

MACHINE LEARNING IN SPECTROSCOPY
ROBUST PROCESS MONITORING FOR BIOPHARMACEUTICAL DOWNSTREAM PROCESSING

Zur Erlangung des akademischen Grades eines
DOKTORS DER INGENIEURWISSENSCHAFTEN (Dr.-Ing.)

von der KIT-Fakultät für Chemieingenieurwesen und Verfahrenstechnik des
Karlsruher Instituts für Technologie (KIT)
genehmigte

DISSERTATION

von
M. Sc. Robin Schiemer
aus Dieburg

Erstgutachter:	Prof. Dr. Jürgen Hubbuch
Zweitgutachterin:	Prof. Dr. Gisela Guthausen
Tag der mündlichen Prüfung:	21.05.2026



This document, excluding Chapter 6 - A novel Raman-chromatography assembly for automated calibration and in-line monitoring in bioprocessing, is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0): <https://creativecommons.org/licenses/by/4.0/deed.en>.

Chapter 6 was previously published under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0) and is not covered by the CC BY 4.0 license granted for the rest of this document.

Abstract

The work presented in this thesis contributes to applied spectroscopy research by tailoring machine learning (ML) towards chemometric applications. Spectroscopic techniques have long served as foundational tools for chemical and biological molecular analysis and have emerged as powerful non-invasive, real-time information sources for biopharmaceutical process monitoring. Biopharmaceuticals now dominate the pharmaceutical market, with 8 of the 10 top-selling products in 2024 being biopharmaceuticals, addressing diseases such as cancer, inflammatory diseases, diabetes, and viral infections. Monoclonal antibodies (mAbs), antibody-drug conjugates (ADCs), and virus-like particles (VLPs) represent key product classes for targeted therapies and gene delivery applications. Due to their structural diversity, most biopharmaceuticals are recombinantly produced and undergo extensive downstream processing (DSP) to remove process- and product-related impurities. Manufacturing processes must consistently deliver products meeting stringent quality specifications, making process development experimentally demanding. The implementation of process analytical technology (PAT) promises direct measurement of product compositions throughout processing, thereby enhancing process understanding and enabling real-time control.

Translating spectral data into actionable process information requires chemometric models capable of handling the inherent complexity of bioprocess matrices while providing sufficient selectivity to distinguish between analytes of interest. Traditional partial least squares (PLS)-based approaches encounter fundamental limitations when confronted with non-linear relationships between spectral features and analyte concentrations, spectral similarities between co-existing species, and complex matrices exhibiting variable background contributions. Modern ML methods present opportunities through hierarchical feature extraction, yet data scarcity and labor-intensive model optimization have impeded their adoption in spectroscopic calibration. These challenges motivated the objective of this thesis: advancing chemometric model development through ML methods that address selectivity and data availability limitations, combined with systematic frameworks for preprocessing, calibration, and validation.

Selectivity limitations become particularly acute when monitoring reactions where spectrally indistinguishable species coexist. ADC conjugation exemplifies this challenge, as absorption spectra of unconjugated antibody and conjugated products overlap extensively. An adaptive soft-sensor was constructed that fuses ultraviolet/visible (UV/Vis)-derived predictions from a Gaussian process regression (GPR) model with forecasts from a kinetic reaction model through an extended Kalman filter (EKF). Time-variant uncertainty quantification in both sensor and process components enabled dynamic weighting between spectroscopic observations and mechanistic predictions as their

respective accuracies evolved during the reaction. The GPR model outperformed PLS by 18% for conjugated drug quantification, while the EKF framework decreased prediction errors by up to 17% relative to the kinetic model alone. Robustness evaluations confirmed effective compensation for sensor noise, initialization errors, and operational mode transitions between batch and fed-batch configurations.

ML methods have gained attention as alternatives to traditional chemometric approaches due to their capacity to model non-linear relationships and extract complex spectral features without explicit preprocessing. Convolutional neural networks (CNNs), originally developed for image recognition, offer a fundamentally different approach by learning hierarchical feature representations directly from raw data. Their convolutional layers act as adaptive filters that can identify local spectral patterns, while pooling operations provide robustness against minor wavelength shifts. This architecture is particularly well-suited for spectroscopic data, where relevant chemical information is encoded in localized band shapes and peak positions rather than global intensity patterns. However, a recurring obstacle in deploying CNNs for spectroscopic applications is the mismatch between the data requirements of deep learning and the limited experimental datasets typically available. A data augmentation strategy was developed that preserves chemical realism by estimating pure component spectra locally within concentration neighborhoods and recombining these estimates to generate synthetic mixtures spanning the experimental space. Coupled with Bayesian hyperparameter optimization (HPO), this approach identified optimal CNN configurations featuring three convolutional layers with dropout and regularization. The resulting models achieved up to 50% lower root mean squared error (RMSE) than PLS regression for mAb size variant quantification. Interpretability assessment through Shapley additive explanations (SHAP) values confirmed the extraction of chemically meaningful features, and enhanced robustness was demonstrated against spectral noise and wavelength perturbations. The optimized architecture generalized successfully across UV/Vis and infrared (IR) datasets without re-optimization, demonstrating transferability beyond the original training domain.

While UV/Vis spectroscopy captures electronic transitions yielding broad absorption bands, vibrational spectroscopy techniques such as Raman and IR spectroscopy probe molecular vibrations and provide a chemical fingerprint. Raman spectroscopy offers particular advantages for bioprocess monitoring such as compatibility with aqueous environments, minimal sample preparation, and sensitivity to molecular structure including protein secondary conformation. However, the transition from UV/Vis to Raman introduces distinct challenges. Raman scattering is inherently weak, requiring longer acquisition times or higher laser powers. Spectra are susceptible to fluorescence interference and baseline distortions. In complex bioprocess matrices, these challenges are compounded by variable background contributions and overlapping spectral features from multiple components. Monitoring VLP precipitation from crude *E. coli* lysates exemplifies these difficulties, as particulate-induced scatter, precipitant signals, and feedstock components dominate observed spectra. A systematic methodology was pursued involving data diversification through conditioned lysate experiments with spiked components across batch and fed-batch modes. Screening of preprocessing combinations identified an effective pipeline incorporating turbidity normalization, asymmetric least squares baseline correction, orthogonal projection to latent structures (OPLS)-based background removal, and Savitzky–Golay (SG) derivative filtering. The resulting PLS models achieved R^2 values exceeding 0.97 for precipitant quantification and 0.64–0.74 for VLP concentration in turbid suspensions. Spectral band ratio analysis provided mechanistic insight into selective precipitation behavior.

The practical deployment of PAT is frequently constrained by calibration bottlenecks, as manual sample preparation and measurement demand substantial time and introduce variability. An automated platform was engineered by integrating a chromatography system with a Raman spectrometer, enabling high-throughput calibration data acquisition with minimal operator intervention. The platform was evaluated in a case study targeting simultaneous monitoring of lysozyme, phosphate, and citrate during ultra-/diafiltration (UF/DF) processing. Calibration datasets were generated using statistical experimental designs to efficiently sample the multi-component concentration space, and the resulting PLS models were subsequently applied to in-line spectral data acquired during UF/DF processes. Systematic spectral discrepancies between off-line calibration and in-line process measurements degraded prediction accuracy, motivating the introduction of a direct standardization approach using synthetic spectra generated through the previously developed augmentation method. This enabled accurate tracking of protein concentration dynamics alongside buffer exchange progression throughout the UF/DF operation.

Protein aggregation represents a critical quality attribute in biopharmaceutical manufacturing, as aggregates can compromise product efficacy and elicit immunogenic responses. Vibrational spectroscopy techniques, particularly Raman and IR spectroscopy, are sensitive to protein secondary structure through characteristic amide band signatures and have been extensively investigated for aggregate detection. Numerous studies have demonstrated spectral changes associated with conformational transitions during aggregation under controlled laboratory conditions, including shifts in amide I band position and alterations in secondary structure composition. However, these investigations were predominantly conducted using purified protein solutions or forced degradation samples, while the practical deployment of such analytical methods during actual bioprocessing remains largely unexplored. Distinguishing protein aggregates from monomers during chromatographic separation poses additional challenges, as subtle conformational differences produce weak spectral signatures that must be detected against variable process backgrounds. Using bovine serum albumin (BSA) as a model protein, systematic validation through stress experiments, orthogonal mixing studies, and in-line anion-exchange chromatography monitoring established the amide I band as a robust structural marker for aggregate co-elution under realistic process conditions. A CNN architecture consistent with earlier optimization results outperformed PLS variants across gradient and step elution conditions with varying loading densities and gradient lengths. Signal limitations at low protein concentrations identified boundaries for practical implementation.

This thesis advances chemometric model development by integrating ML methods with systematic experimental and computational frameworks tailored to spectroscopic process monitoring. The data augmentation methodology addresses deep learning data requirements while maintaining chemical validity, with demonstrated utility in model transfer applications. Mechanistic knowledge integration through sensor fusion extends monitoring capability beyond inherent spectroscopic selectivity. Systematic preprocessing optimization provides solutions for challenging matrices involving turbidity and environmental interference. Automated calibration platforms reduce implementation barriers while enabling robust model transfer. These methodological contributions are transferable across biopharmaceutical products and processes, establishing foundations for robust, interpretable, and automated spectroscopic PAT in biopharmaceutical manufacturing.

Zusammenfassung

Die vorliegende Arbeit leistet einen Beitrag zur angewandten Spektroskopieforschung durch die Anpassung von Machine Learning (ML) für chemometrische Anwendungen. Spektroskopische Techniken bilden seit Langem die Grundlage für die chemische und biologische Molekülanalytik und haben sich als leistungsstarke, nicht-invasive Echtzeitinformationsquellen für die biopharmazeutische Prozessüberwachung etabliert. Biopharmazeutika dominieren mittlerweile den Pharmamarkt, wobei 8 der 10 meistverkauften Produkte im Jahr 2024 zu dieser Klasse zählen und unter anderem zur Behandlung von Krebs, entzündlichen Erkrankungen, Diabetes und Virusinfektionen eingesetzt werden. Monoklonale Antikörper (mAbs), Antikörper-Wirkstoff-Konjugate (ADCs) und virusähnliche Partikel (VLPs) stellen wichtige Produktklassen für zielgerichtete Therapien und Gentherapie-Anwendungen dar. Aufgrund ihrer strukturellen Vielfalt werden die meisten Biopharmazeutika rekombinant hergestellt und durchlaufen eine umfangreiche Aufreinigung (Downstream Processing, DSP), um prozess- und produktbezogene Verunreinigungen zu entfernen. Herstellungsprozesse müssen konsistent Produkte liefern, die strenge Qualitätsspezifikationen erfüllen, was die Prozessentwicklung experimentell anspruchsvoll gestaltet. Die Implementierung von Process Analytical Technology (PAT) ermöglicht die direkte Messung von Probenzusammensetzungen während der gesamten Verarbeitung und verbessert dadurch das Prozessverständnis bei gleichzeitiger Ermöglichung von Echtzeitregelung.

Die Übersetzung spektraler Daten in verwertbare Prozessinformationen erfordert chemometrische Modelle, die in der Lage sind, die inhärente Komplexität von Bioprozessmatrizen zu bewältigen und gleichzeitig ausreichende Selektivität zur Unterscheidung zwischen verschiedenen Analyten zu bieten. Traditionelle Ansätze basierend auf Partial Least Squares (PLS) Regression stoßen an grundlegende Grenzen, wenn sie mit nicht-linearen Beziehungen zwischen spektralen Merkmalen und Analytkonzentrationen, spektralen Ähnlichkeiten zwischen koexistierenden Spezies sowie komplexen Matrizen mit variablen Hintergrundbeiträgen konfrontiert werden. Moderne ML-Methoden bieten durch hierarchische Merkmalsextraktion neue Möglichkeiten, jedoch haben Datenknappheit und arbeitsintensive Modelloptimierung ihre Anwendung in der spektroskopischen Kalibration bisher erschwert. Diese Herausforderungen motivierten das Ziel dieser Arbeit: die Weiterentwicklung chemometrischer Modellentwicklung durch ML-Methoden, die Selektivitäts- und Datenverfügbarkeitslimitierungen adressieren, kombiniert mit Ansätzen zur systematischen Optimierung von Datenverarbeitung, Modellkalibrierung und -validierung.

Selektivitätslimitierungen werden besonders kritisch bei der Überwachung von Reaktionen, in denen spektral nicht unterscheidbare Spezies koexistieren. Die ADC-Konjugation veranschaulicht diese

Herausforderung, da sich die Absorptionsspektren von unkonjugiertem Antikörper und konjugierten Produkten weitgehend überlappen. Ein adaptiver Soft-Sensor wurde entwickelt, der UV/Vis-basierte Vorhersagen eines Gaussian Process Regression (GPR) Modells mit Prognosen eines kinetischen Reaktionsmodells mittels eines Extended Kalman Filters (EKF) fusioniert. Zeitvariante Unsicherheitsquantifizierung in beiden Komponenten ermöglichte eine dynamische Gewichtung zwischen spektroskopischen Beobachtungen und mechanistischen Vorhersagen entsprechend ihrer jeweiligen Genauigkeiten während der Reaktion. Das GPR-Modell übertraf das PLS-Modell um 18% bei der Quantifizierung des konjugierten Wirkstoffs, während das EKF-Framework die Vorhersagefehler um bis zu 17% gegenüber dem kinetischen Modell allein reduzierte. Robustheitsevaluierungen bestätigten eine effektive Kompensation von Sensorrauschen, Initialisierungsfehlern und Betriebsmodusübergängen zwischen Batch- und Fed-Batch-Konfigurationen.

ML Methoden haben als Alternativen zu traditionellen chemometrischen Ansätzen Aufmerksamkeit gewonnen, da sie in der Lage sind, nicht-lineare Beziehungen zu modellieren und komplexe spektrale Merkmale ohne explizite Vorbearbeitung zu extrahieren. Convolutional Neural Networks (CNNs), ursprünglich für die Bilderkennung entwickelt, bieten einen grundlegend anderen Ansatz, indem sie hierarchische Merkmalsrepräsentationen direkt aus Rohdaten erlernen. Ihre Faltungsschichten fungieren als adaptive Filter, die lokale spektrale Muster identifizieren können, während Pooling-Operationen Robustheit gegenüber geringfügigen Wellenlängenverschiebungen gewährleisten. Diese Architektur ist besonders gut für spektroskopische Daten geeignet, bei denen relevante chemische Informationen in lokalisierten Bandenformen und Peakpositionen kodiert sind, anstatt in globalen Intensitätsmustern. Ein wiederkehrendes Hindernis beim Einsatz von CNNs für spektroskopische Anwendungen ist jedoch die Diskrepanz zwischen den Datenanforderungen von Deep Learning und den typischerweise verfügbaren Datensätzen, die in ihrer Größe limitiert sind. Eine Datenaugmentierungsstrategie wurde entwickelt, die die physicochemische Validität der synthetischen Spektren bewahrt, indem Reinkomponentenspektren aus unterschiedlichen Konzentrationsbereichen geschätzt und diese Schätzungen zu synthetischen Mischungen über den experimentellen Raum rekombiniert werden. In Kombination mit Bayes'scher Hyperparameteroptimierung (HPO) identifizierte dieser Ansatz optimale CNN-Konfigurationen mit drei Faltungsschichten sowie verschiedenen Regularisierungsmechanismen. Die resultierenden Modelle erzielten bis zu 50% niedrigere Vorhersagefehler (Root Mean Square Error, RMSE) als PLS Regression bei der Quantifizierung von Größenvarianten von mAbs. Interpretierbarkeitsanalysen mittels Shapley Additive Explanations (SHAP) bestätigten die Extraktion chemisch aussagekräftiger Merkmale, und erhöhte Robustheit wurde gegenüber spektralem Rauschen und Wellenlängenperturbationen nachgewiesen. Die optimierte Architektur generalisierte erfolgreich über UV/Vis- und Infrarot (IR)-Datensätze ohne erneute Optimierung, was die Übertragbarkeit über die ursprüngliche Trainingsdomäne hinaus demonstriert.

Während UV/Vis-Spektroskopie elektronische Übergänge erfasst und breite Absorptionsbanden liefert, untersuchen Raman- und IR-Spektroskopie molekulare Schwingungen und liefern mehr Informationen über die chemische Zusammensetzung. Raman-Spektroskopie bietet besondere Vorteile für die Bioprozessüberwachung wie Kompatibilität mit wässrigen Umgebungen, minimale Probenvorbereitung und Empfindlichkeit gegenüber molekularen Strukturen einschließlich der Proteinsekundärstruktur. Der Wechsel von UV/Vis zu Raman Spektroskopie bringt jedoch spezifische Herausforderungen mit sich. Raman-Streuung ist inhärent schwach und erfordert längere Messzeiten oder höhere Laserleistungen. Spektren sind anfällig für Fluoreszenzinterferenzen und Basislinienverzerrungen. In komplexen Bioprozessmatrizen werden diese Herausforderungen durch variable Hintergrundbeiträge und überlappende spektrale Merkmale mehrerer Komponenten verstärkt. Die Überwachung der VLP-Präzipitation aus unverarbeiteten *E. coli*-Lysaten verdeutlicht diese

Schwierigkeiten, da partikelinduzierte Streuung, Präzipitanzsignale und Feedstock-Komponenten die beobachteten Spektren dominieren. Eine systematische Methodik wurde verfolgt, die Datendiversifizierung durch konditionierte Lysat-Experimente mit zugesetzten Komponenten über Batch- und Fed-Batch-Modi umfasste. Das Screening von Preprocessing-Kombinationen identifizierte eine effektive Pipeline bestehend aus Trübungsnormalisierung, asymmetrischer Least-Squares-Basislinienkorrektur, Orthogonal Projection to Latent Structures (OPLS)-basierter Hintergrundentfernung und Savitzky-Golay (SG)-Ableitungsfilerung. Die resultierenden PLS-Modelle erreichten R^2 -Werte über 0,97 für die Präzipitanzquantifizierung und 0,64–0,74 für die VLP-Konzentration in trüben Suspensionen. Die Analyse spektraler Bandenverhältnisse lieferte mechanistische Einblicke in das selektive Präzipitationsverhalten.

Der praktische Einsatz von PAT wird häufig durch Kalibrierungsengpässe eingeschränkt, da manuelle Probenvorbereitung und -messung erhebliche Zeit erfordern und Variabilität herbeiführen. Eine automatisierte Plattform wurde durch Integration eines Chromatographiesystems mit einem Raman-Spektrometer entwickelt, die eine Hochdurchsatz-Kalibrierdatenerfassung mit minimalem Bedienungsaufwand ermöglicht. Die Plattform wurde in einer Fallstudie evaluiert, die auf die simultane Überwachung von Lysozym, Phosphat und Citrat während der Ultrafiltration/Diafiltration (UF/DF) abzielte. Kalibrierdatensätze wurden mittels statistischer Versuchsplanung generiert, um den Mehrkomponenten-Konzentrationsraum effizient abzudecken, und die resultierenden PLS-Modelle wurden anschließend auf In-line-Spektraldaten angewandt, die während der UF/DF-Prozesse erfasst wurden. Systematische spektrale Diskrepanzen zwischen Off-line-Kalibrierung und In-line-Prozessmessungen verschlechterten die Vorhersagegenauigkeit, was die Einführung eines Standardisierungsansatzes unter Verwendung synthetischer Spektren aus der zuvor entwickelten Augmentierungsmethode motivierte. Dies ermöglichte eine genaue Verfolgung der Proteinkonzentrationsdynamik sowie des Pufferaustauschfortschritts während des gesamten UF/DF-Betriebs.

Proteinaggregation stellt ein kritisches Qualitätsattribut in der biopharmazeutischen Herstellung dar, da Aggregate die Produktwirksamkeit beeinträchtigen und immunogene Reaktionen hervorrufen können. Raman- und IR-Spektroskopie sind empfindlich gegenüber der Proteinsekundärstruktur durch charakteristische Amid-Bandensignaturen und wurden umfassend für die Aggregatdetektion untersucht. Zahlreiche Studien haben spektrale Veränderungen im Zusammenhang mit Konformationsübergängen während der Aggregation unter kontrollierten Laborbedingungen nachgewiesen, einschließlich Verschiebungen der Amid-I-Bandenposition und Änderungen der Sekundärstrukturzusammensetzung. Diese Untersuchungen wurden jedoch überwiegend mit aufgereinigten Proteinlösungen oder Forced-Degradation-Proben durchgeführt, sodass der praktische Einsatz solcher analytischen Methoden während tatsächlicher Bioprozesse weitgehend unerforscht blieb. Die Unterscheidung von Proteinaggregaten und Monomeren während der chromatographischen Trennung stellt zusätzliche Herausforderungen dar, da subtile Konformationsunterschiede schwache spektrale Signaturen erzeugen, die gegen variable Prozesshintergründe detektiert werden müssen. Unter Verwendung von Albumin aus Rinderserum als Modellprotein etablierte eine systematische Validierung durch Stressexperimente, orthogonale Mischstudien und In-line-Anionenaustauschchromatographie-Überwachung der Amid-I Bande als robusten strukturellen Marker für die Aggregat-Ko-Elution unter realistischen Prozessbedingungen. Eine CNN-Architektur konsistent mit früheren Optimierungsergebnissen übertraf PLS-Varianten über Gradienten- und Stufenelutionsbedingungen mit variierenden Beladungsdichten und Gradientenlängen. Signallimitierungen bei niedrigen Proteinkonzentrationen identifizierten Grenzen für die praktische Implementierung.

Diese Arbeit entwickelt die chemometrische Modellentwicklung weiter durch Integration von ML-Methoden mit experimentellen und rechnerischen Ansätzen, die auf die spektroskopische Prozessüber-

wachung zugeschnitten sind. Die Datenaugmentierungsmethodik adressiert die Datenanforderungen von Deep Learning bei gleichzeitiger Wahrung physicochemischer Validität, mit nachgewiesener Anwendbarkeit in Modelltransfer-Anwendungen. Die Integration mechanistischen Wissens durch Sensorfusion erweitert die Überwachungsfähigkeit über die inhärente spektroskopische Selektivität hinaus. Systematische Vorverarbeitungsoptimierung bietet Lösungen für herausfordernde Matrizen mit Trübung und Umgebungsinterferenzen. Automatisierte Kalibrierungsplattformen reduzieren Implementierungsbarrieren und ermöglichen einen robusten Modelltransfer. Diese methodischen Beiträge sind über biopharmazeutische Produkte und Prozesse hinweg übertragbar und etablieren Grundlagen für robuste, interpretierbare und automatisierte spektroskopische PAT in der biopharmazeutischen Herstellung.

Contents

Abstract	i
Zusammenfassung	v
Contents	ix
1 Introduction	1
1.1 Process Analytical Technology (PAT)	2
1.2 PAT Applications in Biopharmaceutical Downstream Processing	3
1.2.1 Monoclonal Antibodies	4
1.2.2 Virus-Like Particles	5
1.2.3 Antibody-Drug Conjugates	6
1.3 Spectroscopic Techniques for Protein Analysis	6
1.3.1 Fundamentals of Light-Matter Interactions	6
1.3.2 Electronic Absorption Spectroscopy (UV-Visible)	8
1.3.3 Vibrational Spectroscopy: Infrared and Raman	10
1.3.4 Fluorescence Spectroscopy	12
1.3.5 Spectral Interferences and Practical Limitations	13
1.4 Preprocessing of Spectroscopic Data	15
1.4.1 Smoothing	16
1.4.2 Normalization	17
1.4.3 Baseline Correction	17
1.4.4 Background Removal	17
1.4.5 Variable Selection	18
1.4.6 Derivative Spectroscopy	18
1.5 Chemometrics	19
1.5.1 Linear Regression	19
1.5.2 Multivariate Data Analysis	20
1.5.3 Machine Learning	22
1.5.4 Enabling Methodologies	29
1.6 Machine Learning Modeling Workflow	31
1.6.1 Data Generation and Experimental Design	31

1.6.2	Data Splitting Strategies	31
1.6.3	Model Evaluation and Performance Metrics	33
1.6.4	Model Interpretation and Validation	34
1.7	Combination of Spectroscopic Sensors with Process Models	35
2	Thesis Outline	43
2.1	Research Proposal	43
2.2	Manuscript Overview	46
3	An adaptive soft-sensor for advanced real-time monitoring of an antibody-drug conjugation reaction	49
	Robin Schiemer, Jan Tobias Weggen, Katrin Marianne Schmitt, and Jürgen Hubbuch	
3.1	Introduction	50
3.2	Materials and methods	51
3.2.1	Experiments	51
3.2.2	Data organization	53
3.2.3	Chemometric model development	53
3.2.4	Kinetic model calibration	54
3.2.5	Extended Kalman filter	54
3.2.6	Study design	57
3.3	Results and discussion	57
3.3.1	Chemometric model development	57
3.3.2	Soft-sensor development	60
3.4	Discussion	64
3.4.1	Chemometric model development	64
3.4.2	Kinetic model calibration	66
3.4.3	Soft-sensor development	66
3.5	Conclusions	67
4	Generative data augmentation and automated optimization of convolutional neural networks for process monitoring	71
	Robin Schiemer, Matthias Rüdte, and Jürgen Hubbuch	
4.1	Introduction	72
4.2	Materials and methods	74
4.2.1	Data and equipment	74
4.2.2	Data augmentation	76
4.2.3	Convolutional neural networks	78
4.2.4	Partial least squares regression	82
4.2.5	Feature importance	82
4.2.6	Robustness and transferability	83
4.3	Results and discussion	84
4.3.1	Generating highly realistic <i>in silico</i> spectra from experimental data	84
4.3.2	Training convolutional networks with augmented spectral data	85
4.3.3	Automating hyperparameter search by Bayesian optimization	88
4.3.4	Understanding model predictions through feature importance	92

4.3.5	Evaluating robustness and transferability	92
4.4	Discussion	94
4.4.1	Data augmentation	94
4.4.2	Model architecture and hyperparameter optimization	96
4.4.3	Model interpretability	97
4.4.4	Robustness	99
4.5	Conclusions	100
5	Raman-based PAT for VLP precipitation: Systematic data diversification and preprocessing pipeline identification	105
	Annabelle Dietrich, Robin Schiemer, Jasper Kurmann, Shiqi Zhang, Jürgen Hubbuch	
5.1	Introduction	106
5.2	Materials and methods	109
5.2.1	Experiments	109
5.2.2	Data analysis and computation	111
5.3	Results and discussion	114
5.3.1	Spiking diversifies experimental data	114
5.3.2	Raman spectra are affected by precipitant and precipitates	116
5.3.3	Raman data reveal structural differences between species	119
5.3.4	Background correction removes interferences	120
5.3.5	Systematic pipeline optimization improves model accuracy	122
5.3.6	Model pipeline captures precipitation trends	122
5.4	Discussion	126
5.4.1	Effects of data diversification	126
5.4.2	Effects of preprocessing operations on Raman data	127
5.4.3	Effects of preprocessing pipeline on model performance	128
5.5	Conclusions	130
6	A novel Raman-chromatography assembly for automated calibration and in-line monitoring in bioprocessing	137
	Jakob Heyer-Müller, Robin Schiemer, Matthias Lopinski, Caty Wang, Franka Willems, Lars Robbel, Michael Schmitt, Jürgen Hubbuch	
6.1	Introduction	138
6.2	Materials and methods	139
6.2.1	Experimental setup	139
6.2.2	Raman spectroscopy	141
6.2.3	Case study: sample preparation	141
6.2.4	Data preprocessing and chemometric modeling	142
6.2.5	Ultrafiltration/Diafiltration	142
6.3	Results and discussion	145
6.3.1	Automating Raman spectroscopy measurements by combining an ÄKTA chromatography system with a Raman spectrometer	145
6.3.2	Calibrating chemometric models off-line by synthetic mixtures	145
6.3.3	Applying chemometric models to quantify protein and buffer concentrations in ultrafiltration/diafiltration processes	147

6.4	Conclusions	148
7	Development of Raman spectroscopy and machine learning methods for protein aggregate quantification: Application to BSA in chromatographic processes.	151
	Jakob Heyer-Müller, Robin Schiemer, Lars Robbel, Michael Schmitt, and Jürgen Hubbuch	
7.1	Introduction	152
7.2	Materials and methods	154
7.2.1	Data analysis	156
7.3	Results and Discussion	161
7.3.1	Forced degradation studies	161
7.3.2	Mixing studies	162
7.3.3	Bind-elute anion-exchange chromatography experiments	164
7.4	Conclusions	171
8	General Discussion and Conclusion	175
	References	179
	List of Figures	199
	List of Tables	203
	Acronyms	207
A	Supplementary Material for Chapter 3	217
B	Supplementary Material for Chapter 4	227
C	Supplementary Material for Chapter 5	239
D	Supplementary Material for Chapter 6	247
E	Supplementary Material for Chapter 7	249
	Curriculum Vitae	259
	Signed Statements of Authorship Contribution	263

Introduction

Biopharmaceuticals have become the most promising class of pharmaceutical products to treat people diagnosed with cancer, inflammatory diseases, diabetes, viral infections or others with 8 of the 10 top selling products in 2024 being biopharmaceuticals (*Top Pharma Products by Global Sales 2024* 2025). From 2018 to 2022, over 29% of products approved by US regulators were novel biological entities with over 50% being monoclonal antibodies (mAbs) (G. Walsh and E. Walsh 2022). The dominance of mAbs in the biopharmaceutical market is owed to their specificity and potency while exhibiting fewer off-target adverse affects than other small molecule drugs (Castelli, McGonigle, and Hornby 2019). Although mAbs have proven to be valuable clinical assets in the treatment of chronic diseases such as cancer, their efficacy in a single regimen is often insufficient (Fu et al. 2022). Antibody-drug conjugates (ADCs), which combine the cytotoxicity of chemotherapeutic agents with the specificity of a mAb, are promising new class of biopharmaceutical products being currently evaluated in over 100 clinical trials and 14 approved products on the market (Fu et al. 2022). While mAbs and ADCs target disease through protein-based interventions, the delivery of therapeutic genetic material such as DNA and RNA offers the potential to intervene at the transcriptional or translational level, opening new avenues to address the underlying molecular basis of disease (Zhu et al. 2022). With the advent of gene therapy, viral vectors emerged as a robust delivery platform for the genetic cargo (Bulcha et al. 2021). Virus-like particles (VLPs) are self-assembled bionanoparticles composed of multiple recombinant proteins originally encoded in a viral capsid (Kheirvari, H. Liu, and Tumban 2023). They represent one of many possible viral vectors and have been used as vaccines or delivery vehicles for nucleic acid-based therapeutics (Mohsen and Bachmann 2022; Kheirvari, H. Liu, and Tumban 2023).

Although the different classes of biopharmaceuticals are structurally diverse and require specific expertise in development and manufacturing, most biopharmaceuticals are recombinantly produced in genetically modified cells. Their production in a bioreactor is commonly referred to as upstream processing (USP). This is followed by harvesting of the bioreactor and a sequence of unit operations to first separate the biologically active product from the host cells (HCs) and purify it from HC-

derived or product-related impurities. This sequence of unit operations is usually referred to as downstream processing (DSP) (Gunter Jagschies et al. 2018).

In process development, USP and DSP operations are optimized to maximize yield of the product, reduce the presence of impurities to a clinically tolerable level, and ensure robustness of the used unit operations to reliably produce product batches with consistent composition and quality. The pharmaceutical industry is regulated by regional authorities such as the United States Food and Drug Administration (FDA) and the European Medicines Agency (EMA) who define the necessary work packages to prove the robustness of manufacturing processes and quality of the pharmaceutical product (ICH 2009b). Since these work packages need to be executed for each product individually, process development activities are experimentally laborious. Hence, the biopharmaceutical industry operates on short timelines to be able to push as many assets through their development pipeline as possible (Schuhmacher, Gassmann, and Hinder 2016).

Current research in DSP not only aims to reduce manufacturing costs to further intensify and optimize individual unit operations, but also to establish computational and analytical techniques to expedite process development (Saleh et al. 2020; Gargalo et al. 2021; Helleckes et al. 2023). Current practices to evaluate the performance of a unit operation with regard to yield and impurity removal are largely based on off-line analytical procedures, i.e. the retrieval of a sample from process materials and their analysis in an analytical laboratory (Flatman et al. 2007). This separation of processing and analysis impedes time delays in development and manufacturing due to material transfer and additional sample preparation. Real-time sensors are being evaluated for their applicability in USP and DSP since the introduction of the process analytical technology (PAT) initiative by the FDA in 2004 (Glassey et al. 2011).

The following sections will establish the concepts and theoretical foundations for understanding the usage of PAT in DSP. First, the concept of PAT (Ch. 1.1) and the production processes of the herein studied biological product classes (Ch. 1.2) are introduced. Second, the theoretical foundations of different spectroscopic techniques as potential real-time sensors are presented (Ch. 1.3). In Ch. 1.4 mathematical techniques to remove unwanted effects from spectroscopic data are reviewed. In Ch. 1.5, the statistical regression theory from conventional linear regression to modern machine learning (ML) methods is introduced. Finally, the experimental and computational workflows required to establish robust chemometric models are provided (Ch. 1.6).

1.1 Process Analytical Technology (PAT)

PAT represents "a systematic framework for designing, analyzing, and controlling pharmaceutical manufacturing processes through timely measurements of critical quality and performance attributes of raw materials, in-process materials, and processes" (FDA 2004). The formal introduction of PAT by the FDA in 2004 marked a paradigm shift in pharmaceutical manufacturing philosophy, transitioning from a quality-by-testing to a quality-by-design approach (FDA 2004). The implementation of PAT in biopharmaceutical manufacturing addresses fundamental limitations inherent in traditional quality control strategies. Conventional approaches rely predominantly on intermediate and end-product testing, which provides limited insight into process dynamics and offers minimal opportunity for real-time process adjustment (Glassey et al. 2011). In contrast, PAT enables continuous in-process monitoring of critical process parameters (CPPs) and critical quality attributes (CQAs), facilitating immediate detection of process deviations and enabling corrective actions before product quality is compromised (A. S. Rathore, Bhambure, and Ghare 2010; E. K. Read et al. 2010).

Current in-process monitoring is based on the usage of univariate sensors like pressure, temperature, or ultraviolet (UV) while intermediate and end-product testing is largely done by off-line analytical

methods such as ultrahigh-performance liquid chromatography (UHPLC), immunoassays or gel electrophoresis (Bakeev 2005). The unavailability of real-time information about CPPs and CQAs in biopharmaceutical production may lead to time delays between batch-wise operations and potentially compromise product quality in case of unforeseen deviations.

Modern PAT tools are built on multivariate sensors, such as UV, infrared (IR) or Raman spectrometers, which can be used to determine the composition of a sample taken from the process or directly in the process stream (Rüdt, Briskot, and Hubbuch 2017; Rolinger, Rüdt, and Hubbuch 2020a). In this way, they facilitate the immediate detection of process deviations and enable real-time process control. Multivariate sensor technologies capture process or product information in the form of high-dimensional spectroscopic data (Bakeev 2005). However, their successful application depends on chemometric and machine learning methods that can extract relevant information from datasets which are often noisy and highly collinear (Otto 2017).

In the pharmaceutical field, the combination of spectroscopic sensors with established chemometric methods has already proven valuable in biopharmaceutical processes like cell culture (Abu-Absi et al. 2011), chromatography (Thakur, Thori, and Anurag S. Rathore 2020; Rüdt, Brestrich, et al. 2017; Brestrich, Hahn, and Hubbuch 2016), filtration (Thakur, Hebbs, and Anurag S. Rathore 2020; Rolinger, Hubbuch, and Rüdt 2023). When applied to biopharmaceutical processes, however, these conventional approaches frequently fall short of the robustness and selectivity needed to satisfy regulatory standards in a Good manufacturing practice (GMP) environment due to high complexity of biologic materials (Rolinger, Rüdt, and Hubbuch 2020a; Hong et al. 2018). To address these limitations, more advanced machine learning and statistical approaches are being adopted (Rolinger, Rüdt, and Hubbuch 2020a; Banner et al. 2021). Yet this shift introduces a new challenge: regulatory guidelines from agencies such as the EMA and FDA require demonstrated selectivity and robustness of analytical methods in GMP environments (EMA 2014), criteria that are difficult to satisfy with black-box models whose internal decision-making remains opaque. Furthermore, such models often demand extensive calibration efforts that may not be feasible in practice. The need for advanced methods to handle complexity is counterintuitive to the regulatory expectations for transparency and reliability and thus underscores the importance of developing calibration strategies and interpretability tools that can bridge the gap between model performance and regulatory acceptance.

1.2 PAT Applications in Biopharmaceutical Downstream Processing

The implementation of PAT in biopharmaceutical manufacturing presents unique challenges and opportunities compared to small molecule pharmaceutical production. Biopharmaceutical products, particularly therapeutic proteins, exhibit inherent structural complexity and heterogeneity that necessitate sophisticated analytical approaches for comprehensive characterization (Roberts 2014). The complexity of biopharmaceutical DSP, which encompasses multiple unit operations including chromatography, filtration, and formulation steps, provides numerous opportunities for PAT implementation (Glasse et al. 2011). Each unit operation presents specific monitoring requirements and analytical challenges that must be addressed through appropriate selection and implementation of PAT tools. The high value of biopharmaceutical products and the critical nature of their therapeutic applications justify the investment in advanced analytical technologies and create strong economic incentives for process optimization through PAT implementation (Gillespie et al. 2022). This chapter examines PAT applications for three major categories of biopharmaceutical products: mAbs, VLPs,

and ADCs, emphasizing the specific monitoring strategies and analytical technologies applicable to each.

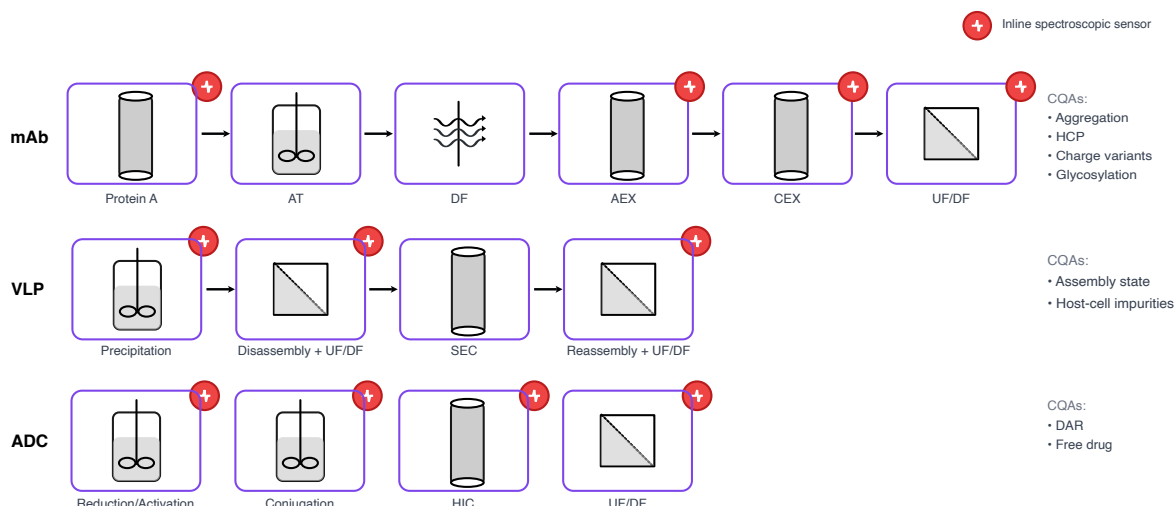


Figure 1.1 Typical downstream processes for biopharmaceutical products: mAbs, VLPs and ADCs. Possible PAT integration sites are highlighted. Abbreviations: Protein A: Protein A chromatography, AT: Acid treatment, DF: Depth filtration, AEX: Anion-exchange chromatography, CEX: Cation-exchange chromatography, UF/DF: ultrafiltration/diafiltration, HCP: Host-cell proteins, SEC: Size-exclusion chromatography, HIC: Hydrophobic interaction chromatograph, DAR: Drug-to-antibody ratio

1.2.1 Monoclonal Antibodies

MAbs represent the largest class of biopharmaceutical products (G. Walsh and E. Walsh 2022), with well-established production platforms and standardized downstream processing strategies (Shukla et al. 2007b). These immunoglobulin molecules, typically 150 kDa glycoproteins composed of two heavy and two light chains, exhibit defined structural features that facilitate platform purification approaches (Carta and Jungbauer 2010).

The downstream processing of mAbs typically follows a platform approach (cf. Fig. 1.1), consisting of multiple sequential chromatography, adjustment and filtration steps. Protein A chromatography is used to selectively capture the mAb molecules from the clarified harvest cell culture material (HCCF), mainly removing HC-derived impurities. This is followed a low pH virus inactivation and depth filtration to inactivate adventitious viruses and clarify the suspension from particulates. The subsequent ion-exchange-based chromatography steps are used to further reduce host cell proteins (HCPs) and product-related impurities such as low molecular weight species (LMW) or high molecular weight species (HMW). The exact sequence and operating mode of the polishing chromatography steps is variable and may be selected based on specific product needs. After polishing, the concentration of the mAb in solution as well as buffer and excipients is adjusted for its final formulation via ultra-/diafiltration (UF/DF). Each unit operation is associated with specific monitoring needs to ensure control over product quality attributes and overall process performance (Shukla et al. 2007a).

CQAs of mAbs include aggregation state, fragmentation levels, charge heterogeneity, glycosylation patterns and HCP content, each of which may influence therapeutic efficacy and safety (Vázquez-Rey and Lang 2011; Flatman et al. 2007). Aggregation can compromise drug stability and trigger immunogenic responses (Rosenberg 2006; Moussa et al. 2016), whereas charge variants, often arising from post-translational modifications such as deamidation or oxidation, can alter both stability and biological activity (Tyther and Jenkins 2011). Glycosylation, particularly the extent of fucosylation and galactosylation, is a key determinant of effector functions such as antibody-dependent cellular cytotoxicity (Thomann et al. 2016). HCPs, by contrast, represent process-related impurities that must be tightly controlled in the ppm range due to their potential to elicit adverse immune responses (Vanderlaan et al. 2018). DSP process design influences the overall impurity profile by selectively enriching or depleting certain CQAs. In this context, PAT offers a unique opportunity to non-invasively detect these critical quality attributes in real-time and thus enable process control, thereby safeguarding both product consistency and patient safety. In summary, all chromatography operations as well as the UF/DF process may benefit from the integration of PAT to monitor the solution composition in real-time to define pooling criteria or adjust operating conditions.

1.2.2 Virus-Like Particles

VLPs are protein-based nanostructures that mimic viral morphology while lacking genetic material to self-replicate (Chackerian 2007; Zeltins 2013). They are typically composed of one or more structural proteins of a virus, which spontaneously self-assemble into particles that are antigenically indistinguishable from infectious virus particles (Qian et al. 2020). Their size typically ranges from 20–200 nm, and their complex assembly into higher-order structures make them attractive vaccine candidates but also present substantial challenges for manufacturing (Kushnir, Streatfield, and Yusibov 2012; Effio and Hubbuch 2015).

Key CQAs for VLPs include assembly state, and nucleic acid encapsidation, as these properties directly determine immunogenicity and vaccine efficacy (Zeltins 2013; Kushnir, Streatfield, and Yusibov 2012). In contrast to mAbs, the structural integrity of VLPs is highly sensitive to process conditions, and disruption of their quaternary structure often leads to irreversible product loss (Moleirinho et al. 2020).

Typical downstream processing of VLPs involves clarification, concentration, chromatographic purification, and final polishing steps (Effio and Hubbuch 2015; Hillebrandt, Vormittag, Bluthardt, et al. 2020). Because VLP preparations often contain partially assembled particles and aggregates, unit operations must balance impurity removal with preservation of structural integrity (Hillebrandt, Vormittag, Bluthardt, et al. 2020). For non-enveloped VLPs, a novel membrane-based process has been established that exploits the reversible nature of VLP assembly (Hillebrandt and Hubbuch 2023; Hillebrandt, Vormittag, Dietrich, and Hubbuch 2022). As illustrated in Fig. 1.1 (middle row), this process begins with precipitation for initial capture and concentration, followed by controlled disassembly of the particles using UF/DF. In the disassembled state, encapsulated impurities such as host-cell nucleic acids become accessible and can be removed via size exclusion chromatography (SEC) (Valentic and Hubbuch 2024). Subsequently, the subunits are reassembled into intact VLPs through a second UF/DF step under conditions that promote proper particle formation. This disassembly–reassembly strategy not only enables more thorough impurity removal but also offers the potential to improve particle homogeneity and structural properties (Hillebrandt, Vormittag, Dietrich, and Hubbuch 2022). Monitoring these assembly transitions is of particular importance, as incomplete disassembly or improper reassembly directly compromises product quality and potency.

1.2.3 Antibody-Drug Conjugates

ADCs combine the specificity of mAbs with the potency of cytotoxic small molecules, creating a new class of biopharmaceuticals with complex quality requirements (Agarwal and Bertozzi 2015). Owing to this dual mechanism, ADCs have emerged as a promising category of anti-cancer therapeutics, reflected by more than 100 molecules currently in clinical trials and several products already on the market (Fu et al. 2022).

In addition to the production of the raw materials, namely the mAb and the cytotoxic small molecule, the manufacturing of ADCs typically involves a conjugation reaction followed by multiple purification steps (cf. Fig. 1.1 (bottom)). Depending on the molecular structure of the mAb, the conjugation reaction may be site-specific or stochastic (Agarwal and Bertozzi 2015; Sadiki et al. 2020). For site-specific conjugations, the conjugation sites are predetermined by specifically engineered side chains and are mediated through the introduction of a chemical linker (Agarwal and Bertozzi 2015). In the case of stochastic conjugations, the conjugation sites are unspecified and need to be activated through chemical reduction prior to conjugation (Sadiki et al. 2020). The conjugated mAbs are then subjected to a sequence of chromatographic and filtration steps to ensure a consistent drug-to-antibody ratio (DAR) and remove the residual chemical agents present in the ADC solution. hydrophobic interaction chromatography (HIC) is often employed to separate species based on DAR, and UF/DF for the removal of free cytotoxic drug and final formulation (Hu et al. 2017; Andris and Hubbuch 2020). Among the mentioned CQAs, the DAR is particularly critical, as it directly affects efficacy, safety, and stability, with optimal ranges depending on the ADC (Hu et al. 2017).

1.3 Spectroscopic Techniques for Protein Analysis

Spectroscopic methods represent a powerful class of analytical tools for real-time process monitoring in biopharmaceutical manufacturing. Their non-invasive nature, rapid acquisition times, and ability to provide molecular-level information make them particularly suitable for PAT applications. This section establishes the theoretical foundations of spectroscopic techniques relevant to protein analysis, beginning with the fundamental physics of light-matter interactions before examining specific measurement modalities and their practical considerations.

1.3.1 Fundamentals of Light-Matter Interactions

Spectroscopy is founded on the interaction between electromagnetic radiation and matter. When light encounters a molecule, several phenomena may occur: the light may be transmitted, absorbed, scattered, or re-emitted. Each of these interactions can provide distinct information about molecular structure, composition, and dynamics. The quantum mechanical description of these interactions is essential for understanding why different spectroscopic techniques probe different molecular properties. Molecules exist in discrete energy states, and transitions between these states can only occur when the energy of the interacting photon matches the energy difference between two states. This resonance condition is expressed as:

$$\Delta E = h\nu = \frac{hc}{\lambda}, \quad (1.1)$$

where ΔE is the energy difference between molecular states, h is Planck’s constant, ν is the frequency of light, c is the speed of light, and λ is the wavelength. This relationship governs all absorption spectroscopy and explains why different wavelength regions probe different molecular phenomena (Gauglitz and Moore 2014).

The electromagnetic spectrum can be divided into regions based on the type of molecular transition that is induced (cf. Fig. 1.2 (A)). At ultraviolet and visible frequencies (approximately 200–800 nm), photon energies are sufficient to promote electrons from occupied to unoccupied molecular orbitals. At longer wavelengths in the infrared region (2,500–25,000 nm, or equivalently $4000\text{--}400\text{ cm}^{-1}$), incoming photons couple instead to nuclear motion, exciting molecular vibrations. The near-infrared region (800–2,500 nm) represents an intermediate regime where overtones and combination bands of fundamental vibrations are observed (Atkins and Paula 2006).

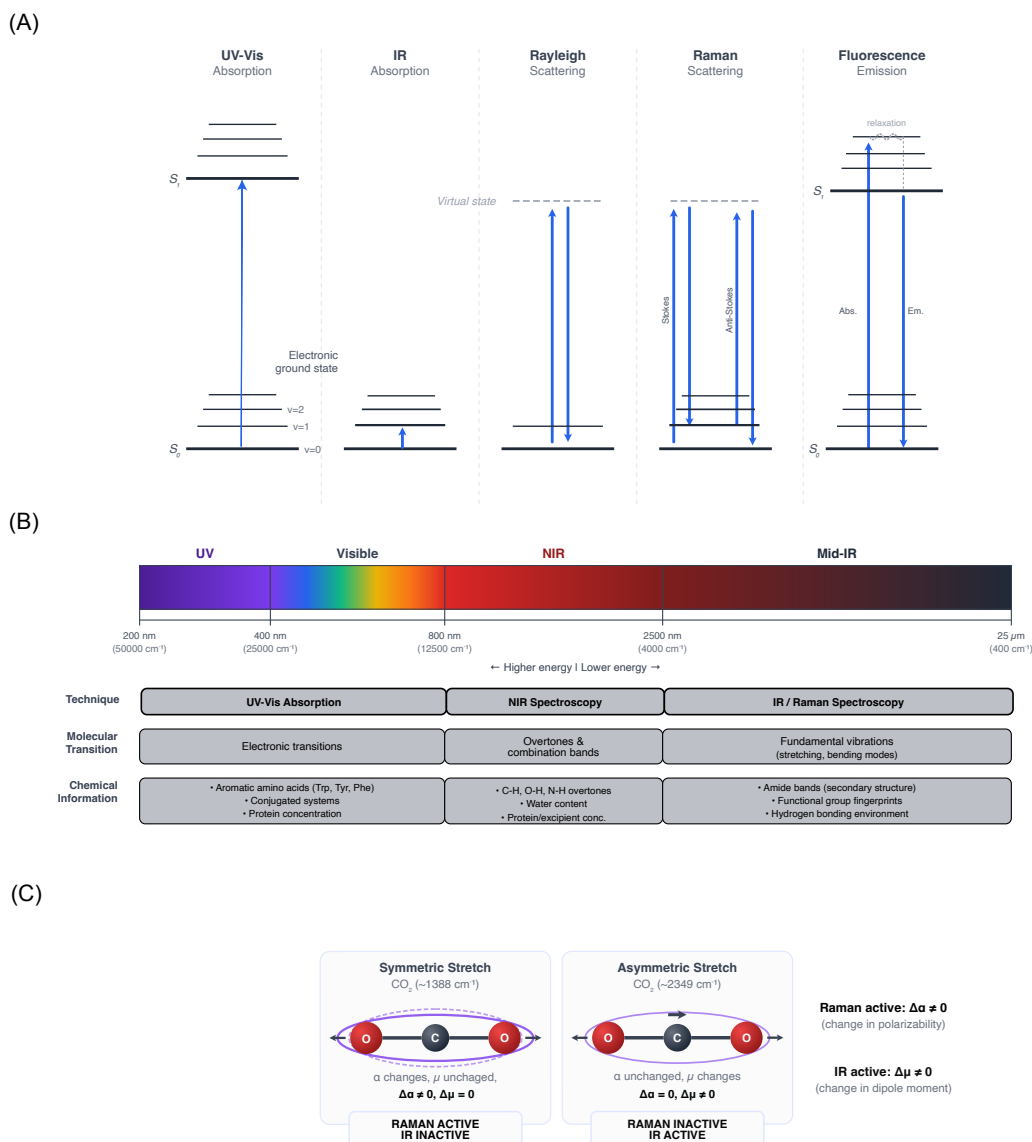


Figure 1.2 Schematic illustration of different spectroscopic effects. (A) Jablonski diagram comparing different spectroscopic techniques and the energy transitions underlying the optical phenomena. (B) Electromagnetic spectrum and light categories depending on the wavelength range. (C) Illustration of complementarity of IR and Raman spectroscopy using the carbondioxide as an example. While the asymmetric stretch is active in IR spectroscopy, the symmetric stretch is not. For Raman spectroscopy, the selection rule cause the opposite behavior.

1.3.2 Electronic Absorption Spectroscopy (UV-Visible)

1.3.2.1 Principles and Quantification

Absorption spectroscopy measures the attenuation of light as it passes through a sample. When a photon's energy matches the energy gap between two molecular states, the photon may be absorbed,

promoting the molecule to an excited state (cf. Fig. 1.2 (B)). The probability of absorption depends on quantum mechanical selection rules that govern whether a particular transition is allowed.

For electronic transitions, the selection rules depend on a variety of quantum mechanical factors, including symmetry considerations and the nature of the molecular orbitals involved. A key quantity is the transition dipole moment, which describes how strongly the electromagnetic field couples to the electronic redistribution during a transition. Transitions with large transition dipole moments generally appear as intense absorption bands, while those with small or zero values are weak or formally forbidden (Atkins and Paula 2006). The absorption of UV-visible light promotes electrons from lower- to higher-energy orbitals. The wavelength at which absorption occurs is related to the energy gap between the involved orbitals, though predicting exact transition energies and intensities for complex molecules remains challenging and typically requires computational methods.

The quantitative relationship between absorption and concentration is described by the Lambert-Beer law (Gauglitz and Moore 2014):

$$A = \varepsilon \cdot c \cdot l, \quad (1.2)$$

where A is the absorbance at a given wavelength (defined as $A = -\log_{10}(I/I_0)$ with I_0 and I being incident and transmitted intensity), ε is the molar absorption coefficient ($\text{L mol}^{-1} \text{cm}^{-1}$), c is the molar concentration of the absorbing species (mol L^{-1}), and l is the optical path length (cm). This linear relationship between absorbance and concentration forms the basis for quantitative spectroscopic analysis, though several phenomena can cause deviations from ideal behavior, as discussed in Section 1.3.5.

1.3.2.2 Protein Chromophores

In proteins, the relevant UV-absorbing groups (chromophores) include the peptide bond and the aromatic amino acid side chains (Schmid 2001). The peptide backbone exhibits strong absorbance in the near-UV region between 190 and 230 nm (Schmid 2001). These transitions provide the basis for protein detection at wavelengths below 230 nm, though measurements in this region are complicated by absorption from oxygen and many buffer components.

The aromatic amino acids—tryptophan, tyrosine, and phenylalanine—dominate protein absorption in the 250–300 nm range due to transitions within their conjugated ring systems (Rüdt, Briskot, and Hubbuch 2017). Disulfide bonds between cysteine residues also contribute weak absorption near 260 nm (Schmid 2001).

The electronic environment of these chromophores modulates their spectral properties. Changes in local polarity, hydrogen bonding, or proximity to other aromatic residues can shift absorption maxima (solvatochromism) or alter intensities. For instance, the spectroscopic properties of aromatic amino acids such as tryptophan are sensitive to their local electronic environment, with residues buried in hydrophobic regions exhibiting different spectral characteristics than those exposed to aqueous solvent (Lakowicz 2006, p. 534). These effects encode information about protein tertiary structure and conformational state, though extracting this information quantitatively requires careful analysis (Rüdt, Briskot, and Hubbuch 2017).

Because each protein possesses a unique amino acid sequence, the number, type, and spatial distribution of aromatic chromophores varies between proteins. This variation produces characteristic UV spectral signatures that can, in principle, distinguish different proteins or protein variants from one another (Hansen, Jamali, and Hubbuch 2013; Rüdt, Briskot, and Hubbuch 2017). Furthermore, when multiple proteins are present in solution, their individual spectral contributions superimpose according to Beer’s law. With appropriate multivariate calibration methods, it becomes possible to

resolve these overlapping signals and quantify each component simultaneously, enabling the analysis of complex protein mixtures without prior separation (Rüdt, Briskot, and Hubbuch 2017).

1.3.3 Vibrational Spectroscopy: Infrared and Raman

Vibrational spectroscopy probes transitions between quantized vibrational energy levels arising from the oscillations of atoms within molecules (cf. Fig. 1.2 (B)). Whereas UV absorption provides information primarily about aromatic and other conjugated chromophores, vibrational spectroscopy yields spectra with contributions from virtually all chemical bonds and functional groups, providing a more comprehensive molecular fingerprint for structural characterization.

1.3.3.1 Molecular Vibrations

Molecular vibrations can be conceptualized using the harmonic oscillator approximation, where atoms connected by chemical bonds behave similarly to masses connected by springs. For a diatomic molecule, the classical vibrational frequency ν is given by (Atkins and Paula 2006)

$$\nu = \frac{1}{2\pi} \sqrt{\frac{k}{\mu}}, \quad (1.3)$$

where k is the force constant (bond strength) and μ is the reduced mass of the two atoms $\mu = m_1 m_2 / (m_1 + m_2)$ with m_1 and m_2 being the mass of the individual atoms. This relationship explains several empirical observations: stronger bonds (larger k) vibrate at higher frequencies, and bonds involving lighter atoms (smaller μ) also vibrate faster. For example, the C=C double bond stretches at higher frequency ($\sim 1700 \text{ cm}^{-1}$) than the C-C single bond ($\sim 1100 \text{ cm}^{-1}$) (Atkins and Paula 2006).

Polyatomic molecules exhibit multiple vibrational modes. A nonlinear molecule with N atoms has $3N - 6$ normal modes of vibration (or $3N - 5$ for linear molecules), each corresponding to a coordinated motion of all atoms at a characteristic frequency. These modes include stretching vibrations (changes in bond length), bending vibrations (changes in bond angle), and more complex motions such as rocking, wagging, and torsional modes (Atkins and Paula 2006).

1.3.3.2 Selection Rules of IR and Raman

The fundamental distinction between infrared and Raman spectroscopy lies in how electromagnetic radiation couples to molecular vibrations, leading to different selection rules that determine which vibrations are observable by each technique.

In infrared absorption spectroscopy, a photon is absorbed when its oscillating electric field resonates with a vibrating molecular dipole. For a vibration to be IR-active, it must produce a change in the molecule's dipole moment during the oscillation (Atkins and Paula 2006). Mathematically, the IR intensity is proportional to:

$$I_{\text{IR}} \propto \left(\frac{\partial \mu}{\partial Q} \right)^2, \quad (1.4)$$

where μ is the molecular dipole moment and Q is the normal coordinate of the vibration. Vibrations that preserve molecular symmetry, such as the symmetric stretch of CO_2 or the C-C stretching in symmetric molecules, do not produce a net dipole change and are therefore IR-silent.

Raman spectroscopy operates through a fundamentally different mechanism: inelastic scattering of light. When monochromatic light, typically from a laser, interacts with a molecule, most photons scatter elastically with unchanged energy (cf. Fig. 1.2 (B)). This phenomenon is called Rayleigh scattering. However, a small fraction of photons ($\sim 10^{-6}$) exchange energy with molecular vibrations through an inelastic process. This energy exchange shifts the scattered photon’s frequency by an amount equal to the vibrational frequency which is called the Raman shift (Gauglitz and Moore 2014).

A vibration is Raman-active if it modulates the molecular polarizability α , which describes how easily the electron cloud deforms in response to an external electric field (Gauglitz and Moore 2014). The Raman intensity is proportional to

$$I_{\text{Raman}} \propto \left(\frac{\partial \alpha}{\partial Q} \right)^2. \quad (1.5)$$

Symmetric stretching vibrations, which compress and expand the electron cloud uniformly, typically produce changes in polarizability and show strong Raman activity. Conversely, asymmetric vibrations that primarily shift electron density often show weak Raman scattering.

Complementarity of IR and Raman These complementary selection rules mean that vibrations silent in IR may be prominent in Raman, and vice versa. For molecules with a center of symmetry, the “rule of mutual exclusion” states that no vibration can be both IR and Raman active (Hollas 2010). This complementarity makes combined IR and Raman analysis powerful for complete vibrational characterization. A prominent example of this behavior is the symmetric and asymmetric stretch of the carbondioxide molecule as schematically illustrated in Fig. 1.2.

1.3.3.3 Protein Vibrational Signatures

Proteins generate rich vibrational spectra due to their diverse chemical functionality. The most diagnostically useful spectral features arise from the peptide backbone, aromatic amino acids, and disulfide bonds (Rygula et al. 2013; Barth 2007).

Amide Bands The peptide bond gives rise to several characteristic vibrational bands, designated amide I, II, and III. The amide I band ($1600\text{--}1700\text{ cm}^{-1}$) is most commonly used for secondary structure analysis. This band arises primarily from C=O stretching of the peptide carbonyl, with minor contributions from C-N stretching and N-H bending. The exact frequency of the amide I band depends sensitively on the hydrogen bonding pattern and geometry of the peptide backbone, making it a powerful probe of secondary structure (Barth 2007). The amide I band is a superposition of multiple vibrational bands originating from multiple secondary structure elements like α -helices ($1650\text{--}1658\text{ cm}^{-1}$), β -sheets ($1620\text{--}1640\text{ cm}^{-1}$ and $1670\text{--}1690\text{ cm}^{-1}$), random coils ($1640\text{--}1650\text{ cm}^{-1}$) and β -turns ($1660\text{--}1685\text{ cm}^{-1}$) (Rygula et al. 2013; Barth 2007).

The amide II ($1510\text{--}1580\text{ cm}^{-1}$) and III ($1230\text{--}1300\text{ cm}^{-1}$) bands result from a combination of N-H bending and C-N stretching (Barth 2007). The amide III band offers an alternative window for secondary structure analysis that avoids the interference with water in the amide I region.

Aromatic Amino Acids Tryptophan and tyrosine exhibit characteristic Raman bands that serve as probes of local environment and tertiary structure (Rygula et al. 2013; Harada, Miura, and Takeuchi 1986). The tryptophan indole ring produces a doublet near 1340 and 1360 cm^{-1} , whose intensity ratio I_{1360}/I_{1340} reflects the strength of hydrogen bonding at the indole N-H group (Barth and Zscherp 2002; Harada, Miura, and Takeuchi 1986). A ratio greater than unity indicates a strong hydrogen bond donor environment, while a ratio less than unity suggests a hydrophobic

environment or weak hydrogen bonding (Harada, Miura, and Takeuchi 1986). Tyrosine exhibits a further characteristic doublet near 830 and 850 cm^{-1} (Rygula et al. 2013).

Disulfide Bonds The S-S stretching vibration of disulfide bonds appears in the 500–550 cm^{-1} region and is exclusively Raman-active due to the symmetric nature of the vibration (Rygula et al. 2013). This sensitivity makes Raman spectroscopy valuable for monitoring disulfide bond integrity during protein processing (Bazylewski, Divigalpitiya, and Fanchini 2017).

Glycosidic Linkages For glycoproteins such as monoclonal antibodies, vibrations from carbohydrate moieties contribute to the spectral signature. C-O-C stretching modes of glycosidic linkages appear in the 1000–1150 cm^{-1} region, while C-H deformation modes of sugar rings contribute around 1300–1450 cm^{-1} (Wiercigroch et al. 2017). Although these features are often weaker than protein backbone vibrations, they can provide information about glycosylation patterns when measured with sufficient sensitivity (M.-Y. Li et al. 2018; Brewster, Ashton, and Goodacre 2011).

1.3.3.4 Measurement Configurations

Conventional transmission IR spectroscopy, where the sample is placed between IR-transparent windows, is problematic for aqueous protein solutions due to the strong absorption of water (discussed in Section 1.3.5) (Tiernan, Byrne, and Kazarian 2020). Several alternative sampling configurations have been developed to address this limitation.

In attenuated total reflectance (ATR) spectroscopy, the IR beam is directed into a crystal of high refractive index such as diamond, zinc selenide, or germanium at an angle exceeding the critical angle for total internal reflection. At the reflection interface, an evanescent wave penetrates a short distance of 0.5–2 μm , depending on wavelength and refractive indices into the sample in contact with the crystal surface (Gauglitz and Moore 2014). This limited penetration depth reduces the effective path length, making measurements of aqueous solutions feasible. ATR is the most common configuration for protein IR spectroscopy in aqueous environments and is well-suited for process applications due to its robust, flow-compatible design (Boulet-Audet, Byrne, and Kazarian 2014).

Raman spectroscopy offers significant advantages for process applications due to its minimal water interference and compatibility with standard fiber optics. Several excitation wavelengths and measurement geometries are employed depending on the application.

The choice of excitation wavelength represents a critical compromise between signal intensity and background interference. Raman scattering intensity scales with λ^{-4} , meaning that shorter excitation wavelengths are favorable for high signal intensity. However, shorter wavelengths also more efficiently excite fluorescence from aromatic residues and impurities, which can overwhelm the Raman signal (see Section 1.3.5) (Goldrick, Lovett, et al. 2018). Common excitation wavelengths include 532, 633, 785 and 1064 nm (Goldrick, Lovett, et al. 2018). For protein process monitoring, 785 nm excitation has emerged as a practical compromise, providing adequate signal with manageable background interference for most applications (Wen 2007).

1.3.4 Fluorescence Spectroscopy

Fluorescence represents a complementary optical phenomenon where absorbed photon energy is re-emitted following internal relaxation processes (cf. Fig. 1.2 (B)). Unlike the instantaneous process of Raman scattering, fluorescence involves absorption to an excited electronic state, followed by vibrational relaxation within that state, and finally radiative decay back to the ground state. The emitted photon has lower energy than the absorbed photon, with the energy difference

corresponding to the energy dissipated during vibrational relaxation (Atkins and Paula 2006, p. 492).

In proteins, intrinsic fluorescence arises predominantly from the aromatic amino acids, with tryptophan being the dominant fluorophore (Lakowicz 2006, p. 530). Tyrosine and phenylalanine also fluoresce but contribute less to protein emission spectra. Tryptophan fluorescence is typically excited at 280–295 nm, with emission maxima ranging from 310 to 350 nm depending on the local environment. A tryptophan buried in a hydrophobic protein core experiences a nonpolar environment that stabilizes the ground state relative to the excited state, resulting in emission at shorter wavelengths (~ 320 nm). Conversely, a tryptophan exposed to aqueous solvent shows emission at longer wavelengths (~ 350 nm) (Lakowicz 2006, p. 534). This environmental sensitivity makes tryptophan fluorescence a powerful probe of protein folding, conformational changes, and ligand binding (Loving, Sainlos, and Imperiali 2010).

1.3.5 Spectral Interferences and Practical Limitations

The successful application of spectroscopic methods for protein analysis requires understanding and mitigating various sources of spectral interference. These interferences can cause deviations from the linear relationships assumed by chemometric models and must be addressed through appropriate experimental design, instrumental configuration, or data preprocessing.

1.3.5.1 Deviations from Lambert-Beer Linearity

The Lambert-Beer law assumes ideal conditions that are often violated in practical applications. Understanding these deviations is essential for developing robust quantitative methods. The Lambert-Beer law is valid only at sufficiently low analyte concentrations. Two physical phenomena contribute to this limitation. First, at elevated concentrations, the absorption cross-sections of neighbouring analyte molecules are no longer additive due to electromagnetic interactions between their induced dipole moments, and local electric field effects further enhance absorption beyond what would be predicted from isolated molecules (Mayerhöfer, Pahlow, and Popp 2020). Second, the refractive index of a solution varies with solute concentration. Since the molar absorption coefficient depends inversely on the refractive index, variations in concentration produce corresponding changes in absorptivity that violate the assumption of a concentration-independent ϵ (Mayerhöfer, Pahlow, and Popp 2020).

Furthermore, instrumental factors contribute to non-linearity. Stray light, radiation reaching the detector without passing through the sample due to imperfections in the wavelength selector, constitutes one such limitation (*Beer's Law* 2023). At high sample concentrations, the transmitted intensity approaches the magnitude of the stray light contribution, resulting in measured absorbance values that are systematically lower than the true values (*Beer's Law* 2023).

1.3.5.2 Light Scattering

Light scattering by particles, aggregates, or bubbles in the sample represents a major interference in absorption spectroscopy (Porterfield and Adam Zlotnick 2010; Basalla and Kendrick 2023). Scattered light that does not reach the detector appears as apparent absorption, elevating the baseline and potentially masking true absorption features.

The wavelength dependence of scattering provides some diagnostic capability. Rayleigh scattering from particles smaller than one-tenth of the wavelength varies as λ^{-4} , producing a steeply rising baseline toward shorter wavelengths (Porterfield and Adam Zlotnick 2010). Larger particles produce Mie scattering with weaker wavelength dependence, becoming essentially wavelength-independent

for particles 10–100 times larger than the wavelength (Basalla and Kendrick 2023). In either case, scattering contributions appear as broad, featureless baseline elevation rather than discrete absorption bands.

