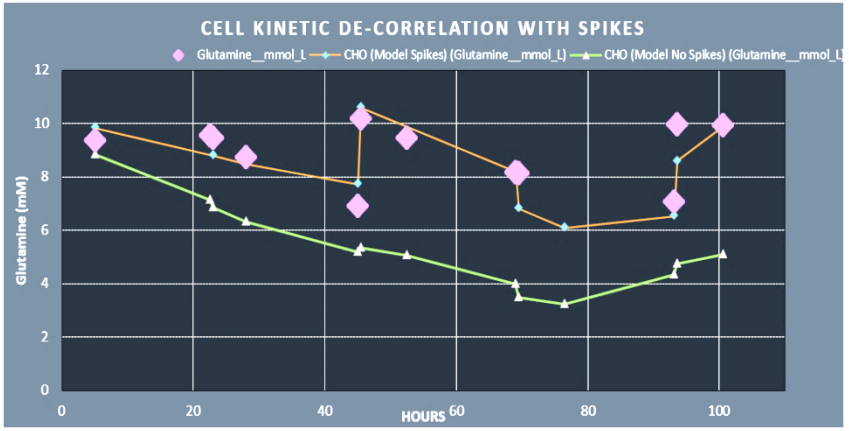
	With data management	Without data management
Database architecture	**5 min**	3h
Calibration batches acquisition (3 batches)	2-3 months	2-3 months
Spectra selection	**2h**	1 day
Pre-processing	**10 min**	10 min
Correlation with reference data	**0 min**	2h (SIMCA required)
Chemometric models building	**5 days**	5 days
Monitoring	**15 min (automatic)**	several hours => Several days

Challenge #2: lack of consistency in model-building steps

Bad consistency in pre-processing steps

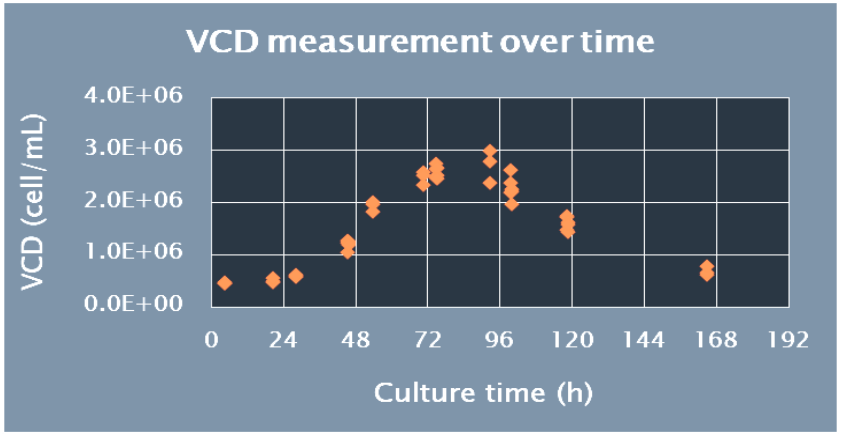


Metabolite bad de-correlation

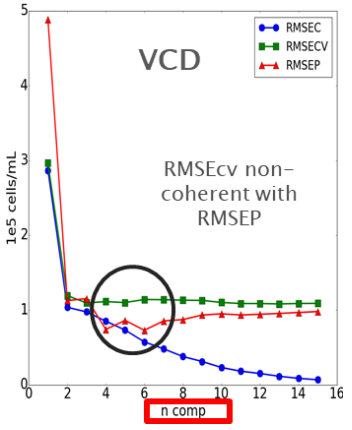


Errors conducting to an inaccurate prediction ➔Training and support help you to avoid traps

Strong variability in reference measurements



Wrong model building parameters choice



Challenge #3: achieving real-time monitoring vs external validation

Needs to overcome methodological issues

- Pre-processing integrity: how to manage one real-time spectrum pre-process?

Example: “normalization” needs consistent spectra dataset

- A step forward with regards to a posteriori processing: a model is generally well done

when performed on the basis of the batch to be predicted!
Important variabilities must be detected soon.

