UNIVERSITÄT
HOHENHEIM

iConsensus WP5: Spectroscopy and chemometrics support

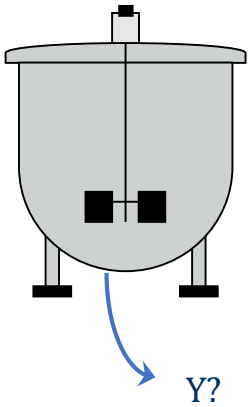
Bioprocess monitoring with Raman spectra Generic data driven models?

O. Paquet-Durand, A. Yousefi-Darani, B. Hitzmann

Department of process analytics und cereal science, Institute for biotechnology and food science, **University of Hohenheim**, Garbenstr. 23, 70599 Stuttgart, Germany

Chemometrics

(and other "…ometrics" such as biometrics, econometrics, psychometrics, …)



Problem:

System or (bio)process with **interesting properties/variables** (Y) that are **not easily accessible**.

Idea:

Use some sort of mathematical model (f) to **calculate** the **interesting part** (Y) **from** other process information that is more **easily accessible** (X).

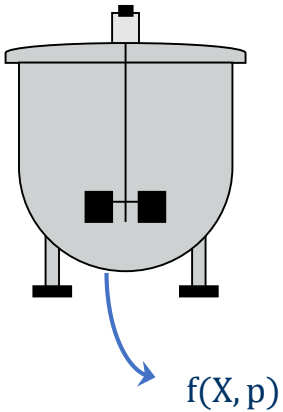
$$Y = f(X, p)$$

Easily available process information (X); wish list:

Easy and (relatively) cheap to measure
Fast, real-time capable measurement
Non-invasive/non-intrusive measurement
…

} Spectroscopy!

Chemometrics



$$Y = f(X, p)$$

What about p?:

Perform experiments; measure spectra (X) as well as difficult to obtain target information (Y)

Calibrate the model = determine p

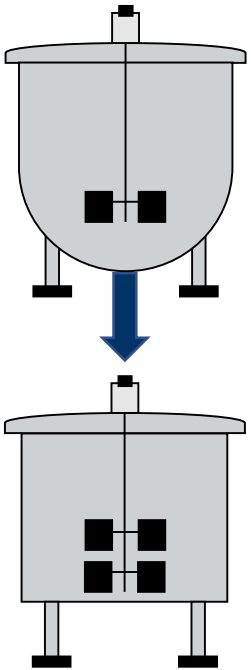
What about f? Which model to chose?:

Keep it simple! (as simple as possible, as complex as necessary to do the job)

Complex models → not necessary or not possible (lack of sufficient data)

Partial least squares regression (PLS-R) with basic pre-processing is fine in most cases!

“Generic” chemometric model?



Key challenge:

Genericity of data driven statistical models is usually very bad.

Small change in the system/(bio)process

→ chemometric models invalidated! New calibration is necessary!

High calibration effort → it might be easier to:

- Measure the variable of interest with the old offline method!
- Ignore the variable if possible!

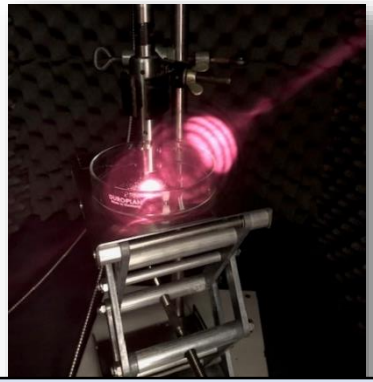
Verdict: Spectroscopy based data-driven models can be awesome, if the work properly **but** they can be very difficult to use in a lot of cases.

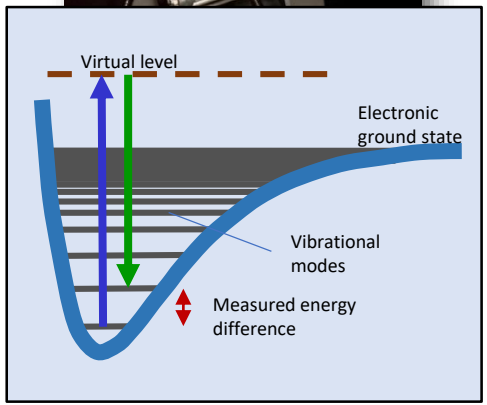
Solutions?

