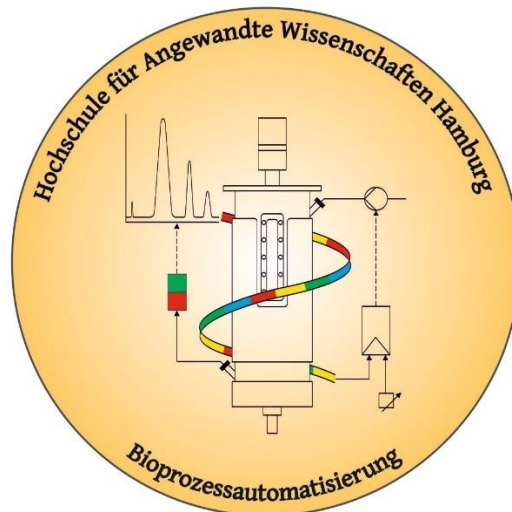


**Investigations in Raman Spectroscopy as a PAT Tool Applied In
Cultivations with *Pichia pastoris***




Master Thesis

Master's Degree M. Sc. Pharmaceutical Biotechnology

Phoebe Chan

Hamburg



First Examiner: Prof. Dr. Gesine Cornelissen (HAW Hamburg)

Second Examiner: Prof. Dr. Christian Kaiser (HAW Hamburg)

Hamburg University of Applied Sciences
Faculty of Life Sciences
Department of Biotechnology
Ulmenliet 20
21033 Hamburg

Author:
Phoebe Chan

██
██

First examiner: Prof. Dr. Gesine Cornelissen (HAW Hamburg)
Second examiner: Prof. Dr. Christian Kaiser (HAW Hamburg)

Declaration of Authorship

Hereby, I declare that I have composed the presented paper independently on my own and without any other resources than the ones indicated. All thoughts taken directly or indirectly from external sources are properly denoted as such.

Place, date

Phoebe Chan

Abstract

The present work investigated the applicability of in-line and off-line Raman spectroscopy for the quantification of glycerol, methanol, cell density, and total protein in fed-batch cultivations of *Pichia pastoris*. For assessment of the multivariate Raman spectra, data pre-processing and orthogonal projections to latent structures were applied. In the context of this work, the optimal pre-processing tools were baseline correction by 1st derivation and linear correction, scatter correction by standard normal variate, and noise removal by Savitzky-Golay filter with 9 points in each moving polynomial. The models yielded root mean square errors of cross-validation of 13.7 %, 13.3 %, 6.24 %, and 11.5 % for glycerol, methanol, cell density, and total protein, respectively. The lowest cross-validation errors for in-line as well as off-line were achieved with prediction of cell concentration. In contrast, the highest error was obtained by glycerol prediction with 18.8 % by in-line Raman spectroscopy. Overall, the in-line Raman probe demonstrated more lower cross-validation errors than off-line Raman spectroscopy. Calibration models were successfully developed for all analytes.

Table of Contents

Abstract	II
Table of Contents	III
List of Figures	V
List of Tables.....	VII
Nomenclature	VIII
1 Introduction	1
2 Objective	2
3 Theoretical Background	3
3.1 Process Analytical Tools	3
3.2 Raman Spectroscopy	4
3.2.1 Principle and Instrumentation.....	5
3.2.2 Applicational Fields.....	7
3.3 Cultivation Strategy.....	8
3.4 Multivariate Data Analysis.....	9
3.4.1 From Data to Information.....	9
3.4.2 Approach to Data Analysis.....	11
3.4.3 Data Preparation	13
3.4.4 Data Pre-Processing.....	13
3.4.5 Principal Component Analysis	19
3.4.6 Projections to Latent Structures.....	23
3.4.7 Orthogonal Projections to Latent Structures	35
4 Material and Methods.....	37
4.1 Cell Line	37
4.2 Medium	37
4.3 Preculture.....	39
4.4 Bioreactor System BIOSTAT® C30.....	39
4.4.1 Cultivation Conditions.....	41
4.4.2 Standard Measurement Systems	43
4.4.3 MFCS/win	44
4.4.4 Turbidity Probe.....	44
4.4.5 Methanol Probe	45
4.5 Raman Spectroscopy	46
4.6 Analytical Methods	47
4.6.1 Cell Dry Weight Concentration.....	47
4.6.2 Optical Density Determination.....	48

4.6.3	Off-line HPLC for Glycerol and Methanol Determination	48
4.6.4	Total Protein Concentration Determination	49
4.6.5	Fluorescence Determination.....	49
4.7	Data Evaluation with MVDA.....	50
4.7.1	Sample Pool.....	50
4.7.2	Data Preparation	50
4.7.3	Data Pre-Processing.....	50
4.7.4	Approach in SIMCA® Environment.....	51
4.7.5	Preliminary Studies on Data Pre-Processing Methods	54
5	Results and Discussion.....	57
5.1	Exemplary Development of an OPLS Model for Glycerol	57
5.1.1	Principal Component Analysis for Outlier Detection.....	59
5.1.2	Multivariate Calibration with OPLS.....	62
5.1.3	Prediction and Validation	65
5.2	Preliminary Studies on Pre-Processing Methods.....	68
5.3	Investigated Process Variables	71
5.4	Prediction of Glycerol Concentration.....	72
5.5	Prediction of Methanol Concentration	75
5.6	Prediction of Cell Concentration	78
5.7	Prediction of Total Protein Concentration.....	81
5.8	Sources of Error in Methodology	83
5.8.1	Critical View on Executed MVDA	84
5.8.2	Cultivation XXPC2622	85
5.9	Process Control and Parameters	86
6	Conclusions	89
7	Future Perspectives.....	90
8	References	91
	Appendix	96

List of Figures

Figure 3.1: Steps for implementation of process analytical technology (Rathore et al., 2010).....	4
Figure 3.2: Energy diagram of vibrational transitions between vibrational energy levels.	5
Figure 3.3: Generalised overview of instrumentation within a dispersive Raman spectroscopy system (modified after Butler et al., 2016).....	6
Figure 3.4: Schematic representation of configurations in Raman spectroscopy after sample excitation.	7
Figure 3.5: Schematic course of three-stage cultivation.....	8
Figure 3.6: MVDA toolbox to solve different data analytical questions.....	10
Figure 3.7: Typical pipeline for chemometric data analysis comprising design, performance, and analysis of experiments.	12
Figure 3.8: Data table with n observations and m variables.....	13
Figure 3.9: Pipeline for pre-processing of spectral data.	13
Figure 3.10: Different pre-processing methods of Raman spectra.	16
Figure 3.11: Graphical representation of PCA.	20
Figure 3.12: Schematic representation of projections to latent structures and involved matrices.....	24
Figure 3.13: Exemplary scores scatter plot with applying Hotelling's T^2 test and a confidence interval of 95 %.	28
Figure 3.14: Visualisation of an observation's distance to model X (DModX) in plane of original data set X.....	30
Figure 3.15: Balance between goodness of fit R^2Y and goodness of prediction Q^2	33
Figure 3.16: Permutation plot for validation and overfit detection.	34
Figure 3.17: Overview of orthogonal projections to latent structures (modified after Trygg & Wold, 2002).....	35
Figure 3.18: Schematic representation of orthogonal projections to latent structures (OPLS).	36
Figure 4.1: Bioreactor BIOSTAT [®] C30 and peripherals.....	40
Figure 4.2: Simplified piping and instrumentation diagram and automation tasks of bioreactor system used.....	41
Figure 4.3: Used turbidity measurement system.	44
Figure 4.4: Used methanol content measurement system.	45
Figure 4.5: Used Raman spectrometer system.	46
Figure 4.6: Flow-chart of multivariate data analytics in SIMCA [®] environment.....	52
Figure 4.7: Graphical user interface of SIMCA [®] in the version 17.0.1.....	53
Figure 4.8: Graphical user interface of <i>Calibration wizard</i> in SIMCA 17.0.1 [®]	54
Figure 5.1: Exemplary off-line Raman spectra of supernatant in cultivations XXPC0922 and XXPC2622 with highlighted observation 9 (yellow).....	58
Figure 5.2: Exemplary off-line Raman spectra for cell suspension of cultivation XXPC0922 and XXPC2622.	59
Figure 5.3: Exemplary summary of fit for PCA model with Autofit or Two first.	60
Figure 5.4: Exemplary scores scatter plots with Hotelling's T^2 test for PCA model ($\alpha = 0.05$).	60
Figure 5.5: Exemplary score contribution plot of observation 9 in a PCA model with $r = 2$ PCs.	61
Figure 5.6: Final scatter plots after exclusion of moderate or high outliers.	62
Figure 5.7: Exemplary prediction error RMSE _{cv} of OPLS model dependent on number of components r	63
Figure 5.8: Exemplary permutation plot for OPLS model with $r = 1 + 4$ components after 100 permutations.	63

Figure 5.9: Exemplary variable importance in projection VIP of 1 + 3 OPLS model.	64
Figure 5.10: Exemplary regression line for cross-validation with cross-validated values plotted against reference for glycerol concentration c_{S1M}	65
Figure 5.11: Exemplary regression line for internal validation with predicted validation set plotted against reference for glycerol concentration c_{S1M}	66
Figure 5.12: Exemplary regression line for external validation set with prediction set plotted against reference for glycerol concentration c_{S1M}	66
Figure 5.13: Exemplary prediction of glycerol concentration c_{S1} in supernatant with external validation set XXPC1722.	67
Figure 5.14: Off-line spectra of XXPC1722 suspension coloured according to process time $t_{process}$	70
Figure 5.15: Course of off-line reference measurement for cultivation XXPC0922.	71
Figure 5.16: Off-line Raman spectra of calibration set for glycerol determination in cell suspension.	72
Figure 5.17: Variable importance in projection (VIP) of models N8 and N12.	73
Figure 5.18: Prediction of glycerol concentration with different measuring types.	74
Figure 5.19: OPLS model parameters of in-line Raman cell suspension for prediction of methanol of model N2.	75
Figure 5.20: Predicted methanol concentration c_{S2M}	77
Figure 5.21: In-line Raman spectra during cultivation XXPC0922.	78
Figure 5.22: OPLS model parameters of in-line Raman cell suspension for prediction of cell density of model N8.	79
Figure 5.23: Predicted cell concentration c_{XL} in cell suspension for N8 by cross-validation using off-line and in-line Raman spectroscopy.	80
Figure 5.24: OPLS model parameters of in-line Raman suspension for prediction of total protein c_{PtotM} of model N11.	82
Figure 5.25: Predicted total protein concentration c_{PtotM} by cross-validation using off-line and in-line Raman spectroscopy.	83
Figure 5.26: Course of cultivation XXPC2622.	86
Figure 5.27: On-line estimation of cell density in cultivation vessel.	87

List of Tables

Table 4.1: Medium composition of FM22.	37
Table 4.2: Composition of trace element solution PTM ₄ stock.....	38
Table 4.3: Reservoirs for cultivation.	38
Table 4.4: Cultivation Conditions.	43
Table 4.5: Process parameters and corresponding measurement systems.....	43
Table 4.6: Overview about pre-processing tools used in the four pre-processing steps.....	51
Table 4.7: Overview about pre-processing methods offered in SIMCA® 17.0.1.....	51
Table 4.8: Combining pre-processing tools into pre-processing methods with one to four pre-processing steps.....	55
Table 4.9: Wavenumber ranges investigated in preliminary studies.....	55
Table 4.10: Criteria and assigned scores for weighted sum model.	56
Table 5.1: Results of preliminary studies for wavenumber range C that were applied onto Raman spectra.....	69
Table 5.2: Overview of pre-processing methods that were applied for upcoming multivariate calibration.....	70
Table 5.3: Summary of predicted glycerol concentration in cell suspension (SUS) and supernatant (SN) with both off-line and in-line Raman spectroscopy.....	73
Table 5.4: Summary of predicted methanol concentration c_{S2M} with both off-line and in-line Raman spectroscopy.....	76
Table 5.5: Summary of predicted cell concentration in cell suspension (SUS) c_{XL} with both off-line and in-line Raman spectroscopy.....	80
Table 5.6: Summary of predicted total protein concentration c_{PtotM} with both off-line and in-line Raman spectroscopy.....	82
Table 5.7: Parameters for cell density estimation with the turbidity probe.....	88

Nomenclature

General Abbreviations

AUC	Area under curve
BSA	Bovine serum albumin
CCD	Charge-coupled device
CDW	Cell dry weight
CPP	Critical process parameters
CQA	Critical quality attributes
DW	Demineralised water
EMA	European Medicines Agency
FDA	United States Food and Drug Administration
FMC	Fermentation Mini Computer
FT	Fourier transformation
HPLC	High-Performance Liquid Chromatography
ICH	International Conference on Harmonisation
ISA	International Society of Automation
MVDA	Multivariate data analysis
NIR	Near infrared
PAT	Process analytical technology
RI	Refractive index
SCADA	Supervisory control and data acquisition
SN	Supernatant
SORS	Spatially-offset Raman spectroscopy
SUS	Suspension

General Variables

λ	Wavelength	nm
ϑ_K	Temperature in subsystem K	°C
μ	Cell-specific growth rate	h ⁻¹
a	Adaption parameter for turbidity measurement	g L ⁻¹
b	Adaption parameter for turbidity measurement	AU ⁻¹
c_{IK}	Mass concentration of component I in subsystem K	g L ⁻¹
D_F	Dilution factor	–
F_K	Volume flow rate from subsystem K	L h ⁻¹
F_{nI}	Aeration rate of component I under standard condition	L h ⁻¹
$K_{X/OD}$	Correlation factor between cell density and OD	–
m_{IK}	Mass of component I in subsystem K	g
N_{St}	Stirrer speed	min ⁻¹
p_K	Total pressure in subsystem K	bar
pO_2	Relative dissolved oxygen partial pressure in liquid phase	%
S_{turb}	Signal of turbidity measurement in liquid phase	AU
$t_{process}$	Process time	h
V_K	Volume of subsystem K	L
vvm	Volume of air sparged per unit volume of liquid phase per minute	–
x_I	Amount-of-substance fraction of component I in gas phase	–
$y_{X/I}$	Yield coefficient of biomass to component I	g g ⁻¹

General Indices

0	Initial condition (time point zero)
AF	Anti-foam
AIR	Air
CO ₂	Carbon dioxide
CDW	Cell dry weight
cal	Calibration condition
est	Estimated
ex	Excitation
G	Gas phase
i, j, k	Incremental variable
in	In-line
L	Liquid phase (cell suspension)
M	Media phase (supernatant)
max	Maximum
meas	Measured
min	Minimum
n	Standard gas conditions
O ₂	Oxygen
off	Off-line
P _{tot}	Total protein
R1	Reservoir 1 (glycerol)
R2	Reservoir 2 (methanol)
Ram	Raman spectroscopy
S1	Substrate 1 (glycerol)
S2	Substrate 2 (methanol)
T1	Titration 1 (acid)
T2	Titration 2 (base)
turb	Turbidity
w	Set point
X	Cell dry weight

MVDA-Related Abbreviations

1stDer	First derivative
2 nd Der	Second order derivative
CS	Calibration set
Ctr	Mean centring
LDA	Linear discriminant analysis
MA	Moving average
MANOVA	Multivariate analysis of variance
MBE	Mean bias error
MLR	Multiple linear regression
MSC	Multiplicative signal correction
NIPALS	Non-linear iterative partial least squares
OPLS	Orthogonal projections to latent structures
PC	Principal component
PCA	Principal component analysis
PLS	Projections to latent structures

PLSR	Partial least squares regression (= PLS)
PS	Prediction set
Q2	Goodness of prediction
R2X	Portion of model variance X
R2Y	Portion of model variance Y
RMSEP	Root mean square errors of predictions
SD	Standard deviation
SG	Savitzky-Golay
SIMCA	Soft independent modelling by class analogy
SNV	Standard normal variate
VIP	Variable importance in projection
VS	Validation set

MVDA-Related Variables

α	Level of significance
a_j	Filter coefficients for Savitzky-Golay filter
d_{ij}	Measured value of i-th row and j-th column
\bar{d}_i	Mean value of i-th row
\bar{d}_j	Mean value of j-th column
DModX	Distance to model X
e_{ij}	Element of residual matrix E in PCA
$F_{(\alpha,r,n-r)}$	Critical value of F-distribution with significance level α and r and n-r degrees of freedom
f_{ih}	Element of residual matrix F
J_l	Quality criterium of component l
l	PCA/OPLS component
m	Number of variables in row i
MBE	Mean bias error
n	Number of observations in column j
NORM	Sum of coefficients of Savitzky-Golay filter
r	Number of components/columns in score matrix T
r_0	For DModX _{ave} : 1 for centred models, otherwise 0
$r_{T,\alpha l}$	Radius of Hotelling's T ² ellipse
R2X	Portion of model variance X
R2Y	Portion of model variance Y
s_X^2	Total variance of data matrix X
$s_{t_l}^2$	Variance of Scores of component l
\bar{t}_l	Mean value of score vector t_l
v	Number of variables (columns) in data matrix Y
\bar{X}	Mean value of data matrix X
y_{ih}	Measurement value of target variable y_h for object i
\hat{y}_{ih}	Estimated value of target variable y_h for object i
\bar{y}_{ih}	Mean value of target variable y_h

MVDA-Related Matrices and Vectors

B	$(m \times v)$	PLS regression coefficient matrix
b_j	$(m \times 1)$	Column vector with PLS regression coefficients
D	$(n \times m)$	Data matrix (spectra)
d_j	$(n \times 1)$	Column vector of data matrix D
E	$(n \times m)$	Residual matrix of X-data space
F	$(n \times v)$	Residual matrix of Y-data space
G	$(n \times v)$	Residual matrix of PLS regression approach
P	$(m \times r)$	Loading matrix of X-data space
p_l	$(m \times 1)$	Loading column vector of component l (X-data)
Q	$(v \times r)$	Loading matrix of Y-data space
q_l	$(v \times 1)$	Loading column vector of component l (Y-data)
T	$(n \times r)$	Score matrix of X-data space
t_l	$(n \times 1)$	Score column vector of component l (Y-data)
U	$(n \times r)$	Score matrix of Y-data space
u_l	$(n \times 1)$	Score column vector of component l (Y-data)
W	$(m \times 1)$	Weight matrix of PLS model
w_l	$(m \times 1)$	Weight column vector of component l
X	$(n \times m)$	Modified data matrix
x_j	$(n \times 1)$	Column vector of data matrix X
Y	$(n \times v)$	Modified measurement matrix (responses in PLS)
y_h	$(n \times 1)$	Column vector of measurement matrix Y

MVDA-Related Indices

abs	Absolute
ave	Average
crit	Critical
CS	Calibration set
CV	Cross validation
i	Observation
j	X-variable
h	Y-variable
l	Principal or OPLS component
MS	Model data set
norm	Normalised
O	orthogonal
P, p	Prediction, predicted
PS	Prediction set
r	Number of principal components
rel	Relative
temp	Temporary
tot	Total
TS	Training data set
UV	Unit variance
VS	Validation data set

1 Introduction

Since the United States Food and Drug Administration (FDA) has outlined the process analytical technology (PAT) initiative in 2004, an innovative approach was established for pharmaceutical development, manufacturing, and quality assurance (European Medicines Agency, 2011; U.S. Department of Health and Human Services Food and Drug Administration, 2004). The framework aims to further improve process understanding and control, generating a greater probability for obtaining high-end product quality (Whelan et al., 2012). Simultaneously, bioprocesses thrive for more advanced, informative, and significant real-time data out of the system in order to control it (Abu-Absi et al., 2014). This requires process analysers, either by off-line, at-line, or in-line measurement approaches. For off-line measurements, the sample is removed and analysed in proximity to the process stream, while in-line measurements are performed within the process without sample removal. Vibrational spectroscopic methods, such as near-infrared (NIR) and Raman spectroscopy, have notably increased in applications as process analysers. Due to their ability of obtaining direct information, they proved to be a rapid and non-destructive analysis method (Nagy et al., 2018). NIR spectroscopy still dominates the pharmaceutical field. However, Raman spectroscopy is becoming more widespread and has demonstrated its feasibility during drug manufacturing processes (Buckley & Ryder, 2017; Goldrick et al., 2020).

Raman spectroscopy, which was first described in the 1920s, has emerged as a viable option for real-time bioreactor monitoring. This option brings measurements, previously limited to off-line analysis, to *in-situ*. The Raman spectroscopy possesses the capability to enlighten the complex bioreactor environment and to measure a number of chemical species simultaneously (Nagy et al., 2018).

2 Objective

Today, Raman spectroscopy is an established tool in bioprocess monitoring. However, before becoming an accessible tool, studies about the applicability of Raman spectroscopy on the used bioreactor system, cultivated strain, and media environment have to be made. In the present work, the potential of Raman spectroscopy for quantification of the compounds glycerol, methanol, cell density, and total protein content is demonstrated.

First, cultivations with the methylotrophic yeast *Pichia pastoris* are executed in a laboratory scale bioreactor. During cultivation, the cells are fed by glycerol, followed by a methanol-fed phase. Frequent probing of the cultivation ensures data mining of Raman spectra and reference values. Both off-line and in-line Raman spectroscopy are subject of this work.

Second, multivariate calibration models are developed for the investigated analytes. In the context of this work, optimal pre-processing methods are scrutinised in a weighted sum model. Principal component analysis is applied for outlier detection, while orthogonal projections to latent structures are employed for regression modelling. The obtained models allow an assessment of the Raman spectra and their potential in monitoring the cultivation of *Pichia pastoris* in a fed-batch culture.

3 Theoretical Background

The following chapter provides scientific and technical fundamentals in order to support the understanding of this work. First, the process analytical tools are further introduced before Raman spectroscopy, its function, and applications are elaborated. Then, the background about multivariate data analysis (MVDA) is thoroughly explained in order to ensure the comprehension of the upcoming results. The working cycle of MVDA is illustrated, pre-processing tools are presented. Finally, the MVDA tools for outlier detection and multivariate calibration are described.

3.1 Process Analytical Tools

The pharmaceutical industry is one of the most regulated industrial fields. Consequently, its production lines must be validated in advance to demonstrate their suitability for commercialisation and for safe human consumption or application (Sandell & Tougas, 2012). This, and the requirements of the industry have led to the emergence of the scientific discipline called PAT. The industry moved from quality-by-inspection to rather quality-by-design, implying full process understanding from development on (Whelan et al., 2012).

Since the FDA initiated the PAT approach, it also has been supported by the European Medicines Agency (EMA) and by the International Conference on Harmonization (ICH) (European Medicines Agency, 2011, 2017). The goal of PAT can be described as the design and development of a well-understood process, consistently ensuring a predefined quality at the end of the manufacturing process. Following criteria account to a well-understood process:

1. All critical sources of variability are both detected and explained,
2. Variability is handled by the process, and
3. Product quality attributes can be reliably and accurately predicted over an established design space for materials, process conditions, manufacturing, and environmental conditions (Rathore et al., 2010).

For implementation of PAT, three major steps can be defined (Figure 3.1). First, the design of the unit operation has to be made in the early phase of process development. Here, the critical process parameters (CPP) are determined which influence the identified critical quality attributes (CQA). Risk assessment tools of ICH guidelines are applied as the understanding of the process is the basis of the upcoming phases. Then, suitable analytical methods for real-time

monitoring of the CQA and CPP are identified (Rathore et al., 2010). These in-process analysers can be either on-line (sample is diverted from process and may be returned to process stream), in-line (sample is not removed from process and analyser can be invasive or non-invasive), at-line (sample is diverted or isolated from process and analysed in close proximity), or off-line (sample is removed from process stream and analysed afterwards) (Cervera et al., 2009; Rathore et al., 2010; Rathore & Gautam, 2009).

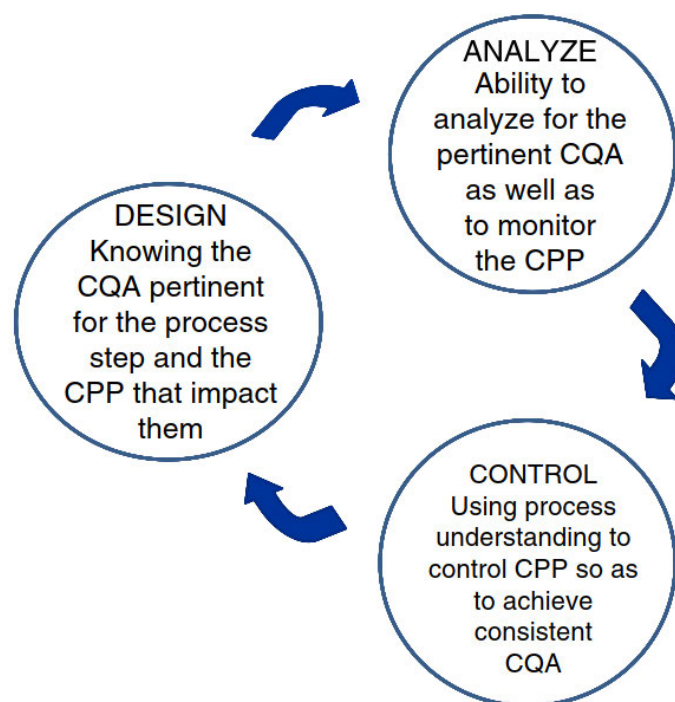


Figure 3.1: Steps for implementation of process analytical technology (Rathore et al., 2010).

For application of PAT, the analytical results have to be available within a time frame where real-time decisions are still possible. This enables a consistent process performance as well as a high product quality at the end of the manufacturing process. Afterwards, the design space has to be re-evaluated regarding its CQAs and CPPs (Rathore et al., 2010).

3.2 Raman Spectroscopy

Raman scattering was first discovered by Chandrasekhara Venkata Raman in 1928 (Raman & Krishnan, 1928). As an optical method, Raman enables non-destructive analysis of chemical composition and molecular structure. Raman applications began to grow in prominence with the advances of improved optics and smaller, more powerful lasers and detectors (Abu-Absi et al., 2014).

3.2.1 Principle and Instrumentation

Raman spectroscopy is a two-photon process, based on energy transfer between illuminated sample and irradiated light (Nagy et al., 2018). When a light beam impinges a sample, a photon is absorbed by the molecule and interacts with the chemical bonds within the molecule. This excites electrons to a higher energy level, denoted virtual state (Figure 3.2). The molecules return to the original energy level by emitting a photon, known as Rayleigh scattering, or it can undergo a shift of energy and return to lower (Stokes) or higher (anti-Stokes) energy state, known as inelastic Raman scattering. Although the predominant mode is elastic Rayleigh scattering, a small proportion of approximately one out of 10^9 photons is scattered inelastically (Wen, 2007). Depending on the spectrometer, either the loss (Stokes) or gain (anti-Stokes) of energy is measured (Butler et al., 2016). The wavelength shift of the scattered light depends on the chemical composition of the molecules. Also, the scattering intensity is proportional to the magnitude of changes in molecular vibrations and can be used for qualitative and quantitative analysis (Whelan et al., 2012).

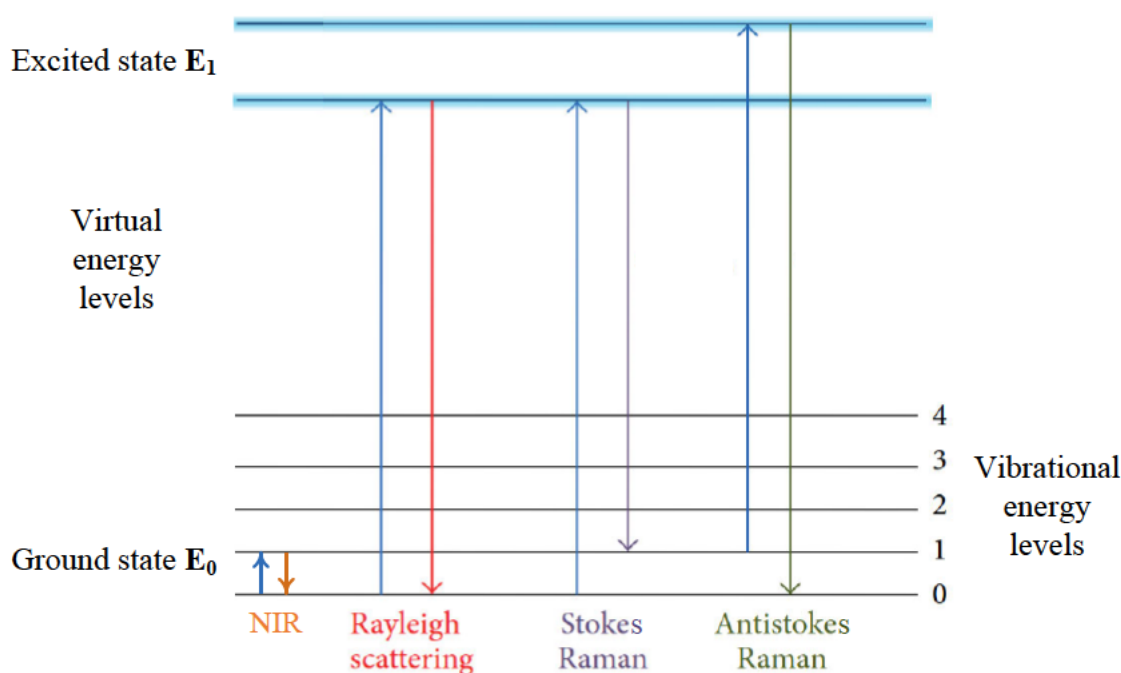


Figure 3.2: Energy diagram of vibrational transitions between vibrational energy levels. E_0 and E_1 are electronic ground and excited states, respectively. In the context of Raman scattering, three types occur after excitation: Rayleigh, Stokes, and anti-Stokes. As a comparison, near-infrared (NIR) energy state is depicted as well (modified after (Kim et al., 2021; Wen, 2007)).

The Raman spectrometer comprises a light source, a filter, and a detector (Figure 3.3). The excitation source can either be gas-based (488 and 514 nm), diode (630 and 780 nm), or helium neon lasers (632 nm), with many more types available (Rostron & Gerber, 2016). NIR lasers at

785 nm and 830 nm have been widely used in biological studies as these lasers have relatively low photon energy and do not cause serious photodamage on viable cells (Shipp et al., 2017). As Rayleigh scattering is more intense than Raman scattering, an optical filter is required to pass only the more informative signals to the detector (Butler et al., 2016). Two major technologies are used for collections of Raman spectra (Vankeirsbilck et al., 2002). Dispersive Raman spectrometers utilise Rayleigh filters, a single monochromator, or a multistage chromator while non-dispersive spectrometers apply Fourier transformation (FT) based on an interferometer. The most common filters are holographic notch and dielectric edge filters. Single monochromators compose a diffraction grating to disperse the Raman scattered light (Butler et al., 2016; Shipp et al., 2017).

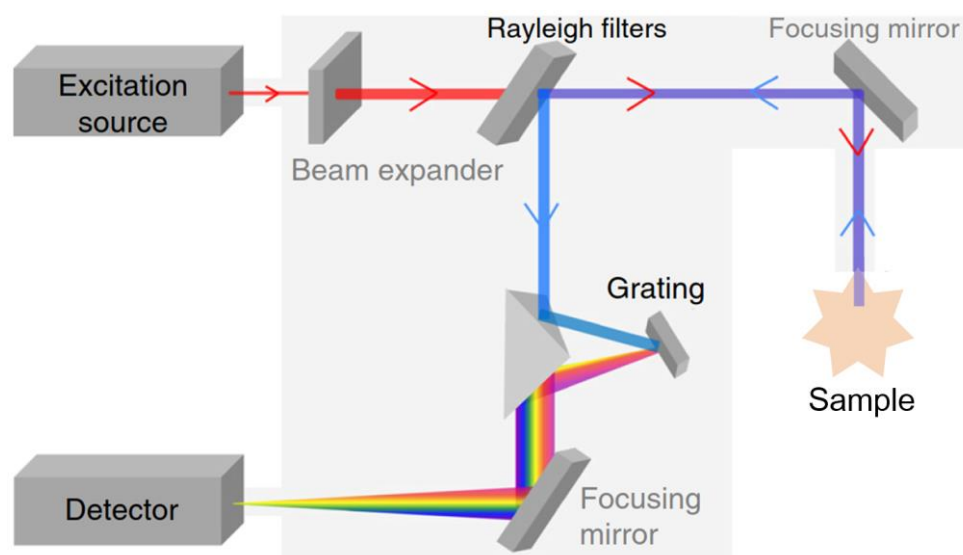


Figure 3.3: Generalised overview of instrumentation within a dispersive Raman spectroscopy system (modified after (Butler et al., 2016)).

In order to identify the weak intensity of scattering, the detection system of the spectrometers needs to be highly sensitive. Charge-coupled devices (CCDs) are widely integrated in Raman systems as they feature high quantum efficiencies with low signal-to-noise ratio. Other detection systems are photomultiplier tubes and photodiode arrays (Butler et al., 2016).

For both dispersive and FT Raman spectrometers, different configurations of the detection systems are available. Configurations can be either backscattering (180° geometry), right-angle scattering (90° geometry), or forward scattering (0° geometry), also denoted transmission (Figure 3.4). Additionally, these configurations can be combined with an optical microscope or optical fibres. Another approach to be mentioned is the spatially-offset Raman spectroscopy (SORS) (Esmonde-White et al., 2017; McCreery, 2000).

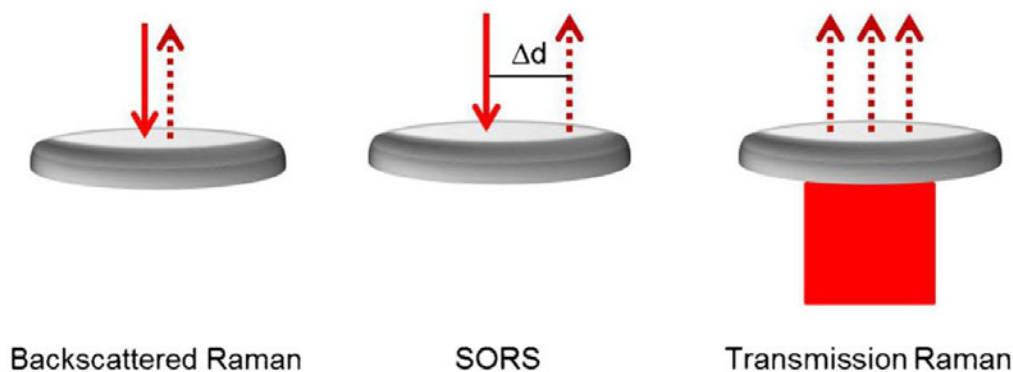


Figure 3.4: Schematic representation of configurations in Raman spectroscopy after sample excitation. Solid line represents the excitation beam while dashed line depicts the type of signal collection afterwards. SORS: spatially-offset Raman spectroscopy (modified after (Esmonde-White et al., 2017)).

Backscattering has been the main configuration for many years. The laser beam irradiates the sample, and the radiation is reflected back to the detector. Here, the small size of laser spot has to be considered. As the sampled volume can be small, the acquired spectrum might not represent the bulk sample. Therefore, it is important to ensure a homogeneous sample or to obtain multiple spectra of the same sample. Today, transmission Raman is growing in application. The sample is excited by a defocused laser and the signal is collected through the sample. This configuration also provides representative sampling and suppresses fluorescence. As backscattering and transmission Raman are used in this work, SORS will not be further described. For this, refer to literature of Esmonde-White and colleagues (Esmonde-White et al., 2017). While backscattering can be applied in-line, on-line, at-line, or off-line, transmission Raman is mostly used as an off-line PAT (Esmonde-White et al., 2017).

3.2.2 Applicational Fields

Raman spectroscopy has demonstrated to be an effective analysis in a range of sciences, such as material science, semi-conductor, geology, medicine, and polymer fields. Either solids, liquids or gases can be analysed (Smith & Dent, 2019). Before its technological advances, NIR spectroscopy was the state of art for pharmaceutical production. As NIR spectra contain bands that are broad and overlapped, Raman spectra show higher specificity (Shipp et al., 2017). Also, Raman spectroscopy generates a weaker water signal, making it an interesting alternative for the application in culture broth and other aqueous solutions (Buckley & Ryder, 2017). Yet, there are shortcomings in the application of dispersive Raman spectroscopy. As Raman scattering is a low-probability event, weak scattering signals can become an issue, especially when working in low concentration ranges of the analyte. Raman signal intensity can be increased by

use of higher power excitation laser. However, sample burning or photoablation can occur (Smith & Dent, 2019). Instead, longer integration times or more efficient detection optics can be used. Another drawback with dispersive Raman spectroscopy is the fluorescent effect. If the laser beam excites molecules into higher energy states, the emitted photon may have a lower energy level than before. This emitted light can completely cover the Raman signal due to fluorescence occurring in the same wavelength range as the anti-Stokes Raman signal and has a stronger signal than Raman scattering. Biological samples such as culture broth or cell culture media contain many fluorophores, which can become a challenge. This can be solved by use of excitation sources with longer wavelengths at 785 nm, 830 nm, or 1064 nm. However, it has to be considered that hereby the Raman signal is further weakened (Buckley & Ryder, 2017).

3.3 Cultivation Strategy

The cultivation of cells to high cell densities can be summarised in a three-stage cultivation course (Figure 3.5) (El-Mansi et al., 2018). For more information regarding the applied cultivation system, see Chapter 4.4.1.

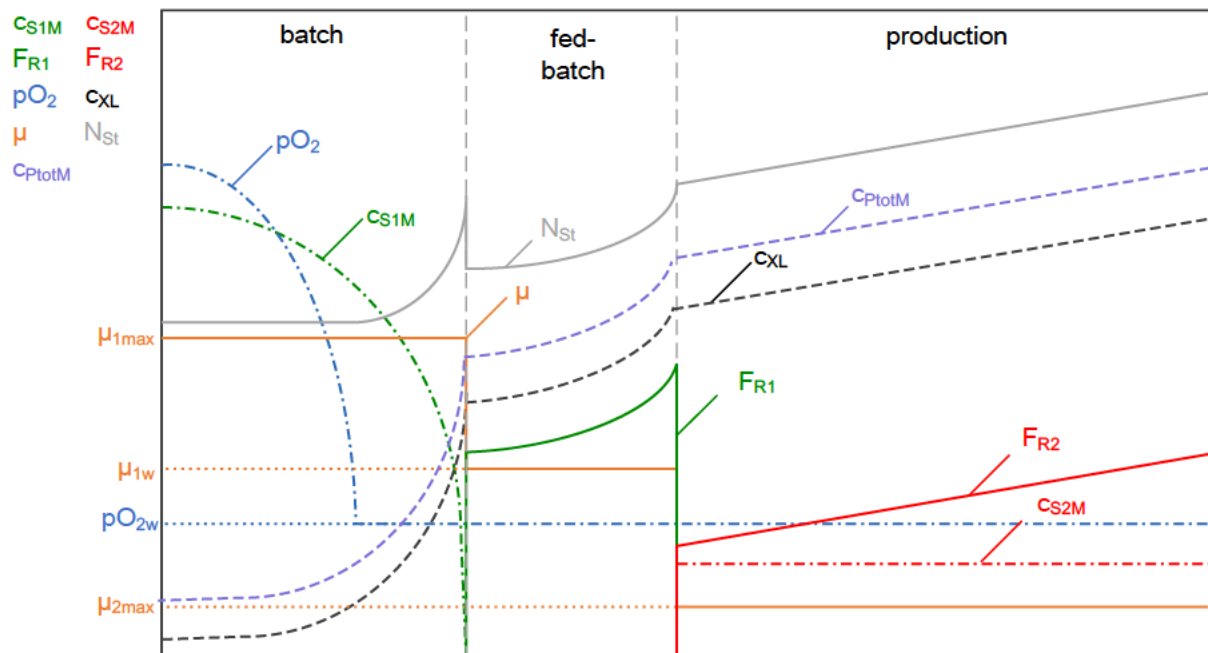


Figure 3.5: Schematic course of three-stage cultivation. S1: substrate 1 (glycerol); R1: reservoir 1 (glycerol); S2: substrate 2 (methanol); R2: reservoir 2 (methanol); Ptot: total protein; X: cell dry weight; St: stirrer.