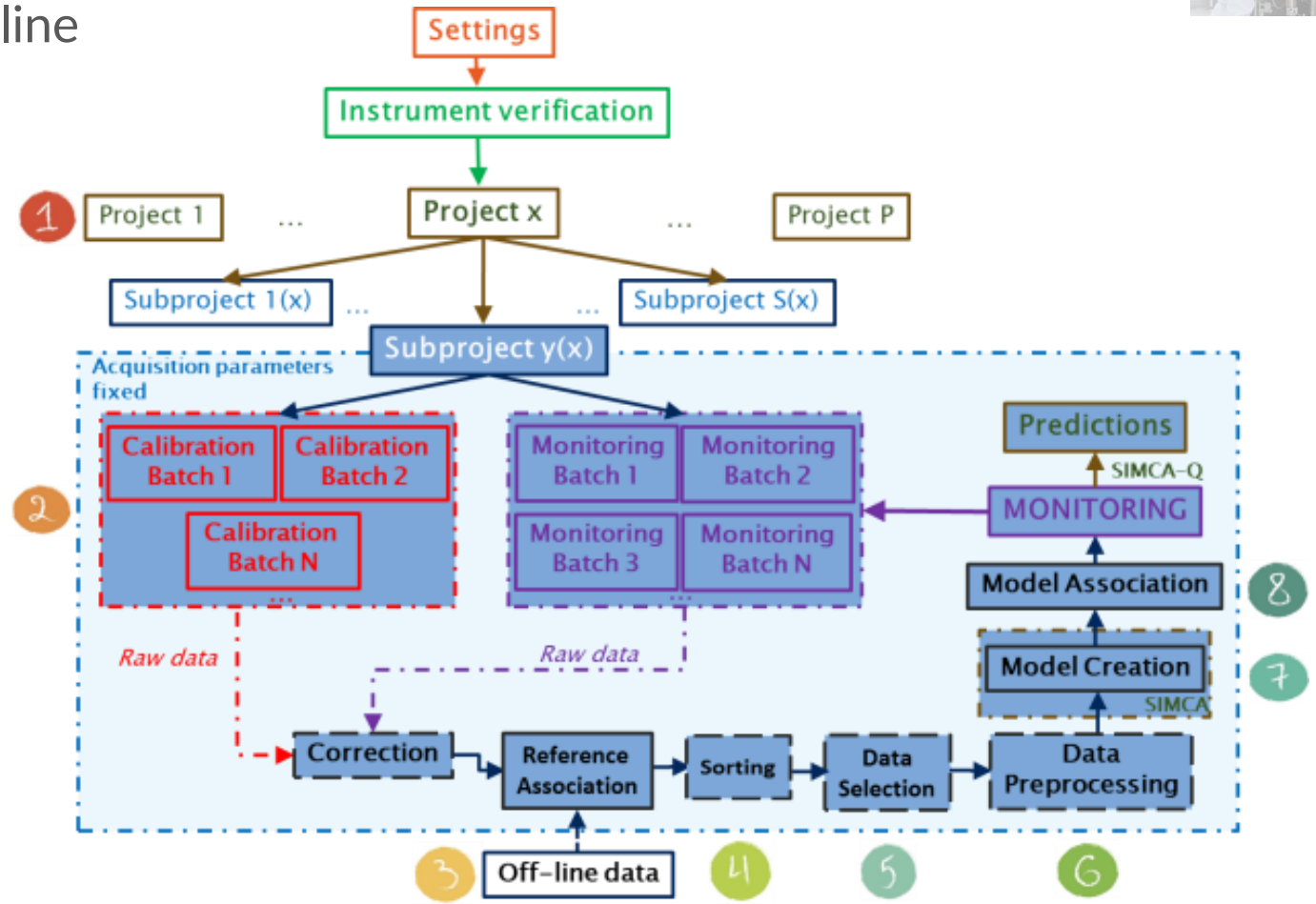
- Useful information: real-time quantitation needs to be accompanied by meta data, comments, links with related spectra, parameters and reference data (analytics needs comparison…)

Achieving real-time monitoring:

Complex software system
or
Integrated software

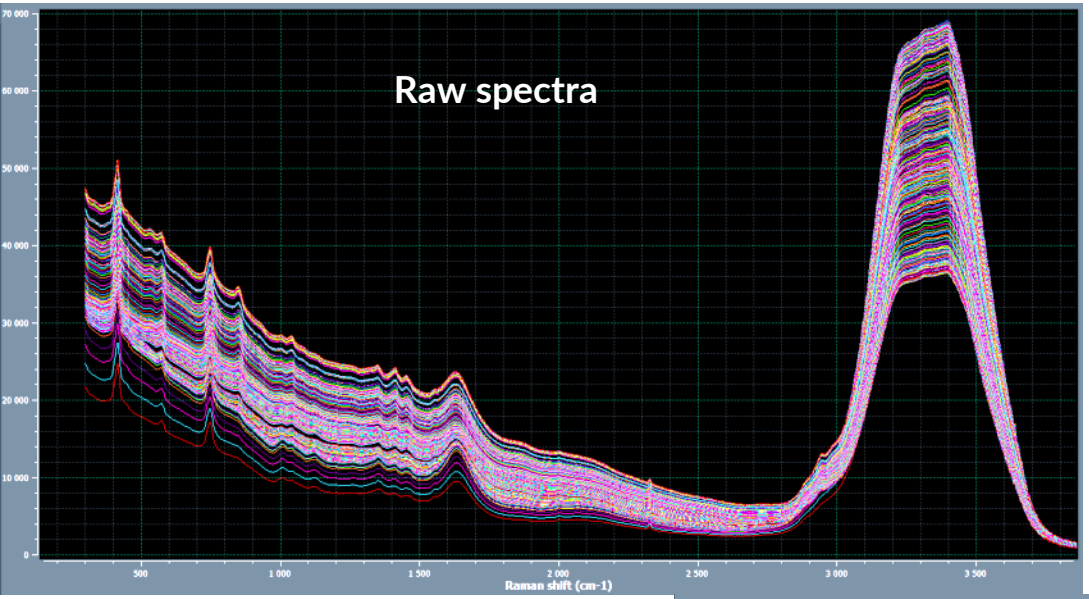
Our answer: data pipeline for real-time monitoring