Trade accuracy for reliability/flexibility

- Use the right spectroscopic tool!
  - Avoid “indirect” measurements!
  - For chemical composition, Raman or TF-MIR is a good start
- Calibration dataset: Have as much variability as possible!
  - User different processes and hardware to collect data
  - Collect data from multiple sites
- Keep the pre-processing to a minimum!
  - Data alignment is necessary (probably)
  - Everything else is optional (depends on the actual data and process)
- Use simple regression model!
  - No need for fancy models here. Simple PLS-R is perfectly adequate.
  - The less parameters, the better!

Raman spectroscopy





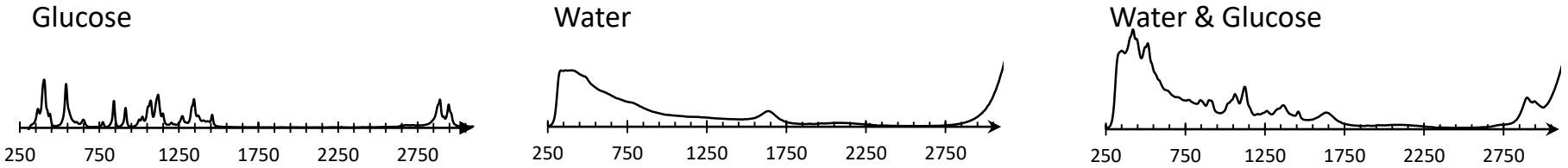
Raman scattering is caused by excitation of vibrational modes in the sample molecules

→Vibrational spectroscopy

Measured energy differences are ~ 0.05 – 0.5 eV

Comparable to MIR spectroscopy (sort of) but shifted to higher energies → water absorption of MIR is no problem in Raman spectroscopy

Raman spectroscopy



Raman effect is linear in nature! A Raman spectrum of a sample is pretty much the sum of all the spectra of the individual compounds in the sample!

→Raman spectra can be used to determine the composition of a mixture different compounds!

But there is **fluorescence!**

Very strong in biological systems, which can make the evaluation of Raman spectra very difficult.

Calibration

- Capture as much variability as possible in the calibration data set!
- Measure Raman spectra as well as target variables/compounds!
- Vary process conditions!
  - Different environmental conditions
  - Different feed strategies
  - Different cell lines
- Measure at different sites (if possible)!
- Use different spectroscopic hardware (if available)!

Calibration data

iConsensus → access to large amounts of process data from 4 companies (Bayer AG, GSK, Sanofi, Rentschler Biopharma)

We picked a small subset (most recent):

- 41 different cultivation batches in total.
- 1699 data points (Raman spectrum & off-line measurement).
- Raman spectra acquired with different spectrometers (Kaiser RXN2, Resolution ProCellics Raman analyzer).

Validation data

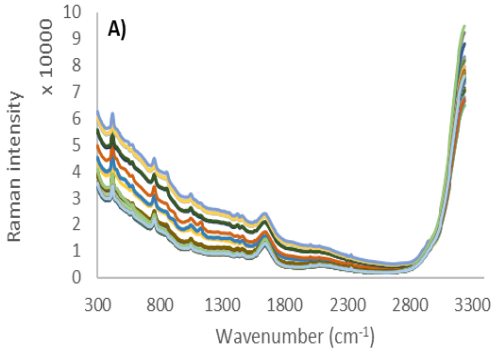
1. Validation data that is **similar** to calibration/training data:

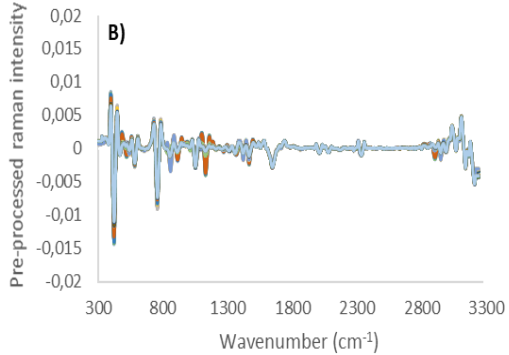
 > 10.000 data points from iConsensus.