1.3.5.3 Interferences in Raman Spectroscopy

Fluorescence represents the most significant interference in Raman spectroscopy of proteins and biological samples (Buckley and Ryder 2017; Wen 2007). Understanding why fluorescence is problematic requires comparing the two phenomena quantitatively.

Raman scattering is an inherently weak process: only approximately one in 10^6 to 10^8 incident photons undergoes inelastic scattering. Fluorescence quantum yields, by contrast, can approach unity for efficient fluorophores. When a sample contains fluorescent species, the fluorescence emission can exceed the Raman signal by several orders of magnitude, completely obscuring the vibrational information (Buckley and Ryder 2017).

The spectral overlap between fluorescence and Raman further compounds this problem. Both signals appear at longer wavelengths compared to the excitation wavelength. While Raman bands are narrow and appear at fixed wavenumber shifts from the excitation, fluorescence emission is broad, typically 50–100 nm, and its spectral position is determined by the fluorophore’s electronic structure rather than the excitation wavelength (Lakowicz 2006, p. 534). For proteins, tryptophan fluorescence excited at visible wavelengths produces broad emission extending across the entire Raman spectral window (Buckley and Ryder 2017).

One approach to mitigate this interference is moving to higher excitation wavelengths such as 633, 785 or 1064 nm which effectively reduces fluorescence excitation faster than it reduces Raman scattering, improving the signal-to-background ratio (Buckley and Ryder 2017). Other strategies to mitigate this effect either employ modern optical technologies such as time-gated detection, where pulsed excitations are used to discriminate fluorescence from Raman scattering, or shifted excitation difference spectroscopy, where Raman spectra are recorded at two slightly different excitation wavelengths and the difference is computed to cancel the fluorescence out (Buckley and Ryder 2017). Once the optical configuration is fixed, computational approaches can model and subtract the broad fluorescence background to recover solely Raman features from the raw measurement. This approach is commonly referred to as baseline correction and will be covered in more detail in Sec. 1.4.

Samples from biopharmaceutical processes usually come in aqueous solutions with solution matrices containing buffers, salts, excipients, and other additives that can interfere with spectroscopic measurements. With water being present in all aqueous solutions, it is the most common interferent. While water is not influencing UV spectroscopy measurements, vibrational spectroscopic measurements are obscured by a water signal. In IR, the and H-O-H bending ($\sim 1645\text{ cm}^{-1}$) modes of water produce intense, broad absorption bands that can completely obscure the sample signal (Barth 2007). Critically, the water bending mode directly overlaps with the protein amide I band, which is the most informative region for secondary structure analysis (Barth 2007). Although Raman is less sensitive to water interference, the same vibrational modes are present Raman spectra (Butler et al. 2016).

In UV spectroscopy, buffer components that contain carboxyl or amino groups such as histidine, trishydroxymethylaminomethane (Tris), or 4-morpholinoethanesulfonic acid (MES) absorb below 230 nm, limiting measurements to the aromatic absorption region (Schmid 2001). In vibrational spectroscopy, buffer components contribute their own spectral features that must be accounted for. To name a few examples: phosphate buffer produces vibrational bands around $800\text{--}1100\text{ cm}^{-1}$

(Preston and Adams 1979); carbohydrate-based excipients contribute in the 900–1200 cm^{-1} region (Wiercigroch et al. 2017); and surfactants such as polysorbate 80 have C-H stretching and deformation modes throughout the fingerprint region (Saggu, Jun Liu, and Patel 2015). These interferences can be addressed through careful reference subtraction if the matrix composition is known, or through multivariate calibration approaches that explicitly model matrix variability.

In summary, spectroscopic techniques provide powerful, non-invasive methods for probing protein structure, composition, and concentration in process environments. Each technique offers distinct advantages and faces specific limitations. UV-visible absorption provides robust protein quantification through well-understood electronic transitions but offers limited structural information and is susceptible to scattering artifacts. Infrared spectroscopy delivers detailed secondary structure information through analysis of amide bands but requires specialized sampling configurations to overcome water interference. Raman spectroscopy combines structural sensitivity with excellent water compatibility, but must contend with inherently weak signals and potential fluorescence interference.

The effective application of these techniques for PAT in downstream processing relies not only on understanding their physical principles but also on careful experimental design to minimize interferences, appropriate preprocessing to correct for instrumental and matrix effects, and robust chemometric methods to extract relevant information from complex, overlapping spectral features. These data analysis approaches will be addressed in the following sections.

1.4 Preprocessing of Spectroscopic Data

The successful application of chemometric models to spectroscopic data critically depends on the quality and consistency of the input spectra. Raw spectra often contain unwanted variability arising from instrumental noise, baseline drift, light scattering, and matrix effects, which can obscure the underlying chemical information (Mehmood et al. 2012). Preprocessing methods aim to remove these artifacts, enhance signal-to-noise ratio, and standardize spectra so that statistical regression models commonly employed in chemometrics can extract meaningful patterns reliably. In the following paragraphs, the most common preprocessing operations are conceptually presented. An illustrative overview of the effect of the most common operations is presented in Fig. 1.3. A spectrum may be preprocessed solely by a single operation or a sequence of operations.

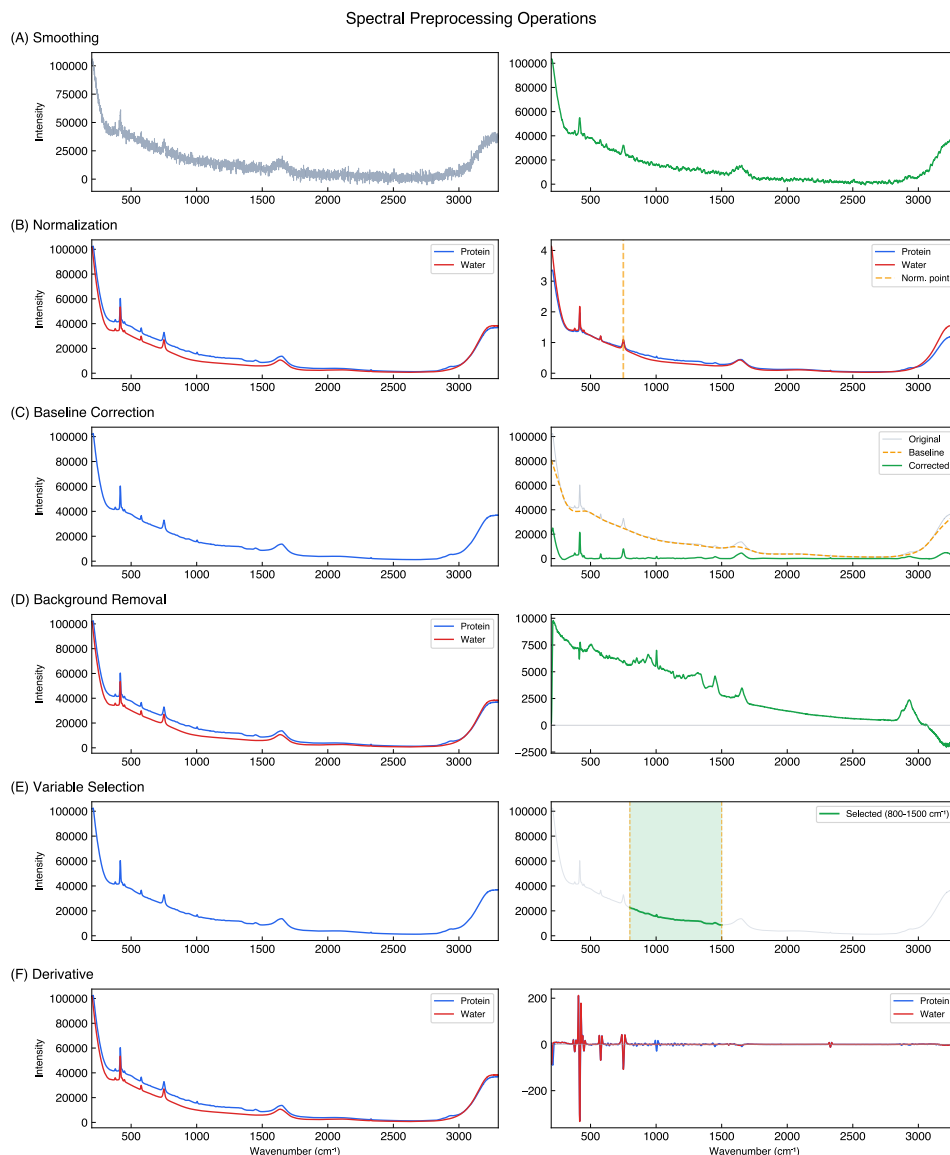


Figure 1.3 Illustration of the effect of common preprocessing operations using an exemplary Raman spectrum of protein at a concentration of 25 g/L. (A) Smoothing using a Savitzky–Golay (SG) filter. White noise has been introduced artificially to emphasize the effect of the SG filtering operation. (B) Normalization at 750 cm⁻¹ is used to overlay the protein spectrum with a water spectrum. (C) Baseline correction is performed using a Whittaker filter based on the iaspls algorithm. (D) Background removal is performed by subtracting the normalized water spectrum from the normalized protein spectrum. (E) Selection of the variables in the fingerprint region (800–1500 cm⁻¹) for protein analysis. (F) First derivative of the protein and water spectrum is performed using a SG filter to amplify spectral differences.

1.4.1 Smoothing

The quality of spectroscopic measurements is often compromised by electronic noise inherent to the detection system. Depending on the device configuration chosen for the measurement, the presence of noise may limit the detectability of the analyte (Gautam et al. 2015). Longer measurement

times usually lead to smoother spectral data. In real-time applications, the sampling rate of the sensor is a key performance indicator and determines the ability for real-time control actions (Rüdt, Briskot, and Hubbuch 2017). In those situations, smoothing techniques can be employed to reduce high-frequency noise. Methods such as moving average smoothing or SG filtering (Savitzky and Golay 1964) replace each data point with using information of neighboring points while preserving the overall spectral shape. While moving average filters replace each data point by the average of neighboring points in a predefined window, SG filters fit a polynomial function to the points in the predefined window and replace the queried data points by the function value at this location. Smoothing helps reduce electronic noise, but bares the potential of distorting fine spectral features corresponding to molecular information. Other methods based on so-called Whittaker filters (Eilers 2003) and deep learning approaches exist which intend to reduce the mentioned limitations (H. He et al. 2024).

1.4.2 Normalization

Spectroscopic measurements for PAT applications are usually conducted over a prolonged period of time or potentially moved from lab-scale to production using multiple different spectrometers or measurement equipment. As changes in the measurement setup, the changing environment or even day-to-day variation can lead to variations in absolute signal intensity, normalization is critical to ensure comparability of multiple data sources. Normalization techniques include vector normalization, area normalization, and standard normal variate (SNV) transformation (Mehmood et al. 2012). Normalization may be insufficient to make spectra recorded from different devices comparable. In this case, standardization approaches can be used to transfer measurements from a source to a target device (Nikzad-Langerodi, Lughofer, et al. 2018; Nikzad-Langerodi and Andries 2021).

1.4.3 Baseline Correction

Baseline drifts, resulting from instrument response or scattering effects, can introduce large variations across the spectral range. Baseline drifts occur in all different kinds of spectroscopy. In Raman spectroscopy of proteins, the high autofluorescence of the biological molecules causes non-linear baseline drifts which dominate the overall signal (Wen 2007). Baseline correction methods, such as polynomial fitting or asymmetric least squares, can be used to remove these trends to extract the relevant information from the spectra (Gautam et al. 2015). The challenges associated with algorithmic baseline correction are the differentiation between peak and non-peak regions which makes it also difficult to quantitatively evaluate whether or which baseline removal algorithm is the optimal choice (S. Guo, Bocklitz, and Popp 2016). Therefore, multiple novel algorithms have been developed to improve the capabilities to perform baseline correction in situations where fluorescence and noise dominate the spectral signals (Bocklitz et al. 2011; S. He et al. 2014; F. Zhang et al. 2020).

1.4.4 Background Removal

Similar to baseline correction, background subtraction removes contributions from solvent, buffer, or other interferents that are not part of the analyte signal (Gautam et al. 2015). While baseline correction is performed to extract the molecular information probed by the spectroscopic method, e.g. vibrational information in Raman spectroscopy, background correction aims to selectively remove the molecular information of certain interferents present in the analyte sample. While background correction can be performed by direct subtraction of a blank spectrum in samples with known composition, algorithmic approaches exist to remove the interferent contributions by

model-based scatter correction (Martens and Stark 1991a) or statistical regression (Trygg and S. Wold 2002; Valensise et al. 2020). In protein spectroscopy, background correction is helpful when measuring in complex matrices, such as cell culture supernatants or formulation buffers to isolate effects of specific components.

1.4.5 Variable Selection

Absorption or vibrational spectra usually carry a mixture of informative and uninformative variables. Spectral regions known to not contain any relevant information or to be corrupted by noise or interferences, may be discarded in order to improve the subsequent statistical analysis. Variable selection can be done algorithmically or manually by a multitude of different methods (Andersen and R. Bro 2010; Balabin and Smirnov 2011). When performed manually, the terms cropping or truncating are often used interchangeably. Manual selection commonly rests on prior knowledge about the spectral fingerprint of the analyte and involves the selection of wider spectral regions, like the 220–300 nm range for proteins in UV spectroscopy or the protein fingerprint region between 800–1800 cm^{-1} in Raman spectroscopy. When performed algorithmically, the importance of individual variables can be judged by custom metrics and subsequently used for selection. The iterative selection of variables is often tightly connected to the optimization of a regression or classification model because the importance of a single variable or a collection of variables, is usually judged by the predictive performance of the studied model. The most popular algorithms to iteratively select spectral variables are forward or backward feature elimination, iterative partial-least squares (iPLS) and genetic algorithms (Xiaobo et al. 2010; Cai, Y. Li, and Shao 2008; Nørgaard et al. 2000; P. Zhang et al. 2021). Since these algorithms are based on repeatedly evaluating model performance, their application to computationally more expensive models is not feasible. Thus, for modern ML models, variable selection needs to be directly integrated into the model training or calibration procedure (Paananen et al. 2019; Y.-y. Chen and Z.-b. Wang 2019).

1.4.6 Derivative Spectroscopy

Derivative preprocessing, often using SG derivatives, is a common tool across all spectroscopic techniques to emphasize spectral features and reduce baseline effects (Rüdt, Briskot, and Hubbuch 2017; Rolinger, Rüdt, and Hubbuch 2020a). First or second derivatives of the spectra can highlight subtle peaks or resolve overlapping bands, which is particularly useful in protein spectra where amide bands may be broad and partially overlapping. Derivative preprocessing, when combined with smoothing, provides enhanced feature resolution for subsequent multivariate modeling. Additionally, derivative spectroscopy is known to remove constant offsets and slope differences at first and second order, respectively.

Summary Preprocessing is therefore a critical step in spectroscopic data analysis. Each method addresses a particular type of spectral distortion, whether random noise, baseline drift, light scattering, or irrelevant variables. Properly preprocessed data enable multivariate chemometric models to operate under the assumptions of linearity and consistent spectral response, improving prediction accuracy and robustness. In the context of protein spectroscopy, preprocessing ensures that models can reliably extract quantitative information from complex, overlapping spectral features, supporting applications such as real-time monitoring of critical quality attributes in downstream processing.

1.5 Chemometrics

Chemometrics represents the interdisciplinary field that applies mathematical, statistical, and computational methods to extract meaningful chemical information from experimental data (Otto 2017, p. 1). The term, first coined by Svante Wold in the 1970s, encompasses the design of experiments, data collection strategies, and the development of mathematical models to understand and predict chemical phenomena (Wold and Sjöström 1977). In the context of analytical chemistry and process monitoring, chemometrics provides a toolbox of statistical methods to build the bridge between raw instrumental measurements and chemical information.

At its core, chemometrics is built upon the principles of linear regression, which forms the mathematical foundation for relating measured instrumental responses to chemical composition and properties of interest (Otto 2017, p. 214). Modern analytical instruments, particularly spectroscopic devices, routinely generate high-dimensional data arrays where the number of variables often exceeds the number of samples, presenting unique challenges that require extensions of basic univariate modeling approaches through multivariate methods (Otto 2017, p. 214).

The field encompasses several major categories of methods, each addressing specific analytical challenges. Linear regression provides the fundamental framework for quantitative relationships. Multivariate extensions address high-dimensional data through dimensionality reduction and latent variable approaches. Exploratory data analysis techniques reveal underlying data structures and identify outliers. Classification methods enable sample categorization, while advanced ML approaches capture non-linear relationships that extend beyond traditional linear assumptions.

In biopharmaceutical applications, the ability to extract specific information about critical quality attributes from noisy, overlapping spectroscopic signals through chemometric methods represents a key enabling technology for real-time process monitoring and control.

1.5.1 Linear Regression

Linear regression forms the mathematical foundation of chemometrics by establishing quantitative relationships between measured variables and properties of interest. Given a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ where each $\mathbf{x}_i$ represents a measured spectrum and y_i is the corresponding response variable (e.g., analyte concentration), linear regression models their relationship through a linear combination of the predictor variables.

For a single predictor variable, the linear model assumes:

$$y = \beta_0 + \beta_1 x + \varepsilon, \quad (1.6)$$

where β_0 is the intercept, β_1 is the slope parameter, and ε represents random error. This univariate framework directly reflects the Lambert–Beer law (cf. Eq. 1.2), where absorbance at a single wavelength is linearly proportional to concentration.

1.5.1.1 Multivariate Linear Regression

Spectroscopic measurements typically involve hundreds or thousands of wavelengths, necessitating extension to the multivariate case. For a dataset with n samples and p variables organized in the data matrix $\mathbf{X}$, multivariate linear regression assumes:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (1.7)$$

where $\mathbf{y}$ is the $n \times 1$ response vector, $\mathbf{X}$ is the $n \times p$ design matrix with rows representing samples and columns representing wavelengths, $\boldsymbol{\beta}$ is the $p \times 1$ parameter vector, and $\boldsymbol{\varepsilon}$ represents the error term.

The data matrix $\mathbf{X}$ has the structure:

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \quad (1.8)$$

where each row represents a spectrum and each column represents measurements at a specific wavelength across all samples.

To solve the multiple linear regression (MLR) for $\boldsymbol{\beta}$, Eq. 1.7 can be rewritten as (Otto 2017, p. 232)

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}. \quad (1.9)$$

This solution requires that $\mathbf{X}^T \mathbf{X}$ is invertible, which becomes problematic in spectroscopic applications where $p \gg n$ (more wavelengths than samples) or when multicollinearity exists among adjacent spectral variables (Otto 2017, p. 233). The high dimensionality and collinearity typical for spectral data require specialized approaches that build upon the MLR model while addressing their limitations. The methods discussed in the following sections represent systematic extensions of MLR to handle multidimensional collinear datasets.

1.5.2 Multivariate Data Analysis

Building upon the linear regression framework, multivariate data analysis provides systematic approaches to extract chemical information from high-dimensional spectroscopic datasets. The fundamental challenge lies in the fact that spectroscopic data violates many assumptions of classical linear regression: the number of variables typically exceeds the number of samples ($p \gg n$) and adjacent wavelengths exhibit strong correlation.

1.5.2.1 Classical Least Squares Regression

Classical least squares (CLS) regression represents the most direct approach to solve the collinearity problem, assuming that the spectral response is a linear combination of known pure component spectra (Miller 2007, p. 257). The CLS model can be expressed as

$$\mathbf{X} = \mathbf{C} \mathbf{S}^T + \mathbf{E}, \quad (1.10)$$

where $\mathbf{C}$ is an $n \times m$ concentration matrix for m chemical components, $\mathbf{S}$ is a $p \times m$ matrix containing the pure component spectra and $\mathbf{E}$ is the residual error. For a new sample with spectrum $\mathbf{x}$, the concentration vector $\mathbf{c}$ is determined by

$$\mathbf{c} = (\mathbf{S}^T \mathbf{S})^{-1} \mathbf{S}^T \mathbf{x}. \quad (1.11)$$

This approach requires that all spectrally active components are known and their pure spectra are available, which is often impractical in complex biological systems .

1.5.2.2 Principal Component Analysis

Principal component analysis (PCA) addresses the limitations of CLS by decomposing the data matrix into orthogonal components that capture the maximum variance in the dataset, rather than assuming spectral profiles *a priori*. The PCA model decomposes $\mathbf{X}$ as (Kessler 2007, p. 38)

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} = \sum_{a=1}^A \mathbf{t}_a \mathbf{p}_a^T + \mathbf{E}, \quad (1.12)$$

where $\mathbf{T}$ is the $n \times A$ score matrix, $\mathbf{P}$ is the $p \times A$ loading matrix, A is the number of principal components retained, $\mathbf{t}_a$ represents the a -th score vector, and $\mathbf{p}_a$ is the a -th loading vector.

The principal components are determined by sequential maximization of the variance captured, where all components have unit norm and are orthogonal to each other. The loadings $\mathbf{p}_a$ are typically determined via the nonlinear iterative partial least squares (NIPALS) algorithm (Kessler 2007, p. 39). The scores $\mathbf{t}_a$ represent the projection of each sample onto the principal component directions and provide a reduced-dimensional representation of the original data. The loadings describe the contribution of each original variable to the principal components and can often be interpreted in terms of underlying chemical phenomena. By breaking down high-dimensional spectral data to an A -dimensional space, PCA is an indispensable tool in the analysis spectroscopic and biomedical data.

1.5.2.3 Principal Component Regression

Principal component regression (PCR) combines the dimensionality reduction capabilities of PCA with linear regression by first decomposing the predictor matrix $\mathbf{X}$ into principal components according to Eq. 1.12, then regressing the response variable $\mathbf{y}$ against the selected components. The PCR model can be written as (Miller 2007, p. 260)

$$\mathbf{y} = \mathbf{T}\mathbf{q} + \mathbf{f} = \mathbf{X}\mathbf{P}\mathbf{q} + \mathbf{f}, \quad (1.13)$$

where $\mathbf{q}$ is the regression coefficient vector for the selected number of principal components and $\mathbf{f}$ is the residual vector. The regression coefficients in the original variable space are given by

$$\mathbf{b} = \mathbf{P}\mathbf{q}. \quad (1.14)$$

While PCR effectively addresses multicollinearity, it suffers from the limitation that the principal components are determined solely based on $\mathbf{X}$ without considering their relationship to $\mathbf{y}$ (Miller 2007, p. 262).

1.5.2.4 Partial Least Squares Regression

Partial least squares (PLS) regression overcomes the limitation of PCR by constructing latent variables that maximize the covariance between $\mathbf{X}$ and $\mathbf{Y}$. Here, the notation of target values is changed from a scalar vector to a matrix $\mathbf{Y}$ as the PLS can handle both scalar and multidimensional target spaces. The PLS model simultaneously decomposes both the predictor and response matrices according to (Otto 2017, p. 236)

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \quad (1.15)$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^T + \mathbf{F} \quad (1.16)$$

where $\mathbf{T}$ contains the X-score vectors, $\mathbf{U}$ the y-scores, $\mathbf{P}$ the X-loadings and $\mathbf{Q}^T$ contains the y-loadings. The residuals of the decomposition of $\mathbf{X}$ and $\mathbf{Y}$ are given by $\mathbf{E}$ and $\mathbf{F}$.

The key distinction from PCR lies in how the latent variables are constructed. In the NIPALS algorithm, X-weights $\mathbf{w}$ are computed by projecting the current y-scores onto the X-matrix, followed by normalization to unit length. These weights define how the original variables are combined to form the X-scores $\mathbf{t} = \mathbf{X}\mathbf{w}$. The algorithm then estimates the Y-loadings from these scores, updates the y-scores, and iterates until convergence. After deflation of both matrices by subtracting the explained variance, the process repeats for subsequent components. The weight vectors from each component are collected column-wise into the matrix $\mathbf{W}$.

The resulting regression model in the original variable space is given by

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{f}, \quad (1.17)$$

where the regression coefficient matrix $\mathbf{B}$ is given by

$$\mathbf{B} = \mathbf{W} (\mathbf{P}^T \mathbf{W})^{-1} \mathbf{Q}^T \quad (1.18)$$

and $\mathbf{W}$ is the matrix of X-weights from the PLS decomposition. The iterative maximization of the covariance between the X-scores and y can be achieved via different algorithms, of which the most commonly chosen ones are the NIPALS or straightforward implementation of a statistically inspired modification of partial least squares (SIMPLS) algorithms (S. De Jong 1993; S. Wold, M. Sjöström, and Eriksson 2001; Andersson 2009). In practice, PLS models can be calibrated as single-output (PLS1) or multi-output (PLS2) models, i.e., predicting one or multiple target variables. For uncorrelated target variables, PLS1 models typically yield more accurate predictions than PLS2 models. In the case of correlated target variables, PLS2 are usually superior to PLS1 models (Kessler 2007, p. 130). Further, the model accuracy in unseen test samples is dependent on the chosen number of principal components or latent variables. The determination of the optimal number of principal components or latent variables represents a critical aspect of multivariate calibration. This problem is more generally known as hyperparameter optimization (HPO) which will be covered in Ch. 1.5.4.

1.5.3 Machine Learning

Building upon the linear regression foundation established previously, ML represents an extension that relaxes the strict linearity assumptions while preserving the goal of learning mappings between predictor and target variables. The fundamental principle of all ML approaches lies in their ability to automatically learn a so-called mapping function given pairs of predictor and target variables. Per definition, ML aims to learn a functional expression f that suffices

$$\mathbf{Y} = f(\mathbf{X}) + \boldsymbol{\varepsilon}, \quad (1.19)$$

where $\mathbf{X}$ and $\mathbf{Y}$ are collections of predictor and target variables, respectively, and $\boldsymbol{\varepsilon}$ is the residual error. While linear regression assumes that f is a linear combination of the input variables, modern ML approaches do not constrain f to a specific parameterized function, allowing for the discovery of complex, non-linear relationships.

In the context of spectroscopy, the previously introduced PCR and PLS regression models represent extensions of linear regression that learn the principal components as new feature maps through covariance maximization and thus can be considered ML models. However, PCR and PLS are subjected to linearity and orthogonality constraints. Modern ML approaches can instead

model complex, nonlinear mappings (I. Goodfellow, Bengio, and Courville 2016; James et al. 2013) without any assumptions about the data that may be necessary when linearity breaks down due to concentration-dependent effects, matrix interactions, or instrumental nonlinearities (T. Chen, Morris, and Martin 2007; Cui and Fearn 2017).

The implementation of ML in chemometrics follows established principles of model development while addressing the specific challenges of spectroscopic data. The preprocessing methods discussed in Ch. 1.4 remain essential, though some advanced models can integrate preprocessing operations directly into the learning process (Jinchao Liu et al. 2017; Kazemzadeh et al. 2022). Proper dataset partitioning into training, validation, and test sets becomes critical for unbiased performance assessment, particularly given the typically limited size of spectroscopic calibration datasets.

In the following sections, two modern ML approaches that have demonstrated particular utility for spectroscopic modeling will be introduced, namely Gaussian process regression (GPR) and artificial neural networks (ANNs) (H. Chen, Bakshi, and Goel 2007; Cui and Fearn 2017; Cui and Fearn 2018). Both methods extend the linear regression framework in complementary ways—GPR through a Bayesian probabilistic approach and ANNs through hierarchical non-linear transformations.

1.5.3.1 Gaussian Process Regression

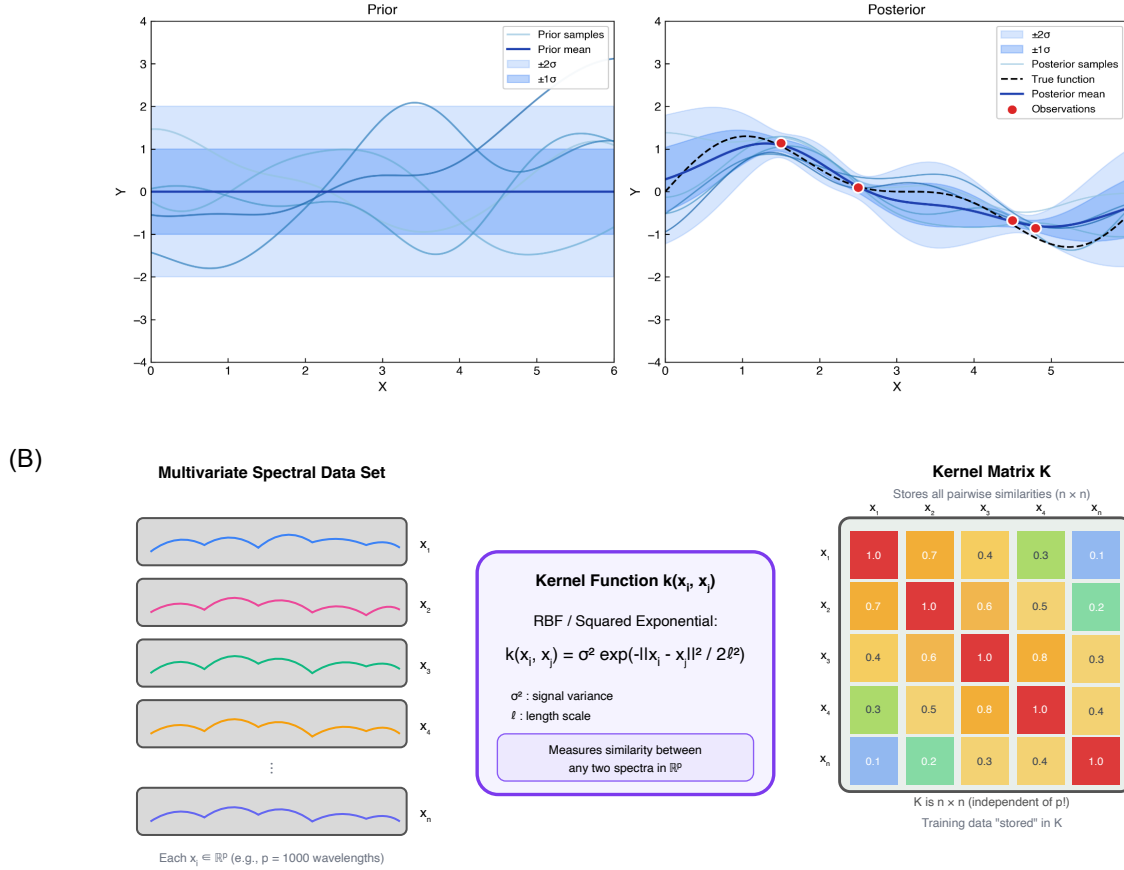


Figure 1.4 Schematic illustration of a GPR. (A) A multivariate spectral data set is stored in the kernel matrix $\mathbf{K}$ during the training process. The radial basis function (RBF) kernel is one of the most commonly used kernel functions. (B) A 1-dimensional example of a GPR using a RBF kernel. The displayed data points were generated from the true function $\sin(x) + 0.5 \sin(2x)$ with Gaussian white noise with a standard deviation of 0.2.

GPR takes a fundamentally different approach to regression by reasoning about entire functions rather than fitting a fixed functional form with adjustable parameters. Unlike parametric methods that seek optimal parameter values for a fixed functional form, GPR directly identifies which functions are plausible given the available data (Rasmussen and C. K. I. Williams 2006, p. 2).

This is achieved by specifying a prior distribution that assigns higher probability to functions with certain desirable properties, such as smoothness (Rasmussen and C. K. I. Williams 2006, p. 4). Before seeing any data, many different functions are considered plausible under this prior as illustrated in Fig. 1.4(A). When training data becomes available, functions that do not pass through (or near) the observed points are ruled out, leaving a posterior distribution over functions that are consistent with both the prior assumptions and the observations (Rasmussen and C. K. I. Williams 2006, p. 3). This is illustrated in Fig. 1.4 (A), where the shaded confidence region is wide in areas

without observations, reflecting high uncertainty, and narrows near the training points where the function is constrained by the data.

Formally, a gaussian process (GP) is specified as:

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')), \quad (1.20)$$

where $m(\mathbf{x})$ is the mean function, typically set to zero, and $k(\mathbf{x}, \mathbf{x}')$ is the covariance function or kernel. The kernel encodes the prior assumption that nearby inputs should produce similar outputs—the closer two points are, the more strongly correlated their function values (Rasmussen and C. K. I. Williams 2006, p. 14). This single design choice controls the smoothness, flexibility, and extrapolation behavior of the model.

For a dataset with training inputs $\mathbf{X}$ and outputs $\mathbf{y}$, and assuming Gaussian noise with variance σ_n^2 , the predictive distribution for a new input $\mathbf{x}_*$ is Gaussian with mean:

$$\mu(\mathbf{x}_*) = \mathbf{k}_*^T (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{y} \quad (1.21)$$

and variance:

$$\sigma^2(\mathbf{x}_*) = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^T (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{k}_*, \quad (1.22)$$

where $\mathbf{K}$ is the $n \times n$ covariance matrix with elements $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$, and $\mathbf{k}_*$ is the $n \times 1$ vector of covariances between the test point and training inputs. This relationship is illustrated in Fig. 1.4 (B). The kernel matrix $\mathbf{K}$ stores all training information as pairwise similarities, independent of the dimensionality of the input vectors.

Kernel Functions and Spectroscopic Applications The choice of kernel function fundamentally determines the model’s behavior. The linear kernel $k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T \mathbf{x}'$ exactly recovers linear regression. The RBF kernel, which is among the most commonly used kernels in practical applications (Rasmussen and C. K. I. Williams 2006, p. 83), is given by

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp \left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\ell^2} \right). \quad (1.23)$$

The RBF kernel enables smooth non-linear relationships by measuring the similarity of the two inputs $\mathbf{x}$ and $\mathbf{x}'$ via the Euclidean distance. The length scale ℓ determines the smoothness of the kernel function, while σ_f^2 controls the amplitude. Fig. 1.4 (A) presents an exemplary regression using the RBF kernel to recover a hypothetical periodical function while providing a confidence over the entire function space.

In spectroscopic applications, this example needs to be extended to the multivariate case. The kernel functions remain unchanged as they compress the multivariate measurements via the Euclidean distance. Multiple kernel functions can also be combined by forming sums or products or individual functions (T. Chen, Morris, and Martin 2007). To increase weight of individual spectral variables, individual length scales can be assigned to each wavelength, enabling automatic relevance determination that identifies the most informative spectral regions (T. Chen and Bo Wang 2010). This feature selection capability extends the variable selection concepts from linear regression to a fully automated, probabilistic framework.

1.5.3.2 Artificial Neural Networks

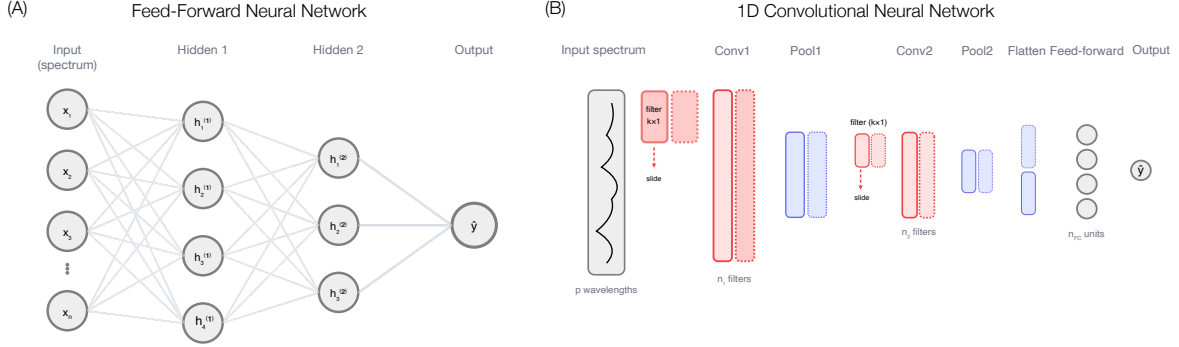


Figure 1.5 Schematic illustration of one-dimensional feed-forward (A) and convolutional neural networks (B).

ANNs represent a fundamentally different extension of linear regression that achieves flexibility through hierarchical composition of simple linear transformations followed by non-linear activation functions (I. Goodfellow, Bengio, and Courville 2016, p. 168). While linear regression directly computes $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$, ANNs create complex non-linear mappings by stacking multiple layers of these basic operations.

The connection between linear regression and ANNs becomes clear when considering a single-layer network without activation functions, which exactly reproduces linear regression. An ANN with one hidden layer and linear activation would compute

$$\mathbf{h} = \mathbf{X}\mathbf{W}_1, \quad \hat{\mathbf{y}} = \mathbf{h}\mathbf{W}_2 = \mathbf{X}\mathbf{W}_1\mathbf{W}_2. \quad (1.24)$$

This reduces to linear regression with parameter matrix $\boldsymbol{\beta} = \mathbf{W}_1\mathbf{W}_2$. The power of ANNs emerges from introducing non-linear activation functions and multiple layers, enabling the approximation of arbitrary continuous functions.

Feed-forward Neural Networks Feed-forward neural networks (FFNNs) are the basic building blocks for modern neural networks (I. Goodfellow, Bengio, and Courville 2016, p. 168). A single hidden layer network with input $\mathbf{x}$, weight matrices $\mathbf{W}_1, \mathbf{W}_2$, bias vectors $\mathbf{b}_1, \mathbf{b}_2$, and nonlinear activation function φ can be mathematically expressed by

$$\mathbf{h} = \varphi(\mathbf{X}\mathbf{W}_1 + \mathbf{b}_1), \quad \hat{\mathbf{y}} = \mathbf{h}\mathbf{W}_2 + \mathbf{b}_2. \quad (1.25)$$

Here, $\mathbf{h}$ represents the hidden layer activations that transform the original spectral features into a learned representation. Extending this to L hidden layers yields the general FFNN:

$$\mathbf{h}^{(l)} = \varphi^{(l)}(\mathbf{h}^{(l-1)}\mathbf{W}^{(l)} + \mathbf{b}^{(l)}), \quad l = 1, \dots, L \quad (1.26)$$

with $\mathbf{h}^{(0)} = \mathbf{X}$ and $\hat{\mathbf{y}} = \mathbf{h}^{(L)}$. In modern ML, FFNNs may be composed of an arbitrary number of layers L with an arbitrary number units per layer. This is exemplified in Fig. 1.5(A) for a FFNN with 2 hidden layers. The size of the weight matrices and thus the number of trainable weights which need to be estimated in the training process, increase with the depth and size of the neural

network. For an one-dimensional input with 1000 variables, a three-layer FFNN with 10 units per layer, amounts to 10230 trainable weights.

Convolutional Neural Networks Convolutional neural networks (CNNs) represent a specialized architecture that exploits the sequential nature of spectroscopic data. Where FFNNs treat each input variable independently, convolutional layers apply shared filters across the input dimension. The fundamental operation underlying these networks is the discrete convolution, which computes a weighted sum of local input values using a filter or kernel function. For a one-dimensional input signal $\mathbf{x}$ and filter $\mathbf{w}$, the convolution operation is defined as (I. Goodfellow, Bengio, and Courville 2016, p. 332)

$$(\mathbf{x} * \mathbf{w})[j] = \sum_{m=0}^{M-1} \mathbf{x}[j+m] \cdot \mathbf{w}[m], \quad (1.27)$$

where M is the filter size and the filter coefficients $\mathbf{w}[m]$ determine what features are detected. For example, a filter with weights $[-1, 0, 1]$ approximates a first derivative. The mathematical transformation through a single convolutional filter in a CNN may thus be written as (Prince 2023, p. 167)

$$h_j^{(l)} = \varphi \left(\sum_{m=0}^{M-1} w_m^{(l)} x_{j \cdot s + m} + b^{(l)} \right), \quad (1.28)$$

where $h_j^{(l)}$ is the output activation at position j in layer l , $w_m^{(l)}$ are the learnable filter weights of size M , $x_{j \cdot s + m}$ represents the input values accessed by the filter, s is the stride parameter controlling the step size between filter applications, and $b^{(l)}$ is the bias term. The filter moves across the input sequence by applying the same weights $w_m^{(l)}$ at different positions: for output position $j = 0$, the filter processes input positions $[0, 1, \dots, M-1]$; for $j = 1$ with stride $s = 1$, it processes $[1, 2, \dots, M]$; and generally for position j , it processes $[j \cdot s, j \cdot s + 1, \dots, j \cdot s + M - 1]$. This sliding window operation enables the network to detect the same spectral pattern regardless of its position in the spectrum, while the stride controls the degree of overlap between consecutive filter applications and determines the output size reduction. This parameter sharing reduces model complexity while enabling detection of local spectral features such as peaks or characteristic band shapes. The convolutional operation hence effectively implements learned spectral filters that can identify relevant patterns across different wavelength regions.

A convolutional layer in a CNN usually consists of multiple such filters and a typical CNN architecture consists of multiple convolutional layers followed by pooling layers after each convolutional layer (Prince 2023, p. 167). This alternating structure is illustrated in Fig. 1.5 (B). Pooling layers, commonly using max-pooling or average-pooling operations, reduce the spatial dimensions while retaining the most important features (I. Goodfellow, Bengio, and Courville 2016, p. 332). The max-pooling operation may be written as

$$h_j^{pool} = \max_{k \in P_j} h_k^{(l)} \quad (1.29)$$

where P_j represents the pooling region, $h_k^{(l)}$ is the output of the previous convolutional layer and h_j^{pool} is the pooled output which reduced dimension. For a pooling window of size 2, the number of variables is effectively halved. This combines the pattern recognition ability with a successive

dimensionality reduction. This down-sampling operation provides translation invariance and reduces computational complexity (Prince 2023, p. 164).

After the convolutional and pooling layers, the feature maps are typically flattened into a one-dimensional vector before being fed to layers of a FFNN (I. Goodfellow, Bengio, and Courville 2016, p. 346). The flattening operation reshapes the outputs of multiple convolutional filters into a single vector by

$$h^{flat} = \text{flatten}(h^{conv}). \quad (1.30)$$

Training To train neural network models, i.e., establish a correlation between predictor and target variables through the previously introduced mathematical transformations, the adjustable weights of each of these transformations need to be optimized. To do so, neural networks require iterative optimization through stochastic gradient descent (I. Goodfellow, Bengio, and Courville 2016, p. 294).

The optimization process minimizes a loss function that quantifies the discrepancy between predictions and target values. For regression problems, the mean squared error is commonly used as a metric in the so-called *loss* function which is given by

$$L = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2, \quad (1.31)$$

where N is the number of samples being evaluated. Gradient descent algorithms require the calculation of the gradient with regard to every adjustable weight which is usually referred to as *backpropagation*. Backpropagation requires the individual layers to be mathematically differentiable to be able to use the chain rule of calculus. This makes the calculation computationally efficient and enables the optimization of networks with millions of parameters (I. Goodfellow, Bengio, and Courville 2016). Each weight w is then updated in the direction that reduces the loss according to

$$w \leftarrow w - \eta \frac{\partial L}{\partial w}, \quad (1.32)$$

where η is the learning rate. In stochastic gradient descent, the computation of the gradient is usually done in minibatches (Prince 2023, p. 85), which makes the individual calculations more efficient and the gradient estimation more robust on average as noise is introduced by choosing a subset of the entire training data set. More sophisticated optimizers such as Adam (Kingma and Ba 2015) additionally compute estimates of gradient moments and adapt the effective learning rate for each parameter individually (Prince 2023, p. 90).

Training typically processes data in mini-batches rather than individual samples or the entire dataset, balancing computational efficiency with gradient estimate quality (Ioffe and Szegedy 2015). Each mini-batch undergoes a forward pass to compute predictions and loss, followed by a backward pass to compute gradients, after which the weights are updated. One complete pass through all training data, processing all mini-batches sequentially, constitutes a *training epoch*. For a training set of N samples and mini-batch size B , this results in $\lceil N/B \rceil$ weight updates per epoch.

After each epoch, model performance is evaluated on a held-out validation set. These validation metrics do not directly influence weight updates but rather serve an observational purpose. They monitor generalization performance and may inform early stopping decisions when the validation

loss is not improving anymore. This separation ensures that validation data provides an unbiased estimate of model performance on unseen data. Training continues for multiple epochs until convergence criteria are met, such as a maximum number of epochs or early stopping triggered by a validation loss plateau.

Regularization Modern neural network training often incorporates regularization techniques. L2 regularization adds a penalty term to the loss function which penalizes large weight values as given by (Prince 2023, p. 140)

$$L_{reg} = L + \lambda \sum_{l=1}^L \|\mathbf{W}^{(l)}\|_2^2, \quad (1.33)$$

where λ controls the strength of regularization. This encourages the network to learn smoother functions (Ng 2004).

Dropout is another regularization technique that randomly sets a fraction of hidden units to zero during training (Srivastava et al. 2014). For a hidden layer with dropout rate p , each unit is retained with probability $(1 - p)$. This improves generalization by forcing the network to learn redundant representations (Srivastava et al. 2014).

Early stopping, as described above, also serves as an implicit regularization strategy by halting training before the model overfits to the training data.

1.5.4 Enabling Methodologies

Several methodological advances have significantly expanded the applicability and effectiveness of ML approaches. These enabling technologies address fundamental challenges in model development, optimization, and deployment that have historically limited the adoption of advanced ML methods.

Data Augmentation

Data augmentation techniques address the persistent challenge of limited training data availability by synthetically generating additional training examples. In order for these synthetically generated data to benefit the model training, they require to preserve the underlying statistical properties of the original dataset (I. Goodfellow, Bengio, and Courville 2016, p. 240). This approach is particularly valuable in applications where collecting large datasets may be expensive, time-consuming, or technically challenging. Data augmentation methods generate new training samples through systematic transformations of existing data, effectively increasing the size and diversity of the training set without requiring additional experimental measurements.

Common augmentation strategies include additive noise injection, and transformations that simulate realistic variations in the analyzed data. Geometric transformations such as shifting, scaling, or warping are common tools in augmenting data sets used for training image recognition or natural language processing (I. Goodfellow, Bengio, and Courville 2016; Feng et al. 2021, p. 240). For spectroscopic data, these techniques are not directly applicable they do not create realistic representations of spectral features which must be proportional to the underlying chemical species. The effectiveness of data augmentation depends critically on the realism of the generated variations. Augmentation strategies must reflect genuine sources of variability that the model will encounter during deployment while avoiding the introduction of unrealistic artifacts that could degrade model performance. Proper validation procedures are essential to ensure that augmented data improves rather than harms generalization capability.

Automated Hyperparameter Optimization

The progression from simple linear models to complex machine learning approaches is commonly accompanied by a larger set of so-called *hyperparameters*, meaning configurable parameters that are not learned during training. These include architecture design choices in ANNs or kernel parameters in GPRs (Hutter, Kotthoff, and Vanschoren 2019, p. 3). While traditional methods like linear regression or PLS typically require tuning only a few parameters, such as the polynomial degree or number of latent variables, GPRs and ANNs introduce substantially more complex parameter spaces that benefit from systematic optimization (Hutter, Kotthoff, and Vanschoren 2019, p. 4). Automated hyperparameter optimization formalizes this challenge as minimizing an objective function $f(\boldsymbol{\lambda})$ over the hyperparameter space Λ as

$$\boldsymbol{\lambda}^* = \arg \min_{\boldsymbol{\lambda} \in \Lambda} f(\boldsymbol{\lambda}), \quad (1.34)$$

where f typically represents a model performance metric such as the residual error on a data set $\mathcal{D}$, and $\boldsymbol{\lambda}$ denotes the hyperparameter configuration. A detailed overview of model performance metrics is provided in Ch. 1.6.

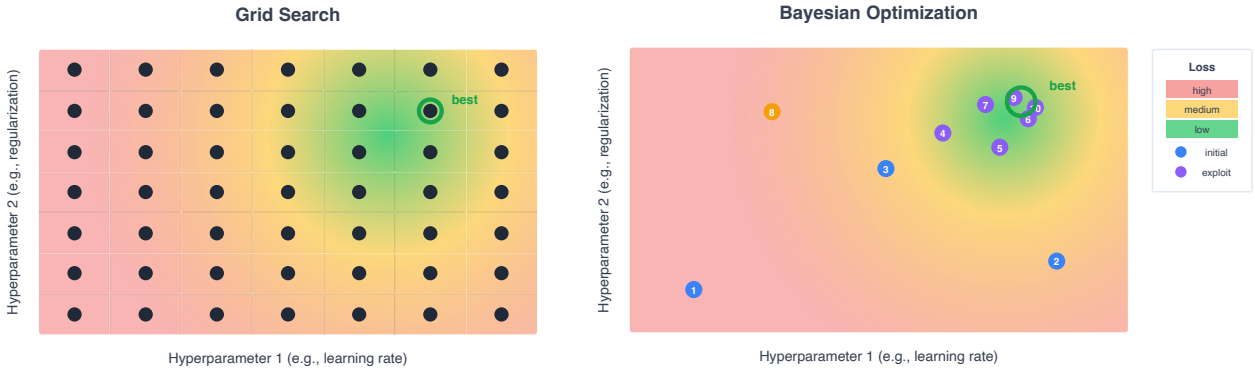


Figure 1.6 Comparison of grid search and Bayesian optimization for HPO. Grid search evaluates every combination of hyperparameters and selects the one with the lowest loss. Bayesian optimization instead builds a surrogate model of the objective function after each evaluation, using it to guide subsequent sampling through an exploration–exploitation trade-off. Here, Bayesian optimization locates the optimal region after only 10 function evaluations.

Two principal strategies exist for hyperparameter search: exhaustive methods such as grid or random search, and surrogate model-based optimization. Grid search exhaustively evaluates all hyperparameter combinations and selects the best according to the chosen criterion (Hutter, Kotthoff, and Vanschoren 2019, p. 7). This approach is only feasible when function evaluations are computationally cheap and the search space is small. In deep learning, however, model training is expensive and the search space grows rapidly due to the large number of design choices. Surrogate model-based optimization addresses this by approximating the objective function with a cheaper-to-evaluate model (Hutter, Kotthoff, and Vanschoren 2019, p. 9). The most widely used technique in this category is Bayesian optimization, which balances exploration of uncertain regions with exploitation of promising ones (Hutter, Kotthoff, and Vanschoren 2019, p. 9). As illustrated in Fig. 1.6, Bayesian optimization can identify optimal hyperparameter regions with far fewer evaluations than grid search.

1.6 Machine Learning Modeling Workflow

The successful application of ML methods to spectroscopic data requires a systematic workflow that integrates the concepts discussed in the previous sections. This section outlines the essential steps for developing robust predictive models, from initial data generation through final model deployment and interpretation. Understanding this workflow is crucial for ensuring reproducible results and avoiding common pitfalls that can compromise model performance and reliability.

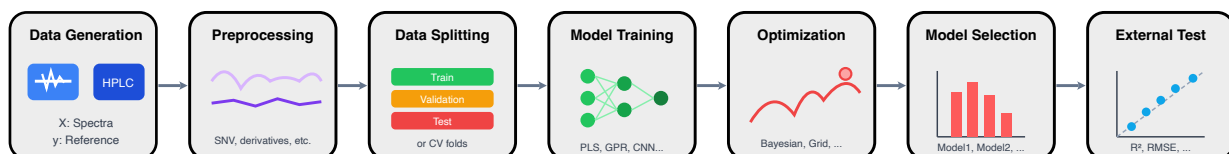


Figure 1.7 Visualization of a generalized ML workflow.

1.6.1 Data Generation and Experimental Design

The foundation of any successful ML model lies in proper experimental design and systematic data collection. In spectroscopic applications, data generation typically involves conducting multiple experiments or measurement campaigns under controlled conditions to capture the relationship between spectroscopic measurements and the target properties of interest.

Data collection often occurs in batches, where each batch represents a distinct experimental campaign conducted under specific conditions. These batches might correspond to different operating conditions or different starting materials. The systematic variation between batches introduces important considerations for model development, as models must be robust enough to generalize across these naturally occurring variations while remaining sensitive to the underlying chemical relationships. The experimental space should be designed to cover the expected range of conditions that the model will encounter during deployment. This includes systematic variation in target concentrations, environmental conditions, and other relevant factors that may affect both the spectroscopic measurements and the target properties. Ideally, the design should incorporate sufficient replicates to characterize measurement precision and include samples that represent edge cases or potential failure modes of the analytical system.

1.6.2 Data Splitting Strategies

Proper partitioning of data into training, validation, and test sets represents a fundamental aspect of robust model development that directly impacts the reliability of performance estimates. The choice of splitting strategy should consider both the statistical requirements of the modeling approach and the practical constraints of the experimental system.

1.6.2.1 Hold-out Validation

The most straightforward approach involves randomly partitioning the dataset into distinct subsets, typically with proportions of 60% for training, 20% for validation, and 20% for testing (Raschka 2018). The training set is used for parameter estimation and model fitting, the validation set guides

hyperparameter optimization and model selection decisions, and the test set provides an unbiased estimate of final model performance on previously unseen data.

Random splitting assumes that all samples are independent and identically distributed, which may not hold when data is collected in batches or exhibits systematic structure. In such cases, stratified splitting ensures that each subset maintains representative distributions of both target variables and potential confounding factors such as batch identity or measurement conditions.

1.6.2.2 Cross-Validation

Cross-validation provides a systematic approach for assessing model performance and making modeling decisions while maximizing the utilization of available data. By repeatedly training and testing models on different subsets of the training data, cross-validation provides more robust performance estimates than single train-test splits, particularly when dataset sizes are limited (Raschka 2018). Different cross-validation schemes exist, some of which are illustrated in Fig. 1.8. The different schemes can generally be distinguished based on the number of iterations they use to rotate the data set between training and validation assignments if whether they maintain the group structure of the training data.



Figure 1.8 Visualization of different cross-validation schemes. (A) K-fold cross-validation partitions the training data into k approximately equal subsets, or folds. The model is trained on $k - 1$ folds and tested on the remaining fold, with this process repeated k times so that each fold serves as the test set exactly once. (B) Stratified cross-validation extends the basic k-fold approach by ensuring that each fold maintains the same distribution of target variables as the complete training set. (C) When data exhibits a group structure, such as multiple batches from experimental campaigns, group-based cross-validation ensures that all measurements from the same group are assigned to the same fold. (D) Nested cross-validation provides a two-level cross-validation with an outer loop for repeated performance estimation and an inner loop for hyperparameter optimization.

K-Fold Cross-Validation K-fold cross-validation partitions the training data into k approximately equal subsets, or folds. The model is trained on $k - 1$ folds and tested on the remaining fold,

with this process repeated k times so that each fold serves as the test set exactly once. The final performance estimate is obtained by averaging the results across all folds.

The choice of k involves a bias-variance tradeoff. Smaller values of k such as $k = 5$ provide computationally efficient estimates but may exhibit higher variance due to the limited size of each training set (Raschka 2018). Larger values such as $k = 10$ reduce this variance but increase computational cost. Leave-one-out cross-validation ($k = n$) minimizes bias but can be computationally expensive for large datasets and may provide poor estimates when strong correlations exist between samples.

Stratified Cross-Validation Stratified cross-validation extends the basic K-fold approach by ensuring that each fold maintains the same distribution of target variables as the complete training set. This is particularly important for classification problems with unbalanced classes. Stratification helps ensure that each fold provides a representative sample for both training and testing, leading to more stable performance estimates.

Group-Based Cross-Validation When data exhibits a group structure, such as multiple batches from experimental campaigns, group-based cross-validation ensures that all measurements from the same group are assigned to the same fold. This prevents information leakage between training and test sets that could lead to overly optimistic performance estimates. In process monitoring, group-based cross-validation is the preferred choice as the purpose of the deployment model is to provide robust estimates of the current state in new batches or experimental campaigns (Rolinger, Rüdte, and Hubbuch 2020a).

Nested Cross-Validation When both hyperparameter optimization and model performance estimation are required, nested cross-validation provides an efficient technique employing two levels of cross-validation: an outer loop for performance estimation and an inner loop for hyperparameter optimization.

In the outer loop, the data may be partitioned into training and test sets via a group-based scheme. For each outer fold, the training portion undergoes complete model development, including hyperparameter optimization via an inner cross-validation loop. The optimized model is then evaluated on the held-out outer fold, providing an unbiased estimate of performance for that particular train-test split.

1.6.3 Model Evaluation and Performance Metrics

Comprehensive model evaluation requires multiple complementary metrics that assess different aspects of predictive performance. The choice of metrics should align with the specific requirements of the analytical application and the nature of the target variable.

1.6.3.1 Regression Metrics

For regression problems, several standard metrics provide different perspectives on model performance. The root mean squared error (RMSE) quantifies the average magnitude of prediction errors and is given by

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}, \quad (1.35)$$

where y_i represents the true values, $\hat{y}_i$ the predicted values, and n the number of samples. RMSE is sensitive to outliers and provides error estimates in the same units as the target variable.

The mean absolute error (MAE) offers a more robust alternative that is less sensitive to outliers

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|. \quad (1.36)$$

The MAE provides an intuitive measure of average prediction error magnitude and is particularly useful when outliers are present in the data.

The coefficient of determination R^2 indicates the proportion of variance in the target variable explained by the model according to

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \quad (1.37)$$

where $\bar{y}$ is the mean of the true values. R^2 ranges from 0 to 1, with values closer to 1 indicating better model fit.

For concentration prediction applications, the mean absolute percentage error (MAPE) provides relative error that is independent of the absolute concentration, which makes it particularly useful for comparing model performance across different concentration ranges. The MAPE is given by

$$\text{MAPE} = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right|. \quad (1.38)$$

1.6.3.2 Cross-Validation Performance Assessment

When using cross-validation, these metrics are calculated for each fold and then aggregated to provide overall performance estimates. For k-fold cross-validation, the cross-validated RMSE (RMSECV) is calculated as

$$\text{RMSECV} = \sqrt{\frac{1}{n} \sum_{k=1}^K \sum_{i \in \text{fold}_k} (y_i - \hat{y}_{i,-k})^2}, \quad (1.39)$$

where $\hat{y}_{i,-k}$ represents the prediction for sample i using a model trained without fold k . The same approach applies to other metrics, providing cross-validated estimates of MAE, R^2 , MAPE.

1.6.4 Model Interpretation and Validation

Model interpretability becomes particularly important in analytical applications where understanding the basis for predictions is essential for method validation and regulatory acceptance. The ability to explain model decisions not only builds trust in the predictions but also provides insights into the underlying chemical relationships and helps identify potential model failures or limitations.

1.6.4.1 Feature Importance

In spectroscopy, feature importance refers to the importance of an individual variable or a group of variables. Importance estimates may either be directly obtained from the employed models or computed via specific methods. Linear methods provide direct feature importance through regression coefficients that indicate the contribution of each spectral variable to the prediction. In PLS, the regression coefficients reveal which wavelengths or spectral regions are most important for predicting the target variable. Additionally, the loadings provide additional insight into how each latent variable relate to the original spectral features, enabling identification of characteristic spectral signatures (Kessler 2007, p. 117).

More complex ML models like neural networks do not provide informative regression coefficients due to their hierarchical structure and nonlinear transformations. ML models therefore require specialized interpretation techniques that can reveal which spectral features contribute most to individual predictions or overall model behavior. Several approaches have been developed to address this challenge, each with distinct advantages and limitations.

Permutation Importance Permutation importance assesses feature relevance by measuring how much model performance degrades when individual features are randomly shuffled while keeping other features unchanged (Štrumbelj and Kononenko 2014a). This model-agnostic approach works by breaking the relationship between a feature and the target variable, allowing assessment of that feature’s contribution to predictive performance.

SHAP Values Shapley additive explanations (SHAP) provide a unified framework for interpreting model predictions by assigning each feature an importance value for a particular prediction (S. M. Lundberg, Allen, and Lee 2017). For spectroscopic applications, SHAP values can reveal which wavelengths contribute positively or negatively to individual predictions, enabling identification of characteristic spectral features. The additive nature of SHAP values means that the sum of all feature contributions equals the difference between the prediction and the expected model output, providing a complete explanation of the prediction.

Gradient-Based Methods Gradient-based attribution methods compute the gradient of the output with respect to the input features, indicating how sensitive the prediction is to changes in each input. Methods such as integrated gradients and guided backpropagation have been developed to address limitations of simple gradient calculations, such as saturation effects and noisy attributions (Selvaraju et al. 2020). For convolutional neural networks applied to spectroscopic data, gradient-based methods can highlight which wavelength regions the network focuses on when making predictions. However, these methods can be sensitive to model architecture and may produce unstable attributions for similar inputs.

1.6.4.2 External Validation and Robustness Assessment

External validation using independently collected data represents the gold standard for demonstrating model generalizability. When possible, validation should include data collected under different conditions to assess model robustness. Any significant performance degradation in external validation compared to cross-validation results indicates potential overfitting or insufficient representation of process variability in the training data. The validation process should also assess model behavior at the boundaries of the training space and identify regions where predictions may be unreliable. Applicability domain assessment helps define the conditions under which the model can be expected to provide reliable predictions, supporting appropriate deployment and usage guidelines.

1.7 Combination of Spectroscopic Sensors with Process Models

While chemometric models enable the extraction of quantitative information from spectroscopic data, their predictions are inherently subject to noise and may lack robustness when applied outside the calibration domain. In parallel, mechanistic process models derived from mass balances and reaction kinetics can provide physically meaningful predictions of process dynamics, yet often suffer from plant-model mismatches due to simplifying assumptions or poorly characterized parameters (Gargalo et al. 2021). The combination of sensor measurements with process models offers a promising strategy to leverage the complementary strengths of both approaches: sensors provide real-time

feedback from the actual process, while models encode prior knowledge about the underlying physicochemical phenomena (Luttmann et al. 2012).

Soft sensors represent a class of inferential estimators that combine hardware sensor signals with mathematical models to estimate process variables that are difficult or impossible to measure directly (Luttmann et al. 2012; Kadlec, Gabrys, and Strandt 2009). In the context of biopharmaceutical manufacturing, soft sensors have been applied to estimate biomass concentrations, substrate levels, and product titers from readily available on-line measurements such as temperature, pH, dissolved oxygen, or spectroscopic signals (Luttmann et al. 2012). Soft sensors can be broadly categorized into data-driven and model-driven approaches, where the former rely on empirical correlations derived from historical data and the latter incorporate mechanistic understanding of the process (Kadlec, Gabrys, and Strandt 2009).

Optimal state estimation algorithms, originally developed for navigation and control applications, provide a principled framework for fusing noisy measurements with process model predictions (Simon 2006). The Kalman filter and its extensions compute a statistically optimal estimate of unmeasured process states by recursively combining sensor observations with model forecasts, weighting each source according to its associated uncertainty (Simon 2006). For non-linear systems, the extended Kalman filter (EKF) approximates the state transition and measurement functions through linearization, enabling its application to biochemical reaction systems that exhibit inherently non-linear dynamics (Haseltine and Rawlings 2005). Alternative formulations such as the unscented Kalman filter or particle filters have been developed to improve estimation accuracy for highly non-linear models, though they come at increased computational cost (Stelzer, Kager, and Herwig 2017).

In biopharmaceutical applications, Kalman filter-based soft sensors have been demonstrated for monitoring cell culture processes (Narayanan et al. 2020; Ohadi, Legge, and Budman 2015), capture chromatography (Feidl et al. 2019), and chemical conjugation reactions (Tuveri et al. 2021). A key advantage of these hybrid approaches is their ability to update not only the directly measured states but also unmeasured variables that are coupled through the model equations. Furthermore, the explicit treatment of measurement and process uncertainties enables the dynamic adjustment of confidence in sensor versus model predictions as process conditions evolve (Narayanan et al. 2020). The integration of uncertainty-aware chemometric models, such as GPR, with mechanistic process models through state estimation frameworks represents a promising direction for robust real-time monitoring in pharmaceutical manufacturing (Tulsyan et al. 2021).

Chapter references

- Abu-Absi, Nicholas R. et al. (05/2011). „Real time monitoring of multiple parameters in mammalian cell culture bioreactors using an in-line Raman spectroscopy probe“. In: *Biotechnology and Bioengineering* 108.5, pp. 1215–1221.
- Agarwal, Paresh and Carolyn R Bertozzi (2015). „Site-specific antibody-drug conjugates: The nexus of bioorthogonal chemistry, protein engineering, and drug development“. In: *Bioconjugate Chemistry* 26.2, pp. 176–192.
- Andersen, C. M. and R. Bro (11/2010). „Variable selection in regression-a tutorial“. In: *Journal of Chemometrics* 24.11-12, pp. 728–737.
- Andersson, Martin (2009). „A comparison of nine PLS1 algorithms“. In: *Journal of Chemometrics: A Journal of the Chemometrics Society* 23.10, pp. 518–529.
- Andris, Sebastian and Jürgen Hubbuch (2020). „Modeling of hydrophobic interaction chromatography for the separation of antibody-drug conjugates and its application towards quality by design“. In: *Journal of Biotechnology* 317.April, pp. 48–58.
- Atkins, Peter and Julio de Paula (2006). *Physical Chemistry*. Oxford, GB: Oxford University Press.
- Bakeev, Katherine A (2005). *Process Analytical Technology*. Oxford, GB: Blackwell Publishing.
- Balabin, Roman M. and Sergey V. Smirnov (04/2011). „Variable selection in near-infrared spectroscopy: Benchmarking of feature selection methods on biodiesel data“. In: *Analytica Chimica Acta* 692.1-2, pp. 63–72.
- Banner, Matthew et al. (12/2021). „A decade in review: use of data analytics within the biopharmaceutical sector“. In: *Current Opinion in Chemical Engineering* 34, p. 100758.
- Barth, Andreas (2007). „Infrared Spectroscopy of Proteins“. In: *Biochimica et Biophysica Acta - Bioenergetics* 1767.9, pp. 1073–1101.
- Barth, Andreas and Christian Zscherp (11/2002). „What Vibrations Tell about Proteins“. In: *Quarterly Reviews of Biophysics* 35.4, pp. 369–430.
- Basalla, Andrew J. and Brent S. Kendrick (12/2023). „Correcting Ultraviolet-Visible Spectra for Baseline Artifacts“. In: *Journal of Pharmaceutical Sciences* 112.12, pp. 3240–3247.
- Bazylewski, Paul, Ranjith Divigalpitiya, and Giovanni Fanchini (01/2017). „In situ Raman spectroscopy distinguishes between reversible and irreversible thiol modifications in L-cysteine“. en. In: *RSC Advances* 7.5, pp. 2964–2970.
- Beer’s Law* (02/2023). [Online; accessed 2025-12-24].
- Bocklitz, Thomas et al. (2011). „How to pre-process Raman spectra for reliable and stable models?“ In: *Analytica Chimica Acta* 704.1-2, pp. 47–56.
- Boulet-Audet, Maxime, Bernadette Byrne, and Sergei G. Kazarian (10/2014). „High-Throughput Thermal Stability Analysis of a Monoclonal Antibody by Attenuated Total Reflection FT-IR Spectroscopic Imaging“. In: *Analytical Chemistry* 86.19, pp. 9786–9793.
- Brestrich, Nina, Tobias Hahn, and Jürgen Hubbuch (03/2016). „Application of spectral deconvolution and inverse mechanistic modelling as a tool for root cause investigation in protein chromatography“. In: *Journal of Chromatography A* 1437, pp. 158–167.
- Brewster, Victoria L., Lorna Ashton, and Royston Goodacre (08/2011). „Monitoring the Glycosylation Status of Proteins Using Raman Spectroscopy“. In: *Analytical Chemistry* 83.15, pp. 6074–6081.
- Buckley, Kevin and Alan G. Ryder (2017). „Applications of Raman Spectroscopy in Biopharmaceutical Manufacturing: A Short Review“. In: *Applied Spectroscopy* 71.6, pp. 1085–1116.
- Bulcha, Jote T. et al. (02/08/2021). „Viral Vector Platforms within the Gene Therapy Landscape“. In: *Signal Transduction and Targeted Therapy* 6.1, p. 53.
- Butler, Holly J. et al. (2016). „Using Raman Spectroscopy to Characterize Biological Materials“. In: *Nature Protocols* 11.4, pp. 664–687.
- Cai, Wensheng, Yankun Li, and Xueguang Shao (2008). „A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra“. In: *Chemometrics and Intelligent Laboratory Systems* 90.2, pp. 188–194.
- Carta, Giorgio and Alois Jungbauer (2010). *Protein Chromatography - Process Development and Scale-Up*. Weinheim, Germany: Wiley-VCH Verlag GmbH und Co. KGaA.
- Castelli, María Sofia, Paul McGonigle, and Pamela J. Hornby (12/2019). „The Pharmacology and Therapeutic Applications of Monoclonal Antibodies“. In: *Pharmacology Research & Perspectives* 7.6, e00535.
- Chackerian, Bryce (2007). „Virus-like particles: Flexible platforms for vaccine development“. In: *Expert Review of Vaccines* 6.3, pp. 381–390.
- Chen, Hongshu, Bhavik R. Bakshi, and Prem K. Goel (2007). „Toward Bayesian chemometrics-A tutorial on some recent advances“. In: *Analytica Chimica Acta* 602.1, pp. 1–16.
- Chen, Tao, Julian Morris, and Elaine Martin (2007). „Gaussian process regression for multivariate spectroscopic calibration“. In: *Chemometrics and Intelligent Laboratory Systems* 87.1, pp. 59–71.
- Chen, Tao and Bo Wang (2010). „Bayesian variable selection for Gaussian process regression: Application to chemometric calibration of spectrometers“. In: *Neurocomputing* 73.13-15, pp. 2718–2726.
- Chen, Yuan-yuan and Zhi-bin Wang (08/2019). „Feature selection based convolutional neural network pruning and its application in calibration modeling for NIR spectroscopy“. In: *Chemometrics and Intelligent Laboratory Systems* 191, pp. 103–108.
- Cui, Chenhao and Tom Fearn (02/2017). „Comparison of partial least squares regression, least squares support vector machines, and Gaussian process regression for a near infrared calibration“. en. In: *Journal of Near Infrared Spectroscopy* 25.1 (1), pp. 5–14.
- Cui, Chenhao and Tom Fearn (11/2018). „Modern practical convolutional neural networks for multivariate regression: Applications to NIR calibration“. In: *Chemometrics and Intelligent Laboratory Systems* 182, pp. 9–20.