Data pipeline



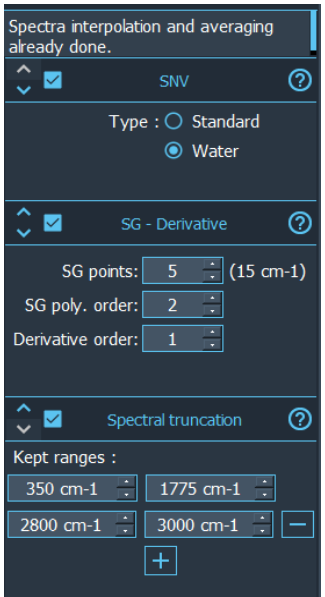
OPC connection to external software suits (SCADA, bioreactor software…)

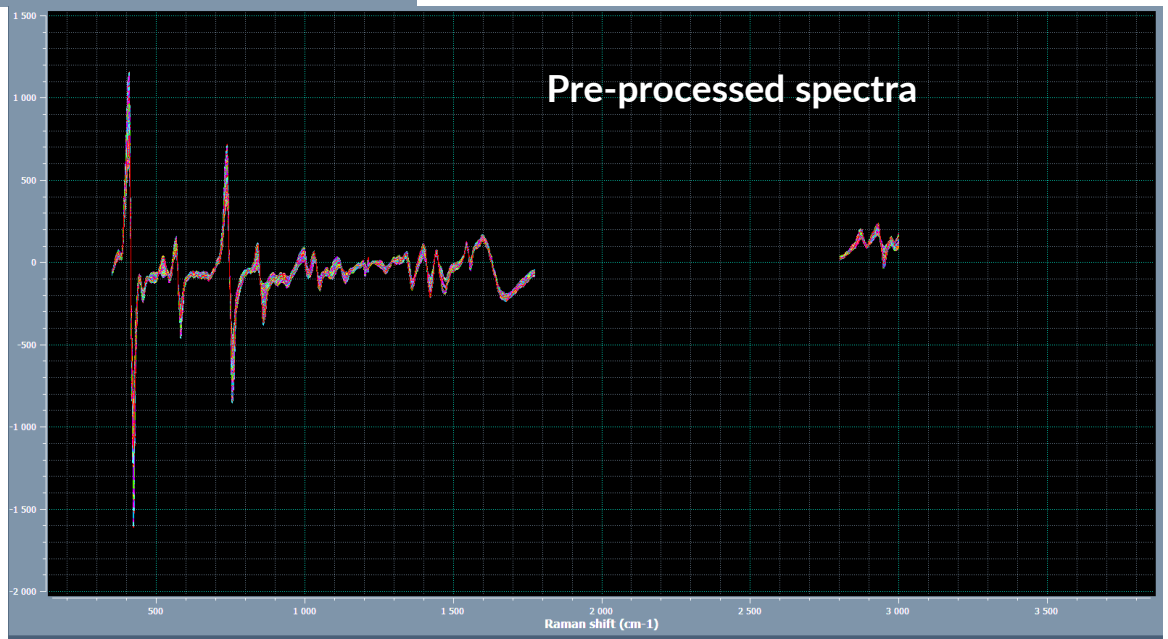
Example of cell culture monitoring: pre-processing steps



Experimental parameters

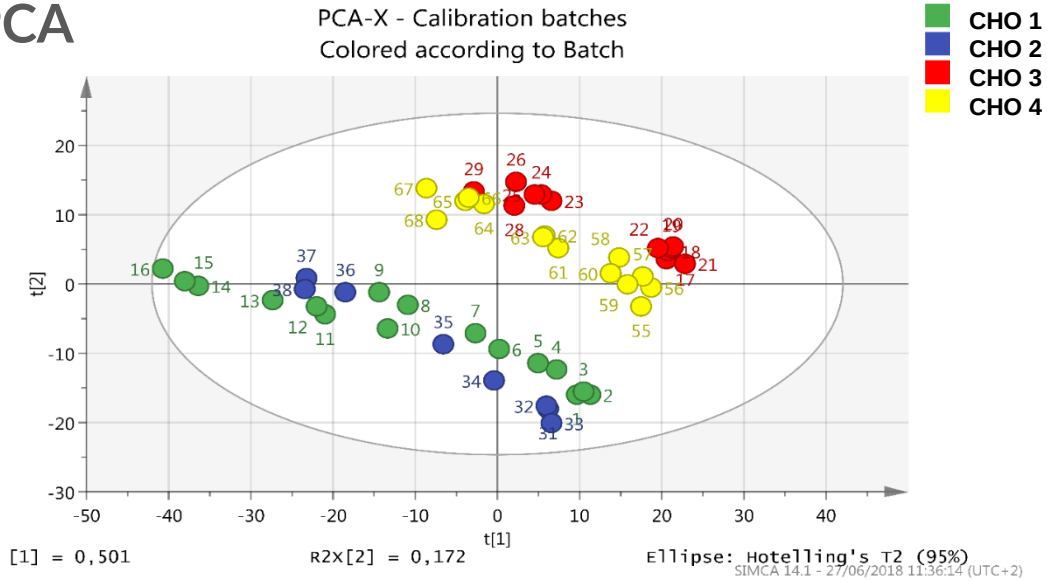
- CHO-S
- ProCho5 medium
- 5L Bioreactor
- Model trained on 4 batches
- Ref. Nova Flex 2 / Luna cell counter