2. Validation data that is **different** to calibration/training data:

- 240 data points form dilution series of major compounds in water and FMX-8 medium
- 550 data points from test cultivations at Royal Institute of Technology in Stockholm, group of Veronique Chotteau

Preprocessing of spectra





Keep the pre-processing to a minimum!

1. Plausibility checking/outlier detection and data alignment:
   - Interpolation
   - Intensity correction

2. Baseline/fluorescence removal:
   - High pass filter
   - First derivative Savitzky-Golay filter

3. Normalization and removal of any remaining linear offset:
   - SNV normalization

Chemometric model

$$Y = L_0 + L_1x_1 + L_2x_2 + L_3x_3 + \cdots + L_nx_n$$

$$Y = f(X, p)$$

Spectra → X:
Pre-processed Raman spectra, 1001 channels

Target variables → Y:
glucose, lactate, glutamine and glutamate

Model f:

Type of regression model	Validation data	
	Similar to calibration set	Independent data set
Neural networks (MLP, CNN,…)	Very good ✓	Very bad ✗
Gaussian process regression	Good ✓	Bad ✗
Linear regression (MLR, PCR, PLSR,…)	Reasonable ✓	Reasonable ✓

"accuracy" decreases

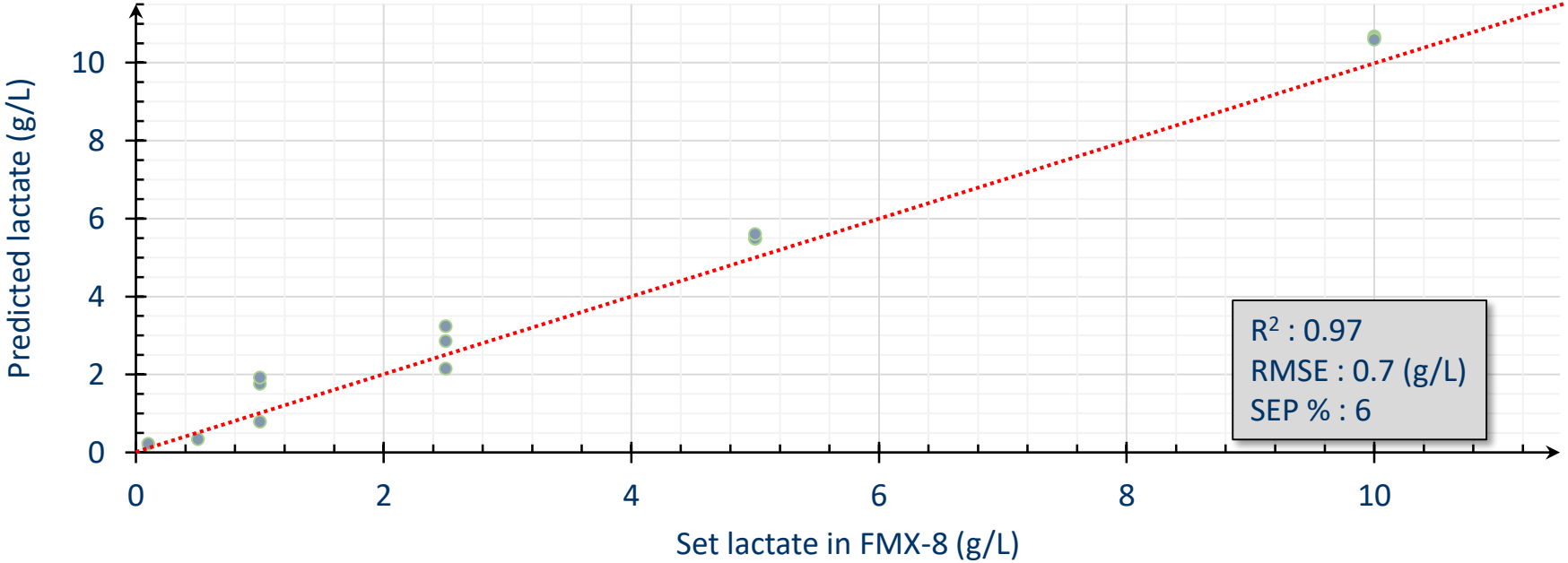
"flexibility" increases

✓ = "good enough"

Results

Independent dataset 1:

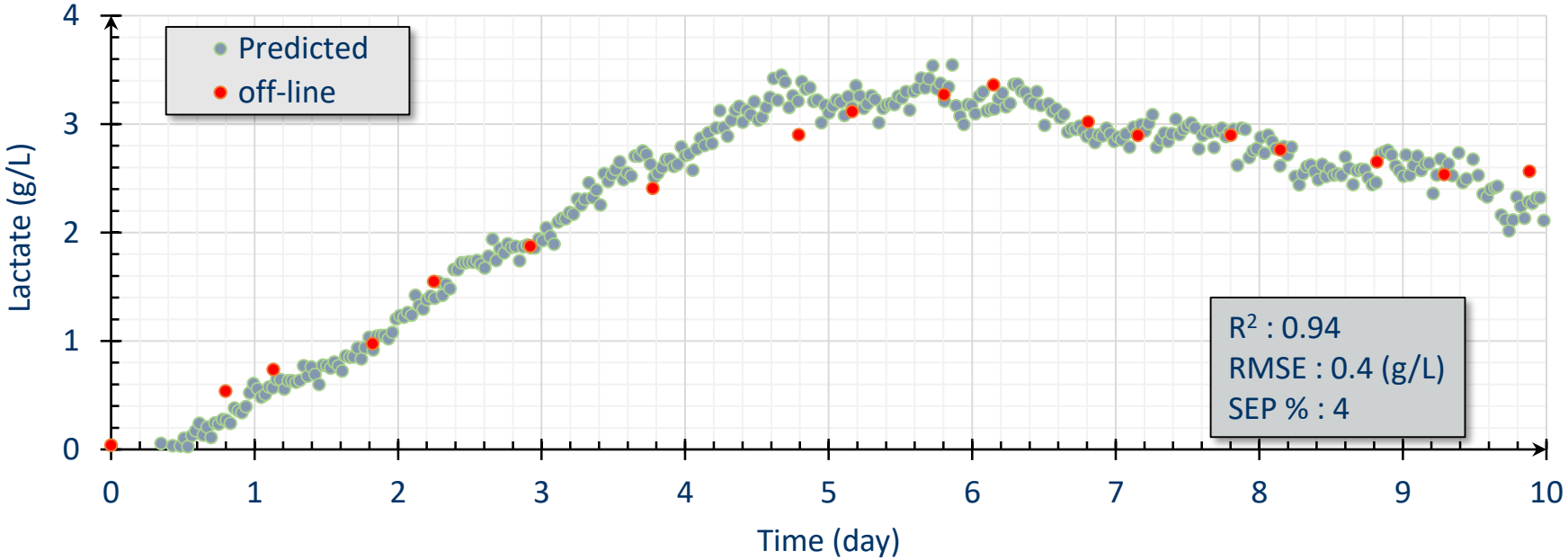
- From University of Hohenheim
- Spectrometer: Tec5 Raman 785 with modified InPhotonics RPS785/FF12 probe
- Dilution series of various compounds in water and FMX-8 mod medium



Results

Independent dataset 2:

- From Royal Institute of Technology in Stockholm
- Spectrometer: Kaiser RXN2 Raman analyzer
- Fed batch cultivation of CHO TurboCell (TM)



Conclusion

If we are willing and able to trade accuracy for reliability/flexibility:

Generic and flexible data driven **models are possible**!

But:
Higher effort for calibration required!
Generic models are often less accurate (but can still be “good enough”)!

Discussion / Questions?

Acknowledgements

The authors acknowledge the Innovative Medicines Initiative 2 Joint Undertaking [grant agreement No 777397] for funding this research. This Joint Undertaking, project iConsensus, receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA Partners Sanofi, GSK, Bayer, Rentschler Biopharma, UCB, Byondis and Pfizer. Rentschler Biopharma has kindly provided the cell line TurboCell producing Rituximab antibody. FudjiFilm Irvine Scientific (CA, USA) kindly provided most of the components for the media BM/FM.

Special thanks to:

J. Traenkle & J. Claasen from **Bayer AG**

M. Mertens & J. Snelders from **Sanofi**

A. Handl, M. Kadisch & D. Lang from **Rentschler Biopharma SE**

P. Dumas from **GlaxoSmithKline**

M.E.L. Makinen & V. Chotteau from **KTH Royal Institute of Technology Stockholm**