Machine Learning in Spectroscopy

- Effio, Christopher Ladd and Jürgen Hubbuch (05/2015). „Next generation vaccines and vectors: Designing downstream processes for recombinant protein-based virus-like particles“. In: *Biotechnology Journal* 10.5, pp. 715–727.
- Eilers, Paul H. C. (2003). „A Perfect Smoother“. In: *Analytical Chemistry* 75.14, pp. 3631–3636.
- EMA (2014). „Guideline on the Use of near Infrared Spectroscopy by the Pharmaceutical Industry and the Data Requirements for New Submissions and Variations“. In: *EMA/CHMP/CVMP/QWP/17760/2009 Rev1*.
- FDA (2004). „Guidance for Industry: PAT—a framework for innovative pharmaceutical development, manufacturing, and quality assurance“. In.
- Feidl et al. (10/2019). „Combining Mechanistic Modeling and Raman Spectroscopy for Monitoring Antibody Chromatographic Purification“. In: *Processes* 7.10, pp. 683–683.
- Feng, Steven Y. et al. (2021). „A Survey of Data Augmentation Approaches for NLP“. In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 968–988.
- Flatman, Stephen et al. (03/2007). „Process analytics for purification of monoclonal antibodies“. In: *Journal of Chromatography B* 848.1, pp. 79–87.
- Fu, Zhiwen et al. (03/22/2022). „Antibody Drug Conjugate: The “Biological Missile” for Targeted Cancer Therapy“. In: *Signal Transduction and Targeted Therapy* 7.1, p. 93.
- Gargalo, Carina L. et al. (2021). „Towards the Development of Digital Twins for the Bio-manufacturing Industry“. In: *Digital Twins: Tools and Concepts for Smart Biomanufacturing*. Ed. by Christoph Herwig, Ralf Pörtner, and Johannes Möller. Cham: Springer International Publishing, pp. 1–34.
- Gaiglitz, Günter and David S. Moore (2014). *Handbook of Spectroscopy*. Ed. by Gaiglitz Günter and David S. Moore. Vol. 4. Weinheim: Wiley-VCH.
- Gautam, Rekha et al. (2015). „Review of multidimensional data processing approaches for Raman and infrared spectroscopy“. In: *EPJ Techniques and Instrumentation* 2.1.
- Gillespie, Christopher et al. (2022). „Systematic assessment of process analytical technologies for biologics“. en. In: *Biotechnology and Bioengineering* 119.2, pp. 423–434.
- Glasse, Jarka et al. (04/2011). „Process analytical technology (PAT) for biopharmaceuticals“. In: *Biotechnology Journal* 6.4, pp. 369–377.
- Goldrick, Stephen, David Lovett, et al. (2018). „Influence of incident wavelength and detector material selection on fluorescence in the application of raman spectroscopy to a fungal fermentation process“. In: *Bioengineering* 5.4, pp. 1–17.
- Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. MIT Press.
- Gunter Jagschies et al. (01/2018). *Biopharmaceutical Processing : Development, Design, and Implementation of Manufacturing Processes*. Amsterdam, Netherlands: Elsevier.
- Guo, Shuxia, Thomas Bocklitz, and Jürgen Popp (2016). „Optimization of Raman-spectrum baseline correction in biological application“. In: *Analyst* 141 (8), pp. 2396–2404.
- Hansen, Sigrid K., Babak Jamali, and Jürgen Hubbuch (02/2013). „Selective high throughput protein quantification based on UV absorption spectra“. eng. In: *Biotechnology and Bioengineering* 110.2, pp. 448–460.
- Harada, Issei, Takashi Miura, and Hideo Takeuchi (01/1986). „Origin of the doublet at 1360 and 1340 cm⁻¹ in the Raman spectra of tryptophan and related compounds“. en. In: *Spectrochimica Acta Part A: Molecular Spectroscopy* 42.2-3, pp. 307–312.
- Haseltine, Eric L. and James B. Rawlings (2005). „Critical evaluation of extended Kalman filtering and moving-horizon estimation“. In: *Industrial and Engineering Chemistry Research* 44.8, pp. 2451–2460.
- He, Hao et al. (01/2024). „Noise learning of instruments for high-contrast, high-resolution and fast hyperspectral microscopy and nanoscopy“. In: *Nature Communications* 15.1, p. 754.
- He, Shixuan et al. (2014). „Baseline correction for Raman spectra using an improved asymmetric least squares method“. In: *Anal. Methods* 6.12, pp. 4402–4407.
- Helleckes, Laura M. et al. (06/2023). „Machine Learning in Bioprocess Development: From Promise to Practice“. In: *Trends in Biotechnology* 41.6, pp. 817–835.
- Hillebrandt, Nils and Jürgen Hubbuch (05/2023). „Size-selective downstream processing of virus particles and non-enveloped virus-like particles“. In: *Frontiers in Bioengineering and Biotechnology* 11.
- Hillebrandt, Nils, Philipp Vormittag, Nicolai Bluthardt, et al. (2020). „Integrated Process for Capture and Purification of Virus-Like Particles: Enhancing Process Performance by Cross-Flow Filtration“. In: *Frontiers in Bioengineering and Biotechnology* 8.May.
- Hillebrandt, Nils, Philipp Vormittag, Annabelle Dietrich, and Jürgen Hubbuch (06/2022). „Process monitoring framework for cross-flow diafiltration-based virus-like particle disassembly: Tracing product properties and filtration performance“. In: *Biotechnology and Bioengineering* 119.6, pp. 1522–1538.
- Hollas, J. Michael (2010). *Modern Spectroscopy*. 4. ed., reprinted. Chichester: Wiley.
- Hong, Moo Sun et al. (2018). „Challenges and opportunities in biopharmaceutical manufacturing control“. In: *Computers & Chemical Engineering* 110, pp. 106–114.
- Hu, Xi et al. (2017). „Development of Commercial-Ready Processes for Antibody Drug Conjugates“. In: *Organic Process Research and Development* 21.4, pp. 601–610.
- Hutter, Frank, Lars Kotthoff, and Joaquin Vanschoren, eds. (2019). *Automated Machine Learning: Methods, Systems, Challenges*. The Springer Series on Challenges in Machine Learning. Cham: Springer International Publishing.
- ICH (2009b). *Q8(R2) Pharmaceutical Development*. Tech. rep. June.
- Ioffe, Sergey and Christian Szegedy (2015). „Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift“. In: *32nd International Conference on Machine Learning, ICML 2015* 1, pp. 448–456.
- James, G. et al. (2013). *An Introduction to Statistical Learning*. 244–251.

- Kadlec, Petr, Bogdan Gabrys, and Sibylle Strandt (04/2009). „Data-Driven Soft Sensors in the Process Industry“. In: *Computers and Chemical Engineering* 33.4, pp. 795–814.
- Kazemzadeh, Mohammadrahim et al. (05/2022). „Deep convolutional neural networks as a unified solution for Raman spectroscopy-based classification in biomedical applications“. In: *Optics Communications* 510, p. 127977.
- Kessler, Waltraud (2007). *Multivariate Datenanalyse für die Pharma-, Bio- und Prozessanalytik - Ein Lehrbuch*. Wiley VCH Verlag GmbH & Co. KGaA.
- Kheirvari, Milad, Hong Liu, and Ebenezer Tumban (05/02/2023). „Virus-like Particle Vaccines and Platforms for Vaccine Development“. In: *Viruses* 15.5, p. 1109.
- Kingma, Diederik P. and Jimmy Lei Ba (2015). „Adam: A method for stochastic optimization“. In: *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pp. 1–15.
- Kushnir, Natasha, Stephen J. Streatfield, and Vidadi Yusibov (12/2012). „Virus-like particles as a highly efficient vaccine platform: Diversity of targets and production systems and advances in clinical development“. In: *Vaccine* 31.1, pp. 58–83.
- Lakowicz, J.R. (2006). *Principles of Fluorescence Spectroscopy*. Springer US.
- Li, Meng-Yao et al. (03/2018). „Real-time monitoring of antibody glycosylation site occupancy by in situ Raman spectroscopy during bioreactor CHO cell cultures“. In: *Biotechnology Progress* 34.2 (2), pp. 486–493.
- Liu, Jinchao et al. (2017). „Deep convolutional neural networks for Raman spectrum recognition: A unified solution“. In: *Analyst* 142.21 (21), pp. 4067–4074.
- Loving, Galen S., Matthieu Sainlos, and Barbara Imperiali (02/2010). „Monitoring Protein Interactions and Dynamics with Solvatochromic Fluorophores“. In: *Trends in Biotechnology* 28.2, pp. 73–83.
- Lundberg, Scott M, Paul G Allen, and Su-In Lee (2017). „A Unified Approach to Interpreting Model Predictions“. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30, pp. 4765–4774.
- Luttmann, Reiner et al. (2012). „Soft sensors in bioprocessing: A status report and recommendations“. In: *Biotechnology Journal* 7.8, pp. 1040–1048.
- Martens, Harald and Edward Stark (01/1991a). „Extended multiplicative signal correction and spectral interference subtraction: New preprocessing methods for near infrared spectroscopy“. In: *Journal of Pharmaceutical and Biomedical Analysis* 9.8, pp. 625–635.
- Mayerhöfer, Thomas G., Susanne Pahlow, and Jürgen Popp (09/2020). „The Bouguer-Beer-Lambert Law: Shining Light on the Obscure“. In: *ChemPhysChem* 21.18, pp. 2029–2046.
- Mehmood, Tahir et al. (08/2012). „A review of variable selection methods in Partial Least Squares Regression“. In: *Chemometrics and Intelligent Laboratory Systems* 118, pp. 62–69.
- Miller, Charles E. (2007). „Chemometrics in Process Analytical Chemistry“. In: *Process Analytical Technology*. Blackwell Publishing, pp. 226–328.
- Mohsen, Mona O. and Martin F. Bachmann (08/12/2022). „Virus-like Particle Vaccinology, from Bench to Bedside“. In: *Cellular & Molecular Immunology* 19.9, pp. 993–1011.
- Moleirinho, Mafalda G. et al. (05/2020). „Current challenges in biotherapeutic particles manufacturing“. In: *Expert Opinion on Biological Therapy* 20.5, pp. 451–465.
- Moussa, Ehab M. et al. (02/2016). „Immunogenicity of Therapeutic Protein Aggregates“. English. In: *Journal of Pharmaceutical Sciences* 105.2, pp. 417–430.
- Narayanan, Harini et al. (2020). „Hybrid-EKF: Hybrid model coupled with extended Kalman filter for real-time monitoring and control of mammalian cell culture“. In: *Biotechnology and Bioengineering* 117.9, pp. 2703–2714.
- Ng, Andrew Y (2004). „L1 and L2 Regularisation Comparisation“. In: *Proceedings of the 21 st International Conference on Machine Learning*.
- Nikzad-Langerodi, Ramin and Erik Andries (2021). „A chemometrician’s guide to transfer learning“. en. In: *Journal of Chemometrics* 35.11, e3373.
- Nikzad-Langerodi, Ramin, Edwin Lughofer, et al. (2018). „Calibration model maintenance in melamine resin production: Integrating drift detection, smart sample selection and model adaptation“. In: *Analytica Chimica Acta* 1013, pp. 1–12.
- Nørgaard, L. et al. (03/2000). „Interval Partial Least-Squares Regression (*i* PLS): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy“. en. In: *Applied Spectroscopy* 54.3, pp. 413–419.
- Ohadi, Kaveh, Raymond L. Legge, and Hector M. Budman (2015). „Development of a soft-sensor based on multi-wavelength fluorescence spectroscopy and a dynamic metabolic model for monitoring mammalian cell cultures“. In: *Biotechnology and Bioengineering* 112.1, pp. 197–208.
- Otto, Matthias (2017). *Chemometrics - Statistics and Computer Application in Analytical Chemistry*. Wiley VCH Verlag GmbH & Co. KGaA.
- Paananen, Topi et al. (12/2019). „Variable selection for Gaussian processes via sensitivity analysis of the posterior predictive distribution“. In: *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)* PLMR.89, pp. 1743–1752.
- Porterfield, J. Zachary and Adam Zlotnick (11/2010). „A simple and general method for determining the protein and nucleic acid content of viruses by UV absorbance“. In: *Virology* 407.2, pp. 281–288.
- Preston, Caroline M. and W. A. Adams (04/1979). „A Laser Raman Spectroscopic Study of Aqueous Orthophosphate Salts“. In: *The Journal of Physical Chemistry* 83.7, pp. 814–821.
- Prince, Simon J. D. (2023). *Understanding Deep Learning*. Cambridge, Massachusetts London, England: The MIT Press.
- Qian, Ciying et al. (03/2020). „Recent Progress on the Versatility of Virus-Like Particles“. In: *Vaccines* 8.1, p. 139.
- Raschka, Sebastian (11/2018). „Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning“. In.
- Rasmussen, Carl Edward and Christopher K. I. Williams (2006). *Gaussian processes for machine learning*. Ed. by Thomas Dietterich. Vol. 38. Cambridge, MA, USA: MIT Press.

Machine Learning in Spectroscopy

- Rathore, A. S., R. Bhambure, and V. Ghare (09/2010). „Process analytical technology (PAT) for biopharmaceutical products“. In: *Analytical and Bioanalytical Chemistry* 398.1 (1), pp. 137–154.
- Read, E. K. et al. (2010). „Process Analytical Technology (PAT) for biopharmaceutical products: Part II. Concepts and applications“. In: *Biotechnology and Bioengineering* 105.2, pp. 285–295.
- Roberts, Christopher J (12/2014). „Protein aggregation and its impact on product quality“. en. In: *Current Opinion in Biotechnology* 30, pp. 211–217.
- Rolinger, Laura, Jürgen Hubbuch, and Matthias Rüdts (02/2023). „Monitoring of ultra- and diafiltration processes by Kalman-filtered Raman measurements“. In: *Analytical and Bioanalytical Chemistry* 415.5, pp. 841–854.
- Rolinger, Laura, Matthias Rüdts, and Jürgen Hubbuch (04/2020a). „A critical review of recent trends, and a future perspective of optical spectroscopy as PAT in biopharmaceutical downstream processing“. In: *Analytical and Bioanalytical Chemistry* 412.9, pp. 2047–2064.
- Rosenberg, Amy S. (09/2006). „Effects of protein aggregates: An immunologic perspective“. en. In: *The AAPS Journal* 8.3, p. 59.
- Rüdts, Matthias, Nina Brestrich, et al. (02/2017). „Real-time monitoring and control of the load phase of a protein A capture step“. In: *Biotechnology and Bioengineering* 114.2 (2), pp. 368–373.
- Rüdts, Matthias, Till Briskot, and Jürgen Hubbuch (03/2017). „Advances in downstream processing of biologics – Spectroscopy: An emerging process analytical technology“. In: *Journal of Chromatography A* 1490, pp. 2–9.
- Rygula, A. et al. (2013). „Raman spectroscopy of proteins: A review“. In: *Journal of Raman Spectroscopy* 44.8, pp. 1061–1076.
- S. De Jong (1993). „SIMPLS: an alternative approach to partial least squares regression“. In: *Chemometrics and intelligent laboratory systems* 18, pp. 251–263.
- Sadiki, Amissi et al. (12/2020). „Site-Specific Conjugation of Native Antibody“. In: *Antibody Therapeutics* 3.4, pp. 271–284.
- Saggu, Miguel, Jun Liu, and Ankit Patel (09/2015). „Identification of Subvisible Particles in Biopharmaceutical Formulations Using Raman Spectroscopy Provides Insight into Polysorbate 20 Degradation Pathway“. In: *Pharmaceutical Research* 32.9, pp. 2877–2888.
- Saleh, David et al. (2020). „Straightforward Method for Calibration of Mechanistic Cation Exchange Chromatography Models for Industrial Applications“. In: *Biotechnology Progress* 800.
- Savitzky, Abraham. and M. J. E. Golay (07/1964). „Smoothing and Differentiation of Data by Simplified Least Squares Procedures“. In: *Analytical Chemistry* 36.8, pp. 1627–1639.
- Schmid, Franz-Xaver (2001). „Biological Macromolecules: UV-visible Spectrophotometry“. In: .
- Schuhmacher, Alexander, Oliver Gassmann, and Markus Hinder (02/2016). „A Review of the Pharmaceutical R&D Efficiency: Costs, Timelines, and Probabilities“. In: *Value Creation in the Pharmaceutical Industry*, pp. 60–79.
- Selvaraju, Ramprasaath R. et al. (2020). „Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization“. In: *International Journal of Computer Vision* 128 (2), pp. 336–359.
- Shukla, Abhinav A. et al. (03/2007a). „Downstream Processing of Monoclonal Antibodies—Application of Platform Approaches“. In: *Polyclonal and Monoclonal Antibody Production, Purification, Process and Product Analytics* 848.1, pp. 28–39.
- Shukla, Abhinav A. et al. (2007b). „Downstream processing of monoclonal antibodies—Application of platform approaches“. In: *Journal of Chromatography B* 848.1, pp. 28–39.
- Simon, Dan (2006). „Nonlinear Kalman filtering“. In: *Optimal State Estimation*. John Wiley & Sons, Ltd. Chap. 13, pp. 393–431.
- Srivastava, Nitish et al. (01/2014). „Dropout: a simple way to prevent neural networks from overfitting“. In: *J. Mach. Learn. Res.* 15.1, pp. 1929–1958.
- Stelzer, Ines V., Julian Kager, and Christoph Herwig (2017). „Comparison of Particle Filter and Extended Kalman Filter Algorithms for Monitoring of Bioprocesses“. In: *Proceedings of the 27th European Symposium on Computer Aided Process Engineering – ESCAPE 27*. Vol. 40. Elsevier Masson SAS, pp. 1483–1488.
- Štrumbelj, Erik and Igor Kononenko (12/2014a). „Explaining Prediction Models and Individual Predictions with Feature Contributions“. In: *Knowledge and Information Systems* 41.3, pp. 647–665.
- Thakur, Garima, Vishwanath Hebbs, and Anurag S. Rathore (2020). „An NIR-based PAT approach for real-time control of loading in Protein A chromatography in continuous manufacturing of monoclonal antibodies“. In: *Biotechnology and Bioengineering* 117.3, pp. 673–686.
- Thakur, Garima, Swapnil Thori, and Anurag S. Rathore (11/2020). „Implementing PAT for single-pass tangential flow ultrafiltration for continuous manufacturing of monoclonal antibodies“. en. In: *Journal of Membrane Science* 613.July, p. 118492.
- Thomann, Marco et al. (2016). „Fc-galactosylation modulates antibody-dependent cellular cytotoxicity of therapeutic antibodies“. In: *Molecular Immunology* 73, pp. 69–75.
- Tiernan, Hannah, Bernadette Byrne, and Sergei G. Kazarian (11/2020). „ATR-FTIR Spectroscopy and Spectroscopic Imaging for the Analysis of Biopharmaceuticals“. In: *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 241, p. 118636.
- Top Pharma Products by Global Sales 2024* (2025). Statista. URL: <https://www.statista.com/statistics/258022/top-10-pharmaceutical-products-by-global-sales-2011/> (visited on 09/12/2025).
- Trygg, Johan and Svante Wold (2002). „Orthogonal projections to latent structures (O-PLS)“. In: *Journal of Chemometrics* 16.3, pp. 119–128.
- Tulsyan, Aditya et al. (2021). „Spectroscopic models for real-time monitoring of cell culture processes using spatiotemporal just-in-time Gaussian processes“. In: *AIChE Journal* 67.5 (5).
- Tuveri, Andrea et al. (2021). „Sensor fusion based on Extended and Unscented Kalman Filter for bioprocess monitoring“. In: *Journal of Process Control* 106, pp. 195–207.

CHAPTER 1. INTRODUCTION

- Tyther, Raymond and Nigel Jenkins (2011). „Quality Issues Arising from Post-translational Modification of Recombinant Antibodies“. In: *Antibody Expression and Production*. Ed. by Mohamed Al-Rubeai. Dordrecht: Springer Netherlands, pp. 293–303.
- Valensise, C. M. et al. (06/2020). „Removing non-resonant background from CARS spectra via deep learning“. In: *APL Photonics* 5.6, p. 061305.
- Valentic, Angela and Jürgen Hubbuch (2024). „Effective removal of host cell-derived nucleic acids bound to hepatitis B core antigen virus-like particles by heparin chromatography“. In: *Frontiers in Bioengineering and Biotechnology* Volume 12 - 2024.
- Vanderlaan, Martin et al. (2018). „Experience with host cell protein impurities in biopharmaceuticals“. In: *Biotechnology Progress* 34.4, pp. 828–837.
- Vázquez-Rey, María and Dietmar A. Lang (07/2011). „Aggregates in monoclonal antibody manufacturing processes“. en. In: *Biotechnology and Bioengineering* 108.7, pp. 1494–1508.
- Walsh, Gary and Eithne Walsh (12/2022). „Biopharmaceutical benchmarks 2022“. en. In: *Nature Biotechnology* 40.12, pp. 1722–1760.
- Wen, Zai-Qing (11/2007). „Raman spectroscopy of protein pharmaceuticals“. In: *Journal of Pharmaceutical Sciences* 96.11, pp. 2861–2878.
- Wiercigroch, Ewelina et al. (10/2017). „Raman and Infrared Spectroscopy of Carbohydrates: A Review“. In: *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 185, pp. 317–335.
- Wold, Svante and Michael Sjöström (06/01/1977). „SIMCA: A Method for Analyzing Chemical Data in Terms of Similarity and Analogy“. In: *Chemometrics: Theory and Application*. Vol. 52. ACS Symposium Series 52. AMERICAN CHEMICAL SOCIETY, pp. 243–282.
- Wold, Svante, Michael Sjöström, and Lennart Eriksson (10/2001). „PLS-regression: a basic tool of chemometrics“. en. In: *Chemometrics and Intelligent Laboratory Systems* 58.2 (2), pp. 109–130.
- Xiaobo, Zou et al. (05/2010). „Variables selection methods in near-infrared spectroscopy“. en. In: *Analytica Chimica Acta* 667.1-2, pp. 14–32.
- Zeltins, Andris (01/2013). „Construction and Characterization of Virus-Like Particles: A Review“. In: *Molecular Biotechnology* 53.1, pp. 92–107.
- Zhang, Feng et al. (03/2020). „Baseline correction for infrared spectra using adaptive smoothness parameter penalized least squares method“. In: *Spectroscopy Letters* 53.3, pp. 222–233.
- Zhang, Pengfei et al. (02/2021). „A novel variable selection method based on combined moving window and intelligent optimization algorithm for variable selection in chemical modeling“. en. In: *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 246, p. 118986.
- Zhu, Yiran et al. (07/23/2022). „RNA-based Therapeutics: An Overview and Prospectus“. In: *Cell Death & Disease* 13.7, p. 644.

Thesis Outline

2.1 Research Proposal

Biopharmaceutical downstream processing requires robust real-time monitoring to ensure product quality and process consistency. However, current practice relies predominantly on offline analytical methods that introduce time delays compromising process control strategies. The FDA's PAT framework promotes real-time monitoring through multivariate spectroscopic sensors, but implementation remains challenging for complex biological products. Optical spectroscopy provides molecular-specific process information through non-invasive techniques. UV/Vis absorption spectroscopy enables rapid measurements but offers limited selectivity due to broad, overlapping absorption bands. Raman spectroscopy provides molecular fingerprinting with sensitivity to protein structure and buffer composition, though challenged by low signal intensity and fluorescence interference. The evaluation of spectroscopic data requires the use of chemometrics, i.e. preprocessing and statistical regression techniques, for the elimination of unwanted spectral effects and the quantitative determination of the corresponding sample composition and hence critical quality attributes. Commonly used regression techniques such as PLS regression work well when linearity holds and spectral variations reflect concentration changes. However, these linear methods fail when concentration-dependent effects violate linearity, spectral overlap prevents component resolution, or environmental factors introduce systematic variations. Modern ML approaches including GPR and CNNs offer increased modeling flexibility, but require careful tuning of the associated hyperparameters as well as proper validation to ensure they capture relevant chemical relationships rather than dataset-specific artifacts. Proper calibration and validation of chemometric models is crucial for robust performance and necessitates the collection of data sets with systematic variation of process parameters and analyte concentrations over the expected ranges. The combination of all previously mentioned approaches promises the development of selective and robust PAT incorporating systematic, efficient experimental design, automatic pipeline optimization, and ML.

This work focuses on developing robust chemometric models for spectroscopic process monitoring in biopharmaceutical manufacturing using modern ML methods. Five studies systematically progress

from basic selectivity challenges to increasingly complex scenarios involving strong spectral overlap, environmental interferences, and demanding process conditions. The work establishes comprehensive frameworks spanning adaptive sensor fusion, deep learning with limited data, systematic preprocessing optimization, automated high-throughput calibration, and validated structural monitoring.

In antibody-drug conjugation reactions, multiple species with similar spectral properties exist simultaneously. UV/Vis spectroscopy cannot directly resolve individual conjugation states because their absorption spectra are too similar for direct chemometric decomposition. To overcome this fundamental selectivity limitation, hybrid approaches that combine spectroscopic measurements with mechanistic process knowledge offer a promising alternative.

In the first study, an adaptive soft-sensor will be developed that combines chemometric model predictions from UV/Vis measurements with kinetic reaction models through an optimal state estimation algorithm. This hybrid system will dynamically estimate all reaction species by fusing spectroscopic observations with mechanistic predictions. The system will make use of a GPR model to obtain time-variant uncertainty estimates for each spectroscopic observation, enabling the balancing between spectroscopic measurements and kinetic model predictions as signal quality and model accuracy evolve. Validation across batch and fed-batch conjugation modes will demonstrate whether mechanistic knowledge integration enables monitoring of spectroscopically indistinguishable species.

Spectral selectivity challenges are common in biopharmaceutical applications as product-related impurities are often structurally similar to the product and hence exhibit a strong spectral overlap. These challenges may also be addressed using deep learning methodologies to extract the fine spectral differences from the spectroscopic measurements. In this regard, CNNs offer potential advantages over linear methods through their hierarchical feature extracting capabilities, yet their application to spectroscopic calibration faces fundamental obstacles. Experimental datasets rarely provide sufficient samples for robust training, and configuring network architectures to maximize performance is a laborious and computationally expensive process. For accurate quantification in biopharmaceutical applications, sufficient training data must be available and the risk of overfitting needs to be minimized.

In the second study, a generative data augmentation method will be introduced to generate synthetic spectra that preserve physicochemical validity and can be used for training deep learning models. This approach will account for concentration-dependent effects and matrix interactions that cause spectra to vary across the experimental space. To automate the HPO and thus automate the neural architecture search, Bayesian methods will be employed. The automated HPO will enable to simultaneously search the combined space of data augmentation parameters, network architecture, and training strategies to identify configurations that maximize generalization performance. Comprehensive interpretability assessment will ensure that optimized models extract relevant spectral features rather than exploiting dataset artifacts.

The transition from UV/Vis to Raman spectroscopy introduces additional challenges related to environmental interference in complex bioprocess matrices. VLP precipitation represents a challenging test case where particulate suspensions originating from *E. coli* lysates and precipitating agents interfere with the Raman measurements and dominate the observed spectra over the VLP-related features. Unlike the relatively clean UV/Vis measurements, Raman spectra are affected by varying baseline shapes and scatter behaviors that continuously change with sample composition. A variety

of preprocessing methods are available for spectral correction, but for their application to particulate systems, these methods do not necessarily achieve optimal performance without systematic evaluation and optimization.

In the third study, a Raman-based PAT will be developed to monitor VLP concentrations during precipitation from crude *E. coli* lysates. To achieve applicability to particulate systems, preprocessing operations will be systematically evaluated, including turbidity correction, baseline correction, interferent removal, and derivative transformations. Systematic experimental design with intentional variation in lysate composition and precipitation dynamics will validate whether optimized pipelines effectively eliminate process-related interferences while preserving chemical information.

Traditional calibration procedures create an experimental bottleneck limiting widespread implementation of Raman spectroscopy for process monitoring. Manual preparation and measurement of calibration samples is labor-intensive and error-prone, while systematic differences between offline calibration and in-line process measurements can severely degrade model performance. In an ideal case, calibration of PAT tools can be automated and transferred to in-line applications.

In the fourth study, an automated platform will be introduced that integrates a commercially available chromatography system with a Raman spectrometer to enable high-throughput generation of spectral calibration data with minimal sample volumes and operator involvement. Statistical experimental designs will be applied to efficiently sample multi-dimensional concentration spaces. Direct standardization methods will be developed to transfer offline-calibrated models to in-line applications by modeling systematic spectral differences between measurement configurations. This workflow will be demonstrated using UF/DF as a case study with the goal of simultaneously monitoring protein and excipient concentrations.

Monitoring protein aggregation during chromatographic separation presents a crucial lever to enable real-time process control for biopharmaceutical downstream processing. The distinction of protein aggregates from monomers is challenging due to the subtle spectral changes between the two species. To use PAT tools in biopharmaceutical operations, regulatory agencies require rigorous validation of the established chemometric model and call for interpretable predictions. Raman spectroscopy presents an opportunity to derive molecular structural features rather than empirical correlations. Several techniques have been reported in the literature to detect protein aggregation using Raman spectroscopy, though these studies typically assess aggregation qualitatively or under idealized conditions. Furthermore, current literature lacks comprehensive validation that spectral features respond specifically to conformational changes rather than confounding factors.

In the fifth study, a Raman-based PAT will be developed for the monitoring of a chromatographic separation of protein monomers and aggregates. To evaluate the molecular structural features derived from Raman spectroscopy measurements for their specificity to protein aggregation, controlled stress experiments, mixing studies where concentration and aggregation vary orthogonally, and in-line detection during anion-exchange chromatography will be performed. Both PLS and CNN models will be trained and systematically validated to enable real-time monitoring of the chromatography process. The capabilities and limitations of both approaches will be compared, with particular attention to deep learning robustness when coupled with augmentation strategies.

In summary, expected outcomes of this research include systematic frameworks for developing robust and selective ML models for spectroscopic process monitoring. The research will provide methodological advances in data augmentation, preprocessing and HPO strategies, ML model

calibration, and interpretability assessment, exemplified through novel and relevant applications in the DSP of biopharmaceutical products such as mAbs, VLPs, and ADCs. The proposed PAT solutions may enable robust spectroscopic real-time monitoring allowing immediate detection of process deviations and implementation of corrective actions before product quality is compromised. Similarly, these tools can be employed to accelerate process development by reducing reliance on time-consuming offline analytics and to enhance process understanding by providing continuous insights into process dynamics that cannot be captured by intermittent sampling. While the research focuses on biopharmaceutical applications, the developed methods are broadly applicable to small molecule pharmaceutical manufacturing, chemical process industries, food and beverage quality control, and environmental monitoring.

2.2 Manuscript Overview

Within the scope of this thesis, five manuscripts were prepared and subsequently published. Each manuscript is presented briefly with a summary and the corresponding citation. As a result, the publications are slightly changed and adapted to the format and layout of this thesis. Moreover, four of the following manuscripts were prepared in collaboration with colleagues to enhance their quality. As a result, the first authorship was shared, i.e., contributed equally. Detailed information on the contributions and the signature of each author is included as a supplement to the examination copy of this thesis. Ch. 3, can also be found in the thesis of the co-author Jan Tobias Weggen. Ch. 5, can also be found in the thesis of the co-author Annabelle Dietrich. Ch. 6 and Ch. 7, can also be found in the thesis of the co-author Jakob Heyer-Müller.

Chapter 3

An adaptive soft-sensor for advanced real-time monitoring of an antibody-drug conjugation reaction.....49

Robin Schiemer, Jan Tobias Weggen, Katrin Marianne Schmitt and Jürgen Hubbuch

published in Biotechnology and Bioengineering, Volume 120(7), 2023,
<https://doi.org/10.1002/bit.28428>

Ch. 3 addresses the challenge of real-time monitoring in ADC processing, where ultraviolet/visible (UV/Vis) spectroscopy lacks the selectivity to directly distinguish all relevant reaction species. To overcome this limitation, an adaptive soft-sensor framework is developed that fuses spectroscopic measurements with mechanistic process knowledge through an optimal state estimation algorithm. This approach combines predictions from an uncertainty-aware GPR sensor model with forecasts from a kinetic reaction model, enabling the indirect estimation of unmeasurable ADC species during site-directed conjugation. By explicitly accounting for time-variant uncertainties in both the spectroscopic sensor and the kinetic model, the framework achieves robust state estimation of all reaction species even in the presence of sensor noise and model errors. The methodology is validated across both batch and fed-batch operation modes, demonstrating its potential for future implementation in model-based process control strategies for ADC manufacturing.

Chapter 4

Generative data augmentation and automated optimization of convolutional neural networks for process monitoring..... 71

Robin Schiemer, Matthias Rüdert and Jürgen Hubbuch

published in Frontiers in Bioengineering and Biotechnology, Volume 12, 2024

<https://doi.org/10.3389/fbioe.2024.1228846>

Ch. 4 explores the use of CNNs as chemometric models for biopharmaceutical process monitoring. To overcome the fundamental constraint of limited availability of experimental data for training CNN models, a novel data augmentation method is developed that generates highly realistic synthetic spectral data. This generative approach is combined with automated HPO using Bayesian optimization to identify optimal model architectures. The resulting CNN models demonstrate substantial improvements in prediction accuracy compared to PLS regression, particularly for the quantification of monoclonal antibody size variants, while simultaneously reducing preprocessing requirements. Feature importance methods reveal that the optimized CNNs exhibit increased robustness against spectral distortions and improved transferability across different experimental conditions, addressing key concerns about the interpretability and reliability of ML approaches in chemometric applications.

Chapter 5

Raman-based PAT for VLP precipitation: systematic data diversification and preprocessing pipeline identification 105

Annabelle Dietrich, Robin Schiemer, Jasper Kurmann, Shiqi Zhang and Jürgen Hubbuch

published in Frontiers in Bioengineering and Biotechnology, Volume 12, 2024

<https://doi.org/10.3389/fbioe.2024.1399938>

Ch. 5 addresses the challenge of real-time monitoring for VLP precipitation processes, where conventional analytical methods are hindered by the turbidity and complexity of precipitate-containing lysates. Raman spectroscopy is evaluated as a PAT for this application, with a particular focus on overcoming spectral interferences caused by flocculates, precipitant, and feedstock components. To generate a diversified experimental dataset, systematic variations are introduced through spiking and conditioning of the starting material. Combinations of preprocessing operations are screened to identify an optimal pipeline that effectively eliminates background variations while preserving information relevant to VLP and precipitant quantification. The resulting PLS regression models successfully capture precipitation dynamics in both batch and fed-batch modes, demonstrating the feasibility of Raman-based monitoring for phase behavior-dependent processes despite the non-linear relationship between spectral features and product concentration in complex biological matrices.

Chapter 6

A novel Raman-chromatography assembly for automated calibration and in-line monitoring in bioprocessing.....137

Jakob Heyer-Müller, Robin Schiemer, Matthias Lopinski, Franka Willems, Caty Wang, Lars Robbel, Michael Schmitt and Jürgen Hubbuch

published in Engineering in Life Sciences, Volume 25(9), 2025
<https://doi.org/10.1002/elsc.70044>

Ch. 6 introduces an automated platform integrating a commercial chromatography system with Raman spectroscopy to enable efficient chemometric model calibration. The novel assembly automates the preparation and analysis of synthetic calibration mixtures with minimal sample volumes, significantly reducing the manual effort required for chemometric model development. UF/DF is used as a case study with the aim to simultaneously monitor protein and excipient concentration in real-time during a filtration experiment. Therefore, calibration datasets are generated off-line using the automated measurement platform, on which PLS models are trained to quantify protein and buffer components. To transfer the models to in-line applications, a direct standardization approach is employed compensating spectral discrepancies between measurement modes. Thus, the combination of automated off-line measurements and in-line usage of the calibrated model establishes a practical workflow for accelerated PAT implementation in biopharmaceutical DSP.

Chapter 7

Development of Raman spectroscopy and machine learning methods for protein aggregate quantification: Application to BSA in chromatographic processes.....151

Jakob Heyer-Müller, Robin Schiemer, Lars Robbel, Michael Schmitt and Jürgen Hubbuch

submitted to Biotechnology and Bioengineering, (accepted)
<https://doi.org/10.1002/bit.70163>

Ch. 7 investigates the capability of Raman spectroscopy to reliably measure protein aggregation in DSP. To study whether protein aggregation can be probed via Raman spectroscopy, bovine serum albumin (BSA) is subjected to heat stress which induces aggregate formation. Raman spectra are collected from forced degradation experiments, independent mixing studies, and chromatography experiments, where protein dimers are separated from native monomers. The collected data is assessed for molecular structural markers derived from protein-linked Raman bands. Based on data from the chromatography experiments, PLS and CNN models are trained to selectively quantify monomer and aggregate concentrations. The models are subsequently validated to ensure they capture the relevant spectral signatures rather than spurious correlations. While both model types demonstrate their ability to quantify BSA size variants, the CNNs achieve superior performance, underlining the advantage of modern ML methods in chemometrics.

3

An adaptive soft-sensor for advanced real-time monitoring of an antibody-drug conjugation reaction

Robin Schiemer^{1,+}, Jan Tobias Weggen^{1,+}, Katrin Marianne Schmitt¹, and Jürgen Hubbuch¹

¹ Institute of Engineering in Life Sciences, Section IV: Biomolecular Separation Engineering, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

⁺ contributed equally

Abstract

In the production of antibody-drug conjugates (ADCs), the conjugation reaction is a central step defining the final product composition and, hence, directly affecting product safety and efficacy. To enable real-time monitoring, spectroscopic sensors in combination with multivariate regression models have gained popularity in recent years. The Extended Kalman filter (EKF) can be used as so-called soft-sensor to fuse sensor predictions with long-horizon forecasts by process models. This enables the dynamic update of the current state and provides increased robustness against experimental noise or model errors. Due to the uncertainty associated with sensor and process models in biopharmaceutical applications, the deployment of such soft-sensors is challenging.

In this study, we demonstrate the combination of an uncertainty-aware sensor model with a kinetic reaction model using an EKF to monitor a site-directed ADC conjugation reaction. As the sensor model, a Gaussian process regression model is presented to realize a time-variant determination of the sensor uncertainty. The EKF fuses the time-discrete predictions of the amount of conjugated drug from the sensor model with the time-continuous predictions from the kinetic model. While the ADC species are not distinguishable by on-line recorded UV/Vis spectra, the developed soft-sensor is able to dynamically update all relevant reaction species. It could be shown that the use of time-variant process and sensor noise computation approaches improved the performance of the EKF and achieved a reduction of the prediction error of up to 23 % compared to the kinetic model. The developed framework proved to enhance robustness against noisy sensor measurements or wrong model initialization and was successfully transferred from batch to fed-batch mode. In future, this framework could be implemented for model-based process control and be adopted for other ADC conjugation reaction types.

3.1 Introduction

Antibody-drug conjugates (ADCs) are one of the fastest growing class of biopharmaceuticals using the target specificity of a monoclonal antibody (mAb) to exclusively deliver a cytotoxic drug to tumor cells (Thomas, Teicher, and Hassan 2016). Many different conjugation strategies exist consisting of multiple reaction steps while one of the greatest challenges remains in controlling the drug-to-antibody ratio (DAR) (Panowski et al. 2014; Joubert et al. 2020). To ensure consistent final product quality, regulatory agencies have been promoting the implementation of advanced monitoring and control strategies according to the concepts of quality by design (QbD) and process analytical technology (PAT) during a decade (ICH 2009a; FDA 2004). In line with the QbD framework, recent publications showed the successful application of model-based techniques in ADC process development, such as molecular dynamics simulations (Coumans et al. 2020), conjugation kinetic models (Andris, Seidel, and Hubbuch 2019) and chromatography models (Andris and Hubbuch 2020; Keller and Wendeler 2021). Mechanistic process models, however, often require simplification of complex phenomena which may lead to plant-model mismatches or poor predictability of the transient behavior of the process (Gargalo et al. 2021). With regards to PAT, spectroscopic sensors have shown to enable real-time measurement of certain critical quality attributes (CQAs) (A. S. Rathore, Bhambure, and Ghare 2010). Consequently, these sensors have been deployed as PAT for various unit operations in downstream processing of mAbs (Großhans et al. 2018; Rolinger, Rüdts, and Hubbuch 2020b; Wei et al. 2022). In the field of ADCs, monitoring of the conjugation reaction using ultraviolet/visible (UV/Vis) spectra (Andris, Rüdts, et al. 2018) and quantification of aggregation using Raman spectroscopy (C. Zhang et al. 2019) are recently established PAT tools. Partial least squares (PLS) models have commonly been applied in chemometrics (S. Wold,

M. Sjöström, and Eriksson 2001; Andersson 2009) and, more recently, Gaussian process regression (GPR) models have been identified as a competitive alternative to PLS (H. Chen, Bakshi, and Goel 2007; Cui and Fearn 2017; Tulsyan et al. 2021). Furthermore, GPR models deliver an uncertainty estimate for each evaluated data point and therefore help to perform risk-based dynamic adjustments during the process (Tulsyan et al. 2021).

The extracted information from sensors can be used to supply process models with direct feedback from the process in near real-time, which enables timely model updates and, hence, more accurate estimates of process- or product-relevant variables (Sinner et al. 2021). To fuse estimates from two noisy sources, various optimal state estimation algorithms have been developed and recently been applied to chemical (Haseltine and Rawlings 2005; Schneider and Georgakis 2013) and pharmaceutical (Simutis and Lübbert 2017; Golabgir and Herwig 2016; Elsheikh et al. 2021) processes. The extended Kalman filter (EKF), as one of the most prominently applied algorithms, has been used to combine non-linear process models and sensor measurements for cell culture processes (Narayanan et al. 2020; Elsheikh et al. 2021) or for capture chromatography (Feidl et al. 2019). In these studies, the EKF computes a weighted average of both source models, but does not affect the long-horizon predictions of the process model. In many cases, however, only a few states are directly available from measurement probes. This problem can be circumvented by utilizing an EKF formulation that combines the measurement with mass or energy balances. Thus, an EKF can be utilized to update all states in the model which are not available for measurement. This approach is considered a model-based soft-sensor according to (Luttmann et al. 2012) and was similarly demonstrated in recent studies for advanced monitoring of bioprocesses (Cabaneros Lopez et al. 2020; Tuveri et al. 2021; Ohadi, Legge, and Budman 2015; Spann et al. 2018).

This study aims to develop a robust and adaptive soft-sensor to monitor the site-directed conjugation reaction of a mAb with a surrogate payload. Therefore, an uncertainty-aware chemometric sensor and the system of equations from a kinetic model are combined using an EKF algorithm. This enables to deliver dynamically updated predictions of all reaction states not available for measurement. The experimental data set encompasses nine batch conjugations and two fed-batch conjugations. Both kinetic model and chemometric sensor model are calibrated on the batch data based on the methodology of previously published works (Andris, Rüdt, et al. 2018; Andris, Seidel, and Hubbuch 2019). For the chemometric sensor, a novel GPR model is evaluated against benchmark alternatives and combined with different preprocessing strategies. Subsequently, an EKF is deployed as a soft-sensor fusing the information of the GPR model with the kinetic model. In detail, this approach recursively updates the unmeasurable ADC species predicted from the kinetic model by using the time-discrete GPR predictions of the conjugated drug concentration. The influence of time-variant uncertainties for the GPR and kinetic model are investigated based on the methods given in literature and the optimal configuration is selected. Finally, the robustness and accuracy of the EKF is assessed by comparison to the two underlying models in different scenarios. The described workflow suggests an unified calibration routine for the underlying models, enables the quantification of model uncertainties and assesses the performance boost of the developed soft-sensor.

3.2 Materials and methods

3.2.1 Experiments

3.2.1.1 Chemicals and buffers

The chemicals used in this study were obtained from Merck KGaA, unless otherwise stated. Standard working buffer was made with NaH_2PO_4 , dihydrate (VWR International GmbH) and

ultrapure water (PURELAB Ultra, ELGA LabWater). The buffer was titrated to pH 7.2 with 4M NaOH and filtered through a 0.2 μm cellulose acetate membrane filter (Sartorius AG). For the mAb reduction and oxidation step tris(2-carboxyethyl)phosphine hydrochloride (TCEP) and (L)-dehydroascorbic acid (DHAA) was used. To mimic a cytotoxic drug, the non-toxic surrogate payload N-(1-pyrenyl)maleimide (NPM) was used. For quenching of the conjugation reaction, unconjugated free drug was quenched with N-acetyl cysteine (NAC). For the reversed-phase ultra-high performance liquid chromatography (RP-UHPLC), acetonitrile (VWR) and trifluoroacetic acid (TFA, Thermo Scientific) was used.

3.2.1.2 mAb functionalization and conjugation procedure

The reactions were conducted using an engineered IgG1 mAb with two inserted cysteines in the hinge region which was generously provided by AstraZeneca. All chemicals were dissolved in 50 mM sodium phosphate buffer, except NPM which was dissolved in dimethyl sulfoxide (DMSO). In the beginning, the mAb was thawed, diluted and three reaction steps were consecutively performed: a reduction reaction to uncap the engineered cysteines, a dialysis to remove excess TCEP and a re-oxidation reaction of the reduced interchain disulfide bonds. This functionalization procedure was conducted under the same methodology and conditions as previously described in (Andris, Seidel, and Hubbuch 2019). The subsequent conjugation reaction was performed in a stirred glass beaker (400 rpm) with a reaction volume of 10 mL. Batch conjugations were started by adding a certain volume of 0.68 mM NPM solution with a pipette to reach the desired molar drug excess. In nine runs, the mAb concentration and drug excess were varied, including some runs in duplicates. Two additional fed-batch runs were performed by adding the NPM solution with a syringe pump (Nemesys S, Cetoni GmbH, Germany) constantly over 5 min and 10 min for run 10 and run 11, respectively. All conducted runs are summarized in Table 3.1.

Table 3.1 Overview of conducted conjugation runs using different mAb concentrations and NPM excesses. Batch runs (1-9) were split into training and test set for model calibration, EKF tuning and performance evaluation. Fed-batch runs (10 and 11) were used later in the study to test the transferability of the designed framework.

Experiment	mAb concentration (mg mL^{-1})	Molar drug excess(-)	Number of samples	Subset
Run 1	1	2x	20	Training
Run 2	1	4x	17	Test
Run 3	1.25	3x	21	Training
Run 4	1.5	2x	21	Test
Run 5	1.5	2x	20	Training
Run 6	1.5	3x	18	Test
Run 7	1.5	4x	21	Training
Run 8	1.75	2x	20	Training
Run 9	1.75	3x	20	Training
Run 10 (Fed-batch)	1.5	3x	22	External
Run 11 (Fed-batch)	1.5	3x	20	External

3.2.1.3 UV/Vis on-line monitoring and off-line RP-UHPLC

To monitor the conjugation reaction, the same set-up as in (Andris, Rüdts, et al. 2018) was used. In detail, an on-line loop was installed where a peristaltic pump (Minipuls 3, Gilson, Middleton,

USA) pumped the reaction mixture continuously through the loop tubing. The flow rate was adjusted to 1 mL min^{-1} by measuring the flow rate with a liquid flow meter SLS-1500 (Sensirion AG, Switzerland). To collect UV/Vis spectra every 0.2 s, a DAD-3000 RS diode array detector with a 0.4 mm path length flow cell (both Thermo Fisher Scientific) was installed in the loop. UV/Vis spectra were recorded between 250 and 390 nm at a resolution of 1 nm. Before starting the conjugation reaction, the entire loop was equilibrated with the oxidized mAb solution for up to 15 min and an autozero of the DAD was performed to collect the pure NPM spectra during the reaction.

Off-line samples were taken at certain time points and immediately quenched with a NAC solution. To determine the concentrations of the single ADC species for the time points, all samples were analyzed with RP-UHPLC. The same device, column and analysis method as described in (Andris, Rdt, et al. 2018) was applied. The RP-UHPLC data will further be treated as reference data for all models.

3.2.2 Data organization

The full data set containing the reaction time-courses from the RP-UHPLC and the spectra for each run was split into training and test subsets. The split was done based on a Kennard-Stone algorithm (Kennard and Stone 1969) using the first-derivative mean spectra of each run to ensure that the choice of the calibration set is chosen objectively and spans the full experimental design space. The training subset was used for hyperparameter tuning of the chemometric models, calibration of the kinetic model parameters and tuning of the EKF. The test subset was used for model performance evaluation and selection. The fed-batch runs were evaluated separately. The subsets used for chemometric model development consisted of 122 and 56 reference data points for training/cross-validation and testing, respectively. An overview of all conducted experimental runs is shown in Table 3.1.

3.2.3 Chemometric model development

Initially, all absorbance spectra were averaged over an interval of 5 s and aligned with the RP-UHPLC reference data. Before regression modeling, the spectra were mean-centered and normalized by the maximum absorbance at 250 nm of the training subset. Several chemometric models were calibrated and compared to determine the best performing option. In the following, all mentions of cross-validation refer to a *leave-one-run-out* rotation.

Partial-least squares regression (PLS) models were calibrated in combination with a Savitzky-Golay derivative filter (Savitzky-Golay (SG)) or a variable selection algorithm. The number of latent variables, the window size and order of derivative for the SG were tuned by employing a cross-validated grid search. The variable selection algorithm was adapted from (Mehmood et al. 2012) using the regression coefficients of the PLS as an importance metric and evaluated using cross-validation.

For GPR models, multiple kernel functions were screened and combined with a SG or a variable selection algorithm analogue to the one described above. Multiple combinations of a linear, a radial basis function (RBF) and an automatic relevance determination (ARD) kernel were evaluated. The final implementation of the kernel given in the Supplementary A.1. More theoretical information on GPR modeling can further be found in Ch. A. The kernel parameters were estimated by expectation maximization as implemented in (Pedregosa et al. 2011). The implemented variable selection algorithm was adapted from (Paananen et al. 2019). For the exact mathematical procedure the reader is referred to the cited literature. The computed variable sensitivities are used as an importance metric, ranked and selected as performed for the PLS variable selection algorithm.

All calibrated models were evaluated based on their root mean square error of cross validation (*RMSECV*) and prediction (*RMSEP*). A detailed definition of the error metric used in this study can be found in the Supplementary A.4. All chemometric models presented in this study were implemented with recent versions of Python 3.8, NumPy ≥ 1.20 (Harris et al. 2020), pandas ≥ 1.3 (team 2020), scikit-learn ≥ 1.1 (Pedregosa et al. 2011) and related packages.

3.2.4 Kinetic model calibration

A previously developed kinetic model for the site-directed conjugation reaction (see (Andris, Seidel, and Hubbuch 2019)) was applied in this study to be adopted in the EKF framework. A detailed explanation of the model and the full ODE system can be found the Ch. A. The kinetic rates k_1 , k_2 and k_3 were estimated by non-linear least squares optimization using the *lsqnonlin* solver in MATLAB 2019b (The Mathworks, Inc.). The initial values were set to 0.8 mM s^{-1} , 1.4 mM s^{-1} and 0.01 s^{-1} for k_1 , k_2 and k_3 , respectively. Assuming that measurement errors in each sample are independent and normally distributed with zero mean and variance σ^2 , the parameter covariance matrix $\mathbf{C}_k$ is computed according to Eq. 3.1,

$$\mathbf{C}_k = \text{cov}(\mathbf{k}) = \sigma^2 (\mathbf{J}^T \mathbf{J})^{-1} \quad (3.1)$$

where $\mathbf{J}$ is the Jacobian with regard to the kinetic rates $\mathbf{k}$ and σ^2 is the estimated error variance from the non-linear estimation. The error variance σ^2 is approximated by $RMSEP^2$ as suggested in literature (Valappil and Georgakis 2000; Feidl et al. 2019).

3.2.5 Extended Kalman filter

3.2.5.1 Terminology and concept

As the herein applied models are calibrated using different target variables, the terminology is shortly introduced: The un-, mono- and biconjugated ADC species will be referred to as mAb, Conj1 and Conj2 for the remainder of this study. It is worth mentioning that 'mAb' corresponds to the sum of the three sub-species with different amount of active cysteines: mAb_{2c}, mAb_{1c} and mAb_{0c}, and 'Conj1' to the sum of the two sub-species Conj1_{1c} and Conj1_{0c} of the ODE system. While the kinetic model continuously predicts all species, the chemometric model converts the absorbance spectra to the amount of conjugated drug at discrete time points t_k . The same values can also be computed from the single species predictions by $c_{\text{drug, Conj}} = c_{\text{Conj1}} + 2c_{\text{Conj2}}$, where c_{Conj1} and c_{Conj2} are the concentration of mono- and biconjugated mAbs. In order to enable the sensor model to update the individual ADC species in the ODE system, a special measurement function was chosen (see chapter 3.2.5.2). Furthermore, both model outputs are thought to be noisy as they are subject to errors caused by model assumptions, plant-model mismatch and experimental handling. An EKF framework was designed to combine the noisy model outputs and generate a more accurate estimate than either of the two sources. Since the two used models are not producing estimates for the same set of species, the EKF acts as a soft-sensor updating several unmeasurable species by an indirect measurement of the sum of conjugated drug. For better understanding, the overall concept and data organization is summarized in Fig. 3.1.

3.2.5.2 Mathematical formulation

The theory of the EKF framework is well established and can be found e.g. in (Simon 2006). In principle, all species are represented in a state vector $\mathbf{x}$. The seven ordinary differential equations (ODEs) of the kinetic model act as state transition function f propagating the state vector $\mathbf{x}$ through time according to Eq. (3.2) given deterministic inputs $u(t)$ and the model parameters $\boldsymbol{\theta}$:

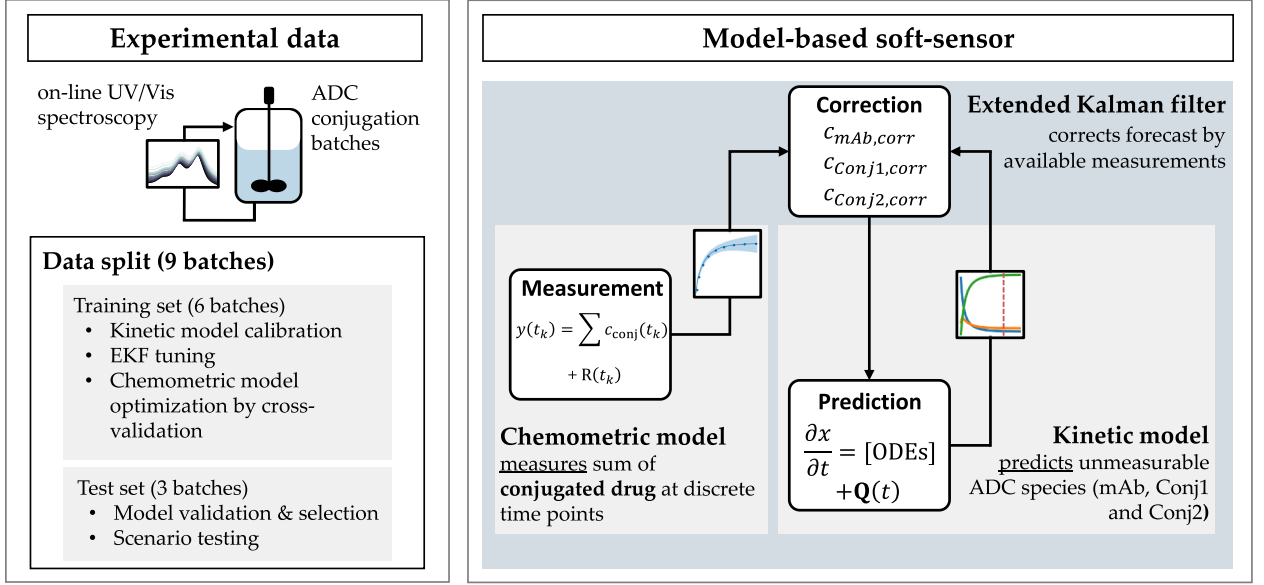


Figure 3.1 Schematic overview of the purposes of the data sets and the methodology of the designed model-based soft-sensor.

$$\frac{\partial \mathbf{x}}{\partial t} = f(\mathbf{x}(t), u(t), \boldsymbol{\theta}) + \mathbf{q}(t) \quad (3.2)$$

with the state vector $\mathbf{x}$ for all seven reacting species in the ODE system defined as:

$$\mathbf{x}(t) = [c_{\text{mAb}_{2c}} \ c_{\text{mAb}_{1c}} \ c_{\text{mAb}_{0c}} \ c_{\text{Conj1}_{1c}} \ c_{\text{Conj1}_{0c}} \ c_{\text{Conj2}} \ c_{\text{NPM}}]^T \quad (3.3)$$

where c_{mAb} denotes the concentration of unconjugated mAbs with the respective number of available cysteines (c) and the conjugates with the respective number of conjugated NPM molecules. The free NPM concentration is given as c_{NPM} . Measurements $y(t_k)$ available at discrete time points t_k are related to states vector using the selected measurement function h which is given in Eq. (3.4). The process noise $\mathbf{q}(t)$ and the measurement noise $r(t)$ are assumed to be additive, zero-mean Gaussian noise:

$$\begin{aligned} y(t_k) &= h(\mathbf{x}(t_k)) + r(t_k) \\ &= [0 \ 0 \ 0 \ 1 \ 1 \ 2 \ 0] \cdot \mathbf{x}(t_k) + r(t_k) \end{aligned} \quad (3.4)$$

When a new measurement $y(t_k)$ is available, the measurement function effectively distributes the measured value to the fourth, fifth and sixth reaction state, i.e. the species with available binding sites. This measurement function was chosen as the available chemometric sensor model solely accounts for the amount of conjugated drug which is equivalent to the sum of the respective reaction states (cf. section 3.2.3). A detailed mathematical formulation of the EKF working principle and its elementary functions can be found in Ch. A. The EKF was implemented using the *ExtendedKalmanFilter* class in MATLAB 2019b.

3.2.5.3 Process noise covariance

The process noise $\mathbf{q}(t)$ is parameterized by the covariance matrix $\mathbf{Q}$ which provides a measure of uncertainty and cross-correlation for the states $\mathbf{x}$. For non-linear process models, as used here, the implementation of a time-variant process noise covariance matrix $\mathbf{Q}(t)$ is crucial to capture process dynamics and allow the EKF to flexibly adjust the weighting between the models. In literature (Valappil and Georgakis 2000), two methods have been suggested to implement a time-variant process noise covariance: A linearized approach that computes a time-dependent process noise covariance matrix $\mathbf{Q}_{\text{lin}}(t)$ based on gradient evaluations with regard to the model parameters. This method uses the Jacobian $\mathbf{J}_{\boldsymbol{\theta}, \text{nom}}$ with regard to the nominal model parameters $\boldsymbol{\theta}_{\text{nom}}$, namely the kinetic rates $\mathbf{k}$, as well as their covariance matrix $\mathbf{C}_{\boldsymbol{\theta}}$ and computes $\mathbf{Q}(t)$ according to Eq. (3.5)

$$\mathbf{Q}_{\text{lin}}(t) = \mathbf{J}_{\boldsymbol{\theta}, \text{nom}} \mathbf{C}_{\boldsymbol{\theta}} \mathbf{J}_{\boldsymbol{\theta}, \text{nom}}^T. \quad (3.5)$$

Alternatively, the Monte Carlo (MC) approach uses linear estimates of the parameter covariance matrix $\mathbf{C}_{\boldsymbol{\theta}}$ to span a normal distribution for all model parameters θ_i with mean μ and variance σ as $\mathcal{N}_{\theta_i}(\mu = \theta_{i, \text{nom}}, \sigma^2 = \mathbf{C}_{\theta_i, \theta_i})$. A definite number of n random samples are drawn from the distribution and the sampled parameter values are used to propagate the states $\mathbf{x}$ to the next time step t_{k+1} . The time-dependent process noise covariance $\mathbf{Q}_{\text{MC}}$ is then computed as the covariance matrix of the mean-centered residuals $\tilde{\mathbf{w}}$ according to Eq. (3.6)

$$\mathbf{Q}_{\text{MC}}(t) = \text{cov}(\tilde{\mathbf{w}}) = \text{cov}(w_n - \bar{\mathbf{w}}) \quad (3.6)$$

with $\bar{\mathbf{w}}$ being the mean of the residuals and the residuals w_n being calculated by

$$w_n = f(t_{k+1}, \boldsymbol{\theta}_n) - f(t_{k+1}, \boldsymbol{\theta}_{\text{nom}}) \quad (3.7)$$

where $\boldsymbol{\theta}_{\text{nom}}$ and $\boldsymbol{\theta}_n$ denote the nominal and the MC samples of the model parameters, respectively.

3.2.5.4 Measurement noise covariance

The measurement noise covariance matrix $\mathbf{R}$ is the sensor-equivalent of $\mathbf{Q}$. As the chemometric sensor in this study only accounts for the amount of conjugated drug, R is a scalar. In contrast to literature as e.g. (Feidl et al. 2019; Narayanan et al. 2020), a time-variant measurement noise covariance $R(t_k)$ may be expressed using the predictive variance of the GPR evaluated at all discrete time points t_k . Further information on the computation of the predictive variance is given in the Ch. A.

3.2.5.5 EKF tuning

When updating the reaction states using a new available measurement, the EKF combines the information stored in $\mathbf{Q}$ and R to produce an improved estimate of all states. Since the absolute values of $\mathbf{Q}$ and R depend on the estimation method, the process noise covariance is scaled to $\mathbf{Q}^* = k_Q \mathbf{Q}$ by a scaling factor k_Q specific for each method. To ensure an equal weighting of all ADC species i , the factor k_Q is optimized with regards to the normalized error $nRMSE$ averaged over all species. The $nRMSE_i$ for each species is calculated according to Eq. 3.8

$$nRMSE_i = \frac{RMSE_i}{\bar{c}_i} \quad (3.8)$$

where $\bar{c}_i$ is the respective mean concentration of each species in each run and $RMSE_i$ being calculated for each species individually as defined in the Ch. A. The total $nRMSE$ is obtained by averaging over all runs and species.

3.2.6 Study design

3.2.6.1 Model-based soft-sensor discrimination

The best performing chemometric model was selected based on the methodology described in section 3.2.3 and a kinetic model was calibrated using the procedures given in section 3.2.4. Based on the selected models, the influence of the different computation approaches for $\mathbf{Q}$ and R were compared. The performance was evaluated based on the prediction error regarding the unmeasurable species in the training and test subsets using the GPR predictions for the conjugated drug concentration. For each approach, the scaling factor k_Q was tuned independently.

3.2.6.2 Challenge scenarios and transfer to fed-batch

The best performing option was selected and further challenged in three scenarios. In scenario 1, the kinetic model was initialized with randomly sampled mAb and NPM concentrations with a standard deviation of 2 % of the nominal mAb/NPM concentration. In scenario 2, the kinetic model was initialized with a randomly sampled cysteine distribution within the experimentally determined standard deviation (see chapter 3.2.4). In the scenario 3, the sensor was fed with noisy absorbance spectra. The spectral data for the test subset was modified by additive Gaussian white noise with a standard deviation of 0.02 % of the respective absorption value. Additionally, a random, normally distributed offset was introduced with a standard deviation of 5 % of the absorbance at 250 nm. In all scenarios, the random sampling and subsequent model evaluation was conducted for 100 repetitions.

Finally, the investigated models were transferred to fed-batch mode. The kinetic model was extended with a feeding term to account for a continuous addition of free drug solution. The GPR model was supplied with the on-line absorbance spectra which were treated as described in section 3.2.3.

3.3 Results and discussion

3.3.1 Chemometric model development

Multiple combinations of PLS and GPR models with different preprocessing strategies were screened in this study and evaluated with regard to their predictive performance on the training and test set. A summary of the results is presented in Tab. 3.2. The PLS model in combination with a SG achieved the lowest $RMSECV$, while the GPR in combination with variable selection (GPR-VS) achieved the lowest $RMSEP$. Overall, the GPR-VS model performed best with regard to both error metrics and improved accuracy by 18 % compared to the PLS model in combination with variable selection (PLS-VS). Hence for further evaluation, only the GPR-VS model is considered. The resulting selected variables for the GPR-VS model are shown in Fig. A.3.

Table 3.2 Overview of predictive performance of selected chemometric models. Both PLS and GPR models were combined with SG and a model-specific VS method.

Model	$RMSECV$ (μM)	$RMSEP$ (μM)
PLS-SG	0.31	0.84
PLS-VS	0.54	0.51
GPR-SG	0.61	0.82
GPR-VS	0.43	0.43

The predictive mean and variance for the (cross-validated) training and test set are shown Fig. 3.2. The predictions for the GPR-VS model show good alignment with the reference data for the training and test set with R^2 of 0.968 and 0.973, respectively. For the cross-validation, the predictions for runs 3, 7 and 9 deviate from the reference data towards the end of the reaction. In run 3, an offset of the predictions from the reference data can be observed starting at roughly 500 s. For runs 7 and 9, the predictions reach a maximum at $17.5 \mu M$ and decrease towards the end of the reaction to $16.0 \mu M$. This tendency is not present in the reference data. The 95 % confidence intervals were approximated by ± 1.96 predictive standard deviations σ^* and show a growing uncertainty with time for all runs in all three subsets. For the cross-validation, the confidence intervals of runs 3, 7 and 9 are considerably wider than for the remaining runs indicating an increased uncertainty of the GPR-VS model. For the test subset, the predictions for run 4 deviate only $1 \mu M$ from the mean at a 95 % confidence as opposed to $2 \mu M$ for run 6.