Example of cell culture monitoring: model building

PCA



Dataset characteristics

- 4 batches
- 51 points
- Biological / experimental variability presence

PLS

Analyzed	#factor	R^2	Q^2	Range
Total Cell Density (TCD)	7	0,981	0,945	2 – 40 e5 cells/mL
Viable Cell Density (VCD)	8	0,989	0,942	1,4 – 33 e5 cells/mL
Ammonium	5	0,976	0,961	0 – 8,8 mM
Glucose	5	0,992	0,983	2,7 – 10 g/L
Lactate	5	0,994	0,989	0 - 52,2 mM
Glutamine	8	0,986	0,899	0,6 – 10,6 mM

Example of cell culture monitoring: in-line monitoring display

Analyzed	Total Cell Density (TCD)	Viable Cell Density (VCD)	Ammonium	Glucose	Lactate	Glutamine
Average % error	12	11	11	3	7	10

OPC connection to external Software (SCADA, bioreactor software…)

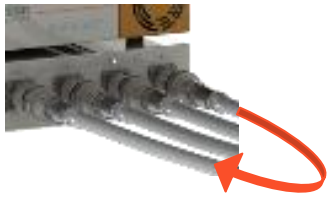
Challenge #4: managing process and setup variations

Raman for Bioprocessing Workshop

resolution spectra systems

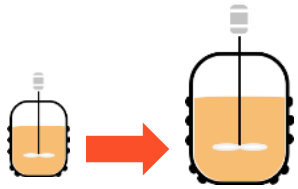
Typical cases of new variabilities (process + measurement)

Switch from ProCellics™ Multi-Channel Unit to ProCellics™ single channel

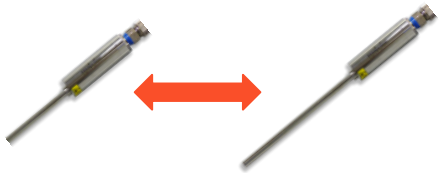


Switch from one probe/channel to another with ProCellics™ Multi-Channel Unit

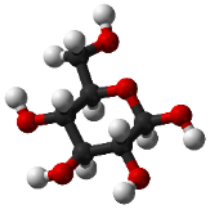
ProCellics™ instrumental transferability



Scale-up



Switch from one probe tube to another

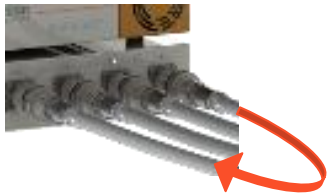


New min or max concentration of a given metabolite / parameter (pH, gas rate…)

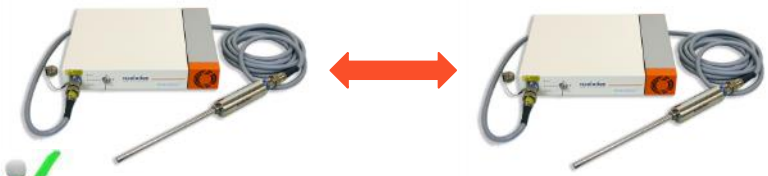
Our chemometric solutions

Variabilities exist and are manageable

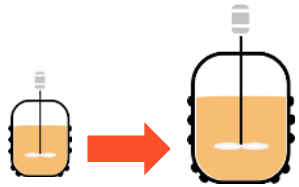
Switch from ProCellics™ Multi-Channel Unit to ProCellics™ single channel



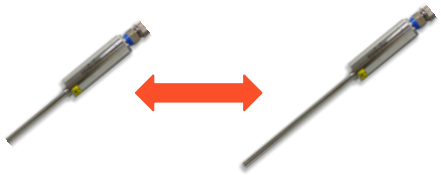
Switch from one probe/channel to another with ProCellics™ Multi-Channel Unit



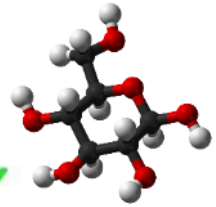
ProCellics™ instrumental transferability