A cultivation is initiated by transferring the inoculum to the bioreactor in the batch phase. Here, the cells (black dashed) first have to adapt to the new environment in the culture vessel. After adaption, the cells grow exponentially with maximum cell-specific growth rate μ_{1max} (orange

solid) as long as substrate 1 (S1) (green dashed) is available. Due to automatic process control of pH and temperature, these process parameters are continuously maintained at its set points. However, the DO level (blue dotted-dashed) is exponentially decreasing until falling below the set point pO_{2w} . To maintain the set point and DO supply for the cells during the whole course of cultivation, agitation (grey solid) controls the DO. The end of the batch phase is characterized by a rapid increase of DO as S1 is not extensively available for all cells. A pre-determined exponential feeding profile where S1 is continuously added to the process (green solid) is run during fed-batch. The feeding enables maintenance of a pre-defined cell-specific growth rate μ_{1w} and supervised growth to high cell densities. When a pre-defined cell concentration was achieved, the next phase of cultivation was initiated. In production phase, S1 feeding is stopped. Instead, substrate 2 (S2) (red solid) is added to the process such that the set point c_{S2Mw} (red dotted-dashed) is maintained (El-Mansi et al., 2018).

3.4 Multivariate Data Analysis

Advances in technology and increasing availability of powerful instruments enable the possibility of obtaining high amounts of data on each sample analysed in a reasonable time frame. In spectroscopy, a single rapid analysis on the sample creates multiple informational data. Its spectrum can be considered as a data vector where the order of the variables, e.g. Raman intensities measured in arbitrary units at consecutive wavenumbers, has a physical meaning (Danzon et al., 2001; Oliveri et al., 2020). Simultaneously, the gathered data is most likely not immediately interpretable, therefore the information is not directly accessible. Rather, a number of steps is required in order to extract and properly interpret the potential information manifested in the data (Eriksson et al., 2006b). The science of extracting chemical information out of complex data is called chemometrics. In fact, disciplines such as applied mathematics, computer science, and multivariate statistics are required (Wold, 1995). MVDA is a sub-discipline of chemometrics and aims to process and evaluate large complex data sets with numerous observations by extracting the relevant information (Eriksson et al., 2006b). In the following, possibilities with MVDA are elaborated, different pre-processing steps followed by MVDA tools are explained before the construction of multivariate models is introduced.

3.4.1 From Data to Information

Statistics offer helpful tools which can be used to turn data into information. Univariate strategies consider one variable at a time, independently from each other. This strategy has been

extensively used for information extraction. However, when intercorrelation between variables occur, it comes to its limits (Eriksson et al., 2006b). In contrast, multivariate statistics, can take this aspect into account, allowing a more complete interpretation of the data. For this, computer-based methods are applied as classical approaches are promptly limited by the human three-dimensional imagination (Danzer et al., 2001).

3.4.1.1 Basic Types of Data Analytical Questions

MVDA serves as a toolbox for three basic types of data analytical problems (Figure 3.6), also representing the major stages of MVDA:

1. Overview of the data set,
2. Classification and discrimination of observations, and
3. Regression analysis between two data blocks (X and Y).

Due to this, MVDA has been widely used in applicational areas such as process monitoring and early failure detection, quality control, data mining, multivariate calibration, and image analysis (Eriksson et al., 2013).

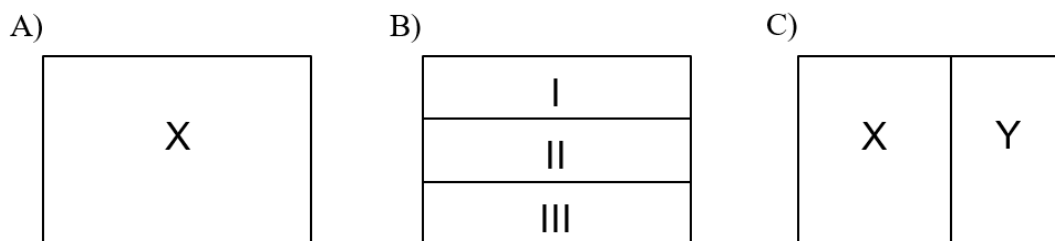


Figure 3.6: MVDA toolbox to solve different data analytical questions. A) Overview; B) Classification and discrimination; C) Regression modelling between two blocks of data (modified after (Eriksson et al., 2013).

In the first stage, a data overview can be obtained in an early phase of a project. This overview is accomplished with principal component analysis (PCA), a tool of MVDA introduced in Chapter 3.4.5. PCA points out how observations are related and if there are any deviating observations. Also, time trends and sudden shifts in the data are revealed. In the second stage, separate models of defined classes can be produced by PCA. Here, PCA aims to predict class membership of additional observations which were not previously considered in the data analysis. In the last stage, often linking two data blocks is desired, referred to as X and Y. This regression modelling enables the prediction of Y from X for new observations and can be achieved by projection to latent structures (PLS) (Eriksson et al., 2013). This will be further elaborated in Chapter 3.4.6.

3.4.1.2 Data and Data Structures

In process modelling and monitoring, the X-variables, also denoted factors or predictors, are signals which are measured frequently in order to monitor the process status. In contrast, the Y-variables, denoted responses, are measured less frequently and represent properties such as quality or yield of a product. The Y-variables are often time-consuming, laborious, and expensive to measure compared with the X-variables (Eriksson et al., 2013).

3.4.1.3 Tools for Data Analysis

With MVDA, different models can be obtained by applying different types of analysis. Some analysis techniques, amongst others, are multivariate analysis of variance (MANOVA), multiple linear regression (MLR), linear discriminant analysis (LDA), soft independent modelling by class analogy (SIMCA), PCA, and PLS (Eriksson et al., 2006b). The last two methods are further introduced in the following chapters as these were subject of this work.

3.4.2 Approach to Data Analysis

The typical data analytical evaluation is divided into several steps (Figure 3.7). In the phase of experimental design, one or more hypotheses are formulated, serving as a basis for data mining during experiments. After data collection, the data have to be prepared. Here, the raw data is transformed into mathematical structures such as matrices or tables. Then, raw data is turned into cleaned data by removing unwanted variations like undesirable scatter effects originating from experimental and instrumental artifacts (Danzer et al., 2001; Engel et al., 2013; Rinnan et al., 2009). If pre-processing steps were not chosen adequately, the results can also introduce unwanted variation. Thus, pre-processing influences the successful outcome of all following steps in the pipeline and with this, the entire experiment (Engel et al., 2013).

After pre-processing, the actual data analysis in chemometrics can be performed. The aim here is the design of a mathematical model which describes the inherent structures and relations of the data. By this, new observations can be predicted which were not previously considered for modelling. The hypotheses originally formulated can then be refuted or substantiated with the help of statistics. If necessary, the pre-processing can be changed to obtain a better model result (Danzer et al., 2001).

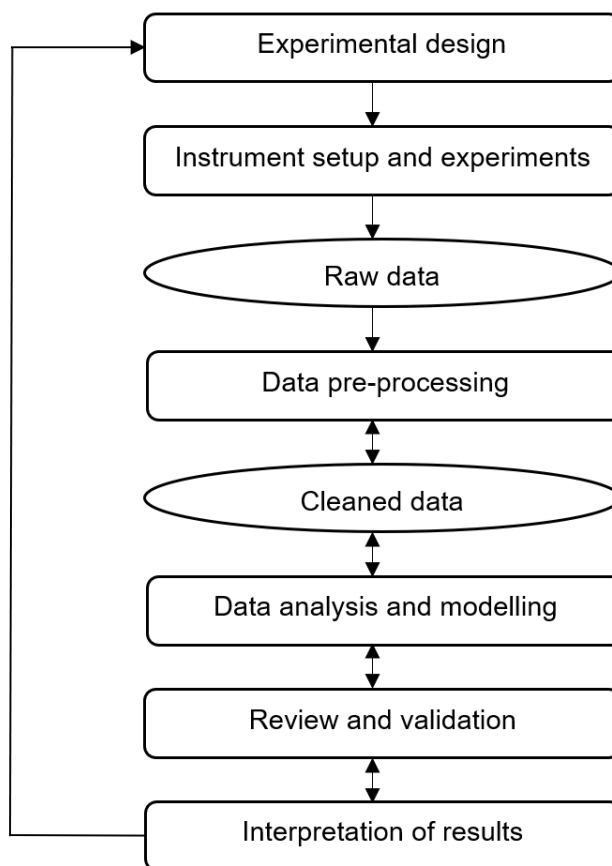


Figure 3.7: Typical pipeline for chemometric data analysis comprising design, performance, and analysis of experiments. Box: action step; circle: condition of data (modified after (Danzer et al., 2001; Engel et al., 2013)).

The quality of the model can be evaluated by two criteria. First, internal validity should be present. This means, the interpretation of the results must be valid and reliable for the prevailing data set, denoted training or calibration set (CS). Second, the results should be transferrable and generalisable to future measurements, referred to as external validity with a prediction set (PS). An external validation is performed with new, unknown data. In the practical sense, a cyclic procedure with alternating model building and validation is typical until the prediction is not improving significantly (Danzer et al., 2001). Due to this, a good understanding of the characteristics of the methods employed for data pre-processing is advantageous and will be further elaborated in Chapter 3.4.4.

However, there is the chance to overfit the model. Since the data set used for modelling depicts a sample of the population, an improved internal validity does not automatically result in a higher external validity. In fact, when the model fits exactly against the CS, the model cannot perform accurately against unknown data, decreasing the prediction quality. To overcome this possibility, a proper validation of models is required (cf. Chapter 3.4.6.5) (Oliveri et al., 2020).

3.4.3 Data Preparation

Before MVDA can be implemented, the measured data is transferred into a $(n \times m)$ data matrix X . This data matrix X comprises $i = 1, 2, \dots, n$ rows and $j = 1, 2, \dots, m$ columns (Figure 3.8). In the context of chemometrics, rows and columns are named observations and variables, respectively. Observations can be analytical samples, chemical compounds, or process time points of a continuous process. To characterize the properties of the observations, the variables are measured in form of spectra, chromatograms, or from sensors and instruments in a process (Eriksson et al., 2006b).

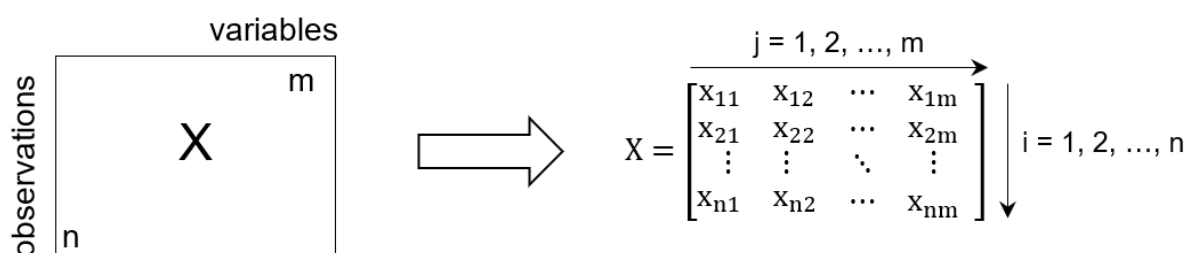


Figure 3.8: Data table with n observations and m variables. The variables are measurements m made in order to capture the properties of the observations n . The data table is transformed into a $(n \times m)$ data matrix X .

3.4.4 Data Pre-Processing

The role of pre-processing for the model outcome is crucial. Furthermore, it deals with challenging data characteristics such as data artifacts or missing values. Artifacts, in contrast, are dependent on the used analytical chemical technique, e.g. baseline shifts in spectroscopy or peak shifts in chromatography. In general, both observations and variables can be pre-processed. Also, several data pre-processing steps can be consecutively employed (Figure 3.9) (Engel et al., 2013).

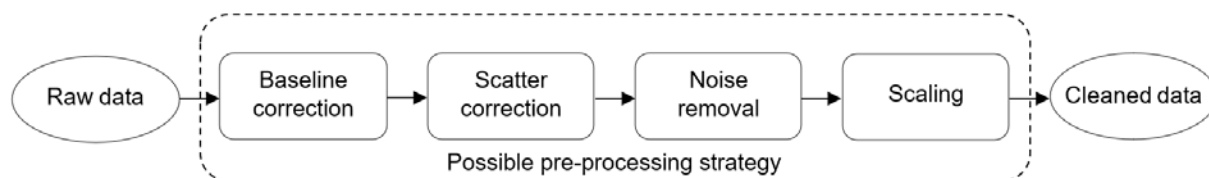


Figure 3.9: Pipeline for pre-processing of spectral data. Steps in the dashed box can be either skipped, added or the order of steps can be changed. Box: action step; circle: condition of data (modified after (Engel et al., 2013)).

Each pre-processing step aims to correct a particular artifact. The choice and order of pre-processing steps should be based on the goal of the data analysis. Simultaneously, changing the order may result in a change of the final result. When working with large, complex data sets,

the effect of each single pre-processing step is not transparent. Due to this, conclusions drawn should be robust for the pre-processing method applied on the CS (Engel et al., 2013).

Many pre-processing methods have been developed for spectral data (Engel et al., 2013; Eriksson et al., 2006b; Kessler, 2006; Rinnan et al., 2009). Still, it is not possible to certainly predict the output of pre-processed data and there are no clear guidelines for or against the use of certain pre-processing methods (Eriksson et al., 2006b). Therefore, different pre-processing methods are subject of this work and will be introduced in the following.

3.4.4.1 Baseline Correction

Baseline removal is useful in eliminating variable background originating from fluorescence or interfering ambient light when clear Raman peaks are still present in a spectrum (Huang et al., 2010). Baseline effects result in signals with a vertical offset or a slope (Engel et al., 2013).

Derivatives

Derivatives offer an effective method for baseline correction and simultaneously improve the spectral resolution. However, the chemical interpretability is impaired as the appearance of the spectra is heavily altered (Kessler, 2006). The 1st order derived (1stDer) element $X_{ij1stDer}$,

$$X_{ij1stDer} = \frac{\partial d_{ij}}{\partial \lambda_{exj}} \approx \frac{d_{ij+1} - d_{ij-1}}{\lambda_{exj+1} - \lambda_{exj-1}}, \quad (3.1)$$

with d_{ij} value of row i and column j
 λ_{exj} excitation wavelength of column j ,

corresponds to the slope at each point of the original spectrum. It peaks where the original spectrum has maximum slope and it passes zero where the original spectrum has peaks (Eriksson et al., 2006b). A 1st order derivative eliminates a constant baseline (offset) while a 2nd order derivative (2ndDer) also eliminates the baseline slope (Engel et al., 2013).

The second derivative spectrum shows the curvature at each point in the original spectrum. This derivative spectrum is more similar to the original and shows peaks at vicinity as the original spectrum, but with an inverse configuration. Higher-order derivatives can amplify unwanted noise. One drawback with derivatives is that signals may be reduced and noise may be increased, producing a noisy spectrum (Eriksson et al., 2006b; Huang et al., 2010).

3.4.4.2 Scatter Correction

Signal differentiation can subtract the signal background by scattering, enhancing the visual resolution (Danzer et al., 2001).

Standard Normal Variate Filter

For spectral data, standard normal variate (SNV) filters are often applied as pre-processing step, working row-wisely. By SNV-filtering, the spectra are normalised and both baseline and wavelength-dependent scatter effects are corrected (Kessler, 2006). The SNV-filtered value x_{ijSNV} ,

$$x_{ijSNV} = \frac{d_{ij} - \bar{d}_i}{s_{di}}, \quad (3.2)$$

with s_{di} standard deviation of row i ,

is calculated by subtracting the row mean \bar{d}_i ,

$$\bar{d}_i = \frac{1}{m} \sum_{j=1}^m d_{ij}, \quad (3.3)$$

with m number of variables in row i ,

by the measured value d_{ij} and then divided by the row standard deviation s_{di} ,

$$s_{di} = \sqrt{\frac{1}{m-1} \sum_{j=1}^m (d_{ij} - \bar{d}_i)^2}. \quad (3.4)$$

Multiplicative Signal Correction

Multiplicative signal correction (MSC) assumes that wavelength-dependent scattering effects can be separated from chemical information. It estimates the coefficients a_i and b_i describing the scattering by fitting the spectrum x_i ,

$$x_i = a_i + b_i \cdot \bar{x} + e_i, \quad (3.5)$$

with a_i scatter difference coefficient
 b_i baseline offset coefficient
 \bar{x} mean spectrum
 e_i polynomial coefficients,

to a reference, the average spectrum \bar{x} , by least squares fit. The chemical information is ideally incorporated in e_i , as scattering and offset are described by a_i and b_i , respectively.

For each spectrum, the MSC correction coefficients a_i and b_i are determined in order to calculate the MSC-corrected spectrum x_{iMSC} ,

$$x_{iMSC} = \frac{x_i - a_i}{b_i}. \quad (3.6)$$

This pre-processing tool is dependent on the mean spectrum as the coefficients are calculated by use of these. If observations are excluded from the data set or further observations are included, the MSC model has to be recalculated (Kessler, 2006).

To exemplify the effects of previously introduced pre-processing step, Raman spectra of the culture broth during cultivation are depicted (Figure 3.10).

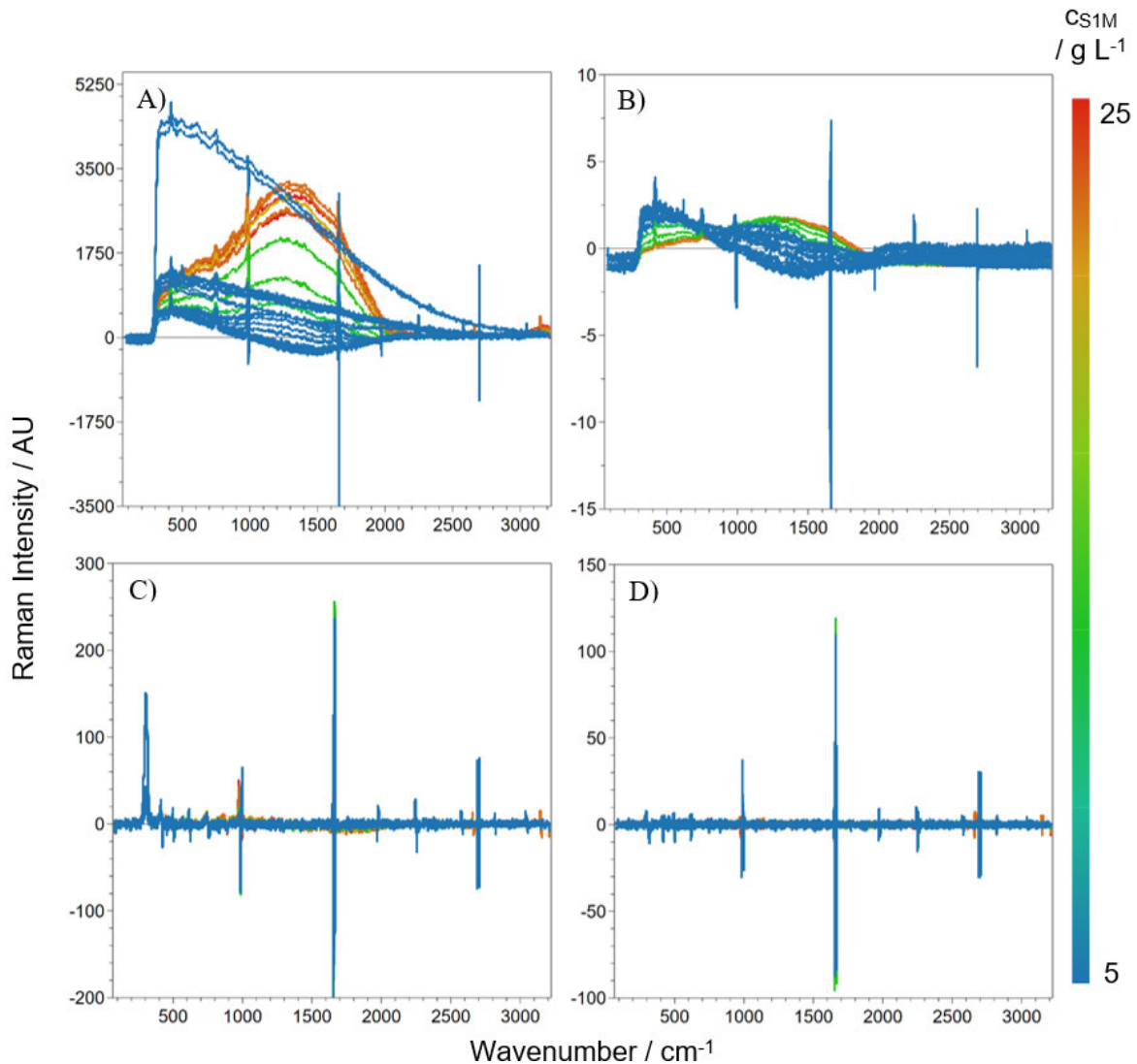


Figure 3.10: Different pre-processing methods of Raman spectra. Spectra are coloured according to glycerol concentration c_{S1M} . A) raw data; B) SNV-filtered; C) 1st derivative; D) 2nd derivative.

3.4.4.3 Noise Removal

Noise is prevalent in almost any analytical techniques and the underlying background differs per analytical technique. For this, mathematical filters can be applied to improve signal-to-noise ratio (Engel et al., 2013).

Savitzky-Golay Filter

The numerical differentiation of spectral data is mostly done with the Savitzky-Golay (SG) algorithm. This filter uses a moving data frame which is consecutively adapted to a polynomial based on the least squares fitting. By differentiation of the polynomial, derivatives of different orders are obtained. The size of the data frame used determines the level of smoothing. Before applying the Savitzky-Golay filter, the data have to be modified to an equidistant time axis (Kessler, 2006).

After setting data frame size, the SG-filtered measured value x_{kSG} ,

$$x_{kSG} = \frac{1}{\text{NORM}} \sum_{j=-m}^m a_j \cdot x_{k+j}, \quad (3.7)$$

with	k	k -th data point, depending on data frame size
	NORM	sum of coefficients of Savitzky-Golay filter
	a_j	filter coefficients

can be calculated. The filter coefficient a_j can be extracted from corresponding tables, while the norm factor NORM is the sum of coefficients (Otto, 1997).

Moving Average Filter

Measured values from probes can also be pre-processed by the moving average (MA) filter. The averaged value x_{kMA} ,

$$x_{kMA} = \frac{1}{2 \cdot m + 1} \sum_{j=-m}^m x_{k+j}, \quad (3.8)$$

is calculated by the raw data x_k within the filter width $2 \cdot m$ (Ross & Heinisch, 2006).

3.4.4.4 Scaling

While previously described artifacts are all related to analytical techniques, the following pre-processing steps are artifacts related to the sample (Engel et al., 2013). For scaling, two methods are introduced, data mean centring (Ctr) and unit variance (UV). Scaling is also known under normalisation (Huang et al., 2010).

Data Mean Centring

Mean centring is applied to reduce model complexity and improves interpretability of multi-variate models. This pre-processing step works column-wisely and is mostly applied on spectral data (Kessler, 2006).

For calculation of the centred measured value $x_{ij\text{Ctr}}$,

$$x_{ij\text{Ctr}} = d_{ij} - \bar{d}_j, \quad (3.9)$$

with \bar{d}_j column mean value of d_j ,

the variable value d_{ij} is subtracted by the column mean value \bar{d}_j ,

$$\bar{d}_j = \frac{1}{n} \sum_{i=1}^n d_{ij}, \quad (3.10)$$

with n number of observations in column j .

Unit Variance

When handling data originating from different physical processes, the numerical dimensions can differ. Consequently, the variance of the data differs as well. This can lead to false weighing of the numerically higher-dimensioned data. To counteract the overweighing, UV (also known as autoscaling or standardisation (Engel et al., 2013)) can be applied. This step normalises the data to a unit variance of 1 (Kessler, 2006).

To calculate an auto-scaled measured value $x_{ij\text{UV}}$,

$$x_{ij\text{UV}} = \frac{d_{ij} - \bar{d}_j}{s_{dj}}, \quad (3.11)$$

the mean centred value is divided by the column standard deviation s_{dj} ,

$$s_{dj} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (d_{ij} - \bar{d}_j)^2}. \quad (3.12)$$

While centring is mostly applied to spectral data, UV scaling cannot be applied here as the high variance of the variables corresponds to the chemical information content. However, it can be applied to pre-process the response variable Y (Kessler, 2006).

3.4.5 Principal Component Analysis

The principal component analysis serves as an important tool for MVDA (Wold et al., 1987). Depending on the specialised field, PCA has many synonyms. In psychology and chemistry, PCA is known as factor analysis, in mathematics as singular value decomposition, and in signal processing as Kosambi–Karhunen–Loève theorem (Kessler, 2006). However, all terms aim for a reduction of the original dimension to a low-dimensional plane to facilitate analysis and interpretation. For this, a high number of observable variables are reduced to a few latent variables, denoted factor or principal component (PC), such that the original information of the uncorrelated PCs is largely preserved. This enables an overview of the data and uncovers trends, outliers, and groups or relationships of observations (Eriksson et al., 2006b).

3.4.5.1 Mathematical and Graphical Model of PCA

For explanation of the mathematical model, an example with a data set of $m = 3$ variables is examined first. The variables x_1 , x_2 , and x_3 represent the columns of the data matrix X and shape a three-dimensional coordinate system. Each observation n can be depicted within the three-dimensional space (Figure 3.11). Prior to PCA, data are pre-processed by mean centring and scaling to unit variance into the modified data matrix X . The vector of means is interpretable as a point in the space positioned in the middle of the swarm of points, denoted centre of gravity. Then, the coordinate system is re-positioned towards the gravity point (Eriksson et al., 2006b).

The first PC t_1 is the line in the m -dimensional space which best approximates the swarm of points in the sense of least squares. Each observation is projected onto this line to obtain a new coordinate value, denoted score, along the PC-line. The line passes the gravity point and represents the maximum variance in the scores. Usually, one PC is insufficient to adequately represent the systematic variation of the data. Thus, a second PC t_2 is calculated. This line is orthogonal to the first PC and has the maximum variance to the data, passing the gravity point and containing the direction of the line. Two PCs define a plane, the sub-space. All projected scores

in the sub-space are summarised in the score plot (Danzer et al., 2001; Eriksson et al., 2006b). If the residual variance of the data is negligible, there is no more need for calculation of the third PC t_3 or further PCs. With this, the dimension reduction was successful (Voß, 2017).

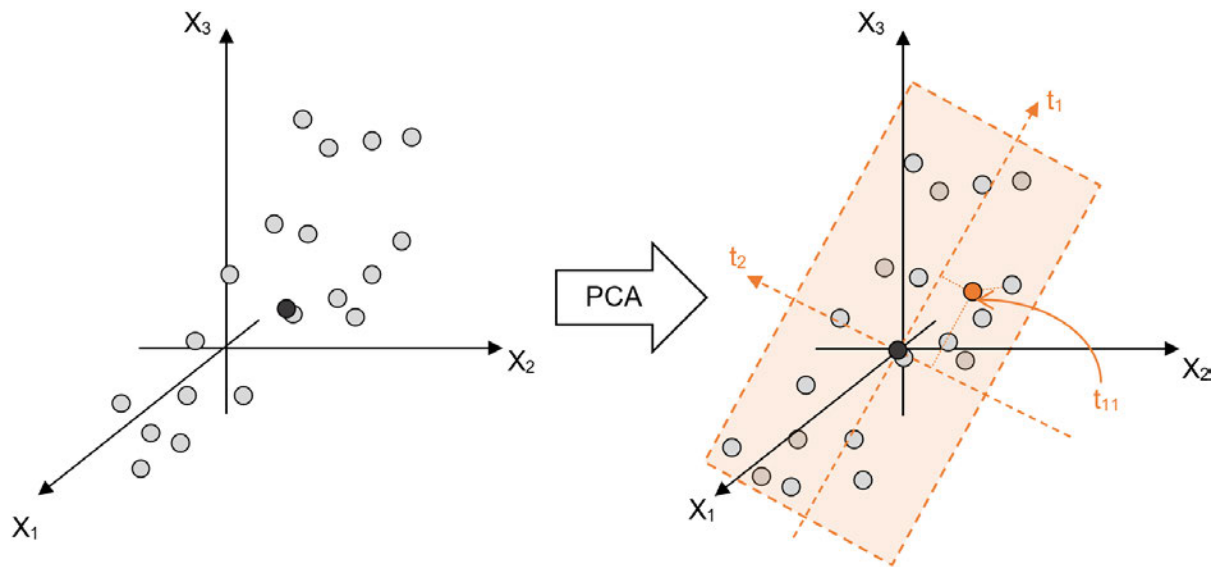


Figure 3.11: Graphical representation of PCA. The matrix X with n observations and m variables is interpreted as a cluster of n points in a m -dimensional space with the gravity point in the middle of the swarm (dark grey circle). PCA fits a line (one-dimensional), plane (two-dimensional), or hyperplane (three-dimensional) to the data. Here, a plane is formed (orange dashed square) out of two principal components. Each observation is projected onto the plane (orange dotted line), e.g. t_{11} (orange circle), resulting in a score for each of the calculated dimensions t_1 and t_2 . The new coordinates of the observations for the plane are the rows of the score matrix T while the directions in space are the m columns of the loading matrix P^T (modified after (Eriksson et al., 2006a).

For the mathematical explanation of a PCA model, the mean centred data matrix X ,

$$X = T \cdot P^T + E, \quad (3.13)$$

with	X	mean centred ($n \times m$) data matrix
	T	($n \times r$) score matrix of X
	r	number of principal components (columns in T)
	P^T	($r \times m$) transposed loading matrix of X
	E	($n \times m$) residual matrix of X ,

can be described by a score matrix T , a loading matrix P , and a residual matrix E (Kessler, 2006). The residual matrix E is equally dimensioned as the data matrix X and contains the remaining variance of the data X which is not described by the uncorrelated PCs (Voß, 2017).

The score matrix T comprises n rows and $l = 1, 2, \dots, r$ columns. For each observation, one score is assigned to each calculated PC, resulting in a new coordinate system. Often, the term

principal component (PC1, PC2, ...) is used for the column vectors (t_1, t_2, \dots , respectively) of score matrix T . In this work, the nomenclature remains at t_1 . The loading matrix P comprises m rows and $l = 1, 2, \dots, r$ columns. For each variable m , one loading is assigned to each calculated PC. The loading vectors p_l correspond to the direction vectors of the PC in the original coordinate system (Voß, 2017).

3.4.5.2 Calculation of Principal Components

To calculate the PCs of the PCA model, the most commonly used method is the non-linear iterative partial least squares (NIPALS) algorithm. This algorithm is an approximation procedure in order to calculate the scores and loadings. The procedure starts with a random value for the first PC t_1 which is iteratively improved until the value falls below a predefined error threshold J_{crit} (Wold, 1973). The exact algorithm steps will be explained in the following.

- 1) Starting point is the assignment of the temporary score vector t_{ltemp} ,

$$t_{ltemp} = x_j, \quad (3.14)$$

with x_j column vector of mean centred data matrix X ,

which index j starts with $l = 1$ and increases with each factor by 1 and has the highest variance $s_{x_j}^2$,

$$s_{x_j}^2 = \frac{1}{n-1} \cdot \sqrt{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}, \quad (3.15)$$

with x_{ij} column vector of data matrix X ,

of the mean centred data matrix X .

- 2) For this score vector, the corresponding temporary loading vector p_{ltemp} ,

$$p_{ltemp} = \frac{X^T \cdot t_{ltemp}}{t_{ltemp}^T \cdot t_{ltemp}}, \quad (3.16)$$

with X^T transposed mean centred data matrix X

t_{ltemp}^T transposed temporary score vector,

is calculated by projecting the data matrix X to the sub-space t_{ltemp} .

The loading vector p_l ,

$$p_l = \frac{p_{ltemp}}{\|p_{ltemp}\|} = \frac{p_{ltemp}}{\sqrt{p_{ltemp}^T \cdot p_{ltemp}}}, \quad (3.17)$$

with p_{ltemp}^T transposed temporary loading vector,

is normalised to 1, providing the direction vector.

- 3) To improve the estimation for the score vector t_l ,

$$t_l = X \cdot p_l, \quad (3.18)$$

the data matrix X is projected to the new loading vector and simultaneously to the subspace p_l .

- 4) By evaluating the convergence criterium J_l ,

$$J_l = \|t_{ltemp} - t_l\| = \sqrt{\sum_{i=1}^n (t_{ltemp} - t_l)^2}, \quad (3.19)$$

the difference between temporary and corrected score vector is compared.

- 5) When the difference exceeds a predefined value J_{crit} , e.g. 10^{-6} , there was no convergence reached and the temporary score vector t_{ltemp} ,

$$t_{ltemp} = t_l, \quad (3.20)$$

becomes the in Step 3) calculated score vector t_l . Then, a new iteration procedure starts with Step 2). Otherwise, when the difference falls below the predefined value J_{crit} , the procedure did converge. The score vector t_l and the corresponding loading vector p_l form the solution for the l -th PC. Then, the procedure continues with Step 6).

- 6) For the calculation of another PC with $l = l + 1$, the residual matrix E ,

$$E = X - t_{ltemp} \cdot p_l^T, \quad (3.21)$$

is calculated by removing the information of the PC t_l from the data matrix X .

- 7) Lastly, the new modified data matrix X ,

$$X = E, \quad (3.22)$$

is assigned to the residual matrix E calculated in Step 6) for the restart of the algorithm at Step 1).

For the NIPALS algorithm, Steps 1) to 7) are repeated such that all potential PCs have been calculated or a certain fraction of the population variance is explained by the PCA model. The maximum number of PCs conforms the maximum number of observations n or variables m , respectively (Kessler, 2006).

3.4.6 Projections to Latent Structures

The second MVDA tool introduced is the projections to latent structures by means of partial least squares, also known as partial least squares regression (PLSR). The aim of PLS is the correlation of information in two blocks of variables, X and Y , by linear multivariate calibration. This enables the prediction of variables which were not previously considered in the model. These variables are characterised by noisiness, incompleteness, of high number, or laborious to measure (Eriksson et al., 2006b). In this work, Raman spectra were used in order to predict bioprocess variables.

3.4.6.1 Mathematical Model of PLS

In PLS, a multivariate approach is applied with a $(n \times m)$ data matrix X containing n observations. Each observation i corresponds to one target value y_i with $i = 1, 2, \dots, n$ rows, forming the vector y . Multiple measured y_i result in the response matrix Y ,

$$Y = X \cdot B + G, \quad (3.23)$$

with Y $(n \times v)$ mean centred and auto-scaled measurement matrix
 B $(m \times v)$ PLS regression coefficient matrix
 G $(n \times v)$ residual matrix of regression approach,

consisting of n rows and $h = 1, 2, \dots, v$ columns. Y is typically the analyte concentration of the CS. The resulting model can then be used to predict the analyte concentration from the spectra of new samples (Eriksson et al., 2006b; Kessler, 2006).

The PLS regression is based on PCA. In fact, two PCA models are simultaneously calculated for data matrix X (equation 1.13) and measurement matrix Y ,

$$Y = U \cdot Q^T + F, \quad (3.24)$$

with U $(n \times r)$ score matrix of Y
 Q^T $(r \times v)$ transposed loading matrix of Y
 F $(n \times v)$ residual matrix of Y .

In order to correlate X and Y, both data spaces need to exchange their information in order to project the score matrix T to the score matrix U (Figure 3.12).

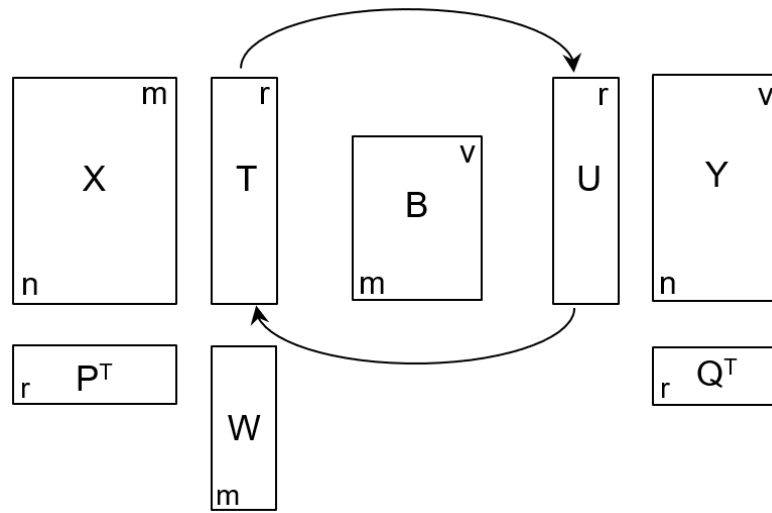


Figure 3.12: Schematic representation of projections to latent structures and involved matrices. Aim is to correlate data matrices X with Y. For this, the score matrices of both data matrices need to exchange information, enabled by the weighted loading matrix W (modified after (Kessler, 2006; Voß, 2017)).

The latent variables calculated by PLS are influenced by both data spaces. Due to this, the score vector T and loading vector P are not similar to the scores and loadings of a plain PCA. In the context of PLS, the latent variables are denoted PLS-component or factor instead of PC (Voß, 2017). Before the exchange of information between X and Y is possible, the $(m \times r)$ weight matrix W is required. This matrix is equally dimensioned as the loading matrix P and is also referred to as weighted loading matrix of X, enabling the correlation between both data spaces (Kessler, 2006). In the following, the calculation of PLS-components is elaborated.

3.4.6.2 Calculation of PLS-Components

The regression coefficient matrix B,

$$B = W \cdot (P^T \cdot W)^{-1} \cdot Q^T, \quad (3.25)$$

with W $(m \times r)$ weighted loading matrix of X
 P^T $(r \times m)$ transposed loading matrix of X
 Q^T $(r \times v)$ transposed loading matrix of Y,

can be calculated from the weight and loading matrices W, P, and Q, respectively (Wold, 1995).

To determine the PLS-components, a modified NIPALS algorithm is applied. The so-called local models for X,

$$X = T \cdot W^T + E', \quad (3.26)$$

and

$$X = U \cdot W^T + E'', \quad (3.27)$$

and for Y,

$$Y = T \cdot Q^T + F', \quad (3.28)$$

are used which also explain the data spaces X and Y. Then, the iterative NIPALS is applied until convergence is reached. This will be explained in the following.

- 1) Starting point is the assignment of the temporary score vector u_{Itemp} ,

$$u_{\text{Itemp}} = y_h, \quad (3.29)$$

with y_h column vector of scaled measurement matrix Y.

- 2) Then, the local model in equation (3.25) is used and the residual matrix E'' is neglected in order to calculate the temporary weight vector w_{Itemp} ,

$$w_{\text{Itemp}} = \frac{X^T \cdot u_{\text{Itemp}}}{u_{\text{Itemp}}^T \cdot u_{\text{Itemp}}}, \quad (3.30)$$

by a least squares fitting.

The temporary weight vector is then normalised to 1, turning into w_1 ,

$$w_1 = \frac{w_{\text{Itemp}}}{\|w_{\text{Itemp}}\|} = \frac{w_{\text{Itemp}}}{\sqrt{w_{\text{Itemp}}^T \cdot w_{\text{Itemp}}}}. \quad (3.31)$$

- 3) Subsequently, the local model in equation (3.24) is used and the residual matrix E' is neglected in order to determine the score vector t_1 ,

$$t_1 = X \cdot w_1, \quad (3.32)$$

by projecting the data matrix X to the sub-space w_1 .

- 4) The local model in equation (3.26) is used and the residual matrix F' is neglected in order to calculate the X-loading vector p_l ,

$$p_l = \frac{X^T \cdot t_l}{t_l^T \cdot t_l}, \quad (3.33)$$

and the Y-loading vector q_l ,

$$q_l = \frac{Y^T \cdot t_l}{t_l^T \cdot t_l}. \quad (3.34)$$

- 5) A new Y-score vector u_l ,

$$u_l = \frac{Y \cdot q_l}{q_l^T \cdot q_l}, \quad (3.35)$$

is determined by projecting the measurement matrix Y to the sub-space q_l .