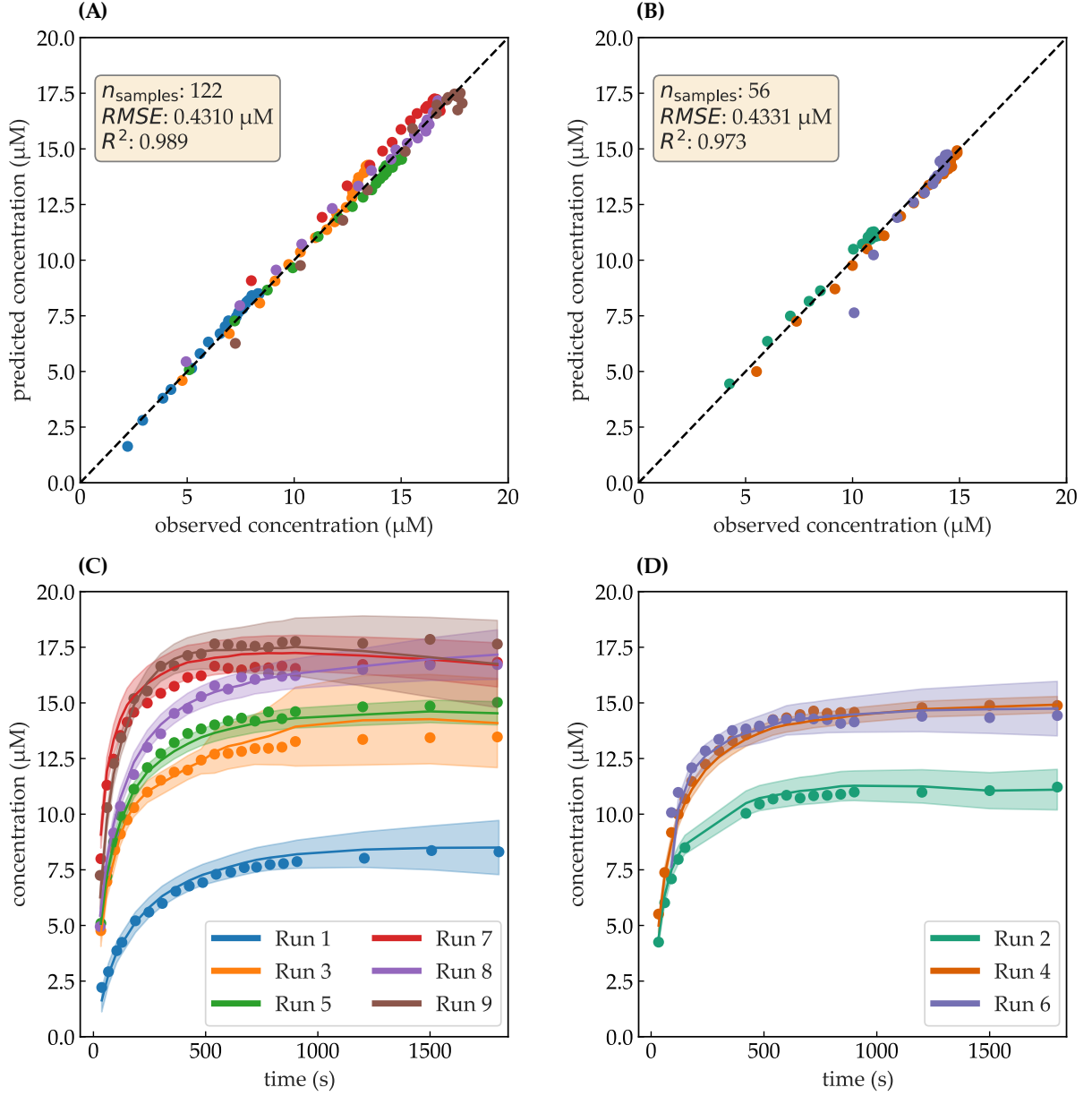


Figure 3.2 Predictions for the GPR-VS model. (A) and (B) show the parity plots for the cross-validation and test subset. The tie line is depicted as dashed line and represents ideal model predictions. The data points are colored by the run numbers. (C) and (D) show the predicted conjugated drug concentration over time. Reference data are illustrated as circles and colored by runs. The GPR-VS predictive mean is showed as solid lines, while the 95 % confidence interval is indicated by the colored areas. The 95 % confidence interval were approximated by 1.96 standard deviations. Note that the GPR-VS model was only evaluated at the reference data points and the lines are linearly interpolated to guide the eye.

3.3.2 Soft-sensor development

3.3.2.1 Kinetic model calibration

Initially, the end-point compositions of runs with a drug excess of 3x or higher were used to determine the percentages of pre-inactivated cysteines as proposed by (Andris, Seidel, and Hubbuch 2019). This resulted in 85.98, 10.14 and 3.93 % of mAbs with two, one and none available cysteines, respectively. Based on this distribution and the training data, the kinetic rates k_1 , k_2 and k_3 were estimated to $0.784 \text{ mM}^{-1} \text{ s}^{-1}$, $1.625 \text{ mM}^{-1} \text{ s}^{-1}$ and 0.0012 s^{-1} . The parameter covariance matrix was calculated by Eq. 3.1 for the computation of the process noise covariance for the EKF. The calibrated model resulted in R^2 values for mAb, Conj1 and Conj2 of 94 %, 95 % and 98 % in the training data, and of 93 %, 94 % and 96 % for the test data.

3.3.2.2 Tuning and selection of covariance matrices

In this section, the performance of the EKF with different approaches for the calculation of process noise $\mathbf{Q}$ and measurement noise R are compared. Each combination was tuned individually by the optimization of the scaling factor k_Q . The effect of the scaling factor on the $nRMSE$ is exemplarily shown for one $\mathbf{Q}/R$ combination in Fig. 3.3. Values in the range of $k_Q = 10^9 \dots 10^{11}$ result in proper tuning so that the EKF beneficially considers both models in order to minimize the $RMSE$ for all species. The optimal k_Q varies slightly in each combination of $\mathbf{Q}$ and R as shown in Tab. 3.3 together with the accomplished $RMSE$ reduction by the EKF with regard to the kinetic model predictions. It could be demonstrated that the linearized method does not lead to an overall improvement, while the MC approach achieves an overall $RMSE$ reduction in both subsets. Only the prediction error for Conj1 in the training set slightly increases. In summary, this approach yields an averaged $RMSE$ reduction of 8.7 % and 17.2 % for the training and test set, respectively. The time-variant computation of R could further improve the performance by up to 4 % for all species compared to the constant setting. In conclusion, the EKF using the $\mathbf{Q}_{MC}$ approach in combination with the R variable setting presents the most valuable combination and is used for the remainder of this study.

Table 3.3 Comparison of the three studied combinations of $\mathbf{Q}$ and R regarding their accomplished $RMSE$ reduction compared to the kinetic model in training and test set.

Q	R	tuned k_Q	Species	$RMSE$ reduction (%)	
				training	test
linearized	variable	9.48	mAb	2.19	2.11
			Conj1	-1.98	3.78
			Conj2	0.19	0.87
MC	variable	10.11	mAb	17.37	11.95
			Conj1	-6.67	23.30
			Conj2	21.83	11.07
MC	constant	10.25	mAb	17.78	8.04
			Conj1	-8.86	21.04
			Conj2	23.24	7.06

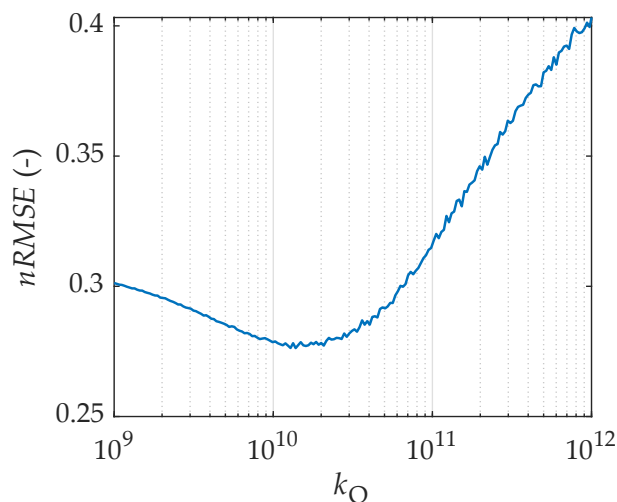


Figure 3.3 EKF prediction error regarding the $nRMSE$ during k_Q tuning

3.3.2.3 EKF soft-sensor performance

The behavior and the soft-sensing capability of the selected EKF combination are exemplarily demonstrated for run 3. Fig. 3.4A and B present the predicted conjugated drug concentration as measurable states from the GPR, kinetic model and EKF, as well as the predicted ADC species as the unmeasurable states from the kinetic model and the EKF, respectively. The reference data is overlayed in both figures. During the initial phase until approx. 500s, it can be observed that the kinetic model predicts the conjugation reaction rate to be faster compared to GPR sensor, which closely follows the reference data. For the unmeasurable states, the same trend is apparent for Conj1 and Conj2. After approximately 200s, the EKF starts to adjust the trajectory by following the GPR estimates. Towards the end, the kinetic predictions converge to the reference data, while the GPR sensor overestimates the conjugation progress after roughly 600s. In the same interval, the EKF switches and clearly follows the kinetic model to be aligned with the reference data. This trend can be similarly followed in Fig. 3.4C and 3.4D, where the normalized time-variant process and sensor noise covariance, and the resulting Kalman gain are displayed. In the beginning, the measurement noise is comparably lower than the process noise regarding all ADC species. This leads the Kalman Gain, and consequently the EKF, to rely more on the GPR predictions with values close to 1. Over the course of the reaction, the process noise is continuously decreasing as the reaction is slowing down and the GPR noise increases conversely relative to the measurement noise. This causes the Kalman gain to shift towards 0 resulting in the EKF relying more strongly on the kinetic model towards the end of the reaction. A similar adjustment dynamic could be observed for all runs in the training and test set and shows the advantage of using time-variant process and measurement noises. The fact that the EKF is converging towards the kinetic model at the end of the reaction, comes along with a decreasing state covariance which may be considered an internal EKF uncertainty. In Fig. A.5, the state covariances for all species are shown for the test subset. For runs, where the Kalman Gain is close to 0 at the end of the reaction, the variance in the EKF estimates is small. Whereas, in Run 4 where the Kalman Gain plateaus at approximately 0.75 due to the high confidence of the GPR sensor, a larger variance is observable. This variance is in agreement with additional data from replicate runs with the same starting conditions which were

not used otherwise within this study. Furthermore, the model and EKF predictions for all training and test runs are depicted in the Figs. A.4 and A.6.

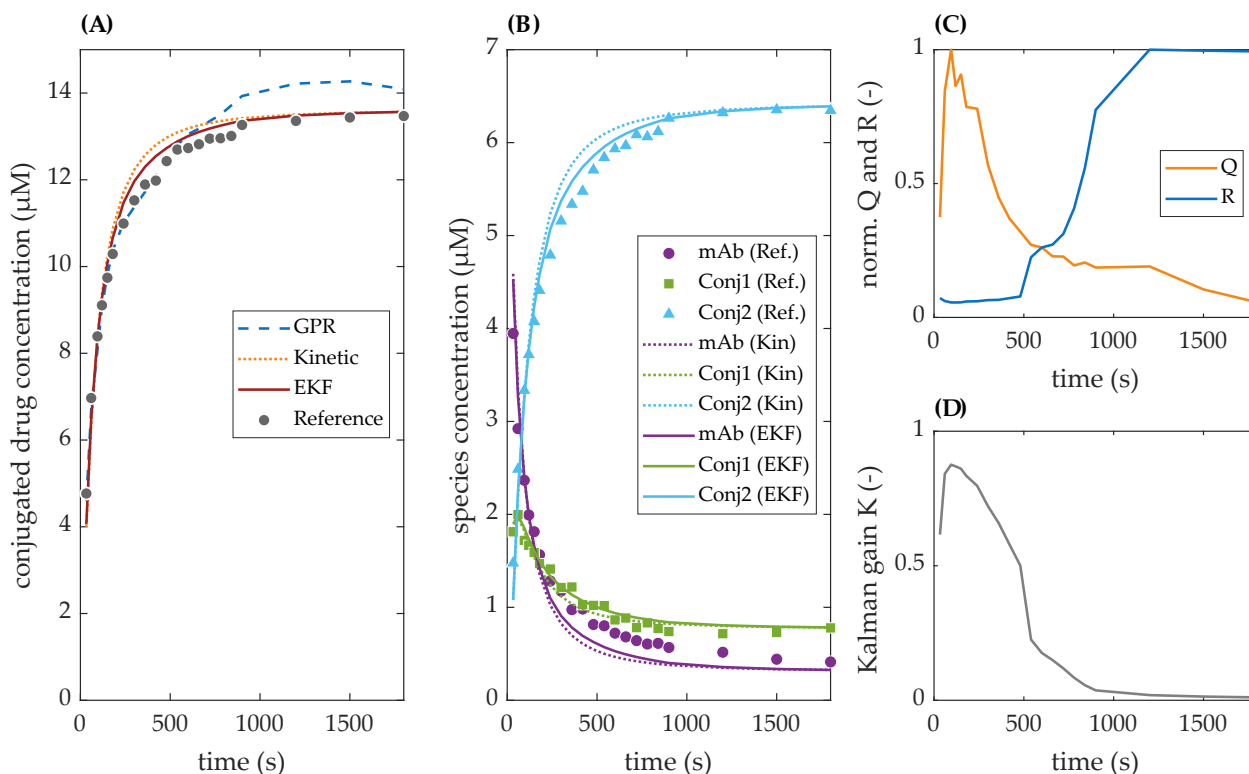


Figure 3.4 Model predictions and calculated process and sensor noise covariance for run 3. (A) Comparison of the conjugated drug concentration as the measured state for the GPR sensor, the kinetic model and the EKF. The reference data from the RP-UHPLC are displayed as circles. (B) Comparison of the predictions for the three ADC species (mAb, Conj1 and Conj2) as the unmeasurable states for kinetic model and the EKF. (C) Time course of the normalized process and sensor noise covariance $\mathbf{Q}_{MC}$ and R_{var} and (D) of the Kalman gain K .

3.3.2.4 Soft-sensor challenge scenarios

Three challenging scenarios were created to further investigate the robustness and accuracy of the designed EKF in comparison to the sole kinetic and sensor model. The resulting statistical distributions of the *RMSEP* after 100 repetitions are displayed as box plots in Fig. 3.5.

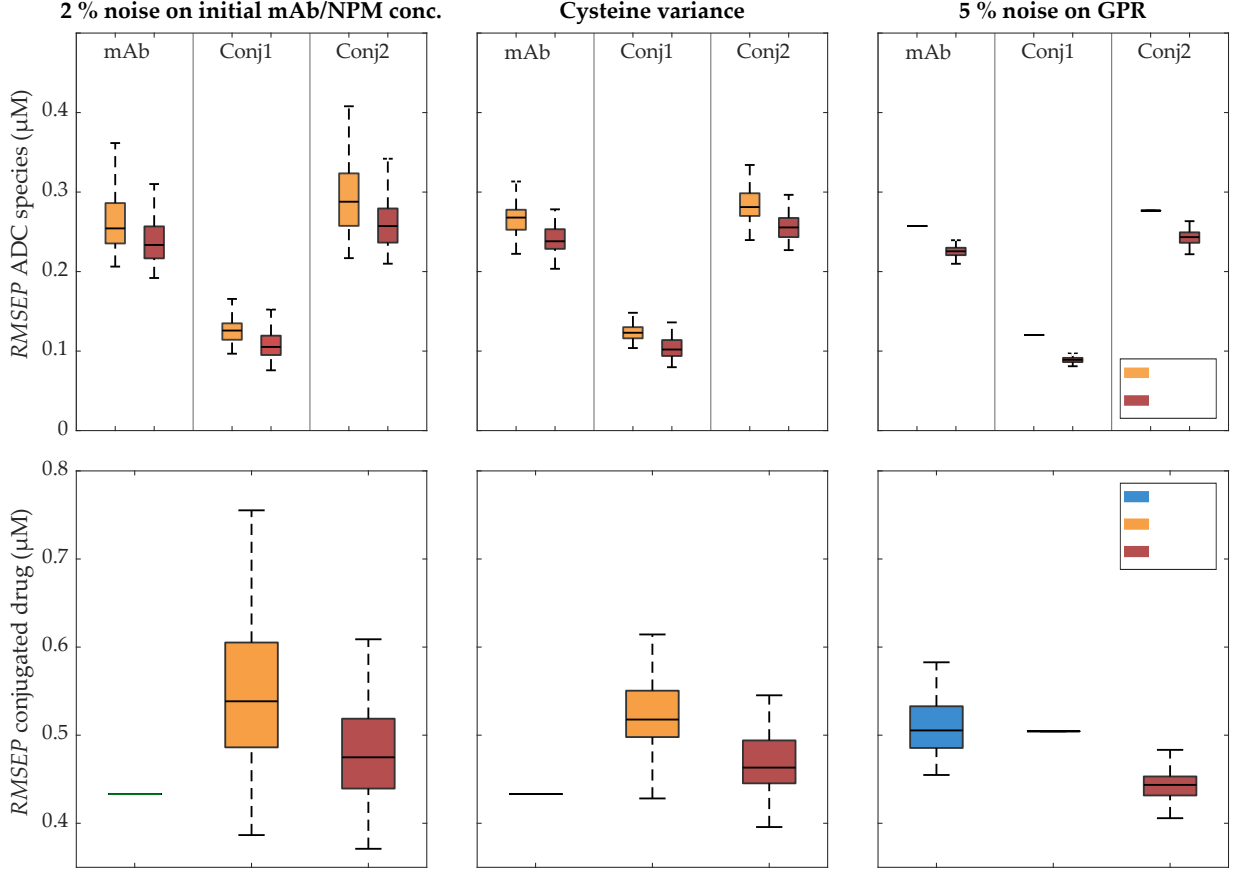


Figure 3.5 Comparison of the *RMSEP* values of the GPR, the kinetic model and the EKF obtained from 100 simulations for each scenario. The upper row shows the *RMSEP* values regarding all ADC species and the lower row the *RMSEP* values regarding the conjugated drug concentration.

The upper and lower row present the comparison of the *RMSEP* distributions with regard to the ADC species and to the conjugated drug, respectively. In the first column, the effect of noisy initial concentrations of mAb and NPM (scenario 1) is depicted. By fusing the information from the unmodified GPR sensor with the faulty initialized kinetic model, the EKF achieves a reduction of the mean prediction error by 10.7, 21.5 and 18.2 % compared to the single species predictions, mAb, Conj1 and Conj2, from the kinetic model. However, the interquartile ranges of the displayed boxes are overlapping and their span is similar for all species indicating that the noisy concentrations are projected to the EKF estimates. Regarding the conjugated drug, the EKF reduces the error by 18.0 % compared to the kinetic model while not reaching the baseline of the GPR error entirely. In the second column, the effect of deviation in the activated cysteines (scenario 2) is shown. All achieved EKF reductions are similar to the results in scenario 1 with overall more dense distributions of the noisy data in both the kinetic model and EKF estimates. The third column demonstrates the effect of noise in the sensor data (scenario 3). With regard to the *RMSEP* of the single species, a reduction in the range of 12-25 % compared to the kinetic model is still accomplished by the EKF.

In this case, the EKF effectively reduces the error in the conjugated drug by 13 % compared to the noisy GPR as measured by the mean of both distributions.

3.3.2.5 Transfer to fed-batch mode

Finally, the EKF was applied to fed-batch runs in order to test the ability of the EKF to perform under different reaction dynamics. In these runs, the relevant wavelengths exhibit a different behavior compared to the experiments operated in batch mode due to a timely overlay of spectral effects from the conjugation reaction and the NPM feed. A comparison of the spectral data from fed-batch and batch mode can be found in the Fig. A.1. To compensate for this changing effect in the absorbance spectra, the GPR was re-calibrated with fed-batch run 10 being part of the training subset. Fig. 3.6 illustrates the model predictions for the fed-batch run 11. The modified kinetic model closely aligns with the reference data until a reaction time of 750 s. Towards the end of the reaction, the modified kinetic model underestimates the conjugated drug concentration, while the GPR sensor follows the reference data. Combining the information of both models, the EKF shows minor improvements during the whole reaction and all species while converging to the predictions of the modified kinetic model towards the end of the reaction. Thus, the EKF errors for all species are improved compared to the kinetic model by 13.2, 0.2 and 16.4 %. For the conjugated drug, the EKF follows the kinetic model after approx. 750 s and hence results in a relative error increase by 37.9 % compared to the GPR model.

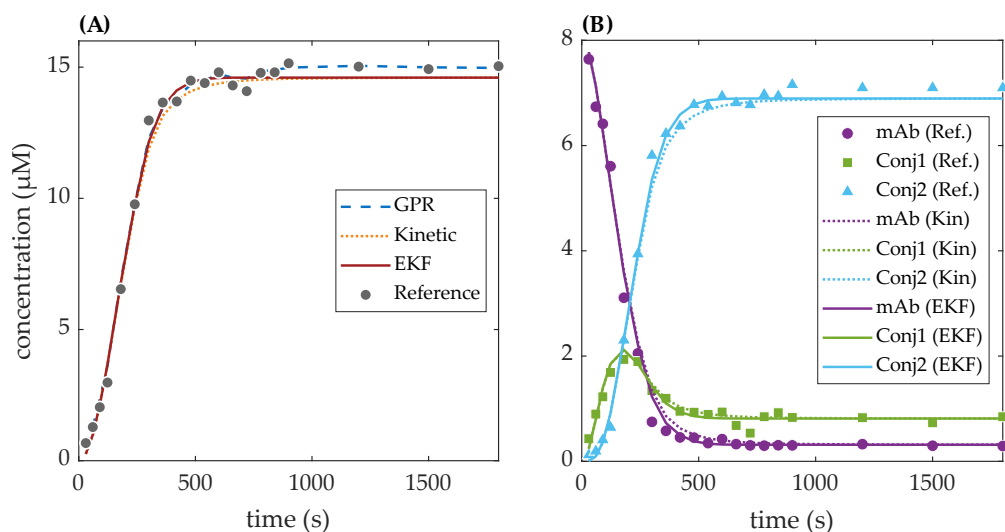


Figure 3.6 Model predictions for fed-batch run 11 for the concentration of conjugated drug (A) and the ADC reaction species (B). To compensate for a changing effect in the absorbance spectra compared to the experiments operated in batch mode, the GPR sensor model was re-calibrated with run 10 being part of the training subset.

3.4 Discussion

3.4.1 Chemometric model development

3.4.1.1 Preprocessing and hyperparameter optimization

The developed GPR model presented in this study uses a composite kernel function and a variable selection algorithm based on the Kullback-Leibler divergence. The model showed elevated predictive

performance compared to commonly applied PLS models and other combinations of kernel functions and preprocessing methods for the GPR (see Tab. 3.2).

SG did not improve the predictive capability for either of the two studied regression models. In Andris, Rüdt, et al. (2018), a baseline correction and subsequent first-derivative filter was found to achieve the best performance, while the herein used GPR model does not rely on any further preprocessing method prior to variable selection. The herein implemented variable selection methods notably improved the predictive accuracy for both the PLS and the GPR model. This suggests that the absorbance spectra contain uninformative regions which can effectively be removed by the employed algorithms. Particularly wavelengths above 350 nm were found to have a negative impact on the sensor model performances due to the presence of spectral artifacts probably caused by the surrogate drug molecule (cf. Fig. A.1).

The implemented kernel function which is a combination of the linear and the RBF kernel has previously been shown to be a viable combination for chemometric purposes in (H. Chen, Bakshi, and Goel 2007). In other cases, the RBF kernel alone was used for GPR models using near-infrared and Raman spectroscopy data (T. Chen, Morris, and Martin 2007; Cui and Fearn 2017; Tulsyan et al. 2021). In general, the authors consider the usage of covariance kernels beneficial to model the non-linear observed spectral shift and baseline drift of the absorbance spectra compared to the linear subspace projections the PLS models are based upon. It has previously been shown, that peak shifts which may be resembled by closely overlapping absorption maxima of multiple components have a strong influence on the regression vector in PLS modeling which may be a reason for the inferior performance compared to kernel-based models (Kjeldahl and Rasmus Bro 2010; Witjes et al. 2000).

3.4.1.2 Quantification of spectra-based uncertainty

When the GPR model is evaluated at an unseen spectrum, the kernel function computes the covariance with all available training instances and uses that information to quantify the prediction estimate and the associated uncertainty. The uncertainty estimate is hence influenced by the distribution of available data points along the process as well as the spectral differences among the samples. As emphasized in Fig. 3.2, the trajectories of run 4 and 6 closely resemble one another. However, the associated uncertainty estimate is considerably higher for run 6. This observation is in agreement with the experimental data as run 4 is a replicate of run 5 and has been conducted on the same day. Run 6, instead, has a different initial condition than all training runs suggesting higher uncertainty due to unseen spectral variation. Similarly, run 3 and run 9 supposedly exhibit spectral features causing the covariance function to produce a wider confidence interval than the other runs. In fact, the variance increases towards the end of the reaction when the conjugation reaction is considerably slowing down. This may be influenced by an irregular chemical degradation of the NPM molecule which has been observed in experiments with pure NPM solutions (cf. Fig. A.2). However, this effect may not occur for other molecules such as real payloads due to incomparable spectral properties (Ouyang 2013).

It should be noted that GPR models are a comparably novel method in chemometrics and the physical validity of the computed confidence intervals has not been extensively studied in the literature. As the confidence intervals are computed with regard to both the experimental variance of the reference and the spectral data, the predictive uncertainty of the GPR can not directly compared to the variance of the reference data alone. Nonetheless, the implemented method provides further insights into the quality of the spectral data and the relation with the training instances and exceeds conventional models in terms of predictive accuracy. The GPR is therefore

considered sufficiently accurate to be used as a sensor within the soft-sensor framework. In future, separate studies specifically designed for the validation of predictive uncertainty estimates should be conducted.

3.4.2 Kinetic model calibration

The identified kinetic reaction rates are in agreement with the ones found in (Andris, Seidel, and Hubbuck 2019) with k_1 and k_2 being 1.6 and 9.2 % below and above the values in the referenced study. The difference might originate from minor differences in the determined cysteine distribution or the overall experimental or analytical procedure. The distribution of pre-inactivated cysteines is a necessary model parameter to account for actual amount of available bindings sites on the oxidized mAb which in practice are lower than two. This phenomenon is already discussed in detail in (Cao et al. 2019).

3.4.3 Soft-sensor development

The selection and appropriate tuning of the process and sensor noise covariance play a crucial role in the development of an optimal state estimator such as an EKF (Elsheikh et al. 2021). It has been observed that the EKF mainly especially selects the Conj1 state to compensate for the deviation when k_Q is set too high. This unfavourable adjustment is due to the underlying physical constraints from the kinetic model ODE system and could be effectively minimized by using a normalized *RMSE* for tuning.

(Valappil and Georgakis 2000) stated that the applicability of the computation method for $\mathbf{Q}(\mathbf{t})$ has to be systemically tested for each individual case. In our study, the MC method outperforms the linearized method in terms of prediction error (see Tab. 3.3). Remarkably, the EKF using the MC method improves the errors for all ADC species in the test set of at least 11 %. It appears that the sampling-based MC method is more adequate and robust in the case of the non-linear ADC reaction. However, it should be noted that the parameter distributions used for the MC method are solely approximations. For increasingly non-linear models, Gaussian distributions may no longer be assumed for all model parameters. Another disadvantage of the MC method may be the diminishing process variance and hence an overconfident EKF state estimate towards the steady-state of the reaction. If applicable, Bayesian inference methods could be used to assess more accurate estimations of the parameter distributions and provide a means for more conservative variance estimation.

The sensor noise covariance R is commonly set to a constant scalar and derived from the calibration error of the sensor model (Feidl et al. 2019; Elsheikh et al. 2021; Narayanan et al. 2020). The usage of the GPR model firstly allows the implementation of a time-varying sensor noise covariance. As the uncertainty derived from the GPR increases with time, the sensor noise resembles this trend and the EKF counteracts to this by relying more on the kinetic model towards the end of the reaction. In comparison to a time-invariant R , the variable approach achieves small improvements, while the setting of $\mathbf{Q}$ majorly contributes to the presented error reduction. However, it should be emphasized that in this case the GPR provides an estimate of the cumulative conjugated drug and does not affect all states individually. It would be hence interesting to evaluate the performance of the variable sensor noise in cases where direct measurements of all states are available. Moreover, another drawback of the implemented EKF are the steady-state constraints, i.e. the experimentally derived cysteine distribution assumed for all batches, which make the EKF rather stiff towards the end of the reaction. This is a known problem of the EKF (Tul’syan et al. 2021) when dealing with only a few states being available for measurement as well as for non-linear models. In theory,

this problem could be circumvented with other more flexible state estimation algorithms, such as particle filters as implemented in (Golabgir and Herwig 2016; Stelzer, Kager, and Herwig 2017) or unconstrained Kalman filter algorithms as used in (Simutis and Lübbert 2017; Kolås, Foss, and Schei 2009) which were beyond the scope of this study.

The three challenging scenarios prove that even for a faulty EKF initialization and a noisy sensor model, the designed soft-sensor consistently leads to more accurate predictions compared to the sole chemometric or kinetic models. In cases of large offsets of the kinetic model, the EKF could be averted to rely on the sensor model instead, which is demonstrated in challenge scenario 1 and 2. In case of noisy sensor inputs, the EKF produces less noisy predictions, also regarding the unmeasured species. As discussed before, the predictions of both the process and the sensor model are dependent on the variations in the absorbance spectra or initial conditions. The EKF was shown to be able to compensate foreseeable deviations and hence demonstrated improved robustness which is essential for the successful deployment of the soft-sensor.

Finally, the adaptability of the soft-sensor was investigated by transfer to fed-batch operation mode. Under fed-batch conditions, the EKF also performed adequately while an extended re-tuning was not required. Despite the necessary addition of the fed-batch run 10 to the training data set for the GPR model (see chapter 3.3.2.5), the configuration of the model remained the same. It is therefore recommended to perform multiple fed-batch experiments with varying initial conditions if the chemometric model is intended to be used for this operation mode. In the studied case, the kinetic model aligns perfectly with the reference data due to accurate initial values and precise feeding. Regarding the transferability to other setups or ADC reactions, a re-tuning of the EKF may be necessary as soon as one of the two underlying models' performances change, as the EKF relies on weighting of the two models. However, a thorough investigation this was out of the scope of this study. The scenarios and the fed-batch transfer showed that the EKF is less sensitive to deviations and could be implemented for model-based process control, e.g. for endpoint determination of the conjugation reaction.

3.5 Conclusions

In the ADC conjugation reaction, a cytotoxic payload is conjugated to a previously functionalized mAb with different levels of available binding sites. Current multivariate measurement approaches only enable the on-line measurement of the amount of conjugated drug as a whole but not the ADC species along the conjugation pathway. In this study, a robust and adaptive model-based soft-sensor for the advanced monitoring of a site-directed ADC conjugation reaction was presented combining a chemometric sensor based on a GPR model with a kinetic reaction model. The devised GPR model exhibited superior accuracy compared to commonly applied PLS models and provided uncertainty estimates for each measurement. The associated uncertainty is directly influenced by the variation in the spectral data and could provide a means for risk-based monitoring approaches. The deployed soft-sensor EKF fuses the information from the GPR model with the timely forecasts of the kinetic model. The EKF is able to dynamically adjust the predictions of all unmeasurable ADC species throughout the entire reaction using time-variant computation approaches for the process and sensor noise covariances. The designed EKF improved the predictions on average by 8.7 and 17.2 % for the training and test set, respectively. The EKF was further shown to be capable to compensate for experimental noise effects, wrong model initialization and poorly determined kinetic model parameters compared the unmodified kinetic or sensor model. In conclusion, the combination of a chemometric sensor and a kinetic model using an EKF-based soft-sensor is highly beneficial for accurate process monitoring and enables a high degree of flexibility with regard to

process deviations. This ultimately helps to facilitate robust process control or early end-point determination. The presented methodology further ensures a unified calibration routine for all underlying components and provides knowledge on sensor and model uncertainties.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author contributions

Conceptualization, D.G.G.; funding acquisition, J.H. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

We would like to thank student A and student B for their valuable contributions in the form of experimental work for this project. We acknowledge support by the KIT-Publication Fund of the Karlsruhe Institute of Technology.

Data availability statement

The raw data supporting the conclusions of this article as well as the written codes for Matlab® will be made available on request, inquiries can be directed to the corresponding author

Chapter references

- Andersson, Martin (2009). „A comparison of nine PLS1 algorithms“. In: *Journal of Chemometrics: A Journal of the Chemometrics Society* 23.10, pp. 518–529.
- Andris, Sebastian and Jürgen Hubbuch (2020). „Modeling of hydrophobic interaction chromatography for the separation of antibody-drug conjugates and its application towards quality by design“. In: *Journal of Biotechnology* 317.April, pp. 48–58.
- Andris, Sebastian, Matthias Rüdert, et al. (12/2018). „Monitoring of antibody-drug conjugation reactions with UV/Vis spectroscopy“. In: *Journal of Biotechnology* 288.October, pp. 15–22.
- Andris, Sebastian, Jonathan Seidel, and Jürgen Hubbuch (2019). „Kinetic reaction modeling for antibody-drug conjugate process development“. In: *Journal of Biotechnology* 306.September, pp. 71–80.
- Cabaneros Lopez, Pau et al. (2020). „Towards a digital twin: a hybrid data-driven and mechanistic digital shadow to forecast the evolution of lignocellulosic fermentation“. In: *Biofuels, Bioproducts and Biorefining* 14.5, pp. 1046–1060.
- Cao, Mingyan et al. (2019). „Site-specific antibody-drug conjugate heterogeneity characterization and heterogeneity root cause analysis“. In: *mAbs* 11.6, pp. 1064–1076.
- Chen, Hongshu, Bhavik R. Bakshi, and Prem K. Goel (2007). „Toward Bayesian chemometrics-A tutorial on some recent advances“. In: *Analytica Chimica Acta* 602.1, pp. 1–16.
- Chen, Tao, Julian Morris, and Elaine Martin (2007). „Gaussian process regression for multivariate spectroscopic calibration“. In: *Chemometrics and Intelligent Laboratory Systems* 87.1, pp. 59–71.
- Coumans, Ruud G.E. et al. (2020). „A Platform for the Generation of Site-Specific Antibody-Drug Conjugates That Allows for Selective Reduction of Engineered Cysteines“. In: *Bioconjugate Chemistry* 31.9, pp. 2136–2146.
- Cui, Chenhao and Tom Fearn (02/2017). „Comparison of partial least squares regression, least squares support vector machines, and Gaussian process regression for a near infrared calibration“. en. In: *Journal of Near Infrared Spectroscopy* 25.1 (1), pp. 5–14.
- Elsheikh, Mohamed et al. (2021). „A Comparative Review of Multi-Rate Moving Horizon Estimation Schemes for Bioprocess Applications“. In: *Computers & Chemical Engineering* 146, p. 107219.
- FDA (2004). „Guidance for Industry: PAT—a framework for innovative pharmaceutical development, manufacturing, and quality assurance“. In.
- Feidl et al. (10/2019). „Combining Mechanistic Modeling and Raman Spectroscopy for Monitoring Antibody Chromatographic Purification“. In: *Processes* 7.10, pp. 683–683.
- Gargalo, Carina L. et al. (2021). „Towards the Development of Digital Twins for the Bio-manufacturing Industry“. In: *Digital Twins: Tools and Concepts for Smart Biomanufacturing*. Ed. by Christoph Herwig, Ralf Pörtner, and Johannes Möller. Cham: Springer International Publishing, pp. 1–34.
- Golabgir, Aydin and Christoph Herwig (2016). „Combining Mechanistic Modeling and Raman Spectroscopy for Real-Time Monitoring of Fed-Batch Penicillin Production“. In: *Chemie-Ingenieur-Technik* 88.6, pp. 764–776.
- Großhans, Steffen et al. (2018). „In-line Fourier-transform infrared spectroscopy as a versatile process analytical technology for preparative protein chromatography“. In: *Journal of Chromatography A* 1547, pp. 37–44.
- Harris, Charles R. et al. (09/2020). „Array programming with NumPy“. In: *Nature* 585.7825, pp. 357–362.
- Haseltine, Eric L. and James B. Rawlings (2005). „Critical evaluation of extended Kalman filtering and moving-horizon estimation“. In: *Industrial and Engineering Chemistry Research* 44.8, pp. 2451–2460.
- ICH (2009a). *ICH guideline Q8 (R2) pharmaceutical development*.
- Joubert, Nicolas et al. (2020). „Antibody–Drug Conjugates: The Last Decade“. In: *Pharmaceuticals* 13.9.
- Keller, William R. and Michaela Wendeler (2021). „Using multimodal chromatography for post-conjugation antibody-drug conjugate purification: A methodology from high throughput screening to in-silico process development“. In: *Journal of Chromatography A* 1653, p. 462378.
- Kennard, R. W. and L. A. Stone (1969). „Computer Aided Design of Experiments“. In: *Technometrics* 11.1, pp. 137–148.
- Kjeldahl, Karin and Rasmus Bro (07/2010). „Some common misunderstandings in chemometrics“. en. In: *Journal of Chemometrics* 24.7–8, pp. 558–564.
- Kolås, S., B. A. Foss, and T. S. Schei (2009). „Constrained nonlinear state estimation based on the UKF approach“. In: *Computers and Chemical Engineering* 33.8, pp. 1386–1401.
- Luttmann, Reiner et al. (2012). „Soft sensors in bioprocessing: A status report and recommendations“. In: *Biotechnology Journal* 7.8, pp. 1040–1048.
- Mehmood, Tahir et al. (08/2012). „A review of variable selection methods in Partial Least Squares Regression“. In: *Chemometrics and Intelligent Laboratory Systems* 118, pp. 62–69.
- Narayanan, Harini et al. (2020). „Hybrid-EKF: Hybrid model coupled with extended Kalman filter for real-time monitoring and control of mammalian cell culture“. In: *Biotechnology and Bioengineering* 117.9, pp. 2703–2714.
- Ohadi, Kaveh, Raymond L. Legge, and Hector M. Budman (2015). „Development of a soft-sensor based on multi-wavelength fluorescence spectroscopy and a dynamic metabolic model for monitoring mammalian cell cultures“. In: *Biotechnology and Bioengineering* 112.1, pp. 197–208.
- Ouyang, Jun (2013). „Drug-to-antibody ratio (DAR) and drug load distribution by hydrophobic interaction chromatography and reversed phase high-performance liquid chromatography“. In: *Antibody-Drug Conjugates*. Springer, pp. 275–283.
- Paananen, Topi et al. (12/2019). „Variable selection for Gaussian processes via sensitivity analysis of the posterior predictive distribution“. In: *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)* PLMR.89, pp. 1743–1752.
- Panowski, Siler et al. (2014). „Site-specific antibody drug conjugates for cancer therapy“. In: *mAbs* 6.1, pp. 34–45.

- Pedregosa, F. et al. (2011). „Scikit-learn: Machine Learning in Python“. In: *Journal of Machine Learning Research* 12, pp. 2825–2830.
- Rathore, A. S., R. Bhambure, and V. Ghare (09/2010). „Process analytical technology (PAT) for biopharmaceutical products“. In: *Analytical and Bioanalytical Chemistry* 398.1 (1), pp. 137–154.
- Rolinger, Laura, Matthias Rüdte, and Jürgen Hubbuch (2020b). „A multi-sensor approach for improved protein A load phase monitoring by conductivity-based background subtraction of UV spectra“. In: *Biotechnology and Bioengineering* October.
- Schneider, René and Christos Georgakis (2013). „How to NOT make the extended kalman filter fail“. In: *Industrial and Engineering Chemistry Research* 52.9, pp. 3354–3362.
- Simon, Dan (2006). „Nonlinear Kalman filtering“. In: *Optimal State Estimation*. John Wiley & Sons, Ltd. Chap. 13, pp. 393–431.
- Simutis, Rimvydas and Andreas Lübbert (03/2017). „Hybrid approach to state estimation for bioprocess control“. In: *Bioengineering* 4 (1).
- Sinner, Peter et al. (2021). „Usage of Digital Twins Along a Typical Process Development Cycle“. In: *Digital Twins: Tools and Concepts for Smart Biomanufacturing*. Ed. by Christoph Herwig, Ralf Pörtner, and Johannes Möller. Cham: Springer International Publishing, pp. 71–96.
- Spann, Robert et al. (2018). „A probabilistic model-based soft sensor to monitor lactic acid bacteria fermentations“. In: *Biochemical Engineering Journal* 135, pp. 49–60.
- Stelzer, Ines V., Julian Kager, and Christoph Herwig (2017). „Comparison of Particle Filter and Extended Kalman Filter Algorithms for Monitoring of Bioprocesses“. In: *Proceedings of the 27th European Symposium on Computer Aided Process Engineering – ESCAPE 27*. Vol. 40. Elsevier Masson SAS, pp. 1483–1488.
- team, The pandas development (02/2020). *pandas-dev/pandas: Pandas*. Version latest.
- Thomas, Anish, Beverly A. Teicher, and Raffit Hassan (2016). „Antibody–drug conjugates for cancer therapy“. In: *The Lancet Oncology* 17.6, e254–e262.
- Tulsyan, Aditya et al. (2021). „Spectroscopic models for real-time monitoring of cell culture processes using spatiotemporal just-in-time Gaussian processes“. In: *AIChE Journal* 67.5 (5).
- Tuveri, Andrea et al. (2021). „Sensor fusion based on Extended and Unscented Kalman Filter for bioprocess monitoring“. In: *Journal of Process Control* 106, pp. 195–207.
- Valappil, Jaleel and Christos Georgakis (2000). „Systematic estimation of state noise statistics for extended Kalman filters“. In: *AIChE Journal* 46.2, pp. 292–308.
- Wei, Bingchuan et al. (12/2022). „Multi-attribute Raman spectroscopy (MARS) for monitoring product quality attributes in formulated monoclonal antibody therapeutics“. In: *mAbs* 14.1 (1).
- Witjes, H. et al. (2000). „Automatic correction of peak shifts in Raman spectra before PLS regression“. In: *Chemometrics and Intelligent Laboratory Systems* 52.1, pp. 105–116.
- Wold, Svante, Michael Sjöström, and Lennart Eriksson (10/2001). „PLS-regression: a basic tool of chemometrics“. en. In: *Chemometrics and Intelligent Laboratory Systems* 58.2 (2), pp. 109–130.
- Zhang, Chi et al. (2019). „Rapid, quantitative determination of aggregation and particle formation for antibody drug conjugate therapeutics with label-free Raman spectroscopy“. In: *Analytica Chimica Acta* 1081, pp. 138–145.

4

Generative data augmentation and automated optimization of convolutional neural networks for process monitoring

Robin Schiemer¹, Matthias Rüdert², and Jürgen Hubbuch¹

¹ Institute of Engineering in Life Sciences, Section IV: Biomolecular Separation Engineering, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

² Institute of Life Technologies, HES-SO Valais-Wallis, Sion, Switzerland

Abstract

Chemometric modeling for spectral data is considered a key technology in biopharmaceutical processing to realize real-time process control and release testing. Machine learning (ML) models have been shown to increase the accuracy of various spectral regression and classification tasks, remove challenging preprocessing steps for spectral data, and promise to improve the transferability of models when compared to commonly applied, linear methods. The training and optimization of ML models require large data sets which are not available in the context of biopharmaceutical processing. Generative methods to extend data sets with realistic *in silico* samples, so-called data augmentation, may provide the means to alleviate this challenge.

In this study, we develop and implement a novel data augmentation method for generating *in silico* spectral data based on local estimation of pure component profiles for training convolutional neural network (CNN) models using four data sets. We simultaneously tune hyperparameters associated with data augmentation and the neural network architecture using Bayesian optimization. Finally, we compare the optimized CNN models with partial least-squares regression models (PLS) in terms of accuracy, robustness, and interpretability.

The proposed data augmentation method is shown to produce highly realistic spectral data by adapting the estimates of the pure component profiles to the sampled concentration regimes. Augmenting CNNs with the *in silico* spectral data is shown to improve the prediction accuracy for the quantification of monoclonal antibody (mAb) size variants by up to 50 % in comparison to single-response PLS models. Bayesian structure optimization suggests that multiple convolutional blocks are beneficial for model accuracy and enable transfer across different data sets. Model-agnostic feature importance methods and synthetic noise perturbation are used to directly compare the optimized CNNs with PLS models. This enables the identification of wavelength regions critical for model performance and suggests increased robustness against Gaussian white noise and wavelength shifts of the CNNs compared to the PLS models.

4.1 Introduction

Driven by the FDA initiative in 2004 (FDA 2004), process analytical technology (PAT) has evolved in the past two decades from niche applications to a frequently applied tool widely used in the biopharmaceutical research and manufacturing (E. K. Read et al. 2010; E. Read et al. 2010; Ündey et al. 2010; Glassey et al. 2011; Rüdts, Brestrich, et al. 2017; Sauer et al. 2019; Wei et al. 2022; Bryan Wang et al. 2022). PAT allows to monitor and control processes efficiently and provides means for real-time release testing or in-process prediction of product quality attributes (Jiang et al. 2017; Markl et al. 2020). Optical spectroscopic techniques such as ultraviolet/visible (UV/Vis), infrared (IR) and Raman spectroscopy have been shown to enable real-time monitoring across a wide range of pharmaceutical processes (Bakeev 2005; Rolinger, Hubbuch, and Rüdts 2023; Romann et al. 2022; Trampuž, Teslić, and Likožar 2020; Feidl et al. 2019). In combination with multivariate data analysis, these techniques are e.g. suitable for quantifying product and impurity species from process data (Capito, Skudas, Stanislawski, et al. 2013; Rüdts, Briskot, and Hubbuch 2017; Brestrich, Hahn, and Hubbuch 2016; Brestrich, Rüdts, et al. 2018), identify unknown sample compositions (Jinchao Liu et al. 2017; Wegner, Zimmermann, and Hubbuch 2022), or determine product modifications (M.-Y. Li et al. 2018; Sanden et al. 2019; C. Zhang et al. 2019) owing to their fast and non-invasive characteristics and high selectivity in protein analysis.

Current approaches to the quantitative analysis of spectroscopic data heavily rely on multivariate linear regression methods such as partial least squares (PLS) (Banner et al. 2021). Due to the

linear behavior, these models typically need a limited number of samples for robust calibration and provide comprehensible metrics for critical model evaluation and interpretation (S. Wold, M. Sjöström, and Eriksson 2001). Machine learning (ML) methods have gradually been applied to the field of chemometrics and have been shown to sometimes outperform linear methods on various regression and classification tasks, employing artificial neural networks (ANNs) (Long, Gregoriou, and Gemperline 1990; V. O. Santos et al. 2005), Gaussian process regression (GPR) (Cui and Fearn 2017; Malek, Melgani, and Bazi 2018), support vector machines (SVMs) (Cui and Fearn 2017), k-nearest neighbor (kNN) (Jiarui Wang et al. 2023a) or convolutional neural networks (CNNs) (Acquarelli et al. 2017; Bjerrum, Glahder, and Skov 2017; Cui and Fearn 2018; Blazhko et al. 2021; D. Passos and Mishra 2021; Rolinger, Rüdte, and Hubbuch 2021; Jiarui Wang et al. 2023a). Next to the increased accuracy, ML models were found to reduce the amount of preprocessing needed prior to spectral modeling (Cui and Fearn 2018; Tulsyan et al. 2021; Rolinger, Rüdte, and Hubbuch 2021; Schiemer, Weggen, et al. 2023) and increase robustness against variability in the data (Cui and Fearn 2018; Yuanyuan and Zhibin 2018). Major obstacles to successfully deploy these models for process monitoring in biopharmaceutical operations are the required amount of data for model calibration (Tulsyan et al. 2021; Banner et al. 2021), the high number of hyperparameters (D. Passos and Mishra 2022) as well as the necessity for universally applicable diagnostic tools to reduce the black-box character of these models (Burkart and Huber 2021).

In other branches of ML, where data is more abundantly available, nonlinear methods are in many applications state-of-the-art. Major advances have been made in natural language processing or image analysis by using generative techniques such as data augmentation to further increase the amount and variability of data for building models (Shorten and Khoshgoftaar 2019; Feng et al. 2021). In Bjerrum, Glahder, and Skov (2017), the authors first introduced a data augmentation method used for chemometric CNN models based on simple mathematical modifications of the underlying spectral data to induce artificial offset or slope effects and wavelength shifts. This method was generalized by Blazhko et al. (2021) using the theory obtained from extended multiplicative scatter correction. Both mentioned approaches solely address the variations in the spectral domain and do not extract component-specific information for augmenting experimental data. Other ML approaches have been tested using generative adversarial networks (GANs) (Wu et al. 2021; Mishra and Herrmann 2021; McHardy et al. 2023) or variational autoencoders (VAEs) (F. Guo, Xie, and B. Huang 2020), where the different input data are projected onto so-called latent structures before they are recombined into *in silico* representations. Both GANs and VAEs involve neural network structures and hence increase the overall complexity of the approach due to additional hyperparameters. Alternatively, the feature dimension of the experimental data may be extended by stacking the outputs of multiple preprocessing methods as proposed in (D. Passos and Mishra 2021; Mishra and D. Passos 2021a), however, not addressing the limitation in the number of samples.

Finding the right architecture for the underlying problem and tuning the hyperparameters remains a challenging and laborious task due to a high-dimensional search space and long computation times compared to linear methods (Feurer and Hutter 2019). While several scholars have proposed rather complex architectures resulting in a large number of trainable parameters (Bjerrum, Glahder, and Skov 2017; Jinchao Liu et al. 2017; Blazhko et al. 2021) for their chemometric CNNs, others chose simple architectures employing solely one convolutional layer to maintain interpretability (Acquarelli et al. 2017; Cui and Fearn 2018). Automating the process of architecture search and hyperparameter tuning, which is commonly referred to as hyperparameter optimization (HPO), reduces the amount of manual work needed to build ML models and helps to identify the best overall configuration. Model-based HPO methods such as Bayesian optimization have been shown

to be more efficient at finding the global optimum for computer vision (Bergstra, Yamins, and Cox 2012) and chemometrics (Rolinger, Rüdert, and Hubbuch 2021; D. Passos and Mishra 2021; D. Passos and Mishra 2022) compared to randomized or grid-based approaches.

While linear methods such as PLS are well understood and many evaluation metrics exist to assess model quality, ML models are often considered black boxes due to the increased amount of parameters and different mathematical principles. For CNNs, various visualization methods exist to understand the trained convolutions and the corresponding feature importance (Zeiler and Fergus 2013; Yosinski et al. 2015). Gradient-weighted class activation maps (GradCAMs) as proposed in Selvaraju et al. (2020) have already been applied to chemometrics (Mishra and D. Passos 2021b; D. Passos and Mishra 2021) to provide quantitative insights into the contributions of a specific wavelength. However, GradCAMs are not directly comparable to conventional evaluation metrics for PLS models such as regression coefficients or otherwise computed PLS-specific importance metrics. Additive feature attribution methods such as Shapley additive explanations (SHAP) (S. M. Lundberg, Allen, and Lee 2017) or Shapley additive global importance (SAGE) (Covert, S. Lundberg, and Lee 2020a; Covert, S. Lundberg, and Lee 2020b) provide model-agnostic frameworks to compute quantitative feature importance based on multivariate permutations.

In this manuscript, we develop and implement a novel data augmentation method for generating synthetic spectral data based on the local estimation of the pure component profiles. We further establish a holistic modeling workflow for chemometric data considering data augmentation, HPO, and interpretation. The herein calibrated CNN models are evaluated using three different data sets from protein chromatography employing UV/Vis spectroscopy as well as one publicly available data set using IR spectroscopy. Firstly, the suitability of the proposed data augmentation method to enlarge small experimental data sets is demonstrated and a systematic tuning of the method is performed. Secondly, the optimal configuration of the CNN model is determined by automated HPO. Thirdly, we assess the interpretability of the optimized models by quantification of the importance of individual wavelengths. Finally, the robustness and transferability of the optimized CNNs are studied by *in silico* perturbations and model transfer to external data sets.

4.2 Materials and methods

The evaluation of chemometric CNNs in this manuscript involves multiple steps which are performed on the basis of four data sets. Fig. 4.1 provides an illustrative overview of the individual steps and the data sets used within each step. In this section, the methodology for the individual steps is explained in detail.

4.2.1 Data and equipment

In this study, four data sets based on spectroscopic data were used for the evaluation of the herein-presented methods. Data sets 1-3 originate from chromatography experiments of which the experimental details were presented elsewhere (Brestrich, Sanden, et al. 2015; Brestrich, Rüdert, et al. 2018). Data set 4 was presented in (Blazhko et al. 2021). A summary of the experimental conditions and the data subsets reserved for training and testing of the developed models are given in Tab. B.1. For data sets 1-3, the training and test sets were chosen as presented in the referred literature, where the rationale was to evaluate the trained models on independent chromatography experiments with varying process conditions. For data set 4, a random split was used as no additional information about the underlying experiments was available. In the following, we will refer to all data points in a data set as samples. One sample consists of an absorption spectrum and the corresponding

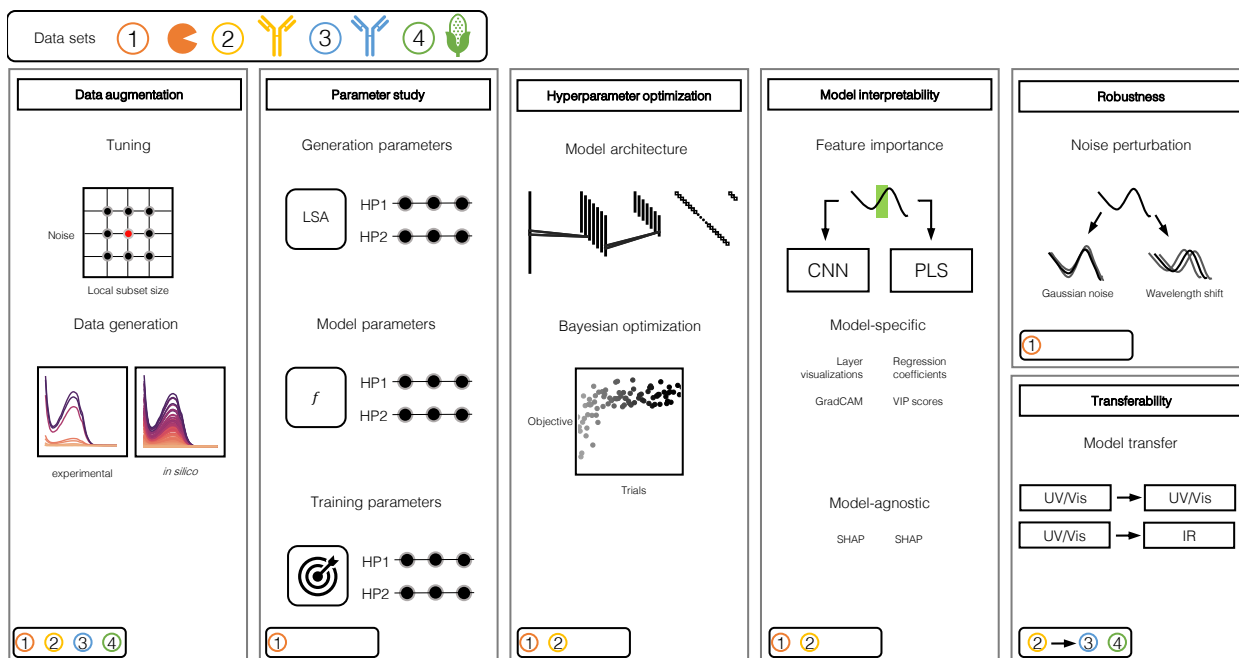


Figure 4.1 Illustration of the study design. The study can largely be divided into 5 steps: 1) Data augmentation, 2) parameter study, 3) automated HPO, 4) model interpretability assessment and 5) robustness and transferability evaluation. The round boxes in the bottom left corners indicate which data set was used for which step. Data set 1 consists of three model proteins, namely ribonuclease A (Rib A), cytochrome C (Cyt C), lysozyme (Lys). Data set 2 and 3 are comprised of monoclonal antibody (mAb) size variants and data set 4 stems from samples of corn.

concentration values obtained by fraction analytics. In data sets 1 to 3, all samples stem from fractions of chromatography elution peaks.

4.2.1.1 Data set 1

Experimental procedures for data set 1 can be found in (Brestrich, Sanden, et al. 2015). The data set consists of 233 samples stemming from five chromatography experiments with varying elution conditions. From each experiment, fractions were collected and analyzed for the concentrations of the three protein components Rib A, Cyt C, and Lys. The experiments were monitored by UV/Vis spectroscopy using a wavelength range of 240 nm to 300 nm at a resolution of 1 nm resulting in 61 features for regression modeling. This data set involves well-studied model proteins and therefore serves for method development within this study.

4.2.1.2 Data set 2

Data set 2 consists of 432 samples stemming from four chromatography experiments with varying elution conditions. The experimental procedures for the data set can be found in (Brestrich, Rüdte, et al. 2018). The concentrations of mAb monomers and aggregates were obtained by fraction analytics. For UV/Vis monitoring, a variable path length spectrometer was used operating at a wavelength range of 240 nm to 340 nm at a resolution of 2 nm, resulting in 51 features for regression modeling.

4.2.1.3 Data set 3

Experimental procedures for data set 3 can be found in (Brestrich, Sanden, et al. 2015). The data set consists of 348 samples stemming from three chromatography experiments with varying elution conditions. The concentrations of the mAb size variants low molecular weight species (LMWS), monomers, high molecular weight species (HMWS)1, and HMWS2 were obtained by fraction analytics. The experiments were monitored by UV/Vis spectroscopy using a wavelength range of 240 nm to 300 nm at a resolution of 1 nm, resulting in 61 features for regression modeling.

4.2.1.4 Data set 4

Data set 4 was obtained from (Blazhko et al. 2021) and consists of in total of 80 samples from IR spectroscopy. The data was obtained from analyzing corn samples and the contents of oil, protein and starch were given as reference values. The spectral range is 1100 cm^{-1} to 2500 cm^{-1} at a resolution of 2 cm^{-1} , resulting in 701 features for regression modeling. The training and test subsets were assigned using a randomized 80:20% split.

4.2.1.5 Hardware and software

Data analysis was done in Python 3.8. Data augmentation was performed using *numpy* (v. 1.19.5), *scikit-learn* (v. 1.1.1) and *scipy* (v. 1.7.3). CNNs were implemented in *tensorflow* (v. 2.5.0). HPO was done in *optuna* (v. 3.1.0) in connection with a MySQLTM8.0 database and *PyMySQL* (v. 1.0.2). SHAP values were computed using *shap* (v. 0.41.0). All computations were done using a workstation equipped with AMD Ryzen 9 3900X 12-core processor and 32GB of memory operating Microsoft Windows 10.

4.2.2 Data augmentation

Before describing the data augmentation method mathematically, the motivation is laid out. In spectroscopy, each molecule is considered to possess a unique spectrum characterized by well-defined extinction coefficients. However, in practical scenarios, various factors such as detector saturation, noise, wavelength shifts, or interfering buffer species can influence the observed absorption spectra. The proposed data augmentation method aims to incorporate these effects by local approximations of the pure component profiles. The method can largely be divided into three consecutive steps: 1) Concentration density approximation, 2) subset selection, and 3) spectra generation. An illustrative overview of the data augmentation method is presented in Fig. 4.2. Mathematically, a given data set can be described as $\mathbf{X}, \mathbf{Y} = \{(\mathbf{x}_i^T, \mathbf{y}_i^T)\}$ for $i \in \{1, M\}$ with M being the total number of samples and $\mathbf{x}_i^T \in \mathbb{R}^{1 \times N}$ and $\mathbf{y}_i^T \in \mathbb{R}^{1 \times P}$ being absorbance spectra with N wavelengths and concentrations of P components, respectively. First, the value distribution in $\mathbf{Y}$ is approximated for each column using a kernel-density estimation as implemented in *scipy.gaussian_kde*. From the approximated distribution, a random concentration vector $\mathbf{y}_*^T$ is sampled. The distance between the sampled vector $\mathbf{y}_*^T$ and all instances in $\mathbf{Y}$ is then computed according to

$$d(\mathbf{y}_*^T, \mathbf{y}_i^T) = \|\mathbf{y}_*^T - \mathbf{y}_i^T\|_l, \text{ for } i \in [1, M] \quad (4.1)$$

with l being the order of the vector norm. As a second step, a number of n_{LSA} samples with the smallest values $d(\mathbf{y}_*^T, \mathbf{y}_i^T)$ are selected to form a local subset of available samples with matrices $\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}$. In the third step, pure-component profiles are estimated based on these local subsets by solving the linear problem as given by

$$\tilde{\mathbf{X}}\tilde{\mathbf{S}} = \tilde{\mathbf{Y}}, \quad (4.2)$$

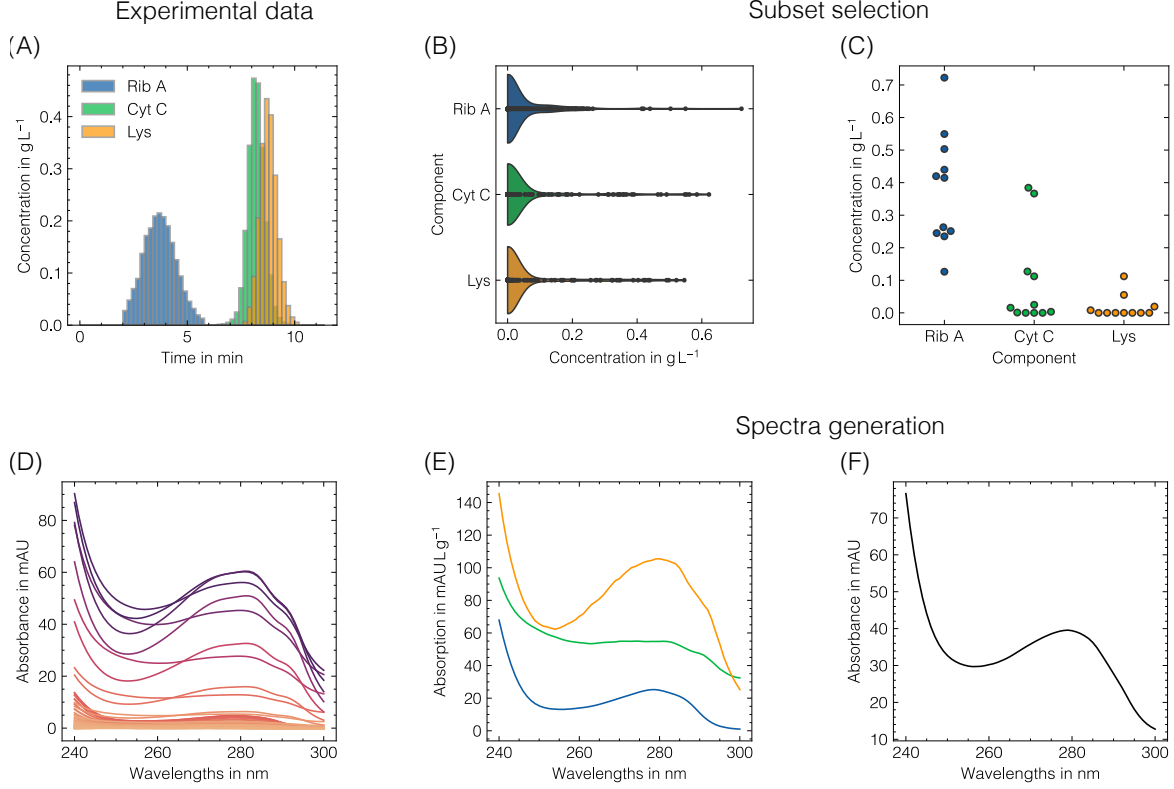


Figure 4.2 Visualization of the local subset augmentation (LSA) data augmentation method. The method is based on experimentally derived response data (A) and spectral data (D). Firstly, the response data distributions are approximated by kernel density functions (B). Second, for a specific sample for which a new spectrum is generated, vector distances for all instances in the supplied experimental data are computed. The n_{LSA} nearest samples in terms of most similar concentration profiles are then selected (C) to form a representative subset of the sampled concentration vector and posing as the basis to derive the pure component spectra (E). Given the sampled concentration vector, the pure component spectra are combined into a newly generated sum spectrum (F).

where $\tilde{\mathbf{S}} \in \mathbb{R}^{N \times P}$ are the estimated pure-component profiles for the local subsets $\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}$. The solution of (4.2) is realized by a ordinary least-squares or non-negative least squares (NNLS) solver to constrain solutions to positive values as implemented in *numpy* or *scipy*, respectively. Thirdly, an *in silico* spectrum $\mathbf{x}_*$ is calculated by

$$\mathbf{x}_* = \tilde{\mathbf{S}}\mathbf{y}^*. \quad (4.3)$$

The motivation behind assembling a subset of samples with closely similar compositions lies in the pursuit of extracting the local differences in the pure component estimations between experimental data points e.g. induced by concentration differences or higher noise contents. By focusing on samples that are closely related in composition, we aim to enhance the quality of the *in silico* spectra. As the generation of *in silico* spectra is based on local subsets of the available data, the method is coined LSA. A more detailed explanation of the approach can be found in the Ch. B. To add more variation to the synthesized spectrum, a Gaussian white noise distributed as $\mathcal{N}(0, \sigma_{\text{noise}})$ is added to each feature and a normally distributed wavelength shift $\mathcal{N}(0, \sigma_{\text{shift}})$ is applied to the entire spectrum $\mathbf{x}_*^T$. In summary, the LSA method presents several configurable parameters, namely $n_{\text{LSA}}, l, \sigma_{\text{noise}}, \sigma_{\text{shift}}$, and the type of solver employed to derive the pure-component profiles. These so-called hyperparameters are automatically tuned using a cross-validation scheme. The residuals between the *in silico* and the measured spectra were used as quality metrics and summarized by the root mean squared error (RMSE) of reconstruction. Hyperparameters associated with the LSA method were screened using a grid-based scheme and the optimal configuration with regard to the cross-validated reconstruction error RMSECV was selected. Depending on the data set, the determined local subset size served as an initial estimate and was further refined within the optimization procedure described in Sec. 4.2.3.3.

4.2.3 Convolutional neural networks

CNNs are multivariate regression models, which may be composed of several convolutional, pooling, fully connected (FC), dropout, and regularization layers. For a comprehensive, theoretical overview of CNNs the reader is referred to pertinent literature (I. J. Goodfellow, Bengio, and Courville 2016; Rosebrock 2018).

4.2.3.1 Neural network architecture

The following design choices in neural network architecture were made based on existing studies (Cui and Fearn 2018; Rolinger, Rüdte, and Hubbuch 2021; D. Passos and Mishra 2022), while aiming to keep model complexity low in order to reduce computational time during HPO. It is worth noting that the UV/Vis data mostly used in this study usually is of lower dimension than Raman or IR spectroscopy data used in other studies and hence the required model complexity may be lower. CNNs were constructed from 1-3 convolutional layers. For each convolutional layer, a number of 1-10 convolutional filters with a customized filter width were defined. After the first and second convolutional layers, a maximum pooling layer was implemented using a window size of 2, effectively halving the number of features generated from the previous layer. As the convolutional filter width is constrained by the output dimension of the previous layer, the maximum allowed filter width is adjusted accordingly after each pooling step. After the convolutional block, a flattening operation was implemented to concatenate the outputs from all convolutional filters of the last convolutional layer into a one-dimensional vector. In the regression block, an FC layer with up to 100 units was used. The output layer is configured to use a rectified linear unit (ReLU) activation function to restrict the prediction to positive values. As activation functions for the convolutional and FC layers, linear and hyperbolic tangent (tanh) functions were used, respectively. Other options may

CHAPTER 4. DATA AUGMENTATION AND CNNs

individually be chosen and several options were tested within this study. Depending on the chosen architecture, the number of configurable hyperparameters may greatly vary and hence a standardized workflow for optimization is required. The base architectures and corresponding hyperparameters for training and data augmentation used for data sets 1 and 2 are listed in Tab. 4.1.

Table 4.1 Overview of hyperparameters used for data augmentation, CNN architecture and training for data sets 1 and 2. The base configuration refers to hyperparameter values used during the single-factor parameter study. Data augmentation parameters were derived from tuning of the LSA method and a simple CNN architecture was assumed. The initial configuration refers to settings adapted after the parameter study and served as a comparison for the optimized models. The optimized configuration refers to values derived from HPO. Here, the square brackets indicate the optimized parameter values. Multiple numbers are given for the determined values for the individual layers. The parenthesis denote the search spaces during HPO. If no search space is given, the parameter was not included in the optimization. For the convolutional filter width, the allowed maximum filter width was configured to halve with each additional layer due to the interposed pooling layers.

Category	Hyperparameter	Base	Initial	Data set 1	Initial	Data set 2
				Optimized		Optimized
Data augmentation	number of gen. samples	1e5	1e5	1e5	1e5	1e5
	local subset size	5	11	11	13	12
	distance norm	2	2	2	1	1
	solver type	NNLS	NNLS	NNLS	NNLS	NNLS
	std. white noise	0.001	0.001	0.001	0.001	0.001
	std. wavelength shift	0.01	0.01	0.01	0.03	0.03
Model architecture	number of convolutional layers	1	1	2 (1-3)	1	3 (1-3)
	number of conv. filters	5	5	[2, 7] (1-10)	5	[2, 10, 8] (1-10)
	filter width	9	9	[3, 9] (3-61)	9	[7, 5, 11] (3-51)
	pooling width	2	2	2	2	2
	number of FC units	12	12	29 (5-100)	12	9 (5-100)
	activation function conv. layer	linear	linear	linear	linear	linear
	activation function FC units	sigmoid	tanh	tanh	tanh	tanh
	initialization function weights	glorot uniform	random uniform	random uniform	random uniform	random uniform
	dropout rate	0	0	0 (0-0.3)	0	0.07
	regularization factor	0	0	0 (1×10^{-9} - 1×10^{-3})	0	3.07×10^{-9} (1×10^{-9} - 1×10^{-3})
Model training	learning rate	1×10^{-3}	1×10^{-3}	1×10^{-3}	1×10^{-3}	1×10^{-3}
	batch size	100	100	100	100	100
	optimizer	Adam	Adam	Adam	Adam	Adam
	patience	4	4	4	4	4

4.2.3.2 Training, cross-validation and testing

In the context of training CNN models, data produced by the LSA method are denoted *in silico* data, while experimental data are split into training and test data as listed in Tab. B.1. For the remainder of the manuscript, we will further refer to *calibration* as the process of fitting the CNN model weights and *training* as an entire cycle of generating *in silico* data, fitting the weights, and evaluation based on the training data.