Scale-up

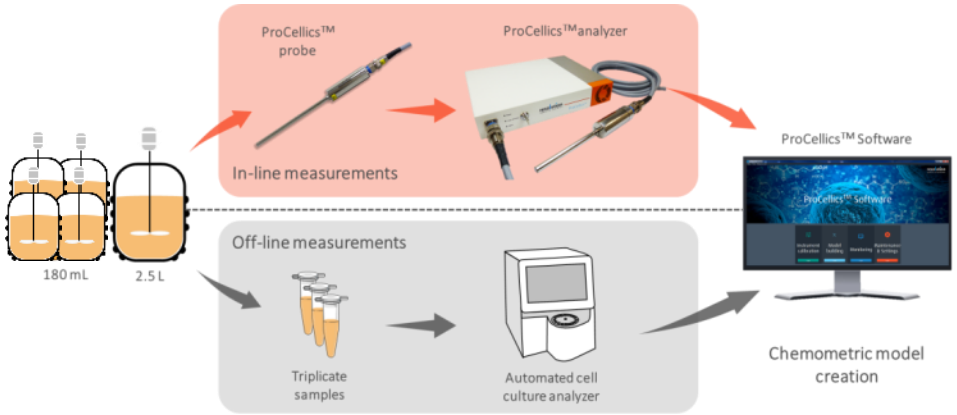


Switch from one probe tube to another

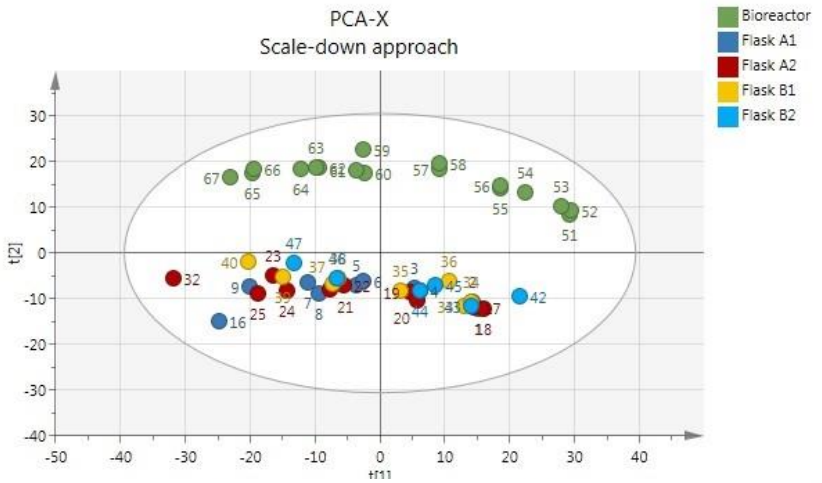


New min or max concentration of a given metabolite / parameter (pH, gas rate…)

Chemometric solutions: scale-up, new concentration



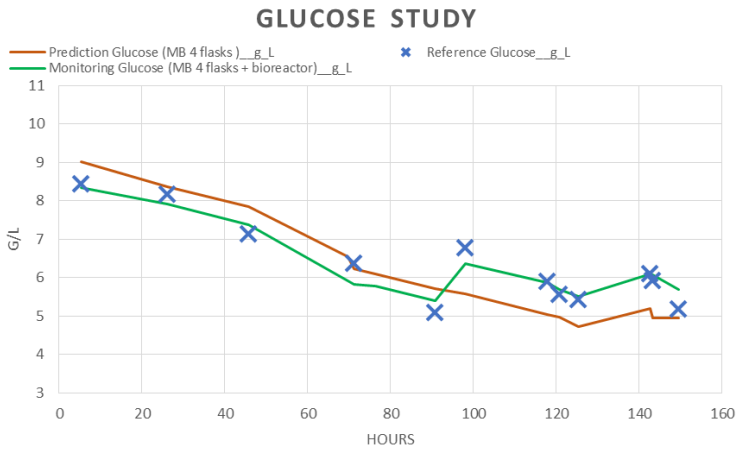
See poster



The variability between scales is visible and limited, but needs to be taken into account

	Model building on 4 flasks only		Model building on 4 flasks + scale-up bioreactor	
	Prediction RMSEP	Error*	Real-time monitoring RMSEP	Error *
TCD (10^6 cells/mL)	1.97	32 %	0.41	6 %
VCD (10^6 cells/mL)	1.64	27 %	0.63	10 %
Glucose (g/L)	0.67	8 %	0.38	5 %
Glutamine (mM)	3.30	33 %	1.34	13 %

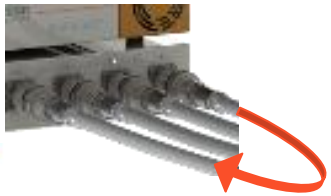
* Real-time monitoring RMSEP as percentage of maximum process value