- 6) Parallel to the NIPALS algorithm for PCA, the convergence criterium J_l ,

$$J_l = \|u_{ltemp} - u_l\| = \sqrt{\sum_{i=1}^n (u_{iltemp} - u_{il})^2}, \quad (3.36)$$

is evaluated.

- 7) When the difference exceeds a predefined value J_{crit} , e.g. 10^{-6} , there was no convergence achieved and the temporary score vector u_{ltemp} ,

$$u_{ltemp} = u_l, \quad (3.37)$$

becomes the in Step 5) calculated score vector t_l . Then, a new iteration procedure starts with Step 1). Otherwise, when the difference falls below the predefined value J_{crit} , the procedure converged. The score vector t_l and the corresponding loading vector p_l form the solution for the l -th PLS-component and the procedure continues with Step 8).

- 8) For calculation of another PLS-component with $l = l + 1$, a new residual matrix E ,

$$E = X - t_l \cdot p_l^T, \quad (3.38)$$

and F ,

$$F = Y - u_l \cdot q_l^T, \quad (3.39)$$

are determined by removing the scores and loadings of the data matrices X and Y , respectively.

9) Lastly, the matrices X,

$$X = E, \quad (3.40)$$

and Y,

$$Y = F, \quad (3.41)$$

are assigned to the residual matrix E and F calculated in Step 8) for the restart of the algorithm at Step 1).

The algorithm is repeated such that the desired number of PLS-components is determined, and the prediction error is sufficiently low. Then, the coefficient matrix B can be calculated by use of equation (3.25) (Kessler, 2006).

3.4.6.3 Detection of Outliers

Due to the approach of the least squares errors used in PCA and PLS, outliers can have a great impact upon the multivariate model (Wold et al., 1987). Due to this, outliers should be excluded from the model. In the following, two distinct outlier detection methods are introduced.

Hotelling's T^2 Test

The first method for outlier detection is the Hotelling's T^2 test. An observation differing strongly from others can be detected. The Hotelling's T^2 test is the multivariate generalisation of the Student's t test and examines observations on normal distributions (Hotelling, 1951).

The Hotelling's T^2 value T_i^2 of an observation i,

$$T_i^2 = \sum_{l=1}^r \frac{(t_{il} - \bar{t}_l)^2}{s_{il}^2}, \quad (3.42)$$

with	t_{il}	score of component l for observation i
	\bar{t}_l	mean of score vector t_l
	s_{il}^2	variance of scores of component l,

describes the normalised distance of an observation to the centre of the model for all calculated PCs. When using centred or scaled data, the mean score \bar{t}_l of all components l equals 0. Hereby, the model centre is found at the origin (Ross & Heinisch, 2006).

Observations with a Hotelling's T^2 value above the critical value $T_{crit\alpha}^2$,

$$T_{crit\alpha}^2 = \frac{r \cdot (n - 1)}{n - r} \cdot F_{(\alpha, r, n-r)}, \quad (3.43)$$

with	α	level of significance
	r	number of principal components
	n	number of observations
	$F_{(\alpha, r, n-r)}$	critical value of F-distribution with significance level α and r and $n-r$ degrees of freedom,

embody with a probability as high as the chosen level of significance α an outlier. Commonly, $\alpha = 5\%$ is used, resulting in a confidence interval of 95%. The required critical $F_{(\alpha, r, n-r)}$ value is derived from the cumulative distribution function of a F-distribution, depending on α , the degrees of freedom n , and $n - r$, or can be read out from tables (Ross & Heinisch, 2006).

To visually detect outliers with Hotelling's T^2 test, scores scatter plots are used. Scores of t_1 are plotted against scores of t_2 , allowing a straightforward method for the detection of outliers. Observations with deviant characteristics occur on the borders of the data points. By displaying the Hotelling's T^2 ellipse for the examined PCs, outliers can be detected (Figure 3.13) (Eriksson et al., 2006b).

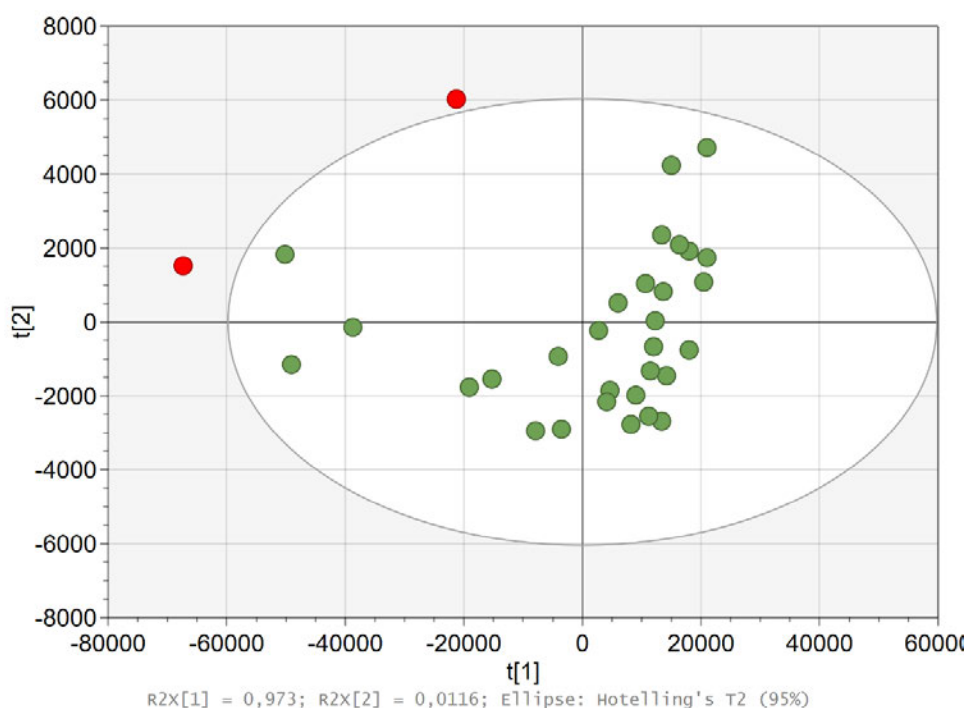


Figure 3.13: Exemplary scores scatter plot with applying Hotelling's T^2 test and a confidence interval of 95%. Principal component 1 t_1 and 2 t_2 are plotted against each other. Green: data points within α ; red: data points out-side level of significance. Figure produced with SIMCA® 17.0.1.

The radius of the Hotelling's T^2 ellipse $r_{T\alpha}$,

$$r_{T\alpha} = \sqrt{\frac{2 \cdot (n - 1)}{n - 2} \cdot F_{(\alpha, 2, n-2)} \cdot s_{tl}} \quad (3.44)$$

with $F_{(\alpha, 2, n-2)}$ critical value of F-distribution with significance level α , 2, and $n-2$ degrees of freedom,

is calculated by the square root of $T_{crit, \alpha}^2$ (equation 3.41) for $r = 2$ and the standard deviation s_{tl} for the examined PCs (Eriksson et al., 2006b).

Distance to Model X

The other method for outlier detection is the parameter Q residuals, describing the orthogonal distance of an observation to the hyperplane in the original data space X. When a data point exceeds a certain distance, it may be an outlier (Eriksson et al., 2013). In the environment of the software SIMCA[®] 17.0.1 (Sartorius Stedim Data Analytics, Sweden), the parameter is denoted Distance to Model X (DModX) which will be used subsequently in this work. DModX is proportional to the residual standard deviation of the model (Eriksson et al., 2006b).

The absolute distance $DModX_{absi}$,

$$DModX_{absi} = \sqrt{\frac{1}{m - r} \sum_{j=1}^m e_{ij}^2} \quad (3.45)$$

with m number of variables in data matrix X
 r number of calculated principal components
 e_{ij} element of residual matrix E in principal component analysis,

corresponds to the residual standard deviation of an observation i for all variables.

The mean distance $DModX_{ave}$,

$$DModX_{ave} = \sqrt{\frac{1}{(n - r - r_0) \cdot (m - r)} \cdot \sum_{i=1}^n \sum_{j=1}^m e_{ij}^2} \quad (3.46)$$

with n number of observations in data matrix X
 r number of calculated principal components
 r_0 1 for centred models, otherwise 0,

describes the pooled residual standard deviation of all observations n (Eriksson et al., 2013).

For comparison of DModX values of different models, DModX is displayed in normalised units. That is $DModX_{absi}$ divided by the pooled residual standard deviation of the model to obtain normalised $DModX_{normi}$,

$$DModX_{normi} = \frac{DModX_{absi}}{DModX_{ave}} = DModX_i. \quad (3.47)$$

In the following, the normalised $DModX_{normi}$ will be simply denoted DModX. As with the Hotelling's T^2 value T_i^2 , a critical value $D_{crit\alpha}$ can be calculated for a chosen level of significance α . $D_{crit\alpha}$ is calculated from the F-distribution and regulates the “envelope” surrounding the data points of the CS (Figure 3.14). Observations twice as large as $D_{crit\alpha}$ are moderate outliers, indicating that these observations are different from the normal observations with respect to the correlation structure of the variables (Eriksson et al., 2013).

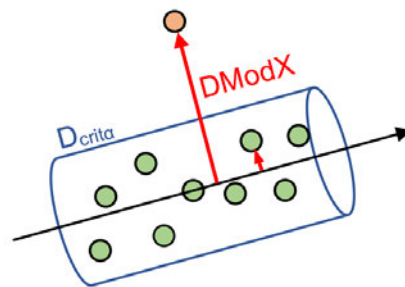


Figure 3.14: Visualisation of an observation's distance to model X (DModX) in plane of original data set X. A value for DModX can be calculated for each observation. These values can be plotted in a control chart where the maximum tolerable distance $D_{crit\alpha}$ is given in order to reveal outliers.

3.4.6.4 Selection of Variables

For multivariate models with many components and a multitude of responses, the interpretation can be challenging. A metric which summarises the importance of the X-variables, both for X- and Y-models, is denoted variable importance in projection (VIP). VIP is a weighted sum of squares of the PLS-weights w_{jlnorm} , taking into account the amount of explained Y-variance (Wold et al., 2001).

In order to calculate the metric VIP_j ,

$$VIP_j = \sqrt{\frac{m \cdot \sum_{l=1}^r (SSY_l \cdot w_{jlnorm}^2)}{SSY_{tot}}}, \quad (3.48)$$

with m number of X-variables
 r number of calculated PLS-components,

of a X-variable j , the fraction of PLS-components l describing Y-variance SSY_l ,

$$SSY_l = c_l^2 \cdot t_l^T \cdot t_l, \quad (3.49)$$

with t_l $(n \times 1)$ score vector of PLS-component in X
 c_l coefficient of intrinsic relation of the PLS model for component l ,

is multiplied with the normalised PLS-weights $w_{jl\text{norm}}$,

$$w_{jl\text{norm}} = \frac{w_{jl}}{\sqrt{w_l^T \cdot w_l}}, \quad (3.50)$$

with w_{jl} weight of component l for variable j
 w_l $(m \times 1)$ weighted column vector of component l ,

and divided by the total model-describing Y-variance SSY_{tot} ,

$$SSY_{\text{tot}} = \sum_{l=1}^r SSY_l. \quad (3.51)$$

The intrinsic coefficient c_l ,

$$c_l = \frac{u_l^T \cdot t_l}{t_l^T \cdot t_l}, \quad (3.52)$$

with u_l $(n \times 1)$ Y-score vector of PLS-component l ,

can be calculated with the score vector of both X and Y of the model (Eriksson et al., 2006b).

All VIP-values larger than 1 indicate an important variable and values lower than 0.5 indicate unimportant variables. The interval in-between is a grey area, depending on the data set size (Eriksson et al., 2006b).

3.4.6.5 Validation of Multivariate Model

The process of validation is one of the most important steps in multivariate modelling for evaluating the number r of calculated PCs. The portion of model variance R^2X ,

$$R^2X = 1 - \frac{1}{n \cdot m \cdot s_X^2} \cdot \sum_{i=1}^n \sum_{j=1}^m e_{ij}^2, \quad (3.53)$$

of the total variance s_X^2 ,

$$s_X^2 = \frac{1}{n \cdot m} \cdot \sum_{i=1}^n \sum_{j=1}^m (x_{ij} - \bar{X})^2, \quad (3.54)$$

of data matrix X , is a parameter which indicates the quality of the model describing the data matrix X , with the mean of the data matrix \bar{X} ,

$$\bar{X} = \frac{1}{n \cdot m} \cdot \sum_{i=1}^n \sum_{j=1}^m x_{ij} \quad (3.55)$$

Additionally, the portion of model variance R^2Y ,

$$R^2Y = 1 - \frac{1}{n \cdot v \cdot s_Y^2} \cdot \sum_{i=1}^n \sum_{h=1}^v f_{ih}^2, \quad (3.56)$$

with f_{ih} element of $(n \times v)$ residual matrix F ,

describes the quality of prediction. In PLS, the parameter R^2Y is more relevant than R^2X , as the prediction power is more important than the fit to the model X .

The quality of prediction R_{Ph}^2 for a variable h ,

$$R_{Ph}^2 = 1 - \frac{\sum_{i=1}^{n_{VS}} (y_{VSih} - \hat{y}_{VSih})^2}{(y_{VSih} - \bar{y}_{VSih})^2}, \quad (3.57)$$

with n_{VS} number of observations of validation set (VS)
 y_{VSih} measurement value of target variable y_h for object i in VS
 \hat{y}_{VSih} estimated value of target variable y_h for object i in VS
 \bar{y}_{VSih} mean value of target variable y_h in VS,

corresponds to the coefficient of determination R^2 and uses values of the validation data set (VS). This is a parameter for linear fittings of estimated model values plotted against the reference measurements.

The root mean square error of predictions $RMSEP_h$,

$$RMSEP_h = \sqrt{\frac{1}{n_{VS}} \cdot \sum_{i=1}^{n_{VS}} (y_{VSih} - \hat{y}_{VSih})^2}, \quad (3.58)$$

is the most important metric for the evaluation of the validation of the model in order to predict the target variable y_h . In order to compare different models based on different scales, relative $RMSEP$ $RMSEP_{rel}$,

$$RMSEP_{rel} = \frac{RMSEP}{Y_{CSmax} - Y_{CSmin}} \cdot 100\%, \quad (3.59)$$

with Y_{CSmax} maximum value of target variable y_h in calibration set (CS)
 Y_{CSmin} minimum value of target variable y_h in CS,

for external validation of PLS regression models is used. If there is no appropriate validation set available, an internal validation is used, also denoted cross validation (CV). Each sample of the CS is excluded once from the modelling process, and then predicted with the generated sub-model. When cross-validating, the introduced metrics for model validation are denoted R_{CV}^2 or $RMSE_{cv}$, respectively (Eriksson et al., 2006b; Martens & Naes, 1989).

3.4.6.6 Check for Overfit

A model is desired which serves on one side an appropriate goodness of fit R^2Y , and on the other side an appropriate goodness of prediction Q^2 (Figure 3.15). Since the fit is a measure for the mathematical reproducibility of the CS, the predictive ability describes the reliability of the prediction. With increasing calculated PLS-components, the R^2Y usually steadily increases until unity 1. Q^2 on the contrary, does not increase continuously and will not automatically approach 1 with increasing model complexity. In fact, at a certain number of components, the predictive ability does not improve any further and reaches a plateau. After this component, the Q^2 decreases (Eriksson et al., 2006b). The theoretical $R^2Y = 1$ mostly occurs with an overfitted model due to systematic variance originating from, e.g. measured noise. It comprises very accurate data description on one side, but a low predictive power R_{ph}^2 on the other side (Eriksson et al., 2006a; Eriksson et al., 2006b).

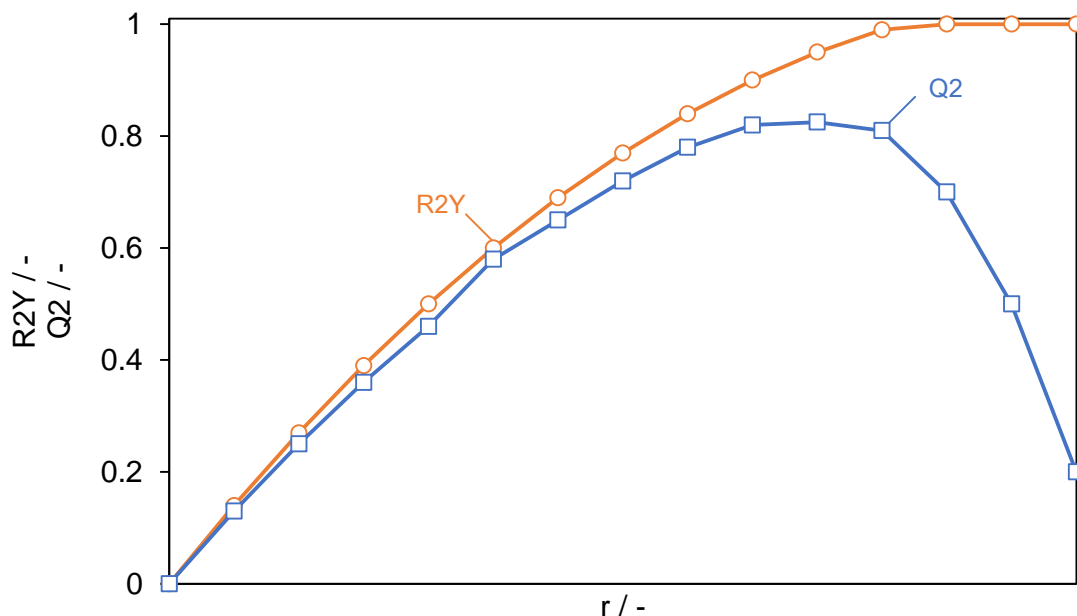


Figure 3.15: Balance between goodness of fit R^2Y and goodness of prediction Q^2 . Vertical axis corresponds to the amount of explained or predicted variation, R^2Y and Q^2 , respectively. Horizontal axis displays the model complexity, number r of PLS-components. At a certain number of PLS-components the most valid model is obtained with an optimal balance between fit and predictive ability (modified after (Eriksson et al., 2006b)).

For visual investigation of overfitting or validation, the permutation plot can be used (Figure 3.16). One limitation of cross-validation is the fact that it assesses only the predictive power, while no information about the statistical significance of the estimated predictive power is provided. In order to obtain an estimate of the significance of a Q^2 value, numerous parallel models based on fit to randomly reordered Y -data is developed. Then, the real Q^2 value is compared with distributed Q^2 values of the reordered response data Y . In the CS, the X -data are left intact while the Y -data are randomly shuffled, appearing in different order. A PLS model is then fitted to the permuted data. By using cross-validation, both R^2Y and Q^2 values are calculated and compared with the R^2Y and Q^2 values of the real model. In the next step, a second PLS model is fitted to another permuted version of the Y -data and compared with the real values. By repeating this cycle a number of times, ideally between 25–100 times, distributions based on random data can be obtained (Eriksson et al., 2006b).

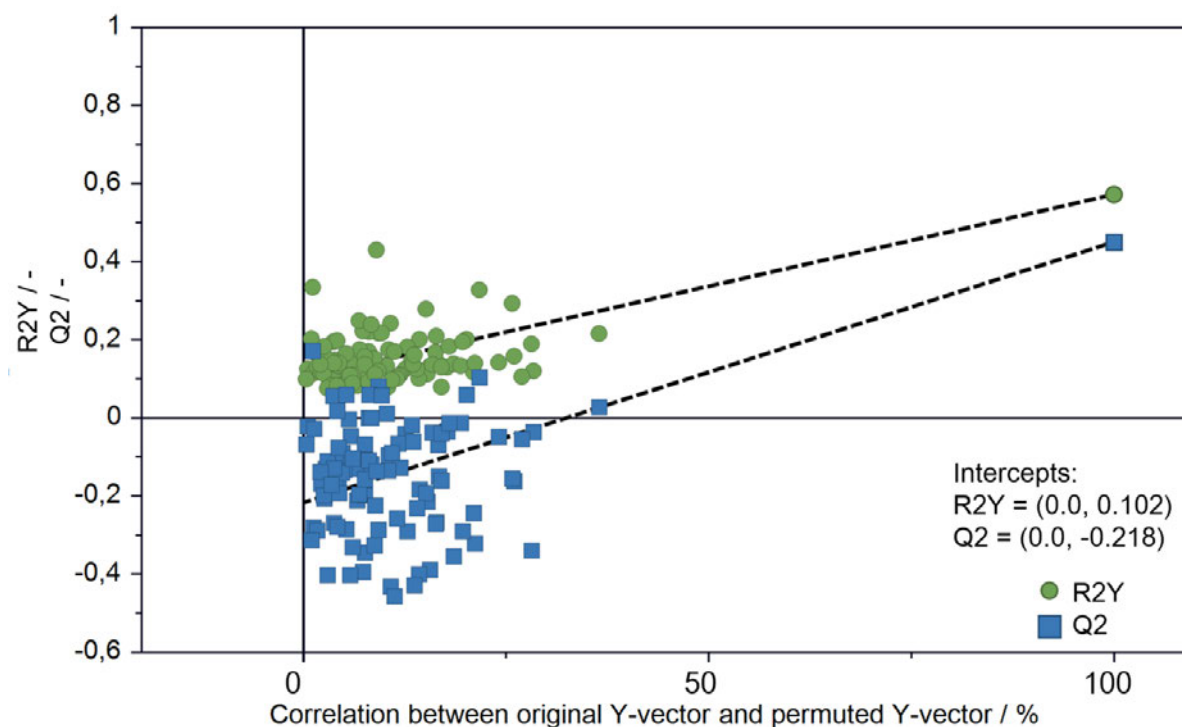


Figure 3.16: Permutation plot for validation and overfit detection. For valid models, R^2Y -intercept should not exceed 0.3–0.4 and Q^2 -intercept should not exceed 0.05.

A permutation plot comprises a vertical scale representing the R^2Y and Q^2 values of all models and a horizontal scale representing the correlation coefficients between permuted and original response. The original R^2Y and Q^2 are located in the right part of each plot at 1. One is the coefficient obtained when correlating a variable with itself. A regression line is fitted amongst the permuted and original R^2Y and Q^2 values, respectively. The intercepts of the regression lines are interpretable as measures of this plot. R^2Y -intercept should not exceed 0.3–0.4 and

Q2-intercept should not exceed 0.05. Intercepts below these limits are considered valid and well fitted models (Eriksson et al., 2006b).

Another measure for overfitting is the mean bias error MBE,

$$MBE = \frac{1}{n} \cdot \sum_{i=1}^n (x_{pi} - x_{oi}), \tag{3.60}$$

with x_{pi} predicted value for observation i
 x_{oi} observed value for observation i ,

representing the systematic error of a prediction model. The sign indicates the direction of the error. Negative MBE implies underprediction and positive MBE indicates overprediction on an average (Pal, 2017).

3.4.7 Orthogonal Projections to Latent Structures

The orthogonal projections to latent structures by means of partial least squares regression (OPLS) is a variation of the previously introduced PLS (Wold et al., 1998). This method was proposed as a filtration method that could either replace or complement signal correction methods such as MSC or SNV, facilitating the interpretation of predictive variation. While traditional filter methods are useful and often also necessary, it is usually not obvious whether predictive or orthogonal systematic variation is filtered out (Stenlund, 2011). Therefore, its aim is to improve the interpretation of PLS models and reduce model complexity (Figure 3.17) (Trygg & Wold, 2002).

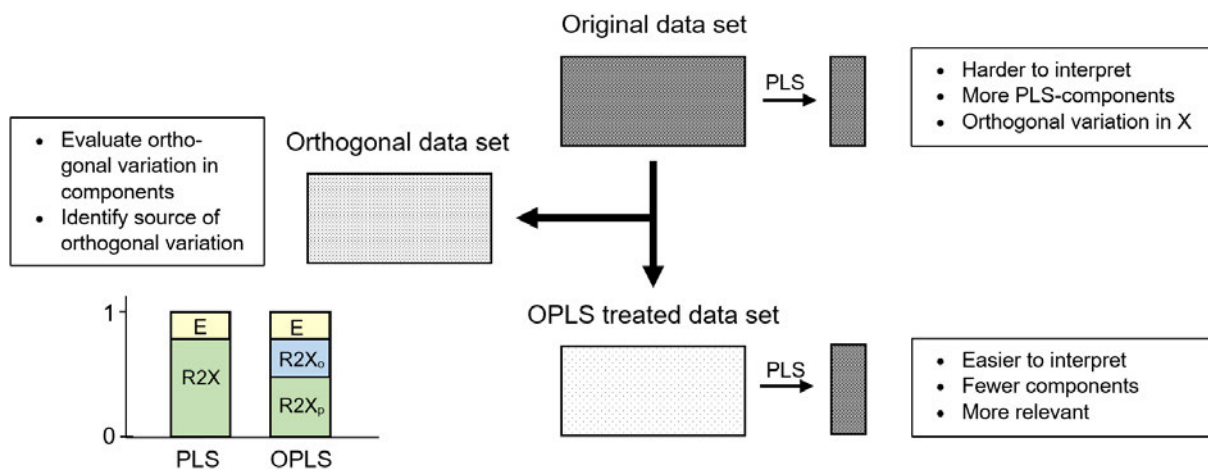


Figure 3.17: Overview of orthogonal projections to latent structures (modified after (Trygg & Wold, 2002)).

OPLS modelling separates the systematic variation from input data X into a part that is related to Y and a part that is unrelated (orthogonal) to Y (X - Y joint variation) (Figure 3.18). The

unrelated part originates from experimental or analytical nature, e.g. calibration transfers, detection limits, or time trends. The other part corresponds to variations in identified biochemical trends (Stenlund, 2011).

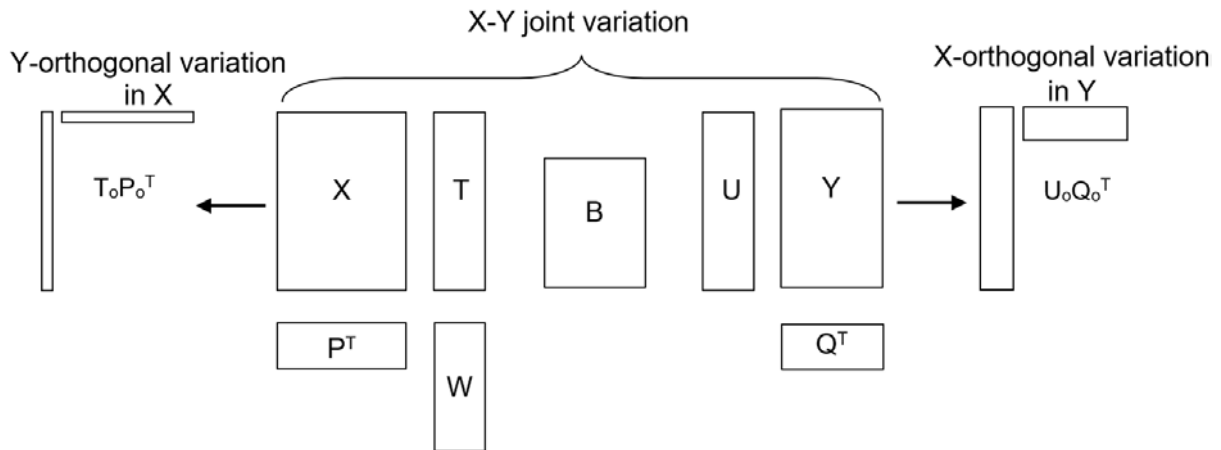


Figure 3.18: Schematic representation of orthogonal projections to latent structures (OPLS). OPLS is an extended version of the projections to latent structures. Index o: orthogonal, index p: predictive (modified after (Gabrielsson et al., 2006)).

Mathematically, the model for PLS is extended by the calculation of the orthogonal parts to Y. This is done by applying the NIPALS algorithm. For further information on the mathematical background see literature of Trygg and Wold (Trygg & Wold, 2002).

4 Material and Methods

The cultivations and analysis were executed in the laboratories for bioprocess automation at the Hamburg University of Applied Sciences. The following chapter deals with the used materials and applied methods.

4.1 Cell Line

The yeast strain *Pichia pastoris* (*P. pastoris*) BSYBG11 was used during this project for production of enhanced green fluorescent protein (eGFP). However, the measurement of eGFP was not scope of this work. Stock cultures ($OD_{600} = 5.55$) of this strain were stored in 25 % glycerol at $-80\text{ }^{\circ}\text{C}$.

4.2 Medium

The cultivations were executed with defined minimal medium FM22 which is described in Table 4.1. This medium was modified after Stratton and colleagues (Stratton et al., 1998).

Table 4.1: Medium composition of FM22.

Component	Article No.	Manufacturer	Concentration / g L^{-1}
$\text{C}_3\text{H}_8\text{O}_3$	3783.2	Carl Roth	$20.0^{(a)}/25.0^{(b)}$
KH_2PO_4	3904.3	Carl Roth	25.79
$(\text{NH}_4)_2\text{SO}_4$	3746.4	Carl Roth	5.00
K_2SO_4	P022.3	Carl Roth	8.60
$\text{CaSO}_4 \cdot 2\text{H}_2\text{O}$	P741.3	Carl Roth	1.40
$\text{MgSO}_4 \cdot 7\text{H}_2\text{O}$	T888.3	Carl Roth	16.4
$\text{Na}_3\text{-citrate} \cdot 2\text{H}_2\text{O}$	A12274	Alfa Aesar	6.81
0.2 g L^{-1} biotin stock	3822.1	Carl Roth	8 mL L^{-1}
PTM ₄ stock	–	–	4 mL L^{-1}

(a) preculture; (b) main culture

PTM₄ is a trace element solution containing components mentioned in Table 4.2.

Table 4.2: Composition of trace element solution PTM₄ stock.

Component	Article No.	Manufacturer	Concentration / g L ⁻¹
CuSO ₄ · 5 H ₂ O	209198	Sigma-Aldrich	2.00
NaI	8783.1	Carl-Roth	0.0800
MnSO ₄ · H ₂ O	7347.1	Carl Roth	3.00
Na ₂ MoO ₄ · 2 H ₂ O	31439	Riedel de Haën	0.200
H ₃ BO ₃	31146	Riedel de Haën	0.0200
CaSO ₄ · 2 H ₂ O	T888.3	Carl Roth	0.500
CoCl ₂ · 6 H ₂ O	255599	Sigma-Aldrich	0.500
ZnSO ₄ · 7 H ₂ O	221376	Sigma-Aldrich	7.00
FeSO ₄ · 7 H ₂ O	P015.1	Carl Roth	22.0
96 % H ₂ SO ₄	30743	Sigma-Aldrich	1.00 mL

Both PTM₄ stock and biotin stock were sterile filtered with 0.22 µm pore size cellulose nitrate filters and stored until use at -4 °C.

For media preparation, FM22 medium was autoclaved (Systec VX-150, Linden, Germany) at 121 °C for 20 min and both biotin and PTM₄ stocks were aseptically combined with FM22 medium in the laminar flow cabinet (Heraeus Instruments, Germany). The pH value was adjusted with 1 M and 25 % ammonium hydroxide to 4.8 in precultures and to 5.0 in the bioreactor, respectively.

For acid, base, and anti-foam titration during cultivation, the reservoirs and corresponding concentrations can be found in Table 4.3. The reservoirs with exception of T2 (ammonia) and R2 (methanol) were autoclaved at 121 °C for 20 min.

Table 4.3: Reservoirs for cultivation.

Reservoir	Designation	Component	Art. No.	Manufacturer	Concentration
R1, glycerol	C _{S1R1}	C ₃ H ₈ O ₃	3783.2	Carl Roth	630 g L ⁻¹
R2, methanol	C _{S2R2}	CH ₃ OH	4627.1	Carl Roth	790 g L ⁻¹
T1, acid	–	H ₃ PO ₄	6366.1	Carl Roth	1.5 M
T2, base	–	NH ₃	5460.3	Carl Roth	25 %
Anti-foam agent	–	Struktol®	J673	Schill+Seilacher	100 %

4.3 Preculture

In order to cultivate *P. pastoris* BSYBG11 to high cell densities, precultures grown in small volumes are required for inoculation of the bioreactor. This prevents excessive lag-phases, maintains cell viability, and decreases process costs (Keil et al., 2019). 3.00 mL of stock cultures were divided into ten 1 L shake flasks containing FM22 medium such that the inoculum volume totals 10 % of the initial bioreactor working volume. The shake flasks were incubated at 30 °C and 150 min⁻¹ (Certomat® BS-1, B. Braun Biotech, Germany) for 28 h before they were aseptically transferred to the bioreactor via a transfer bottle. Prior to inoculation, the optical density at 600 nm (OD₆₀₀) of the transfer bottle content was measured.

4.4 Bioreactor System BIOSTAT® C30

The main cultivation was carried out in the *in-situ* sterilisable bioreactor BIOSTAT® C30 (Sartorius Stedim Biotech, Germany) with a total volume of 42 L and an aspect ratio h/d of 2:1 (Figure 3.6). The bioreactor is equipped with a stainless-steel lid comprising six 19 mm ports for a safety valve, an agitation system with double mechanical seal and 6-blade disk impeller, an exhaust cooler, a ring sparger aeration, a level probe, and a sight glass for illumination. The ports for acid, base, and anti-foam titration are also located on the lid through connections and mountings. The jacketed stainless-steel vessel comprises five 25 mm connection ports for feeding medium such as glycerol and methanol. Also, a longitudinal viewing window is positioned on the upper vessel wall. Both sampling and drain valve and another seven ports are located at the lower vessel wall for in-line electrodes. These are applied for determination of pH value, relative dissolved oxygen, temperature, turbidity, methanol, and for Raman spectroscopy.

For automatic data acquisition and managing activities of the peripherals, the electrodes, balances (Sartorius, Germany), peristaltic pumps (Watson Marlow 101U, UK), and off-gas analyser BlueInOne (BlueSense gas sensor, Germany) are connected to the digital control unit (DCU) DCU 3 (Sartorius Stedim Biotech, Germany). The DCU allows manual operation such as adjusting set points or controller settings and is connected to a software for supervisory bio-process control and data acquisition, enabling automatic operation. This software, denoted multi-fermenter control system (MFCS/win 3.0 (Sartorius Stedim Biotech, Germany)), will be further described in Chapter 4.4.3. The control system Simatic PCS7 (Siemens, Germany) was used for emptying the vessel after each run.

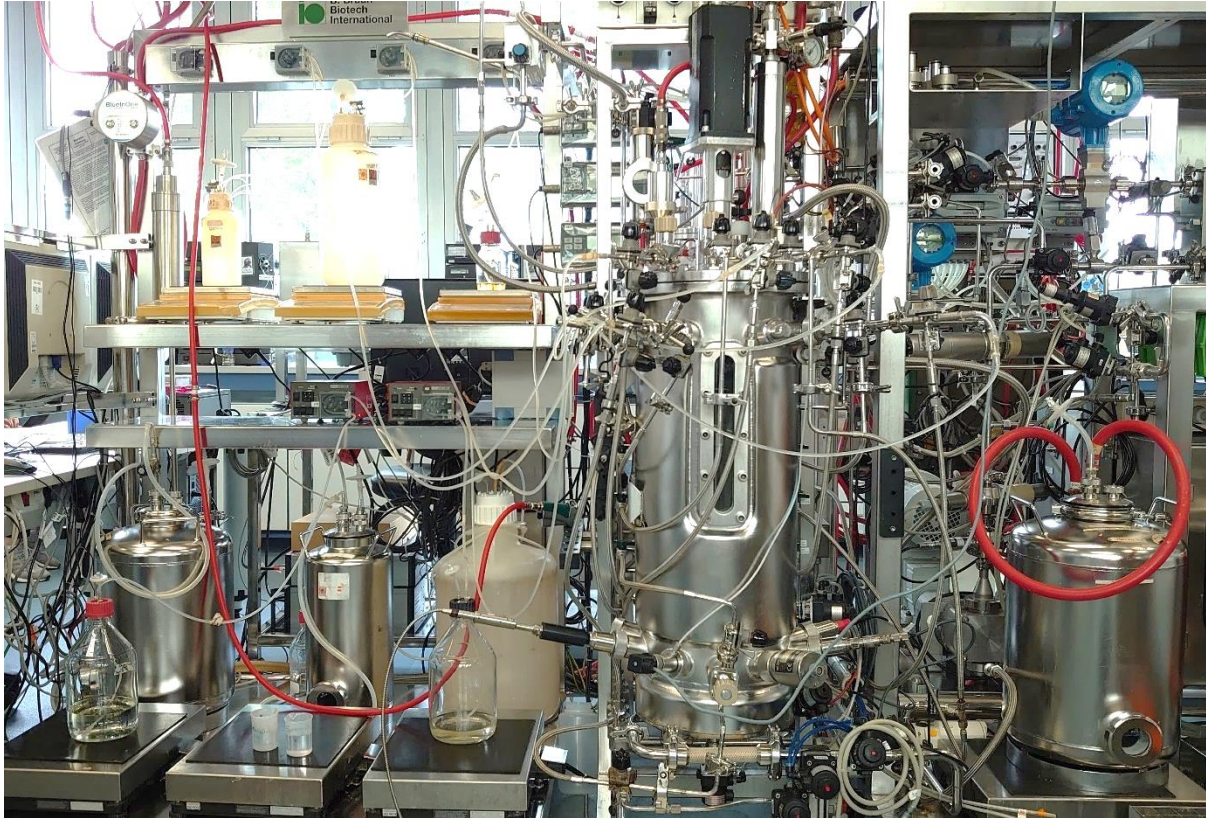


Figure 4.1: Bioreactor BIostat® C30 and peripherals.

The double mechanical seal was continuously maintained 0.5 bar above the pressure of the cultivation vessel in order to avoid any fluids entering the seal. Before and after the culture vessel, the supply and exhaust air, respectively, were sterile filtered with 0.2 μm Sartofluor® filters (Sartorius, Germany). The steam generator Steamboy-9 (ZIRBUS technology, Germany) was used in order to allow sterilisation of titration agent ports before the cultivation process and frequent sterilisation of sampling valve during cultivation.

Prior to sterilisation of the culture vessel, pH and methanol probes were calibrated. The pH was calibrated with commercial buffer solutions of pH 4.01 and 7.0 (Carl Roth, Germany). Then, the fully mounted culture vessel filled with culture medium was *in-situ* sterilised. Here, the zero-point calibration of the pO_2 probe was applied. After ending the sterilisation, pO_2 slope calibration was executed and the heat-labile stocks were aseptically added to the culture medium.

4.4.1 Cultivation Conditions

For cultivation in BIostat[®] C30, the initial working volume valued 12 L such that all three blade impellers were covered with culture broth. The relative dissolved oxygen was controlled at 25 % by the agitation master controller (Figure 4.2).