To train a CNN model, the *in silico* subset is generated solely based on the training data, and the CNN model is solely calibrated on the *in silico* data. For cross-validation, a similar data setup was used. Here, the experimental training data is rotated in a *leave-one-group-out* scheme, i.e. holding out one of the assigned training experiments. During each rotation, the experimental training data are split into rotation-specific training and test sets. Again, the *in silico* data are generated based on the rotation-specific training data and are further used to calibrate the CNN model. In both cases, training and cross-validation, the assigned training data serve to evaluate the stopping criteria. The CNN models were calibrated for a maximum of 100 epochs using the mean squared error (MSE) of all responses as the loss function and the stochastic gradient-based optimizer referred to as Adam (Kingma and Ba 2015). For all parameter studies and HPO, the number of generated samples was set to 1×10^5 , and model calibration was stopped when the loss of the assigned validation set did not improve for 4 consecutive epochs further referred to as *patience*. Finally, the CNN models were evaluated on the independent test set, which has not been used for *in silico* data generation or cross-validation.

To study the effect of hyperparameters associated with data generation, CNN architecture, or training on model performance, hyperparameters were varied in a one-factor-at-a-time scheme while all other parameters remained constant. Model performance was measured using the cross-validation error RMSECV across all response variables and the optimal settings were adapted as the base configuration for subsequent HPO. This parameter study was solely conducted with data set 1. The findings for the initial configuration were then also used for HPO for data set 2.

4.2.3.3 Hyperparameter optimization

HPO routines were implemented in *optuna* for data sets 1 and 2. For both data sets, a combination of a randomized and a tree-structured parzen estimator (TPE)-based sampler was used. A random sampling of hyperparameters in pre-configured ranges was performed for the first 100 trials when the optimizer switched to the TPE. The TPE can be used to optimize continuous, discrete, and categorical variables at the same time using a Bayesian approach based on kernel-density estimations. For the technical details and the theory of the method, the reader is referred to Bergstra, Yamins, and Cox (2012) and Akiba et al. (2019). The sum of the component-specific cross-validated coefficients of determination $\sum_{i=1}^P R_{CV,i}^2$ was used as the objective value. Automated pruning of unpromising hyperparameter combinations was configured to set in after 100 trials and was triggered when $\sum_{i=1}^P R_{train,i}^2$ was lower than the median of all previously reported trials. This effectively reduces the computation time as cross-validation does not need to be performed for the pruned trials. The search spaces for the optimizer were determined based on previously conducted individual parameter studies and are listed in Tab. 4.1. A MySQLTM database was used to facilitate distributed computation to accelerate HPO. In total, the optimization was run over 500 trials. However, the effective number of finished trials differs due to automated pruning.

Among the top 5 models, the best candidate was selected based on quantitative metrics such as training and cross-validation performance as well as qualitative metrics such as model complexity.

The selected model was retrained for 10 repetitions using a modified patience of 10. The obtained performance metrics were compared with optimized PLS models as measured by the normalized error $\text{NRMSE} = \text{RMSE}/\bar{y}_i$ with $\bar{y}_i$ being the arithmetic mean of the observed concentration for the respective data subset. As a second baseline comparison, the optimized CNN models were trained without any prior data augmentation. Therefore, the models were trained for 300 epochs and the early stopping criteria were disabled. The training data were divided randomly into 80/20% calibration-validation subsets to determine after which epoch the best performance was achieved.

4.2.4 Partial least squares regression

PLS models were implemented in *scikit-learn* using the non-linear iterative partial least squares (NIPALS) algorithm. While the CNNs were used as multi-response models i.e. predicting all target species using the same model, single-response PLS models were used for each component. Spectral data were preprocessed using a Savitzky-Golay filter (SGF) and mean-centered. HPO for PLS models was performed using a grid-based scheme. Therefore, the number of PLS components (1-10), the order of derivative (0-2) and the width of the smoothing window of the SGF (3-31) were varied in pre-configured ranges as stated in parenthesis. The SGF was used with a second-degree polynomial. The optimal configuration was chosen using the cross-validated and scaled sum of squared errors $\text{SSECV}_{\text{scaled}}$ according to S. Wold, M. Sjöström, and Eriksson (2001) as given by

$$\text{SSECV}_{\text{scaled}} = \frac{\sum_{i=1}^M (\hat{y}_i - y_i)^2}{M - n_{\text{PLS}} - 1}, \quad (4.4)$$

where $\hat{y}_i$ and y_i denote the predicted and observed response values for a sample i , respectively, and n_{PLS} designates the number of PLS components.

4.2.5 Feature importance

To quantitatively evaluate the importance of individual wavelengths, GradCAM and SHAP were employed. While GradCAM can solely be applied to the CNNs, SHAP is model-agnostic and can therefore be used to directly compare CNN and PLS models. For the PLS models, the regression coefficients and variable importance in projection (VIP) scores were used as evaluation metrics. Feature importance techniques were solely employed using the optimized models from HPO for data set 1 and 2.

4.2.5.1 Gradient-weighted class activation maps

Guided GradCAM is a response-discriminative localization technique which was proposed in Selvaraju et al. (2020) and was implemented according to Rosebrock (2018). GradCAM can be largely divided into three steps: 1) Computation of backward gradients with respect to each response variable and the last convolutional layer for one specific input spectrum, 2) global average pooling of the computed gradients along the wavelength dimension to obtain a single weighting value for each filter in the last convolutional layer, and 3) computation of the GradCAM estimate. In cases, where pooling layers are used in between convolutional layers, the localization estimate is of reduced dimension compared to the original input spectrum and is thus linearly interpolated to match the original dimension.

4.2.5.2 Shapley additive explanations

SHAP is a model-agnostic additive feature attribution method derived from economic game theory and can be used to quantify feature importance which are in turn referred to as SHAP values. To compute a SHAP value for a specific wavelength, the absorbance values in a given data set are

randomly permuted and replaced by absorbance values sampled from a conditional distribution. Every sample in the given data set is permuted d times and passed through the regression model to obtain the model prediction for the permuted input spectrum. The permuted model prediction $w(S)$ is compared to the prediction using the original data $w(S \cup \{i\})$. The SHAP value $\phi_i(w)$ for a feature i and a permutation cycle w with d permutations $(w_1, \dots, w_d)$ is according to S. M. Lundberg, Allen, and Lee (2017) then defined as

$$\phi_i(w) = \frac{1}{d} \sum_{S \subseteq D \setminus \{i\}} \binom{d-1}{|S|}^{-1} (w(S \cup \{i\}) - w(S)) \quad (4.5)$$

where D and S denote the entire feature set and the permuted feature subset, respectively. To incorporate inter-dependency due to collinearity between multiple wavelengths, conditional sampling is performed for multiple wavelengths at the same time and repeated for a fixed number of permutations. For a deeper overview of the theory and discrepancies to closely related methods, we refer to (S. M. Lundberg, Allen, and Lee 2017; Covert, S. Lundberg, and Lee 2020a; Covert, S. Lundberg, and Lee 2020b; Belle and Papantonis 2021). To compute SHAP values within this study, the test subsets for data sets 1 and 2 were used to obtain permuted input spectra. All computations were done as implemented in *shap* using the *PermutationExplainer*. In total, 1×10^5 permutations per input spectrum were used.

4.2.5.3 Variable importance in projection

The VIP scores are a common metric to assess variable importance in PLS models next to the regression coefficients. According to (Mehmood et al. 2012), the VIP score v_j for a wavelength $j \in [1, N]$ is defined as

$$v_j = \sqrt{N \sum_{a=1}^A \left[(q_a^2 t_a^T t_a) (w_{aj} / \|w_a\|)^2 \right] / \sum_{a=1}^A (q_a^2 t_a^T t_a)}, \quad (4.6)$$

where w_a , q_a , and t_a denote the loading weights, the y-loadings and the scores vector corresponding to the PLS component $a \in [1, A]$, respectively. The total number of wavelengths is given by N .

4.2.6 Robustness and transferability

The robustness and transferability of the LSA method and the optimized CNN were evaluated by an *in silico* noise perturbation study and a model transfer to two external data sets based on UV/Vis and IR spectroscopy.

4.2.6.1 *In silico* noise perturbation

To compensate for increasingly noisy data, both model types were evaluated with modified generated data sets with an increasing level of white noise $\mathcal{N}(0, \sigma_{\text{noise}})$ with σ_{noise} of $\{0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1\}$ mAU and an increasing level of axial wavelength shifts $\mathcal{N}(0, \sigma_{\text{shift}})$ with σ_{shift} of $\{0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$ nm. We differentiate between models solely being trained on the noise-free data set and models being retrained for each noise level. The latter models are referred to as 'retrained'.

4.2.6.2 Model transfer to data set 3

To evaluate model performance on a data set without additional HPO, the optimized model from data set 2 was transferred to data set 3. The CNN input and output dimensions were adjusted

according to the experimental data. The CNN models were trained in the optimized configuration for 10 repetitions and evaluated against the PLS models which were optimized as described in Sec. 4.2.4.

4.2.6.3 Model transfer to data set 4

To compare the LSA method to the augmentation method presented in Blazhko et al. (2021), which will be referred to as extended multiplicative scatter augmentation (EMSA), both methods were applied to data sets 2 and 4. The CNN architecture found for data set 2 was therefore transferred to data set 4 without additional HPO. To account for the higher dimension of the IR data, the pooling window size of the first CNN layer was adjusted to 4 and the number of FC units was raised to 25. The EMSA method was obtained from Blazhko et al. (2021) and used with its default configurations. The LSA method was tuned as described in Sec. 4.2.2. IR spectra were preprocessed using a second derivative SGF with a window size of 19 and second-order polynomial, as this was reported to improve the performance of the EMSA method (Blazhko et al. 2021).

4.3 Results and discussion

4.3.1 Generating highly realistic *in silico* spectra from experimental data

The LSA method was used to generate *in silico* UV/Vis absorbance spectra based on the assigned training data as previously described in Sec. 4.2.2. To systematically evaluate the suitability of the proposed data augmentation method and tune the corresponding hyperparameters, the LSA method was used to reconstruct the experimental data in a cross-validation scheme. The reconstruction accuracy for the cross-validation and test subsets for both data sets as measured by the RMSECV are displayed in Fig. 4.3 with regard to the local subset size and the standard deviation of the applied wavelength shift. The local subset size strongly affects the reconstruction RMSECV and shows optima at sizes of 5 and 13 samples for data sets 1 and 2, respectively. In both cases, the RMSECV remains stable for small wavelength shifts and grows exponentially starting at 0.1 nm. The Cityblock norm is observed to slightly improve the RMSECV for data set 2 compared to the Euclidean norm while the overall impact on RMSECV is considerably small in comparison to the local subset size. While the Cityblock norm uses absolute differences, the Euclidean norm is based on squared differences and hence can affect the selection of local subsets depending on the concentration ranges in the samples. The residuals for the spectra in the test sets are displayed in Fig. 4.3 and show maximal deviations of 5 and 15 % for data sets 1 and 2, respectively, as measured by the maximum deviation normalized by the maximum absorbance in the corresponding run. As LSA is based on local subsets of the experimental data, the pure component profiles differ depending on the selected data points. Fig. 4.4 shows the local pure component profiles for all components from data sets 1 and 2 for concentration samples stemming from the test data. The color of the lines indicates the concentration of the respective component in the corresponding sample. The dashed lines indicate the global pure component profiles using the entire training data for estimation instead of the local subsets. For all components, the local profiles are scattered around the global profiles with larger deviations for samples in extraordinarily high or low concentration regimes. Particularly for data set 2, these effects are visible for the aggregate component, where the local pure component profiles strongly deviate from the global estimates for low concentration regimes by a factor >10 . Contrarily, for the monomer species low concentration regimes cause the spectra to capture an increased level of noise in the data.

In summary, LSA provides a concentration-adaptive data augmentation method by leveraging

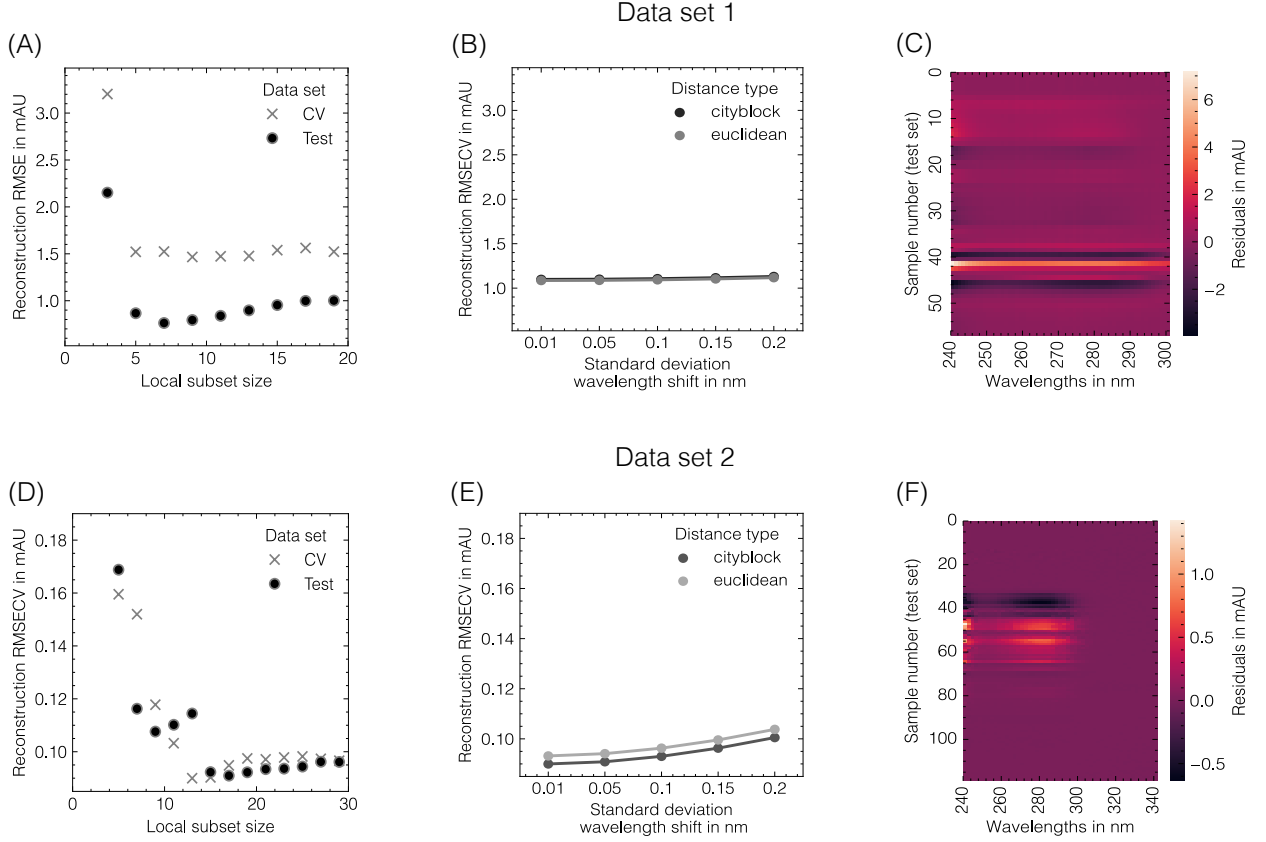


Figure 4.3 Tuning of the LSA data augmentation method for data set 1 (A-C) and data set 2 (D-F). From left to right, the reconstruction RMSE in dependence of the local subset size, the reconstruction RMSE in dependence of the standard deviation of the wavelength shift and a heatmap of wavelength-specific residuals for the respective test subsets are shown. The errors in (A),(B),(D) and (E) are shown for fixed levels of white noise $\sigma_{\text{noise}}=0.001$.

variations in the spectral and concentration domain. The tuned LSA method can generate highly realistic *in silico* spectra and can hence be used to augment experimental data sets.

4.3.2 Training convolutional networks with augmented spectral data

To study the effect of hyperparameters associated with data augmentation, CNN training, and model architecture on predictive performance, hyperparameters were varied in a one-factor-at-a-time scheme while all other parameters remained constant. The base configuration of the CNN for this parameter study can be found in Tab. 4.1. The obtained performance as measured by the predictive RMSECV for all response variables for data set 1 are displayed in Fig. 4.5.

Considering the data augmentation hyperparameters in Fig. 4.5 (A-C), the kernel-density estimation (KDE) sampling surpasses the performances obtained by uniform and normal sampling. The local subset size shows a stable performance for Cyt C and Lys between 5 and 15 samples, while the RMSECV for Rib A is reduced by approximately 25 % by increasing n_{LSA} from 5 to 11, reaching a more balanced performance between all three components. This suggests that it may be beneficial to include the local subset size during HPO to ensure finding the optimal solution for all response

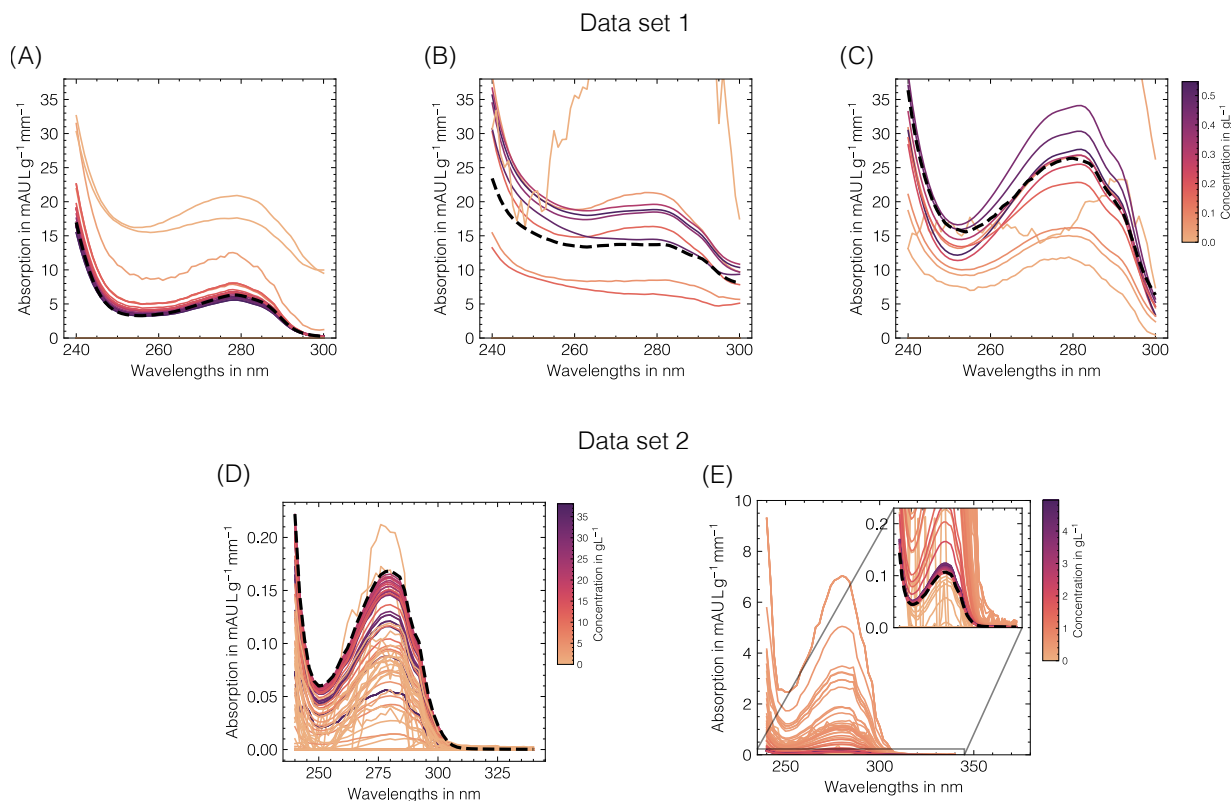


Figure 4.4 Local estimations of the pure component profiles for Rib A (A), Cyt C (B), Lys (C) for data set 1 and monomer (D) and aggregate (E) species for data set 2. The pure component spectra are shown for each sample in the test data (solid lines) with a local subset size of 11 for the data augmentation method and are colored according to the concentration of the corresponding component with darker colors denote high concentrations. The pure component spectra estimates using the entire training data are shown as dashed black lines.

variables. Hence, the local subset size was subsequently incorporated during HPO for both data sets 1 and 2. The number of *in silico* generated samples positively affects model performance and starts to plateau at 1×10^5 . As increasing the number beyond that only minorly affects model performance while increasing computational time considerably, 1×10^5 was adopted for the base configuration for HPO.

The influence of initialization and activation functions for convolutional and FC layers are shown in Fig. 4.5 (D-F). Initialization only minorly affects model performance with random uniform providing the best option. While the linear function performs best for the convolutional layer, non-linear activation functions show superior accuracy for the FC layers with tanh and the exponential linear unit (ELU) returning the lowest RMSECV values for all components. Increasing the complexity of the CNN by increasing the number of convolutional layers, the number of FC units, or the size of the convolutional window, does not directly improve model performance. Performance gains are not equally distributed among all response variables and no overall trends can be extracted as depicted in Fig. 4.5 (G-I). Regarding the training parameters, it is suggested that a learning rate of 1×10^{-3} ,

CHAPTER 4. DATA AUGMENTATION AND CNNs

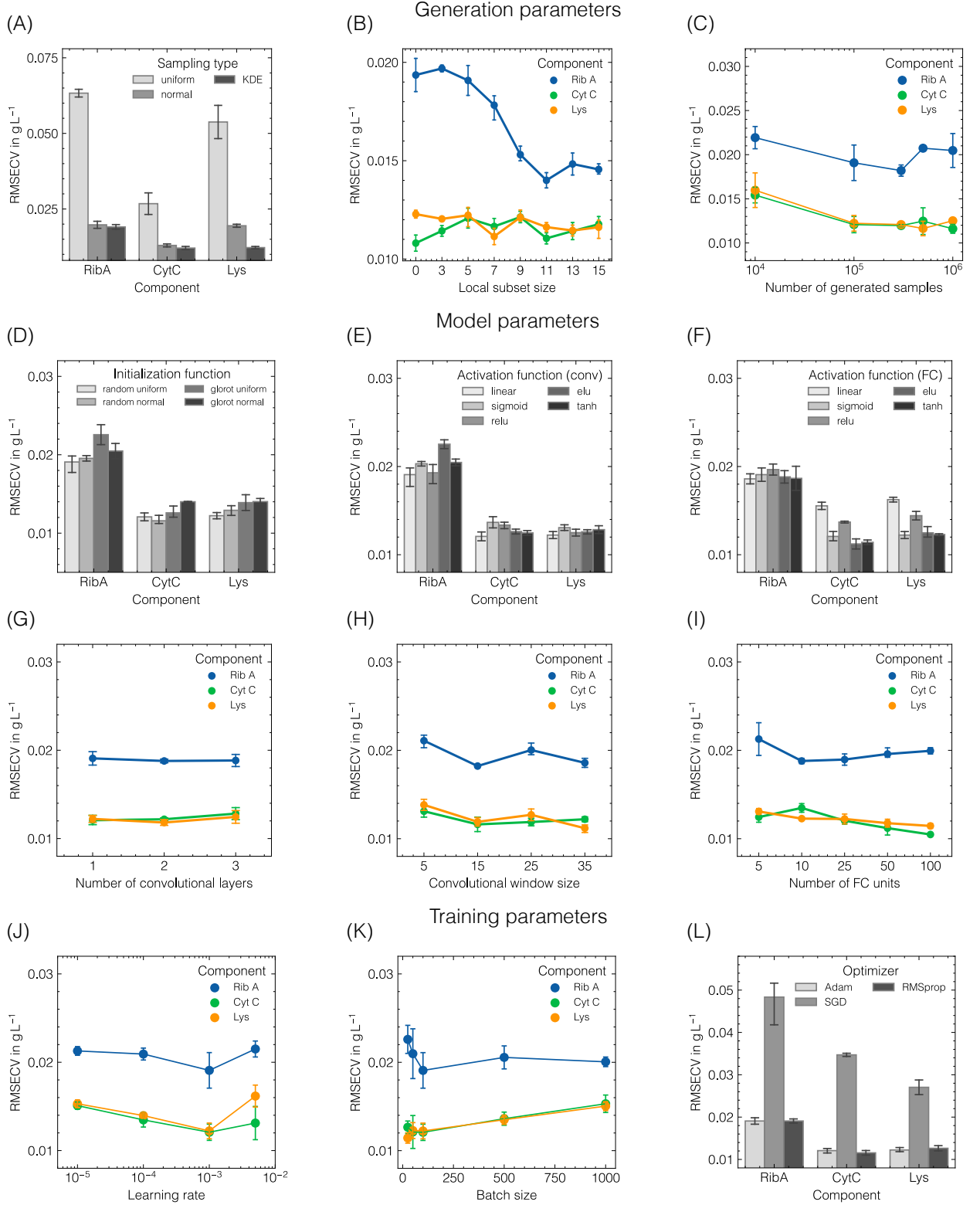


Figure 4.5 Influence of selected hyperparameters on the predictive cross-validation performance (RMSECV) for each component in data set 1. The selected hyperparameters can be categorized in three groups: generation parameters (A-C), model parameters (D-I) and training parameters (J-L).

a batch size of 100, and the Adam optimizer provide the best options as shown in Fig. 4.5 (J-L). In summary, it was possible to identify optimal configurations for hyperparameters for data augmentation and model training. However, given the high dimensionality of the search space of the remaining hyperparameters and unequally distributed effects on model performance it is suggested to use automated HPO to identify the optimal model architecture for each data set individually.

4.3.3 Automating hyperparameter search by Bayesian optimization

As hyperparameter search is a multi-dimensional, computationally expensive problem, automated HPO was performed using a TPE-based Bayesian optimizer as implemented in *optuna*. The optimizer considers local and global hyperparameters which enables the solution of optimization problems with multiple decision levels such as the choice of the number of convolutional layers and the optimization of a set of layer-specific hyperparameters. Fig. 4.6 presents the evolution of objective values with regard to all studied hyperparameters throughout the optimization process for data set 2 exemplarily. As shown in Fig. 4.6 (A), the transition between random and TPE-based sampling can be captured after 100 trials. The objective values form a band centered around 1.9 with scattered maxima at 1.93, with 2 being the maximum achievable objective value. The color of the points indicates the number of the optimization trials with dark blue being the end of the optimization process. The evolution profiles of global parameters (cf. Fig. 4.6 (B-F)) suggest an optimum at 3 convolutional layers with less than 25 FC units. Both regularization methods, governed by the regularization factor and the dropout rate, were found to positively influence model performance although weight regularization using a fairly low regularization factor in the range of 1×10^{-8} - 1×10^{-6} was employed in later trials. In contrast to data set 1, where the local subset size needed to be adjusted for optimal model performance, here, the previously determined value was found to be suitable for the given data set and only varied slightly between 9 and 15 in later trials. According to Fig. 4.6 (G-L), the three-layer CNN achieves the highest accuracy for all components with a convolutional window size between 1 and 10 for all three layers. The number of convolutional filters shows no clear optimum for the first and second layers, while higher counts are found beneficial for the third layer.

For data set 1, optimal performance with an objective value close to 2.98 was achieved by employing a two-layer CNN with 2 and 7 convolutional filters, a window size of 3 and 9, and 29 FC units. Contrarily, dropout and regularization were both found to be disadvantageous. The evolution plot can be found in Fig. B.1. The exact hyperparameters and visualization of the architecture for both optimized models can be found in Tab. 4.1 and Fig. B.2, respectively.

The optimized CNN models were retrained for 10 repetitions using random initialization of the weights and an increased patience of 10 epochs. The obtained model predictions for training and test subsets for both data sets are summarized by the NRMSE and presented in Fig. 4.7. The boxplots show the distribution of errors for the 10 repetitions as a result of random initializations and the stochastic nature of the training process. CNN models using the initial configuration obtained from tuning the augmentation method are also included for reference. As a baseline comparison, the NRMSE obtained from optimized single-response PLS models are indicated by the dashed lines. Optimized PLS hyperparameters can be found in Tab. B.2. Additionally, the optimized architectures were trained without using any prior data augmentation as described in Sec. 4.2.3.3. For reference, timely predictions for the optimized CNN and the PLS can be found in Fig. B.3.

For data set 1 (cf. Fig. 4.7 (A)), initial and optimized CNN models show generally lower NRMSE on average than the PLS models for the training subset for all components. In the test subset, the prediction error for Rib A is reduced by up to 50%, while the test errors for Cyt C and Lys increase. This increase can be attributed to erroneous predictions of Cyt C and Lys during the elution of Rib

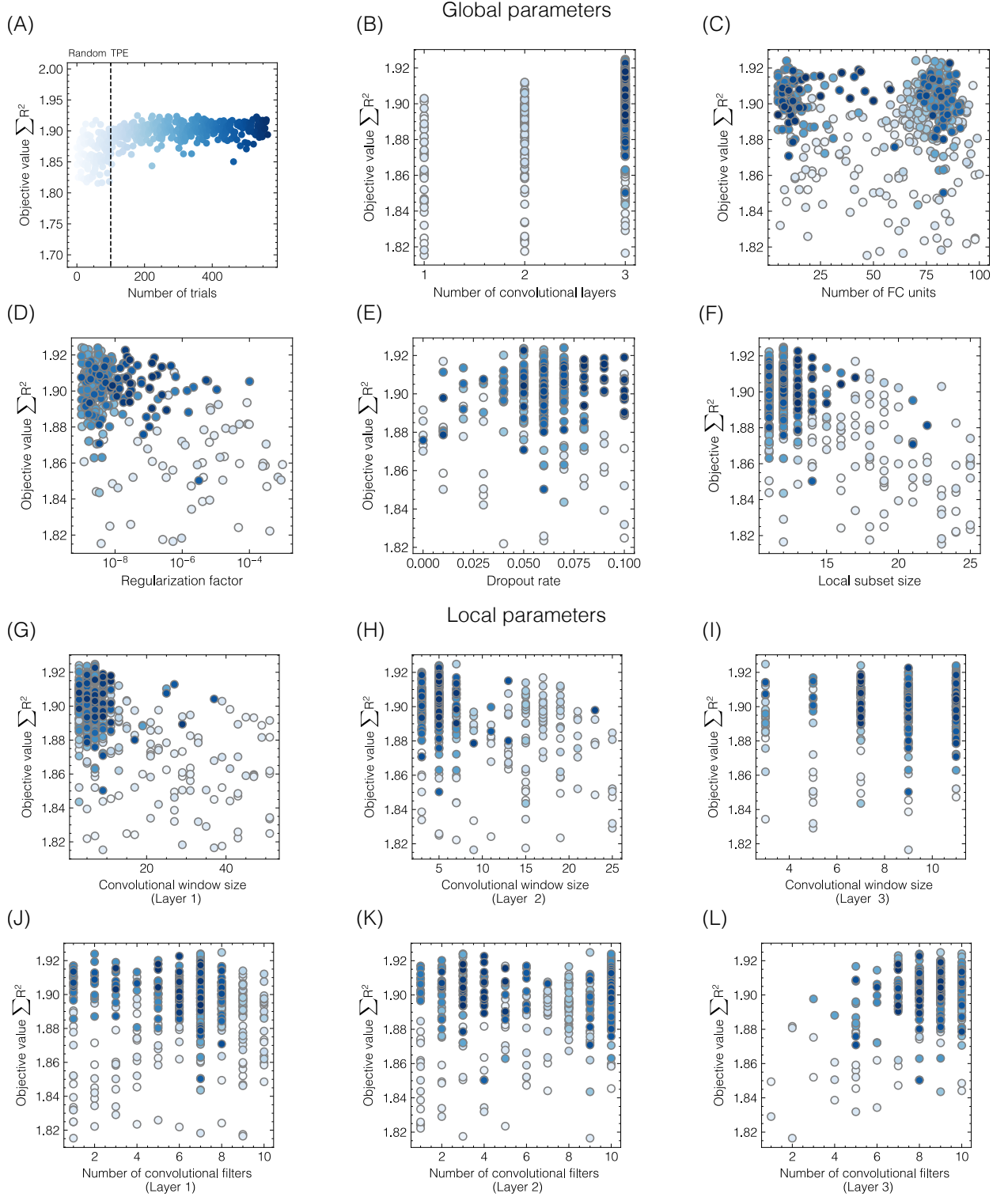


Figure 4.6 Hyperparameter evolution profiles during optimization for data set 2. The objective values $\sum R^2$ are shown for all trials (A) and each hyperparameter individually (B – L). The colors of the circles indicate the number of HPO trials with darker shades of blue corresponding to later sampling points. Random sampling was performed for the first 100 trials after which median pruning and the TPE-based optimization were enabled.

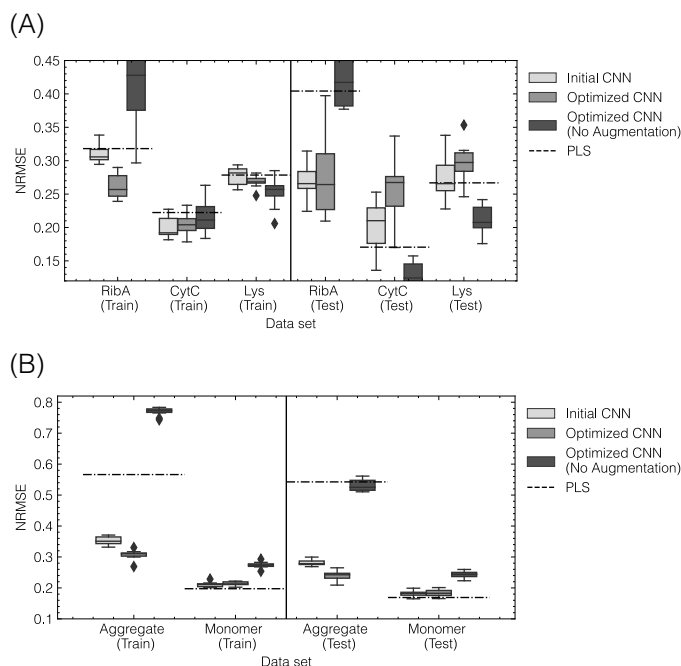


Figure 4.7 Model performance for initial and optimized CNN models for data set 1 (A) and data set 2 (B). The normalized error NRMSE for training and test data are shown on the left and right, respectively. For both data sets, the initial and optimized CNN model were trained 10 times with an elevated patience of 10. The arithmetic mean of the obtained performances is indicated by the solid lines within the boxes. Outliers are shown as diamonds and were defined as such when the error exceeded 1.5 times the interquartile range. The dashed lines correspond to the NRMSE obtained from an optimized PLS model using a SGF for preprocessing as a benchmark.

As can be seen in the timely predictions in Fig. S3. The initial CNN architecture performs slightly better on the test subset than the optimized CNN architecture as indicated by a lower average and variance. Interestingly, CNN models without prior data augmentation perform similarly to the PLS models achieving higher accuracy for Cyt C and Lys and significantly lower accuracy for Rib A compared to the CNNs using data augmentation. For data set 2, the optimized CNN model reduces the NRMSE compared to the initial architecture and the PLS model by up to 50% for the aggregate species. The accuracy for the monomer species slightly decreases compared to the PLS model with a 5% increase in the NRMSE. The reduction for the aggregate species can be attributed to the improved capture of the onset of the elution peak as can be seen in the timely prediction profiles in Fig. S3. In contrast to data set 1, the optimized CNN without augmentation results in considerably higher NRMSE than both augmented CNN models and the PLS. In summary, HPO enabled the automated identification of the optimal model architecture for data sets 1 and 2, leading to improved quantification of Rib A and mAb aggregates, respectively, while sometimes reducing the accuracy for the other species compared to the benchmark methods.

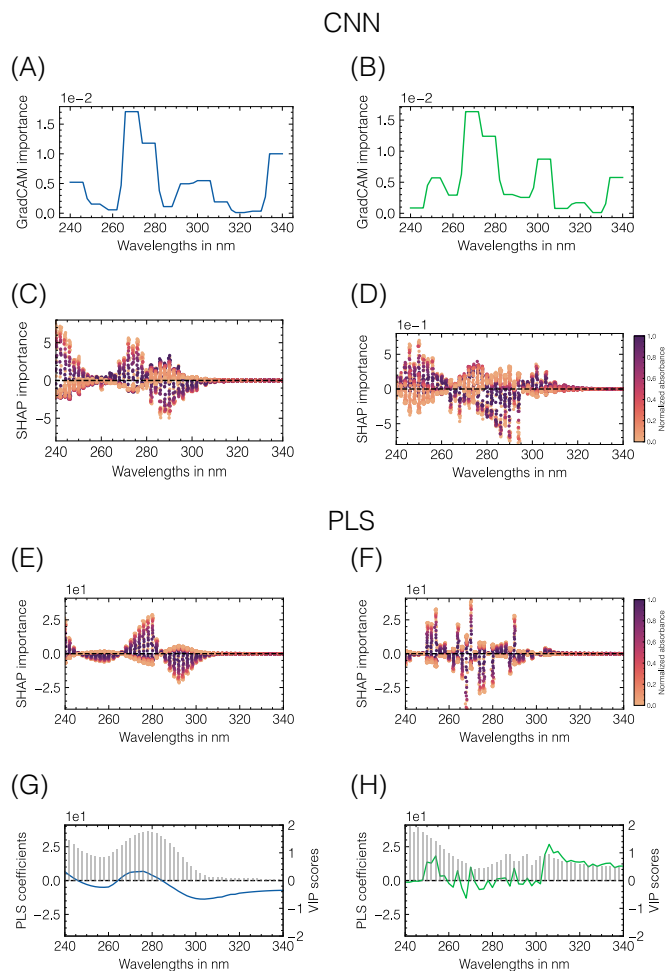


Figure 4.8 Model-specific and model-agnostic importance measures for data set 2. CNN-specific GradCAM importance values (A-B), SHAP values for CNN and PLS (C-D) and (E-F), respectively. Each dot in (C-F) corresponds to the feature contribution for one specific model output in the test data set, which are colored according to the measured absorbance. The absorbances were normalized by the maximum value at 280 nm. PLS-specific regression coefficients (shown as lines) and VIP scores (shown as bars) are presented in (G-H). The left and right columns correspond to importance metrics for the monomer and aggregate species, respectively.

4.3.4 Understanding model predictions through feature importance

To investigate the differences in model performance, model-specific and model-agnostic importance metrics were used. Fig. 4.8 presents an overview of multiple importance metrics for data set 2. CNN-specific GradCAM localization maps are shown in Fig. 4.8 (A-B). PLS regression coefficients and VIP scores are shown in Fig. 4.8 (G-H). To enable a direct comparison between the two model types, SHAP values are illustrated in Fig. 4.8 (C-F). The left and right columns correspond to the feature importances for the monomer and aggregate components, respectively. For the monomer species, both CNN- and PLS-SHAP values closely resemble the PLS regression coefficients with wavelengths between 260-280 nm positively contributing to the model output and wavelengths between 280-300 nm being assigned negative values. A similar behavior can be observed for the GradCAM estimates with both wavelength ranges being assigned positive importance as GradCAM is constrained to positive values per definition. Additionally, GradCAM identifies the border areas at the beginning and the end of the spectrum as important. The SHAP values generally confirm those observations for small wavelengths, whereas wavelengths above 315 nm are shown to not affect the model output for both the CNN and PLS models. This is in accordance with the PLS-VIP scores as well as the spectral data as no absorbance is detected at wavelengths above 315 nm (cf. Fig. B.4). For the aggregate species, GradCAM provides a fairly similar profile compared to the monomer species showing a shift in the importance peak from 280-305 to 300-305 nm. The PLS regression coefficients support this observation with a maximum of 315 nm for the aggregate species, while the VIP scores for all wavelengths between 280-340 nm are between 0.5-1. Regression coefficients for smaller wavelengths appear comparably noisy and without any structural integrity in contrast to the coefficients for the monomer species. While CNN-SHAP values indicate a similar profile as seen for the monomer with an inversed importance for wavelengths between 295-305 nm, PLS-SHAP suggests a high degree of noise and considerably lower importance for the previously identified wavelength region relative to the remainder of the spectrum. As a means for comparison, the visualization of the convolutional layers for data set 2 can be found in Fig. B.5. For each convolutional layer and each corresponding filter, the convoluted input spectra for the test set are shown. To emphasize the differences between the monomer and aggregate components, the samples with the maximum concentration of monomers and aggregates are marked in green and blue, respectively. While the convoluted signals in layer 1 largely resemble the original input spectra, more fluctuations between positive and negative outputs are introduced with layers 2 and 3, with some filters showing fairly similar profiles within one layer. Due to the interposed pooling layers, the resolution of the presented profiles is gradually decreasing across layers 1 to 3.

In summary, the employed feature importance methods point to an elevated importance for higher wavelengths for the aggregates compared to those for the monomer species and identify this area as critical for differentiation between both species. The filter visualization further provides insights into the working principle of the CNNs. Analogous to the presented comparison for data set 2, the feature importance evaluation and filter visualization for data set 1 can be found in Fig. B.6 and Fig. B.7, respectively.

4.3.5 Evaluating robustness and transferability

To evaluate the capability of the CNN and PLS models to compensate for increasingly noisy data, both model types were retrained using modified generated data sets with 1) an increasing level of Gaussian noise and 2) an increasing level of axial wavelength shifts. The model performance as summarized by the sum of NRMSE for all three components in data set 1 is shown in Fig. 4.9 (A-B).

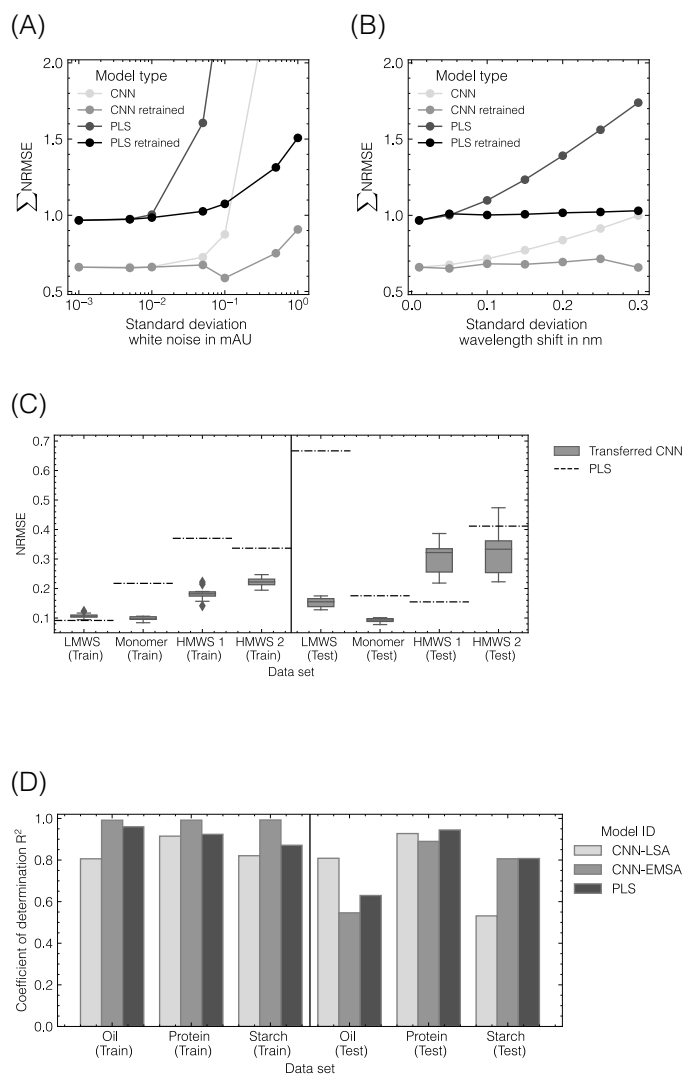


Figure 4.9 Evaluation of CNN robustness and transferability. Comparison of CNN and PLS model robustness for increasing noise (A) and wavelength shifts (B) in spectral data for data set 1 as measured by the normalized error NRMSE. CNN performance after model transfer to data set 3 (C). Comparison of model performance with data augmentation conducted by the EMSA and LSA method for data set 4 (D).

It is observed that the PLS model does generally not reach the same level of accuracy as the CNN model being trained on the generated data. For both cases, 1) and 2), the PLS model error starts to increase at lower levels of white noise and wavelength shift than the CNN model. Similarly, when both models are retrained at each noise level, the CNN model is observed to be more robust against noisy data and can slightly decrease the model error at a noise level of 0.01 mAU or a wavelength shift of 0.03 nm.

Using the hyperparameters determined during HPO for data set 2, the CNN was trained on data set 3 including data augmentation by the LSA method to assess the generalizability of the identified architecture. All hyperparameters were obtained from data set 2 and remained unchanged. Fig. 4.9 (C) presents a comparison of CNN and PLS models as measured by the NRMSE. The transferred CNN achieves NRMSE values lower than 0.25 for all components for the training subset and decreases the observed errors for the monomer, HMWS1 and HMWS2 species by up to 50 %. For the test subset, the variance of the transferred CNN increases and the error for HMWS1 increases compared to the PLS model. The PLS model fails to predict the LMWS component which was observed to depend on the cross-validation split used for optimizing the PLS models.

Finally, the herein-developed LSA data augmentation method was compared to the EMSA method (Blazhko et al. 2021). Both methods were evaluated based on data sets 2 and 4 to assess the applicability of EMSA on UV/Vis data as well as the transferability of the LSA method to IR data. Therefore, the optimized CNN model from data set 2 was trained with *in silico* data from the LSA and the EMSA method for both data sets 2 and 4. A detailed overview of the results for both data sets including the performance of PLS models can be found in Tab. B.3. Fig. 4.9 (D) provides a summary of the obtained performances using the LSA method versus the EMSA method for data set 4. The LSA method in combination with a three-layer CNN could successfully be transferred to data set 4 using IR data. When compared with the EMSA method as employed in Blazhko et al. (2021), the LSA method achieves accuracies of +32 %, +4 %, and -34 % in terms of the relative difference in R^2 for oil, protein, and starch content in the test subset, respectively. However, the training errors for the CNN-EMSA model are considerably higher for all three components indicating potential overfitting of the model. When applied to data set 2, the EMSA method fails to generate high-quality data to be used for chemometric modeling resulting in a two-fold increase in RMSE for both components compared to the LSA method.

In summary, the optimized CNN models in combination with the LSA method present a more flexible alternative to commonly used PLS models due to their increased robustness against noisy data and good generalization capability to other UV/Vis spectroscopy data. They may further be universally applied to other types of spectroscopic data without prior HPO.

4.4 Discussion

4.4.1 Data augmentation

Data augmentation can be used to enlarge experimental data sets by extracting and recombining information from the collected data. This is considered to avoid overfitting and thus improve the generalizability of ML models (Shorten and Khoshgoftaar 2019). The herein presented LSA method builds upon this idea and leverages latent spectral information in terms of pure component profiles. By employing a local subset selection, these pure component profiles are locally approximated and flexibly adapted to the concentration regime of the sample composition for which a new spectrum is generated. The local subset selection further induces the mixing of spectral information from different experimental runs. We consider the designed working principle beneficial to leverage locally

available spectral information. While other methods solely take variations in the spectral domain into account, e.g. by modifying the spectra based on variations in model coefficients (Blazhko et al. 2021; Bjerrum, Glahder, and Skov 2017), the LSA method projects the information to a latent space in the spectral domain and also accounts for variations in the concentration domain. This idea is similar to algorithms employed in variational autoencoding or GANs (F. Guo, Xie, and B. Huang 2020; Mishra and Herrmann 2021), while maintaining interpretability by employing simple numerical methods. By using a cross-validation-based reconstruction approach, we showed that the LSA method can be tuned by various hyperparameters and is able to reconstruct the hold-out test set of the experimental data with a maximum error of roughly 5 % and 15 % for data sets 1 and 2, respectively. While pure component profiles for samples within the center of the experimental data range showed slight variations around the global estimates, samples with exceptionally low or high concentration regimes were shown to incorporate more noise or considerably diverge from the global estimates. The shape and magnitude of the profiles is dependent on the selected subsets. For low-concentrated samples, the selected spectra exhibit a low signal-to-noise ratio potentially causing distortions in the spectral shape or numerical inflation of the estimated profiles. Whereas for high-concentrated samples, the density of similar data points is lower resulting in local adaptations of the pure component profiles. In the case of data set 2, there are several samples almost solely containing monomer or aggregate species due to the working principle of the underlying separation process the data is taken from. When one component concentration is close to 0, the estimation algorithm will numerically inflate the component’s profile and hence a higher absorption is observed. The described handling of different concentration regimes is considered to introduce more variance in the *in silico* data. Although the estimates obtained for the extreme data points are not realistic, their impact is considered to be low. It was further observed that the local subset size determined by the tuning method did not provide the optimal value in the case of data set 1. This may be explained by the low absorbance signal by Rib A causing a reduced influence on the reconstruction error compared to Cyt C and Lys (Hansen, Skibsted, et al. 2011; Rüdte, Vormittag, et al. 2019). When comparing the LSA and the EMSA method for training CNNs, it was observed that CNN-EMSA models tend to overfit the training data, potentially caused by solely depending on variations in the spectral domain. Whereas the CNN-LSA model resulted in consistent training and test performances. When applied to UV/Vis data, the CNN-EMSA models were not able to achieve the same level of accuracy as the CNN-LSA models. As the scatter correction method, on which EMSA is based, is routinely used with vibrational spectroscopy data (Martens and Stark 1991b; Afseth and Achim Kohler 2012), the lower accuracy may be explained by insufficient accuracy of the estimated model or an overestimation of the variations in the UV/Vis spectra. However, it should be noted that the EMSA method was used with its default configuration (Blazhko et al. 2021) and may potentially be tuned to be applicable to UV/Vis data.

Many studies published in the literature do not use any type of data augmentation before chemometric modeling with CNNs (Acquarelli et al. 2017; Malek, Melgani, and Bazi 2018; Cui and Fearn 2018; C. Zhang et al. 2019). This can either be realized by using comparably large data sets in combination with simple model architectures e.g. in Cui and Fearn (2018) or by longer training on the same data (C. Zhang et al. 2019). For the herein evaluated data sets, training CNNs without prior data augmentation resulted in unbalanced model performance with regard to all components which was comparable to the performance of the PLS model. On the one hand this suggests that non-linear models such as CNNs do not necessarily improve the accuracy compared to PLS models. In theory, the CNNs without augmentation have little leverage over PLS models as they are potentially overparameterized given the number of available training data. On the other hand

this suggests that data augmentation effectively helps to extract more information from the data to enable more accurate multi-response modeling. This may especially be beneficial in cases where one of the target components contributes a comparably weak spectral signal or when all target components exhibit highly similar spectral profiles. While data set 1 comprises proteins with clearly differentiable UV/Vis absorbance spectra (Hansen, Skibsted, et al. 2011), the size variants of a mAb are more challenging to distinguish (Brestrich, Rüdte, et al. 2018). Based on the given results, we hypothesize that by augmenting spectral data, minor differences in spectral profiles or signals from underrepresented components can be amplified and therefore increase regression performance. Other spectroscopic methods such as IR or Raman can provide higher selectivity depending on the analytes and hence the benefits of data augmentation should be studied in more detail for those spectroscopic methods. Blazhko et al. (2021) pointed out that for the classification of yeast and mould species using IR data, the usage of data augmentation positively affected the classification performance.

In conclusion, the combination of CNNs and data augmentation via the LSA method provides a flexible, generally applicable approach for chemometric modeling of multiple quality attributes. Given the findings mentioned above, it would further be interesting to study the effect of spectral mixing introduced by the LSA method with data sets considerably larger than the ones used within this study, possibly including data from multiple spectrometers, target proteins, or cell lines.

4.4.2 Model architecture and hyperparameter optimization

Finding the optimal configuration for complex ML models such as CNNs is challenging due to the high-dimensional search space and potentially long computation times (Feurer and Hutter 2019). In this study, automated HPO based on Bayesian optimization has been used to optimize CNN configurations for data sets 1 and 2. Here, the aim was to maximize the predictive accuracy of the CNN models, while keeping model complexity low. To weight all response variables, i.e. molecular species, equally, the optimizer used the sum of R^2 of the cross-validation residuals. While equal weighting of all components may also be realized by other evaluation metrics, the sum of R^2 was chosen to provide a simple figurative metric. During the optimization, the structure of the neural network was varied simultaneously with other hyperparameters such as the local subset size n_{LSA} as well as regularization parameters, namely the regularization factor and the dropout rate. In general, increasing the number of convolutional layers implicitly increases the number of hyperparameters to optimize as each layer is assigned an individual set of parameters. At the same time, the intermediate number of features is reduced by the pooling layers thus potentially reducing the number of units in the FC layer.

For data set 1, the identified architecture using two convolutional layers was found to achieve only marginally improved performance compared to the initial configuration. For the second data set, the optimized configuration could improve the initial configuration considerably with regard to the contaminating aggregate species. The ineffectiveness observed for data set 1 may be caused by a suboptimal splitting of the experimental runs or non-exhaustive exploration of the search space as well as potential overfitting on the training subset. It should however be noted that the achieved accuracy lies above 0.97 in terms of the R^2 for all components and the differences between the studied models are considered marginal. In the context of chemometrics, TPE-based Bayesian optimization has previously been applied to the HPO of CNNs (D. Passos and Mishra 2021; D. Passos and Mishra 2022). While previous studies have used pruning and optimization based on the simple training-validation splits, the herein-used cross-validation approach is considered more robust and less prone to overfitting (Rolinger, Rüdte, and Hubbuch 2020a). A collection of more

detailed practical remarks for using HPO can be found in Ch. B.

Furthermore, it is unclear from previous studies which hyperparameters should be included during HPO. While D. Passos and Mishra (2021) vary the depth of the regression block by incorporating more FC layers, an increased number of convolutional layers was found beneficial within this study. Additionally, it was found that critical hyperparameters requiring optimization are the number of filters in each convolutional layer, the convolutional window size per layer, and the number of FC units in the regression block. Increasing the number of filters increases the model’s capacity to learn complex features from the input data. However, a larger number of filters potentially requires a higher number of FC units and thus can cause slower training time (Szegedy et al. 2015). The convolutional window size determines the size of the receptive field, which affects the model’s ability to capture relevant features from the input data (I. J. Goodfellow, Bengio, and Courville 2016, chap. 9). Smaller sizes capture local features, while larger sizes capture more global features (Gu et al. 2018). Here, the optimal window sizes were found to be as small as 3 for the first convolutional, with a general increasing trend for additional layers. With regard to CNN design choices, multiple suggestions have been made in literature including simple architectures with only one convolutional layer (Cui and Fearn 2018; Acquarelli et al. 2017), or more complex models using multiple FC layers and a comparably high number of convolutional kernels (Bjerrum, Glahder, and Skov 2017; D. Passos and Mishra 2021; D. Passos and Mishra 2022). Whereas, others adapted design choices made for widely applied computer vision architectures such as the so-called vgg-block design (Blazhko et al. 2021) or inception modules (C. Zhang et al. 2019) or even tried to combine the convolutional block with other types of regression models such as Gaussian process regression (Malek, Melgani, and Bazi 2018). Given the variety of studied architectures, it is difficult to draw a conclusion on which architecture works best for which type of spectroscopic data as they are often combined with prior preprocessing. Transferring the model architecture identified for data set 2 to data sets 3 and 4 in this study, was found suitable for modeling UV/Vis and IR spectroscopic data. The CNN with three convolutional and one FC layer, using weight and dropout regularization, was shown to work well with previously unknown data sets. Especially for data sets with few available experiments or samples, the transfer of CNN architectures is preferred over potentially unstable HPO. It has been shown that multiple factors may affect the successful transfer of chemometric CNNs (Mishra and D. Passos 2021c; Mishra and D. Passos 2021b), leading to a high degree of freedom in HPO.

While HPO has shown great potential for improving the performance of chemometric models in this study and several other cases (D. Passos and Mishra 2021; Brunel, Alsamad, and Piot 2021), further studies are needed to investigate which CNN architectures are the most suitable for different types of spectroscopic data such as UV/Vis, IR, and Raman spectroscopy. In this regard, the authors advocate for a large-scale HPO study using multiple data sets at the same time. Such research could help identify the most viable CNN architectures for chemometrics in terms of accuracy, generalizability, and robustness, and pave the way to dedicated pre-trained models for biotechnological applications. Pre-trained models facilitate easier transfer from case to case by computationally less expensive transfer learning (Mishra and D. Passos 2021c).

4.4.3 Model interpretability

Explainability of machine learning models has become an important topic of research in recent years (Belle and Papantonis 2021; Burkart and Huber 2021). In the context of spectroscopy, it is critical to understand which elements of the spectral data are rendered important and contribute to the model’s predictions. In this study, an interpretability assessment compared the predictions of CNNs with PLS models. Four feature importance metrics were used to compare the models:

Guided-GradCAM, SHAP values, PLS regression coefficients, and PLS VIP scores.

The results for data set 2 showed that three of four methods were able to identify a small wavelength region that was relevant for differentiating between the two target components. For data set 1, the distinction between the three target components was less pronounced as the importance profiles for Cyt C and Lys were found to be fairly similar for CNN and PLS models. For Rib A, it was possible to resolve clear differences between both model types which supports the improved performance of the CNN model. In general, the studied methods differ in their level of directness in assessing the importance of individual wavelengths. Established methods such as GradCAM or other visualization methods for CNNs (Zeiler and Fergus 2013) are generally applicable to all types of model architectures (Selvaraju et al. 2020) and enable a fast inspection of the chemometric model. In other studies (D. Passos and Mishra 2022; Mishra and D. Passos 2021a), it has been shown that GradCAM was able to identify the most suitable section of features from a preprocessing-based extension method. However, as the variable importance is computed from the feature maps of the final convolutional layer, it is directly influenced by previous normalization and pooling operations causing lower resolution profiles or potentially distorted peak locations (Blazhko et al. 2021). Similar to GradCAM, PLS VIP scores provide a positively constrained importance metric. In combination with the PLS regression coefficients, the VIP scores are a commonly applied metric to evaluate the chemical information used by the PLS model (Goldrick, Umprecht, et al. 2020; Wei et al. 2022) or conduct variable selection (Mehmood et al. 2012). While the regression coefficients display positive and negative attributions, the VIP scores help determine which variables mostly contribute to the prediction. Despite their practicality, the VIP scores are specific to PLS models and do not contain any information about the variable sensitivity.

Other chemometric studies exist where SHAP estimation was used to identify the most important variables for various chemometric models (Haghi, Pérez-Fernández, and Robertson 2021; Mahynski et al. 2022; Guindo et al. 2023). By using SHAP values in this study, it was possible to identify the most relevant features in a direct comparison with the PLS model. Using SHAP or other closely related feature removal-based importance quantification techniques is computationally more expensive than analyzing regression coefficients or GradCAM importance (Covert, S. Lundberg, and Lee 2020b). Especially for spectroscopic data, where multiple wavelengths are highly correlated, permutation-based feature importance techniques should be handled carefully by ensuring a sufficient number of permutations and permuting a coalition of wavelengths rather than single wavelengths (Štrumbelj and Kononenko 2014b). By using a single-wavelength approach based on infinitesimal perturbations in Cui and Fearn (2018), the studied CNNs were observed to have increased robustness compared to PLS models. A similar single-wavelength perturbation approach was used in Schiemer, Weggen, et al. (2023) to perform automated variable selection for a Gaussian process regression model for monitoring an antibody-drug conjugation reaction.

Next to the identification of important features, removal-based methods such as SHAP may also be used to determine the sensitivity of specific features (Covert, S. Lundberg, and Lee 2020b) which could potentially be applied to detect failures in chemometric models. In the future, it would be interesting to employ such methods in real-time during the optimization or maintenance of chemometric models. When new data are available from a manufacturing process, individual feature contributions may be beneficial to identify potential process or model failures and hence trigger model updating mechanisms (Nikzad-Langerodi and Andries 2021; Krause et al. 2021; FDA 2023). In summary, we consider the usage of multiple orthogonal variable importance methods crucial to build explainable and robust chemometric models, and to ensure that the most relevant structural and chemical information present in the spectroscopic data are used for prediction.

4.4.4 Robustness

Spectral data are often corrupted by noise or baseline and peak shifts (Bjerrum, Glahder, and Skov 2017; Blazhko et al. 2021). Routinely applied operations to counteract these effects are among others spectral smoothing, derivation, or corrections (Tulsyan et al. 2021; Anderson et al. 2020). Spectral data may contain different levels of noise or baseline effects due to new data being recorded with a different spectrometer (Mishra and D. Passos 2021b), biological variability (Tulsyan et al. 2021) or as part of a technology transfer to another site (Christler et al. 2021).

To evaluate the robustness of CNNs compared to PLS models, an *in silico* noise perturbation study was performed. In summary, the CNNs were found to be more robust against increasing noise and wavelength shifts than PLS models. CNNs have been shown to be less susceptible to noise in images and electrocardiograms depending on the architecture used (Venton et al. 2021; Rodríguez-Rodríguez et al. 2021). In general, the increased robustness may be attributed to various features inherent to CNNs, such as convolution operations and in-built regularization. Each input is transformed by a convolution operation followed by a non-linear activation and a max-pooling operation. By stacking multiple convolutional blocks in sequence, a funnel-like structure is created which is considered to increase robustness. Secondly, regularization methods like dropout and weight regularization can lead to increased robustness (Krizhevsky, Sutskever, and Hinton 2017). In chemometrics, Cui and Fearn (2018) have shown that using dropout reduced noise in the trained weights of their CNNs, while Mishra and D. Passos (2021b) could show that tuning the regularization factor led to improved performance on newly acquired data.

In summary, the CNNs used in this study appear to be more robust against different sorts of noise compared to PLS models, whereas it is not clear which of the mentioned aspects are the main contributors to the robustness of the model. It is thus suggested to evaluate the robustness of chemometric models in an independent study specifically designed for that purpose.

4.5 Conclusions

This study demonstrates the augmentation, optimization, and interpretation of CNNs for process monitoring using spectroscopic data. The herein proposed LSA data augmentation method was shown to generate realistic *in silico* spectra for three data sets based on UV/Vis spectroscopy of model proteins and mAb size variants, and one data set based on IR spectroscopy. By augmenting CNN models with the *in silico* data, the prediction accuracy for the detection of mAb size variants could be improved by up to 50 % compared to conventional PLS models. Through automated optimization, the neural network architecture and the configuration of other model elements could be simultaneously tuned to maximize model performance. The combined optimization led to neural network structures of with multiple convolutional blocks similar to previously published models, while most of the accuracy boost could be attributed to the data augmentation approach for UV/Vis spectroscopy data. Transferring the optimized architecture without prior HPO to a data set with multiple mAb size variants resulted in comparable or superior accuracy compared to PLS models, proving good generalizability of the optimized CNN model. Although the CNN model in combination with the LSA data augmentation method provided promising results for four data sets, further studies are required to validate the applicability of the LSA method to cases where more background effects are present such as IR or Raman spectroscopy data. Finally, the optimized CNN and PLS models were directly compared with regard to the importance of different wavelength regions and robustness against spectral noise. While model-agnostic methods such as GradCAM and VIPs were able to provide specific importance estimates for the CNN and PLS models, respectively, SHAP was able to resolve differences between the two model types directly and suggested an improved capability to discriminate between mAb size-variants. Additionally, the studied CNNs appear to be more robust against synthetic white noise and peak shifts in the spectral data. However, this property should be evaluated in more detail with real data involving biological and spectral variability.

In summary, this study expands on previous works on CNNs for chemometrics and proves the applicability of the suggested methods for the quantification of model proteins and mAb size variants. The deployment of the demonstrated workflow is considered to improve the accuracy, generalizability, and scalability of chemometric models used for process monitoring and control.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author contributions

J.H. initiated and supervised the work. J.H. and M.R. were involved in funding acquisition. M.R. derived the idea for and prototyped the data augmentation method. R.S. evolved the concepts and methods presented in this manuscript, advanced the data augmentation method, performed hyperparameter optimization and interpretability assessment of the established models. R.S. analyzed and interpreted the data, drafted the figures and wrote the manuscript. All authors reviewed and approved the final manuscript.

Acknowledgments

The authors would like to thank Christina H. Wegner and Annabelle Dietrich for careful proof-reading of the manuscript. We would further like to thank the Karlsruhe Graduate School for

Computational and Data Science (KCDS) and particularly Sebastian Krumscheid for the valuable scientific discussions. We acknowledge support by the KIT-Publication Fund of the Karlsruhe Institute of Technology.

Data availability statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Chapter references

- Acquarelli, Jacopo et al. (02/2017). „Convolutional neural networks for vibrational spectroscopic data analysis“. In: *Analytica Chimica Acta* 954.June 2018 (June 2018), pp. 22–31.
- Afseth, Nils Kristian and Achim Kohler (2012). „Extended multiplicative signal correction in vibrational spectroscopy, a tutorial“. In: *Chemometrics and Intelligent Laboratory Systems* 117, pp. 92–99.
- Akiba, Takuya et al. (2019). *Optuna: A Next-generation Hyperparameter Optimization Framework*.
- Anderson, N.T. et al. (10/2020). „Achieving robustness across season, location and cultivar for a NIRS model for intact mango fruit dry matter content“. en. In: *Postharvest Biology and Technology* 168, p. 111202.
- Bakeev, Katherine A (2005). *Process Analytical Technology*. Oxford, GB: Blackwell Publishing.
- Banner, Matthew et al. (12/2021). „A decade in review: use of data analytics within the biopharmaceutical sector“. In: *Current Opinion in Chemical Engineering* 34, p. 100758.
- Belle, Vaishak and Ioannis Papantonis (07/2021). „Principles and Practice of Explainable Machine Learning“. In: *Frontiers in Big Data* 4.July, pp. 1–25.
- Bergstra, J., D. Yamins, and D. D. Cox (2012). *Making a Science of Model Search*.
- Bjerrum, Esben Jannik, Mads Glahder, and Thomas Skov (10/2017). „Data Augmentation of Spectral Data for Convolutional Neural Network (CNN) Based Deep Chemometrics“. In: pp. 1–10.
- Blazhko, Uladzislau et al. (08/2021). „Comparison of augmentation and pre-processing for deep learning and chemometric classification of infrared spectra“. en. In: *Chemometrics and Intelligent Laboratory Systems* 215, p. 104367.
- Brestrich, Nina, Tobias Hahn, and Jürgen Hubbuch (03/2016). „Application of spectral deconvolution and inverse mechanistic modelling as a tool for root cause investigation in protein chromatography“. In: *Journal of Chromatography A* 1437, pp. 158–167.
- Brestrich, Nina, Matthias Rüdte, et al. (02/2018). „Selective protein quantification for preparative chromatography using variable pathlength UV/Vis spectroscopy and partial least squares regression“. In: *Chemical Engineering Science* 176.C, pp. 157–164.
- Brestrich, Nina, Adrian Sanden, et al. (07/2015). „Advances in inline quantification of co-eluting proteins in chromatography: Process-data-based model calibration and application towards real-life separation issues“. eng. In: *Biotechnology and Bioengineering* 112.7, pp. 1406–1416.
- Brunel, Benjamin, Fatima Alsamad, and Olivier Piot (01/2021). „Toward automated machine learning in vibrational spectroscopy - use and settings of genetic algorithms for pre-processing and regression optimization“. In: *Chemometrics and Intelligent Laboratory Systems* 219.104444, pp. 1–10.
- Burkart, Nadia and Marco F. Huber (01/2021). „A Survey on the Explainability of Supervised Machine Learning“. In: *Journal of Artificial Intelligence Research* 70, pp. 245–317.
- Capito, Florian, Romas Skudas, Bernd Stanislawski, et al. (01/2013). „Matrix effects during monitoring of antibody and host cell proteins using attenuated total reflection spectroscopy“. In: *Biotechnology Progress* 29.1, pp. 265–274.
- Christler, Anna et al. (10/2021). „Technology transfer of a monitoring system to predict product concentration and purity of biopharmaceuticals in real-time during chromatographic separation“. In: *Biotechnology and Bioengineering* 118.10, pp. 3941–3952.
- Covert, Ian, Scott Lundberg, and Su-In Lee (11/2020a). „Explaining by Removing: A Unified Framework for Model Explanation“. In: *Journal of Machine Learning Research* 22, pp. 1–90.
- Covert, Ian, Scott Lundberg, and Su-In Lee (04/2020b). „Understanding Global Feature Contributions With Additive Importance Measures“. In: *Advances in Neural Information Processing Systems* 2020-Decem.NeurIPS.
- Cui, Chenhao and Tom Fearn (02/2017). „Comparison of partial least squares regression, least squares support vector machines, and Gaussian process regression for a near infrared calibration“. en. In: *Journal of Near Infrared Spectroscopy* 25.1 (1), pp. 5–14.
- Cui, Chenhao and Tom Fearn (11/2018). „Modern practical convolutional neural networks for multivariate regression: Applications to NIR calibration“. In: *Chemometrics and Intelligent Laboratory Systems* 182, pp. 9–20.
- FDA (2004). „Guidance for Industry: PAT—a framework for innovative pharmaceutical development, manufacturing, and quality assurance“. In.
- FDA (2023). *Artificial Intelligence in Drug Manufacturing*.
- Feidl et al. (10/2019). „Combining Mechanistic Modeling and Raman Spectroscopy for Monitoring Antibody Chromatographic Purification“. In: *Processes* 7.10, pp. 683–683.
- Feng, Steven Y. et al. (2021). „A Survey of Data Augmentation Approaches for NLP“. In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 968–988.
- Feurer, Matthias and Frank Hutter (2019). „Hyperparameter Optimization“. In: *Automated Machine Learning: Methods, Systems, Challenges*. Ed. by Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren. Cham: Springer International Publishing, pp. 3–33.
- Glassey, Jarka et al. (04/2011). „Process analytical technology (PAT) for biopharmaceuticals“. In: *Biotechnology Journal* 6.4, pp. 369–377.
- Goldrick, Stephen, Alexandra Umprecht, et al. (09/2020). „High-Throughput Raman Spectroscopy Combined with Innovative Data Analysis Workflow to Enhance Biopharmaceutical Process Development“. In: *Processes* 8.9 (9), pp. 1179–1179.
- Goodfellow, Ian J., Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. Cambridge, MA, USA: MIT Press.
- Gu, Jiuxiang et al. (05/2018). „Recent advances in convolutional neural networks“. In: *Pattern Recognition* 77, pp. 354–377.
- Guindo, Mahamed Lamine et al. (01/2023). „Chemometric Approach Based on Explainable AI for Rapid Assessment of Macronutrients in Different Organic Fertilizers Using Fusion Spectra“. In: *Molecules* 28.2.
- Guo, Fan, Ruimin Xie, and Biao Huang (02/2020). „A deep learning just-in-time modeling approach for soft sensor based on variational autoencoder“. In: *Chemometrics and Intelligent Laboratory Systems* 197, p. 103922.