Our instrumental solutions

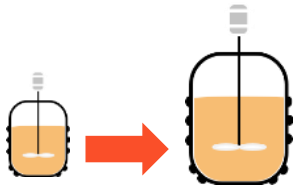
Variabilities exist and are manageable

Switch from ProCellics™ Multi-Channel Unit to ProCellics™ single channel

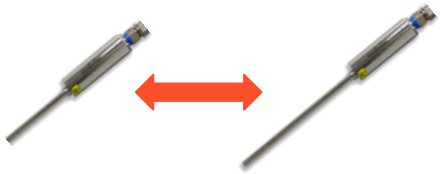


Switch from one probe/channel to another with ProCellics™ Multi-Channel Unit

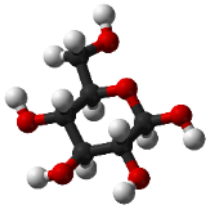
ProCellics™ instrumental transferability



Scale-up

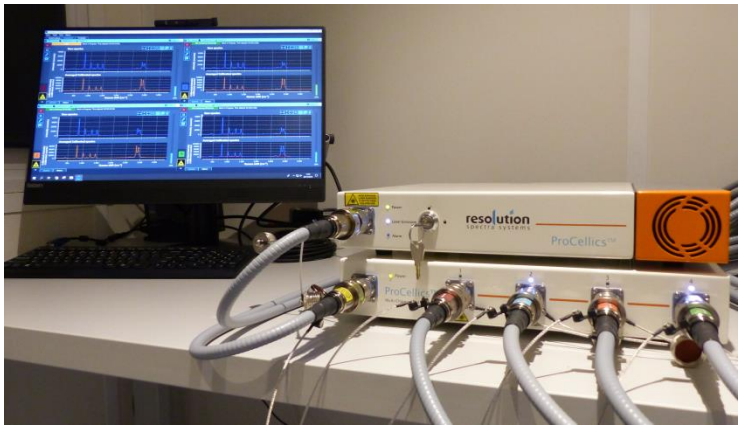


Switch from one probe tube to another

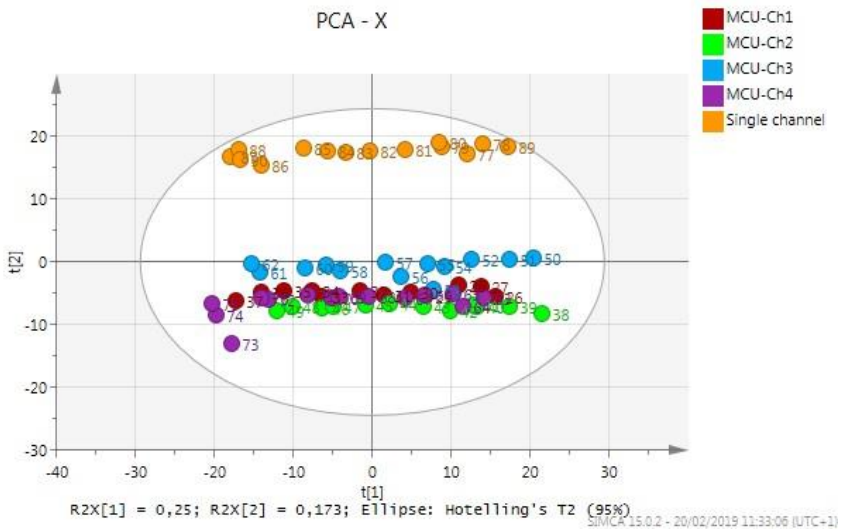


New min or max concentration of a given metabolite / parameter (pH, gas rate…)

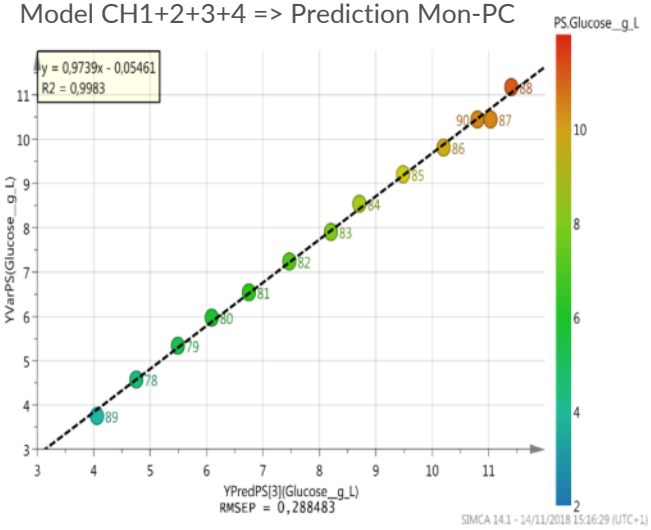
Instrumental solutions: ProCellics™ Multi-Channel Unit