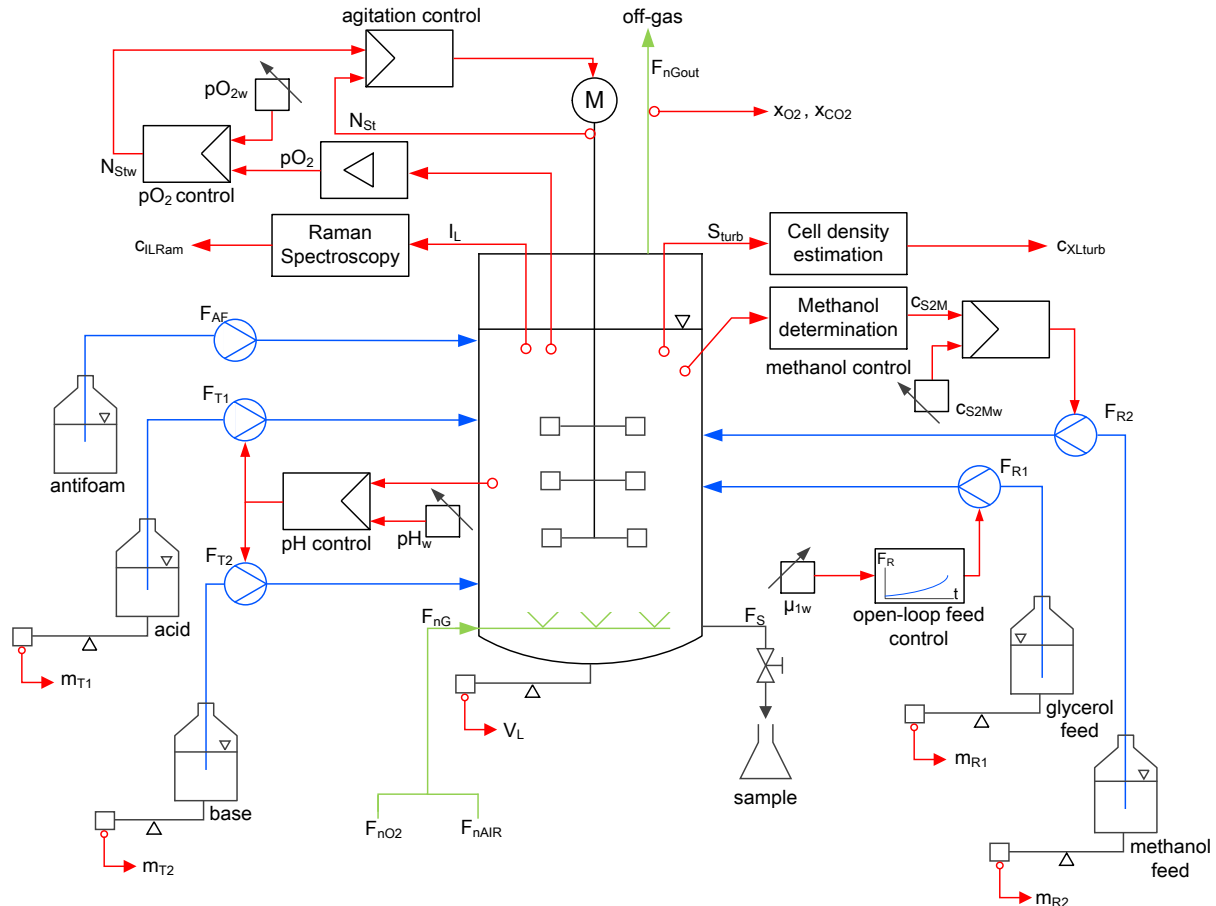


Figure 4.2: Simplified piping and instrumentation diagram and automation tasks of bioreactor system used.

The batch phase was initiated by addition of the preculture to the cultivation vessel. In this phase, the cells were growing with maximum cell-specific growth rate $\mu_{1\max}$ until glycerol depletion.

In this work, an automatic batch end detection tool (Voß, 2017) was used, based on the fact that substrate depletion causes a decrease of cell growth during pO_2 /agitation control. With this, a lower oxygen is demanded. Therefore, the ongoing stirrer speed causes a sudden pO_2 increase which in turn means that the stirrer speed is decreased by the controller. The change of both process parameters in a pre-defined time interval is observed by the software MFCS-Tool in order to induce a change of cultivation phase.

In order to determine the end of the batch phase, the maximum cell density c_{XLmax} is required,

$$c_{XLmax} = c_{XL0} + Y_{X/S} \cdot c_{S1L0} , \quad (4.1)$$

with	c_{XLmax}	maximum cell concentration of batch phase	$g L^{-1}$
	c_{XL0}	initial cell concentration of batch phase	$g L^{-1}$
	$Y_{X/S}$	substrate yield coefficient	–
	c_{S1L0}	initial substrate concentration of batch phase	$g L^{-1}$,

to calculate the batch end time $t_{batchend}$,

$$t_{batchend} = \frac{\ln \frac{c_{XLmax}}{c_{XL0}}}{\mu_{1max}} , \quad (4.2)$$

with	μ_{1max}	max. cell-specific growth rate in batch phase	h^{-1} ,
------	--------------	---	------------

After this, the fed-batch phase was started with a pre-defined glycerol feeding rate F_{R1w} . The feeding rate was exponentially increased such that a cell-specific growth rate $\mu_w < \mu_{1max}$ was maintained, resulting in a substrate limited growth.

The relative glycerol feeding rate F_{R1relw} ,

$$F_{R1relw}(t) = \frac{\mu_w(t_j) \cdot V_L(t_j) \cdot c_{XL}(t_j)}{y_{X/S} \cdot c_{S1R1}} \cdot \frac{100 \%}{F_{R1max}} \cdot e^{\mu_w(t_j) \cdot (t-t_j)} , \quad (4.3)$$

with	μ_w	set point of cell-specific growth rate	h^{-1}
	t_j	time point of fed-batch start	h
	V_L	volume of liquid phase	L
	c_{XL}	cell concentration	$g L^{-1}$
	$Y_{X/S}$	substrate yield coefficient	–
	c_{S1R1}	glycerol concentration in reservoir	$g L^{-1}$
	F_{R1max}	maximum glycerol feeding rate	$L h^{-1}$,

was implemented in the MFCS-Tool in order to control the glycerol feeding pump for the exponential feeding course.

The fed-batch phase was then followed by the production phase when cell density reached $> 40 g L^{-1}$. Here, the glycerol feeding pump rate was set to $F_{R1} = 0 L min^{-1}$ and the methanol feeding pump rate F_{R2} was initially set to 100 % (= $0.0245 L min^{-1}$) for 2 min to induce a metabolism switch of the yeast cells. This led to an initial methanol concentration of $3 g L^{-1}$ in the culture broth. The methanol concentration set point was maintained at $c_{S2Mw} = 1.5 g L^{-1}$ by on-line monitoring with the methanol probe and methanol reservoir pump. Other cultivation conditions can be found in Table 4.4.

Table 4.4: Cultivation Conditions.

Process Set Points	Designation	Batch	Fed-Batch	Production
Methanol concentration	c_{S2Mw} / $g L^{-1}$	0.0	0.0	1.5
Aeration rate	F_{nGw} / vvm	1.5	1.5	1.5
Pressure	p_{Gw} / $mbar$	500	500	500
pH value	pH_w / $-$	5	5	5
Relative dissolved oxygen	pO_{2w} / $\%$	25	25	25
Temperature of liquid phase	ϑ_{Lw} / $^{\circ}C$	30	30	22
Cell-specific growth rate	μ / h^{-1}	μ_{1max}	0.10	μ_{2max}

For simplification, different cultivations are named after following scheme in this work. The first two letters describe the BIOSTAT® C30 bioreactor system, followed by initials of the operator. Then, the calendar week and the year is transcribed, e.g. XXPC0922, describing a cultivation with the BIOSTAT® C30 bioreactor by Phoebe Chan in the ninth week of year 2022.

4.4.2 Standard Measurement Systems

Both on-line and in-line measurement systems were used for the cultivation. Standard measurement probes include relative dissolved oxygen partial pressure pO_2 (in the following simply denoted dissolved oxygen, DO), pH value, temperature ϑ_L , and high level and foam alarm (Table 4.5). The bioreactor comprises an aeration system ($F_{nAIRmax} = 30 L min^{-1}$) and an external gas mixer in order to allow oxygen sparging ($F_{nO2max} = 5 L min^{-1}$). The pressure of the bioreactor is monitored via a pressure probe above the exhaust cooler. For determination of oxygen and carbon dioxide in the off-gas, the off-gas analyser BlueInOne was used.

Table 4.5: Process parameters and corresponding measurement systems.

Process Variables	Designation	Time Response	Measurement System
pH value	pH	In-line	Redox electrode
Temperature	ϑ_L	In-line	Pt-100
Pressure	p_G	On-line	Piezoelectric membrane
Stirrer speed	N_{St}	On-line	-
Volume	V_L	On-line	Mass balances
Off-gas O_2 mole fraction	x_{O2}	On-line	Galvanic cell
Off-gas CO_2 mole fraction	x_{CO2}	On-line	Non-dispersive infrared
High level/foam alarm	-	On-line	Capacitance switch
Dissolved oxygen	pO_2	In-line	Clark electrode

4.4.3 MFCS/win

MFCS/win 3.0 is a bioprocess data management and automation software (Sartorius Stedim Biotech, Germany) for supervisory control and data acquisition (SCADA). This software uses International Society of Automation (ISA)-88 recipes, a standardised procedure to integrate, communicate, and configure batches. Alarms and set points can be implemented by MFCS/win in order to ensure robust automation performance and to minimise batch-to-batch variations.

The device, which MFCS/win is implemented in, is connected to other physical devices such as the DCU, external balances, pumps, and off-gas analysers to allow the control and data acquisition.

4.4.4 Turbidity Probe

For on-line biomass monitoring, turbidity probes are mostly used in bioreactors. The measurement principle is based on the attenuation of light when entering the liquid phase and is dependent on dispersed particles in the suspension. However, there is no distinction between active and dead biomass while measuring (Madrid & Felice, 2005).

The used turbidity probe ASD25-BT-N-5 (optek-DANULAT, Germany) works with an optical path length of 5 mm and is a single-channel absorption photometer using light in the NIR spectrum of 840-910 nm (Figure 4.3). The signal is transmitted to the transducer Control 4000 (optek-DANULAT, Germany) for indication, saving, and editing the measured values. The data acquisition of the analogue signal S_{turb} is possible between $0 < \text{AU} < 4$ by MFCS/win.



Figure 4.3: Used turbidity measurement system. A) Turbidity probe; B) Transducer Control 4000 (optek-DANULAT, 2022)

The cell concentration determined by the turbidity probe c_{XLturb} was calculated by

$$c_{XLturb}(t) = a \cdot (e^{b \cdot S_{turb}(t)} - 1) \cdot \left(\frac{N_{St}(t)}{N_{Stmax}} \right)^c, \quad (4.4)$$

with	a	adaption parameter	$g L^{-1}$
	b	adaption parameter	AU^{-1}
	S_{turb}	turbidity signal	AU
	c	adaption parameter	–.

The adaption parameters were post-experimentally determined by fitting c_{XLCDW} against c_{XLturb} by using the simplex algorithm of Nelder-Mead with MATLAB.

4.4.5 Methanol Probe

For determination of methanol concentration, the on-line probe Alcoline® (Biotechnologie Kempe, Germany) was used (Figure 4.4). The probe comprises a permeable silicone membrane, allowing volatile substances such as methanol or ethanol to pass the membrane. An inert carrier gas transports the volatile substance to a gas sensor, causing a change of electrical capacity. Hereby, the electrical resistance of the sensor is decreased and information about the methanol content can be read off the Fermentation Mini Computer (FMC) (Biotechnologie Kempe, Germany). The output signal was transmitted to MFCS/win in order to enable methanol concentration control by the corresponding reservoir pump.

Prior to cultivation, the methanol probe was calibrated by a three-point calibration at initial cultivation conditions. After each addition of defined volume of methanol, the system was equilibrated for 10 min before reading off the electrical resistance from the FMC.

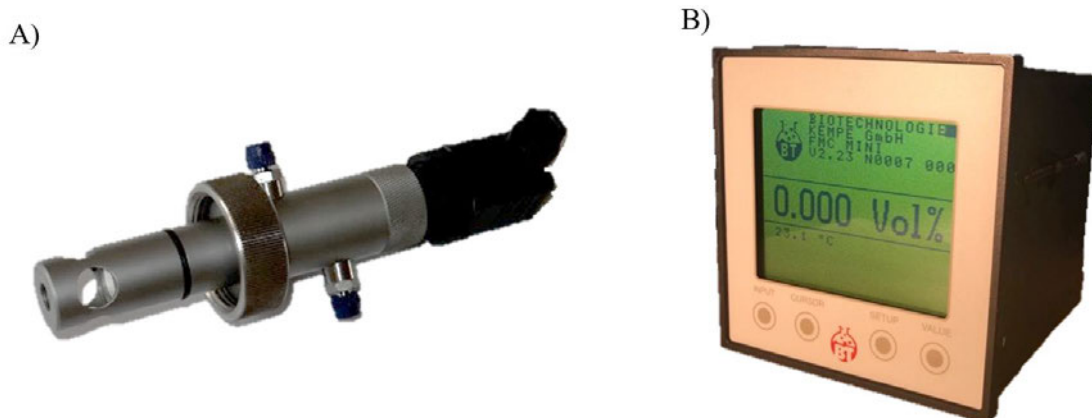


Figure 4.4: Used methanol content measurement system. A) Methanol probe; B) FMC Mini (modified after (Biotechnologie Kempe, 2022)).

4.5 Raman Spectroscopy

In this work, the Raman devices RamanProbe™ (InPhotonics, USA) and tecRaman Sonde 785 Küvette (tec5, Germany) were used (Figure 4.5). The Raman probe was adapted for insertion into a standard 19 mm bioreactor port. Both in-line and off-line devices were coupled to the light source MultiSpec® Raman Spektrometer (tec5, Germany), and to the multiplexer MUX-4P (tec5, Germany), allowing sequential selection of analogue or digital signals of up to four different input channels. While the immersion probe is based on Stokes backscattering configuration, the off-line spectrometer used transmission Raman.

The light source comprises a NIR diode laser of class 3B with an excitation wavelength of $\lambda_{\text{ex}} = 785 \text{ nm}$. The maximum power values 500 mW. For the in-line Raman probe, the laser focus is 5 mm ahead the probe tip. For the off-line probe, however, the focal plane values 25 mm. The laser output of the probes is collected in a CCD detector for spectra recording in the range of $75\text{--}3215 \text{ cm}^{-1}$ where a resolution of up to 1 cm^{-1} is possible. The spectra were visually displayed in the corresponding software MultiSpec® Pro II (tec5, Germany).

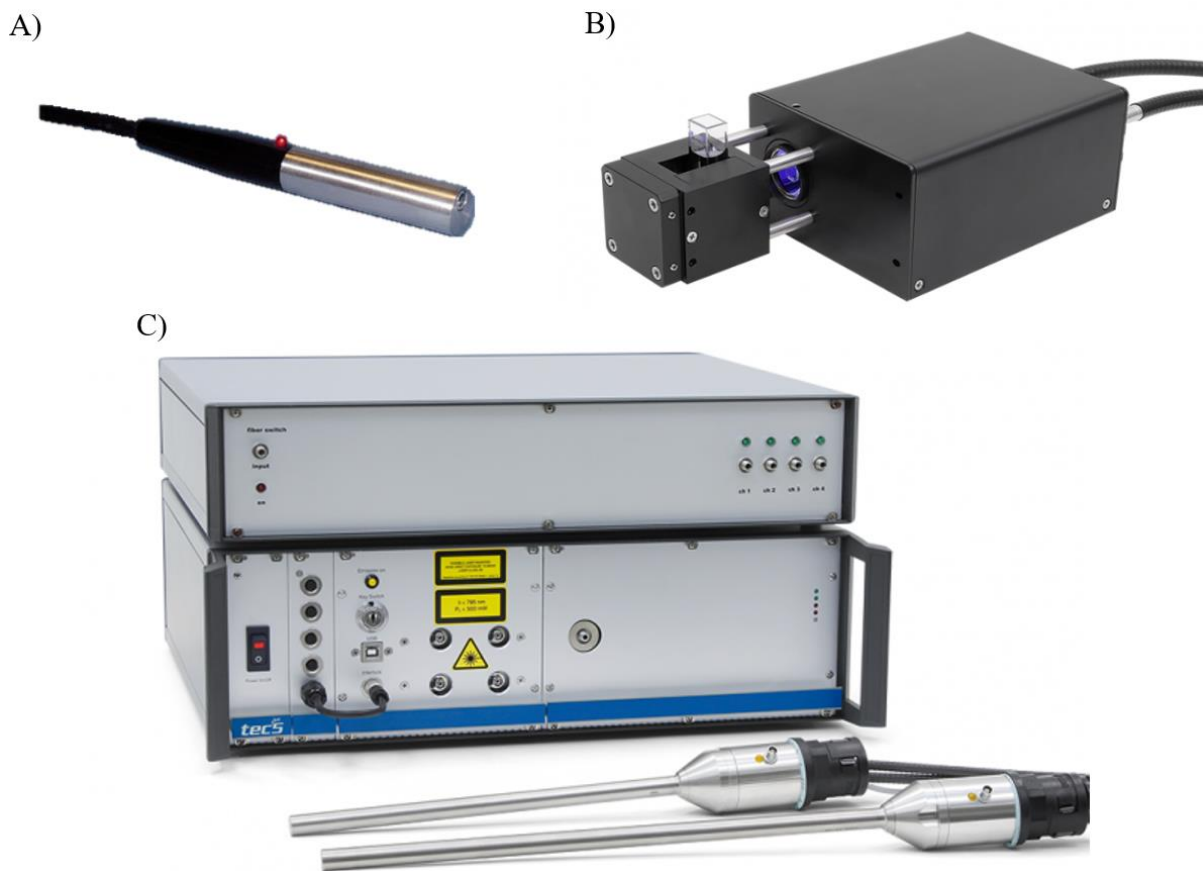


Figure 4.5: Used Raman spectrometer system. A) In-line Raman probe without bioreactor port mounting (modified after (InPhotonics, 2022)); B) Off-line Raman spectrometer; C) Light source and multiplexer (modified after (tec5, 2022))

Measurements of the off-line Raman system were carried out in 10 mm layered 3.5 mL quartz cuvettes. Both culture broth and supernatant were measured with the off-line system while the in-line probe measured only the culture broth. The glass window of the bioreactor was covered with a metal cover in order to sustain complete darkness inside the reactor vessel and for safety reasons with the laser class 3B.

4.6 Analytical Methods

During cultivation, samples were taken for analysis on the concentration of cell dry weight (CDW), glycerol, methanol, for fluorescence, and for optical density (OD) determination. The following chapter deals with the applied methods.

For measurement of cell density, two distinct methods were used. In order to obtain direct estimation about cell growth, the OD was measured while CDW allows more accurate values due to higher independence of external influences. If not mentioned explicitly, all data were measured in duplicates.

For probing of the bioreactor, the first 20 mL of cell suspension was disposed before another 20 mL served as sample to be examined.

4.6.1 Cell Dry Weight Concentration

To determine the CDW concentration c_{XLCDW} , 1 mL of cell suspension was transferred into a dried, weighed microreaction tube and centrifuged for 10 min at 14,462 g (Centrifuge 5417, eppendorf, Germany). The supernatant was frozen at $-20\text{ }^{\circ}\text{C}$ for determination of metabolites and medium components at a later time point. The pellet was put in the drying cabinet (Heraeus, Germany) for 24 h at $104\text{ }^{\circ}\text{C}$ until weight constancy was reached and was then weighed.

The CDW concentration c_{XLCDW} ,

$$c_{XLCDW}(t) = \frac{m_{Xdry}(t) - m_0}{V_{sample}}, \quad (4.5)$$

with	m_{Xdry}	weight of loaded tube after drying	g
	m_0	weight of empty tube	g
	V_{sample}	sample volume in tube	L,

was calculated using the quotient of weight difference between before and after loading the tube, and the loaded sample volume V_{sample} .

4.6.2 Optical Density Determination

The OD measurement was executed during cultivation at 600 nm with a spectrophotometer (Ultra Spec 3000 pro, Amersham Pharma Biotech, UK) where demineralised water (DW) was used as blank. The OD must be in the range of $0.1 < OD < 0.6$ in order to be in the linear range of the spectrophotometer and to ensure reliable measurement. If the limit was exceeded, the cell suspension was diluted with DW. Incorporating the dilution factor, the actual OD_{600} ,

$$OD_{600}(t) = D_F \cdot OD_{meas}(t) , \quad (4.6)$$

with	OD_{meas}	diluted optical density at 600 nm	g
	D_F	dilution factor	–,

can be calculated.

Based on the correlation factor $K_{X/OD}$, the cell density c_{XLOD} ,

$$c_{XLOD}(t) = K_{X/OD} \cdot OD_{600}(t) , \quad (4.7)$$

with	OD_{600}	optical density at 600 nm	–
	$K_{X/OD}$	correlation factor between cell density and OD	$g L^{-1}$,

was determined.

4.6.3 Off-line HPLC for Glycerol and Methanol Determination

For off-line quantification of glycerol and methanol, the High-Performance Liquid Chromatography (HPLC) system LaChrom[®] (Hitachi High Technologies, USA) was used. The HPLC comprises pump system L-7100, autosampler L-7250, column oven L-7360, diode array detector L7455, and refractive index (RI) detector L-7490. The column Rezex[™] RHM-Monosaccharide H+ (8 %) 300 x 7.8 mm (Phenomenex, USA) was utilised with the preceding security guard cartridge Carbo-H 4 x 3 mm (Phenomenex, USA) in order to protect the column from impurities.

The working principle is based on ions exchanging on the column where sulfonic groups are fixed on the surface of the polystyrene-divinylbenzene resin in order to form a negatively charged shield, denoted Donnan membrane. This membrane allows the passing of non-ionic particles and therefore the elution from the column (Han, 1999).

For measurement, both standards and supernatant samples were acidified to a concentration of 10 mM H_2SO_4 . A mix of glycerol and methanol served as external standard. Then, the samples

were centrifuged at 14,462 g for 10 min and the supernatant was filtered using 0.2 μm syringe filters (Minisart, Sartorius, Germany) and transferred into single-use glass vials with 200 μL micro inserts (VWR, Germany). The column oven temperature was set to 60 $^{\circ}\text{C}$ and 10 mM H_2SO_4 was used as mobile phase. Before sample measurement, the HPLC system was washed with mobile phase at 0.1 mL min^{-1} for > 3 h. For measurement, the flow rate was increased to 0.6 mL min^{-1} and each sample was measured for 22 min.

The outcoming chromatograms were visualised and evaluated by the corresponding software D-7000 HSM (Hitachi High Technologies, USA). For peak evaluation of chromatograms, the integration of the peak is considered. This area under the curve (AUC) was investigated in order to overcome baseline drifts appearing.

4.6.4 Total Protein Concentration Determination

For quantification of the total proteins in the culture broth, the Roti[®]-Quant assay (Carl Roth, Germany) was performed according to Bradford (Bradford, 1976). The working mechanism is based on the interaction between the blue dye Coomassie Brilliant Blue-G250 (Carl Roth, Germany) with primary amino acids of the protein. Upon binding, the ionic state of the dye is changed from cationic to anionic form. With this, the absorption maximum is shifted from 470 nm to 595 nm (Bradford, 1976).

For high-throughput screening, the microplate reader Infinite[®] F Plex M200 Pro (Tecan, Switzerland) was used. Bovine serum albumin (BSA) served as standard. All samples were measured in triplicates in order to compensate pipetting errors in small volumes. The standard and sample preparation steps were handled according to the assay's instructions for microplates (Carl Roth, 2021). Finally, the samples were transferred to transparent 96 well plates (Sigma Aldrich, Germany) and the absorption was measured at 595 nm.

4.6.5 Fluorescence Determination

For quantification of fluorescence in the samples, the microplate reader Infinite[®] NanoQuant M200 Pro (Tecan, Switzerland) was used. 200 μL of supernatant were transferred to black 96 well plates (Sigma Aldrich, Germany). Then, the samples were excited at 485 nm and the emitted light at 535 nm was measured.

4.7 Data Evaluation with MVDA

The gathered data was evaluated using the software SIMCA[®] 17.0.1 (Sartorius Stedim Data Analytics, Sweden). Prior to data import into SIMCA[®], the raw data measured was prepared with MATLAB version 2022a and Excel. The MVDA-related plots were produced with SIMCA[®].

The aim of this work was the comparison of both off-line and in-line Raman spectroscopy by using OPLS. For this, the quality of the data set is crucial for the development of a multivariate data model. The data set requires representative observations, describing the whole process as much as required. E.g., in order to quantify the substrate concentration by an OPLS model and to avoid weighing of the data model, the whole relevant concentration range in a uniform distribution must be included in the data set.

4.7.1 Sample Pool

Three cultivations, XXPC0922, XXPC1722, and XXPC2622, were used for multivariate calibration. Both off-line and in-line Raman spectra were subject to this work. For in-line, only cell suspension (SUS) while for off-line both cell suspension and supernatant (SN) were examined. For SUS, the samples were measured as doublets while for SN, only a single measurement was executed.

4.7.2 Data Preparation

Before MVDA was implemented, the measured spectra were transferred into a $(n \times m)$ data matrix D . The smallest possible technical resolution was used in order to evaluate the necessity of the high resolution afterwards. With this, a data set of, e.g. 30 Raman spectra containing a wavenumber range of 200–3200 cm^{-1} is made of $n = 30$ observations and $m = 3000$ variables.

4.7.3 Data Pre-Processing

In order to investigate a set of pre-processing methods, a selection of 12 pre-processing tools were used (Table 4.6). For Savitzky-Golay (SG) smoothing, two options were applied. The first option included 9 cm^{-1} (SG9), the second option 15 cm^{-1} (SG15) points spacing in each moving polynomial. Both filters, however, use quadratic polynomials.

Table 4.6: Overview about pre-processing tools used in the four pre-processing steps.

Step 1:	Step 2:	Step 3:	Step 4:
Baseline correction	Scatter correction	Noise removal	Scaling
None	None	None	Ctr
1 st Der	SNV	SG, 9-point, 2 nd order	
2 nd Der	MSC	SG, 15-point, 2 nd order	
LinC		WDS	

1stDer: 1st order derivative, 2ndDer: 2nd order derivative; LinC: linear correction; SNV: standard normal variate; MSC: multiplicative scatter correction; SG: Savitzky-Golay; WDS: wavelet denoise spectral; Ctr: mean centred.

4.7.4 Approach in SIMCA[®] Environment

SIMCA[®] is a MVDA software for data mining, multivariate calibration, and predictive modeling. In the version 17 and above, spectroscopy modelling for PAT was improved, enabling more data pre-processing for pharmaceutical quality control. SIMCA[®] allows data analysis of multiple Y-variables simultaneously. In this work only single response evaluation was used.

The software offers the opportunity to easily pre-process data via the *Preprocessing wizard*. Four sections can be found comprising various pre-processing tools. The four sections are denoted *Smoothing*, *Baseline correction*, *Normalization*, and *Other* (Table 4.7).

Table 4.7: Overview about pre-processing methods offered in SIMCA[®] 17.0.1.

Category	Tool
Smoothing	Savitzky-Golay filter (SG)
	Exponentially weighted moving average
	Wavelet denoise spectral (WDS)
	Moving window
	Asymmetric least squares smoothing
Baseline correction	Row-centre
	Offset
	Linear correction (LinC)
	Asymmetric least squares smoothing correction
Normalization	Standard normal variate (SNV)
	Peak height
	Peak area
Other	Multiplicative signal correction (MSC)
	Derivatives (1 st Der or 2 nd Der)

An overview explains the analysis cycle in SIMCA[®] environment and will be further described in the following (Figure 4.6). A practical demonstration will be introduced in Chapter 5.1.

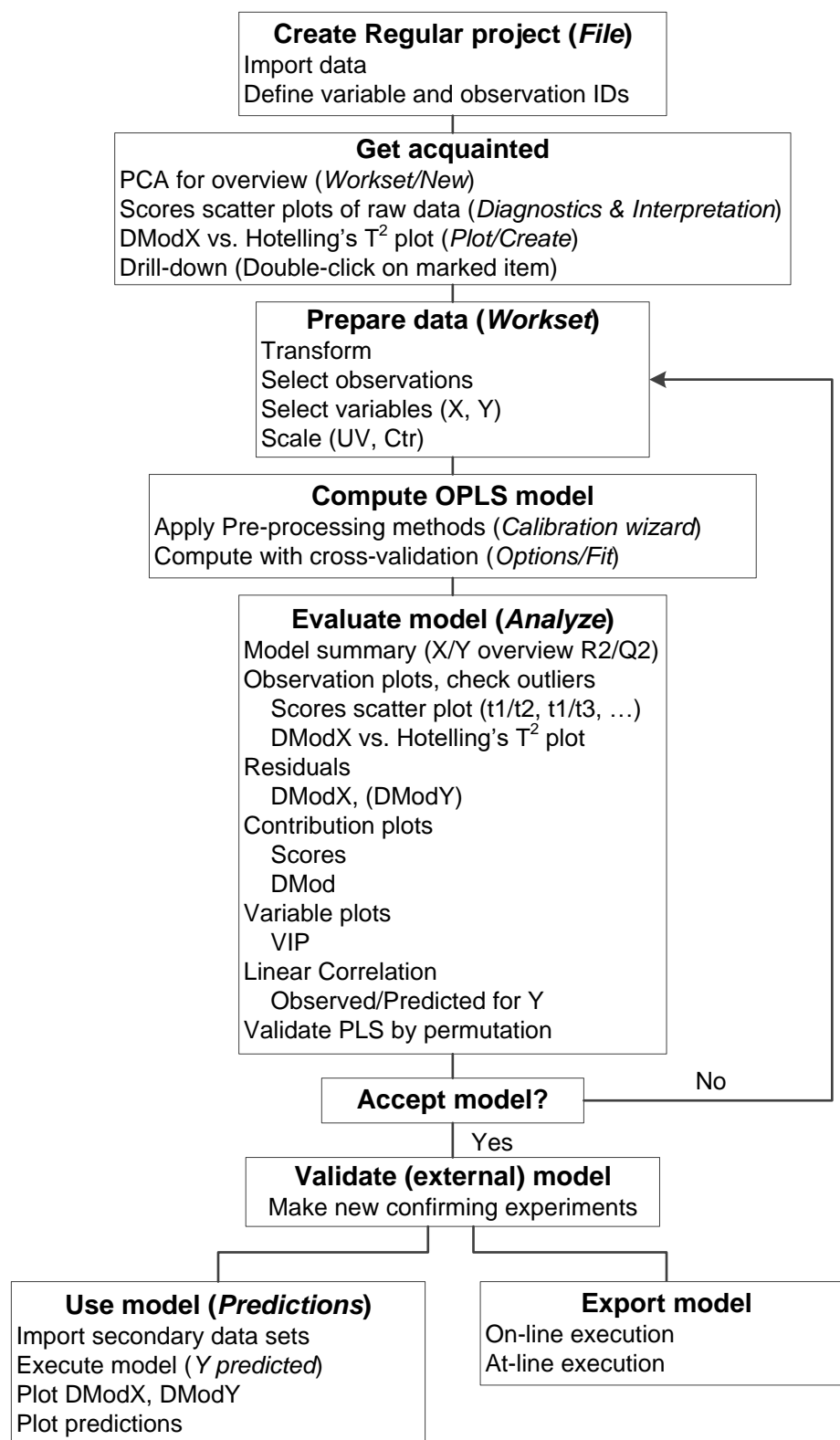


Figure 4.6: Flow-chart of multivariate data analytics in SIMCA® environment. Italic words describe menus/actions in SIMCA® environment (modified after (Eriksson et al., 2006b).

After data import, the data is selected in the *Dataset menu*. Here, it is necessary to define *Primary* and eventual *Secondary observations* and *variables*, respectively. These *Identifiers* have to be identical for both X- and Y-variables to induce sample affiliations. When desired, multiple

pre-processing steps can be executed under *Data/Data Preprocessing* (cf. Table 4.7). The pre-processed data is then stored in a new data set. Then, a new model was developed from the data set in the *Workset menu*. Desired variables and observations of the data set can be either excluded or included for the new model. Also, the scaling (UV or mean centring) and model type (PCA, PLS, or OPLS) can be chosen. Spectral data was mean centred while responses were auto-scaled. After choosing the desired settings, a new work set is created.

Then, the number r of components is chosen. Here, SIMCA[®] offers the opportunity to either process an *Autofit*, to use the *Two first* components, or to manually adjust the desired number of components (Figure 4.7). The number r can be changed again afterwards and becomes an inherent factor for the iterative modelling in OPLS.



Figure 4.7: Graphical user interface of SIMCA[®] in the version 17.0.1.

An outlier can be further interpreted by double clicking on it. This action is denoted *drill-down* in SIMCA[®] environment and opens up the corresponding score contribution plot. When an outlier was detected, it was excluded from the work set. This can be conveniently executed by clicking on the observation to be ignored and the marked item can be excluded in a pop-up menu. When doing so, a new model is automatically produced without the excluded observation. Then, the number r of components had to be chosen again.

After data preparation, pre-processing of the spectral data is a common practice. In SIMCA[®] environment, the *Calibration wizard* eases the procedure of applying a number of different pre-processing methods onto the same work set. In the first tab, the data sets have to be chosen.

Then, the variables and observations could be chosen, when exclusion of wavenumbers or samples was desired. Here, the choice of variables and observations can be copied of already existing models, avoiding to have to exclude each previously detected outlier of the PCA by hand again. Also, the number of observations n can be randomly split into CS (80 %) and VS (20 %), resulting in $n \cdot 0.8$ observations for CS and $n \cdot 0.2$ observations for VS. Finally, in the next tab *Filter & compare*, a set of 14 different pre-processing steps can be applied and chained. This tab automatically compares and highlights the best RMSEcv values of the created OPLS models.

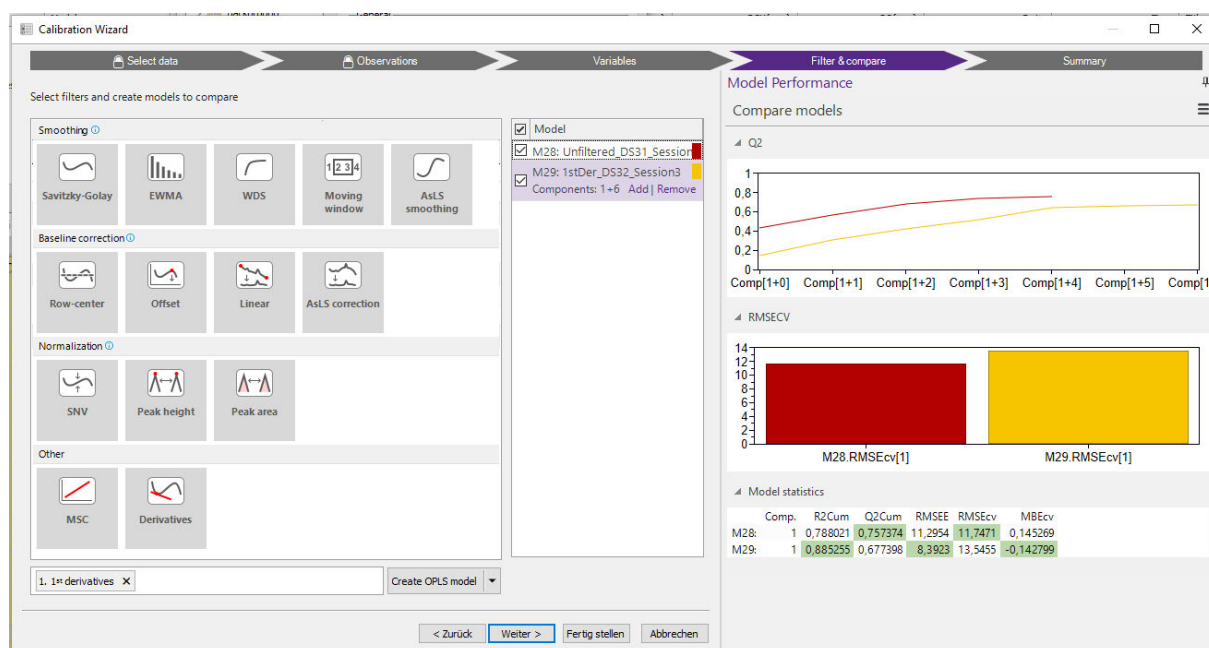


Figure 4.8: Graphical user interface of *Calibration wizard* in SIMCA 17.0.1®.

4.7.5 Preliminary Studies on Data Pre-Processing Methods

To limit the number of applicable pre-processing methods on all four analytes, a set of methods was examined in advance. The 12 introduced pre-processing tools were chained such that in total 48 different pre-processing methods were applied (Table 4.8). Ctr was not explicitly mentioned in the table as scaling was always applied to spectral data in the last step. Other scaling methods were not applied for spectral data due to the loss of information (cf. Chapter 3.4.4.4). The off-line Raman spectra were used for prediction of glycerol concentration c_{SIL} in the cell suspension. Cultivations XXPC1722 and XXPC2622 were used as CS while XXPC0922 was used as PS.

Table 4.8: Combining pre-processing tools into pre-processing methods with one to four pre-processing steps. Different pre-processing tools were chained, expressed by a hyphen between the corresponding abbreviations. Not explicitly mentioned here, the last step always comprises mean centring.

No.	Method	No.	Method	No.	Method
1	Unfiltered	17	2 nd Der-SG9	33	LinC-SNV-SG15
2	1 st Der	18	2 nd Der-SG15	34	LinC-SNV-WDS
3	1 st Der-SNV	19	2 nd Der-WDS	35	LinC-MS-C-SG9
4	1 st Der-MS-C	20	2 nd Der-SNV-SG9	36	LinC-MS-C-SG15
5	1 st Der-SG9	21	2 nd Der-SNV-SG15	37	LinC-MS-C-WDS
6	1 st Der-SG15	22	2 nd Der-SNV-WDS	38	SNV
7	1 st -WDS	23	2 nd Der-MS-C-SG9	39	SNV-SG9
8	1 st Der-SNV-SG9	24	2 nd Der-MS-C-SG15	40	SNV-SG15
9	1 st Der-SNV-SG15	25	2 nd Der-MS-C-WDS	41	SNV-WDS
10	1 st Der-SNV-WDS	26	LinC	42	MS-C
11	1 st Der-MS-C-SG9	27	LinC-SNV	43	MS-C-SG9
12	1 st Der-MS-C-SG15	28	LinC-MS-C	44	MS-C-SG15
13	1 st Der-MS-C-WDS	29	LinC-SG9	45	MS-C-WDS
14	2 nd Der	30	LinCSG15	46	SG9
15	2 nd Der-SNV	31	LinC-WDS	47	SG15
16	2 nd Der-MS-C	32	LinC-SNV-SG9	48	WDS

1stDer: 1st order derivative, 2ndDer: 2nd order derivative; LinC: linear correction; SNV: standard normal variate; MS-C: multiplicative scatter correction; SG: Savitzky-Golay; WDS: wavelet denoise spectral; Ctr: mean centred.

Furthermore, three spectral ranges were investigated in the preliminary studies (Table 4.9). Wavenumber range A valued 300–1840 cm⁻¹, range B 1841–2973 cm⁻¹, and range C combined both ranges, 300–2973 cm⁻¹.

Table 4.9: Wavenumber ranges investigated in preliminary studies.

Designation	Wavenumbers / cm ⁻¹
A	450–1840
B	1841–2973
C	450–2973

In order to guarantee a systematic and reasonable procedure for the selection of the most suitable pre-processing methods, a weighted sum model, also known as weighted linear combination or simple additive weighting, was applied (Churchman & Ackoff, 1954; Fishburn, 1967). Furthermore, a rapid screening of all 48 methods per wavenumber range was enabled.

The decision matrix is based on following three criteria:

- (1) Value of Q2,
- (2) Difference between R2Y and Q2, and
- (3) Number r of OPLS component.

In general, all three criteria are mainly based on experience in multivariate calibration (Chin & Marcoulides, 1998; Eriksson et al., 2006b; Peng & Lai, 2012). A further differentiation was made within the criteria in order to assign each criterium to a corresponding score (Table 4.10).

Table 4.10: Criteria and assigned scores for weighted sum model.

No.	Criterium	Weight / %	Sub-criterium	Assigned score / -
(1)	Value of Q2	40	$Q2 \geq 0.5$	1
			$Q2 < 0.5$	6
(2)	Difference between R2Y and Q2	30	$ R2Y - Q2 \leq 0.2$	1
			$ R2Y - Q2 < 0.3$	2
			$ R2Y - Q2 < 0.5$	3
			$ R2Y - Q2 \geq 0.3$	4
(3)	Number r of OPLS components	30	$r \leq 4$	1
			$r \leq 5$	2
			$r > 5$	3

The highest weight with 40 % was assigned to criterium (1), the value of Q2. With this, the predictive power for the model is evaluated and indicates how well VS and PS perform. For criterium (2), the lower the difference between R2Y and Q2, the more robust the model. The last criterium addresses the likeliness of overfitting the model. With high number r of components, the model is more likely to be overfitted.