CHAPTER 4. DATA AUGMENTATION AND CNNs

- Haghi, R. K., E. Pérez-Fernández, and A. H.J. Robertson (08/2021). „Prediction of various soil properties for a national spatial dataset of Scottish soils based on four different chemometric approaches: A comparison of near infrared and mid-infrared spectroscopy“. In: *Geoderma* 396.
- Hansen, Sigrid K., Erik Skibsted, et al. (11/2011). „A label-free methodology for selective protein quantification by means of absorption measurements“. eng. In: *Biotechnology and Bioengineering* 108.11, pp. 2661–2669.
- Jiang, Mo et al. (11/2017). „Opportunities and challenges of real-time release testing in biopharmaceutical manufacturing“. In: *Biotechnology and Bioengineering* 114 (11), pp. 2445–2456.
- Kingma, Diederik P. and Jimmy Lei Ba (2015). „Adam: A method for stochastic optimization“. In: *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pp. 1–15.
- Krause, Julius et al. (04/2021). „New active learning algorithms for near-infrared spectroscopy in agricultural applications“. In: *at - Automatisierungstechnik* 69.4, pp. 297–306.
- Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton (05/2017). „ImageNet classification with deep convolutional neural networks“. en. In: *Communications of the ACM* 60.6, pp. 84–90.
- Li, Meng-Yao et al. (03/2018). „Real-time monitoring of antibody glycosylation site occupancy by in situ Raman spectroscopy during bioreactor CHO cell cultures“. In: *Biotechnology Progress* 34.2 (2), pp. 486–493.
- Liu, Jinchao et al. (2017). „Deep convolutional neural networks for Raman spectrum recognition: A unified solution“. In: *Analyst* 142.21 (21), pp. 4067–4074.
- Long, James R., Vasilis G. Gregoriou, and Paul J. Gemperline (09/1990). „Spectroscopic calibration and quantitation using artificial neural networks“. In: *Analytical Chemistry* 62.17 (17), pp. 1791–1797.
- Lundberg, Scott M, Paul G Allen, and Su-In Lee (2017). „A Unified Approach to Interpreting Model Predictions“. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30, pp. 4765–4774.
- Mahynski, Nathan A. et al. (10/2022). „Building Interpretable Machine Learning Models to Identify Chemometric Trends in Seabirds of the North Pacific Ocean“. In: *Environmental Science and Technology* 56.20, pp. 14361–14374.
- Malek, Salim, Farid Melgani, and Yakoub Bazi (2018). „One-dimensional convolutional neural networks for spectroscopic signal regression“. In: *Journal of Chemometrics* 32.5 (5), pp. 1–17.
- Markl, Daniel et al. (05/2020). „Review of real-time release testing of pharmaceutical tablets: State-of-the art, challenges and future perspective“. In: *International Journal of Pharmaceutics* 582, p. 119353.
- Martens, Harald and Edward Stark (01/1991b). „Extended multiplicative signal correction and spectral interference subtraction: New preprocessing methods for near infrared spectroscopy“. In: *Journal of Pharmaceutical and Biomedical Analysis* 9.8, pp. 625–635.
- McHardy, Rose G. et al. (08/2023). „Augmentation of FTIR spectral datasets using Wasserstein generative adversarial networks for cancer liquid biopsies“. en. In: *Analyst* 148.16, pp. 3860–3869.
- Mehmood, Tahir et al. (08/2012). „A review of variable selection methods in Partial Least Squares Regression“. In: *Chemometrics and Intelligent Laboratory Systems* 118, pp. 62–69.
- Mishra, Puneet and Ittai Herrmann (08/2021). „GAN meets chemometrics: Segmenting spectral images with pixel2pixel image translation with conditional generative adversarial networks“. In: *Chemometrics and Intelligent Laboratory Systems* 215, p. 104362.
- Mishra, Puneet and Dário Passos (05/2021a). „A synergistic use of chemometrics and deep learning improved the predictive performance of near-infrared spectroscopy models for dry matter prediction in mango fruit“. In: *Chemometrics and Intelligent Laboratory Systems* 212.March (March), p. 104287.
- Mishra, Puneet and Dário Passos (2021b). „Deep chemometrics: Validation and transfer of a global deep near-infrared fruit model to use it on a new portable instrument“. en. In: *Journal of Chemometrics* 35.10, e3367.
- Mishra, Puneet and Dário Passos (05/2021c). „Realizing transfer learning for updating deep learning models of spectral data to be used in new scenarios“. In: *Chemometrics and Intelligent Laboratory Systems* 212, p. 104283.
- Nikzad-Langerodi, Ramin and Erik Andries (2021). „A chemometrician’s guide to transfer learning“. en. In: *Journal of Chemometrics* 35.11, e3373.
- Passos, Dário and Puneet Mishra (08/2021). „An automated deep learning pipeline based on advanced optimisations for leveraging spectral classification modelling“. In: *Chemometrics and Intelligent Laboratory Systems* 215.May (May), p. 104354.
- Passos, Dário and Puneet Mishra (04/2022). „A tutorial on automatic hyperparameter tuning of deep spectral modelling for regression and classification tasks“. In: *Chemometrics and Intelligent Laboratory Systems* 223, p. 104520.
- Read, E. K. et al. (2010). „Process Analytical Technology (PAT) for biopharmaceutical products: Part II. Concepts and applications“. In: *Biotechnology and Bioengineering* 105.2, pp. 285–295.
- Read, E.K. et al. (02/2010). „Process analytical technology (PAT) for biopharmaceutical products: Part I. concepts and applications“. In: *Biotechnology and Bioengineering* 105 (2), pp. 276–284.
- Rodríguez-Rodríguez, José A. et al. (2021). „The Impact of Linear Motion Blur on the Object Recognition Efficiency of Deep Convolutional Neural Networks“. en. In: *Pattern Recognition. ICPR International Workshops and Challenges*. Ed. by Alberto Del Bimbo et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 611–622.
- Rolinger, Laura, Jürgen Hubbuch, and Matthias Rüdts (02/2023). „Monitoring of ultra- and diafiltration processes by Kalman-filtered Raman measurements“. In: *Analytical and Bioanalytical Chemistry* 415.5, pp. 841–854.
- Rolinger, Laura, Matthias Rüdts, and Jürgen Hubbuch (04/2020a). „A critical review of recent trends, and a future perspective of optical spectroscopy as PAT in biopharmaceutical downstream processing“. In: *Analytical and Bioanalytical Chemistry* 412.9, pp. 2047–2064.
- Rolinger, Laura, Matthias Rüdts, and Jürgen Hubbuch (11/2021). „Comparison of UV- and Raman-based monitoring of the Protein A load phase and evaluation of data fusion by PLS models and CNNs“. In: *Biotechnology and Bioengineering* 118.11 (11), pp. 4255–4268.

- Romann, Patrick et al. (11/2022). „Advancing Raman model calibration for perfusion bioprocesses using spiked harvest libraries“. In: *Biotechnology Journal* 17.11, e2200184.
- Rosebrock, A. (2018). *Deep Learning for Computer Vision with Python*. pyimagearch.
- Rüdt, Matthias, Nina Brestrich, et al. (02/2017). „Real-time monitoring and control of the load phase of a protein A capture step“. In: *Biotechnology and Bioengineering* 114.2 (2), pp. 368–373.
- Rüdt, Matthias, Till Briskot, and Jürgen Hubbuch (03/2017). „Advances in downstream processing of biologics – Spectroscopy: An emerging process analytical technology“. In: *Journal of Chromatography A* 1490, pp. 2–9.
- Rüdt, Matthias, Philipp Vormittag, et al. (06/2019). „Process monitoring of virus-like particle reassembly by diafiltration with UV/Vis spectroscopy and light scattering“. In: *Biotechnology and Bioengineering* 116.6, pp. 1366–1379.
- Sanden, Adrian et al. (12/2019). „Fourier-transform infrared spectroscopy as a process analytical technology for near real time in-line estimation of the degree of PEGylation in chromatography“. In: *Journal of Chromatography A* 1608, p. 460410.
- Santos, Vianney O. et al. (08/2005). „A comparative study of diesel analysis by FTIR, FTNIR and FT-Raman spectroscopy using PLS and artificial neural network analysis“. In: *Analytica Chimica Acta* 547.2 (2), pp. 188–196.
- Sauer, Dominik Georg et al. (08/2019). „Real-time monitoring and model-based prediction of purity and quantity during a chromatographic capture of fibroblast growth factor 2“. In: *Biotechnology and Bioengineering* 116.8 (8), pp. 1999–2009.
- Schiemer, Robin, Jan Tobias Weggen, et al. (05/2023). „An adaptive soft-sensor for advanced real-time monitoring of an antibody-drug conjugation reaction“. en. In: *Biotechnology and Bioengineering* 120.7, pp. 1914–1928.
- Selvaraju, Ramprasaath R. et al. (2020). „Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization“. In: *International Journal of Computer Vision* 128 (2), pp. 336–359.
- Shorten, Connor and Taghi M. Khoshgoftar (2019). „A survey on Image Data Augmentation for Deep Learning“. In: *Journal of Big Data* 6.1.
- Štrumbelj, Erik and Igor Kononenko (12/2014b). „Explaining prediction models and individual predictions with feature contributions“. In: *Knowledge and Information Systems* 41.3 (3), pp. 647–665.
- Szegedy, Christian et al. (2015). „Going deeper with convolutions“. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9.
- Trampuž, Marko, Dušan Teslić, and Blaž Likozar (2020). „Process analytical technology-based (PAT) model simulations of a combined cooling, seeded and antisolvent crystallization of an active pharmaceutical ingredient (API)“. In: *Powder Technology* 366, pp. 873–890.
- Tulsyan, Aditya et al. (2021). „Spectroscopic models for real-time monitoring of cell culture processes using spatiotemporal just-in-time Gaussian processes“. In: *AIChE Journal* 67.5 (5).
- Ündey, Cenk et al. (2010). „Applied advanced process analytics in biopharmaceutical manufacturing: Challenges and prospects in real-time monitoring and control“. In: *Journal of Process Control* 20.9, pp. 1009–1018.
- Venton, J. et al. (12/2021). „Robustness of convolutional neural networks to physiological electrocardiogram noise“. en. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 379.2212, p. 20200262.
- Wang, Bryan et al. (2022). „Process analytical technologies in cell therapy manufacturing: State-of-the-art and future directions“. In: *Journal of Advanced Manufacturing and Processing* 4.1, pp. 1–17.
- Wang, Jiarui et al. (2023a). „Automated calibration and in-line measurement of product quality during therapeutic monoclonal antibody purification using Raman spectroscopy“. In: *Biotechnology and Bioengineering*.
- Wegner, Christina H., Ines Zimmermann, and Jürgen Hubbuch (02/02/2022). „Rapid Analysis for Multicomponent High-Throughput Crystallization Screening: Combination of UV–Vis Spectroscopy and Chemometrics“. In: *Crystal Growth & Design* 22.2, pp. 1054–1065.
- Wei, Bingchuan et al. (12/2022). „Multi-attribute Raman spectroscopy (MARS) for monitoring product quality attributes in formulated monoclonal antibody therapeutics“. In: *mAbs* 14.1 (1).
- Wold, Svante, Michael Sjöström, and Lennart Eriksson (10/2001). „PLS-regression: a basic tool of chemometrics“. en. In: *Chemometrics and Intelligent Laboratory Systems* 58.2 (2), pp. 109–130.
- Wu, Man et al. (12/2021). „Deep learning data augmentation for Raman spectroscopy cancer tissue classification“. en. In: *Scientific Reports* 11.1, p. 23842.
- Yosinski, Jason et al. (06/2015). „Understanding Neural Networks Through Deep Visualization“. In.
- Yuanyuan, Chen and Wang Zhibin (10/2018). „Quantitative analysis modeling of infrared spectroscopy based on ensemble convolutional neural networks“. In: *Chemometrics and Intelligent Laboratory Systems* 181, pp. 1–10.
- Zeiler, Matthew D. and Rob Fergus (11/2013). „Visualizing and Understanding Convolutional Networks“. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 8689 LNCS (PART 1), pp. 818–833.
- Zhang, Chi et al. (2019). „Rapid, quantitative determination of aggregation and particle formation for antibody drug conjugate therapeutics with label-free Raman spectroscopy“. In: *Analytica Chimica Acta* 1081, pp. 138–145.

5

Raman-based PAT for VLP precipitation: Systematic data diversification and preprocessing pipeline identification

Annabelle Dietrich^{1,+}, Robin Schiemer^{1,+}, Jasper Kurmann¹, Shiqi Zhang¹, and Jürgen Hubbuch¹

¹ Institute of Engineering in Life Sciences, Section IV: Biomolecular Separation Engineering, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

⁺ contributed equally

Abstract

Virus-like particles (VLPs) are a promising class of biopharmaceuticals for vaccines and targeted delivery. Starting from clarified lysate, VLPs are typically captured by selective precipitation. While VLP precipitation is induced by step-wise or continuous precipitant addition, current monitoring approaches do not support the direct product quantification, and analytical methods usually require various, time-consuming processing and sample preparation steps. Here, the application of Raman spectroscopy combined with chemometric methods may allow the simultaneous quantification of the precipitated VLPs and precipitant owing to its demonstrated advantages in analyzing crude, complex mixtures. In this study, we present a Raman spectroscopy-based process analytical technology (PAT) tool developed on batch and fed-batch precipitation experiments of Hepatitis B core Antigen VLPs. We conducted small-scale precipitation experiments providing a diversified data set with varying precipitation dynamics and backgrounds induced by initial dilution or spiking of clarified *Escherichia coli* (*E. coli*)-derived lysates. For the Raman spectroscopy data, various preprocessing operations were systematically combined allowing the identification of a preprocessing pipeline, which proved to effectively eliminate initial lysate composition variations as well as most interferences attributed to precipitates and the precipitant present in solution. The calibrated partial least squares (PLS) models seamlessly predicted the precipitant concentration with R^2 of 0.98 and 0.97 in batch and fed-batch experiments, respectively, and captured the observed precipitation trends with R^2 of 0.74 and 0.64. Although the resolution of fine differences between experiments was limited due to the observed non-linear relationship between spectral data and the VLP concentration, this study provides a foundation for employing Raman spectroscopy as a PAT sensor for monitoring VLP precipitation processes with the potential to extend its applicability to other phase-behavior dependent processes or molecules.

5.1 Introduction

Virus-like particles (VLPs) as non-viral vectors have emerged as class of protein nanoparticles for vaccines, surface antigen presentation and targeted delivery (Qian et al. 2020). VLPs are composed of viral protein subunits mimicking the structure of the virus they descend from (Chackerian 2007; Zeltins 2013). Since VLPs lack the active viral genome, they are considered safer than other viral nanoparticles such as adeno-associated virus or lentivirus (Chackerian 2007; Nooraei et al. 2021). The recombinant production of VLPs in eukaryotic or prokaryotic cells (Kushnir, Streatfield, and Yusibov 2012) lead to process material of high variability and complexity, which is even more pronounced for intracellular production as a lysis step is required (Effio and Hubbuch 2015). Hence, in early downstream processing (DSP), the removal of the majority of process-related host cell impurities is particularly advised (Effio and Hubbuch 2015). Native precipitation was found highly selective for VLPs and is induced by addition of a precipitation agent, using either polyethylene glycol (Tsoka et al. 2000; Koho et al. 2012) or ammonium sulfate (AMS) (Kim et al. 2010; Zahin et al. 2016; Kazaks et al. 2017). Whereas steric exclusion is typically associated with polyethylene glycol (Iverius and Laurent 1967; Poison 1977), the effect of protein surface charge predominates for precipitation induced by AMS as sulphate is strongly kosmotropic (Curtis, Prausnitz, and H. W. Blanch 1998). The protein surface charge is in turn affected by its structural properties, which may impact the precipitation behaviour (Hämmerling et al. 2017) and hence, the precipitation agent concentration required for precipitation.

The Hepatitis B core Antigen (HBcAg) VLP, a recombinantly produced internal protein capsid of the Hepatitis B virus, is subject to intensive research for diverse medical purposes (Cooper

CHAPTER 5. RAMAN-BASED PAT FOR VLP PRECIPITATION

and Shaul 2005; Porterfield and Adam Zlotnick 2010; Klamp et al. 2011; Mobini et al. 2020; Moradi Vahdat et al. 2021; Petrovskis et al. 2021; Hassebroek et al. 2023). In previous capture studies, the AMS-induced precipitation of HBcAg VLPs was found highly selective (Valentic, J. Müller, and Hubbuch 2022), while co-precipitation of impurities was only observed for higher AMS concentrations (Wegner, Zimmermann, and Hubbuch 2022). Moreover, the required AMS concentration for precipitation was found to be dependent on the surface properties comparing HBcAg VLPs with and without foreign epitopes rather than on internal structural differences comparing various lengths of nucleic acid binding sites (Hillebrandt, Vormittag, Bluthardt, et al. 2020; Valentic, J. Müller, and Hubbuch 2022). In the context of process development, the conventional centrifugation-based HBcAg VLP capture has been replaced by an innovative setup involving fed-batch precipitation and diafiltration-based wash and redissolution steps (Hillebrandt, Vormittag, Bluthardt, et al. 2020). This development aimed to establish a size-based platform process applicable to various protein nanoparticles. However, the development of a standardized platform process for protein nanoparticle DSP remains uncommon, primarily due to their extensive variety, which presents significant challenges in processing (Moleirinho et al. 2020).

To be able to operate such a capture platform process for different protein nanoparticles, the implementation of process analytical technology (PAT) for process monitoring is crucial. Since 2004, the FDA has underscored the importance of real-time process monitoring for its role in enhancing process understanding, ensuring process robustness, and guaranteeing product safety within the biopharmaceutical industry (FDA 2004; A. S. Rathore, Bhambure, and Ghare 2010; Glassey et al. 2011). Besides the optical spectroscopic techniques ultraviolet/visible (UV/Vis) and infrared (IR) spectroscopy, Raman spectroscopy coupled with chemometrics has found extensive applications in monitoring various processes for biopharmaceutical products, including raw material testing (B. Li et al. 2010), cell culture (Abu-Absi et al. 2011; Berry et al. 2015; Golabgir and Herwig 2016), chromatography (Feidl et al. 2019; Rolinger, Rüdte, and Hubbuch 2021; Jiarui Wang et al. 2023b), filtration (Rolinger, Hubbuch, and Rüdte 2023), freezing (Roessl et al. 2014; D. Weber and Hubbuch 2021), or formulation (Wei et al. 2022). For VLPs in particular, recent studies demonstrated the real-time monitoring of a baculovirus cultivation for the production of rabies VLPs (Guardalini, Silva Cavalcante, et al. 2023; Guardalini, Aragão Tejo Dias, et al. 2023; Guardalini, Rangel, et al. 2023) as well as the cross-flow filtration-based polishing operations such as dis- and reassembly of the HBcAg-VLPs (Hillebrandt, Vormittag, Dietrich, and Hubbuch 2022; Rüdte, Vormittag, et al. 2019). Despite the broad applicability of spectroscopic methods in biopharmaceutical processing, their application to precipitation processes is rather unexplored. While multiple studies have used near-infrared (NIR) spectroscopy for monitoring of acid (Sun, Fan, et al. 2020) or alcohol precipitation (L. Li et al. 2014; Sun, Jiayue Wang, et al. 2018; H. Huang and Qu 2018), Raman spectroscopy has only been used to track a precipitation polymerization reaction (Meyer-Kirschner et al. 2016) and monoclonal antibody precipitation (Lohmann and Strube 2021). A major challenge when working with precipitate-containing solutions is the turbidity of the media, which was reported to affect the overall signal intensity of Raman spectra (Van Brink, Pepers, and Van Herk 2002; Sinfield and Monwuba 2014; Meyer-Kirschner et al. 2016). The loss in signal intensity may either be corrected by isolated Raman bands (Sinfield and Monwuba 2014; Meyer-Kirschner et al. 2016) or be directly used for correlation to a target quality attribute (Meyer-Kirschner et al. 2016). Similarly, Zelger et al. (2016) have used a turbidity sensor for monitoring protein precipitation, foregoing the molecular information that could have been attained through the utilization of spectroscopic techniques.

Among the available spectroscopic techniques, Raman spectroscopy stands out for its selectivity to monitor multiple quality attributes (R. M. Santos et al. 2018; Wei et al. 2022; Jiarui Wang et al. 2023b) as well as its capability to analyze complex solutions such as fermentation broths (Abu-Absi et al. 2011; Berry et al. 2015; Goldrick, Umprecht, et al. 2020) or clarified cell lysates (Hassoun et al. 2018; Genova et al. 2018). However, Raman spectroscopy comes with a set of challenges, such as low signal intensity and strong background interferences requiring extensive data preprocessing and model optimization (Rinnan et al. 2009; Bocklitz et al. 2011; Gautam et al. 2015). Due to this sensitivity, undesired data variability can arise by changes in process material or sample processing, compromising comparability and accuracy in the recorded data. Popular approaches for the removal of undesirable systematic variation include signal corrections, filtering, or variable selection (Rinnan et al. 2009; Bocklitz et al. 2011; Mehmood et al. 2012). Most commonly, baseline (S. He et al. 2014; F. Zhang et al. 2020), background (S. Wold, Antti, et al. 1998; Trygg and S. Wold 2002; Boelens et al. 2004), and scatter (Martens, Nielsen, and Engelsen 2003; A. Kohler et al. 2008; Liland, Achim Kohler, and Afseth 2016) correction methods are employed to eliminate interferences caused by contaminating species or scattering effects such as fluorescence or Mie scattering. Since the direct subtraction of reference spectra as a basic approach for background correction might result in insufficient- or over-corrections (Boelens et al. 2004), multiple approaches have been proposed to identify and remove the undesirable background variation (Lorber 1986; S. Wold, Antti, et al. 1998; Trygg and S. Wold 2002; Ben-David and Ren 2005). Among those the most commonly applied algorithm is the orthogonal projection to latent structures (OPLS) as proposed by Trygg and S. Wold 2002 effectively removing systematic background variation not correlated with the target quality attribute (C. P. Passos et al. 2010; Abdallah et al. 2019). To further optimize model performance, additional steps like spectral cropping, derivative filtering, or filter-based variable selection may be used (Gautam et al. 2015; R. M. Santos et al. 2018). In principle, the sequence of preprocessing operations, or preprocessing pipeline, should be designed to remove irrelevant variations from the recorded data and provide the multivariate model with consistent information to enable robust and accurate real-time monitoring. Finally, a careful evaluation of the extracted information by the preprocessing pipeline and the selected multivariate model is demanded by the regulatory agencies (FDA 2015). For linear models, such as partial least squares (PLS), multiple approaches such as variable importance in projection (VIP)-based feature importance (Mehmood et al. 2012) have been proposed to perform this quality check.

In this study, we present a PAT approach based on Raman spectroscopy for a precipitation-based capture process of VLPs. Based on multiple small-scale precipitation experiments with material from various lysate batches and spiking materials, we evaluate the suitability of Raman spectroscopy for analyzing precipitate-containing lysates. By introducing a pretreatment strategy for the UV/Vis reference measurements and identifying a suitable preprocessing pipeline for Raman spectroscopy data, we demonstrate the reduction of variance in both the Raman and the reference data and enable the prediction of the precipitated VLPs from precipitate-containing lysates. We systematically combine numerous preprocessing operations and isolate the effects of the included operations on model performance. Eventually, we transfer this approach to a fed-batch precipitation process, ensuring its adaptability and effectiveness across different operational paradigms.

5.2 Materials and methods

5.2.1 Experiments

5.2.1.1 Virus-like particles

In this study, Cp149, a C-terminally truncated wild-type HBcAg protein (A. Zlotnick et al. 1996) was used, for which the plasmid was kindly provided by Prof. Adam Zlotnick (Indiana University, US). The intracellular expression of Cp149 in *Escherichia coli* (*E. coli*), cell harvest, cell lysis, and lysate clarification were conducted as previously described by Hillebrandt, Vormittag, Bluthardt, et al. 2020. Clarified lysate solutions (hereinafter termed lysate) were thawed, 0.2 μm -filtered, directly used raw or conditioned by dilution or spiking depending on the precipitation experiment, and adjusted to 0.25% (v/v) polysorbate 20.

To systematically diversify the lysate for the precipitation experiments, different lysate batches were used and initial lysate conditioning was performed by dilution and spiking for eight batch experiments B1-B8 and three fed-batch experiments F1-F3. Raw lysates of two lysate batches were used for batch experiments B1 and B4, while a third lysate batch served as lysate for all fed-batch experiments. Dilution of the lysates introduced further changes in initial lysate concentration (B2-B3, F1, F3), which is expressed as volume percent lysate content and was simply measured as 280 nm absorbance ($A_{280\text{L}}$) due to the lysates complexity.

To establish a foundation for systematic spiking, the initial composition of raw lysate was first estimated in pre-experiments (data not shown) to estimate a VLP content despite the lysates complexity. For this pre-experiment, selective VLP precipitation was conducted by adjusting raw lysate to 1.1 M AMS and the supernatant was analyzed by reference analytics. The difference between the ultraviolet (UV) absorbance at 280 nm of the initial lysate ($A_{280\text{L}}$) and the volume-corrected supernatant after precipitation ($A_{280\text{S}}$) account for the absorbance of selectively precipitated VLP ($A_{280\text{VLP}}$). The calculated ratio $A_{280\text{VLP}}/A_{280\text{L}}$ expressing the estimated VLP content resulted in approximately 10% VLP present in the raw lysate. This ratio served as foundation for experimental spiking design under the assumption of purified VLP- or host cell protein (HCP)-enriched spiking material. While the estimated VLP content with 10% remained constant using raw lysate (B1, B4) or during lysate conditioning by dilution (B2-B3, F1, F3) or spiking with salt (B8), the estimated VLP content was systematically varied by VLP- or HCP-spiking (B5-B7, F2). Note that spiking the lysates also resulted in a form of dilution; however, the primary focus was here on systematically adjusting the lysate composition. The final lysate conditioning settings of lysate batch, lysate content, and estimated VLP content are summarized for all batch experiments B1-B8 and fed-batch experiments F1-F3 in Table 1 and Table 2, respectively.

More details on preparation and composition of buffers, solutions, and VLP- and HCP-enriched spiking material are stated in Supplementary Sec. C.

5.2.1.2 Batch precipitation experiments

Conditioned lysate varied in lysate batch, concentration or composition across the eight batch VLP precipitation experiments as summarized in Tab. 5.1. For each experiment, eleven solutions were prepared in 1040 μL -scale with precipitant concentrations in the range of 0 to 1.2 M AMS. In general, the solution compositions were designed to maximize the lysate content under the limitation of 4 M AMS stock solution to set the highest target AMS concentration of 1.2 M, resulting in 728 μL lysate and 312 μL 4 M AMS stock solution. Maintaining 728 μL lysate to prevent a dilution effect within the experiments in batch mode, the remaining 312 μL was composed of proportionally 4 M

Table 5.1 Overview of the eight batch precipitation experiments B1-B8 comprising parameters for initial lysate conditioning and the final, curve fit-derived, apparent concentration of precipitated VLPs.

Batch	Condition	Lysate conditioning		Precipitated VLP
		Batch	Estimated VLP content	
		-	% (v/v)	g L ⁻¹
B1	raw		100	4.35
B2	dilution	lysate 1	52.5	2.96
B3	dilution		35	2.06
B4	raw		100	3.92
B5	VLP spike		50	4.15
B6	VLP spike	lysate 2	50	4.16
B7	HCP spike		31	1.49
B8	NaCl spike		80	4.69

Table 5.2 Overview of the three fed-batch precipitation experiments F1-F3 comprising parameters for initial lysate conditioning and fed-batch processing. Lysate batch 3 was used for all experiments.

Fed-batch	Condition	Lysate conditioning		Processing	
		Lysate content	Estimated VLP content	Time	Feed rate
		% (v/v)	% (A280 _{VLP} /A280 _L)	min	mL min ⁻¹
F1	dilution	43	10	30	0.16
F2	VLP spike	32	23	30	0.16
F3	dilution	45	10	20	0.30

AMS stock solution and ultrapure water to cover the other precipitant concentrations below 1.2 M AMS. Solutions were incubated at 22 °C on a thermo-shaker ThermoMixer Comfort (Eppendorf, Hamburg, DE) at 500 rpm for 30 min. For Raman measurements, 200 μ L samples of the turbid precipitate solutions were taken and immediately analyzed. To enable measuring the particulate-free supernatants, the remaining turbid precipitate solutions were centrifuged at 12000 rcf for 8 min in a Pico 17 tabletop centrifuge (Thermo Fisher Scientific Inc., Waltham, US). The supernatants were analyzed via Raman spectroscopy and UV reference analytics.

5.2.1.3 Fed-batch precipitation experiments

The setup for fed-batch VLP precipitation consisted of a stirred reservoir equipped with a Minipuls 3 peristaltic pump (Gilson, Villiers le Bel, FR) for a 4 M AMS feed. For online monitoring, a second peristaltic pump (Gilson), a SLS-1500 flow meter (Sensirion), and a flow cell for Raman measurements were connected with PEEK capillaries with an inner diameter of 0.25 mm in an on-line loop. The on-line loop flow rate was adjusted to 1.4 mL/min. Fed-batch precipitation experiments were performed with 12 mL conditioned lysate. The lysate and processing conditions are listed in Table 5.2. During fed-batch precipitation, 200 μ L samples were taken at each time step of in total 30 time steps, from which 22 selected samples were centrifuged at 12000 rcf for 3 min facilitating UV reference analytics of the supernatants. For each time step, lysate and AMS content were calculated considering drawn sample volume and a mean feed volume, which was based on a mean feed flow rate considering total AMS amount and feed density.

5.2.1.4 Raman spectroscopy

The Raman BioReactor BallProbe inserted in the FlowCell Adapter (both MarqMetrix, Seattle, US) was connected to a HyperFlux™ PRO Plus 785 with the software SpectralSoft 3.2.6 (Tornado Spectral Systems, Toronto, CA). All measurements were performed in the flow cell with a laser power of 495 mW and with 50 acquisitions per spectrum. For precipitate and supernatant samples, the exposure time was set to 275 ms and 185 ms, respectively, as with precipitate being present, a higher exposure was necessary to achieve a similar intensity level due to the dampening effect of the precipitates in solution. The entire spectral range from 200 to 3300 cm⁻¹ was recorded with a spectral resolution of 1 cm⁻¹. For each of the eleven AMS conditions per experiment, the 50 recorded Raman spectra were averaged. For the fed-batch experiments, the 50 recorded spectra closest to the sampling time point were averaged.

5.2.1.5 Reference analytics

All supernatant samples were analyzed via UV spectroscopy with a high-performance liquid chromatography (HPLC) system equipped with a RS diode array detector and controlled by Chromeleon 6.8 (Dionex Ultimate 3000 RS, Sunnyvale, US). A 0.5 µm pre-column filter cartridge (OPTI-SOLV EXP, Supelco, Bellefonte, US) but no column was installed. A buffer containing 50 mM Tris, 100 mM NaCl at pH 8.0 was used as mobile phase with a flow rate of 50 µL min⁻¹ and an injection volume of 20 µL. Samples were diluted 100-fold in triplicates in lysis buffer and UV spectra in the wavelength range from 220 to 400 nm were recorded.

5.2.2 Data analysis and computation

Data analysis and computation was performed in Python 3.8.

5.2.2.1 Reference data pretreatment

Several pretreatments steps for the UV reference data were conducted in order to transform the scattered data into reliable reference data for the chemometric modeling. The pretreatment comprised scatter correction, scaling, precipitation curve estimation, outlier removal, and a final conversion to the apparent precipitated VLP concentration. Scatter correction for the 280 nm peak areas (A280) was performed to remove nucleic acid contributions according to Porterfield and Adam Zlotnick (2010), whereas AMS content-dependent scatter correction was not required due to the 100-fold sample dilution.

Each batch experiment consisted of A280-triplicates for 11 AMS conditions, resulting in 33 data points. To reduce the effect of outliers in the data within one experiment, a robust least squares fit of the precipitation curve using the Boltzmann function was performed. This model-based approximation assumes that the precipitation can be divided into three distinct phases, namely an initial plateau until the onset of the precipitation, a sigmoidal decrease, and a final plateau once all target molecules have been precipitated. In particular, the optimization problem was set up as

$$\min_{\mathbf{x}} = 0.5 \sum_{i=1}^N \delta(f(c_{\text{AMS},i}, \mathbf{x}) - y_i)^2 \quad (5.1)$$

where y_i denote the discrete UV measurements at a certain concentration of AMS i and $f(c_{\text{AMS},i}, \mathbf{x})$ being the Boltzmann function given as

$$f(c_{\text{AMS},i}, \mathbf{x}) = \frac{x_2 - x_4}{1 + \exp\left(-(x_1 c_{\text{AMS},i} + x_2)\right)} + x_3. \quad (5.2)$$

The parameter vector $\mathbf{x}$ is defined as $(x_1, x_2, x_3, x_4)^T$. By employing the non-linear transformation of the squared residuals z , a smooth approximation of the absolute value loss by $\delta(z) = 2(\sqrt{(1+z)} - 1)$ was generated to effectively reduce the impact of outliers. To generalize the problem where the threshold between inliers and outliers is different from 1, the formula $\hat{\delta}(r^2) = C^2\delta((r/C)^2)$ was used with C being a scaling factor that determines the threshold between outliers and inliers in the data set. In our case, C was set to 5 for robust parameter estimation across all batch and fed-batch experiments.

The generated Boltzmann fits were used to remove outliers in each experiment. The distribution of the residuals between the Boltzmann curves and the original measurements were analyzed. Data points with residuals outside of the interval $[p_{0.25} - 1.5IQR; p_{0.75} + 1.5IQR]$ were excluded from the data set with the interquartile range IQR being defined as $p_{0.75} - p_{0.25}$ (Wilcox 2005) and $p_{0.25}$ and $p_{0.75}$ the 25 and 75% percentiles.

Within one experiment, the conversion to apparent precipitated VLP concentration was done by subtracting the minimum observed A280 mean from all data points, taking the absolute values, and scaling by HPLC device parameters and the theoretical extinction coefficient of the VLP at 280 nm of $1.764 \text{ L g}^{-1} \text{ cm}^{-1}$ as provided by the ProtParam tool (Gasteiger et al. 2005). This procedure assumes that solely VLP is precipitating under given precipitation conditions as discussed in Wegner, Zimmermann, and Hubbuch 2022 and the absorption is only caused by VLP in solution. Under consideration of other HCPs and nucleic acids in the solution, this procedure is biased and hence the derived VLP content is termed apparent precipitated VLP concentration.

5.2.2.2 Spectral data processing

Spectral data processing of averaged spectra was done in order to reduce unwanted differences in the collected Raman data and covered spectral outlier removal, turbidity correction, baseline correction, background correction, difference spectra, derivative and smoothing, and cropping. Multiple combinations of the listed operations were screened to find the optimal configuration. Hence, the individual operations will be explained in detail.

Spectral outlier removal Spectral outlier were removed manually based on visual inspection regarding defective spectra. Defective spectra were found for experiment B2 corresponding to the conditions 1.0, 1.1 and 1.2 M AMS and occurred mainly due to sample handling and hence the inclusion of air bubbles in the flow cell.

Turbidity correction To remove interfering Raman effects caused by turbidity, the spectra were turbidity-corrected by normalization at 3299 cm^{-1} .

Baseline correction All spectra were baseline-corrected using a Whittaker filter employing the adaptive smoothness penalized least squares (asPLS) according to F. Zhang et al. 2020 and implemented in *pybaselines* (v. 1.0.0). The λ parameter was determined manually by visual inspection of the estimated baseline. The optimal λ value of 6×10^{-7} was chosen based on the most consistent baseline estimation over the entire spectral range across all samples. Furthermore, a second-order difference matrix, a maximum number of iterations of 100, and a tolerance of 1×10^{-3} were used.

Background correction To remove the interfering Raman effects of the buffer system and the precipitating agents, two different background correction methods were evaluated. The first one which we will refer to as scaled subtraction (SS) uses reference measurements of the buffer system at the respective concentration levels of AMS. The reference spectra were scaled to the Raman spectra collected in the precipitation experiments by normalization at 750 cm^{-1} . As the SS method can only

be applied when reference measurements are available, a second background correction method was evaluated, namely the OPLS as described in Trygg and S. Wold 2002. The OPLS method extracts the orthogonal systematic variation in the spectral data which is not correlated to the target quality attribute and was used as implemented in *pyopls* (v 20.02) using 3 latent variables.

Difference spectra Since only the precipitated material contributes to the changes in Raman intensity, difference spectra were computed per experiment by subtracting the first spectrum of each experiment. This is analogous to using the autozero function of other spectroscopic detectors at the beginning of the experiments. In essence, this is supposed to remove the varying background effects of the lysate matrix of which the exact composition is unknown.

Derivative and smoothing Derivative and smoothing was applied using the Savitzky-Golay filter (SGF) (Savitzky and Golay 1964). Unless stated otherwise, the SGF was used with a second-degree polynomial and a window size of 11 as implemented in *scipy* (v. 1.11.4).

Cropping Spectra were cropped to reduce the influence of fluorescent background, scattering, buffers, and precipitant. Multiple intervals were evaluated to study the effect of individual contributions. The selected intervals for cropping of the Raman spectra were 800-1800, 1020-1800 and 1200-1500 cm^{-1} , which were motivated by the exclusion of residual baseline variance, the largest sulfate contribution and the majority of buffer interference.

Preprocessing pipeline The individual preprocessing operations are used in sequence as presented above. In some cases, a differing order may be viable, however the authors decided to keep the order fixed for this study in order to maintain comparability of the different approaches. The preprocessing pipeline configuration identified as optimal for the batch data was utilized for the fed-batch experiments.

Quantitative evaluation To compare preprocessing operations quantitatively without the need to calibrate a multivariate regression model, the metric signal-to-noise ratio (SNR) was used to evaluate the correlation between individual variables with the target quantity. Here, we define the SNR as

$$SNR = \frac{\text{Var}(\mathbf{x}_\lambda \hat{\boldsymbol{\beta}})}{\hat{\sigma}^2} \quad (5.3)$$

with $\hat{\boldsymbol{\beta}}$ being the regression coefficients of a univariate linear model of type $\mathbf{y} = \mathbf{x}_\lambda \boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with normally distributed errors $\boldsymbol{\varepsilon}$ and $\hat{\sigma}^2$ being the residual variance of linear model for wavenumber λ according to Soch and Allefeld (2018). To calculate SNR for the Raman spectroscopy data using Eq. 7.2, a linear model was built for each wavenumber individually. The SNR was used to compare multiple combinations of preprocessing operations. All combinations were applied to the precipitate and supernatant data sets. A preprocessing operation is considered beneficial if the metric at a certain variable increases compared to the un-preprocessed state. Further, the variable is considered predictive if it shows high SNR in precipitate and supernatant samples among all experiments.

5.2.2.3 Regression modelling

Data splitting To make maximum use of the recorded data, the batch data (B1-B8) were split using a nested cross-validation scheme. The eight experiments are divided into six training and two test experiments according to a *Leave-Two-Experiments-Out* cross-validation scheme which serves as the outer cross-validation. In total, this results in $8!/(2! * 6!) = 28$ independent test sets on which the calibrated models are evaluated. For the inner cross-validation, the six training experiments are further rotated using a *Leave-One-Experiment-Out* cross-validation scheme to

perform hyperparameter optimization in each of the iterations of the outer cross-validation. The fed-batch data were split in two training experiments (F1, F3) and one independent test experiment (F2). Here, the hyperparameters were optimized using a randomized split of the training set with eleven validation data points.

Regression models For the comparison of different combinations of preprocessing operations, two model types were evaluated as multivariate regressors for the quantification of the apparent precipitated VLP concentration, namely multiple linear regression (MLR) and PLS models as implemented in *scikit-learn* (v. 1.3.2). The different model types were selected to resolve potential differences between the working principles of these models with regard to preprocessing pipelines. For the MLR models, 6 wavenumbers of all observed wavenumbers were selected which were known to be protein-related but were not affected by AMS, namely 830, 850, 1241, 1314, 1341 and 1617 cm^{-1} (Spinner 2003; Maiti et al. 2004; Rygula et al. 2013). PLS models were trained using the NIPALS algorithm (S. Wold, M. Sjöström, and Eriksson 2001) and the number of latent variables was optimized using a grid-search in the range of 2 to 10 based on the inner cross-validation. Before being passed to the regression models, all spectral data were mean-centered and column-wise scaled to unit variance. PLS models for quantification of AMS were built based on a pipeline including turbidity correction, baseline correction, difference spectra and no further optimization was performed.

Error metrics Accuracy of the calibrated models was assessed using the root mean squared error (RMSE) and the coefficient of determination R^2 . For the inner cross-validation, the RMSE was calculated using the left-out experiments. When PLS models were used, the RMSE was scaled by the number of latent variables according to S. Wold, M. Sjöström, and Eriksson 2001. The importance of individual wavenumbers in PLS models was assessed quantitatively by VIP scores. The VIP score v_j (Mehmood et al. 2012) is defined as

$$v_j = \sqrt{N \sum_{a=1}^A \left[(q_a^2 t_a^T t_a) (w_{aj} / \|w_a\|)^2 \right] / \sum_{a=1}^A (q_a^2 t_a^T t_a)}, \quad (5.4)$$

for the individual wavenumber $j \in [1, N]$, where N denotes the total number of wavenumbers. The loading weights, the y-loadings, and the score vector corresponding to the PLS component $a \in [1, A]$ are given by w_a , q_a , and t_a , respectively.

5.3 Results and discussion

5.3.1 Spiking diversifies experimental data

Experiments in batch and fed-batch mode were conducted in 1 mL and 12 mL scale, respectively. To generate variance in precipitation data, multiple lysate batches and defined spiking solutions were used. In total, 8 batch experiments and 3 fed-batch experiments were conducted (see Tab. 5.1 and 5.2) and samples, taken at different concentrations of AMS, were analyzed by Raman spectroscopy and reference analytics. To extract the apparent precipitated VLP concentration from the reference UV absorbance measurements, the UV data were treated as described in 5.2.2.1. The UV pretreatment procedure is schematically depicted in Fig. 5.1 (A)-(D). Fig. 5.1 (A) shows the scatter-corrected UV data for experiment B1 over the precipitation range from 0 to 1.2 M AMS. To approximate the precipitation trend, a Boltzmann function which is shown in Fig. 5.1 (B) was fitted to the displayed data points. Based on the residuals between the Boltzmann fit and the observed data, outliers

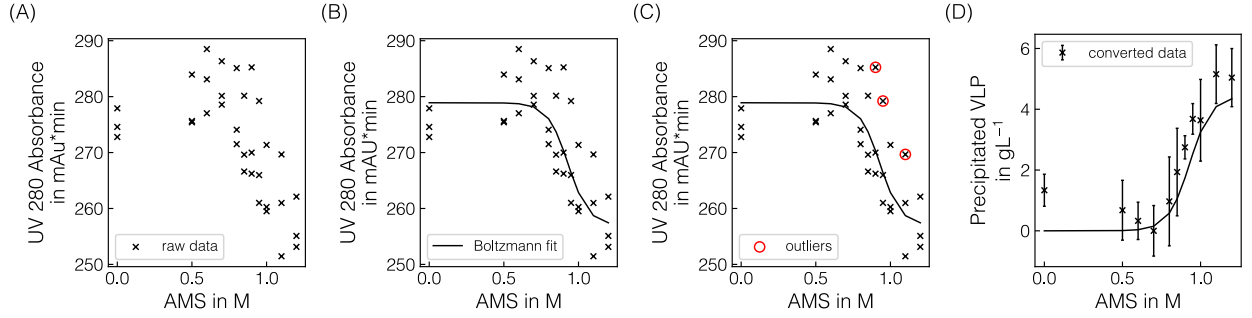


Figure 5.1 Illustration of pretreatment of the UV absorbance measurements. The UV absorbance at 280 nm is exemplarily shown over precipitant concentration for selected pretreatment steps for experiment B1. Raw triplicate UV measurements are shown as black crosses (A) and the fitted Boltzmann function is shown as a solid black line (B). The detected outliers are represented by red circles (C). The converted precipitated VLP concentrations are shown as black crosses with one standard deviation from averaging the three replicates per condition (D).

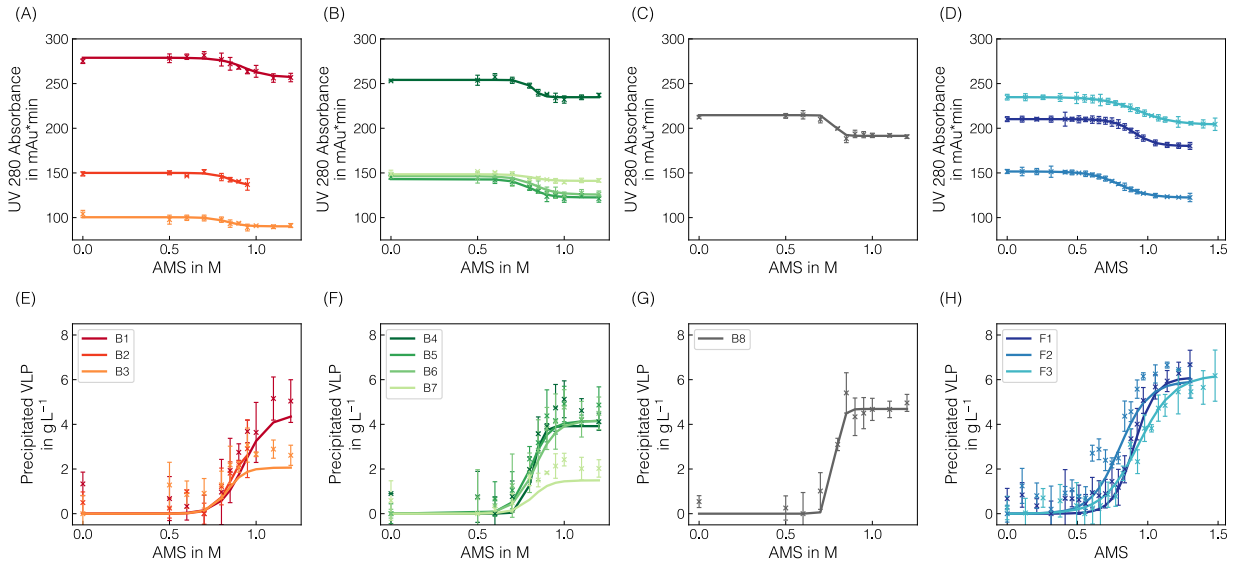


Figure 5.2 Comparison of pretreated UV absorbance data and VLP precipitation curves. Total UV absorbance at 280 nm after scatter- and outlier correction (A)-(D) and converted precipitation curves (E)-(H) are shown for batch experiments B1-B8 and fed-batch experiments F1-F3. Solid lines represent the fitted Boltzmann functions for outlier detection. The experiments are grouped according to the mode of spiking, where B1-B3, B4-B7 and B8 show the influence of lysate dilution, protein-composition regarding or salt content, respectively. For B2, the data points above 0.95 M were excluded due to defective Raman spectra in the precipitate samples.

were excluded which reduced the spread of observed data as visible in Fig. 5.1 (C). Finally, the UV data were converted to the apparent precipitated VLP concentration as presented in Fig 5.1 (D). The resulting outlier-corrected UV absorbance progression and the corresponding apparent VLP concentration are shown for all experiments in Fig. 5.2. For the batch experiments, three distinct groups are illustrated, as the experimental conditions were induced by different methodologies. For each experiment, the outlier-corrected absorbance measurements are displayed with one standard deviation and the fitted Boltzmann functions. While the Boltzmann functions aligned well with the UV absorbance for all experiments, the height of the final plateau for the precipitated VLP concentration was underestimated in the case of B3 and B7. Experiments B1 to B3 are shown in Fig. 5.2 (A) and (E) and were generated by diluting the lysate material using lysis buffer. For these curves over the course of the AMS concentration, the reductions of the total UV absorbance as well as the amount of precipitated VLP indicate the reduced concentration of all components by dilution. Experiments B1 and B4 used lysate material from different batches and showed comparable total UV absorbance and precipitated amounts of VLP. Compared to B4, experiments B5-B7 (cf. Fig. 5.2 (B) and (F)) were generated by specifically enriching the VLP or HCP content in the lysate materials by adding spiking material. This resulted in varying total UV absorbance and precipitated amounts of VLP for these individual curves. While a maximum of 2 gL^{-1} of precipitated VLP was reached in B7, the remaining experiments showed precipitated amounts of VLP up to 5 gL^{-1} . In contrast to B4, experiments B5-B7 varied in the range of 125 and 150 mAU total UV absorbance. In experiment B8 (cf. Fig. 5.2 (C) and (G)), the NaCl concentration was increased to 270 mM, which caused a lower UV absorbance compared to B4 which was based on the same lysate material and reached a maximum amount of precipitated VLP of over 4 gL^{-1} . The final, apparent concentrations of precipitated VLPs derived from the fitted Boltzmann functions are listed in Table 5.1.

In the fed-batch experiments (cf. Fig. 5.2 (D) and (H)), the differences in the precipitation trends were induced by dilution, adding VLP-enriched spiking material and modulating the feeding rate of the precipitant solution (cf. Table 5.2). Here, similar trends to the batch experiments were observed in terms of varying UV absorbance ranges. Note that the observed data include the dilution caused by the addition of the precipitant and hence a more gradual increase in the precipitated VLP concentration is expected compared to the batch experiments. Moreover, the AMS content calculation for F3 revealed a higher mean AMS flow rate than expected, exceeding the initially chosen final concentration of 1.3 M AMS.

In summary, the generated precipitation data provide insights into the precipitation dynamics in the presence of varying backgrounds induced by the addition of spiking material or buffer solution and serve as a diverse data set for chemometric model development.

5.3.2 Raman spectra are affected by precipitant and precipitates

Raman spectra are commonly affected by the matrix, buffers and other components present in the sample. Before using the collected Raman data for chemometric modeling, the data require preprocessing to properly evaluate potential interferences from buffers or excipients. A schematic illustration of the effect of the employed preprocessing operations is presented in Fig. 5.3. From left to right, the schematic pipeline shows the averaged raw Raman spectra for each sample in experiments B1, the turbidity- and baseline-corrected spectra using the 3299 cm^{-1} -normalization and asPLS Whittaker filter, the SS-background-corrected spectra and the difference spectra. The combination of turbidity correction with the Whittaker filter removed turbidity effects and baseline drifts reliably over the whole spectral range (cf. Fig. 5.3 (A), (B)). The SS-background correction subtracts the contribution of the AMS, which caused the AMS-related band at 980 cm^{-1} to become

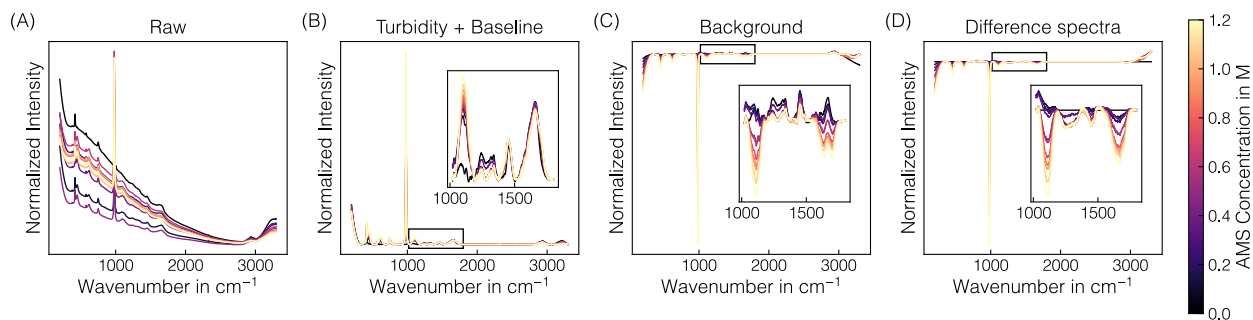


Figure 5.3 Illustration of the preprocessing of the Raman spectra. The preprocessing procedure is exemplarily depicted for experiment B1. The normalized raw Raman spectra (A) and selected preprocessing steps (B-D) are shown. The spectra are colored according to the AMS concentration with brighter colors denoting higher concentrations.

negative and the protein-related bands to show larger variations (cf. Fig. 5.3 (C)). Finally, the difference spectra operation sets the Raman intensity in the beginning of the experiments to 0 and caused negative bands in the wavenumber region $1200\text{--}1700\text{ cm}^{-1}$ (cf. Fig. 5.3 (D)). For a visual interpretation of the difference spectra, the reader is advised to regard positive and negative bands as substances being added and being removed from the solution, respectively. For evaluating the interferences of the precipitant, buffers or excipients, the Raman spectra were firstly treated by turbidity- and baseline correction only.

Fig. 5.4 presents a comparison of the collected Raman spectra in precipitant-containing buffer samples, supernatant samples, precipitate samples from batch experiment B1 without and with turbidity correction (cf. Sec. 6.2.4) from top to bottom. Fig. 5.4 (A) shows the raw Raman spectra and Fig. 5.4 (B)-(D) show the baseline-corrected data within selected wavenumber regions. In the raw spectra (cf. Fig. 5.4 (A)), different baseline effects are visible between all sets of spectra with the precipitate data showing the strongest baseline variations. Baseline variations in the precipitate data were less pronounced for the turbidity-corrected spectra. At low concentrations of AMS, the baseline increased while at high AMS concentrations, the baseline decreased again. This effect was accompanied by the increasing contributions of AMS in the Raman spectra with the most prominent Raman band being located at 980 cm^{-1} . Additionally, the increasing AMS contents are visible in the Raman bands at 450 , 618 and 1106 cm^{-1} (cf. Fig. 5.4 (B) and (C)) as well as near 1435 and 1693 cm^{-1} (cf. Fig. 5.4 (D)), which can be attributed to sulphate and ammonium ions, respectively (Spinner 2003; Fontana, Ben Mabrouk, and Kauffmann 2013). Additional components of the lysis buffer which were contained in all samples can be traced to the bands at 1249 and 1470 cm^{-1} indicative for Tris and EDTA (Socrates 2004). These bands remained roughly constant over the course of a precipitation experiment as only the AMS concentration was gradually increasing. As expected, the sapphire bands at 379 , 418 , 430 , 450 , 577 and 750 cm^{-1} (Watson, Daniels, and C. S. Wang 1981), the band of molecular oxygen at 1556 cm^{-1} (A. Weber et al. 1967) and the broad water band at 1650 cm^{-1} (Spinner 2003) remained roughly constant as well. Protein-related contributions are located between $600\text{--}880\text{ cm}^{-1}$ and $1200\text{--}1800\text{ cm}^{-1}$ with amide bands, CH_2 -deformation bands and bands of aromatic amino residues (Maiti et al. 2004; Rygula et al. 2013). Amide III and I bands are visible near 1241 and 1660 cm^{-1} , respectively. Tyrosine (Tyr) bands appear at 830 , 850 , 1205 and 1617 cm^{-1} and the bands originating from tryptophan (Trp) and phenylalanine (Phe) appear at 1340 and 1605 cm^{-1} , respectively. CH_2 -deformations are visible at 1314 , 1340 and 1448 cm^{-1} . While these

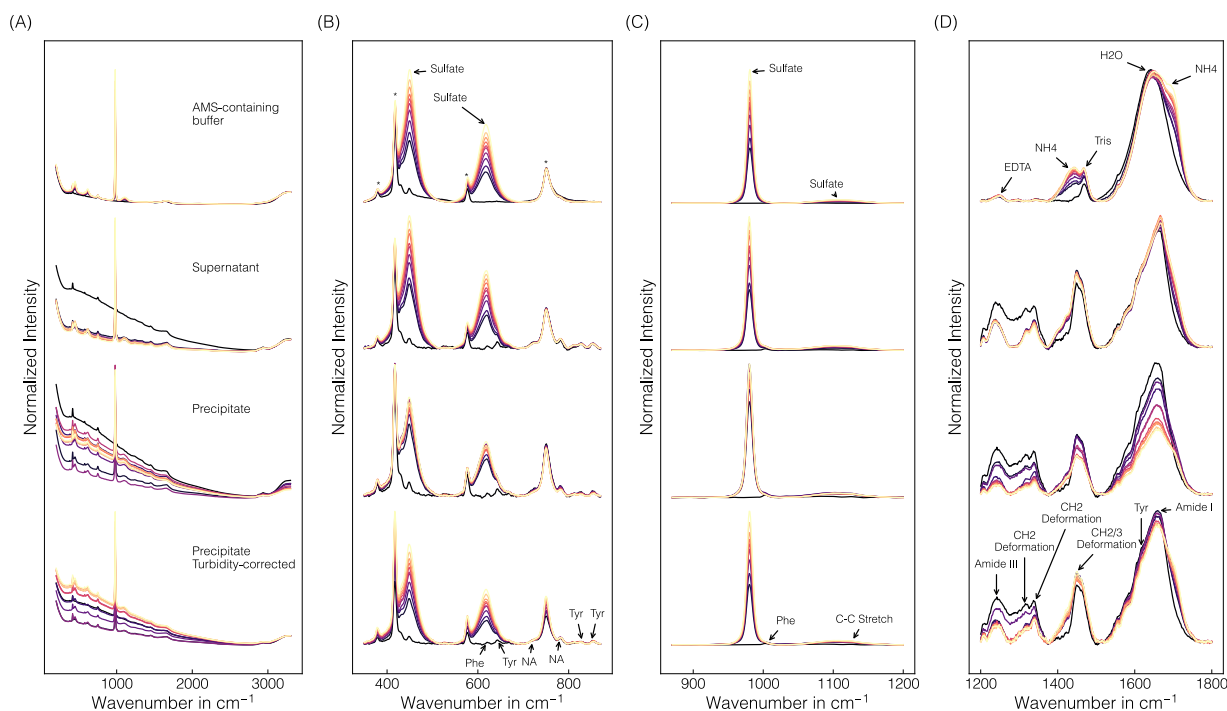


Figure 5.4 Comparison of collected Raman spectra in precipitant-containing buffer, supernatant and precipitate samples. Precipitate and supernatant data are exemplarily shown for batch experiment B1. Full raw spectra **(A)**, and baseline-corrected spectra of selected wavenumber regions **(B)-(D)** are illustrated. Spectra with incorporated turbidity correction are shown in the bottom row for the precipitate samples. The spectra are normalized in each row and subfigure for visual purposes and colored according to the AMS concentration with brighter colors representing higher concentrations. Wavenumber regions indicated by arrows correspond to buffer-, protein- and nucleic acid (NA)-related contributions. Sapphire bands are marked with an asterisk.

protein-related contributions were largely unaffected by the AMS or lysis buffer induced Raman activity, the precipitant interfered with the protein-related contributions in several locations (cf. 620, 643, 1004, and 1127 cm^{-1}). Tyr, Phe, and C-C stretching can be associated with these bands at 643 cm^{-1} , 620 and 1004 cm^{-1} , and 1127 cm^{-1} , respectively (Peticolas 1995; Maiti et al. 2004; Rygula et al. 2013). In addition to the protein-related contributions, the bands at 724 and 781 cm^{-1} are indicative for the presence of nucleic acids (Peticolas 1995).

To study the effect of the precipitates in solution formed during the precipitation experiment, the baseline-corrected Raman spectra from precipitate and supernatant samples were compared. In general, similar bands are visible in both the precipitate and the supernatant spectra. The protein bands showed a decreasing trend with increasing AMS concentrations. This effect was more pronounced in the precipitate spectra compared to the supernatant spectra indicating an overall protein band-specific intensity decrease caused by the precipitates in solution. Similarly, the nucleic acid bands at 724 and 781 cm^{-1} were visibly decreasing with increasing concentrations of AMS. By introducing the turbidity correction, the overall intensity decrease was reduced in all mentioned wavenumber regions, effectively making the precipitate spectra more comparable to the

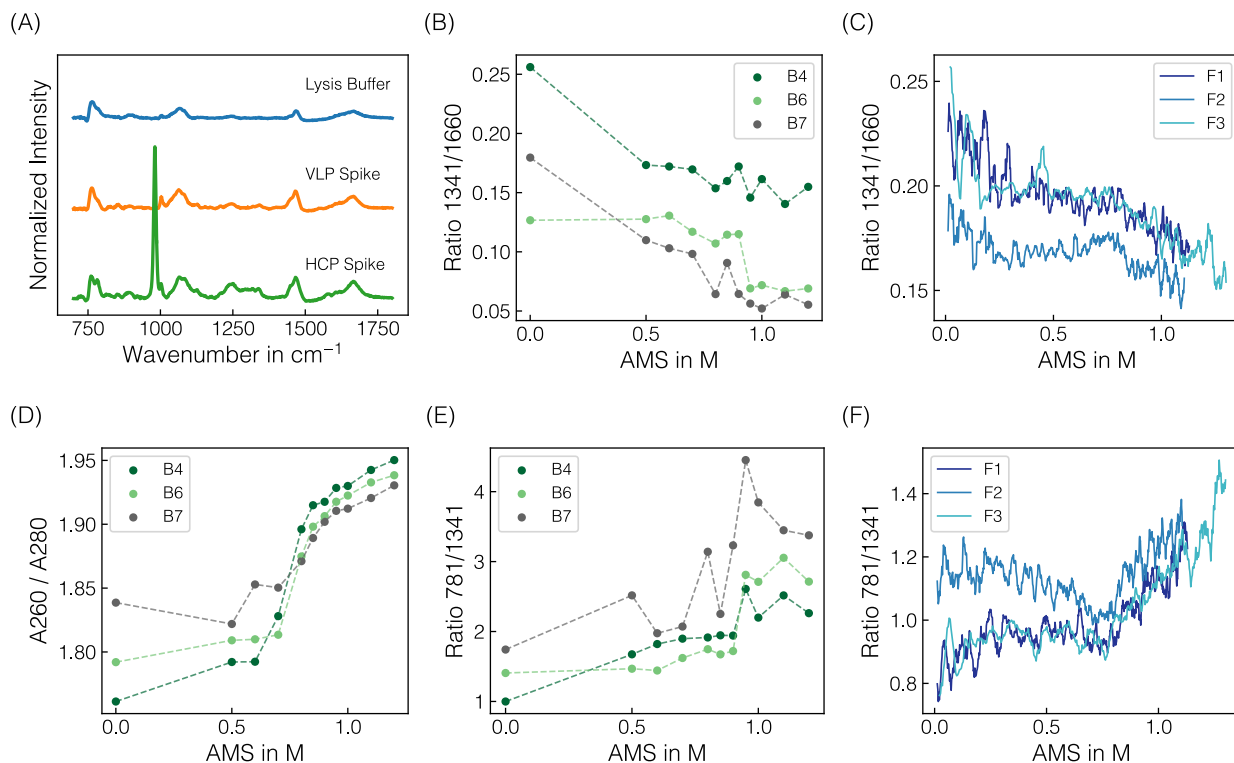


Figure 5.5 Raman spectra of the spiking solutions and comparison of selected wavenumber ratios. Normalized Raman spectra of lysis buffer and VLP- and HCP-enriched spiking materials after baseline-correction and subtraction of water blank are illustrated in (A). The protein-band ratio 1341/1660 is depicted over the course of AMS for batch experiments B4, B6, B7 differing in the protein composition of the lysate batch (B) and all fed-batch experiments (C). The UV absorbance-derived A260/A280 ratio (D) is shown alongside the band ratio 781/1341 comparing the nucleic acid-associated Raman band with the protein-associated Raman band for batch experiments B4, B6, B7 (E) and fed-batch experiments (F). All ratios were calculated with turbidity- and baseline-corrected Raman spectral data. The ratios for fed-batch experiments were calculated using a moving average of 10 spectra to smooth the signal. Dashed lines are only shown for visual purposes.

supernatant spectra, especially in the bands of the nucleic acids. However, individual Raman bands were adversely effected such as the sapphire band at 750 cm⁻¹, showing minor overcorrection.

5.3.3 Raman data reveal structural differences between species

To investigate whether structural differences between the target molecules and the other species in the lysates are observable via Raman spectroscopy, the Raman spectra from the VLP- and HCP-enriched spiking materials were compared. Normalized Raman spectra of the lysis buffer and the purified VLP- and the HCP-enriched spiking solutions are presented in Fig. 5.5 (A) after baseline-correction and water blank subtraction. Both solutions exhibited distinct Raman bands in the protein fingerprint region between 800 and 1800 cm⁻¹ with slight differences in the shape of the most prominent peaks between 1000-1100 cm⁻¹, 1200-1300 cm⁻¹ and 1600-1700 cm⁻¹. Additionally, the nucleic acid-related Raman band at 781 cm⁻¹ and the precipitant-related Raman band at 980 cm⁻¹ were observed for the HCP spiking solution. As the HCP spiking solution was prepared

by recovering the precipitation supernatant and subsequent dialysis, it contained both HCPs and host cell nucleic acids, which was further underlined by a A260/A280 ratio of 1.88. The AMS contribution was caused by the residual amounts of AMS after buffer exchange. As the peak profiles of the two spiking solutions in the protein fingerprint region differed slightly from one another and the amide bands are commonly used Raman markers for higher-order structures (Socrates 2004), the ratio of selected amide bands are presented in Fig. 5.5 (B)-(C) for selected batch and fed-batch experiments, respectively. The data indicated a shift in the ratio of 1341 and 1660 cm^{-1} over the course of a precipitation experiment depending on the initial conditions of the respective experiment. Exemplarily, the VLP-spike experiments (B6, B7, F2) exhibited slightly lower ratios than the experiments with similar lysate compositions (B4, F1, F3). To compare the intensity decrease observed for the nucleic acid and protein-related bands, the ratio of 781 and 1341 cm^{-1} is shown for batch and fed-batch experiments over increasing AMS content in Fig. 5.5 (E)-(F). The curves show a sigmoidal increase indicating a stronger decrease of protein-related Raman bands than observed for the bands associated with nucleic acids. A comparable trend was observed in the A260/A280 ratio for supernatant samples of the batch experiments as presented in Fig. 5.5 (D). In summary, the differences in protein-associated Raman bands of the two spiking solutions underline structural differences between VLPs and HCPs. Selected band ratios could be used to track these spectral changes over the course of the precipitation, suggesting the selective precipitation of specific molecular species, namely the VLP.