Normalized 1st derivative spectra



Model CH1+2+3+4 => Prediction Mon-PC



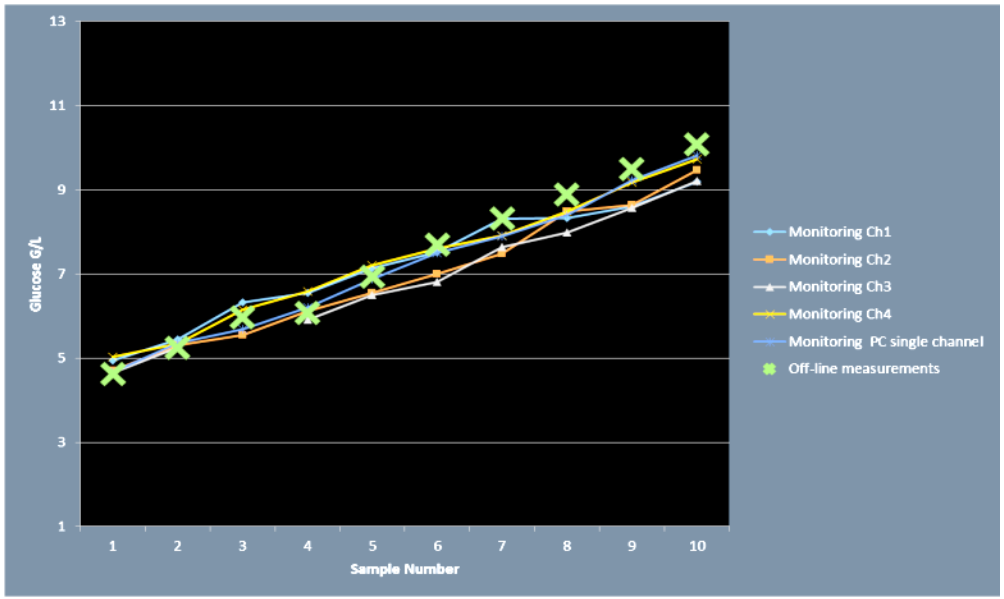
Average % error: 4 %

- Little variability between channels
- Successful model transfer to a single channel system

Instrumental solutions: switch from one probe/channel to another

Model Building with CH1	Range	Real-time monitoring RMSEP	Error*
Glucose channel 1 (control)		0.49	5 %
Glucose channel 2	4.62-10.08 g/L	0.53	5 %
Glucose channel 3		0.62	6 %
Glucose channel 4		0.33	3 %

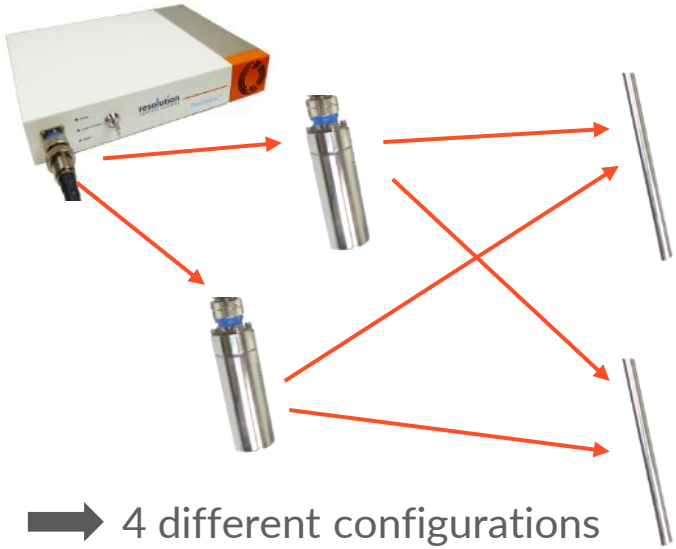
*Real-time monitoring RMSEP as percentage of maximum process value