For evaluation, the best models yield the lowest sum while the worst models comprise a high sum. All models with a sum greater than 1.6 were disposed. E.g., model number 1A addresses the unfiltered wavenumber range A. This model comprises 1 + 4 components yielding $R2Y = 0.730$ and $Q2 = 0.691$. The exemplary calculated weighted sum of 1.6 would consist of the following weights:

$$(1) \quad Q2 \geq 0.5 \quad \rightarrow 1 \cdot 0.4 = 0.4$$

$$(2) \quad |R2Y - Q2| < 0.5 \quad \rightarrow 3 \cdot 0.3 = 0.9$$

$$(3) \quad r \leq 4 \quad \rightarrow 1 \cdot 0.3 = 0.3$$

5 Results and Discussion

The following chapter deals with the results of the experiments and data analysis introduced in Chapter 4. An exemplification of OPLS modelling is introduced first to facilitate the understanding of the upcoming results. To evaluate the applicability of in-line and off-line Raman spectroscopy, results of the preliminary studies on pre-processing methods are elaborated, followed by the results in MVDA for the compounds glycerol, methanol, cell density, and total protein concentration.

Note that due to software issues, plots of SIMCA[®] and process parameters are displayed with decimal comma for decimal separation instead of the decimal point.

5.1 Exemplary Development of an OPLS Model for Glycerol

The presence or absence of outliers, the selected pre-processing method, the used spectral range, and the number of OPLS components are important factors influencing the predictive power of a model. The challenge in modelling lies in the fact that all aspects mentioned above are dependent on each other. E.g., the choice of spectral range can influence whether an observation is identified as an outlier. Therefore, the development or optimization of an OPLS model with the use of spectral data is a sophisticated and most of all iterative process (Buckley & Ryder, 2017). Consequently, it is hardly possible to follow the same pattern for different analytes as it depends on the current problem. However, with knowledge about the prevailing data set and its underlying bioprocess, assumptions can be made to ease the process.

In the following, a general approach for the prediction of glycerol concentration c_{SIM} in the supernatant by off-line Raman spectroscopy is introduced in order to have an insight on the complex process of OPLS modelling. The demonstrated approach shows a combination of the general MVDA pipeline (cf. Figure 3.7) and the flow-chart in SIMCA[®] environment (cf. Figure 4.6).

For multivariate calibration, diverse samples representing the variation of interest are important. Thus, more than one cultivation was used during this work. As reference values, HPLC measured glycerol concentration $c_{SIMHPLC}$ was applied. Raman spectra of cultivations XXPC0922 and XXPC2622 were used as CS. The third cultivation XXPC1722 should be used for external validation as PS. With this, $n = 149$ observations were available (Figure 5.1). Wavenumbers from $75\text{--}3215\text{ cm}^{-1}$ with a resolution of 1 cm^{-1} were displayed, resulting in $m = 3141$ variables.

However, also samples of the production phase were inside these 149 observations where glycerol concentration valued 0 g L^{-1} . In order to obtain an uniformly distributed data set, only observations from batch and fed-batch phases were used. Hereby, the number of observations was reduced to 86.

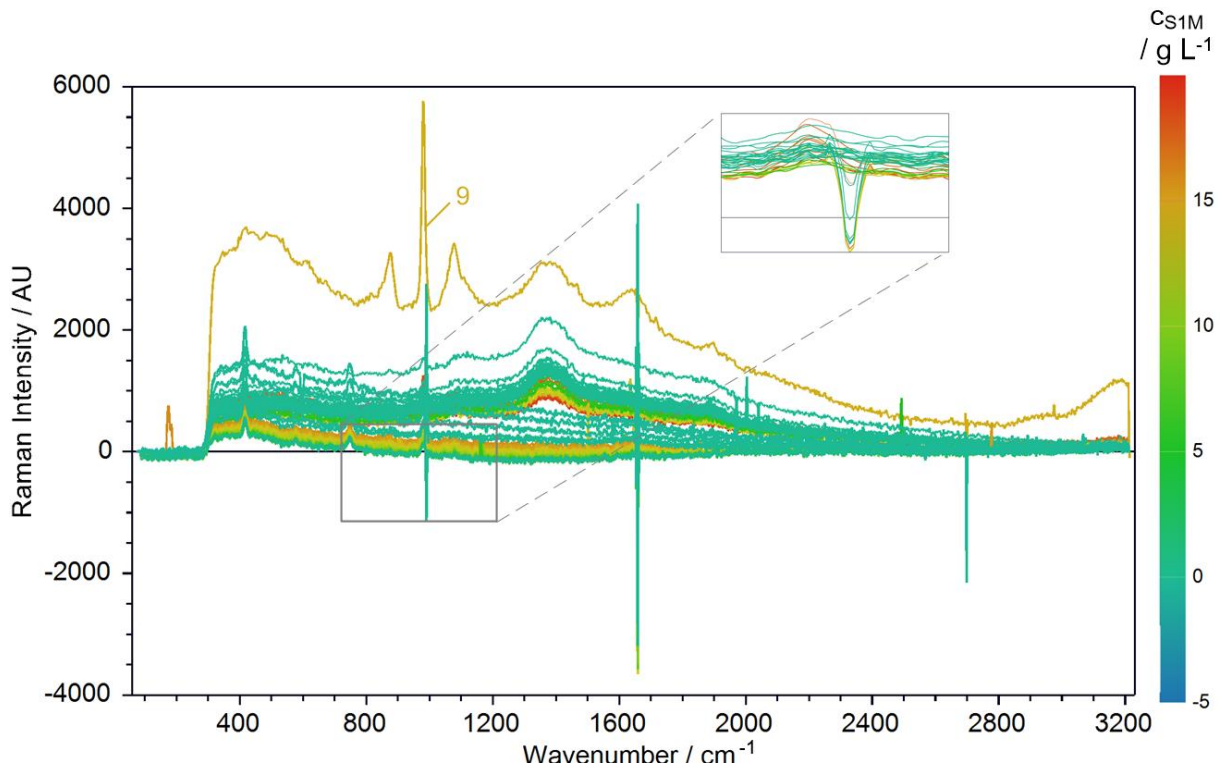


Figure 5.1: Exemplary off-line Raman spectra of supernatant in cultivations XXPC0922 and XXPC2622 with highlighted observation 9 (yellow). Spectra were coloured according to their glycerol concentration c_{S1M} . Zoom-in: samples of XXPC0922.

The spectra were coloured according to their glycerol concentration c_{S1M} . A correlation between Raman intensity of the spectra and glycerol concentration can be specifically observed in the highlighted area. Also, observation 9 deviates from the other spectra.

Inspecting the spectra coloured according to their batch number, there are two differently spectral courses visible (Figure 5.2). Between both cultivations, different courses during nearly the whole spectral range are visible. Although the integration time for both cultivations is 70 s, XXPC2622 shows spectra with a continuously decreasing trend and contains more clear bands. This will be further examined in the following sections.

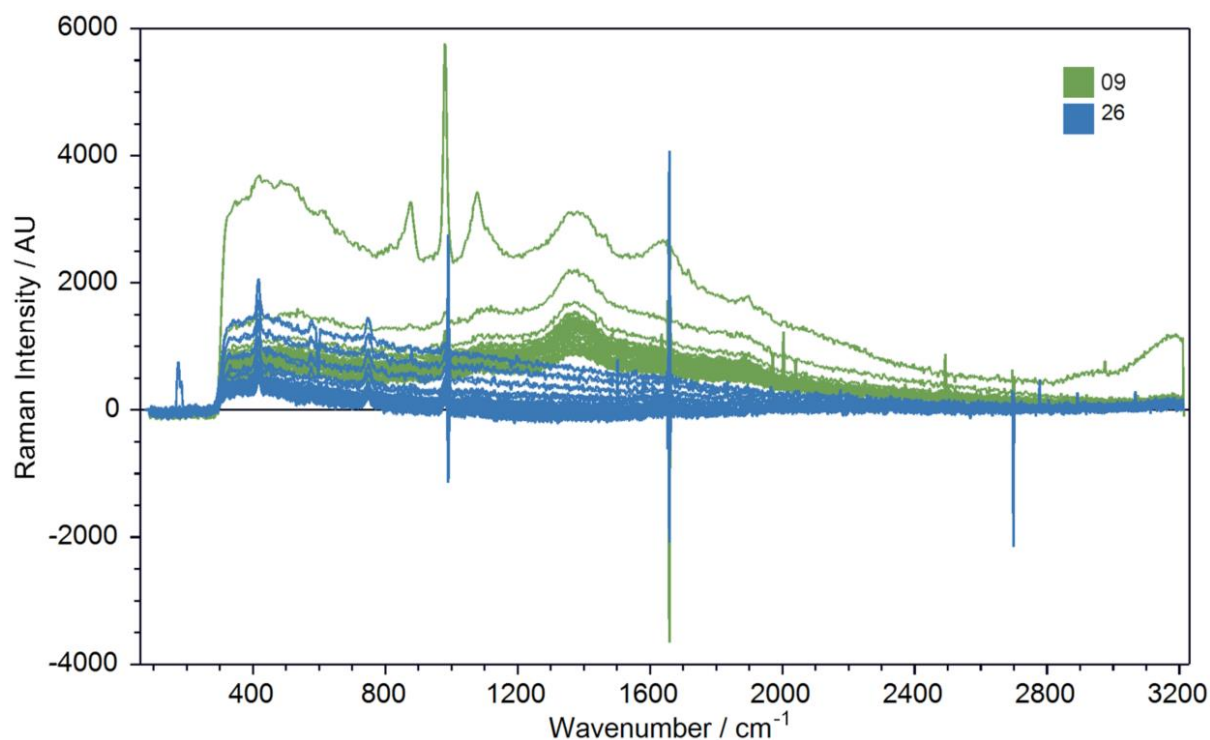


Figure 5.2: Exemplary off-line Raman spectra for cell suspension of cultivation XXPC0922 and XXPC2622. Spectra were coloured according to batch number. Green: XXPC0922; blue: XXPC2622.

5.1.1 Principal Component Analysis for Outlier Detection

As the first step, PCA modelling was done to visualise any outliers and overcome false weighing of the model. As mentioned in Chapter 3.4.4.4, centring was used for spectral data in order to preserve the characteristics of bands while Y-variables (responses) were auto-scaled. For comparison, both options in SIMCA[®] environment, *Autofit* and *Two first*, is introduced (Figure 5.3). When autofitting, the addition of seven more PCs resulted in an increase of the explained variance, denoted cumulative R2X ($R2X_{cum}$) from 0.963 to 0.997. However, PCA models were chosen such that at least 95 % of the variance in the spectral data were described in order to preserve model complexity. Therefore, PCA modelling was continued with $r = 2$ PCs.

After calculating the PCs, the Hotelling's T^2 test was executed. The result is illustrated in form of a scores scatter plot with a total of 86 observations where two different PCs are plotted against each other (Figure 5.4). Observations close to each other have similar properties, whereas those far from each other are dissimilar with respect to spectral profiles.

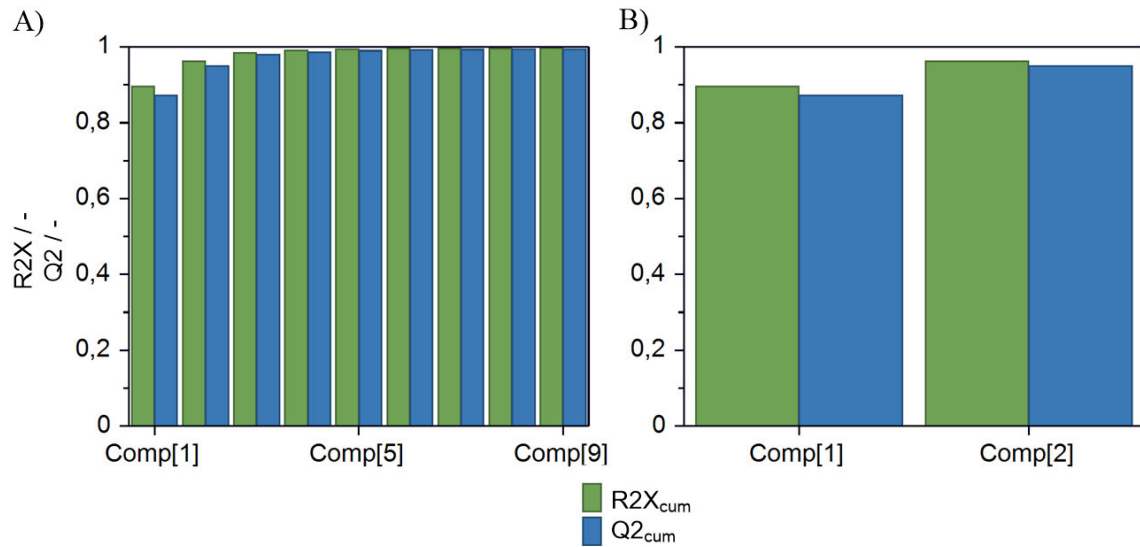


Figure 5.3: Exemplary summary of fit for PCA model with *Autofit* or *Two first*. A) *Autofit* yields nine principal components in total, with $R2X_{cum} = 0.997$ (green bar) and $Q2_{cum} = 0.995$ (blue bar); B) Fitting method of *Two first* results in two principal components with a $R2X_{cum} = 0.963$ and $Q2_{cum} = 0.950$.

In the scores scatter plot, PC2 t_2 is plotted against PC1 t_1 (Figure 5.4A). Observation 9, representing $t_{process} = 4$ h of XXPC0922, depicts with a probability of 95 % an outlier as the level of significance $\alpha = 5$ %. This observation pulls the whole model towards itself. Figure 5.4B shows the DModX of the last PC plotted against the Hotelling's T^2 range of all PCs. Although there are more observations outside $D_{crit0.05} = 1.17$, not all observations outside this limit were considered outliers. For DModX, values twice as large as $D_{crit0.05}$ are considered moderate outliers. With this, only observation 9 was an outlier whereas observations 114 and 115 were still in a reasonable range for this stage of PCA modelling.

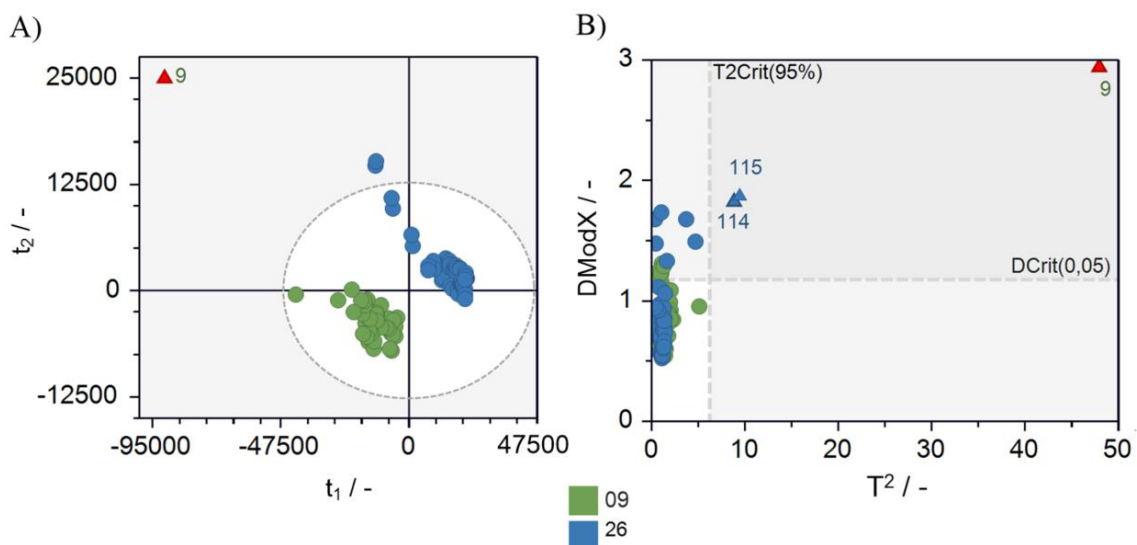


Figure 5.4: Exemplary scores scatter plots with Hotelling's T^2 test for PCA model ($\alpha = 0.05$). A) Scores scatter plot of PCA model with $r = 2$, $R2X = 0.963$. Observation 9 (red triangle) possesses a high Hotelling's T^2 value far above the confidence interval of 95 %. B) Plot of DModX versus Hotelling's T^2 value with highlighted observation 9 (red triangle) above $T^2_{crit0.05}$ and $D_{crit0.05}$.

A *drill-down* of observation 9 was executed (Figure 5.5). The contribution plot shows where a point in a Hotelling's T^2 plot deviates from the average or from another point in X-space. The plot shows the weighted difference between the data of the point (centred as the work set) and the average of the model. The horizontal scale corresponds to the wavenumber, the vertical scale to the scaling of X. The dominating variables deviate by greater than three standard deviations (SDs) from the reference point. The sign of the line indicates in which direction the variable deviates.

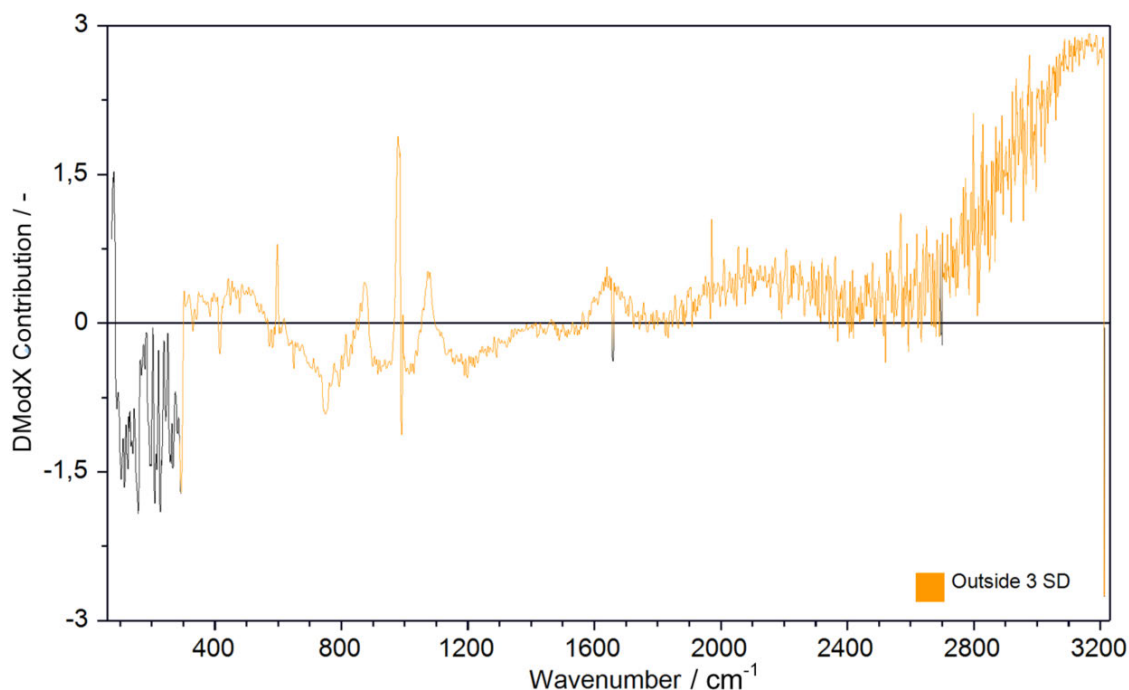


Figure 5.5: Exemplary score contribution plot of observation 9 in a PCA model with $r = 2$ PCs. Orange: deviating from average by three standard deviations (SD).

In this case, nearly the whole spectrum was contributing to the deviation from the average. In fact, the top yellow spectrum represents observation 9 (cf. Figure 5.1). Due to this, observation 9 was considered an outlier and was consequently excluded from the work set. This process of outlier detection was repeated until no more moderate observation remained outside the confidence interval. With a total number of initially 86 observations, there is statistically a value of $86 \cdot 0.05 = 4.30$ observations expected to be outside the Hotelling's T^2 tolerance ellipse.

This was achieved after another six exclusion steps (Figure 5.6). In this case, 77 observations were remaining and used for OPLS modelling. A clear differentiation between both cultivations is visible in the scatter plot. Cultivation XXPC0922 resides on the negative part of t_1 while XXPC2622 does the opposite. This property can be observed in the original spectrum (cf. Figure 5.2). By outlier detection, the goodness of fit R^2X was increased from initially 0.963 to

0.980 while the goodness of prediction Q2 increased from 0.950 to 0.980. Also, exclusion of observation 9 prevented pulling the model towards one observation.

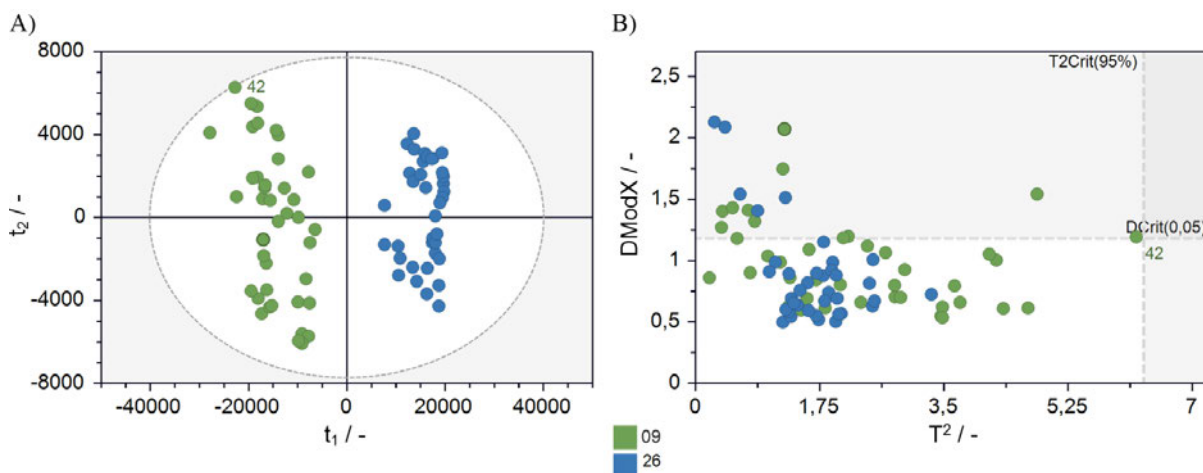


Figure 5.6: Final scatter plots after exclusion of moderate or high outliers. A) Scores scatter plot of principal component analysis with $r = 2$, $R^2X = 0.980$. Observation 42 is highlighted as it possesses a fairly high Hotelling's T^2 value close to the confidence interval of 95 %. B) Plot of $DModX$ versus Hotelling's T^2 with labelled observation 42, resulting not to be an outlier.

5.1.2 Multivariate Calibration with OPLS

Pre-processing is a crucial step in multivariate modelling as the performance of the model stands or falls with a clean dataset (Engel et al., 2013). In this work, several pre-processing methods were investigated in order to find the most prominent candidates for this work. However, for exemplification, it will be continued with non-pre-processed data.

After construction of an OPLS model, the number r of components had to be set. To evaluate this, the root mean square error of prediction RMSEP through an external validation set, the PS, is ideally used. However, if there is no prediction set available, an alternative predictivity measure can be used through summarising the cross-validation residuals of the observations in the work set.

The root mean square error of cross-validation $RMSE_{cv}$ was calculated for the work set, indicating predictive power. For OPLS models with $1 + r$ components, the evolution of $RMSE_{cv}$ across the model components is displayed (Figure 5.7). OPLS representation comprises following components. There is at least one predictive, one orthogonal in X, and one orthogonal in Y component. The predictive component captures the variation found both in X and Y, while the orthogonal component captures the variation found in X or Y, respectively. When there is one response (one Y-vector), there will only be one predictive component.

The model with 1 + 4 components was further investigated. Here, RMSEcv valued 2.77 g L^{-1} . With increasing orthogonal components, the RMSEcv decreased. However, the decrease was negligible as the model would only become more complex.

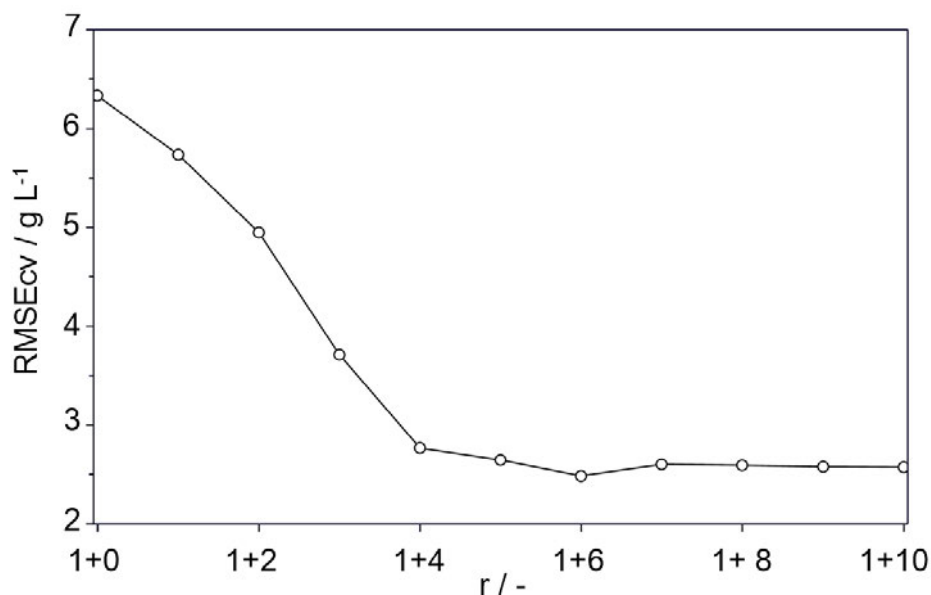


Figure 5.7: Exemplary prediction error RMSEcv of OPLS model dependent on number of components r .

In order to avoid overfitting, further investigation on the model with 1 + 4 components was necessary. This can be done by evaluating the permutation plot (Figure 5.8). In this work, 100 permutations were constantly applied onto each investigated model.

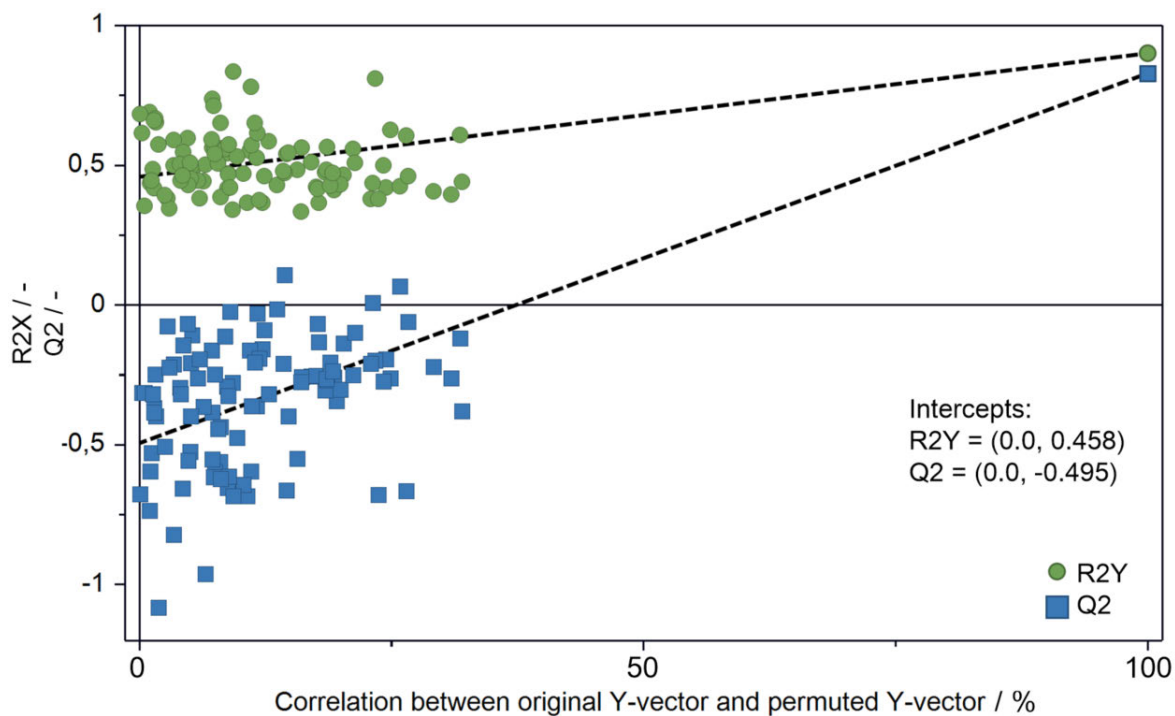


Figure 5.8: Exemplary permutation plot for OPLS model with $r = 1 + 4$ components after 100 permutations. Green: R2Y values, blue: Q2 value.

R²_Y and Q² of the original model were always higher than the permuted corresponding values. The intercept of the R²_Y regression line was slightly above the proposed limit of 0.4 and shows a model which could be overfitted or even not valid (Eriksson et al., 2006b). In contrast, the intercept for Q² was below the proposed limit of 0.05. In general, with increasing components, the R²_Y-intercept increases as well, indicating an overfit. Therefore, the number of components was reduced to 1 + 3 components. This process of permuting data and comparing permutation plots for each relevant component is repeated such that the intercepts were below the proposed limits. Simultaneously, the plot for predicted vs. reference was kept in high quality which will be addressed later.

For variable selection, the model's VIP value was examined, summarising the importance of variables both to explain X and to correlate Y. The beginning and the end of the spectra did not contribute to the model interpretability as $VIP < 0.5$ (Figure 5.9). However, the spectral range 300–1957 cm^{-1} has a $VIP > 1$, indicating important variables. In order to overcome overfitting, the spectral range was extended to 300–2200 cm^{-1} . With this, the variables could be reduced from 3141 to 1901 variables.

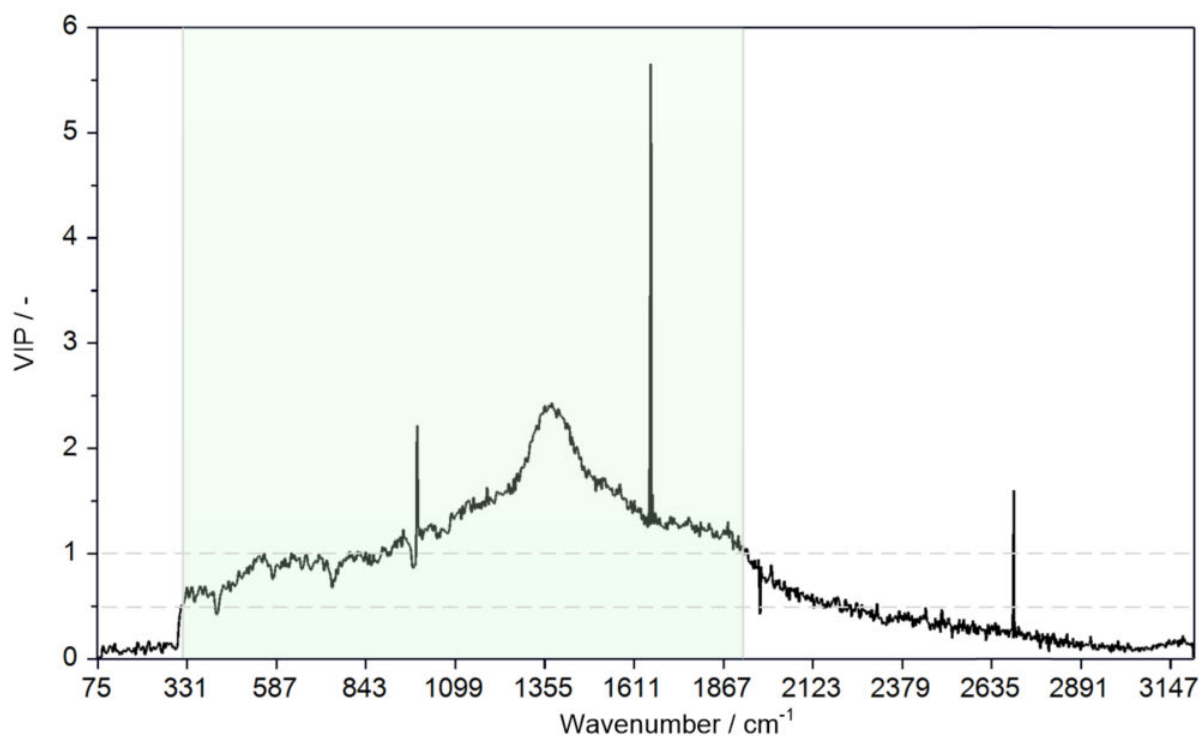


Figure 5.9: Exemplary variable importance in projection VIP of 1 + 3 OPLS model. VIP values greater than 1 imply important X-variables and values lower than 0.5 indicate unimportant X-variables.

5.1.3 Prediction and Validation

For validation, the OPLS model with decreased variables and excluded outliers was further investigated in terms of the predicted values. Three different methods of validation are demonstrated in the following, before giving explanation about the methods.

First, cross-validated Y-values was plotted against the reference $c_{S1MHPLC}$ in the predicted vs. reference plot (Figure 5.10). This option of cross-validation is used when there is no VS neither PS available.

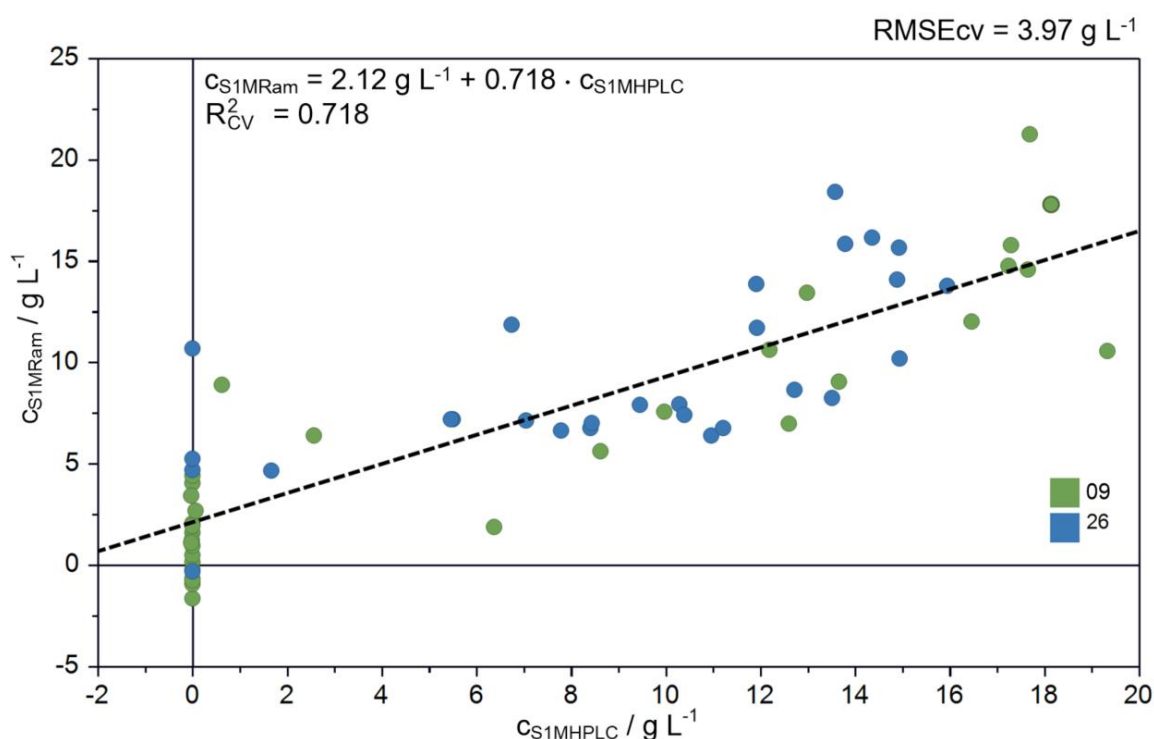


Figure 5.10: Exemplary regression line for cross-validation with cross-validated values plotted against reference of glycerol concentration c_{SIM} . Green: XXPC0922; blue: XXPC2622.

The second method for validation is applied using the VS which was created by separating 20 % of all observations into this sub-group (Figure 5.11).

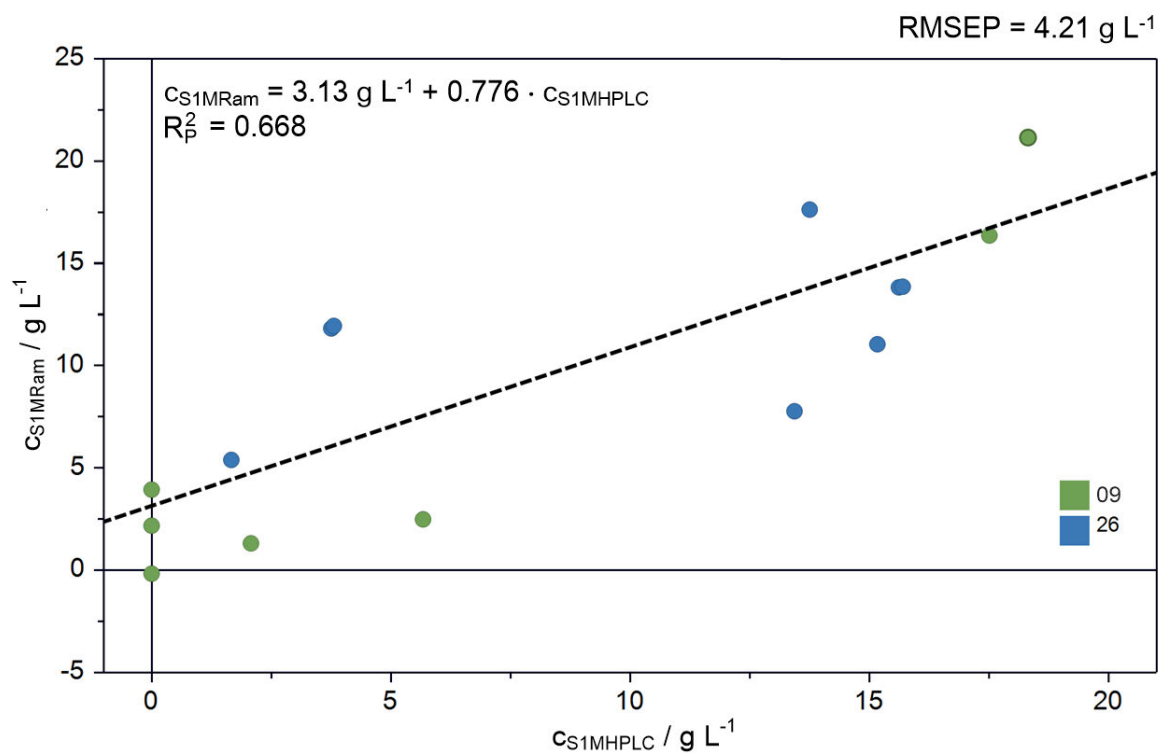


Figure 5.11: Exemplary regression line for internal validation with predicted validation set plotted against reference of glycerol concentration c_{SIM} . Green: XXPC0922; blue: XXPC2622.

The third method is external validation by use of PS XXPC1722 (Figure 5.12).