5.3.4 Background correction removes interferences

The batch experiment data were used to evaluate the suitability of preprocessing pipelines with regard to their universal applicability to all experiments, and their capacity for removing background effects such as buffer and precipitant interferences. Fig. 5.6 (A)-(C) present the difference spectra after turbidity and baseline correction, incorporated SS-background correction or incorporated OPLS-background correction, respectively, for the precipitate samples of all batch experiments B1-B8. In Fig 5.6 (D)-(F), the SNR is presented to quantify the effect of the baseline and background correction methods on the correlation with the precipitated amount of VLP. The turbidity- and baseline-corrected spectra showed evolving negative bands between 1200 and 1400 cm^{-1} and 1500 and 1700 cm^{-1} . These negative bands exhibited SNR values below 0.5. The Raman bands between 1020 and 1200 cm^{-1} and between 1700 and 1720 cm^{-1} showed a clear trend with increasing levels of AMS. The SNR confirmed this observation with the highest values close to 1 for these intervals. Similar trends was observed between 1400 and 1500 cm^{-1} with SNR values up to 0.9. By subtracting the reference spectra comprising solely buffers and precipitants, an additional negative band between 1020 and 1200 cm^{-1} was unveiled and the peak around 1700 cm^{-1} was removed (cf. Fig. 5.6 (A),(B)), while the wavenumber region between 1200 and 1400 cm^{-1} remained largely unaffected. The SNR confirmed these observations and showed a strong correlation of the bands between 1020 and 1200 cm^{-1} and 1550 and 1750 cm^{-1} , where protein-related Raman bands are located. Bands between 1400 and 1500 cm^{-1} also showed elevated correlation according to SNR compared to sole baseline correction. By incorporating an OPLS-based background correction, the spectra mostly retained positive bands between 1020 and 1200 cm^{-1} and the peaks around 1450 and 1700 cm^{-1} (cf. Fig. 5.6 (C)). However, the SNR revealed that the correlation increase in the range from 1020 to 1200 cm^{-1} and from 1550 to 1750 cm^{-1} is lower than for SS-background correction. Almost no correlation increase was observed near 1600 cm^{-1} . Despite the visual differences between the SS- and OPLS-background correction, the SNR assessment showed a similar profile with correlations between 1020 and 1200 cm^{-1} , near 1450 cm^{-1} and near 1700 cm^{-1} . An analogous illustration of difference spectra

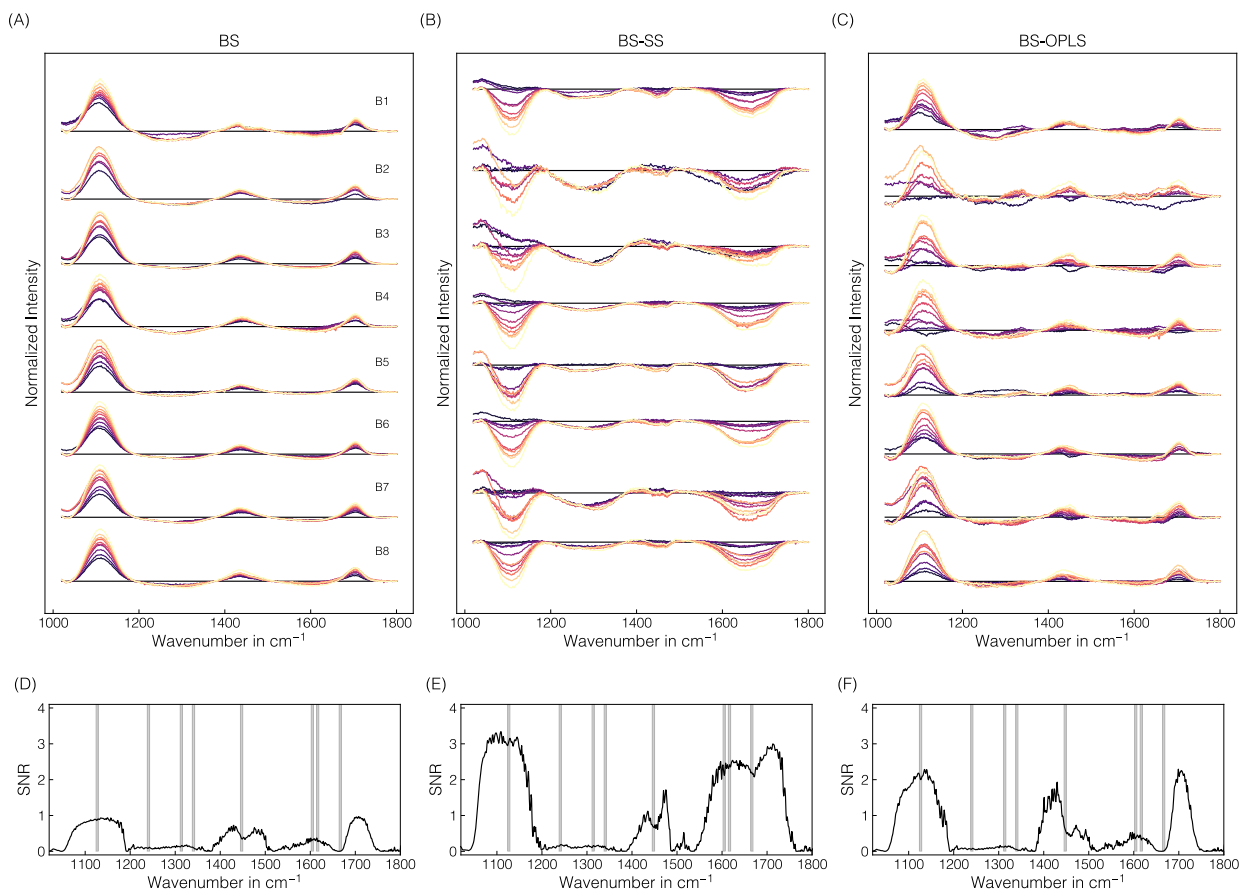


Figure 5.6 Comparison of the effects of preprocessing operations on Raman spectra in precipitate samples for the wavenumber region 1020-1800 cm^{-1} . In (A)-(C), difference spectra for all batch experiments B1-B8 are shown after turbidity and baseline correction, with incorporated SS-background correction, or with incorporated OPLS-background correction, respectively. The spectra are colored according to the AMS concentration with brighter colors denoting higher concentrations. In (D)-(F), the SNR for the respective preprocessing operations are shown. Gray shaded areas indicate protein-related Raman regions according to literature (Maiti et al. 2004; Rygula et al. 2013).

and corresponding SNR for the supernatant samples can be found in the Supplementary (Fig. C.1), where comparable but substantially less pronounced trends were observed.

5.3.5 Systematic pipeline optimization improves model accuracy

Upon retrieving information about correlative structures in the collected Raman spectra, the data were used to build multivariate regression models taking the Raman spectra as inputs and predicting the apparent precipitated VLP concentration. Multiple combinations of preprocessing operations and MLR or PLS models were screened with the aim to find the optimal pipeline configuration with respect to model accuracy and robustness. While a strong focus lied on evaluating previously assessed operations such as SS and OPLS-based background correction, their combinations with cropping and derivative filters were also evaluated in this Section. In Fig. 5.7 (A), the distributions of RMSE of the held-out test sets are shown for all tested model and pipeline configurations, sorted by the mean of the RMSEs. Fig. 5.7 (B)-(G) additionally show the direct comparison of selected model configurations and aim to emphasize the effect of individual changes to the model and preprocessing pipeline on model performance. While the best models achieved RMSE values of approx. 0.8 g L^{-1} on average, the errors increased to up to approx. 2.0 g L^{-1} on average. Model configurations with higher average errors also showed larger variance. Among the 4 best performing model configurations (cf. Fig. 5.7 (A)), all pipelines included SS or OPLS-background correction, second order SGFs and PLS regression models. While the best and the model ranking 4th used cropping to $800\text{-}1800 \text{ cm}^{-1}$, the remaining used the full wavenumber range. The PLS model applied to turbidity- and baseline-corrected difference spectra achieved better performance than MLR (cf. Fig. 5.7 (B)). In general, PLS models benefited from using background correction method with the lowest average RMSE and variance being achieved when applying the OPLS method (cf. Fig. 5.7 (C)). For both SS- and OPLS-corrected data, increasing the order of derivative from 0 to 2 marginally improved the median RMSE and reduced variance (cf. Fig. 5.7 (D)). Cropping the Raman spectra generally showed adverse effects on model accuracy using OPLS-corrected data (cf. Fig. 5.7 (E)). As pointed out above, this was not true for all pipeline configurations but may be pointed out as a general tendency in this case study. To validate that the identified pipelines can not only be applied to Raman spectra recorded from the precipitate-containing samples, but also to the particulate-free supernatant samples, separate models were calibrated on the supernatant Raman spectra as shown in Fig. 5.7 (F). While the PLS and MLR models achieved similar accuracy after turbidity and baseline correction in precipitate and supernatant samples, OPLS-background-corrected models showed considerably increased RMSE on average in the supernatant samples. In Fig. 5.7 (G), separate models were calibrated without turbidity correction to verify that model performance benefit from the incorporated turbidity correction. Although MLR and PLS models achieved comparable performance when applied to solely baseline-corrected difference spectra, MLR performed worse than PLS models when combined with turbidity correction. All tested model and pipeline configurations without turbidity correction can be found in the Supplementary (Fig. C.2).

5.3.6 Model pipeline captures precipitation trends

The previously identified best model and preprocessing pipeline PLS(BS-OPLS-800/1800-SG2) was finally evaluated on a representative split of batch data and further transferred to the fed-batch data. Fig. 5.8 presents PLS model predictions for the batch test set (B4, B6) and fed-batch experiments F1-F3 as well as VIP scores for both calibrated models. Despite having different VLP to impurity ratios in the starting material (cf. Tab. 5.1), the PLS model predictions for batch experiments B4 and B6 aligned well with the experimental data (cf. Fig. 5.8 (A), (B)) with an R^2 of 0.83.

CHAPTER 5. RAMAN-BASED PAT FOR VLP PRECIPITATION

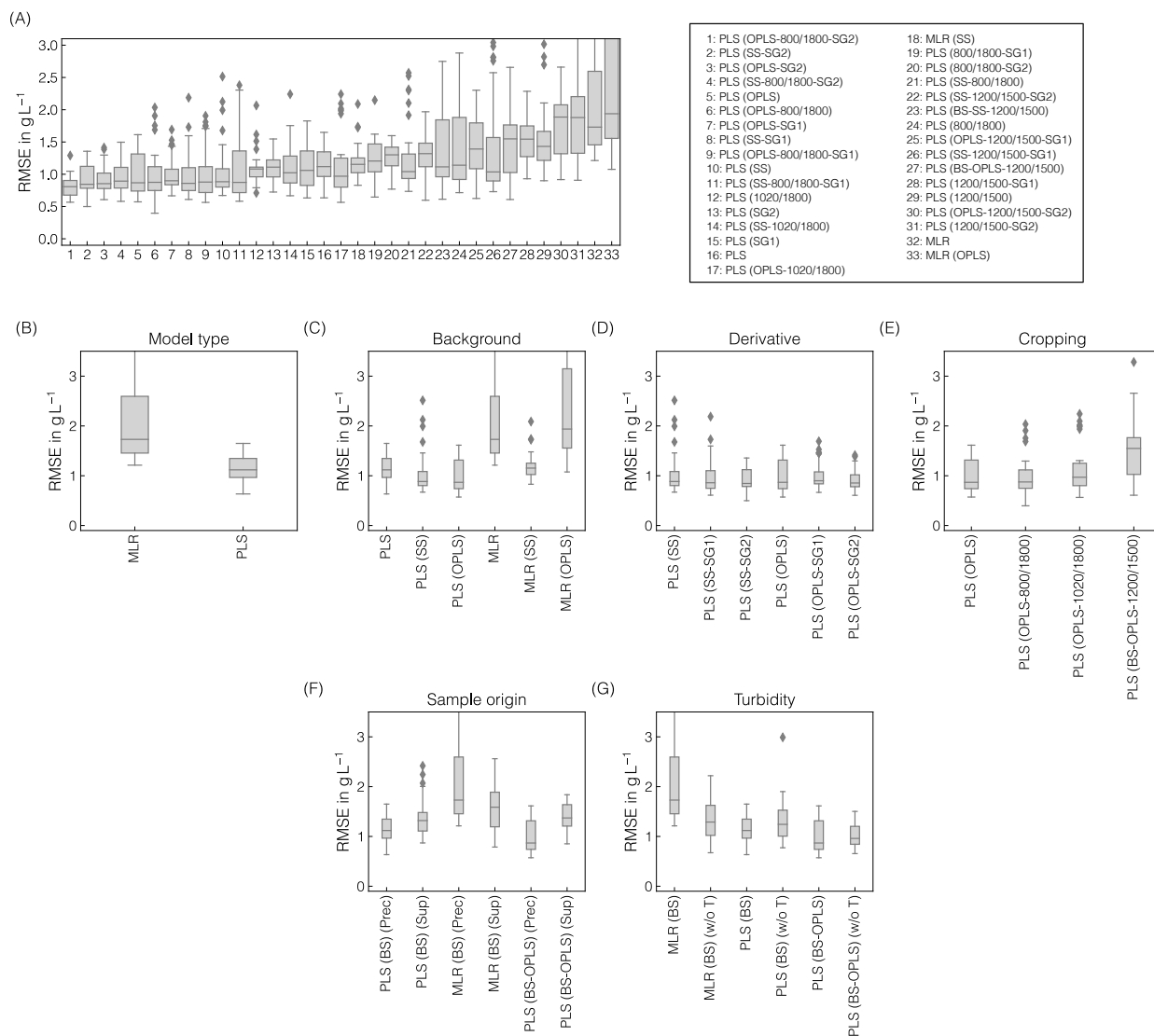


Figure 5.7 Comparison of preprocessing pipelines and model types. The distributions of RMSE of the outer cross-validation are presented and ranked by the mean of the RMSEs for all tested model configurations (A). All tested model configurations comprised turbidity correction, baseline correction, and difference spectra. The solid lines within the boxes represent the median of the obtained performances. Outliers were characterized by errors surpassing 1.5 times the interquartile range and are symbolized by diamonds. RMSEs of selected model configurations are illustrated separately, namely for comparison of model types (B), background correction methods (C), SGF derivative (D), spectral cropping (E), sample origin (F), and models with and without incorporated turbidity correction (G).

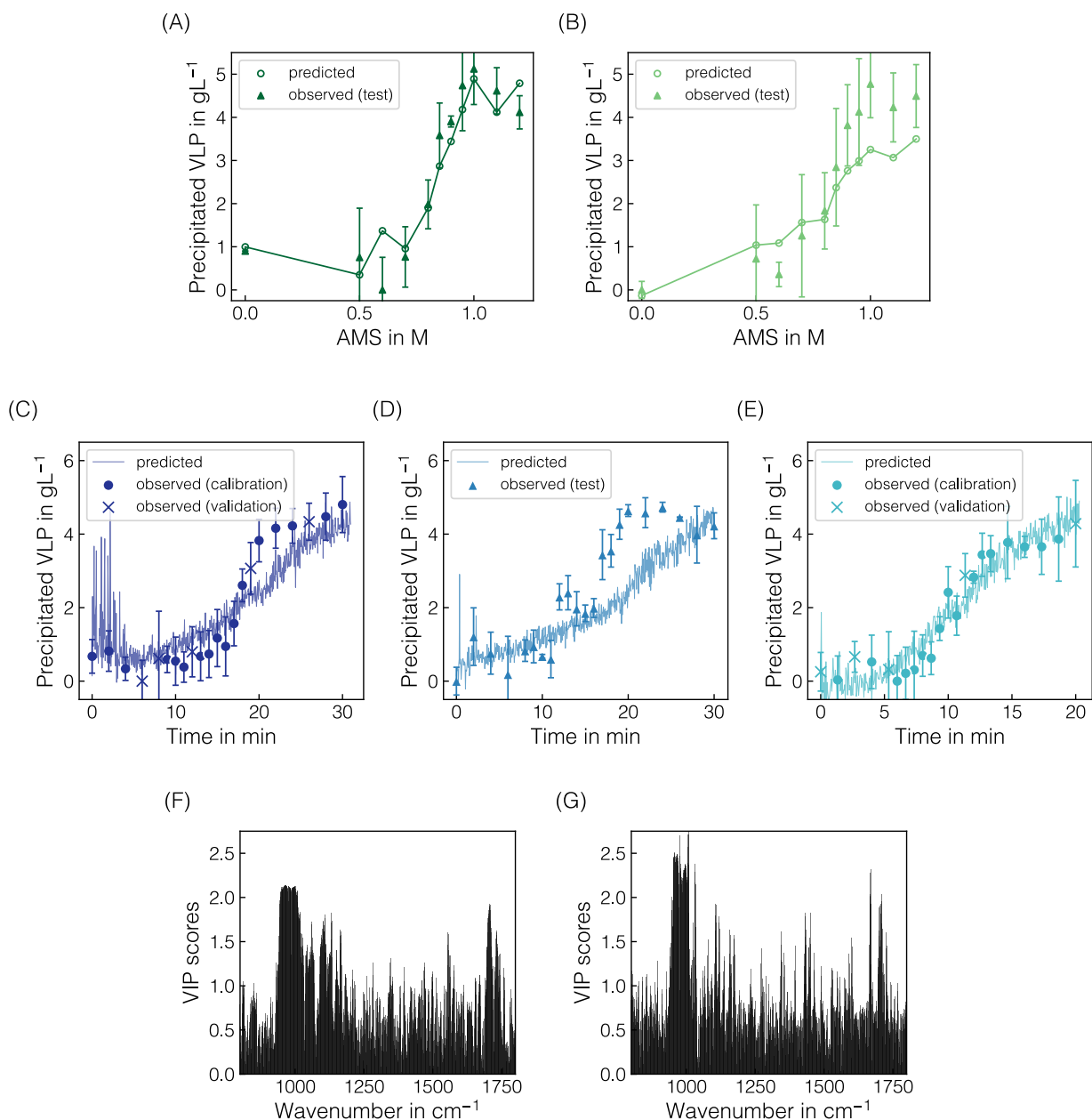


Figure 5.8 PLS model predictions with regard to precipitated VLP for batch and fed-batch experiments and the corresponding VIP scores for both models. For the batch experiments B4 (A) and B6 (B), the predicted and observed precipitated VLP concentrations are shown over increasing AMS concentrations. For visual purposes, solid lines are shown to linearly connect the PLS model predictions. For the fed-batch experiments F1-F3 (C)-(E), the observed apparent VLP concentrations are shown as symbols with their respective assignment to calibration, validation and test data. Model predictions from the real-time data after baseline correction and difference spectra are presented as solid lines. The VIP scores are illustrated for both batch (F) and fed-batch (G) models in the wavenumber region 800-1800 cm^{-1} .

Error metrics for calibration, cross-validation and test sets are presented in the Supplementary in Fig. C.3. Alternatively, PLS model predictions for all calibration runs are shown over AMS content in the Supplementary in Fig. C.4. While for B4, the sigmoidal shape of the precipitation curve was represented in the model predictions, the model predicted an early onset of the VLP precipitation for B6 and saturate at lower final concentration of approx. 3.5 g L^{-1} , effectively smoothing out the sigmoidal shape. For the calibration experiments (cf. Fig. C.4), the sigmoidal shape and onset of the precipitation were predicted accurately while predictions for the initial plateau differed from the observations. Most of the predictions, however, lay within the range of one standard deviation of the observed data points.

In Fig. 5.8 (C)-(E), the predictions of the calibrated PLS model are shown for fed-batch experiments F1, F2 and F3 as symbols with their respective assignment to calibration, validation and test sets. Fig. C.5 in the Supplementary presents error metrics for all data sets. The real-time data used for prediction were not averaged and hence the model predictions were scattered around the observed data points and were available in between reference measurements. While the trajectories of the expected precipitation curves were well represented in the predictions for the training experiments F1 and F3, the trajectory for F2 was underestimated in the range from 15 to 25 min which corresponded to approx. 0.5 to 1 M AMS. Using the averaged Raman spectra, this yielded an R^2 of 0.64 (cf. Fig. C.5). For F1, the predictions for the spectra during the first 3 minutes strongly scattered and deviated from the expected values. Similar spikes for individual spectra were observed in the start-up phase for F2 and F3. In all cases, the effect reduced over the first 1 to 3 min after which constant scatter can be observed for all experiments until the end of the precipitation process. To evaluate the origin of these scatter, a step-wise visualization of the Raman spectra for all fed-batch experiments is presented in Fig. C.8 in the Supplementary. The affected Raman spectra show slight distortions to lower intensity in the protein-related and water Raman bands, which are subsequently enhanced through the sequence of preprocessing operations. Towards the end of the process, no plateauing of the predictions based on the Raman data was observed whereas the observed data signals tended to stabilize with individual spikes being present.

Finally, Fig. 5.8 (F), (G) present the VIP-based feature importance for the PLS models based on batch and fed-batch data, respectively, in the wavenumber range between 800 to 1800 cm^{-1} . Both models relied on multiple wavenumber regions as indicated by elevated VIP scores. Most prominently, the interval between 950 and 1000 cm^{-1} originating from the sulfate ions contributed to the model predictions. Additionally, local maxima appeared at wavenumber intervals of 800-900, 1050-1150, 1400-1500 and 1550-1650 cm^{-1} for both the batch and the fed-batch models, mainly encoding protein-related information.

PLS models were further used to quantify the AMS concentration in batch and fed-batch using the same methodology as laid out for the precipitated VLP concentration. In the case of AMS the preprocessing only included turbidity and baseline correction and the formation of difference spectra. The predictions for the batch and fed-batch data are presented in the Supplementary in Fig. C.6 and Fig. C.7, respectively. For both data sets, the models showed almost perfect alignment for the test sets with R^2 of 0.98 and 0.97 over the concentration range from 0 to 1.2 M. For the fed-batch runs deviations from the expected values were observed above AMS concentrations of 1.2 M.

5.4 Discussion

5.4.1 Effects of data diversification

Data diversification is crucial for generating representative data sets for developing chemometric sensors suitable for PAT in biomanufacturing. The precipitation experiments were designed to mimic multiple sources of experimental variance which may arise in process development studies or manufacturing campaigns. Data were diversified using different lysate batches and conditioning of the lysates by dilution, salt addition, or spiking with purified VLP or HCP solutions. Similar strategies for data diversification were employed in the literature by spiking purified monoclonal antibodies with mock solutions to mimic complex feedstocks (Großhans, Suhm, and Hubbuch 2019), spiking with nutrients during cell culture (R. M. Santos et al. 2018), spiking with cells during inoculum maturation (Guardalini, Silva Cavalcante, et al. 2023) or mixing available samples (Jiarui Wang et al. 2023b). In general, diversification strategies were reported to improve model robustness and interpretability (R. M. Santos et al. 2018) or enlarge the experimental data set (Jiarui Wang et al. 2023b). Alternative strategies may include synthetic enlargement of the available data sets by data augmentation (Schiemer, Rüdte, and Hubbuch 2024) or the collection of data from multiple products, formulation components and sensors (Wei et al. 2022) for the diversification of experimental data sets.

In essence, the diversification strategy in this study yielded versatile precipitation experiments covering different precipitation trajectories, varying levels of total concentration of precipitated VLP and impurity content. However, the UV absorbance data exhibited high variance for the triplicate measurements which was enhanced by the conversion to the apparent precipitated VLP concentration, resulting in difficulties in determining the precipitation trajectories. Particularly, samples in the plateau region of the precipitation curves showed unexpected fluctuations. The variance between replicates or individual data points within one experiments, i.e. variable onsets of the precipitation or the initial and final plateau concentration, may be attributed to the precipitation procedure, sampling procedure and sample handling. Since the batch experiments involved the manual addition of the precipitant as well as sampling of the precipitate-containing solution, the heterogeneity of the precipitate solution may introduce analytical variance between data points. Moreover, inadequate mixing during precipitant addition or during the subsequent incubation time could also contribute to variance within a batch experiment. Variance between replicates and individual data points were less pronounced for fed-batch than batch experiments, which can be attributed to the inherent continuity of the fed-batch approach as well as effective mixing in the reservoir, thus, supporting uniform precipitate formation while avoiding co-precipitation of other species (Pons Royo et al. 2022).

Eventually, the observed variances led to the employment of a pretreatment strategy based on experiment-based curve fitting using the Boltzmann function. This strategy assumed the selective precipitation of VLPs which has been shown for AMS concentrations below approx. 1.3 M (Wegner, Zimmermann, and Hubbuch 2022). The pretreatment strategy was shown to effectively reduce single-point variation by removing outliers and resolving differences between the precipitation experiments. Treatment of the reference data is uncommon in chemometrics but certainly improved the quality of the data in this study and enabled more robust model calibration (data not shown). In summary, it is crucial to manage data variance within individual experiments, e.g. by increasing the number of replicates or incorporating a physically-motivated pretreatment, while highly diversifying

data across experiments, as the obtained batch and fed-batch data sets served as basis for the development of a chemometric modeling pipeline for the quantification of precipitated VLPs.

5.4.2 Effects of preprocessing operations on Raman data

The Raman spectra are affected by the composition of the lysate, the addition of the precipitant and the formation of precipitates in solution causing multiple independent changes in the Raman spectra to occur simultaneously. While the addition of the precipitant induced the emergence of the sulfate and ammonium-associated Raman bands, the formation of precipitates reduced the overall intensity due to turbidity of the solution. Turbidity is commonly observed in particulate-containing separation processes such as crystallization (Moreno et al. 2000; Grön et al. 2003; Harner et al. 2009; Y. Liu, Pietzsch, and Ulrich 2014) or precipitation (Zelger et al. 2016; Meyer-Kirschner et al. 2016). While Zelger et al. (2016) directly used the turbidity measurements for monitoring the precipitation progress, Meyer-Kirschner et al. (2016) demonstrated that a simple correction using the OH-stretching Raman band is feasible. Since regulatory agencies (FDA 2015) recommend rigorous qualification of PAT sensors, the authors consider the use of Raman spectroscopy in combination with turbidity correction a viable PAT sensor for monitoring protein precipitation processes. Here, normalization at the OH-stretching Raman band reliably compensated for turbidity in the raw spectra. However, it was also pointed out that the turbidity correction adversely affects other wavenumber regions in the herein recorded, baseline-corrected spectra such as the sapphire band at 750 cm^{-1} . This may be connected to the non-linear correlation between turbidity and Raman intensity decrease (Sinfield and Monwuba 2014).

After turbidity and baseline correction, the variation in the protein-associated Raman bands was comparable to what has been observed for the supernatant spectra while the inexplicable decrease in nucleic acid-associated bands was eliminated. This supports the assumption of selectivity of the precipitation process for VLPs. A direct quantification of the amount of HCPs was not possible due to the unavailability of quantitative reference measurements in this study. The spectral data of the HCP- and VLP-enriched solutions, however, suggest that the VLPs are structurally different from the HCPs contained in the clarified cell lysate which is supported by different amide band ratios at the start and the end of a precipitation experiment.

While this study was performed using HBcAg VLPs, the applicability of Raman spectroscopy for HBcAg variants or other types of VLPs is expected. To apply the presented workflow to other VLP systems, the generation of new experimental data is strictly necessary. Furthermore, by incorporating additional reference analytics such as HCP-enzyme-linked immunosorbent assay (ELISA) as done in Großhans, Suhm, and Hubbuch 2019, the prediction of HCP content via Raman spectroscopic measurements may hence be facilitated. However, the prediction of HCPs using chemometric methods has only rarely been reported (Capito, Skudas, Kolmar, et al. 2015; J. Chen et al. 2024), most likely due to the diversity of species classified as HCPs (Falkenberg et al. 2019). The same applies to nucleic acids and endotoxins, which might affect VLP precipitation, but their analysis exceed the scope of this study.

To prepare spectral data for multivariate modeling, turbidity, baseline and background correction followed by the difference spectra operation were evaluated for their correlation with the amount of precipitated VLP and their capacity to eliminate interferences in the Raman spectra caused by the addition of precipitant. Baseline correction of turbidity-corrected spectra proved to remove baseline drifts reliably across the entire wavenumber range. As the applied asPLS combines baseline estimation with smoothing of the estimated function while preserving the underlying features of

interest (F. Zhang et al. 2020), baseline drifts caused by both sample compounds and instrumental noise can be addressed.

The SNR-based correlation analysis of the baseline-corrected spectra suggested that next to protein-associated Raman bands, wavenumber regions encoding buffer and precipitant-associated information were also correlated with the amount of precipitated VLP. As the increase in precipitated VLP is expected to follow a sigmoidal curve (Wegner, Zimmermann, and Hubbuch 2022; Valentic, J. Müller, and Hubbuch 2022), the precipitant concentration in the solution and the amount of precipitated VLP is not linearly correlated. Hence, the contributions of the precipitant should be removed from Raman spectra via background correction. The SS-background correction reliably removed the background signal from the Raman spectra, while overcorrections were visible for samples with AMS concentrations above 0.6 M (cf. Fig. 5.3). This is caused by the discrepancy with regards to the turbidity between the precipitate spectra and the reference spectra where the solution remains clear even at the highest AMS concentrations of 1.2 M. The OPLS method as proposed by Trygg and S. Wold 2002 demonstrated improved correlations by removing the non-correlated systematic variation in the Raman spectra without the need for reference measurements. However, also the OPLS-treated spectra showed residual contributions in the dominant 980 cm^{-1} sulfate band and hence could not fully remove the precipitant-associated effects. As the OPLS method extracts the non-correlated variation with respect to the target variable (Trygg and S. Wold 2002), it is likely that the AMS-associated information was partially contained due to its correlation with the amount of precipitated VLP as mentioned before. Alternatively background variations may be modeled using indirect hard modeling (IHM) (Alsmeyer, Koß, and Marquardt 2004; Kriesten et al. 2008; Meyer-Kirschner et al. 2016). Furthermore, the extended multiplicative signal correction (EMSC) algorithm supports the removal of an interferent spectrum (Martens and Stark 1991a). However, Liland, Achim Kohler, and Afseth (2016) reported its propensity for overfitting and its implementation has been found ineffective in this study (data not shown).

Finally, all Raman data were treated using the difference spectra operation by subtracting the first Raman spectrum in each experiment. Difference spectra are a common method to emphasize temporal changes in spectroscopic data (Andris, Rüdt, et al. 2018) or differences between individual samples (Zhou et al. 2015; Gautam et al. 2015). The difference spectra operation was shown to effectively remove all variations in the initial composition of the studied lysate using different conditioning approaches. This improved comparability between experiments and specificity of the Raman signals for the amount of precipitated VLPs which is crucial for generating robust chemometric models (R. M. Santos et al. 2018).

In summary, the preprocessing pipeline comprising turbidity and baseline correction, a background correction technique and difference spectra contributes to the overall quality of the precipitate Raman data, improving model building. The identified preprocessing pipeline may further be transferable to other precipitants such as polyethylene glycol (PEG). In the case of PEG-based systems, it may be more challenging to resolve the differences between the precipitants and protein-related signals due to organic Raman contributions of the PEG polymers (Kuzmin et al. 2020).

5.4.3 Effects of preprocessing pipeline on model performance

Chemometric models based on Raman spectroscopy data often employ multiple preprocessing operations in sequence and require optimization of the associated hyperparameters (Bocklitz et al. 2011; Gerretzen et al. 2015; Brunel, Alsamad, and Piot 2021). To study the effect of individual preprocessing operations on model performance, multiple pipeline configurations were screened and evaluated using a nested cross-validation approach. The nested cross-validation enabled the

comparison of the pipeline configurations, not only with regard to the overall accuracy, but also their robustness when permuting the training and test sets. By further incorporating interval-based cropping of the Raman spectra and the comparison of MLR with PLS models, the understanding for the relevance of specific features in the Raman data could be increased.

In general, PLS models showed considerably better performance and lower variance in cross-validated test sets than MLR models. This may be explained by the high complexity of Raman data and the experimental variance which was observed for the batch data set. The MLR models solely rely on VLP-associated Raman bands and hence require high-quality data. The PLS models may also use other correlative information available in the Raman spectra (R. M. Santos et al. 2018; Goldrick, Umprecht, et al. 2020) which is not necessarily specific for the precipitation of VLPs such as the turbidity or precipitant markers. Nevertheless, the comparison with MLR as a benchmark is a useful strategy as in less complex cases, the relevant information may be reliably contained in a few individual variables (Hillebrandt, Vormittag, Dietrich, and Hubbuch 2022).

As anticipated from the SNR analysis, both background correction approaches improved the performance of the PLS models due to the removal of precipitant-related Raman bands. As the AMS contribution dominated the Raman spectra and was recovered in the first latent variables of the PLS models, models without prior background correction may rely too strongly on background information and hence not robustly predict the precipitated amount of VLP. Recent studies with biopharmaceuticals often rely on standardization, derivative filtering or baseline correction (R. M. Santos et al. 2018; Wei et al. 2022; Jiarui Wang et al. 2023b; Rolinger, Hubbuch, and Rüdts 2023). Others performed background corrections by subtracting blank spectra but did not further investigate the importance of the respective features on model predictions (Großhans, Suhm, and Hubbuch 2019; D. Weber and Hubbuch 2021; Rolinger, Rüdts, and Hubbuch 2021). Alternatively, Jiarui Wang et al. (2023b) performed blank chromatography runs with the elution buffer system and appended the recorded data to the calibration data set.

To assess whether the focus on spectral regions of interest improves model predictions, multiple cropping intervals were explored within this study. Cropping the spectra to $800\text{-}1800\text{ cm}^{-1}$ generally improved model performance, whereas further cropping decreased the overall accuracy of the PLS models. While most of spectral information is still contained in the $800\text{-}1800\text{ cm}^{-1}$ region, narrowing the wavenumber region to $1200\text{-}1500$ solely focused on the protein-related information (Maiti et al. 2004; Rygula et al. 2013). The narrow bandwidth was not suitable to reliably predict precipitated VLP concentrations and notably broadened the distribution of RMSE while increasing the mean error by over 50%. This may be explained by the strong variances in the sensor data between experiments leading to non-linear correlations between the spectral data and the amount of precipitated VLP. When using B4 and B6 as the hold-out test set, a RMSE of 0.74 gL^{-1} was reached, supporting the VLP-associated information being stored in the respective wavenumber regions. Because the stated RMSE lies in the center of the RMSE distribution of the held-out test sets (cf. Fig. 5.7), the data split is considered representative.

The optimal performance in the permuted batch data was found for the PLS model for the combination of turbidity, baseline, and OPLS-based background corrections with cropping to $800\text{-}1800\text{ cm}^{-1}$ and second derivative filtering. When applied to the representative splits of the batch and fed-batch data, the identified model pipeline recovered the observed trends in the experimental data but lacked to resolve the correct sigmoidal shape for experiments B7 and F2. While the pipeline proved robust against background variations induced by lysate conditioning and spiking, the sigmoidal shape of the VLP precipitation trajectories and the height of the final plateau were not exactly met. Due to the overlaying effects occurring in the Raman spectra, the insufficient removal

of the background contributions by the OPLS method and the high single-point variance in the reference data, the relationship between the VLP-related Raman bands and the target variable is considered to be non-linear. By inspecting the importance of different wavenumber regions through VIP, the model’s reliance on residual precipitant information and the VLP-associated bands became apparent. Moreover, the VIP scores suggested a strong influence of noise in the spectra on the model predictions as no Raman band is clearly recovered. Despite being robust to background variations, the preprocessing pipeline may also be influenced by unforeseen changes in the Raman spectra due to equipment failure or other experimental factors. In the case of the fed-batch experiments, the scattered predictions for the spectra during the first 3 minutes are likely due to a combination of factors. These include experimentally introduced air bubbles or large floculates, which can distort the Raman signal, and overcorrections from preprocessing operations that may perform differently when the Raman spectra show unforeseen changes. Local concentration gradients due to possible heterogeneities in precipitant concentration could also cause scattered predictions, but there is no evidence for this in the AMS-related Raman bands.

To further study the effect of noise on the model, perturbation studies (Jiarui Wang et al. 2023b) could be employed or the distribution of PLS weights could be assessed (Cui and Fearn 2018). To effectively reduce the noise in the model, a larger data set preferably from fed-batch experiments would be required. Eventually, due to the non-linear relationship of the Raman spectra and the VLP concentration, non-linear regression models should be evaluated such as kernel-based methods (Thissen et al. 2004; Barman et al. 2010; Zavala-Ortiz et al. 2020; Schiemer, Weggen, et al. 2023) or neural networks (Cui and Fearn 2018; Jiarui Wang et al. 2023a; Schiemer, Rüdts, and Hubbuch 2024).

5.5 Conclusions

In conclusion, we present a Raman spectroscopy-based PAT for real-time monitoring of VLP precipitation from clarified *E. coli*-derived lysate as well as the precipitant concentration. The generated precipitation data provide a challenging data set with varying precipitation dynamics and backgrounds which were induced by spiking and are desirable for robust PAT sensor development. High experimental variance is successfully managed by employing a pretreatment approach for the UV absorbance data. The preprocessing pipeline including turbidity, baseline and OPLS-based background correction, difference spectra, cropping and derivative filtering is identified as optimal for removing all variations in the initial composition of the studied lysates as well as most interferences caused by precipitates and precipitant in solution. The final PLS models recover the observed trends in the batch and fed-batch data but lack to resolve fine differences between experiments owing to the non-linear relationship between spectral data and the precipitated VLP concentration. Additionally, the Raman data reveal structural differences between VLPs and HCPs and qualitatively support the selective precipitation of VLPs while nucleic acids and HCPs remain in solution. Overall, the developed preprocessing pipeline provides a foundation for integrating Raman spectroscopy as a PAT sensor for monitoring particulate-containing bioprocesses and bears the potential to be applied to other phase behavior-dependent processes for protein purification.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author contributions

AD: Conceptualization, Methodology, Investigation, Formal Analysis, Visualization, Writing - Original Draft, RS: Conceptualization, Methodology, Software, Formal Analysis, Visualization, Writing - Original Draft, JK: Investigation, Software, Writing - Review & Editing, SZ: Investigation, Writing - Review & Editing, JH: Conceptualization, Supervision, Funding Acquisition, Writing - Review & Editing.

Acknowledgments

The authors would like to thank Jan T. Weggen and Christina H. Wegner for careful proof-reading of the manuscript. We acknowledge support by the KIT-Publication fund of the Karlsruhe Institute of Technology.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Chapter references

- Abdallah, Fatma F. et al. (2019). „Orthogonal projection to latent structures and first derivative for manipulation of PLSR and SVR chemometric models’ prediction: A case study“. In: *PLoS ONE* 14.9, pp. 1–13.
- Abu-Absi, Nicholas R. et al. (05/2011). „Real time monitoring of multiple parameters in mammalian cell culture bioreactors using an in-line Raman spectroscopy probe“. In: *Biotechnology and Bioengineering* 108.5, pp. 1215–1221.
- Alsmeyer, Frank, Hans Jürgen Koß, and Wolfgang Marquardt (2004). „Indirect spectral hard modeling for the analysis of reactive and interacting mixtures“. In: *Applied Spectroscopy* 58.8, pp. 975–985.
- Andris, Sebastian, Matthias Rüdert, et al. (12/2018). „Monitoring of antibody-drug conjugation reactions with UV/Vis spectroscopy“. In: *Journal of Biotechnology* 288.October, pp. 15–22.
- Barman, Ishan et al. (02/2010). „Development of Robust Calibration Models Using Support Vector Machines for Spectroscopic Monitoring of Blood Glucose“. In: *Analytical Chemistry* 82.23, pp. 9719–9726.
- Ben-David, Avishai and Hsuan Ren (2005). „Comparison between orthogonal subspace projection and background subtraction techniques applied to remote-sensing data“. In: *Applied Optics* 44.18, pp. 3846–3855.
- Berry, Brandon et al. (2015). „Cross-scale predictive modeling of CHO cell culture growth and metabolites using Raman spectroscopy and multivariate analysis“. In: *Biotechnology Progress* 31.2, pp. 566–577.
- Bocklitz, Thomas et al. (2011). „How to pre-process Raman spectra for reliable and stable models?“ In: *Analytica Chimica Acta* 704.1-2, pp. 47–56.
- Boelens, Hans F M et al. (2004). „New background correction method for liquid chromatography with diode array detection, infrared spectroscopic detection and Raman spectroscopic detection“. In: *Journal of Chromatography A* 1057.1-2, pp. 21–30.
- Brunel, Benjamin, Fatima Alsamad, and Olivier Piot (01/2021). „Toward automated machine learning in vibrational spectroscopy - use and settings of genetic algorithms for pre-processing and regression optimization“. In: *Chemometrics and Intelligent Laboratory Systems* 219.104444, pp. 1–10.
- Capito, Florian, Romas Skudas, Harald Kolmar, et al. (2015). „At-line mid infrared spectroscopy for monitoring downstream processing unit operations“. In: *Process Biochemistry* 50.6, pp. 997–1005.
- Chackerian, Bryce (2007). „Virus-like particles: Flexible platforms for vaccine development“. In: *Expert Review of Vaccines* 6.3, pp. 381–390.
- Chen, Jingyi et al. (03/2024). „Application of Raman spectroscopy during pharmaceutical process development for determination of critical quality attributes in Protein A chromatography“. In: *Journal of Chromatography A* 1718, p. 464721.
- Cooper, Arik and Yosef Shaul (02/2005). „Recombinant viral capsids as an efficient vehicle of oligonucleotide delivery into cells“. In: *Biochemical and Biophysical Research Communications* 327.4, pp. 1094–1099.
- Cui, Chenhao and Tom Fearn (11/2018). „Modern practical convolutional neural networks for multivariate regression: Applications to NIR calibration“. In: *Chemometrics and Intelligent Laboratory Systems* 182, pp. 9–20.
- Curtis, R. A., J. M. Prausnitz, and H. W. Blanch (01/1998). „Protein-protein and protein-salt interactions in aqueous protein solutions containing concentrated electrolytes“. In: *Biotechnology and Bioengineering* 57.1, pp. 11–21.
- Effio, Christopher Ladd and Jürgen Hubbuch (05/2015). „Next generation vaccines and vectors: Designing downstream processes for recombinant protein-based virus-like particles“. In: *Biotechnology Journal* 10.5, pp. 715–727.
- Falkenberg, Heiner et al. (05/2019). „Mass spectrometric evaluation of upstream and downstream process influences on host cell protein patterns in biopharmaceutical products“. In: *Biotechnology Progress* 35.3.
- FDA (2004). „Guidance for Industry: PAT—a framework for innovative pharmaceutical development, manufacturing, and quality assurance“. In: .
- FDA (2015). „Development and Submission of Near Infrared Analytical Procedures Guidance for Industry“. In: March, p. 24.
- Feidl et al. (10/2019). „Combining Mechanistic Modeling and Raman Spectroscopy for Monitoring Antibody Chromatographic Purification“. In: *Processes* 7.10, pp. 683–683.
- Fontana, Marc D., Kawther Ben Mabrouk, and Thomas H. Kauffmann (2013). „Raman spectroscopic sensors for inorganic salts“. In: *Spectroscopic Properties of Inorganic and Organometallic Compounds*. Ed. by Jack Yarwood, Richard Douthwaite, and Simon Duckett. Vol. 44. RSC Publishing, pp. 40–67.
- Gasteiger, Elisabeth et al. (2005). „Protein Identification and Analysis Tools on the ExPASy Server“. In: *The Proteomics Protocols Handbook*, pp. 571–607.
- Gautam, Rekha et al. (2015). „Review of multidimensional data processing approaches for Raman and infrared spectroscopy“. In: *EPJ Techniques and Instrumentation* 2.1.
- Genova, E. et al. (2018). „SERS of cells: What can we learn from cell lysates?“ In: *Analytica Chimica Acta* 1005, pp. 93–100.
- Gerretzen, Jan et al. (09/2015). „Simple and Effective Way for Data Preprocessing Selection Based on Design of Experiments“. In: *Analytical Chemistry* 87.24, pp. 12096–12103.
- Glasse, Jarka et al. (04/2011). „Process analytical technology (PAT) for biopharmaceuticals“. In: *Biotechnology Journal* 6.4, pp. 369–377.
- Golabgir, Aydin and Christoph Herwig (2016). „Combining Mechanistic Modeling and Raman Spectroscopy for Real-Time Monitoring of Fed-Batch Penicillin Production“. In: *Chemie-Ingenieur-Technik* 88.6, pp. 764–776.
- Goldrick, Stephen, Alexandra Umrecht, et al. (09/2020). „High-Throughput Raman Spectroscopy Combined with Innovative Data Analysis Workflow to Enhance Biopharmaceutical Process Development“. In: *Processes* 8.9 (9), pp. 1179–1179.
- Grön, Heidi et al. (2003). „Dynamic in-process examination of particle size and crystallographic form under defined conditions of reactant supersaturation as associated with the batch crystallization of monosodium glutamate from aqueous solution“. In: *Industrial and Engineering Chemistry Research* 42.20, pp. 4888–4898.
- Großhans, Steffen, Susanna Suhm, and Jürgen Hubbuch (2019). „Precipitation of complex antibody solutions: influence of contaminant composition and cell culture medium on the precipitation behavior“. In: *Bioprocess and Biosystems Engineering*.

CHAPTER 5. RAMAN-BASED PAT FOR VLP PRECIPITATION

- Guardalini, Luis Giovanni de Oliveira, Vinícius Aragão Tejo Dias, et al. (01/2023). „Comparison of chemometric models using Raman spectroscopy for offline biochemical monitoring throughout the VLP-making upstream process“. In: *Biochemical Engineering Journal* 198, pp. 109013–109013.
- Guardalini, Luis Giovanni de Oliveira, Rafaela Morais Rangel, et al. (12/2023). „Monitoring by Raman spectroscopy of rabies virus-like particles production since the initial development stages“. en. In: *Journal of Chemical Technology & Biotechnology* n/a.n/a.
- Guardalini, Luis Giovanni de Oliveira, Paulo Eduardo da Silva Cavalcante, et al. (2023). „Biochemical monitoring throughout all stages of rabies virus-like particles production by Raman spectroscopy using global models“. eng. In: *Journal of Biotechnology* 363, pp. 19–31.
- Hämmerling, Frank et al. (01/2017). „Investigation and prediction of protein precipitation by polyethylene glycol using quantitative structure–activity relationship models“. In: *Journal of Biotechnology* 241, pp. 87–97.
- Harner, Richard S. et al. (2009). „Use of a fiber-optic turbidity probe to monitor and control commercial-scale unseeded batch crystallizations“. In: *Organic Process Research and Development* 13.1, pp. 114–124.
- Hassebroek, Anna M. et al. (02/2023). „A hepatitis B virus core antigen-based virus-like particle vaccine expressing SARS-CoV-2 B and T cell epitopes induces epitope-specific humoral and cell-mediated immune responses but confers limited protection against SARS-CoV-2 infection“. In: *Journal of Medical Virology* 95.2.
- Hassoun, Mohamed et al. (2018). „A droplet-based microfluidic chip as a platform for leukemia cell lysate identification using surface-enhanced Raman scattering“. In: *Analytical and Bioanalytical Chemistry* 410.3, pp. 999–1006.
- He, Shixuan et al. (2014). „Baseline correction for Raman spectra using an improved asymmetric least squares method“. In: *Anal. Methods* 6.12, pp. 4402–4407.
- Hillebrandt, Nils, Philipp Vormittag, Nicolai Bluthardt, et al. (2020). „Integrated Process for Capture and Purification of Virus-Like Particles: Enhancing Process Performance by Cross-Flow Filtration“. In: *Frontiers in Bioengineering and Biotechnology* 8.May.
- Hillebrandt, Nils, Philipp Vormittag, Annabelle Dietrich, and Jürgen Hubbuch (06/2022). „Process monitoring framework for cross-flow diafiltration-based virus-like particle disassembly: Tracing product properties and filtration performance“. In: *Biotechnology and Bioengineering* 119.6, pp. 1522–1538.
- Huang, Hongxia and Haibin Qu (2018). „In-situ monitoring of saccharides removal of alcohol precipitation using near-infrared spectroscopy“. In: *Journal of Innovative Optical Health Sciences* 11.5, pp. 1–12.
- Iverius, P.H. and T.C. Laurent (02/1967). „Precipitation of some plasma proteins by the addition of dextran or polyethylene glycol“. In: *Biochimica et Biophysica Acta (BBA) - Protein Structure* 133.2, pp. 371–373.
- Kazaks, Andris et al. (2017). „Production and purification of chimeric HBc virus-like particles carrying influenza virus LAH domain as vaccine candidates“. In: *BMC Biotechnology* 17.1, pp. 1–11.
- Kim, Hyoung Jin et al. (2010). „One-step chromatographic purification of human papillomavirus type 16 L1 protein from *Saccharomyces cerevisiae*“. In: *Protein Expression and Purification* 70.1, pp. 68–74.
- Klamp, Thorsten et al. (01/2011). „Highly Specific Auto-Antibodies against Claudin-18 Isoform 2 Induced by a Chimeric HBcAg Virus-Like Particle Vaccine Kill Tumor Cells and Inhibit the Growth of Lung Metastases“. In: *Cancer Research* 71.2, pp. 516–527.
- Kohler, A. et al. (03/2008). „Estimating and Correcting Mie Scattering in Synchrotron-Based Microscopic Fourier Transform Infrared Spectra by Extended Multiplicative Signal Correction“. In: *Applied Spectroscopy* 62.3, pp. 259–266.
- Koho, Tiia et al. (2012). „Purification of norovirus-like particles (VLPs) by ion exchange chromatography“. In: *Journal of Virological Methods* 181.1, pp. 6–11.
- Kriesten, E. et al. (2008). „Fully automated indirect hard modeling of mixture spectra“. In: *Chemometrics and Intelligent Laboratory Systems* 91.2, pp. 181–193.
- Kushnir, Natasha, Stephen J. Streatfield, and Vidadi Yusibov (12/2012). „Virus-like particles as a highly efficient vaccine platform: Diversity of targets and production systems and advances in clinical development“. In: *Vaccine* 31.1, pp. 58–83.
- Kuzmin, V. V. et al. (10/2020). „Raman spectra of polyethylene glycols: Comparative experimental and DFT study“. In: *Journal of Molecular Structure* 1217, p. 128331.
- Li, Boyan et al. (2010). „Rapid characterization and quality control of complex cell culture media solutions using Raman spectroscopy and chemometrics“. In: *Biotechnology and Bioengineering* 107.2, pp. 290–301.
- Li, Lian et al. (2014). „The relevance study of effective information between near infrared spectroscopy and chondroitin sulfate in ethanol precipitation process“. In: *Journal of Innovative Optical Health Sciences* 7.6, pp. 1–7.
- Liland, Kristian Hovde, Achim Kohler, and Nils Kristian Afseth (06/2016). „Model-based pre-processing in Raman spectroscopy of biological samples“. en. In: *Journal of Raman Spectroscopy* 47.6, pp. 643–650.
- Liu, Yi, Markus Pietzsch, and Joachim Ulrich (2014). „Determination of the phase diagram for the crystallization of L-asparaginase II by a turbidity technique“. In: *Crystal Research and Technology* 49.4, pp. 262–268.
- Lohmann, Lara Julia and Jochen Strube (2021). „Process analytical technology for precipitation process integration into biologics manufacturing towards autonomous operation–mab case study“. In: *Processes* 9.3.
- Lorber, Avraham (1986). „Error Propagation and Figures of Merit for Quantification by Solving Matrix Equations“. In: *Analytical Chemistry* 58.6, pp. 1167–1172.
- Maiti, Nakul C. et al. (03/2004). „Raman Spectroscopic Characterization of Secondary Structure in Natively Unfolded Proteins: α -Synuclein“. In: *Journal of the American Chemical Society* 126.8, pp. 2399–2408.
- Martens, Harald, Jesper Pram Nielsen, and Seren Bailing Engelsen (2003). „Light scattering and light absorbance separated by extended multiplicative signal correction. Application to near-infrared transmission analysis of powder mixtures“. In: *Analytical Chemistry* 75.3, pp. 394–404.

- Martens, Harald and Edward Stark (01/1991a). „Extended multiplicative signal correction and spectral interference subtraction: New preprocessing methods for near infrared spectroscopy“. In: *Journal of Pharmaceutical and Biomedical Analysis* 9.8, pp. 625–635.
- Mehmood, Tahir et al. (08/2012). „A review of variable selection methods in Partial Least Squares Regression“. In: *Chemometrics and Intelligent Laboratory Systems* 118, pp. 62–69.
- Meyer-Kirschner, Julian et al. (2016). „In-line Monitoring of Monomer and Polymer Content during Microgel Synthesis Using Precipitation Polymerization via Raman Spectroscopy and Indirect Hard Modeling“. In: *Applied Spectroscopy* 70.3, pp. 416–426.
- Mobini, Saeed et al. (08/2020). „Computational Design of a Novel VLP-Based Vaccine for Hepatitis B Virus“. In: *Frontiers in Immunology* 11.
- Moleirinho, Mafalda G. et al. (05/2020). „Current challenges in biotherapeutic particles manufacturing“. In: *Expert Opinion on Biological Therapy* 20.5, pp. 451–465.
- Moradi Vahdat, Maryam et al. (03/2021). „Hepatitis B core-based virus-like particles: A platform for vaccine development in plants“. In: *Biotechnology Reports* 29, e00605.
- Moreno, A. et al. (2000). „Turbidity as a useful optical parameter to predict protein crystallization by dynamic light scattering“. In: *Journal of Molecular Structure* 519.1-3, pp. 243–256.
- Nooraei, Saghi et al. (02/2021). „Virus-like particles: preparation, immunogenicity and their roles as nanovaccines and drug nanocarriers“. In: *Journal of Nanobiotechnology* 19.1, p. 59.
- Passos, Cláudia P. et al. (2010). „Application of Fourier transform infrared spectroscopy and orthogonal projections to latent structures/partial least squares regression for estimation of procyanidins average degree of polymerisation“. In: *Analytica Chimica Acta* 661.2, pp. 143–149.
- Peticolas, Warner L. (1995). „Raman spectroscopy of DNA and proteins“. In: *Methods in Enzymology*. Vol. 246. Elsevier Inc., pp. 389–416.
- Petrovskis, Ivars et al. (01/2021). „Production of the HBc Protein from Different HBV Genotypes in E. coli. Use of Reassociated HBc VLPs for Packaging of ss- and dsRNA“. In: *Microorganisms* 9.2, p. 283.
- Poison, A. (01/1977). „A Theory for the Displacement of Proteins and Viruses with Polyethylene Glycol“. In: *Preparative Biochemistry* 7.2, pp. 129–154.
- Pons Royo, Maria del Carme et al. (2022). „Mode and dosage time in polyethylene glycol precipitation process influences protein precipitate size and filterability“. In: *Process Biochemistry* 114, pp. 77–85.
- Porterfield, J. Zachary and Adam Zlotnick (11/2010). „A simple and general method for determining the protein and nucleic acid content of viruses by UV absorbance“. In: *Virology* 407.2, pp. 281–288.
- Qian, Ciyang et al. (03/2020). „Recent Progress on the Versatility of Virus-Like Particles“. In: *Vaccines* 8.1, p. 139.
- Rathore, A. S., R. Bhambure, and V. Ghare (09/2010). „Process analytical technology (PAT) for biopharmaceutical products“. In: *Analytical and Bioanalytical Chemistry* 398.1 (1), pp. 137–154.
- Rinnan, Åsmund et al. (2009). „Data Pre-processing“. In: *Infrared Spectroscopy for Food Quality Analysis and Control*. Vol. 3. Academic Press, pp. 29–50.
- Roessl, Ulrich et al. (2014). „In situ protein secondary structure determination in ice: Raman spectroscopy-based process analytical tool for frozen storage of biopharmaceuticals“. In: *Journal of Pharmaceutical Sciences* 103.8, pp. 2287–2295.
- Rolinger, Laura, Jürgen Hubbuch, and Matthias Rüdert (02/2023). „Monitoring of ultra- and diafiltration processes by Kalman-filtered Raman measurements“. In: *Analytical and Bioanalytical Chemistry* 415.5, pp. 841–854.
- Rolinger, Laura, Matthias Rüdert, and Jürgen Hubbuch (11/2021). „Comparison of UV- and Raman-based monitoring of the Protein A load phase and evaluation of data fusion by PLS models and CNNs“. In: *Biotechnology and Bioengineering* 118.11 (11), pp. 4255–4268.
- Rüdert, Matthias, Philipp Vormittag, et al. (06/2019). „Process monitoring of virus-like particle reassembly by diafiltration with UV/Vis spectroscopy and light scattering“. In: *Biotechnology and Bioengineering* 116.6, pp. 1366–1379.
- Rygula, A. et al. (2013). „Raman spectroscopy of proteins: A review“. In: *Journal of Raman Spectroscopy* 44.8, pp. 1061–1076.
- Santos, Rafael M. et al. (2018). „Monitoring mAb cultivations with in-situ raman spectroscopy: The influence of spectral selectivity on calibration models and industrial use as reliable PAT tool“. In: *Biotechnology Progress* 34.3, pp. 659–670.
- Savitzky, Abraham. and M. J. E. Golay (07/1964). „Smoothing and Differentiation of Data by Simplified Least Squares Procedures“. In: *Analytical Chemistry* 36.8, pp. 1627–1639.
- Schiemer, Robin, Matthias Rüdert, and Jürgen Hubbuch (2024). „Generative data augmentation and automated optimization of convolutional neural networks for process monitoring“. In: *Frontiers in Bioengineering and Biotechnology* 12.
- Schiemer, Robin, Jan Tobias Weggen, et al. (05/2023). „An adaptive soft-sensor for advanced real-time monitoring of an antibody-drug conjugation reaction“. In: *Biotechnology and Bioengineering* 120.7, pp. 1914–1928.
- Sinfield, Joseph V. and Chike K. Monwuba (2014). „Assessment and correction of turbidity effects on Raman observations of chemicals in aqueous solutions“. In: *Applied Spectroscopy* 68.12, pp. 1381–1392.
- Soch, Joram and Carsten Allefeld (08/2018). „MACS – a new SPM toolbox for model assessment, comparison and selection“. In: *Journal of Neuroscience Methods* 306, pp. 19–31.
- Socrates, G. (2004). *Infrared and Raman Characteristic Group Frequencies: Tables and Charts*. Wiley.
- Spinner, E. (05/2003). „Raman-spectral depolarisation ratios of ions in concentrated aqueous solution. The next-to-negligible effect of highly asymmetric ion surroundings on the symmetry properties of polarisability changes during vibrations of symmetric ions“. In: *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 59.7, pp. 1441–1456.
- Sun, Zhongyu, Jiajin Fan, et al. (2020). „Assessment of the human albumin in acid precipitation process using NIRS and multi-variable selection methods combined with SPA“. In: *Journal of Molecular Structure* 1199, p. 126942.

CHAPTER 5. RAMAN-BASED PAT FOR VLP PRECIPITATION

- Sun, Zhongyu, Jiayue Wang, et al. (2018). „Calibration transfer of near infrared spectrometers for the assessment of plasma ethanol precipitation process“. In: *Chemometrics and Intelligent Laboratory Systems* 181.August, pp. 64–71.
- Thissen, U et al. (2004). „Comparing support vector machines to PLS for spectral regression applications“. In: *Chemometrics and Intelligent Laboratory Systems* 73.2, pp. 169–179.
- Trygg, Johan and Svante Wold (2002). „Orthogonal projections to latent structures (O-PLS)“. In: *Journal of Chemometrics* 16.3, pp. 119–128.
- Tsoka, Sophia et al. (01/2000). „Selective Flocculation and Precipitation for the Improvement of Virus-Like Particle Recovery from Yeast Homogenate“. In: *Biotechnology Progress* 16.4, pp. 661–667.
- Valentic, Angela, Jakob Müller, and Jürgen Hubbuch (07/2022). „Effects of Different Lengths of a Nucleic Acid Binding Region and Bound Nucleic Acids on the Phase Behavior and Purification Process of HBcAg Virus-Like Particles“. In: *Frontiers in Bioengineering and Biotechnology* 10.
- Van Brink, Mark Den, Michel Pepers, and Alex M. Van Herk (2002). „Raman spectroscopy of polymer latexes“. In: *Journal of Raman Spectroscopy* 33.4, pp. 264–272.
- Wang, Jiarui et al. (2023a). „Automated calibration and in-line measurement of product quality during therapeutic monoclonal antibody purification using Raman spectroscopy“. In: *Biotechnology and Bioengineering*.
- Wang, Jiarui et al. (12/2023b). „In-line product quality monitoring during biopharmaceutical manufacturing using computational Raman spectroscopy“. In: *mAbs* 15.1, p. 2220149.
- Watson, G. H., W. B. Daniels, and C. S. Wang (02/1981). „Measurements of Raman intensities and pressure dependence of phonon frequencies in sapphire“. In: *Journal of Applied Physics* 52.2, pp. 956–958.
- Weber, Alfons et al. (01/1967). „High-Resolution Raman Spectroscopy of Gases with cw-Laser Excitation*“. In: *JOSA* 57.1, pp. 19–28.
- Weber, Dennis and Jürgen Hubbuch (2021). „Raman spectroscopy as a process analytical technology to investigate biopharmaceutical freeze concentration processes“. en. In: *Biotechnology and Bioengineering* 118.12, pp. 4708–4719.
- Wegner, Christina H., Ines Zimmermann, and Jürgen Hubbuch (02/02/2022). „Rapid Analysis for Multicomponent High-Throughput Crystallization Screening: Combination of UV–Vis Spectroscopy and Chemometrics“. In: *Crystal Growth & Design* 22.2, pp. 1054–1065.
- Wei, Bingchuan et al. (12/2022). „Multi-attribute Raman spectroscopy (MARS) for monitoring product quality attributes in formulated monoclonal antibody therapeutics“. In: *mAbs* 14.1 (1).
- Wilcox, R. R (2005). *Introduction to Robust Estimation and Hypothesis Testing*. Netherlands: Elsevier Science, p. 100.
- Wold, Svante, Henrik Antti, et al. (1998). „Orthogonal signal correction of near-infrared spectra“. In: *Chemometrics and Intelligent Laboratory Systems* 44.1-2, pp. 175–185.
- Wold, Svante, Michael Sjöström, and Lennart Eriksson (10/2001). „PLS-regression: a basic tool of chemometrics“. en. In: *Chemometrics and Intelligent Laboratory Systems* 58.2 (2), pp. 109–130.
- Zahin, Maryam et al. (2016). „Scalable production of HPV16 L1 protein and VLPs from tobacco leaves“. In: *PLoS ONE* 11.8, pp. 1–16.
- Zavala-Ortiz, Daniel Arturo et al. (2020). „Support Vector and Locally Weighted regressions to monitor monoclonal antibody glycosylation during CHO cell culture processes, an enhanced alternative to Partial Least Squares regression“. In: *Biochemical Engineering Journal* 154.August 2019, p. 107457.
- Zelger, Monika et al. (2016). „Real-time monitoring of protein precipitation in a tubular reactor for continuous bioprocessing“. In: *Process Biochemistry* 51.10, pp. 1610–1621.
- Zeltins, Andris (01/2013). „Construction and Characterization of Virus-Like Particles: A Review“. In: *Molecular Biotechnology* 53.1, pp. 92–107.
- Zhang, Feng et al. (03/2020). „Baseline correction for infrared spectra using adaptive smoothness parameter penalized least squares method“. In: *Spectroscopy Letters* 53.3, pp. 222–233.
- Zhou, Chen et al. (03/2015). „Concomitant Raman spectroscopy and dynamic light scattering for characterization of therapeutic proteins at high concentrations“. en. In: *Analytical Biochemistry* 472, pp. 7–20.
- Zlotnick, A. et al. (01/1996). „Dimorphism of Hepatitis B Virus Capsids Is Strongly Influenced by the C-Terminus of the Capsid Protein“. In: *Biochemistry* 35.23, pp. 7412–7421.

6

A novel Raman-chromatography assembly for automated calibration and in-line monitoring in bioprocessing

Jakob Heyer-Müller^{1,+}, Robin Schiemer^{1, +}, Matthias Lopinski¹, Caty Wang¹, Franka Willems¹, Lars Robbel², Michael Schmitt², and Jürgen Hubbuch¹

¹ Institute of Engineering in Life Sciences, Section IV: Biomolecular Separation Engineering, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

² CSL Innovation GmbH Germany, Michelbacher Straße, 35041 Marburg, Hessen, Germany

⁺ contributed equally

Abstract

Optical spectroscopic techniques have been successfully employed in bioprocessing as Process Analytical Technology for real-time process monitoring in numerous applications. The implementation of spectroscopy-based PAT techniques commonly necessitates the generation of representative process data used for calibration and validation of multivariate statistical models for analysing the sample composition in real-time. To automate the generation of such data, we present a novel assembly of a commercially available chromatography system in combination with a Raman spectrometer for fast and accurate acquisition of Raman spectra. Using the ultra-/diafiltration (UF/DF) process as a case study, our methodology involved the preparation of representative calibration and validation mixtures of phosphate and citrate buffer and lysozyme as a model protein. Chemometric PLS models were calibrated and validated using these datasets, and applied to in-line recorded Raman spectra during an UF/DF experiment. The primary results demonstrated that the novel assembly provides robust and precise off-line measurement of Raman spectra which directly compare with in-line record data. The chemometric PLS models showed good alignment in calibration and validation datasets (R^2 and Q^2), and could be used to simultaneously monitor the buffer and protein concentrations in real-time during UF/DF. This study provides a simple, commercially available setup for automated acquisition of Raman spectra and demonstrates its straightforward application to bioprocess monitoring.

6.1 Introduction

Optical spectroscopic techniques are powerful tools for chemical analysis due to its ability to provide detailed molecular information without the need for extensive sample preparation. Optical spectroscopy is often applied in combination with multivariate statistical modeling as a Process Analytical Technology (PAT) for the analysis of sample composition in chemical or pharmaceutical processes. In biopharmaceutical downstream processing, the application of Raman spectroscopy has been successfully applied, e.g. for the analysis of chromatography (Feidl et al. 2019), filtration (Rolinger, Hubbuch, and Rüdts 2023), or precipitation (Dietrich et al. 2024) processes. The implementation of spectroscopy-based PAT tools into downstream processes requires the generation of representative data for training, optimisation and testing of statistical chemometric models which is time-consuming and labor-intensive.

Representative data sets may be generated by systematic variation of the operating conditions of the process of interest (Dietrich et al. 2024; Brestrich, Rüdts, et al. 2018), artificial spiking with the target analytes or synthetic mixtures (R. M. Santos et al. 2018; Romann et al. 2022; Klaverdijk, Ottens, and Klijn 2025). The use of synthetic mixtures, when available, facilitates an off-line model calibration (Thakur, Hebbs, and Anurag S. Rathore 2021) or can be used to supplement process-derived data sets (Klaverdijk, Ottens, and Klijn 2025).

To streamline off-line model calibration and improve reproducibility, automated sample preparation and data collection are essential. While ultraviolet/visible (UV/Vis) spectra can easily be acquired in a high-throughput mode using plate readers (Hansen, Jamali, and Hubbuch 2013), vibrational spectroscopy data such as Raman spectra strongly depend on acquisition settings and the experimental setup. If the chemometric model is intended for on-line or in-line use, off-line model calibration using the same experimental setup is desirable to avoid error-prone data correction methodologies (Jiarui Wang et al. 2023b). Recent advancements in automation and data processing have shown promise in addressing these limitations. For example, a custom integrated Raman-Tecan system was developed to automate the calibration process, generating Raman-based models suitable for

in-line usage (J. Chen et al. 2024).