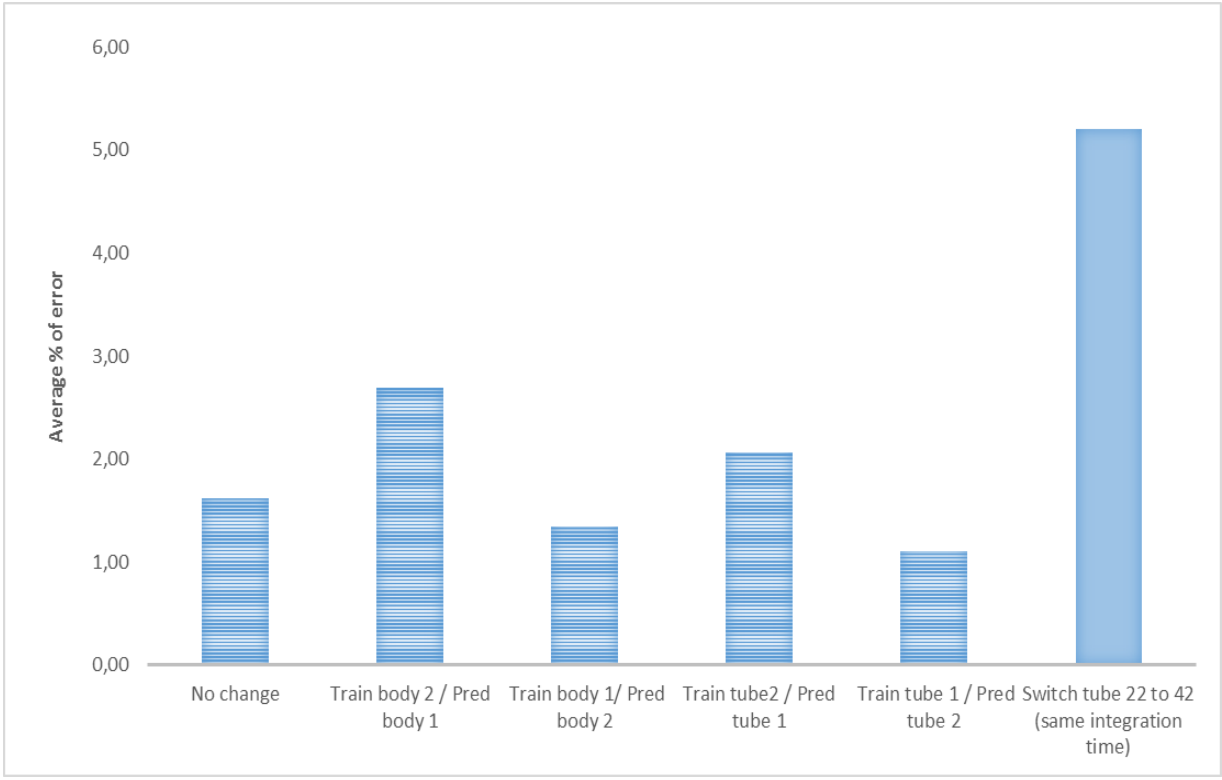
Little variability between the different channels of ProCellics™ Multi-Channel Unit was observed

Instrumental solutions: tube/probe change

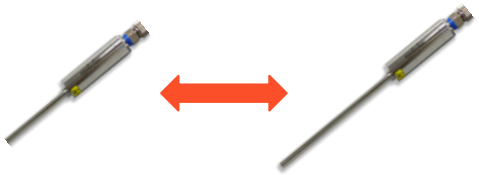
- Switch of tubes and probes



➡ 4 different configurations



- Switch from 22 cm tube to 42 cm tube

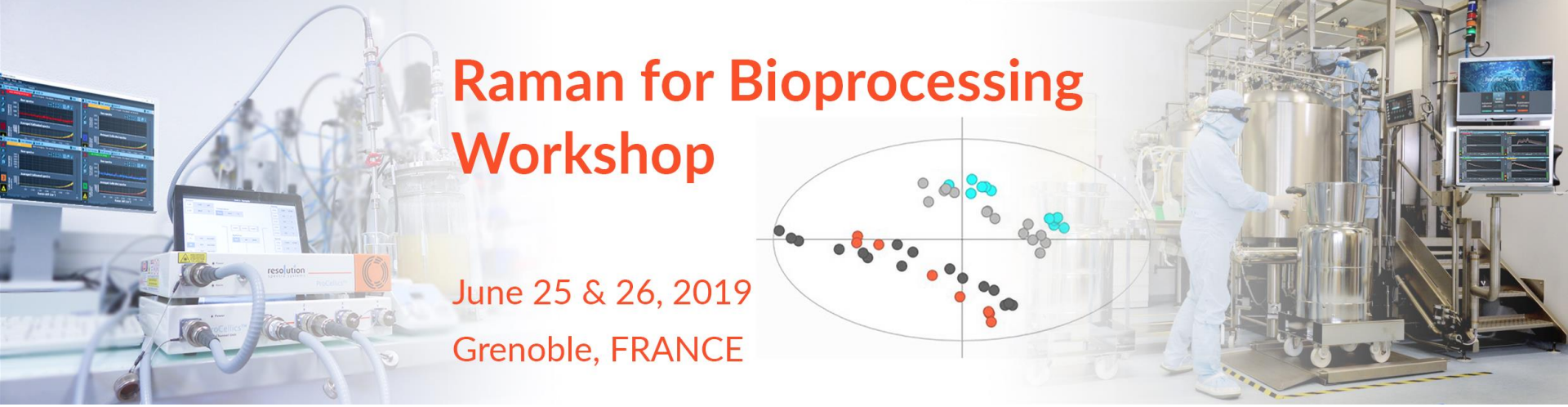


- Limited impact on predictions due to probe or tube (same size) hardware variability
- Switch from tube 22 cm to 42 cm brought variability: change of integration time must be considered

Conclusion

- Chemometrics is not rocket science!
- Chemometrics allows to deeply understand and analyze your process!
- Challenges and traps can be easily avoided with basic knowledge. Don't be scared to try!
- Instrumental and process variability must be considered carefully but are easily manageable!

Models Bioprocess Data Management
Speed **Efficient prediction** Pipeline
It's not rocket science! Real-Time Chemometrics



Thank you!

www.resolutionspectra.com

resolution
spectra systems