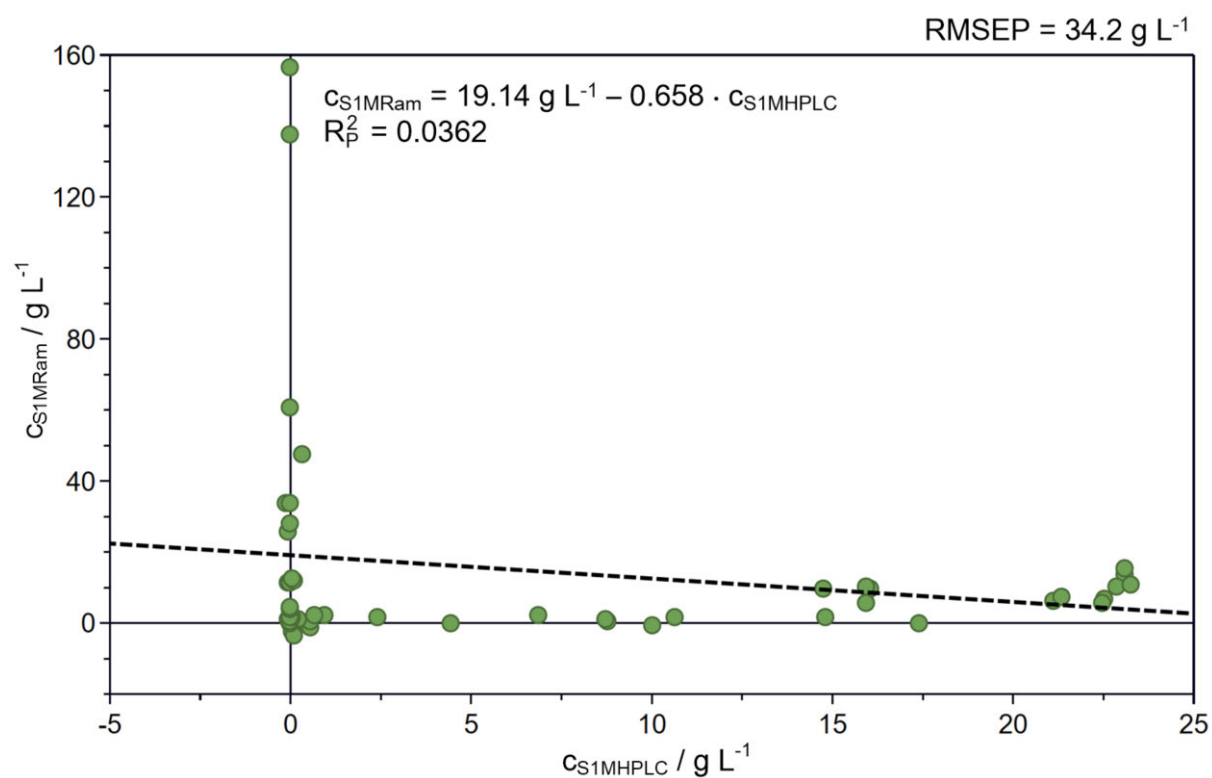


Figure 5.12: Exemplary regression line for external validation set with prediction set plotted against reference of glycerol concentration c_{SIM} for XXPC1722.

In the ideal case, the points result in a straight line through the origin with the slope of one. When the scores imply a non-linear course, following steps can be altered and evaluated: change of pre-processing method, need for further OPLS component(s), applying data transformation (e.g. log-transformation for exponential course), or change of applied regression modelling method (Eriksson et al., 2006b).

The lowest goodness of prediction was obtained by using the external validation with $R^2Y = 0.0362$. There is no correlation between the reference values and prediction. This property becomes apparent when plotting both reference and prediction against process time t_{process} (Figure 5.13).

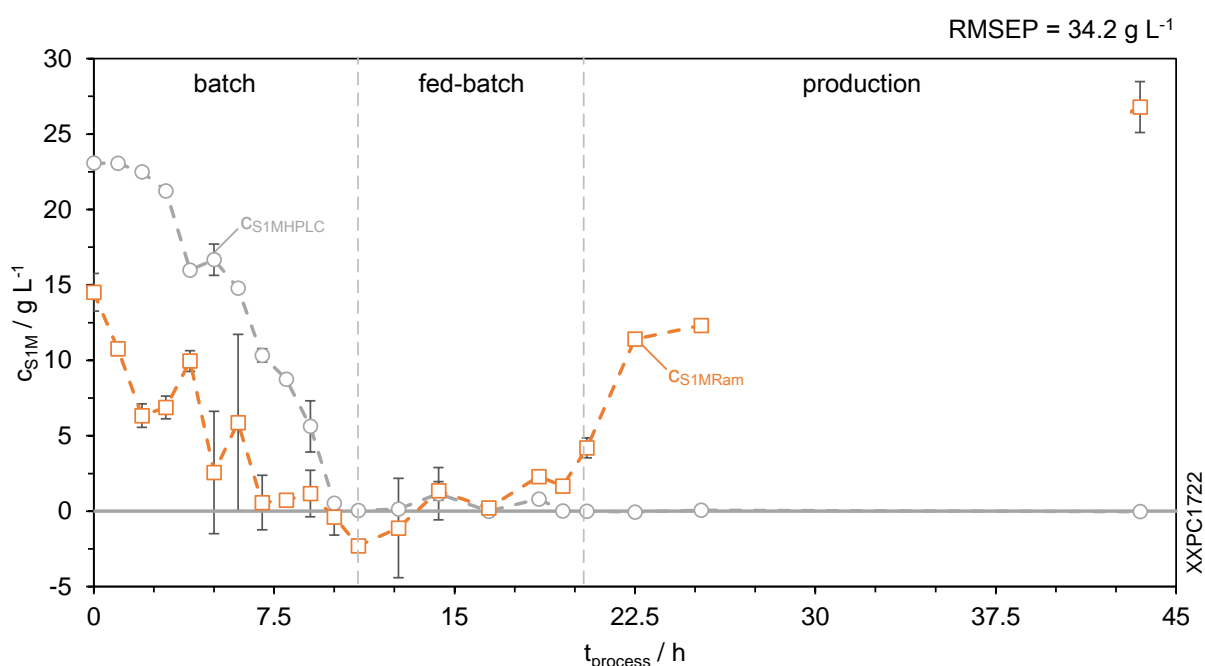


Figure 5.13: Exemplary prediction of glycerol concentration c_{S1} in supernatant with external validation set XXPC1722. Grey: HPLC reference measurement; orange: off-line supernatant Raman measurement.

Typically, the cross-validated predictions have a higher goodness of fit R^2Y than the predicted values of the VS. Working with time-dependent process data typically comes with auto-correlated process data, leading to the fact that the predictive power may not be reliably estimated. This is due to observations (time points), adjacent to those eliminated in a CV-round, carrying information similar to the eliminated data points. Therefore, CV with auto-correlated process data may lead to overrated Q^2 values (Eriksson et al., 2006b). This was addressed in this work by evaluating the permutation plot (cf. Chapter 5.1.2). Alternatively, raw data can be sorted according to their Y-value in order to break up the auto-correlation among neighbouring samples. For prediction of VS, $R^2Y = 0.668$ and $RMSEP_{VS} = 4.21 \text{ g L}^{-1}$. These values were lower than the cross-validated values of $R^2Y = 0.718$ and $RMSE_{cv} = 3.97 \text{ g L}^{-1}$. Since cell

suspension was investigated and no pre-processing method was applied, interfering substances in the cell suspension represent a major factor for the moderate predictive power.

Because both cultivations of the CS were different in spectral data (cf. Figure 5.6A), upcoming calibrations are done with all three available cultivations XXPC0922, XXPC1722, and XXPC2622. Thus, there is no external validation available but only a PS comprising excluded samples of the cultivation phases.

5.2 Preliminary Studies on Pre-Processing Methods

This section focuses on the evaluation of applied pre-processing methods in order to limit the number of pre-processing steps in the upcoming chapters. Also, the range of wavenumber was investigated. 144 OPLS model (3 wavenumber ranges · 48 pre-processing methods) were constructed. Without the preliminary studies, the number of total OPLS models would exceed the scope of this work (144 OPLS models · 4 analytes · 3 measuring types = 1,728 models in total). Initially, 48 different pre-processing methods were examined which comprised 12 pre-processing tools (cf. Table 4.6). Out of these 12 tools a set of 48 different pre-processing methods was constructed (cf. Table 4.8).

For the evaluation of all models, a catalogue of criteria based on experimental values was developed which were based on experimental values. It is important to mention that the weighted sum model applied here does not allow a proof of concept regarding its evidence on model performance. However, these criteria were developed to give an overview on the perfunctory performance of a model, allowing a time-effective screening. In parallel, the RMSEcv and RMSEP were evaluated.

The results showed that there are a number of pre-processing methods which fall out of consideration for the upcoming analysis (Table 5.1). It was observed that models with a weighted sum ≤ 1.3 yielded a low RMSEcv. Models with a sum between $1.3 < \text{sum} \leq 1.6$ needed further inspection in terms of permutation plot or predicted Y-value. If the permutation plot showed intercepts below the limit, the model was considered well. All models with a sum > 1.9 yielded in consequence a high RMSEcv. These models were considered unsuitable for prediction. The methods with 1stDer or 2ndDer in combination with MSC failed in correlating Y-value with the spectra at all (numbers 11–13 and 23–25). Generally, MSC-based models retrieved high weighted sums. In literature, it is proposed that MSC should be applied to smaller and well-

selected parts of spectral region (Eriksson et al., 2006b; Stenlund, 2011) which was not the case in this work. Only the result for range C is shown as exemplification.

Table 5.1: Results of preliminary studies for wavenumber range C that were applied onto Raman spectra.

No.	r	R2Y	Q2	Weighted sum	No.	r	R2Y	Q2	Weighted sum
1A	1+4+0	0.750	0.664	1.0	25A	0+0+0	–	–	3.0*
2A	1+4+0	0.767	0.594	1.0	26A	1+5+0	0.780	0.659	1.3
3A	1+6+0	0.844	0.637	1.9	27A	1+6+0	0.788	0.702	1.6
4A	1+0+0	0.039	0.005	3.0	28A	1+5+0	0.688	0.610	1.3
5A	1+4+0	0.766	0.590	1.0	29A	1+4+0	0.719	0.630	1.0
6A	1+6+0	0.939	0.736	1.9	30A	1+5+0	0.825	0.678	1.3
7A	1+5+0	0.674	0.582	1.3	31A	1+4+0	0.653	0.584	1.0
8A	1+5+0	0.786	0.630	1.3	32A	1+6+0	0.797	0.717	1.6
9A	1+5+0	0.783	0.628	1.3	33A	1+6+0	0.836	0.719	1.6
10A	1+5+0	0.718	0.611	1.3	34A	1+6+0	0.661	0.579	1.6
11A	0+0+0	–	–	3.0*	35A	1+5+0	0.690	0.615	1.3
12A	0+0+0	–	–	3.0*	36A	1+5+0	0.680	0.621	1.3
13A	0+0+0	–	–	3.0*	37A	1+5+0	0.625	0.563	1.3
14A	1+5+0	0.852	0.545	1.9	38A	1+4+0	0.740	0.241	3.6
15A	1+6+0	0.782	0.531	1.9	39A	1+6+0	0.825	0.777	1.6
16A	0+0+0	–	–	3.0	40A	1+6+0	0.833	0.752	1.6
17A	1+6+0	0.915	0.564	2.2	41A	1+5+0	0.676	0.632	1.3
18A	0+0+0	–	–	3.0*	42A	1+5+0	0.755	0.672	1.3
19A	0+0+0	–	–	3.0*	43A	1+5+0	0.748	0.687	1.3
20A	0+0+0	–	–	3.0*	44A	1+5+0	0.738	0.670	1.3
21A	0+0+0	–	–	3.0*	45A	1+5+0	0.660	0.605	1.3
22A	0+0+0	–	–	3.0*	46A	1+4+0	0.748	0.666	1.0
23A	0+0+0	–	–	3.0*	47A	1+4+0	0.762	0.680	1.0
24A	0+0+0	–	–	3.0*	48A	1+4+0	0.676	0.601	1.0

* Models could not be calculated for corresponding pre-processing method as there was no correlation at all.

Bold letters: lowest achievable weighted sum.

In the Raman spectra of XXPC1722, a baseline shift is evident and follows the time course change (Figure 5.14). With increasing time, the baseline shift increases as well.

Due to this, 1stDer proved to be most accessible for baseline shift correction. SG and SNV, in contrast, were the best tools for smoothing and scatter correction, respectively. Since two different Savitzky-Golay filters were examined, no significant differences between 9 points and 15 points per moving polynomial could be identified. Discrepancies in the order of 10⁻³ for R2Y and Q2 were present. However, SG9 predominantly yielded higher R2Y and Q2 values than SG15. Thus, analysis was continued with only SG9.

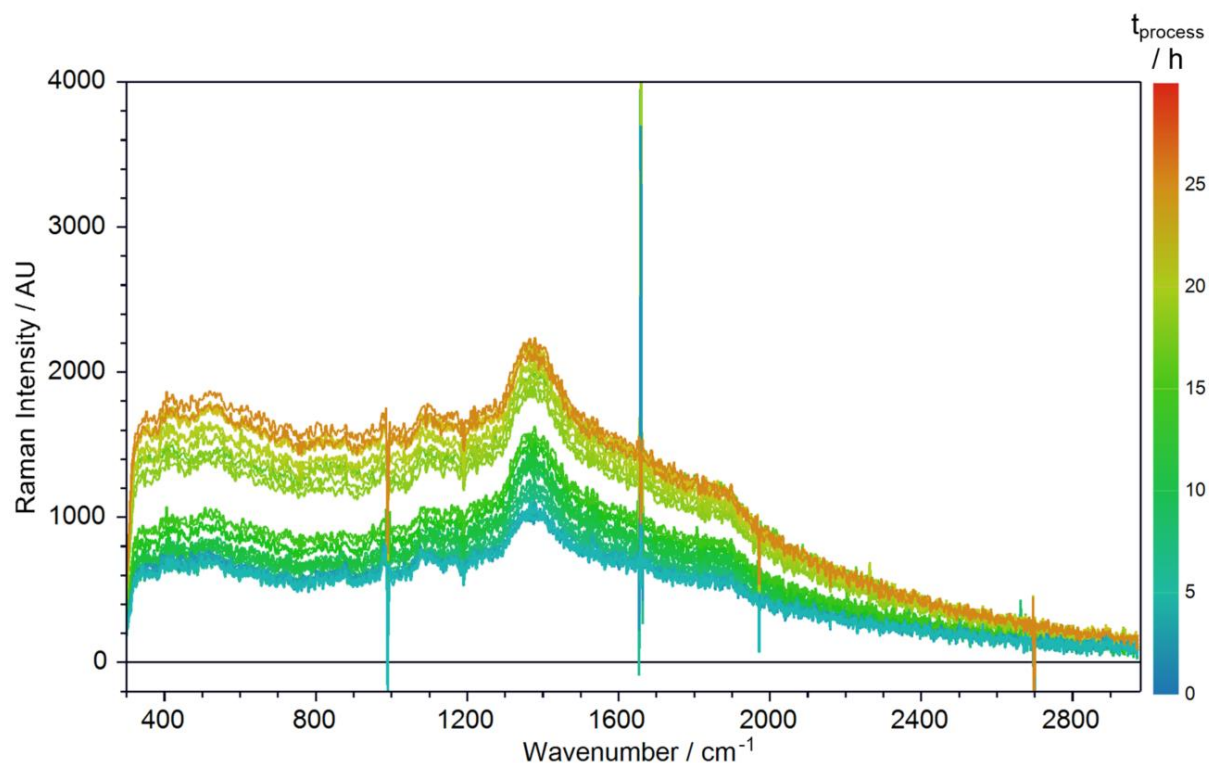


Figure 5.14: Off-line spectra of XXPC1722 suspension coloured according to process time t_{process} .

As a result, the following pre-processing methods could be limited from initially 48 methods to 12 methods and will be used in the upcoming sections (Table 5.2).

Table 5.2: Overview of pre-processing methods that were applied for upcoming multivariate calibration.

Methods were applied upon each analyte. As scaling was included in each method, it was not explicitly mentioned in the designation.

No.	Step 1: Baseline	Step 2: Scatter	Step 3: Noise	Step 4 Scaling	Designation
1	–	–	–	Ctr	Unfiltered
2	1 st Der	–	–	Ctr	1 st Der
3	1 st Der	SNV	–	Ctr	1 st Der-SNV
4	1 st Der	–	S-G, 9 pt, 2 nd ord	Ctr	1 st Der-SG
5	1 st Der	SNV	S-G, 9 pt, 2 nd ord	Ctr	SNV
6	LinC	–	–	Ctr	LinC
7	LinC	SNV	–	Ctr	LinC-SNV
8	LinC	–	S-G, 9 pt, 2 nd ord	Ctr	LinC-SG
9	LinC	SNV	S-G, 9 pt, 2 nd ord	Ctr	LinC-SNV-SG
10	–	SNV	–	Ctr	SNV
11	–	SNV	S-G, 9 pt, 2 nd ord	Ctr	SNV-SG
12	–	–	S-G, 9 pt, 2 nd ord	Ctr	SG

During investigation of wavenumber ranges A and B, both yielded higher RMSE_{cv} and RMSEP values than C. This was due to dependent variables which are based on overlapping signal bands

occurring in vibrational spectroscopy of organic solutions. These are highly collinear across similar functional groups of analytes within the mixture (Gosselin et al., 2010). Therefore, the wavenumber range will be extended to 450–3000 cm^{-1} for further calibration studies.

For easier reading, the models with its corresponding pre-processing methods are addressed by the initial letter N for number and the corresponding number, e.g. model N9 for the pre-processing method LinC-SNV-SG. Also, during preliminary studies, it was observed that spectra between XXPC1722 and XXPC2622 differed. Therefore, all three cultivations were used for calibration, unless otherwise stated.

5.3 Investigated Process Variables

The subject of this work was the cultivation of *P. pastoris* BSYBG11. For this, three data sets were available. One of the three cultivations used for multivariate calibration is depicted for demonstration (Figure 5.15).

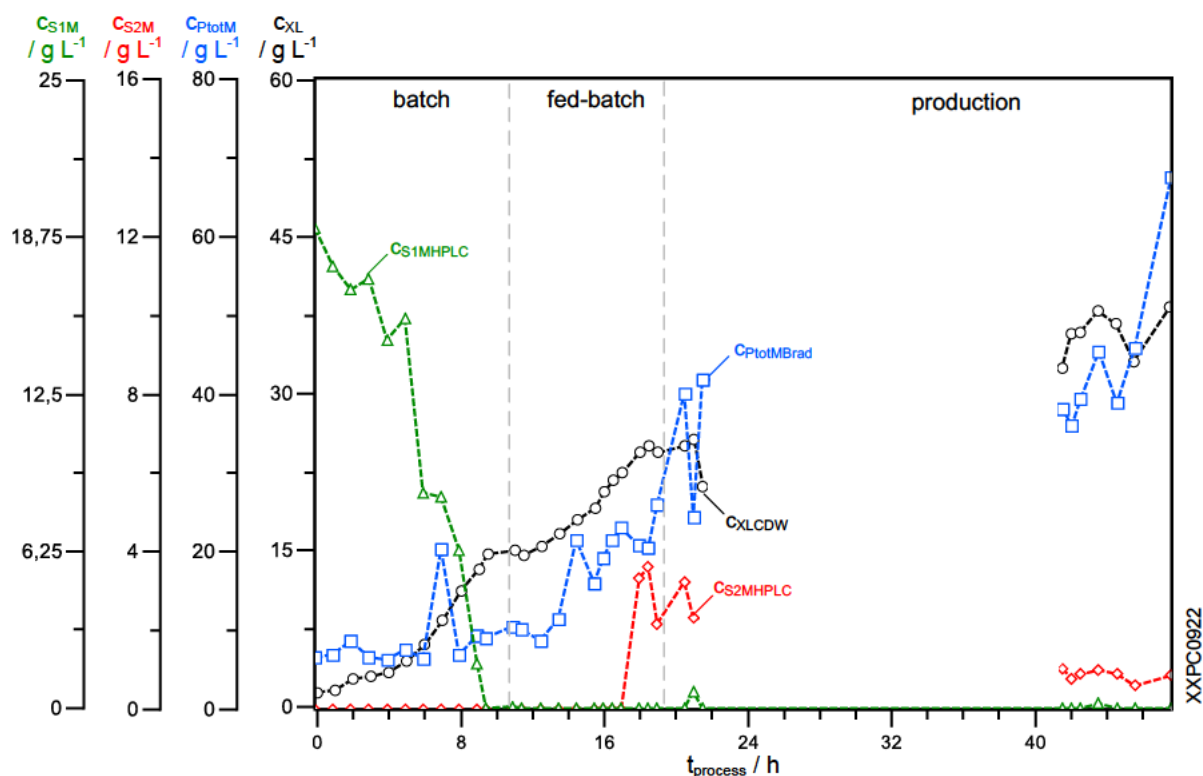


Figure 5.15: Course of off-line reference measurement for cultivation XXPC0922. Black: cell concentration by cell dry weight; blue: total protein concentration by Bradford assay; red: methanol concentration by HPLC; green: glycerol concentration by HPLC.

The analytes off-line cell concentration by CDW c_{XLCDW} (black), glycerol concentration by HPLC $c_{S1MHPLC}$ (green), methanol concentration by HPLC $c_{S1MHPLC}$ (red), and total protein

concentration by Bradford assay $c_{\text{PtotMBrad}}$ (blue) are illustrated. 80 % of the used data set served as CS, while the other 20 % were randomly grouped into the validation set (VS). The goal was to quantify these analytes by Raman spectroscopy.

5.4 Prediction of Glycerol Concentration

For multivariate calibration, samples from batch and fed-batch phase were used. Reference c_{SIMHPLC} values originated from HPLC. For PS, all excluded samples from the production phase were used.

Comparing off-line Raman spectra of the SN (Figure 5.16) with spectra of the SUS (cf. Figure 5.1), the area around 1000 cm^{-1} is similar like for XXPC0922.

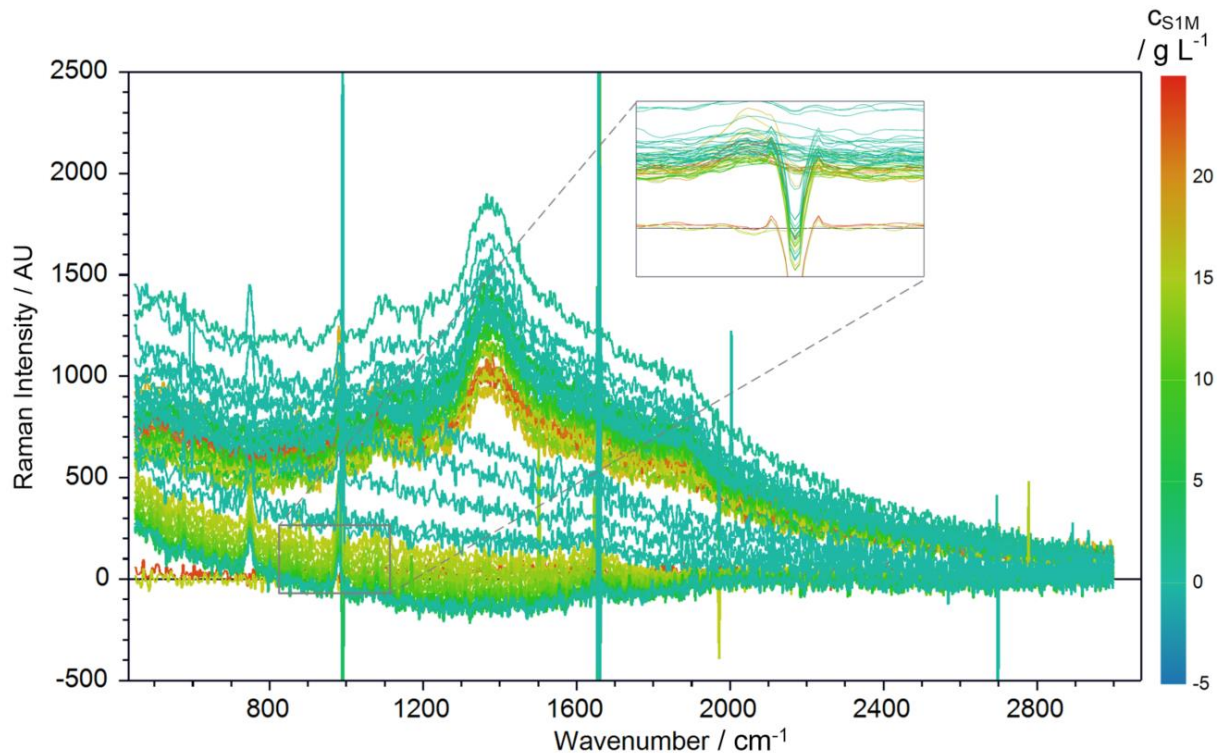


Figure 5.16: Off-line Raman spectra of calibration set for glycerol determination in cell suspension. Spectra are coloured according to the glycerol concentration c_{S1M} . Zoom-in: cultivations XXPC0922 and XXPC1722.

The most prominent bands for calibration are located at wavenumbers 992 cm^{-1} , 1378 cm^{-1} , and 1660 cm^{-1} . In fact, the off-line Raman VIP plots of N8 and N12 also show that these wavenumbers have a $\text{VIP} > 1$, indicating a large influence on the model for glycerol prediction (Figure 5.17).

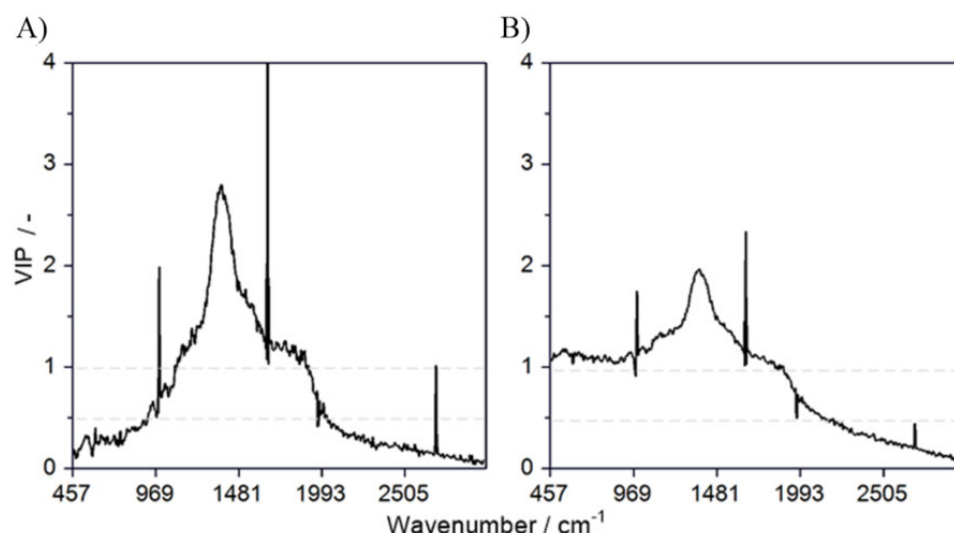


Figure 5.17: Variable importance in projection (VIP) of models N8 and N12. A) Model N8 with LinC-SNV-SG; B) Model N12 with SG. The two models yielded lowest prediction errors in off-line glycerol prediction for cell suspension.

For OPLS of glycerol in the SUS, maximal five components were required in order to explain approximately 85 % of the spectral information yielding RMSE_{cv} of 13.7–18.8 %, depending on data pre-processing (Table 5.3). For a better overview, only the best model of each measuring type is displayed while other model results can be found in the Appendix.

Table 5.3: Summary of predicted glycerol concentration in cell suspension (SUS) and supernatant (SN) with both off-line and in-line Raman spectroscopy.

		N12 (SG)	N1 (Unfiltered)	N2 (1 st Der)
Monitoring	/ –	off-line	off-line	in-line
Fluid	/ –	SUS	SN	SUS
n _{CS}	/ –	93	47	95
n _{VS}	/ –	24	12	24
r	/ –	1+4+0	1+5+0	1+3+0
R ² _Y	/ –	0.871	0.865	0.821
Q ²	/ –	0.837	0.745	0.740
RMSE _{cv}	/ g L ⁻¹	3.16	3.31	3.87
RMSE _{cv,rel}	/ %	13.7	14.3	18.8
MBE _{cv}	/ –	0.0323	0.0670	-0.0311
RMSEP _{VS}	/ g L ⁻¹	2.58	2.06	3.21
RMSEP _{VS,rel}	/ %	11.1	8.92	16.1
MBEP _{VS}	/ –	0.335	0.223	-0.916
RMSEP _{PS}	/ g L ⁻¹	1.58	4.63	4.11
RMSEP _{PS,rel}	/ %	6.83	20.1	38.1
MBEP _{PS}	/ –	-0.85	-2.96	1.43

In CV, the off-line suspension N12 performed best for glycerol determination with $RMSE_{cv} = 13.7\%$. However, all three models perform quite similar in CV. The difference of RMSEP becomes more apparent for VS and PS. $RMSEP_{VS}$ is for all three models lower than $RMSE_{cv}$, with unfiltered N1 showing the lowest $RMSEP_{VS} = 8.92\%$. When applying the PS, the SG-filtered N12 yields the lowest $RMSEP = 6.83\%$ with simultaneously lowest $MBEP_{PS}$. In PS, it is important to consider the influence of fluorescence, especially in the later process stages when eGFP was produced. This may explain why N1 and N2 have such high prediction errors which were not appropriately compensated by pre-processing, especially for the unfiltered model N1.

The results of all three models demonstrate that only off-line SUS prediction (blue) does not follow an exponential decrease during batch-phase while the other predictions adapt well to the reference (grey circle) (Figure 5.18). The measurement of SN resulted in a model which performed best without any pre-processing (orange rhombus). This indicates that the cells in the cell suspension have an influence on the performance of glycerol prediction. This issue was addressed in other works as well (Avila et al., 2012; Voß et al., 2017).

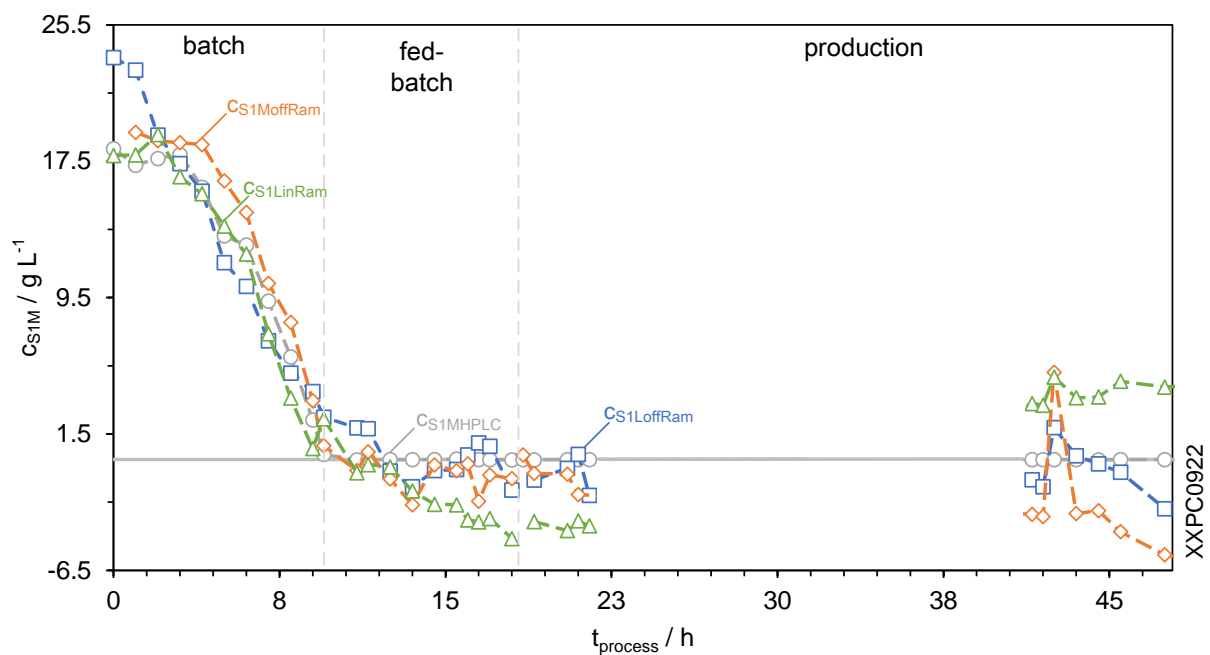


Figure 5.18: Prediction of glycerol concentration with different measuring types. Grey circle: reference by HPLC; blue square: off-line Raman spectroscopy with cell suspension predicted by N12; orange rhombus: off-line Raman spectroscopy with supernatant predicted by N1; green triangle: in-line Raman spectroscopy with cell suspension predicted by N2.

Other than in the exemplification, the cultivation XXPC0922 predicted here was partly involved in modelling. The exemplification, in contrast, used an external PS which originated from

another cultivation not involved in modelling. Thus, the results here show predictions which are closer to the reference.

Overall, the off-line measurement resulted in lower RMSE_{cv} compared to the in-line probe. This could also be observed in the work of Voß (Voß, 2017). With both measuring types, the determination of glycerol was possible.

5.5 Prediction of Methanol Concentration

For quantification of methanol, samples of the production phase were used for calibration as batch- and fed-batch phase contained no methanol and to ensure uniform distribution. The reference values were derived from HPLC off-line measurements.

The VIP plot of the in-line probe shows that only specific wavenumber ranges were significant (Figure 5.19). Wavenumbers in the range of 1649–1668 cm⁻¹, at approximately 1978 cm⁻¹, and 2690–2708 cm⁻¹ comprise a VIP > 1. The spike around 2700 cm⁻¹ is due to the interaction of cosmic rays with the sensitive CCD detector and was consecutively excluded in modelling (Shaw et al., 1999). Due to the fact that model N8 was pre-processed with 1st derivative filter, the VIP plot appears noisy.

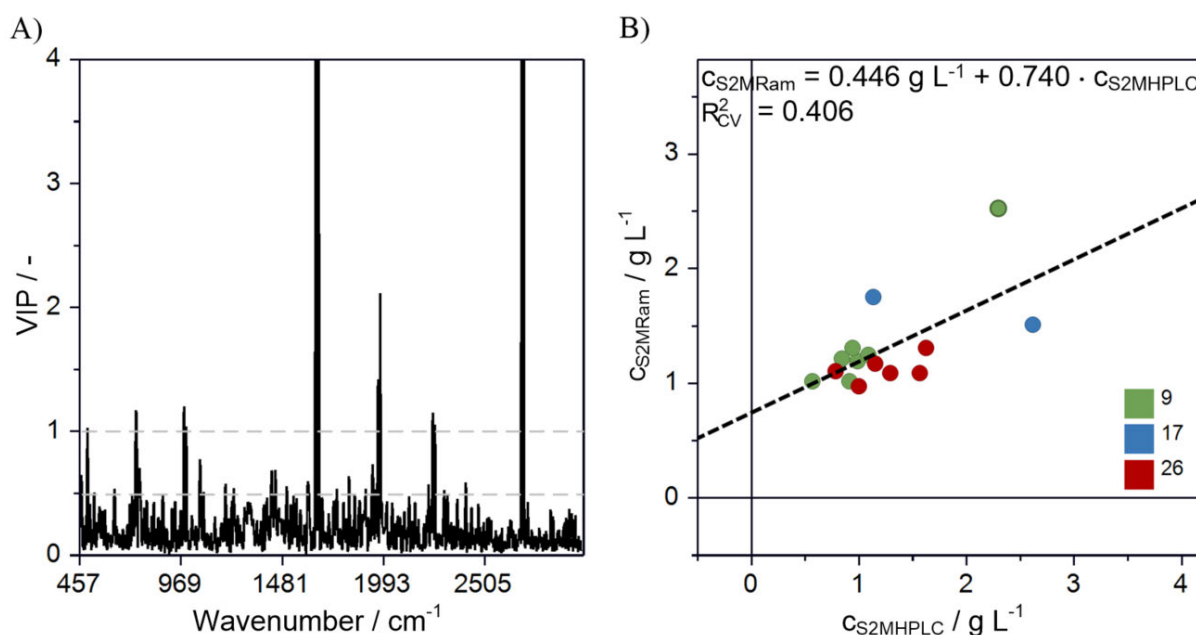


Figure 5.19: OPLS model parameters of in-line Raman cell suspension for prediction of methanol of model N2. A) Variable in projection (VIP), indicating wavenumbers above 1 being important. Three sharp spikes are visible: 1649–1668 cm⁻¹, around 1978 cm⁻¹, and 2690–2708 cm⁻¹; B) Observed vs. predicted cell density by cross-validation. Green: XXPC0922; blue: XX1722; red: XXPC2622.

After exclusion of the band around 2700 cm^{-1} , predictive power did not significantly improve. In literature, typical bands for methanol should be around 1035 cm^{-1} , 1450 cm^{-1} , 1461 cm^{-1} , and 2840 cm^{-1} (Emin et al., 2020; Voß et al., 2017). However, this could not be detected in this work. A reason for this could be the interfering fluorescence which occurs especially in production phase where eGFP is synthesized by the cells (Shaw et al., 1999). This could not be improved by any pre-processing methods.

The best methods were obtained when 1st derivative filter was included in pre-processing method (Table 5.4). Two to three OPLS components were required to achieve a RMSE_{cv} of 13.3–16.0 %. The lowest prediction error for methanol concentration was achieved by in-line Raman, with RMSEP_{VS} = 7.15 %.

As it could also be observed for glycerol prediction, in-line Raman performed best with 1stDer (cf. Table 5.3). This could be due to technical reasons of the Raman probe. Since the immersion probe is based on backscattering measurement, it is prone for fluorescent background. This issue is best corrected using the 1st derivative of the original spectra. For both compounds, glycerol and methanol, this can be observed for all models based on 1stDer or together with SG (cf. Appendix).

Table 5.4: Summary of predicted methanol concentration c_{S2M} with both off-line and in-line Raman spectroscopy.

		N9 (LinC-SNV-SG)	N4 (1 st Der-SG)	N2 (1 st Der)
Monitoring	/ –	off-line	off-line	in-line
Fluid	/ –	SUS	SN	SUS
n_{CS}	/ –	37	29	25
n_{VS}	/ –	10	8	7
r	/ –	1+3+0	1+2+0	1+3+0
R2Y	/ –	0.797	0.427	0.603
Q2	/ –	0.491	0.163	0.348
RMSE _{cv}	/ g L ⁻¹	0.849	0.817	0.705
RMSE _{cv,rel}	/ %	16.0	15.4	13.3
MBE _{cv}	/ –	0.0735	-0.0180	-0.00609
RMSEP _{VS}	/ g L ⁻¹	0.934	0.590	0.380
RMSEP _{VS,rel}	/ %	17.6	11.1	7.15
MBEP _{VS}	/ –	0.225	-0.0396	0.0770
RMSEP _{PS}	/ g L ⁻¹	2.86	2.05	2.19
RMSEP _{PS,rel}	/ %	53.79	38.5	41.2
MBEP _{PS}	/ –	0.271	1.89	2.08

It should be noted that OPLS of methanol was restrained by the fact that only 25–37 observations were available for CS. Furthermore, only two levels of methanol concentration were representing CS. First, the time range directly after production phase initiation where methanol content valued around 2.5 g L^{-1} . Second, the time during production phase where the set point was maintained at $c_{S2Mw} = 1.5 \text{ g L}^{-1}$. As sampling frequency was lowered during production phase, there was no representative distribution of values for the CS available. Simultaneously, the low number of available observations of the VS is not representative for RMSEP_{VS} .

Evaporation of the volatile methanol during the course of sample handling and measurement may have caused deviations between in-line and off-line values, leading to a higher variance of RMSE_{cv} and RMSEP . This was observed by comparing RMSEP_{VS} off-line and in-line values. For the in-line probe, $\text{RMSEP}_{VS} = 0.380 \text{ g L}^{-1}$. This value is more than twice the error of the off-line suspension measurement of $\text{RMSEP}_{VS} = 0.934 \text{ g L}^{-1}$. Also, the supernatant has a higher RMSEP_{VS} than the in-line measured methanol concentration. Another influence on the difference between supernatant and suspension could lie in the presence of cells. However, when comparing off-line supernatant with off-line suspension, the difference between both RMSE_{cv} values is low.

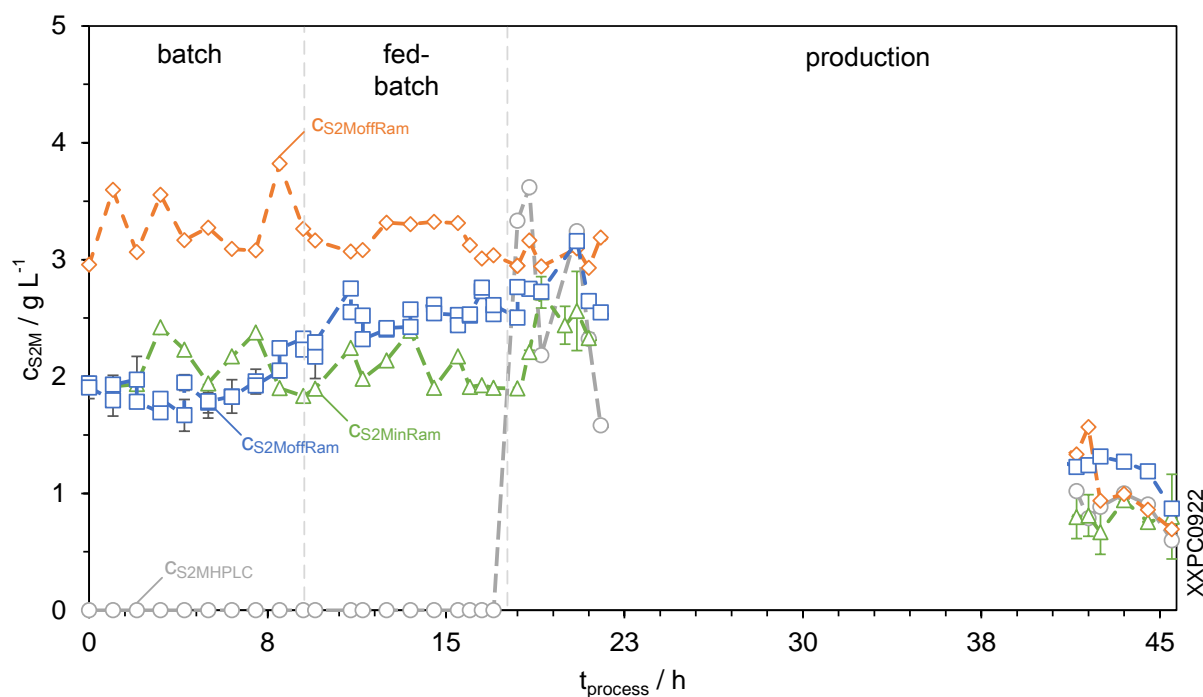


Figure 5.20: Predicted methanol concentration c_{S2M} . Grey circle: reference by HPLC; blue square: off-line Raman spectroscopy with cell suspension predicted by N9; orange rhombus: off-line Raman spectroscopy with supernatant predicted by N4; green triangle: in-line Raman spectroscopy with cell suspension predicted by N2.

Additionally, high discrepancies among $RMSE_{cv}$ and $RMSEP_{PS}$ were observed, implying an unstable model when removing single samples. Further investigation with more taken samples would be necessary for a final evaluation of the applicability of methanol determination with Raman spectroscopy with this Raman probe. However, this potential, was proven in other works (Paul et al., 2016; Voß, 2017).

5.6 Prediction of Cell Concentration

Measurement techniques based on optical approaches are established methods for determination of cell density. As mentioned in Chapter 4.4.4, an in-line turbidity probe was used in parallel during this work.

For prediction of cell density, all three phases of cultivation were applied for calibration. Therefore, only CV was available. As there were no cells in the supernatant due to centrifugation, only cell suspension was investigated. The reference derived from determination of cell dry weight.

The in-line Raman spectra are inherently different in appearance than off-line spectra (Figure 5.21) (cf. Figure 5.16).

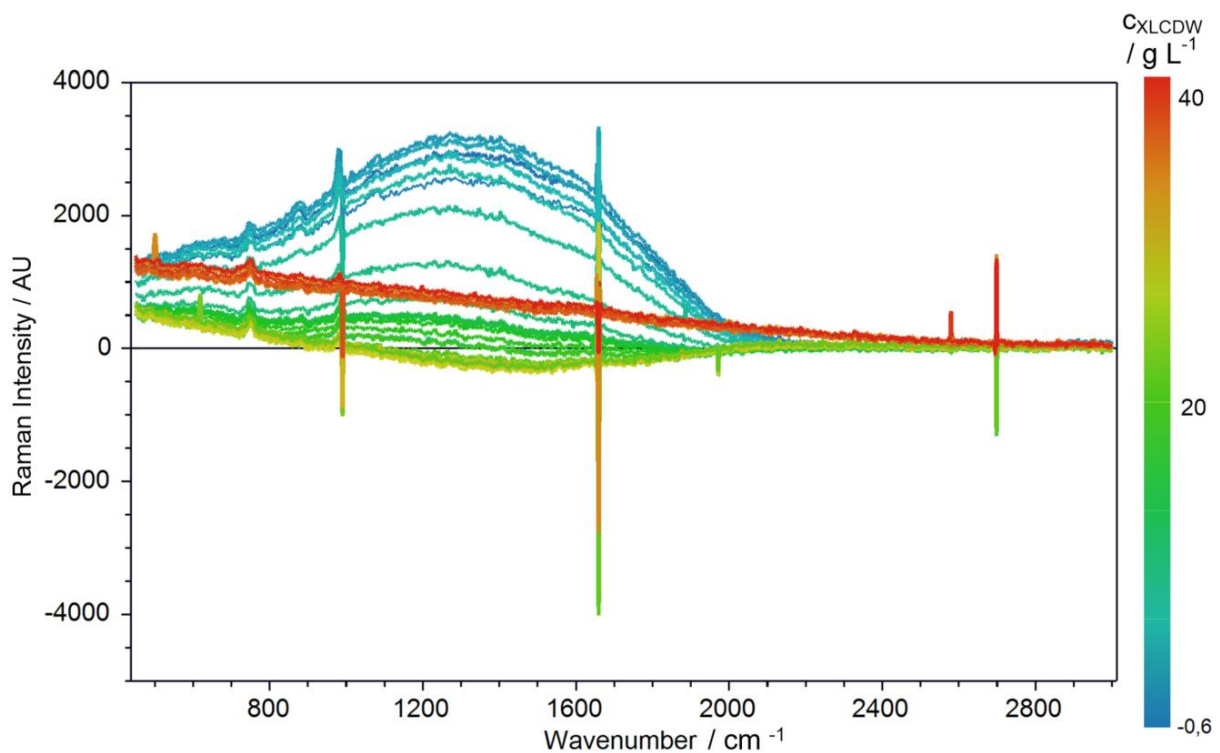


Figure 5.21: In-line Raman spectra during cultivation XXPC0922. Spectra are coloured according to their cell dry weight c_{XLCDW} .

The sharp bands at 582 cm^{-1} and 748 cm^{-1} correspond to sapphire material interference caused by the window of the immersion probe tip (Berry et al., 2015). It could be observed that with increase of cell concentration, the relative amount of backscattered light reaching the detector also increased due to the accumulation of scattering material in the culture broth. For Raman probes based on the backscattering principle, the increased scattering leads to a rise in the overall absorption, causing baseline shifts (Cervera et al., 2009). This can be seen when comparing the lower concentrated orange spectra with the highly concentrated red spectra.

In lower cell concentrations, especially XXPC0922 (green circle) shows a non-linear behaviour (Figure 5.22B).

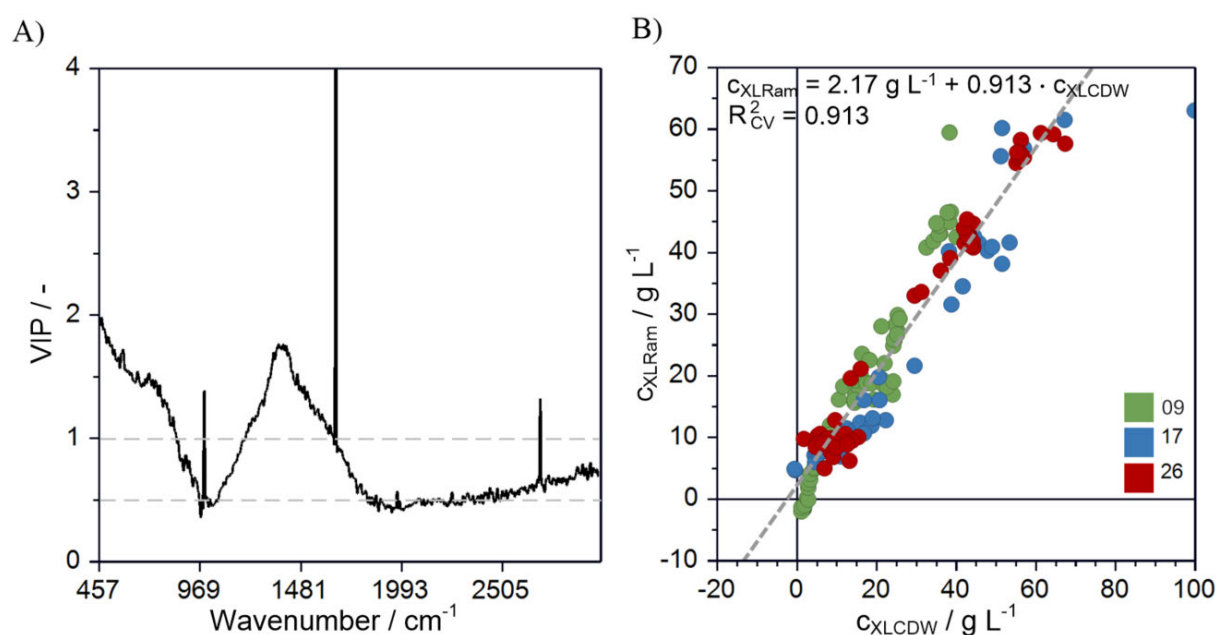


Figure 5.22: OPLS model parameters of in-line Raman cell suspension for prediction of cell density of model N8. A) Variable in projection (VIP), indicating wavenumbers above 1 being important; B) Observed vs. predicted cell density by cross-validation. Green: XXPC0922; blue: XX1722; red: XXPC2622.

The non-linear behaviour could not be corrected by log-transformation. Thus, the low-concentration range is worse predicted compared to higher values. This can be observed in the predicted cell concentration plotted against time (Figure 5.23).

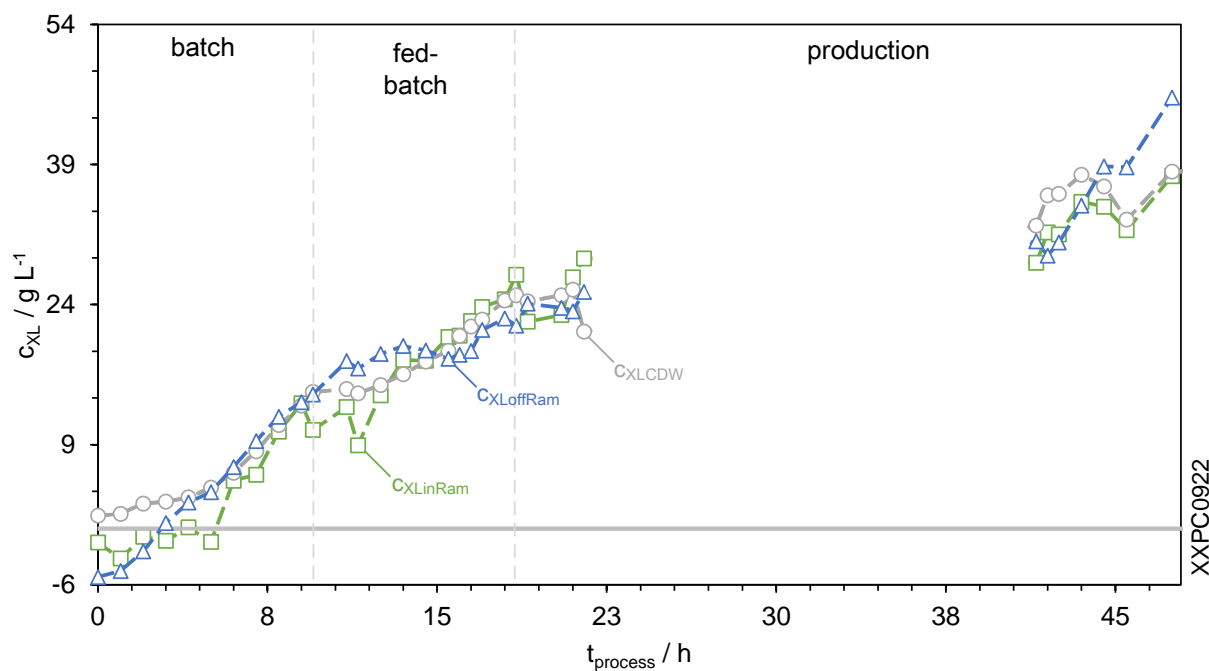


Figure 5.23: Predicted cell concentration c_{XL} in cell suspension for N8 by cross-validation using off-line and in-line Raman spectroscopy. Grey circle: HPLC reference; blue triangle: off-line Raman spectroscopy by cross-validation; green square: in-line Raman spectroscopy.

The best OPLS models for cell density c_{XL} was obtained by using a combination of linear correction and SNV (Table 5.5). Six OPLS components were used to describe about 93 % of the spectral variance R^2Y and led to a $RMSE_{cv}$ of 4.69 g L^{-1} and 5.22 g L^{-1} for off-line and in-line, respectively. These correspond to a relative error of 4.69 % and 5.22 %, respectively.

Table 5.5: Summary of predicted cell concentration in cell suspension (SUS) c_{XL} with both off-line and in-line Raman spectroscopy.

		N8 (LinC-SNV)	N8 (LinC-SNV)
Monitoring	/ –	off-line	in-line
Fluid	/ –	SUS	SUS
n_{CS}	/ –	95	148
n_{VS}	/ –	24	37
r	/ –	1+6+0	1+6+0
R^2Y	/ –	0.931	0.929
Q^2	/ –	0.909	0.896
$RMSE_{cv}$	/ g L^{-1}	4.69	5.22
$RMSE_{cv,rel}$	/ %	7.05	6.24
MBE_{cv}	/ –	–0.0270	0.164
$RMSEP_{VS}$	/ g L^{-1}	4.87	9.48
$RMSEP_{VS,rel}$	/ %	4.93	9.60
$MBEP_{VS}$	/ –	–0.110	3.85

Low cell concentrations were generally underpredicted. This was also observed in other works (Berry et al., 2015; Voß, 2017; Whelan et al., 2012). Other than the other analytes of this work, cell density as such is not a chemical species. However, prediction was still possible. It is likely that the model is based on Raman signals of chemical species which are correlated to cell growth but not directly measured. As many organic compounds naturally fluoresce, the increase of fluorescent background during cultivation may be due to steady accumulation of metabolism products (Berry et al., 2015). This may also be the reason why low cell densities are poorly predicted compared to higher values due to the lower content of by-products.

The quality of the models is sufficient over the whole course of cultivation. For this analyte, the in-line probe yielded a lower RMSE_{cv}. However, both values, off-line and in-line, are similar in value. Therefore, the difference is negligible, resulting in tolerable performance for both measurement types. Further investigations on quantification with Raman immersion probe are proposed to evaluate whether the Raman immersion probe can become a competitive alternative for the turbidity probe.

5.7 Prediction of Total Protein Concentration

For prediction of total protein concentration C_{PtotM} , two cultivations were used instead of three. XXPC2622 was excluded from calibration as the reference measurement by Bradford assay was in poor quality (not shown here). All three phases of cultivation were used, batch-, fed-batch, and production phase. Therefore, prediction with PS was not possible.

Like the other analytes, RMSE_{cv} yielded in the order of 9.93–12.0 %. A linear correlation can be observed between predicted vs. reference (Figure 5.24B). Its sharp band at 1650 cm^{-1} is typical for the functional group amide I, indicating the presence of proteins (Figure 5.24A) (Sivakesava et al., 2001).

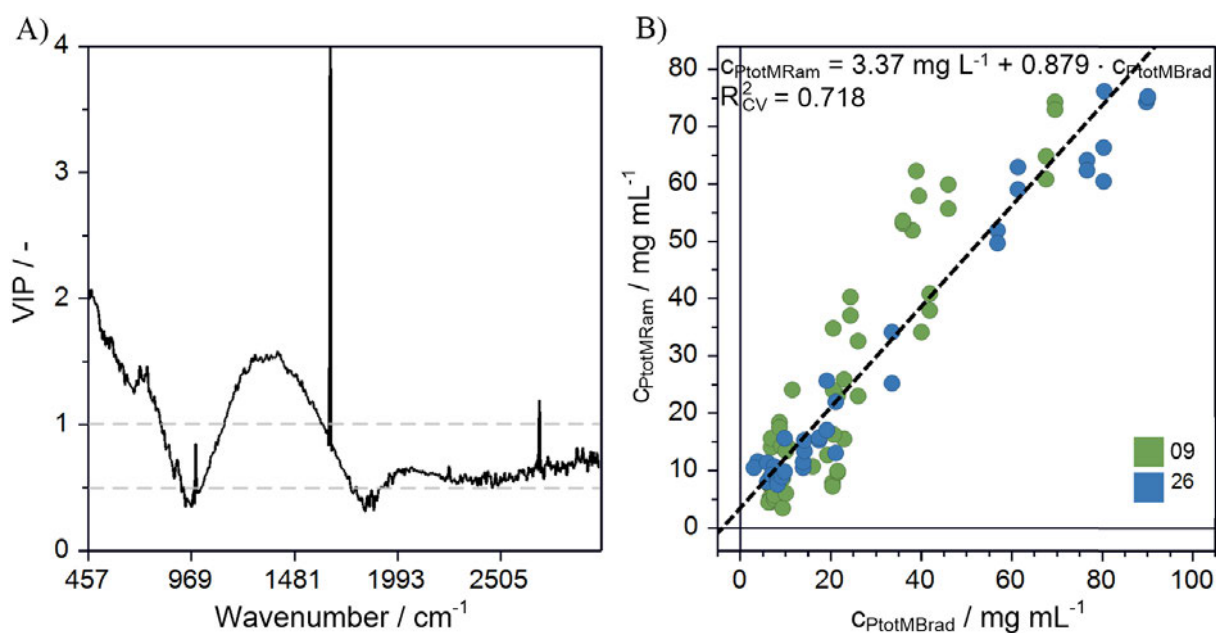


Figure 5.24: OPLS model parameters of in-line Raman suspension for prediction of total protein c_{PtotM} of model N11. A) Variable in projection (VIP), indicating wavenumbers above 1 being important; B) Observed vs. predicted cell density by cross-validation. Green: XXPC0922; blue: XX1722.

Three to five OPLS component yielded about 83 % of the variation (Table 5.6). As the values between RMSEcv and RMSEP have low discrepancies among the models, it is expected that the model is stable. SNV was required to achieve lowest error predictions for all models.

Table 5.6: Summary of predicted total protein concentration c_{PtotM} with both off-line and in-line Raman spectroscopy.

		N10 (SNV)	N10 (SNV)	N11 (SNV-SG)
Monitoring	/ -	off-line	off-line	in-line
Fluid	/ -	SUS	SN	SUS
n_{CS}	/ -	80	40	87
n_{VS}	/ -	21	10	22
r	/ -	1+4+0	1+3+0	1+5+0
R^2Y	/ -	0.784	0.805	0.879
Q^2	/ -	0.740	0.728	0.836
RMSEcv	/ mg L ⁻¹	10.1	12.0	9.93
RMSEcv _{rel}	/ %	11.7	14.0	11.5
MBEcv	/ -	-0.0715	-0.406	0.0248
RMSEP _{VS}	/ mg L ⁻¹	14.0	13.8	10.2
RMSEP _{VSrel}	/ %	16.3	16.1	11.9
MBEP _{VS}	/ -	-4.26	-1.85	2.85

It can be observed that SNV was required when using in-line CS based on all phases of cultivation (batch, fed-batch, and production phase) (cf. Table 5.5). Comparing this fact with in-line

glycerol and methanol prediction, 1stDer was required due to fluorescent background. It can be concluded that the scattering effect prevails the fluorescent effect during the whole course of cultivation. Thus, SNV filter is required to correct the scatter effects.

For prediction of the total protein concentration course, all three models perform well. A quantification of the total protein concentration was possible. However, in the higher concentration ranges, the model lacks in precision. It is likely that the total protein concentration determination happens through indirect measurement of another analyte. This could be caused by overlapping signals, masking the small signals of this low-concentrated analyte (Paul et al., 2016).

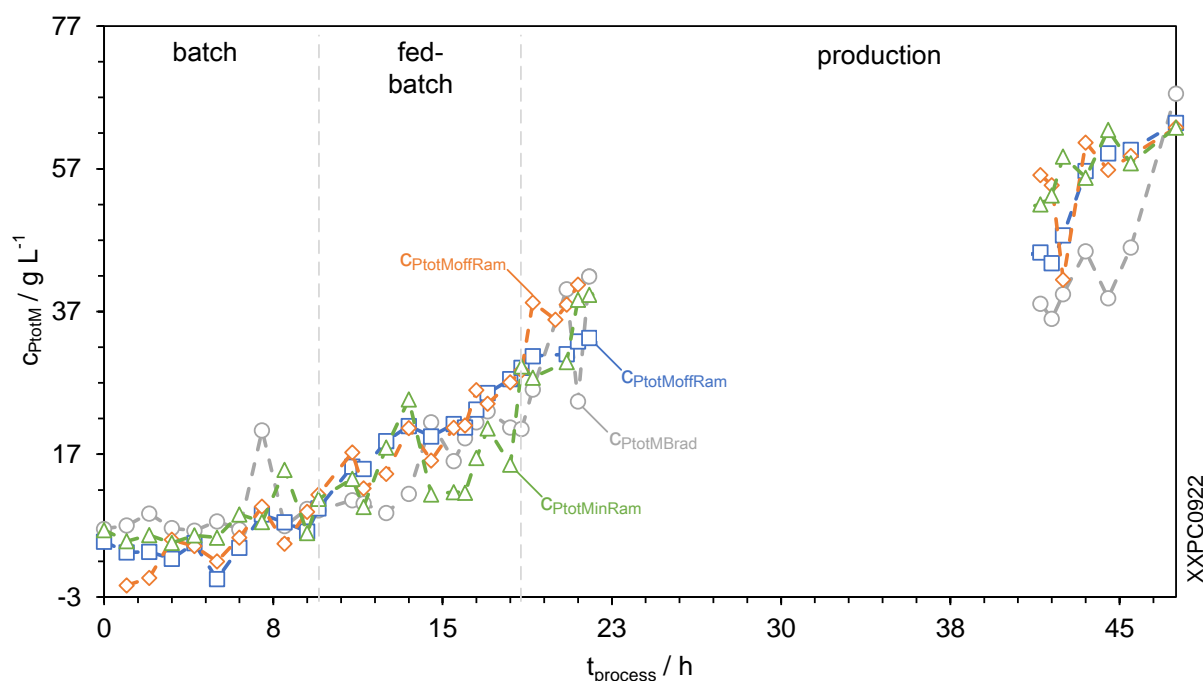


Figure 5.25: Predicted total protein concentration c_{PtotM} by cross-validation using off-line and in-line Raman spectroscopy. Grey circle: Bradford assay reference; blue triangle: off-line Raman spectroscopy of suspension; orange rhombus: off-line Raman spectroscopy of supernatant; green square: in-line Raman spectroscopy of suspension.

For prediction of total protein, the in-line probe performed slightly better than the off-line measurement. However, both RMSEcv values are of same order, concluding that both measurement types give reason to use Raman spectroscopy for determination of total protein concentration.

5.8 Sources of Error in Methodology

This chapter deals with potential error sources which occurred especially during MVDA. Also, the cultivation XXPC2622 is further described due to its outstanding spectra.

5.8.1 Critical View on Executed MVDA

The Hotelling's T^2 method identifies outliers, removing them prevents skewing of the model. However, the threshold for removal must be considered carefully. When sample points at the ends of the calibration range are removed, the scope over which the model is valid is reduced. Despite the fact that the removal of data outside the 95 % confidence interval adds more restrictions on the process ranges where the model is valid, it also increases the model's adherence to these bounds (Berry et al., 2015).

OPLS modelling procedure is adapted such that it generates models which are most accurate during main portion of the run. E.g., for glycerol concentration determination, only batch- and fed-batch phases were considered for modelling. However, this led to high RMSEP for production phase. This trade-off must be kept in mind and adapted, depending on the current problem. Residual data must be considered in model building as these residuals give information about which data remains unexplained by the model. Large residuals in X- or Y-data are indicative of poor models (Eriksson et al., 2006b).

Although time-sensitive data was evaluated by use of permutations, the tailored models were in contrast sensitive to data outside these models. Errors based on CV displays the simplest method for implementation and is least time-consuming. Still, the quality of model evaluation is limited. This could be observed in the comparison of RMSE_{CV} and RMSEP. E.g., in Chapter 5.4, N12 showed the lowest RMSE_{CV}, its RMSEP_{VS} does not remain lowest. Although 20 % of the data was left out from modelling, the data remains highly correlated. With this, similar data was used for validating the model. It was observed that RMSE_{CV} tends to understate how much a model has overfitted the CS. This could be seen by comparing RMSE_{CV} with the corresponding RMSE_{VS}. Most modelling combinations yielded a higher RMSEP than RMSE_{CV} (cf. Appendix), proposing RMSE_{CV} underestimates the actual error. To counteract this, more cultivations would be necessary to, 1. have a higher variability of data and, 2. use whole cultivations as PS. However, this exceeded the scope of this work.

The declared PS in this work, meaning all samples excluded within one cultivation, did not function as such. As the detection of, e.g., glycerol during production phase or methanol during batch phase, should be expectedly 0 g L⁻¹, the inclusion of further cultivations would be interesting to see whether the predictions in these phases are representatively moving in tolerable areas. In industry, this is referred to as golden batch. This enables an early failure detection during cultivation when certain limits are exceeded. Therefore, the high RMSEP_{PS} are to be considered with care in this work.

When comparing in-line with off-line Raman spectroscopy, several factors have to be kept in mind. In general, the culture broth progresses from translucent to increasingly turbid as the cells multiply and by-products or products accumulate. Therefore, light scattering increases which could be observed in Raman spectra over time. This effect can be corrected by sampling and off-line analysing the supernatant. However, in a bioreactor, centrifugation of the culture broth is not possible. Additionally, gas sparging is applied in order to supply the cells with oxygen. These air bubbles and the increasing biomass lead to further attenuation of light scattering (Lee et al., 2004). This aspect contributes to the discrepancies between in-line and off-line results. In Voß' work, the same pre-processing methods were applied for both off-line and on-line Raman measurements (Voß, 2017). However, this work indicated high discrepancies when applying the same method for off-line and in-line. Spectra showed that in-line and off-line result were in different appearance. These differences originate from gas sparging, agitation, and temperature maintenance in the bioreactor (Ghita et al., 2018; Lee et al., 2004; Zobeiri et al., 2022). Therefore, different pre-processing methods were applied in order to compensate this environment.

Another fact to consider is that cultivation XXPC2622 appeared differently in spectra (cf. Figure 5.2, Figure 5.6). The causes of the difference will be further discussed in Chapter 5.8.2. From MVDA view, the difference might have an influence on the model outcome. All RMSE_{cv} obtained in this work ranged in the order around 10 %. In other projects with the same bioreactor systems, RMSE_{cv} values of 1–3 % were achieved (Paul et al., 2016; Voß et al., 2017). The Raman probe used originated from a different manufacturer. However, it implies that lower RMSE_{cv} values are realisable. Therefore, the multivariate calibration with inclusion of XXPC2622 could have led to the increased overall prediction error. For further improvement, the faulty cultivation should have been already excluded in the stage of PCA to enable representation of the relevant variation of the process.

5.8.2 Cultivation XXPC2622

As previously described, cultivation number XXPC2622 deviated from the other two used cultivations XXPC1722 and XXPC0922. This section serves to describe the prevailing situation (Figure 5.26).

The difference in spectra could already be observed during the first samples of batch-phase (cf. Figure 5.2). Consequently, the fault must be prior to inoculation. In fact, due to circumstances, the inoculation was executed 1 h later than initially planned. This could imply cells reaching

stationary phase or even cells decaying. Another reason could be the fact that preculture was contaminated with foreign substances. The OD_{600} of the preculture was half that of precultures in XXPC0922 and XXPC1722 at preculture start. At cultivation begin, the cell density c_{XLCDW} strongly fluctuated which could not be observed for the other two cultivations. Both hypotheses are substantiated by the fact that in XXPC2622 additional anti-foam agent had to be added at the beginning of fed-batch phase. The increased foam formation can be due to contamination or cell debris (Delvigne & Lecomte, 2009). This in contrast, leads to interference with Raman measurement, resulting in a distorted spectral data set for multivariate calibration (Whelan et al., 2012). Due to the anti-foam addition, the volumetric oxygen transfer coefficient to the cells decreased. Thus, a higher stirrer speed was required to maintain the DO level at its set point.

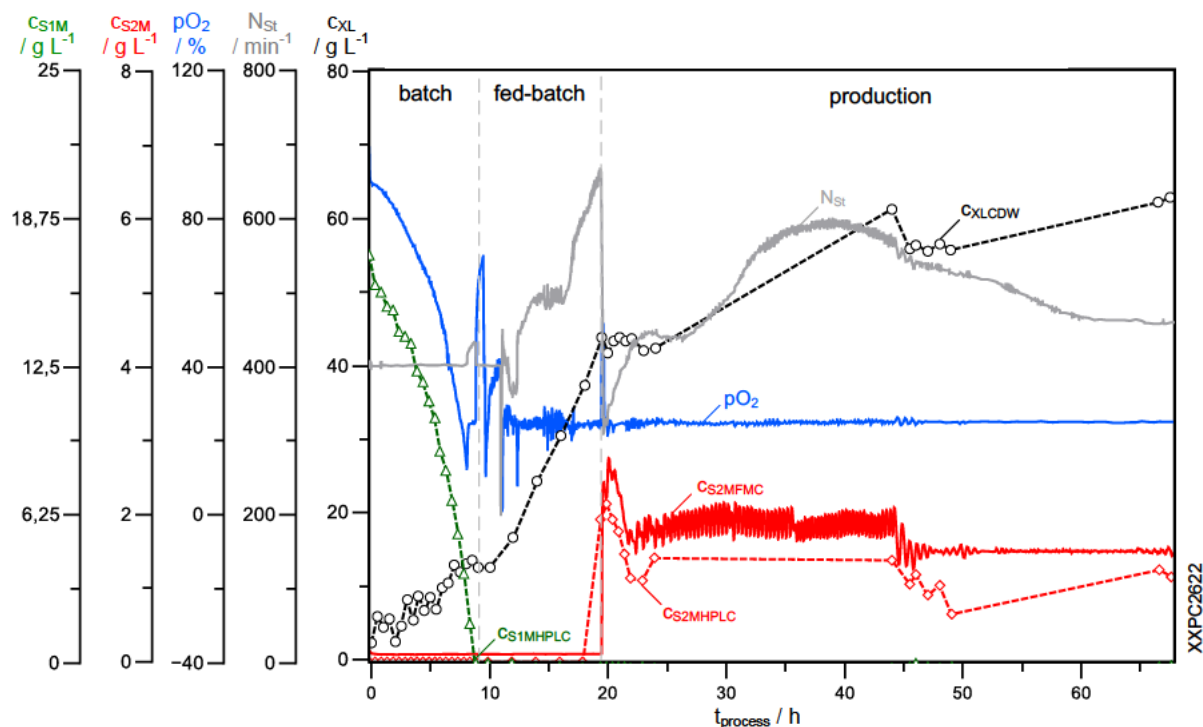


Figure 5.26: Course of cultivation XXPC2622. Compared to the theoretical course of cultivation (cf. Figure 3.5), both courses look similar. The pO_2 at the end of batch and beginning of fed-batch phase indicates issues with pO_2 control. CS_{1M} : glycerol concentration; CS_{2M} : methanol concentration; pO_2 : partial oxygen pressure; c_{XL} : cell dry weight concentration; N_{St} : stirrer speed.

5.9 Process Control and Parameters

One key element of PAT is the surrounding periphery of a bioreactor. Therefore, parameter estimation of devices is necessary. In the following, parameter estimation for the turbidity probe and OD measured cell density is introduced.

First, the parameters for the turbidity probe are presented. The parameters were determined by the Nelder-Mead method in MATLAB (also named downhill simplex method). The corresponding MATLAB script can be found in the Appendix.

Equation (4.4) was extended by the term of stirrer speed N_{st} (Cornelissen, 2004). This enables to counteract the influence of the agitation on the turbidity signal. Also, only measurements during batch and fed-batch phase were considered due to the fluorescent background and change of morphology while production phase.

In Figure 5.27, a comparison of incorporation (Figure 5.27A) and exclusion (Figure 5.27B) of agitation, respectively, is depicted for the cultivation XXPC0922. First, for parameter determination, c_{XL} is plotted against the turbidity signal S_{turb} . For this, the measurements S_{turb} are interpolated to equal time points as the off-line determined cell density c_{XLCDW} . Then, the Nelder-Mead method is applied in order to determine the parameters a , b , and c . When the parameters are known the cell density can be estimated by monitoring the turbidity signal.

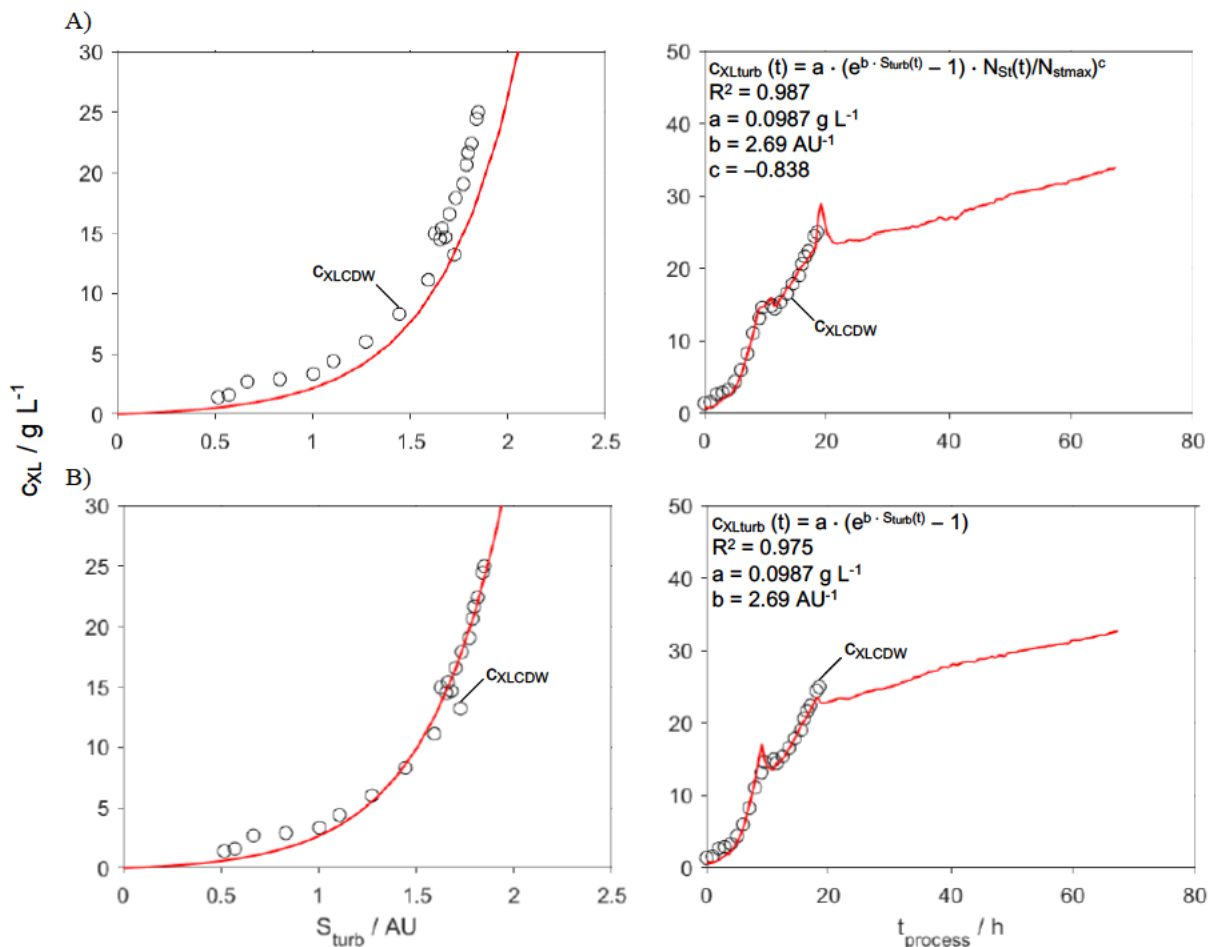


Figure 5.27: On-line estimation of cell density in cultivation vessel. A) With incorporation of stirrer speed; B) Without stirrer speed. On left-hand side, c_{XL} is plotted against S_{turb} . On right-hand side, c_{XL} is plotted against process time $t_{process}$. Red solid line: fitted c_{XL} .

It can be observed that by incorporating the actual stirrer speed, the estimation is more precise. This leads to an increase of the goodness of fit R^2 from 97.5 % to 98.7 %. The corresponding parameters for equation (4.4) can be found in Table 5.7. The cell density estimation by the turbidity probe allows a more precise calculation of the feeding pump rate.

Table 5.7: Parameters for cell density estimation with the turbidity probe.

Setting	a / g L ⁻¹	b / AU ⁻¹	c / -	R ² / -
Incorporation of stirrer speed	0.0987	2.69	-0.838	0.987
Exclusion of stirrer speed	0.234	2.51	-	0.975

6 Conclusions

The objective of the present work was to investigate the applicability of both in-line and off-line Raman spectroscopy for the prediction of glycerol, methanol, cell, and total protein concentration. In the context of PAT, cultivations were executed with *Pichia pastoris*, comprising batch, fed-batch, and production phases. The OPLS models were evaluated regarding their predictive power towards the compounds analysed.

Application of MVDA demonstrated that correlations between Raman spectra and analytes could be made. Preliminary studies on the pre-processing tools using the weighted sum model yielded a selection of 12 pre-processing methods. The methods comprised the pre-processing tools SG9, LinC, SNV, and 1stDer. Depending on the analyte, different pre-processing methods led to varying RMSE_{cv} and RMSEP_{VS}. However, the application of pre-processing method(s) is crucial for the extraction of information out of Raman spectra. The most often applied pre-processing tool was SNV. Its scatter correcting property proved to be very useful for determination of total protein, cell, and methanol concentrations.

The lowest RMSE_{cv} with 6.24 % could be achieved with in-line prediction of cell concentration. However, the highest RMSE_{cv} with 18.8 % was also obtained with the immersion probe for glycerol prediction. Throughout the measurement types, in-line Raman probe generated more minimum RMSE_{cv} and RMSEP_{VS} values than off-line Raman spectroscopy. Methanol prediction demonstrated that a smaller calibration set may lead to instabilities of the model.

Both off-line and in-line Raman spectroscopy appears to be well suited for predictions of glycerol, methanol, cell density, and total protein in culture broths and supernatant.

7 Future Perspectives

The present work demonstrated that Raman spectroscopy was able to quantify the investigated analytes by use of MVDA. Further effort is required in order to develop strategies that result in even more reliable and capable models. Here, the step of early outlier detection plays a vital role. Within the scope of this work, the secreted protein eGFP was not investigated. This could be subject of more research to assess the influence of fluorescence upon Raman spectroscopy. Also, further examination on the influence of sparging and agitation in the culture vessel can enlighten potential disturbances of Raman spectroscopy. Alternatively, an approach with at-line Raman can be considered. Here, the culture broth is diverted from the culture vessel and may be returned to the process. By this means, potential influences of bubbles and turbulences by stirring are prevented. Also, investigations on other medium compounds or metabolites of *P. pastoris* would be interesting such as target protein, amino acids, viable cells, or ammonium consumption.

Subsequently, the process could be adapted even more to PAT approaches by implementing on-line monitoring to the bioprocess. The Raman immersion probe could substitute the turbidity probe. Instead, the vacant port of the bioreactor could be used for other electrodes where Raman comes to its limits. Feed-back loops using Raman spectroscopy for, e.g., glycerol or methanol feeding is possible. For this, SIMCA[®]-online is available. Also, the used process software MultiSpec[®] Pro II offers extensions for SIMCA-Q integration.

8 References

- Abu-Absi, N. R.; Martel, R. P.; Lanza, A. M.; Clements, S. J.; Borys, M. C.; Li, Z. J. Application of spectroscopic methods for monitoring of bioprocesses and the implications for the manufacture of biologics. *Pharmaceutical Bioprocessing* **2014**, *2* (3), 267–284. DOI: 10.4155/pbp.14.24.
- Avila, T. C.; Poppi, R. J.; Lunardi, I.; Tizei, P. A. G.; Pereira, G. A. G. Raman spectroscopy and chemometrics for on-line control of glucose fermentation by *Saccharomyces cerevisiae*. *Biotechnology progress* **2012**, *28* (6), 1598–1604. DOI: 10.1002/btpr.1615.
- Berry, B.; Moretto, J.; Matthews, T.; Smelko, J.; Wiltberger, K. Cross-scale predictive modeling of CHO cell culture growth and metabolites using Raman spectroscopy and multivariate analysis. *Biotechnology progress* **2015**, *31* (2), 566–577. DOI: 10.1002/btpr.2035.
- Biotechnologie Kempe. Homepage. 2022. – URL: <http://biotechnologie-kempe.eu/html/>.
- Bradford, M. M. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Analytical Biochemistry* **1976**, *72* (1-2), 248–254. DOI: 10.1016/0003-2697(76)90527-3.
- Buckley, K.; Ryder, A. G. Applications of Raman Spectroscopy in Biopharmaceutical Manufacturing: A Short Review. *Appl Spectrosc* **2017**, *71* (6), 1085–1116. DOI: 10.1177/0003702817703270.
- Butler, H. J.; Ashton, L.; Bird, B.; Cinque, G.; Curtis, K.; Dorney, J.; Esmonde-White, K.; Fullwood, N. J.; Gardner, B.; Martin-Hirsch, P. L.; Walsh, M. J.; McAinsh, M. R.; Stone, N.; Martin, F. L. Using Raman spectroscopy to characterize biological materials. *Nature protocols* **2016**, *11* (4), 664–687. DOI: 10.1038/nprot.2016.036.
- Carl Roth. Gebrauchsanweisung - ROTI@Quant: Proteinbestimmung nach Bradford. 2021. – URL: <https://www.carlroth.com/medias/BA-K015-DE.pdf?context=bWFzdGVyfGluc3Ry-dWN0aW9uc3w2NzE2OTR8YXBwbGljYXRpb24vcGRmfGluc3Ry-dWN0aW9ucy9oYTcvaDAyLzkwMzg2MTU1NzY2MDYucGRmfDAyNjdjYzNi-YzQ5N2ZjZWNIM-TIxYjFhMDhlZmYxYzVjN2EzYzZiZTg2M2Y0NWZkMjhjMjBhNzU4NDI2NGEyZjY> (accessed October 17, 2022).
- Cervera, A. E.; Petersen, N.; Lantz, A. E.; Larsen, A.; Gernaey, K. V. Application of near-infrared spectroscopy for monitoring and control of cell culture and fermentation. *Biotechnology progress* **2009**, *25* (6), 1561–1581. DOI: 10.1002/btpr.280.
- Chin, W. W.; Marcoulides, G. The Partial Least Squares Approach to Structural Equation Modeling. *Advances in Hospitality and Leisure* [Online] **1998**.
- Churchman, C. W.; Ackoff, R. L. An Approximate Measure of Value. *OR* **1954**, *2* (2), 172–187. DOI: 10.1287/opre.2.2.172.
- Cornelissen, G. *Integrierte Bioprozessentwicklung zur Herstellung pharmakologischer wirksamer Proteine mit Pichia pastoris*. Zugl.: Hannover, Univ., Diss., 2004 u.d.T.: Cornelissen, Gesine: Integrierte Bioprozessentwicklung zur Herstellung pharmakologisch

-
- wirksamer Proteine mit *Pichia pastoris*; Fortschritt-Berichte VDI Reihe 17, Biotechnik, Medizintechnik 249; VDI-Verl.: Düsseldorf, 2004.
- Danzer, K.; Fischbacher, C.; Hobert, H.; Jagemann, K.-U. *Chemometrik*; Springer eBook Collection Life Science and Basic Disciplines; Springer Berlin Heidelberg: Berlin, Heidelberg, 2001.
- Delvigne, F.; Lecomte, J. Foam Formation and Control in Bioreactors. In *Encyclopedia of Industrial Biotechnology*; Flickinger, M. C., Ed.; John Wiley & Sons, Inc: Hoboken, NJ, USA, 2009. DOI: 10.1002/9780470054581.eib326.
- El-Mansi, E. M. T.; Nielsen, J.; Mousdale, D.; Allman, T.; Carlson, R.; Carlson, R. P. *Fermentation Microbiology and Biotechnology, Fourth Edition*; CRC Press: Fourth edition. | Boca Raton : Taylor & Francis, 2018., 2018.
- Emin, A.; Hushur, A.; Mamtimin, T. Raman study of mixed solutions of methanol and ethanol. *AIP Advances* **2020**, *10* (6), 65330. DOI: 10.1063/1.5140722.
- Engel, J.; Gerretzen, J.; Szymańska, E.; Jansen, J. J.; Downey, G.; Blanchet, L.; Buydens, L. M. Breaking with trends in pre-processing? *TrAC Trends in Analytical Chemistry* **2013**, *50*, 96–106. DOI: 10.1016/j.trac.2013.04.015.
- Eriksson, L.; Andersson, P. L.; Johansson, E.; Tysklind, M. Megavariate analysis of environmental QSAR data. Part I--a basic framework founded on principal component analysis (PCA), partial least squares (PLS), and statistical molecular design (SMD). *Molecular diversity* **2006a**, *10* (2), 169–186. DOI: 10.1007/s11030-006-9024-6.
- Eriksson, L.; Byrne, T.; Johansson, E.; Trygg, J.; Vikström, C. *Multi-and Megavariate Data Analysis Basic Principles and Applications Third Revised Edition*. Chapter 18: Process Analytical Technology (PAT) and Quality by Design (QBD), 2006b.
- Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Trygg, J.; Wikström, C.; Wold, S. *Multi- and megavariate data analysis: Part I Basic Principles and Applications. Second revised and enlarged edition*, 2013.
- Esmonde-White, K. A.; Cuellar, M.; Uerpmann, C.; Lenain, B.; Lewis, I. R. Raman spectroscopy as a process analytical technology for pharmaceutical manufacturing and bioprocessing. *Analytical and bioanalytical chemistry* **2017**, *409* (3), 637–649. DOI: 10.1007/s00216-016-9824-1.
- European Medicines Agency. Q8 (R2) Step 5 Pharmaceutical Development **2017**.
- Fishburn, P. C. Letter to the Editor—Additive Utilities with Incomplete Product Sets: Application to Priorities and Assignments. *Operations Research* **1967**, *15* (3), 537–542. DOI: 10.1287/opre.15.3.537.
- Gabrielsson, J.; Jonsson, H.; Airiau, C.; Schmidt, B.; Escott, R.; Trygg, J. OPLS methodology for analysis of pre-processing effects on spectroscopic data. *Chemometrics and Intelligent Laboratory Systems* **2006**, *84* (1-2), 153–158. DOI: 10.1016/j.chemolab.2006.03.013.
- Ghita, A.; Matousek, P.; Stone, N. Sensitivity of Transmission Raman Spectroscopy Signals to Temperature of Biological Tissues. *Scientific reports* **2018**, *8* (1), 8379. DOI: 10.1038/s41598-018-25465-x.
- Goldrick, S.; Umprecht, A.; Tang, A.; Zakrzewski, R.; Cheeks, M.; Turner, R.; Charles, A.; Les, K.; Hulley, M.; Spencer, C.; Farid, S. S. High-Throughput Raman Spectroscopy
-

-
- Combined with Innovate Data Analysis Workflow to Enhance Biopharmaceutical Process Development. *Processes* **2020**, 8 (9), 1179. DOI: 10.3390/pr8091179.
- Gosselin, R.; Rodrigue, D.; Duchesne, C. A Bootstrap-VIP approach for selecting wavelength intervals in spectral imaging applications. *Chemometrics and Intelligent Laboratory Systems* **2010**, 100 (1), 12–21. DOI: 10.1016/j.chemolab.2009.09.005.
- Han, W.-J. *HPLC a practical guide*; RSC chromatography monographs; Royal Society of Chemistry: Cambridge, 1999.
- Hotelling, H. A Generalized T Test and Measure of Multivariate Dispersion [Online] **1951**, 23–41.
- Huang, J.; Romero-Torres, S.; Moshgbar, M. Practical Considerations in Data Pre-treatment for NIR and Raman Spectroscopy. *American Pharmaceutical Review* [Online] **2010**, 13, 116–127. <https://www.americanpharmaceuticalreview.com/Featured-Articles/116330-Practical-Considerations-in-Data-Pre-treatment-for-NIR-and-Raman-Spectroscopy/> (accessed September 7, 2022).
- InPhotonics. Homepage. 2022. – URL: <https://www.inphotonics.com/probes.htm> (accessed October 17, 2022).
- Keil, T.; Landenberger, M.; Dittrich, B.; Selzer, S.; Büchs, J. Precultures Grown under Fed-Batch Conditions Increase the Reliability and Reproducibility of High-Throughput Screening Results. *Biotechnology journal* **2019**, 14 (11), e1800727. DOI: 10.1002/biot.201800727.
- Kessler, W. *Multivariate Datenanalyse*; Wiley, 2006.
- Kim, E. J.; Kim, J. H.; Kim, M.-S.; Jeong, S. H.; Du Choi, H. Process Analytical Technology Tools for Monitoring Pharmaceutical Unit Operations: A Control Strategy for Continuous Process Verification. *Pharmaceutics* **2021**, 13 (6). DOI: 10.3390/pharmaceutics13060919.
- Lee, H. L.; Boccazzi, P.; Gorret, N.; Ram, R. J.; Sinskey, A. J. In situ bioprocess monitoring of *Escherichia coli* bioreactions using Raman spectroscopy. *Vibrational Spectroscopy* **2004**, 35 (1-2), 131–137. DOI: 10.1016/j.vibspec.2003.12.015.
- Madrid, R. E.; Felice, C. J. Microbial biomass estimation. *Critical reviews in biotechnology* **2005**, 25 (3), 97–112. DOI: 10.1080/07388550500248563.
- Martens, H.; Naes, T. *Multivariate calibration*, [Reprinted]; J. Wiley & Sons: Chichester, 1989.
- McCreery, R. L. *Raman Spectroscopy for Chemical Analysis*; John Wiley & Sons, Inc: Hoboken, NJ, USA, 2000.
- Nagy, B.; Farkas, A.; Borbás, E.; Vass, P.; Nagy, Z. K.; Marosi, G. Raman Spectroscopy for Process Analytical Technologies of Pharmaceutical Secondary Manufacturing. *AAPS PharmSciTech* **2018**, 20 (1), 1. DOI: 10.1208/s12249-018-1201-2.
- Oliveri, P.; Malegori, C.; Casale, M. Chemometrics: multivariate analysis of chemical data. *Chemical Analysis of Food*; Elsevier, 2020; pp 33–76. DOI: 10.1016/B978-0-12-813266-1.00002-4.
- optek-DANULAT. Homepage. 2022. – URL: <https://www.optek.com>.
- Otto, M. *Chemometrie. Statistik u. Computereinsatz in d. Analytik*; VCH: Weinheim, 1997.
- Pal, R. Validation methodologies. *Predictive Modeling of Drug Sensitivity*; Elsevier, 2017; pp 83–107. DOI: 10.1016/B978-0-12-805274-7.00004-X.
-

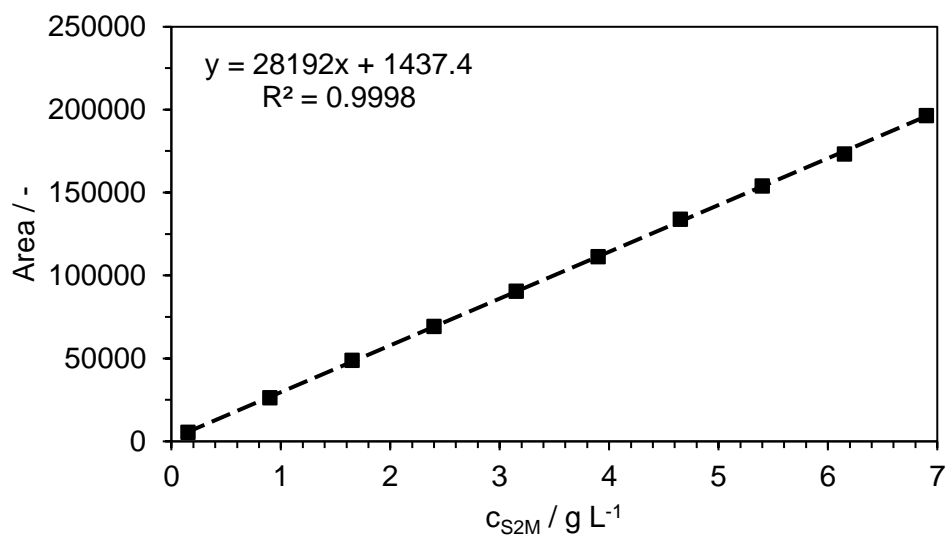
-
- PAT— A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance*, 2004.
- Paul, A.; Carl, P.; Westad, F.; Voss, J.-P.; Maiwald, M. Towards Process Spectroscopy in Complex Fermentation Samples and Mixtures. *Chemie Ingenieur Technik* **2016**, *88* (6), 756–763. DOI: 10.1002/cite.201500118.
- Peng, D. X.; Lai, F. Using partial least squares in operations management research: A practical guideline and summary of past research. *Journal of Operations Management* **2012**, *30* (6), 467–480. DOI: 10.1016/j.jom.2012.06.002.
- Raman, C. V.; Krishnan, K. S. A New Type of Secondary Radiation. *Nature* **1928**, *121* (3048), 501–502. DOI: 10.1038/121501c0.
- Rathore, A. S.; Bhambure, R.; Ghare, V. Process analytical technology (PAT) for biopharmaceutical products. *Analytical and bioanalytical chemistry* **2010**, *398* (1), 137–154. DOI: 10.1007/s00216-010-3781-x.
- Rathore, A. S.; Gautam, K. Process Analytical Technology: Strategies for Biopharmaceuticals. In *Encyclopedia of Industrial Biotechnology*; Flickinger, M. C., Ed.; John Wiley & Sons, Inc: Hoboken, NJ, USA, 2009. DOI: 10.1002/9780470054581.eib652.
- Rinnan, Å.; van Berg, F. den; Engelsen, S. B. Review of the most common pre-processing techniques for near-infrared spectra. *TrAC Trends in Analytical Chemistry* **2009**, *28* (10), 1201–1222. DOI: 10.1016/j.trac.2009.07.007.
- Ross, S. M.; Heinisch, C. *Statistik für Ingenieure und Naturwissenschaftler*, 3. Aufl.; Elsevier/Spektrum: München, 2006.
- Rostron, P.; Gerber, D. Raman Spectroscopy, a review. *International Journal of Engineering and Technical Research* [Online] **2016**, *6* (1), 50–64. https://www.researchgate.net/publication/309179824_Raman_Spectroscopy_a_review.
- Sandell, D.; Tougas, T. Quality Considerations in the Establishment of Specifications for Pharmaceuticals. *Statistics in Biopharmaceutical Research* **2012**, *4* (2), 125–135. DOI: 10.1198/sbr.2010.10333.
- Shaw, A. D.; Kaderbhai, N.; Jones, A.; Woodward, A. M.; Goodacre, R.; Rowland, J. J.; Kell, D. B. Noninvasive, On-Line Monitoring of the Biotransformation by Yeast of Glucose to Ethanol Using Dispersive Raman Spectroscopy and Chemometrics. *Appl Spectrosc* **1999**, *53* (11), 1419–1428. DOI: 10.1366/0003702991945777.
- Shipp, D. W.; Sinjab, F.; Notingher, I. Raman spectroscopy: techniques and applications in the life sciences. *Adv. Opt. Photon.* **2017**, *9* (2), 315. DOI: 10.1364/AOP.9.000315.
- Sivakesava, S.; Irudayaraj, J.; Demirci, A. Monitoring a bioprocess for ethanol production using FT-MIR and FT-Raman spectroscopy. *Journal of industrial microbiology & biotechnology* **2001**, *26* (4), 185–190. DOI: 10.1038/sj.jim.7000124.
- Smith, E.; Dent, G. *Modern Raman spectroscopy. A practical approach*, Second edition; Wiley: Hoboken, NJ, Chichester, West Sussex, UK, 2019.
- Stenlund, H. *Improving interpretation by orthogonal variation. Multivariate analysis of spectroscopic data*; Kemiska institutionen, Umeå universitet: Umeå, 2011.
- Stratton, J.; Chiruvolu, V.; Meagher, M. High cell-density fermentation. *Methods in molecular biology (Clifton, N.J.)* **1998**, *103*, 107–120. DOI: 10.1385/0-89603-421-6:107.
-

-
- tec5. Homepage. 2022. – URL: <https://www.tec5usa.com/product/tec5-probes/> (accessed October 17, 2022).
- Trygg, J.; Wold, S. Orthogonal projections to latent structures (O-PLS). *J. Chemometrics* **2002**, *16* (3), 119–128. DOI: 10.1002/cem.695.
- van Nederkassel Anne-Marie. *Guideline on the Use of Near Infrared Spectroscopy by the Pharmaceutical Industry and the Data Requirements for New Submissions and Variations.*, 2011.
- Vankeirsbilck, T.; Vercauteren, A.; Baeyens, W.; van der Weken, G.; Verpoort, F.; Vergote, G.; Remon, J. Applications of Raman spectroscopy in pharmaceutical analysis. *TrAC Trends in Analytical Chemistry* **2002**, *21* (12), 869–877. DOI: 10.1016/S0165-9936(02)01208-6.
- Voß, J.-P. *Anwendung spektroskopischer Messverfahren und multivariater Datenanalyse zur Bewertung und Beobachtung von Bioprozessen*, Als Manuskript gedruckt; Fortschrittberichte VDI : Reihe 17, Biotechnik, Medizintechnik Nr. 293; VDI Verlag: Düsseldorf, 2017.
- Voß, J.-P.; Mittelheuser, N. E.; Lemke, R.; Luttmann, R. Advanced monitoring and control of pharmaceutical production processes with *Pichia pastoris* by using Raman spectroscopy and multivariate calibration methods. *Engineering in life sciences* **2017**, *17* (12), 1281–1294. DOI: 10.1002/elsc.201600229.
- Wen, Z.-Q. Raman spectroscopy of protein pharmaceuticals. *Journal of pharmaceutical sciences* **2007**, *96* (11), 2861–2878. DOI: 10.1002/jps.20895.
- Whelan, J.; Craven, S.; Glennon, B. In situ Raman spectroscopy for simultaneous monitoring of multiple process parameters in mammalian cell culture bioreactors. *Biotechnology progress* **2012**, *28* (5), 1355–1362. DOI: 10.1002/btpr.1590.
- Wold, H. Nonlinear Iterative Partial Least Squares (NIPALS) Modelling: Some Current Developments. In *Multivariate analysis - III: Proceedings of the Third International Symposium on Multivariate Analysis held at Wright State University, Dayton, Ohio, June 19-24, 1972*; Kishnaiah, P. R., Ed.; Academic Press: New York [u.a.], 1973; pp 383–407. DOI: 10.1016/B978-0-12-426653-7.50032-6.
- Wold, S. Chemometrics; what do we mean with it, and what do we want from it? *Chemometrics and Intelligent Laboratory Systems* **1995**, *30* (1), 109–115. DOI: 10.1016/0169-7439(95)00042-9.
- Wold, S.; Antti, H.; Lindgren, F.; Öhman, J. Orthogonal signal correction of near-infrared spectra. *Chemometrics and Intelligent Laboratory Systems* [Online] **1998**, *44* (1-2), 175–185. <https://www.sciencedirect.com/science/article/pii/S0169743998001099>.
- Wold, S.; Esbensen, K.; Geladi, P. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems* **1987**, *2* (1-3), 37–52. DOI: 10.1016/0169-7439(87)80084-9.
- Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems* **2001**, *58* (2), 109–130. DOI: 10.1016/S0169-7439(01)00155-1.
- Zobeiri, H.; Hunter, N.; Xu, S.; Xie, Y.; Wang, X. Robust and high-sensitivity thermal probing at the nanoscale based on resonance Raman ratio (R3). *Int. J. Extrem. Manuf.* **2022**, *4* (3), 35201. DOI: 10.1088/2631-7990/ac6cb1.
-

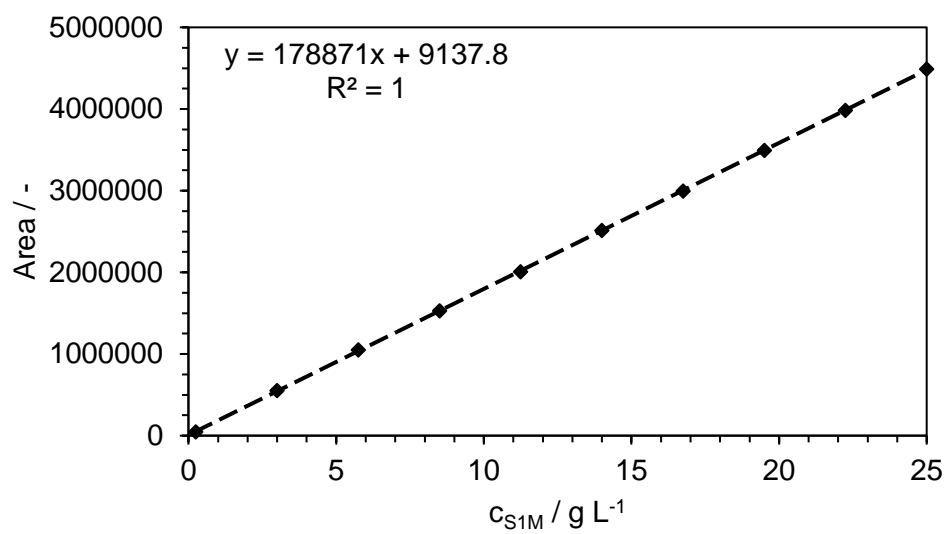
Appendix

A HPLC Calibration

Methanol Calibration



Glycerol Calibration



B MVDA

Prediction of Glycerol Concentration with off-line Suspension

No.	r	R2Y / -	Q2 / -	RMSEcv / g L ⁻¹	MBEcv / -	RMSEP / g L ⁻¹	MBEP / -
1	1+5+0	0.882	0.835	3.19	0.0247	3.05	0.585
2	1+5+0	0.884	0.727	4.09	0.0263	4.55	-0.704
3	1+5+0	0.780	0.648	4.65	0.0410	5.68	-0.364
4	1+5+0	0.886	0.732	4.05	0.0221	4.43	-0.761
5	1+4+0	0.748	0.642	4.69	0.121	5.17	-0.345
6	1+4+0	0.844	0.773	3.73	0.0309	4.19	0.542
7	1+6+0	0.889	0.790	3.59	-0.0607	3.68	0.148
8	1+5+0	0.895	0.846	3.07	0.0896	3.58	0.289
9	1+5+0	0.848	0.753	3.89	-0.0563	3.73	-0.0702
10	1+5+0	0.837	0.764	3.80	-0.107	3.18	-0.0598
11	1+5+0	0.819	0.755	3.88	-0.0622	3.31	-0.0853
12	1+4+0	0.871	0.837	3.16	0.0323	2.58	0.335

Prediction of Glycerol Concentration with off-line Supernatant

No.	r	R2Y / -	Q2 / -	RMSEcv / g L ⁻¹	MBEcv / -	RMSEP / g L ⁻¹	MBEP / -
1	1+5+0	0.865	0.745	3.31	0.0670	2.06	0.223
2	1+5+0	0.910	0.700	3.59	0.0295	3.42	-0.708
3	1+6+0	0.909	0.598	4.16	0.1193	5.07	-1.854
4	1+5+0	0.906	0.705	3.56	0.0198	3.34	-0.615
5	1+6+0	0.906	0.611	4.09	0.0631	4.91	-1.830
6	1+5+0	0.856	0.744	3.32	0.0493	2.55	0.172
7	1+5+0	0.940	0.795	2.97	0.0030	2.77	-0.841
8	1+5+0	0.867	0.710	3.53	0.0256	3.03	0.350
9	1+6+0	0.943	0.845	2.58	0.0578	2.53	-0.670
10	1+6+0	0.838	0.735	3.38	0.1903	2.99	-0.925
11	1+5+0	0.815	0.711	3.52	-0.0158	2.84	-1.056
12	1+4+0	0.756	0.638	3.95	0.1833	2.56	-0.184

Prediction of Glycerol Concentration with in-line Suspension

No.	r	R2Y / -	Q2 / -	RMSEcv / g L ⁻¹	MBEcv / -	RMSEP / g L ⁻¹	MBEP / -
1	1+4+0	0.755	0.707	4.11	0.0612	3.73	-0.625
2	1+3+0	0.821	0.740	3.87	-0.0311	3.21	-0.916
3	1+3+0	0.674	0.605	4.77	0.0447	4.38	-0.763
4	1+3+0	0.816	0.739	3.88	-0.0280	3.32	-0.908
5	1+3+0	0.682	0.603	4.78	0.0573	4.28	-0.707
6	1+2+0	0.692	0.674	4.33	0.0191	4.42	-1.064
7	1+2+0	0.685	0.661	4.42	0.0026	4.61	-1.705
8	1+2+0	0.690	0.673	4.34	0.0278	4.44	-1.087
9	1+3+0	0.695	0.666	4.39	0.0084	4.57	-1.518
10	1+2+0	0.650	0.628	4.63	0.0534	4.50	-1.427
11	1+3+0	0.676	0.644	4.53	0.1008	4.60	-1.246
12	1+3+0	0.714	0.698	4.17	-0.0069	4.18	-1.156

Prediction of Methanol with off-line Suspension

No.	r	R2Y / -	Q2 / -	RMSEcv / g L ⁻¹	MBEcv / -	RMSEP / g L ⁻¹	MBEP / -
1	1+1+0	0.356	0.313	0.99	2.28E-04	0.79	0.360
2	1+1+0	0.394	0.304	0.99	4.71E-03	0.83	0.132
3	1+3+0	0.777	0.455	0.88	-3.42E-02	1.06	0.156
4	1+1+0	0.392	0.302	1.00	4.63E-03	0.83	0.138
5	1+3+0	0.770	0.452	0.88	-3.65E-02	1.05	0.180
6	1+2+0	0.511	0.415	0.91	1.70E-02	0.69	0.263
7	1+3+0	0.541	0.390	0.93	6.44E-03	0.57	0.172
8	1+3+0	0.651	0.455	0.88	2.26E-02	0.76	0.162
9	1+3+0	0.797	0.491	0.85	7.35E-02	0.93	0.225
10	1+0+0	0.188	0.161	1.09	1.14E-04	0.96	0.301
11	1+0+0	0.183	0.160	1.09	9.09E-05	0.96	0.306
12	1+1+0	0.352	0.315	0.99	2.16E-05	0.80	0.366

Prediction of Methanol Concentration with off-line Supernatant

No.	r	R2Y / -	Q2 / -	RMSEcv / g L ⁻¹	MBEcv / -	RMSEP / g L ⁻¹	MBEP / -
1	1+1+0	0.346	0.110	0.84	-2.36E-02	0.83	0.574
2	1+2+0	0.426	0.177	0.81	-8.43E-03	0.60	-0.058
3	1+1+0	0.261	0.078	0.86	5.02E-03	0.70	0.079
4	1+2+0	0.427	0.163	0.82	-1.80E-02	0.59	-0.040
6	1+1+0	0.331	0.098	0.85	-1.94E-02	0.83	0.590
7	1+1+0	0.382	0.197	0.80	-1.56E-02	0.76	0.445
8	1+1+0	0.321	0.108	0.84	-1.97E-02	0.84	0.594
9	1+1+0	0.355	0.188	0.80	-1.18E-02	0.78	0.479
10	1+1+0	0.390	0.230	0.78	-6.46E-02	0.73	0.355
11	1+0+0	0.205	0.122	0.84	1.07E-02	0.82	0.587
12	1+1+0	0.341	0.118	0.84	-2.49E-02	0.83	0.574
0	0	0.000	0.000	0.00	0.00E+00	0.00	0.000

Prediction of Methanol Concentration with in-line Suspension

No.	r	R2Y / -	Q2 / -	RMSEcv / g L ⁻¹	MBEcv / -	RMSEP / g L ⁻¹	MBEP / -
1	1+1+0	0.142	0.093	0.83	-6.59E-03	0.70	0.279
2	1+3+0	0.603	0.348	0.70	-6.09E-03	0.38	0.077
3	1+2+0	0.401	0.141	0.81	4.86E-03	0.48	0.121
4	1+2+0	0.533	0.242	0.76	1.16E-02	0.44	0.040
5	1+3+0	0.499	0.148	0.81	4.44E-03	0.48	0.018
6	1+0+0	0.143	0.112	0.82	-2.93E-03	0.65	0.324
7	1+2+0	0.396	0.310	0.72	1.63E-02	0.45	0.028
8	1+1+0	0.179	0.124	0.82	-8.67E-03	0.71	0.227
9	1+1+0	0.351	0.265	0.75	3.54E-03	0.56	0.029
10	1+0+0	0.262	0.232	0.76	-4.30E-03	0.56	0.240
11	1+0+0	0.257	0.237	0.76	-2.29E-03	0.55	0.240
12	1+1+0	0.139	0.095	0.83	-6.53E-03	0.70	0.277

Prediction of Cell Concentration with off-line Suspension

No.	r	R2Y / -	Q2 / -	RMSEcv / g L ⁻¹	MBEcv / -	RMSEP / g L ⁻¹	MBEP / -
1	1+1+0	0.803	0.799	7.78	0.0211	10.88	-1.017
2	1+5+0	0.896	0.847	6.80	-0.0703	12.20	-2.212
3	1+5+0	0.786	0.665	10.06	-0.0116	11.85	1.310
4	1+5+0	0.898	0.856	6.59	-0.0754	12.36	-2.019
5	1+4+0	0.749	0.649	10.29	-0.0755	12.61	0.495
6	1+6+0	0.932	0.906	5.33	-0.0309	5.58	-0.122
7	1+1+0	0.710	0.703	9.47	0.0470	10.83	0.077
8	1+6+0	0.931	0.909	5.24	-0.0270	4.87	-0.110
9	1+5+0	0.860	0.839	6.96	0.1397	6.48	0.435
10	1+5+0	0.902	0.887	5.84	0.1701	4.86	-0.289
11	1+5+0	0.899	0.888	5.81	0.0497	5.03	-0.431
12	1+5+0	0.925	0.911	5.18	-0.0067	7.15	-0.888

Prediction of Cell Concentration with in-line Suspension

No.	r	R2Y / -	Q2 / -	RMSEcv / g L ⁻¹	MBEcv / -	RMSEP / g L ⁻¹	MBEP / -
1	1+6+0	0.927	0.894	6.21	0.1745	10.33	3.869
2	1+3+0	0.884	0.838	7.66	-0.0560	9.31	3.395
3	1+4+0	0.711	0.634	11.53	-0.4063	12.47	0.366
4	1+3+0	0.884	0.846	7.49	-0.0432	9.29	3.546
5	1+4+0	0.715	0.643	11.39	-0.3548	12.36	0.721
6	1+5+0	0.868	0.832	7.81	0.2057	10.48	3.874
7	1+7+0	0.928	0.885	6.46	0.1967	10.51	3.733
8	1+6+0	0.929	0.896	6.16	0.1637	9.48	3.855
9	1+6+0	0.923	0.882	6.55	0.0317	11.03	3.656
10	1+6+0	0.902	0.881	6.57	0.1767	9.88	1.697
11	1+6+0	0.913	0.896	6.14	0.0906	9.70	2.615
12	1+4+0	0.840	0.811	8.29	0.1977	11.33	3.785

Prediction of Total Protein Concentration with off-line Suspension

No.	r	R2Y / -	Q2 / -	RMSEcv / g L ⁻¹	MBEcv / -	RMSEP / g L ⁻¹	MBEP / -
1	1+2+0	0.774	0.715	10.56	-0.3291	14.57	-5.212
2	1+3+0	0.736	0.570	12.96	-0.1930	15.43	-5.655
3	1+2+0	0.331	0.170	18.02	-0.0819	23.28	-0.271
4	1+3+0	0.748	0.599	12.53	-0.1662	14.99	-5.519
5	1+1+0	0.243	0.180	17.91	-0.1512	26.22	0.313
6	1+3+0	0.755	0.683	11.13	-0.3156	14.93	-4.649
7	1+3+0	0.673	0.583	12.77	0.2053	17.26	-3.303
8	1+3+0	0.756	0.681	11.18	-0.3418	14.82	-5.027
9	1+3+0	0.686	0.589	12.68	-0.1433	17.31	-3.035
10	1+4+0	0.784	0.740	10.09	-0.0715	14.04	-4.257
11	1+4+0	0.782	0.741	10.06	0.0195	14.27	-3.281
12	1+1+0	0.735	0.702	10.79	-0.1819	15.76	-4.625

Prediction of Total Protein Concentration with off-line Supernatant

No.	r	R2Y / -	Q2 / -	RMSEcv / g L ⁻¹	MBEcv / -	RMSEP / g L ⁻¹	MBEP / -
1	1+3+0	0.739	0.642	13.82	0.0124	17.48	-1.300
2	1+1+0	0.314	0.179	20.93	-0.4330	27.00	-0.589
3	1+0+0	0.087	0.049	22.52	-0.2951	28.98	-5.136
4	1+1+0	0.315	0.178	20.94	-0.4087	27.06	-0.678
5	1+3+0	0.704	0.275	19.67	-0.7451	19.09	-1.702
6	1+3+0	0.688	0.589	14.80	0.4721	20.13	-2.499
7	1+1+0	0.603	0.537	15.72	-0.1956	20.68	-4.155
8	1+2+0	0.628	0.514	16.09	0.5408	21.36	-1.773
9	1+3+0	0.787	0.634	13.97	-1.0016	14.20	-3.883
10	1+3+0	0.805	0.728	12.05	-0.4062	13.82	-1.851
11	1+3+0	0.803	0.738	11.81	-0.3382	14.38	-2.616
12	1+2+0	0.675	0.588	14.83	0.1562	19.65	-0.810

Prediction of Total Protein Concentration with in-line Suspension

No.	r	R2Y / -	Q2 / -	RMSEcv / g L ⁻¹	MBEcv / -	RMSEP / g L ⁻¹	MBEP / -
1	1+4+0	0.786	0.744	12.39	0.0726	9.73	0.547
2	0+0+0	0.000	0.000	0.00	0.0000	0.00	0.000
3	0+0+0	0.000	0.000	0.00	0.0000	0.00	0.000
4	0+0+0	0.000	0.000	0.00	0.0000	0.00	0.000
5	0+0+0	0.000	0.000	0.00	0.0000	0.00	0.000
6	1+4+0	0.779	0.734	12.63	0.0537	10.54	0.075
7	1+5+0	0.887	0.831	10.06	-0.0948	10.88	-0.250
8	1+4+0	0.779	0.733	12.65	0.0926	10.44	0.845
9	1+5+0	0.887	0.831	10.06	-0.0438	10.99	0.337
10	1+3+0	0.804	0.781	11.45	-0.2364	12.30	3.266
11	1+5+0	0.879	0.836	9.93	0.0248	10.23	2.849
12	1+3+0	0.749	0.722	12.90	0.0364	10.29	1.921

C Parameter Estimation

1.1) MAIN-FILE

```
% Goal:          Determination of correction factor for Sturb
% Description:   Simple regression with NSt consideration
% Author:       Phoebe Chan
% Date:         18.08.2022
% Version:      4.0
% Related files: funSturbNSt.m

clc, clear
clf

% Data point of CDW mustn't be at t0 = 0 h but must be a number > 0!
dtpth = '3 Forschungsdaten\XXPC0922_FBII\Cultivation Data\';
CDW    = dlmread([dtpth 'BTM_part.txt'],'\t',[1 0 22 1]);
z      = CDW(:,2);
Sturb  = dlmread([dtpth 'Sturb_processed.txt'],'\t',[1 0 11847 1]);
ty     = Sturb(1:3661,1); % Data points from batch and fed-batch phase
y      = Sturb(1:3661,2);
NSt    = dlmread([dtpth 'NSt_processed.txt'],'\t',[1 0 1101 1]);
tx     = NSt(1:584,1);
x      = NSt(1:584,2);

% plot(t,y,'ko') % first, plot data to have an idea about initial values for param-
eters

% Interpolation of Sturb and NSt for equidistant time points
Sturbi = interp1(ty,y,CDW(:,1),'linear');
NSti    = interp1(tx,x,CDW(:,1),'linear');

% Optimisation and regression function
F_opt = @(c) fun_SturbNSt(Sturbi,NSti,CDW(:,2),c);
C0 = [0.03 3.5 5.1]; % Enter initial guessing values for parameters
C_opt = fminsearch(F_opt,C0);

% Regression function:  $c_{XLturb} = a \cdot (e^{(b \cdot S_{turb})} - 1) \cdot (N_{St}/N_{Stmax})^c$ 
F_reg = @(C_opt,y,x) C_opt(1) .* (exp(C_opt(2) .* y) - 1) .* (x./800).^C_opt(3);
Sturbs = (0:0.14:3)'; % smooth interval of Sturb for plotting
NSts = (300:23:800)'; % smooth interval of NSt for plotting
cXLturb = F_reg(C_opt,Sturbs,NSts);

figure(1)
hold on
plot(Sturbs,cXLturb,'r','LineWidth',1)
hold off
xlim([0 2.5])
ylim([0 30])
box on

% Legend formatting
xlgnd = Sturbs(1)+0.2; % offset in x-dir. for legend
ylgnd = max(CDW(:,2))-3; % offset in y-dir. for legend, can be alternated
text(xlgnd,ylgnd,{'c_{XLturb} = a*(e^{b*S_{turb}}(t) - 1)*(N_{St}/N_{St,max})^c',...
    'with', ['a = ', num2str(C_opt(1)), ' g L^{-1}'],...
    ['b = ', num2str(C_opt(2)), ' AU^{-1}'], ['c = ', num2str(C_opt(3))]])

% Title formatting
% Enter here Cultivation No. (and cultivation phase):
CultNo = XXPC0922;
xlabel('S_{turb} / AU')
ylabel('c_{XL} / g L^{-1}')
title(['Nelder-Mead Algorithm for Determination of S_{turb} Parameters'],...
    ['Cultivation No.: ', num2str(CultNo)])

% Set ticks outside
set(gca,'TickDir','out'); % The only other option is 'in'
```

```

% Interpolation of Sturb_whole and NSt_whole for equidistant time points for Sa-
vitzky-Golay filter
ti_turb = (min(Sturb(:,1)):0.478:max(Sturb(:,1)))';
ti_NSt = (min(NSt(:,1)):0.476:max(NSt(:,1)))';
Sturbwi = interp1(Sturb(:,1),Sturb(:,2),ti_turb,'linear');
NStwi = interp1(NSt(:,1),NSt(:,2),ti_NSt,'linear');

% Set smooth time interval for plot 2
F_regwhole = @(C_opt,Sturb,NSt) C_opt(1).*(exp(C_opt(2).*Sturb)-
1).*(NSt./800).^C_opt(3);
W = F_regwhole(C_opt,Sturbwi,NStwi);

figure(2)
clf
hold on
plot(CDW(:,1),CDW(:,2),'ko')
plot(ti_turb,W,'r','LineWidth',1)
hold off
box on
xlim([0 80])
ylim([0 50])

% Title formatting
xlabel('t_{process} / h')
ylabel('c_{XL} / g L^{-1}')

% Set ticks outside
set(gca,'TickDir','out'); % The only other option is 'in'

```

1.2) FUNCTION-FILE

```

function S = fun_SturbNSt(X,Y,Z,c)
%Revised FUNCTION FILE with consideration of Stirrer speed
%Version 1
fun = @(Sturbi,NSti) c(1).*(exp(c(2).*Sturbi)-1).*(NSti./800).^c(3);
x = X;
y = Y;
z = Z;
plot(x,z,'ko')
S = sum((fun(x,y)-z).^2); % Sum of squared residuals
SStotal = (length(fun(x,y))-1)*var(fun(x,y)); % Sum of S
% Calculate Goodness of Fit R^2
Rsq = 1 - S/SStotal;
text(0.2,max(Z)+2.5,{'R^2 = ',num2str(Rsq)})
end

```