Monitoring biopharmaceutical ultra-/diafiltration (UF/DF) processes through analytical tools in order to quantify or control the Donnan effect have long been a goal in industry and academia (Abel et al. 2018; West et al. 2021; Thakur, Hebbs, and Anurag S. Rathore 2021). Monitoring the solution composition in UF/DF processes including protein and excipient concentrations have been realized through combinations of multiple sensors (Rolinger, Rüdts, Diehm, et al. 2020), near-infrared (NIR) spectroscopy (Thakur, Hebbs, and Anurag S. Rathore 2021) or more recently Raman spectroscopy (Rolinger, Hubbuch, and Rüdts 2023). Thakur, Hebbs, and Anurag S. Rathore (2021) have shown the suitability of off-line calibrated chemometric models to monitor UF/DF processes in real-time. (Rolinger, Hubbuch, and Rüdts 2023) combined Raman-based chemometric models with a simple mass-balance model through a Kalman filtering approach to increase accuracy of their model's predictions.

Our study introduces a novel assembly of a commercially available ÄKTA chromatography system in combination with an autosampler and a Raman spectrometer to automate Raman measurements with a minimal sample volume of 400 μ L. A custom software is used to control the ÄKTA chromatography system via its open platform communication - unified architecture (OPC-UA) interface. Two datasets involving synthetic mixtures of phosphate- and citrate-buffered lysozyme solutions are prepared using different statistical designs. Partial least squares (PLS) regression models are calibrated using the data from the synthetic mixtures. The calibrated PLS models are then applied to in-line Raman spectra from UF/DF involving the same buffer-protein matrix.

The goals of this study are to demonstrate the feasibility and effectiveness of automating Raman spectroscopy measurements using a chromatography system for straightforward model calibration. The article outlines the process from system integration and sample preparation to model calibration and validation, showcasing the potential of this automated approach to enhance the efficiency and accuracy of Raman spectroscopy measurements. Our results indicate that the off-line calibrated, chemometric PLS models showed good in-line performance, following expected trends of the UF/DF experiment, demonstrating the capability of the automated system and the developed models to accurately monitor and predict the concentration of analytes in real-time.

6.2 Materials and methods

6.2.1 Experimental setup

An ÄKTApurifier system equipped with an autosampler A-905 was attached to a Raman spectrometer to automatically inject samples into the flow cell of the Raman spectrometer. A schematic depiction of this setup is shown in Fig. 6.1 (A). The system was controlled using a custom-made software communicating via OPC-UA with the ÄKTA chromatography system. A schematic communication protocol is depicted in Fig. 6.1 (B). To enable automated equilibration, loading, injection, measurement and washing, the valves and tubing connections were chosen as visualized in Fig. 6.1 (C)-(G). For all injections in this study, an injection volume of 490 μ L was used. The protocol for a single injection involved 8 ml of system equilibration with ultrapure water (PURELAB Ultra, ELGA LabWater), followed the injection of the sample solution into the Raman spectrometer flow cell. While the measurement was on-going, the flow rate of the system was set to 0 ml/min. Each measurement was performed for 29s totaling to an acquisition of 24 Raman spectra (see Sec. 6.2.2 for more information on the Raman spectroscopy measurements). Afterwards, the sample was

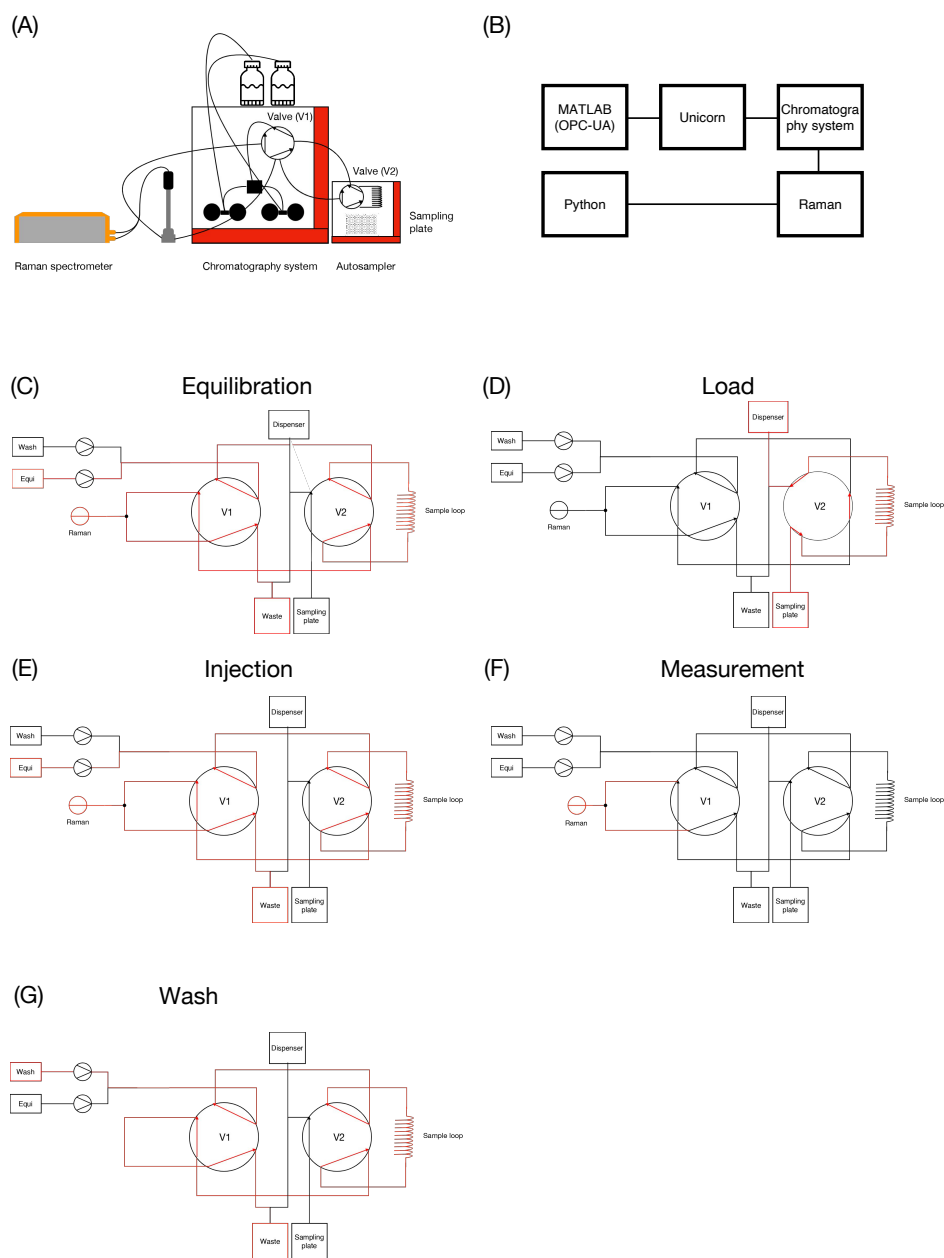


Figure 6.1 Experimental setup connecting ÄKTA purifier, A905 autosampler and Tornado Hyperflux PRO Plus. Schematic illustration of the experimental setup (A). Schematic depiction of software interfaces (B). Flow paths for different phases: equilibration (C), sample load (D), sample injection (E), measurement (F) and wash (G). Red and black lines indicate active and non-active flow paths, respectively.

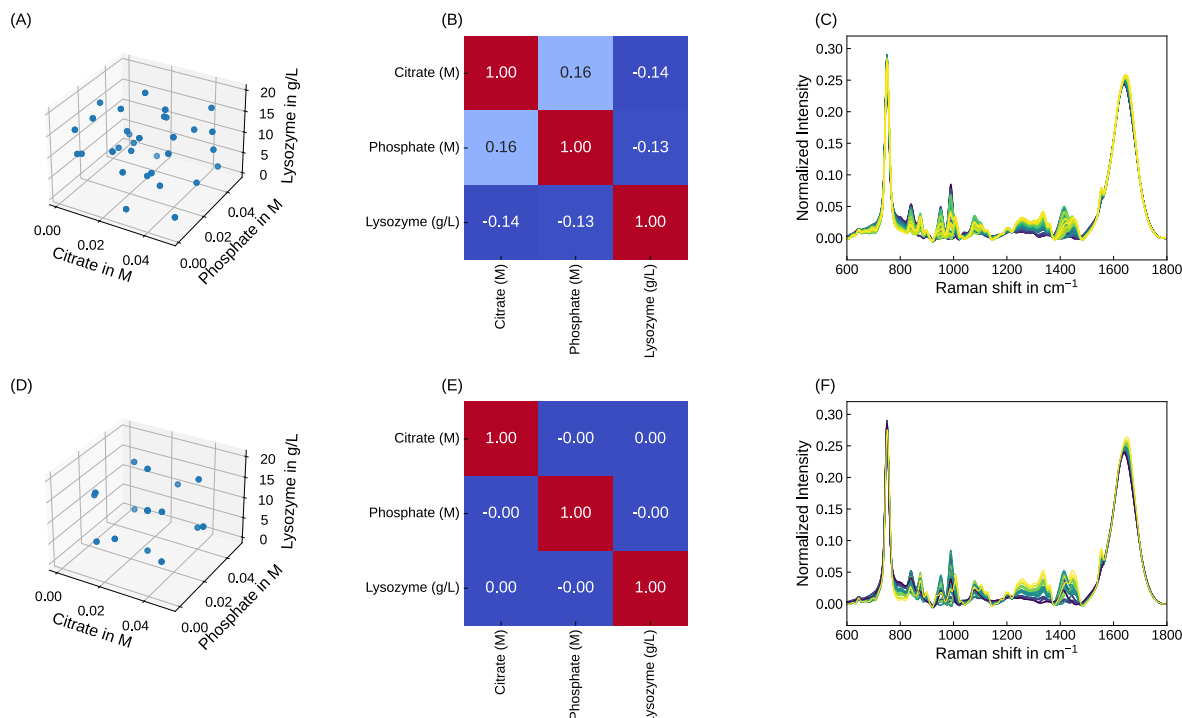


Figure 6.2 Sample distribution (A),(D), correlation coefficients (B),(E), and preprocessed Raman spectra (C), (F) for latin hypercube sampling (LHS) and central composite design (CCD) datasets in the upper and lower row, respectively.

washed out using 0.1 M NaOH for 2 ml and a subsequent reequilibration with ultrapure water for 3 ml. All phases were performed using a flow rate of 2 ml/min.

6.2.2 Raman spectroscopy

A Hyperflux PRO Plus (Tornado Spectral Systems, Toronto, Canada) equipped with a 785 nm laser, a Bioreactor BallProbeTM and a 200 μL flow cell (both MarqMetrix, Seattle, USA) were used. Raman spectra were recorded at a laser power of 495 mW and an exposure time of 1200 ms. For the off-line data, from the 24 exposures the first and last two were excluded leading to a total of 20 exposures per spectrum. For the in-line data, the same setting as for the off-line data were used, but only 10 exposures were averaged per spectrum.

6.2.3 Case study: sample preparation

Samples of 600 μL were automatically prepared in 2 ml deep well plates using a Freedom Evo liquid handling station (Tecan, Männedorf, Switzerland), containing different concentrations of citrate and phosphate buffer and lysozyme. Two datasets were prepared for calibration and validation of chemometric models using different experimental designs, namely a LHS and a CCD. The LHS and CCD sets were composed of 32 and 18 conditions with 3 replicates per condition, totalling at 96 and 54 samples, respectively. The compositions of all samples of the LHS and CCD datasets are illustrated in Fig. 6.2 (A) and (C). All samples were prepared using stock solutions of 0.3 M sodium phosphate buffer (pH 7.0), 0.4 M sodium citrate buffer (pH 5.8) and 50 mg/mL lysozyme

dissolved in 0.1 M sodium chloride. All solutions were prepared with ultrapure water (PURELAB Ultra, ELGA LabWater).

6.2.4 Data preprocessing and chemometric modeling

Raman spectra were preprocessed by a sequence of operations. First, all spectra were cropped to wavenumbers from 500 to 1800 cm^{-1} . The spectra were then smoothed using a Savitzky–Golay (SG) filter with a window size of 21 and polynomial order of 3. All spectra were subsequently normalized using the sapphire band at 750 cm^{-1} . Due to discrepancies in absolute intensity between off-line and in-line data, which most likely occurred due to differences in the lysozyme batch, a direct standardization protocol was used according to Nikzad-Langerodi and Andries (2021). As no direct calibration samples were available, synthetic spectra representing the starting condition of the in-line experiments were generated using the local subset augmentation (LSA) algorithm presented in Schiemer, Rüdts, and Hubbuch (2024). The direct standardization problem was formulated as

$$\|\mathbf{X}_T\mathbf{W} - \mathbf{X}_S + \lambda_{\text{reg}} * \mathbf{D}^T\mathbf{D}\| \quad (6.1)$$

, where $\mathbf{X}_T$ and $\mathbf{X}_S$ are the target and source spectra, respectively. The equation is further regularized using the second-order difference matrix $\mathbf{D}$ and a regularization factor $\lambda_{\text{reg}} = 0.001$. The residual between the target and source spectra are minimized by determining the weight matrix $\mathbf{W}$. The direct standardization procedure is applied for each experimental run individually. The weight matrix $\mathbf{W}$ determined by solving the direct standardization problem for the first spectrum were then used to transform the remaining in-line spectra via $\mathbf{X}'_T = \mathbf{X}_T\mathbf{W}$ of the corresponding experiment. After standardization, baseline drifts were corrected in both the off-line and in-line data using a derivative peak-screening asymmetric least squares algorithm (DERPSALSA)-based Whittaker filter as implemented in the pybaselines library (v1.1.0) and proposed in Korepanov (2020). The Whittaker filter was configured to $\lambda = 10^6$, $k = 0.05$ and $p = 0.01$. The parameters were determined by visual comparison and yielded comparable results for both the in-line and off-line data. An illustrative overview of the preprocessing and standardization procedure can found in the Supplementary Material in Fig. S1.

Chemometric models were calibrated using the LHS dataset and their performance was evaluated using the CCD sets. All PLS models were used as implemented in scikit-learn (v1.5.2). The number of latent variables was optimized by K-Fold cross-validation of the LHS data using 3 folds. The optimal number of latent variables was chosen based on the scaled RMSECV according to S. Wold, M. Sjöström, and Eriksson (2001). Before being fed into the PLS models, all preprocessed spectra were treated by a first-derivative SG filter with a window size of 11 and a second-order polynomial and subsequently mean-centered. For the lysozyme model, the spectra were finally cropped to the wavenumber range from 900 to 1300 cm^{-1} to reduce the co-influence of the other species. Separate PLS models were used for each target species, namely phosphate, citrate and lysozyme. All computations were done in Python 3.10.

6.2.5 Ultrafiltration/Diafiltration

To evaluate the calibrated PLS models for in-line use, two UF/DF experiments were conducted at starting concentrations of approx. 10 and 5 g/L lysozyme. To further increase variability, the starting concentration of phosphate and the citrate concentration in the diafiltration (DF) buffer were varied alongside the protein concentration. A summary of the process conditions is presented in Tab. 6.1.

CHAPTER 6. AUTOMATED RAMAN MODEL CALIBRATION

Table 6.1 Experimental conditions for different runs.

Experiment	Lysozyme [g/L]	Phosphate [M]	Citrate [M]	Pressure [bar]
Run 1	11.5	37.5	25	0.4
Run 2	5.7	35.0	75	0.8

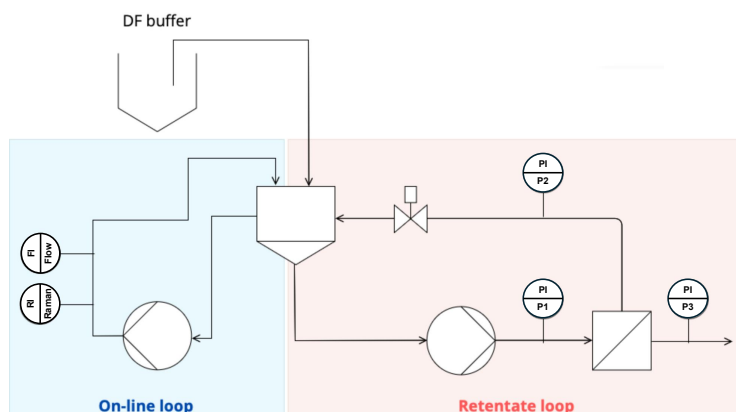


Figure 6.3 Process flow diagram of the CFF system with an external loop for on-line Raman measurements.

For UF/DF, a KrosFlo KRIII (Spectrum Labs) cross-flow filtration (CFF) system was used. A MicroKros Hollow Fiber filter module with a molecular weight cut-off (MWCO) of 3 kDa and a membrane area of 20 cm² was used. The transmembrane pressure (TMP) was controlled using an automated valve in the retentate stream. The retentate flow was set to 45 mL/min. The Raman spectrometer was connected via a separate on-line loop with an Minipuls 3 peristaltic pump (Gilson, Villiers le Bel, France). The flow rate in the on-line loop was set to 1 mL/min and monitored using a SLS-1500 flow meter (Sensirion, Stäfa, Switzerland). The on-line loop circulated the process solution from and back into the retentate reservoir container. A detailed scheme of the experimental setup is shown in Fig. 6.3.

During each experiment, the starting solution was prepared and circulated in the CFF system before applying the TMP. For the protein runs, the first ultrafiltration (UF) phase was conducted until a concentration factor of approx. 2 has been reached. The DF phase was initiated by manually attaching the DF buffer. After approx. 3 diafiltration volumes (DVs), the second UF phase was initiated. The process was terminated when 10 g permeate mass have been collected after start of the second UF phase.

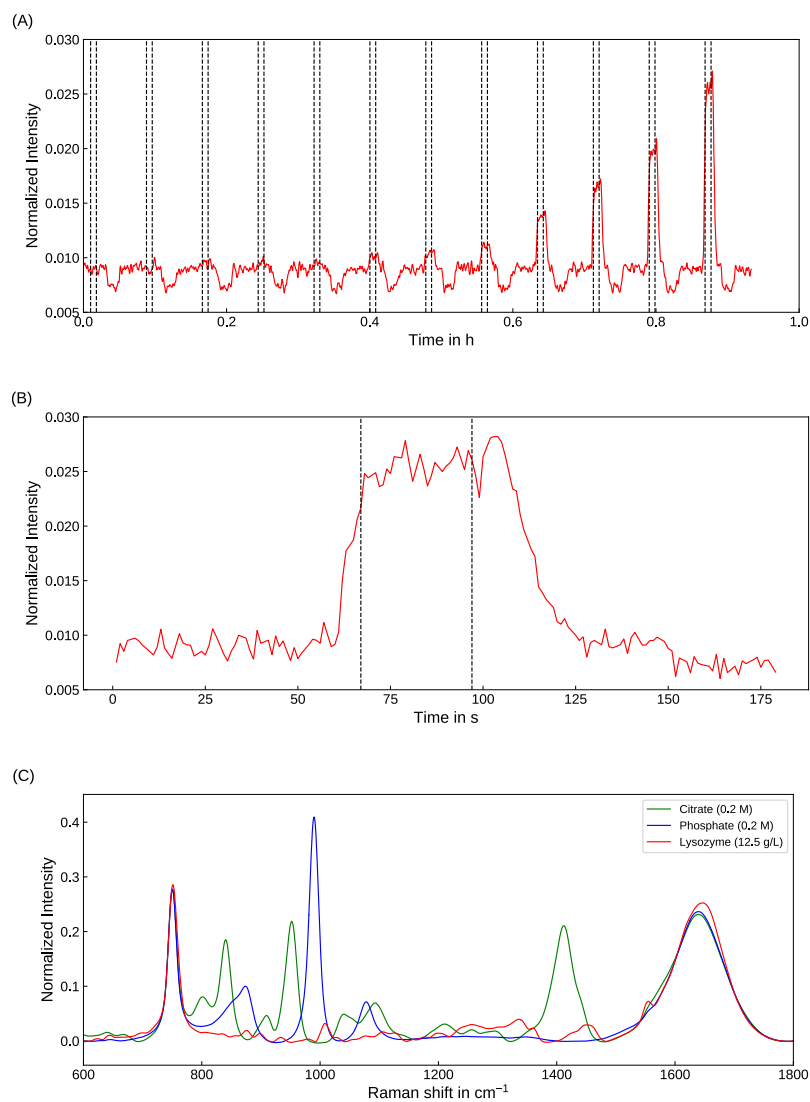


Figure 6.4 Injection sequence of selected samples with increasing lysozyme concentration depicted by the normalized Raman intensity using the 1004 cm^{-1} band after baseline correction as an indicator (A) Dashed vertical lines indicate the time intervals for the static measurements. Zoomed Raman intensity profile for a single injection using the 1004 cm^{-1} band after baseline correction (B). Baseline-corrected Raman spectra for dilution series of lysozyme (red), sodium phosphate pH 7.0 (blue) and sodium citrate pH 5.6 (green) (C).

6.3 Results and discussion

6.3.1 Automating Raman spectroscopy measurements by combining an ÄKTA chromatography system with a Raman spectrometer

To characterize the experimental setup for Raman measurements and evaluate robustness of the system, multiple injection sequences were performed. In Fig. 6.4 (A), the Raman measurements of a lysozyme dilution series are exemplarily depicted. The sequence shows multiple subsequent injections at different lysozyme concentrations with intermittent washing cycles to re-equilibrate the flow cell. The depicted signal at 1004 cm^{-1} was extracted from Raman spectra after preprocessing according to Sec. 6.2.4. After each injection the signal reduces to a stable baseline, supporting the assumption of a sufficiently long washing phase and complete re-equilibration. In Fig. 6.4 (B), a single injection is depicted in detail. The sample is injected into the flow cell using the autosampler. The vertical lines indicate the time interval during which the Raman spectra are collected for the injected sample. During this phase, the flow is halted (0 ml/min) for the collection of the Raman spectra and is thus considered a static measurement. Halting the flow is a strict necessity to keep the injection volume minimal. With the current setup, dynamic measurements would only be possible by applying a flow by the system pump which increases the necessary injection volume. By integration of an external pump in the measurement loop, dynamic measurements would be possible. During the injection and wash-out of the samples, a continuous flow is applied to the analyte causing a small overshoot in the averaged Raman signal. This overshoot is mainly caused by an amplified baseline in the Raman spectra. Noticeably, the 1004 cm^{-1} band after baseline correction is not affected by this phenomenon. Within the measurement interval, the Raman signal intensity is scattering due to the intrinsic variability of the Raman spectrometer in combination with the algorithmic preprocessing. This variability is known to originate from the detector and shot noise (Smulko et al. 2014) which influence both the total intensity and the preprocessing performance. The static measurement phase was configured to last 24 exposures for each sample measured in this study, to ensure stable signals after averaging. In Fig. 6.4 (C), an overlay of Raman spectra for dilution series of sodium phosphate (pH 7), sodium citrate (pH 5.8) and lysozyme are depicted in shades of blue, green and red, respectively. The spectra are shown after preprocessing. In the spectral region between 600 and 1800 cm^{-1} , all analytes show specific Raman profiles with strongly overlapping peaks. While phosphate and citrate dominate the Raman spectra at Raman shifts of 989 cm^{-1} and 1417 cm^{-1} , the Raman bands of lysozyme are comparably smaller. The lysozyme band at 1350 cm^{-1} is the least interfered by the other species in the mixture.

The proposed setup provides an easy way to repurpose chromatography systems for automated spectral acquisition. While other laboratory systems, like high-performance liquid chromatography (HPLC) systems come with similar functionality built-in, the flow cell of the Raman spectrometer was built for preparative lab-scale experiments and hence is not compatible with the tubing dimensions of a HPLC system. Furthermore, the setup allows for measurement with other spectrometers or incorporation of an external pump to allow for dynamic measurements, which would increase representativeness of the acquired data with in-process data.

6.3.2 Calibrating chemometric models off-line by synthetic mixtures

After successful characterization, the presented assembly was used to record Raman data for the monitoring of a UF/DF process using lysozyme as a model protein. To build PLS models for process monitoring, the data used for training and testing require the coverage of the process design space while ensuring statistical independence of all target species (Rolinger, Rüdts, and Hubbuch 2020a).

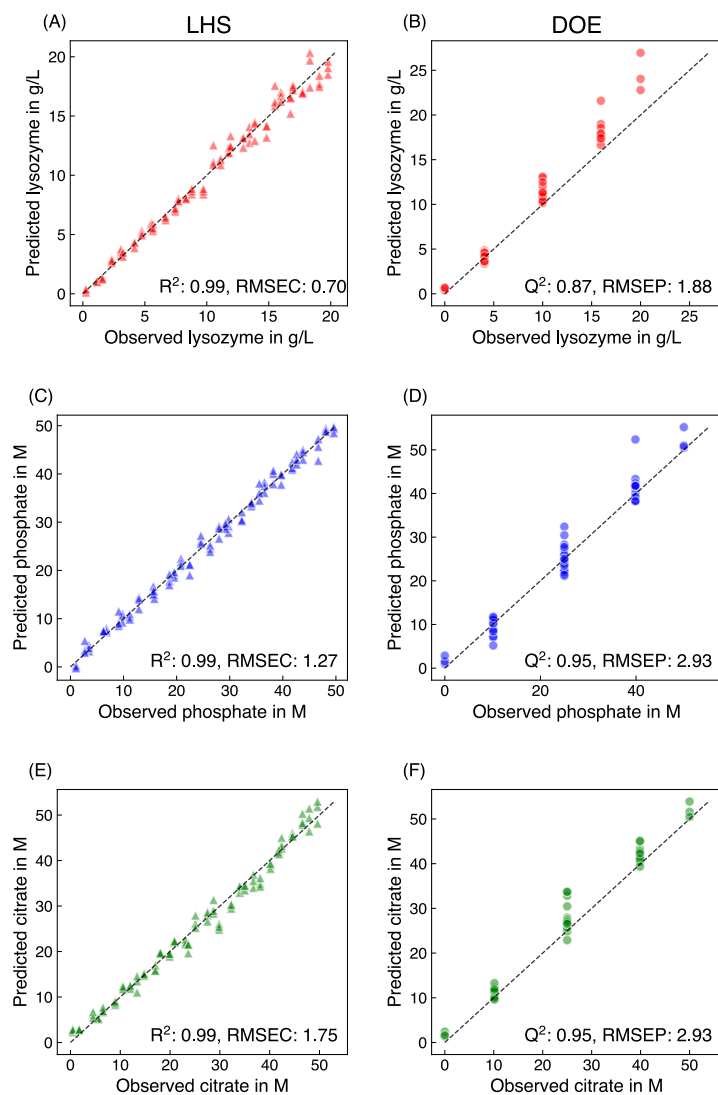


Figure 6.5 Regression errors for calibration datasets LHS and CCD. The errors are displayed for lysozyme, phosphate and citrate from top to bottom. R^2 and $RMSEC$ values refer to the error metrics obtained the data assigned to training, while Q^2 and $RMSEP$ refer to the data assigned to test.

In this case, the target species were sodium phosphate, sodium citrate and lysozyme. For calibration and validation of the PLS models, the LHS and CCD data sets were used, respectively. Fig. 6.2 presents the sample composition, correlation matrices of the target species and preprocessed Raman spectra for the LHS and CCD datasets. The LHS and CCD dataset contain 32 and 18 different sample compositions, respectively, with 3 replicates each. While the LHS data set, as a randomized space-filling design, covers the process design space with minimal target correlation ($R^2 = 0.2$), the CCD dataset perfectly isolates the effects of the target species ($R^2 = 0$).

Fig. 6.5 presents parity plots for training and test data for all three target species. In all cases, the PLS models show good alignment between theoretical and predicted values for training and test data. The lysozyme model was trained in range of 0 to 20 g/L with an R^2 of 0.99. The accuracy for the test set reduces with a Q^2 of 0.87 and a RMSEP of 1.88 corresponding to 2.5-fold increase compared to the training data. This is mainly due to a consistent overestimation of the lysozyme concentration for samples with concentrations above 10 g/L. The reason for this offset is most likely a discrepancy between the training and test set itself as they were prepared using two different lysozyme stock solutions, suggesting that for robust model calibration it would be advisable to repeat the procedure multiple times to include potential variations. Often, chemometric models are validated on random splits of the same dataset (Thakur, Hebhi, and Anurag S. Rathore 2021), which may lead to too optimistic estimation of model errors and hence may negatively affect model performance when used in their intended application. We thus recommend to always record multiple independent data sets to test chemometric models. In this scenario, despite the overestimation of the lysozyme concentrations, the CCD data set additionally evaluates the influence of the other target species which are varied independently from the lysozyme concentration.

The phosphate and citrate model predictions show almost perfect alignment with the theoretical concentrations with RMSEC 1.27 and 1.75 mM, respectively. Contrarily to the lysozyme model, the test errors only increased approx. 2-fold compared to the training errors to RMSEP of 2.93 mM for both species. In comparison to Thakur, Hebhi, and Anurag S. Rathore (2021), we achieve comparable accuracy as has been demonstrate for other common buffer systems like histidine and acetate. For quantification of monoclonal antibody (mAb) concentration, the authors state a RMSEP of 3.25 g/L, while their model was calibrated over a wider design space, spanning 0 to 200 g/L. Noticeably, the changing pH in the range of 5.8 to 7.0 in the analyzed sample within the LHS and CCD data sets did not negatively affect the model performance, even though pH affects the Raman spectra of buffer substances (D. H. Müller, Börger, et al. 2023).

6.3.3 Applying chemometric models to quantify protein and buffer concentrations in ultrafiltration/diafiltration processes

The established PLS models were applied to Raman data, which were recorded on-line through an external loop sourcing from the UF/DF reservoir as depicted in Fig. 6.3. Data acquisition was started once the reservoir and retentate loop were filled with the load material. In total, two runs at starting concentrations of 11.5 and 5.7 g/L were performed. The detailed experimental conditions are listed in Tab 6.1. Fig. 6.6 presents model predictions for Runs 1 and 2 in (A) and (B), respectively. It needs to be noted that for these experiments no off-line reference measurements were available to quantify the actual content of the substances. However, from the predicted trajectories a qualitative evaluation can be done.

The UF/DF runs 1 and 2 can be broadly separated into the UF1, DF and UF2 phase. During UF1 and UF2, the concentrations of phosphate and citrate are expected to remain roughly constant while the concentration of lysozyme is expected to increase. During the DF phase, the lysozyme

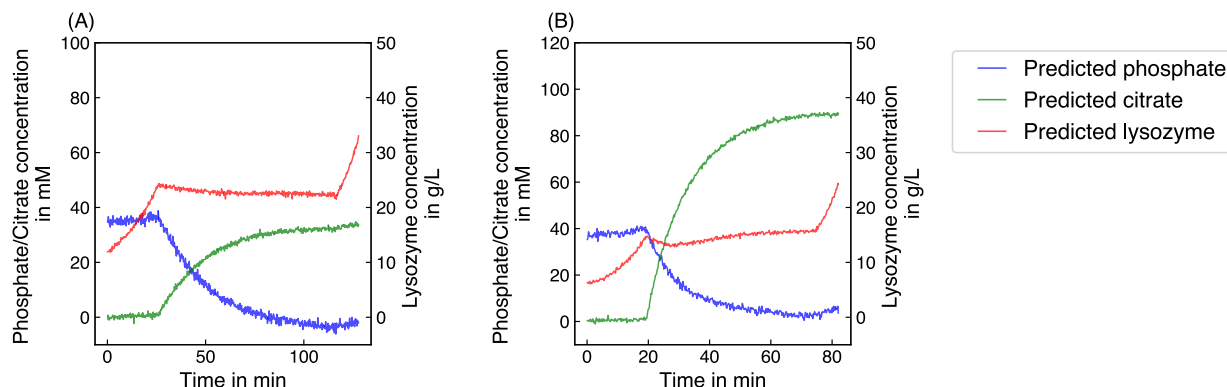


Figure 6.6 Model predictions for in-line data from UF/DF experiments.

concentration is expected to remain constant while a buffer exchange from phosphate to citrate is performed. The three phases and the described process dynamics are well represented in the Raman model predictions. In both runs, the PLS models predict an increase of the lysozyme concentration from initially 11 and 5 g/L to approx. 25 and 15 g/L after UF1 and 30 and 20 g/L after UF2, respectively. Predictions for phosphate and citrate remain mostly constant during UF1 and UF2. During the DF phase, the phosphate concentration quickly decreases from initially 35 and 40 mM until a complete depletion is reached. Inversely, the citrate concentration is predicted to increase and reaches a plateau of 35 and 90 mM at the end of the DF phase in run 1 and 2, respectively. The lysozyme concentration remains mostly stable over the DF phase with minor variations such as an overall decrease in run 1 and a initial drop in run 2. These effects may be caused by product sieving through the UF/DF membrane or the sensitivity of the model to other species. Although the PLS models produce realistic predictions of the investigated analyte concentrations, several limitations can be pointed out. Due to the discrepancy between the off-line and on-line collected data, which could not be entirely eliminated by the direct standardization approach, the phosphate model predicts negative concentrations, and the citrate model likely overestimates the citrate content at the end of the DF phase. This suggests that for successful implementation of Raman spectroscopy for UF/DF monitoring, on-line data should be incorporated into model training. Further, the chosen experimental conditions do not match realistic operating conditions in biopharmaceutical UF/DF processes, where concentrations of up to 200 g/L are typical for formulated mAb solutions (Rischawy et al. 2023). Under these conditions, the Donnan effect causes macroscopic pH shifts and retention of buffer substances (Briskot et al. 2022). These effects were out of scope for this study and the capability of Raman spectroscopy to monitor buffer concentrations influenced by the Donnan effect will be evaluated in future studies.

6.4 Conclusions

This study demonstrates the feasibility of automating Raman spectroscopy measurements using a chromatography system with a minimal sample volume of 490 μ L. A custom software interface enabled seamless integration of the ÄKTA chromatography system with a Raman spectrometer, allowing for efficient data acquisition and chemometric model calibration.

The developed system was evaluated using synthetic mixtures of phosphate- and citrate-buffered lysozyme solutions, with calibration performed through PLS regression models. The PLS models

trained on off-line data demonstrated good alignment with theoretical values, with minor deviations in high-concentration lysozyme samples. When applied in-line to UF/DF processes, the models accurately reflected expected concentration trends of lysozyme and buffer components. However, discrepancies between off-line and on-line data led to limitations in phosphate and citrate predictions, highlighting the need for on-line data incorporation in future model calibration.

Despite these limitations, the system successfully captured the key dynamics of UF/DF processes, demonstrating its potential for real-time monitoring of protein and buffer concentrations. Further research should address the impact of process conditions such as high protein concentrations and Donnan effects to enhance model robustness and applicability in biopharmaceutical settings.

Conflict of interest

The authors declare no conflicts of interest.

Author contributions

Jakob Heyer-Müller and Robin Schiemer both evolved the concepts and methods presented in this manuscript, supervised experimental work, analyzed and interpreted the data, drafted the figures and wrote the final manuscript. Matthias Lopinski developed automation software and performed initial experiments. Franka Willems and Caty Wang both conducted experiments for UF/DF case study.

Acknowledgments

Open Access funding enabled and organized by Projekt DEAL.

Data availability statement

The data supporting this study's findings, as well as the OPC-UA-based communication software, are available from the corresponding author upon reasonable request.

Chapter references

- Abel, Jeff et al. (2018). „A Small-Scale Process for Predicting Donnan and Volume Exclusion Effects During Ultrafiltration/Diafiltration Process Development“. In: *Journal of Pharmaceutical Sciences* 107.5, pp. 1296–1303.
- Brestrich, Nina, Matthias Rüdtt, et al. (02/2018). „Selective protein quantification for preparative chromatography using variable pathlength UV/Vis spectroscopy and partial least squares regression“. In: *Chemical Engineering Science* 176.C, pp. 157–164.
- Briskot, Till et al. (04/2022). „Modeling the Gibbs–Donnan effect during ultrafiltration and diafiltration processes using the Poisson–Boltzmann theory in combination with a basic Stern model“. In: *Journal of Membrane Science* 648, p. 120333.
- Chen, Jingyi et al. (03/2024). „Application of Raman spectroscopy during pharmaceutical process development for determination of critical quality attributes in Protein A chromatography“. In: *Journal of Chromatography A* 1718, p. 464721.
- Dietrich, Annabelle et al. (05/2024). „Raman-based PAT for VLP precipitation: systematic data diversification and preprocessing pipeline identification“. en. In: *Frontiers in Bioengineering and Biotechnology* 12, p. 1399938.
- Feidl et al. (10/2019). „Combining Mechanistic Modeling and Raman Spectroscopy for Monitoring Antibody Chromatographic Purification“. In: *Processes* 7.10, pp. 683–683.
- Hansen, Sigrid K., Babak Jamali, and Jürgen Hubbuch (02/2013). „Selective high throughput protein quantification based on UV absorption spectra“. eng. In: *Biotechnology and Bioengineering* 110.2, pp. 448–460.
- Klaverdijk, Maarten, Marcel Ottens, and Marieke E. Klijn (2025). „Single compound data supplementation to enhance transferability of fermentation specific Raman spectroscopy models“. In: *Analytical and Bioanalytical Chemistry*, pp. 1–12.
- Korepanov, Vitaly I. (10/2020). „Asymmetric least-squares baseline algorithm with peak screening for automatic processing of the Raman spectra“. en. In: *Journal of Raman Spectroscopy* 51.10, pp. 2061–2065.
- Müller, David Heinrich, Marieke Börger, et al. (12/2023). „The Good pH probe: non-invasive pH in-line monitoring using Good buffers and Raman spectroscopy“. en. In: *Analytical and Bioanalytical Chemistry* 415.29, pp. 7247–7258.
- Nikzad-Langerodi, Ramin and Erik Andries (2021). „A chemometrician’s guide to transfer learning“. en. In: *Journal of Chemometrics* 35.11, e3373.
- Rischawy, Federico et al. (09/2023). „Modeling of biopharmaceutical UF/DF from laboratory to manufacturing scale“. In: *Computers & Chemical Engineering* 177, p. 108337.
- Rolinger, Laura, Jürgen Hubbuch, and Matthias Rüdtt (02/2023). „Monitoring of ultra- and diafiltration processes by Kalman-filtered Raman measurements“. In: *Analytical and Bioanalytical Chemistry* 415.5, pp. 841–854.
- Rolinger, Laura, Matthias Rüdtt, Juliane Diehm, et al. (02/2020). „Multi-attribute PAT for UF/DF of Proteins–Monitoring Concentration, particle sizes, and Buffer Exchange“. In: *Analytical and Bioanalytical Chemistry* 412.9, pp. 2123–2136.
- Rolinger, Laura, Matthias Rüdtt, and Jürgen Hubbuch (04/2020a). „A critical review of recent trends, and a future perspective of optical spectroscopy as PAT in biopharmaceutical downstream processing“. In: *Analytical and Bioanalytical Chemistry* 412.9, pp. 2047–2064.
- Romann, Patrick et al. (11/2022). „Advancing Raman model calibration for perfusion bioprocesses using spiked harvest libraries“. In: *Biotechnology Journal* 17.11, e2200184.
- Santos, Rafael M. et al. (2018). „Monitoring mAb cultivations with in-situ raman spectroscopy: The influence of spectral selectivity on calibration models and industrial use as reliable PAT tool“. In: *Biotechnology Progress* 34.3, pp. 659–670.
- Schiemer, Robin, Matthias Rüdtt, and Jürgen Hubbuch (2024). „Generative data augmentation and automated optimization of convolutional neural networks for process monitoring“. In: *Frontiers in Bioengineering and Biotechnology* 12.
- Smulko, Janusz M et al. (02/2014). „Anatomy of Noise in Quantitative Biological Raman Spectroscopy“. en. In: *Bioanalysis* 6.3, pp. 411–421.
- Thakur, Garima, Vishwanath Hebhi, and Anurag S. Rathore (05/2021). „Near Infrared Spectroscopy as a PAT tool for monitoring and control of protein and excipient concentration in ultrafiltration of highly concentrated antibody formulations“. en. In: *International Journal of Pharmaceutics* 600.December 2020, p. 120456.
- Wang, Jiarui et al. (12/2023b). „In-line product quality monitoring during biopharmaceutical manufacturing using computational Raman spectroscopy“. In: *mAbs* 15.1, p. 2220149.
- West, Jay M. et al. (2021). „Process analytical technology for on-line monitoring of quality attributes during single-use ultrafiltration/diafiltration“. In: *Biotechnology and Bioengineering* 118.6, pp. 2293–2300.
- Wold, Svante, Michael Sjöström, and Lennart Eriksson (10/2001). „PLS-regression: a basic tool of chemometrics“. en. In: *Chemometrics and Intelligent Laboratory Systems* 58.2 (2), pp. 109–130.

Development of Raman spectroscopy and machine learning methods for protein aggregate quantification: Application to BSA in chromatographic processes.

Jakob Heyer-Müller^{1,+}, Robin Schiemer^{1, +}, Lars Robbel², Michael Schmitt², and Jürgen Hubbuch¹

¹ Institute of Engineering in Life Sciences, Section IV: Biomolecular Separation Engineering, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

² CSL Innovation GmbH Germany, Michelbacher Straße, 35041 Marburg, Hessen, Germany

⁺ contributed equally

Abstract

Protein aggregation poses a significant risk to biopharmaceutical product quality, as even minor amounts of oligomeric species can compromise efficacy and safety. Rapid and reliable detection of protein aggregates thus remains a major challenge in biopharmaceutical manufacturing. Although traditional offline methods such as size-exclusion chromatography provide accurate results, their inherent time delays limit real-time process control capabilities. Consequently, there is an urgent scientific need for inline analytical techniques capable of selectively quantifying protein monomers and aggregates in real-time to facilitate immediate corrective actions and enhance overall process robustness. Raman spectroscopy, as a tool for a Process Analytical Technology application, is especially suitable due to its molecular specificity, rapid data acquisition, and compatibility with aqueous solutions commonly used in biopharmaceutical manufacturing. Addressing this need, this study establishes a Raman spectroscopy-based strategy for the selective detection and quantification of monomeric and aggregated forms of a model protein (bovine serum albumin). Controlled stress conditions were applied to generate aggregated species reproducibly, and a Latin Hypercube sampling design was used to independently vary protein concentration and aggregate fraction, ensuring that observed spectral effects were attributable to aggregation rather than concentration differences. Furthermore, spectral markers identified in spectra acquired from multiple chromatographic runs were qualitatively compared with off-line reference measurements from size-exclusion chromatography. This limitation in real-time applicability was circumvented by chemometric machine learning approaches. The use of convolutional neural networks enabled the selective quantification of the protein monomers and aggregates and delivered superior predictive performance and robustness across cross-validation, independent testing and synthetic perturbation scenarios compared to traditional chemometric approaches. Collectively, these results demonstrate that the selected Raman spectral markers, combined with advanced chemometric modeling, enable reliable, real-time monitoring of protein size variants in biopharmaceutical downstream processes.

7.1 Introduction

Protein aggregation is a critical quality attribute (CQA) and a major challenge in the manufacturing of biopharmaceuticals. Aggregates can form throughout the product lifecycle - including upstream production, downstream purification, formulation, storage (Vázquez-Rey and Lang 2011; Roberts 2014; Pham and Meng 2020) - potentially compromising both the safety and efficacy of therapeutic proteins (Rosenberg 2006; Moussa et al. 2016). To address this issue, there is an increasing need for real-time monitoring of protein structural integrity under application-relevant conditions. Implementing such monitoring, particularly in downstream processing, is essential to ensure consistent product quality and meet regulatory requirements.

The introduction of the process analytical technology (PAT) initiative by the United States Food and Drug Administration (FDA) aimed to enable real-time, in-line measurement and control of critical process parameters and quality attributes (FDA 2004). Univariate sensors are commonly applied to monitor critical process parameters in upstream and downstream unit operations. Modern PAT frameworks integrate spectroscopic tools and chemometric models (Rolinger, Rüdts, and Hubbuch 2021) to enable real-time impurity quantification (Brestrich, Sanden, et al. 2015; Capito, Skudas, Kolmar, et al. 2015; Wei et al. 2022; J. Chen et al. 2024), protein structure assessment (Rüdts, Vormittag, et al. 2019), reaction monitoring (Schiemer, Rüdts, and Hubbuch 2024) or precipitation behavior (Dietrich et al. 2024). These innovations support a shift toward Quality by Design (QbD)

CHAPTER 7. MONITORING PROTEIN AGGREGATES IN DSP

approaches in biopharmaceutical production (Glassey et al. 2011).

Early studies using vibrational spectroscopy techniques, such as Raman and FTIR, reported changes in the amide I and III bands and the appearance of intermolecular β -sheet structures for proteins subjected to heat stress (Alizadeh-Pasdar, Nakai, and Li-Chan 2002; Militello et al. 2004). It has been shown that environmental stresses (e.g. pH, temperature, ionic strength) induce specific conformational changes that can be spectroscopically monitored and are commonly associated with protein aggregation (Ettah and Ashton 2018). Later studies reported similar spectral changes in the amide bands for different natural occurring proteins as well as for common therapeutic proteins such as immunoglobulin Gs (IgGs) (Gómez de la Cuesta, Goodacre, and Ashton 2014; Lewis et al. 2014; Barnett et al. 2015; Zhou et al. 2015). Aside from the changes in the amide bands, multiple studies also report the shifts in the vibrational bands of aromatic amino acid side chains like tryptophan, tyrosine or phenylalanine which undergo changes in their solvation during protein aggregation (Barnett et al. 2015; Zhou et al. 2015). However, the vibrational spectra are not only affected by protein aggregation but also by other environmental factors such as pH, protein concentration or excipients (Ota et al. 2016; Sato et al. 2023; Makki et al. 2021) as well as intrinsic fluorescence and Rayleigh scattering of the protein analytes requiring targeted pre-processing of the collected Raman spectra (Rolinger, Rüdts, and Hubbuch 2020a; Dietrich et al. 2024).

Due to fine differences between the native and the aggregated proteins and the influence of environmental factors on the Raman spectra, multivariate data analysis is commonly used to monitor purification processes and in particular protein aggregation during process development and early formulation stages (Gómez de la Cuesta, Goodacre, and Ashton 2014; C. Zhang et al. 2019; Wei et al. 2022; Jiarui Wang et al. 2023b). While these methods often indicate a correlation between spectral features and aggregation state, the specific spectral changes directly linked to aggregation and the detection limits for low aggregate levels remain uncertain. Hence, despite the reported evidence of Raman effects corresponding to protein aggregation, a reliable method for real-time quantification of protein aggregates in downstream process operations using structural markers is not yet fully established.

To be used for real-time process monitoring via spectroscopic techniques, multivariate regression models have been established for various separation processes such as precipitation (Dietrich et al. 2024), chromatography (Rüdts, Brestrich, et al. 2017), or filtration (Rolinger, Hubbuch, and Rüdts 2023). Raman spectroscopy has previously been used in bioseparations, but has been shown to have limited accuracy due to the inherent variation in spectral signals (Feidl et al. 2019; Rolinger, Hubbuch, and Rüdts 2023; Rolinger, Rüdts, and Hubbuch 2021). Previous studies have shown that these limitations can be circumvented when combining chemometric models with mechanistic process models (Rolinger, Hubbuch, and Rüdts 2023; Schiemer, Weggen, et al. 2023) or using machine learning models such as convolutional neural networks (CNNs) (Schiemer, Rüdts, and Hubbuch 2024).

To address these limitations, in this study, protein aggregates are generated under controlled conditions, using bovine serum albumin (BSA) as a model system, to isolate and characterize their spectral signatures. The objective is to identify reliable spectral markers capable of distinguishing between native and aggregated BSA species. These markers should enable a quantitative assessment of monomer and aggregate content. Subsequently, these markers are evaluated for their suitability in in-line monitoring of the separation of BSA monomers and dimers via anion-exchange chromatography

(AEX). Finally, the recorded spectral data are used to train chemometric partial least squares (PLS) and CNN models that enable real-time discrimination and quantification of BSA size variants.

7.2 Materials and methods

7.2.0.1 Forced degradation by heat incubation

Forced degradation was performed to generate BSA aggregates and to identify Raman spectral features associated with aggregation. A BSA stock solution at a concentration of approximately 60 g/L was prepared by dissolving lyophilized BSA (Sigma-Aldrich, St. Louis, Missouri, United States) in ultrapure water (PURELAB Ultra, ELGA LabWater, Celle, Germany). The stock solution was subsequently diluted 1:2 using a mixture of 500 mM NaCl solution and ultrapure water to prepare BSA samples with final NaCl concentrations of 0, 50, 100, and 250 mM, maintaining a total protein concentration of approximately 30 mg/mL. Based on these experiments, a NaCl concentration of 150 mM was identified as providing reproducible aggregation kinetics and was therefore selected for all subsequent studies. Aliquots of these BSA solutions were transferred into 1 mL Eppendorf tubes (Eppendorf SE, Hamburg, Germany), which were incubated at 65 °C for durations of 0, 30, 60, 90, and 120 minutes in a stationary state (without agitation) using a heat block (Eppendorf SE, Hamburg, Germany). Each condition was prepared and incubated in triplicate. Following incubation, samples were filtered using 0.2 µm cellulose acetate syringe filters (Sigma-Aldrich, St. Louis, Missouri, United States). Raman spectroscopy measurements were conducted offline and automated using a custom-built automated Raman measurement platform (ARMP) (described in Sec. 7.2.0.4). Residual samples were diluted 1:100 and analyzed by ultrahigh performance - size exclusion chromatography (UHP-SEC) for reference analytics.

7.2.0.2 Mixing studies

Mixing studies were conducted to evaluate whether previously identified spectral features are sustained while varying the total protein concentration and aggregate content in an independent manner. Therefore, an aggregate-enriched BSA solution was prepared by heat incubation. A 300 mL BSA stock solution was prepared (225 mL centrifuge tubes, Thermo Scientific, Waltham, USA) at a protein concentration of 30 g/L and a NaCl concentration of 50 mM. All subsequent experiments were conducted at a NaCl concentration of 150 mM, chosen as it provided reproducible aggregation kinetics, while heating time was systematically varied to control aggregate formation. A 200 mL aliquot was incubated at 65 °C for 24 hours (Incubated orbital shaker, Thermo Scientific, Waltham, USA) to induce aggregation while the rest was left unstressed and stored at 4 °C until further use. After incubation, the stressed solution was filtered using a 0.2 µm cellulose acetate vacuum filtration membrane (Sigma-Aldrich, St. Louis, Missouri, United States).

The methodology resulted in two BSA stock solutions with concentrations of 30 g/L and 19.85 g/L, containing 25 % and 50 % aggregates, respectively, as determined by the ratio of the aggregate peak area to the total peak area in the UHP-SEC analysis. This aggregate yield in the stressed stock is considerably lower than for the forced degradation experiments. This is believed to be due to the scaling of the volume from 1 mL to 200 mL. The lower aggregate yield restricted the covered aggregate range in the mixing study.

For the remainder of this manuscript, we will refer to these as the unstressed stock and stressed stock. To prepare samples where the protein concentration and aggregate fraction are varied independently, a latin hypercube sampling (LHS) design with two dimensions, with 48 samples, was used. The LHS was scaled to cover the ranges from 0 to 15 g/L total protein concentration and 25 to 50 %

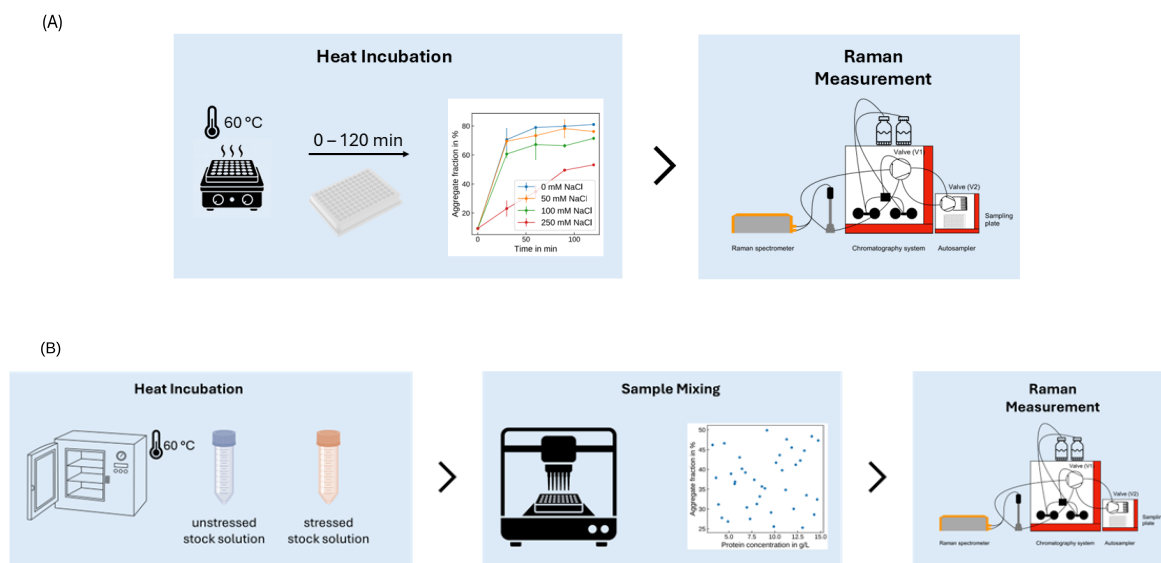


Figure 7.1 Schematic representation of the forced degradation (A) and mixing study workflows (B). (A) Heat-induced BSA aggregation was generated by incubating samples at 60 °C for 0–120 min at defined NaCl concentrations, followed by offline Raman spectroscopy and UHP-SEC reference analytics. (B) To assess spectral features under independently varied protein concentration and aggregate levels, unstressed and heat-stressed BSA stock solutions were combined according to a Latin hypercube design spanning 0–15 g/L protein and 25–50 % aggregates, with Raman spectra and UPSEC data collected for all conditions.

aggregates. Each condition was prepared in duplicate at a total sample volume of 800 μ L, resulting in 96 samples in total. For each sample a Raman spectrum was collected using the ARMP, with 20 acquisitions averaged in the mixing experiments to improve the signal-to-noise ratio. The aggregate content was determined via UHP-SEC. Comprehensive information on sample compositions, mixing ratios, concentrations, and the replicate structure is provided in Table S1. The corresponding experimental workflow is depicted in Figure 7.1.

7.2.0.3 Bind-elute anion-exchange chromatography experiments

To generate process data for the separation of BSA monomers and aggregates, bind-elute AEX chromatography experiments with in-line Raman measurements were conducted using a 5 mL Eshmuno Q column (Merck KGaA, Darmstadt, Germany) on an Äkta Pure system (Cytiva, Uppsala, Sweden). Unstressed BSA stock solutions at a concentration of approx. 15 g/L and 25 % aggregate content, prepared in 20 mM Tris, pH 8, were used as feedstock. The column was equilibrated for 5 column volumes (CVs) using 20 mM Tris, pH 8 prior to sample loading. Following sample loading, the column was washed for 5 CV using equilibration buffer. Gradient elution was performed at different lengths ranging from 0 to 500 mM NaCl, followed by a strip phase at 1 M NaCl for 5 CVs. The column was subsequently regenerated using 1 M NaOH for 5 CVs and reequilibrated using equilibration buffer. During the elution phase using a flow rate of 1 ml/min, fractions were collected at intervals of 200 μ L, and Raman spectra were recorded at an exposure time of 500 ms at a laser power of 495 mW resulting at 24 acquisitions per fraction. In contrast to the spectra generated during the mixing studies, where 20 acquisitions were averaged to improve the signal-to-noise ratio, single 500 ms acquisitions were used during in-line AEX chromatography

Table 7.1 Experimental conditions for bind-elute AEX chromatography experiment used in in-line Raman monitoring of BSA monomer and aggregate separation.

Run ID	Gradient length (CV)	Load density (mg/ml resin)	Total feed concentration (mg/ml)	HMWC content (%)	Data set (-)
Batch 1 5CV	5	15	15	25.33	Train
Batch 1 15CV	15	22.5	15	25.33	Train
Batch 2 10CV	10	30	15	15.67	Train
Batch 1 10CV	10	17.5	15	25.33	Test
Batch 1 Step	Step	22.5	15	25.33	Test
Batch 2 5CV	5	22.5	15	15.67	Test
Batch 2 15CV	15	45	15	15.67	Test

to preserve temporal resolution. These single acquisitions were subsequently averaged over the time intervals corresponding to each UPLC-analyzed fraction. The Raman spectrometer was incorporated into the flow path between the outlet valve and the fractionator. In total, 7 AEX experiments were performed using different loading densities, gradient lengths and feedstocks which were prepared from two different batches. An overview of the experimental conditions is presented in Tab. 7.1.

7.2.0.4 Automated Raman measurement platform

Off-line Raman spectroscopy measurements were conducted using an in-house developed system referred to as ARMP. The platform consists of an Äkta purifier system equipped with an autosampler A-905 combined with a Raman spectrometer. The Raman BioReactor BallProbe, inserted a Raman flow cell with a dead volume of 240 μ L (both MarqMetrix, Seattle, Washington, USA), was connected to a HyperFlux™ PRO Plus 785 spectrometer, operated by SpectralSoft 3.2.600.1 software (Tornado Spectral Systems, Mississauga, Ontario, Canada). The ARMP automatically injects samples into the Raman flow cell and collects Raman spectra using predefined configurations. For technical details about the ARMP, we refer the reader to (Heyer-Müller et al. 2025).

7.2.0.5 Raman spectroscopy - acquisition settings

All measurements were conducted using the maximum laser power of 495 mW. For the forced degradation experiments and the mixing studies, the exposure time was set to 1200 ms per acquisition. Each collected spectrum was averaged from 20 acquisitions. For in-line measurements during AEX chromatography experiments, the exposure time was set to 500 ms per acquisition. The raw Raman signal was treated by a Y-correction and a cosmic ray filter as directly implemented in the operating software SpectralSoft. The Raman spectrometer was calibrated using a white light source at regular intervals.

7.2.0.6 Ultrahigh performance - Size exclusion chromatography

To enable precise determination of size variant distributions, samples were analyzed via UHP-SEC. Therefore, a TSKgel SuperSW mAb HTP column (4 μ m particle size, 4.6 $\times$ 150 mm) operated at a flow rate of 0.3 mL/min was used. The mobile phase consisted of 15 mM sodium phosphate buffer at pH 6.2. Before injection, all samples were diluted to achieve a protein concentration within the range of 0–1 mg/mL. Analysis was conducted using a Vanquish UHPLC system controlled by Chromeleon software (version 7.2) (both Thermo Fisher Scientific, Waltham, MA, USA).

7.2.1 Data analysis

All data analysis and computations were performed in Python 3.12.7.

7.2.1.1 Spectral preprocessing

Raman spectral processing consisted of multiple steps, including cropping, normalization, baseline and background correction, smoothing, and derivative computation. Each operation and its rationale are detailed below, with processing chains optimized based on empirical testing across different data types.

All spectra were first cropped to $500\text{--}3250\text{ cm}^{-1}$, followed by normalization to the water band at 3250 cm^{-1} . Truncation was performed to remove irrelevant spectral regions and improve the quality of correction steps downstream. The lower bound of 500 cm^{-1} was chosen as smaller wavenumbers solely contain baseline drift and bands stemming from the sapphire glass built into the probe head. The upper bound of 3250 cm^{-1} corresponds to the maximum of the water band in the covered bandwidth (Furić, Ciglenečki, and Čosović 2000) and was hence used for normalization. Normalization was done to compensate for small variations in total intensity potentially induced by turbidity (Dietrich et al. 2024). Subsequently, background subtraction was performed. For forced degradation and mixing experiments, a water spectrum was subtracted. For AEX experiments, a buffer spectrum originating from the start of the respective elution phase was subtracted. Background subtraction was performed to remove the overlapping water and buffer bands and enable better visualization of the spectral effects. Furthermore, baseline correction was performed using a Whittaker filter through the derivative peak-screening asymmetric least squares algorithm (DERPSALSA) algorithm, as implemented in the *pybaselines* library (v. 1.1.0), with parameters $\lambda = 10^5$, $k = 0.02$, and $d = 2$ (Erb n.d.). Finally, the Savitzky-Golay filter as implemented in *scipy* (v. 1.14.1) was employed both for spectral smoothing and derivative computation to enhance subtle spectral variations. Here, a second-order polynomial and a window size of 21 was used in all cases. In the following, spectral data are displayed either after smoothing only or after second derivative computation.

7.2.1.2 Evaluation of spectral bands and correlation

To evaluate the changes of spectral bands and their correlation with the aggregate content, multiple spectral markers or metrics were used throughout this study.

It is known that the amide I and III vibrations carry the most structural information of proteins which are located at $1500\text{--}1800\text{ cm}^{-1}$ and $1200\text{--}1400\text{ cm}^{-1}$, respectively Lippert, Tyminski, and Desmeules 1976; R. W. Williams 1986; Rygula et al. 2013. While the amide I region is largely overlapped with the water bending vibration, the amide III region is largely unaffected by water and depending on the used buffer system free of other interfering species. In addition, the Fermi doublet of tryptophan side chains is known to be located at the frequencies of approx. 1340 and 1360 cm^{-1} Rygula et al. 2013 and to provide a informative marker of local solvent environment Harada, Miura, and Takeuchi 1986; Eisenberg and Juszczak 2012. Given the local superposition of the amide III and tryptophane vibrations, we chose to observe the ratio of the intensities at 1341 and 1320 cm^{-1} as the most prominent bands in this region. We will refer to this intensity ratio as the amide III ratio for the remainder of this manuscript.

To compress the information contained in the amide I band into a single spectral marker, the center of mass was calculated as the intensity-weighted average wavenumber within the chosen spectral range ($1500\text{--}1800\text{ cm}^{-1}$),

$$\bar{\nu} = \frac{\sum_i I_i \nu_i}{\sum_i I_i} \quad (7.1)$$

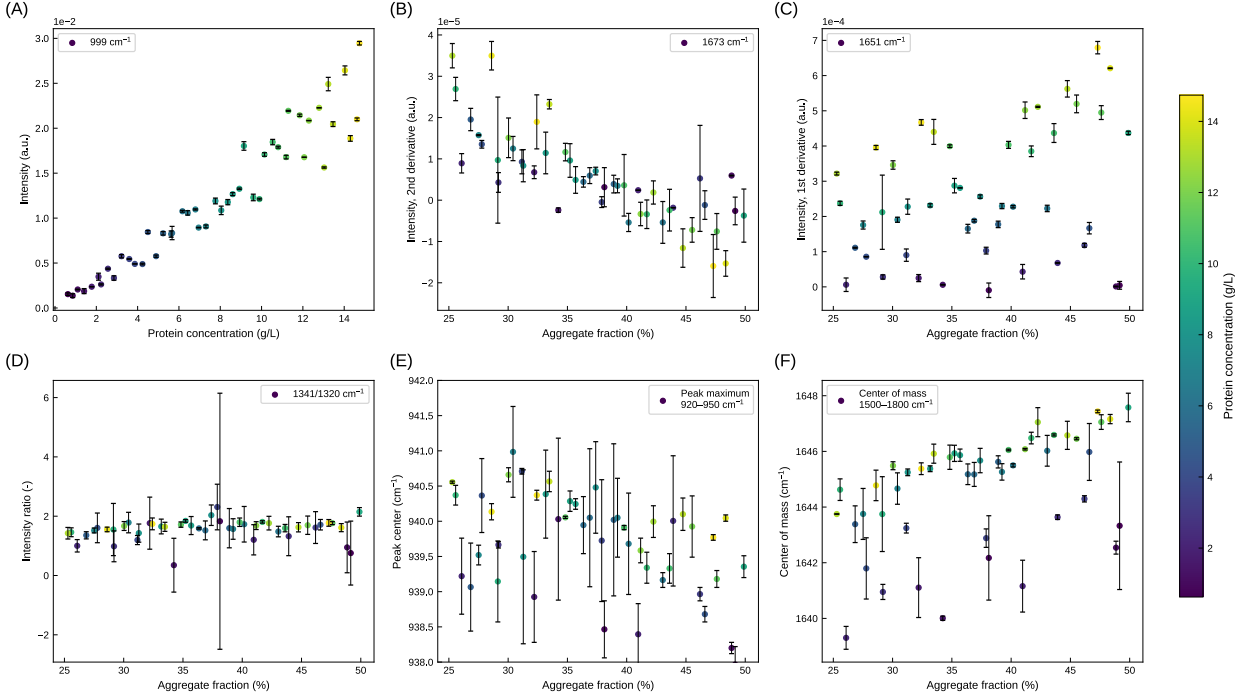


Figure 7.2 Spectral features observed in offline-recorded Raman spectra of BSA obtained during mixing studies, preprocessed as described in Section 7.2.1.1. Raman intensities, peak shifts, and band ratios are plotted as functions of total protein concentration (g/L), aggregate fraction (%), and high molecular weight component (HMWC) content (%). Error bars represent standard deviations of replicate measurements. Color scale from dark to light indicates increasing protein concentration (g/L).

where I_i and ν_i are the intensity and wavenumber at point i , respectively. The center of mass is a common marker summarizing observed spectral shifts in a easy to comprehend manner. To improve the precision of the calculation, the acquired spectra were interpolated to a resolution of 1000 steps in the given spectral range.

Finally, the signal-to-noise ratio (SNR) was used as correlation metric between the spectral intensity and the aggregate content. The SNR was calculated as

$$SNR = \frac{\text{Var}(\mathbf{x}_\lambda \hat{\beta})}{\hat{\sigma}^2} \quad (7.2)$$

with $\hat{\beta}$ being the regression coefficients of a univariate linear model of type $\mathbf{y} = \mathbf{x}_\lambda \beta + \epsilon$ with normally distributed errors ϵ and $\hat{\sigma}^2$ being the residual variance of linear model for wavenumber λ according to (Soch and Allefeld 2018). To calculate SNR for the Raman spectroscopy data using Eq. (7.2), a linear model was built for each wavenumber individually. A spectral feature is considered to be correlated with aggregate content if it shows high SNR. Low values indicate a high degree of noise and no linear correlation. For all spectral markers in the mixing study (Fig. 7.2), error bars represent the standard deviation of two independent replicates, each derived from an averaged spectrum comprising 20 acquisitions.

7.2.1.3 Regression modelling

Data split Regression models were calibrated using in-line collected data from AEX experiments with their respective assignments to training and test sets presented in Tab. 7.1. Raman spectra were preprocessed as laid out in Sec. 7.2.1.1 and used as inputs to the regression models. As outputs, the absolute concentrations of monomers and aggregates were chosen as determined by UHP-SEC for all collected fractions. A leave-one-batch-out (LOBO) scheme was chosen for cross-validation for all regression models. The LOBO cross-validation iterates through all experiments included in the training set, leaving each of them out once. The LOBO scheme has been found to provide more realistic estimates of prediction errors than more common K-fold cross-validation schemes and was hence selected for model evaluation (Rolinger, Rüdts, and Hubbuch 2020a; Dietrich et al. 2024).

Partial least squares regression PLS regression models were calibrated as single-output (PLS1) and multi-output (PLS2) models. To improve model accuracy, additional preprocessing was applied to the input data. Additional preprocessing involved truncating the Raman spectra to 600-1800 cm^{-1} , followed by Savitzky–Golay (SG) filtering with a window size of 15 and varying derivative orders (0, 1, or 2). Finally, the input data was mean-centered. The number of latent variables and the SG derivative order were optimized during cross-validation by grid-search in the ranges 2-14 and 0-2, respectively. In total, 8 different PLS model variants were calibrated. For each combination of preprocessing operations, a PLS1 and PLS2 model was calibrated. Aside from the previously mentioned preprocessing operations, there are two PLS variants which are calibrated using synthetic data. For more details on the generation of these synthetic data, we refer to Sec. 7.2.1.3

Convolutional neural network CNN model were calibrated as multi-output models, predicting the monomer and aggregate concentrations simultaneously. The CNNs were implemented in *tensorflow* (v. 2.11). The model architecture and training methodology were similar to Schiemer, Rüdts, and Hubbuch 2024. The model architecture consisted of three convolutional layers with defined kernel sizes of 15, 7 and 3 and a rectified linear unit (ReLU) activation function. Average pooling with a pool size of 2 was employed after each convolutional layer. Each convolutional layer was configured to use 5 kernels. After the convolutional block, a flattening operation was implemented to project all kernel outputs on a single vector. A fully connected (FC) layer with 50 units and a ReLU activation function was used as the subsequent layer. Finally, the output layer was configured to use 2 FC units and a ReLU activation function.

CNN models were calibrated with and without synthetic data augmentation, which will be denoted CNN and CNN-Aug for the remainder of this manuscript. For non-augmented CNN and the augmented CNN, training was conducted over 500 and 50 epochs, respectively, with a batch size of 32 using the mean squared error (MSE) as the loss function. The target variables were standardized using a min-max scaler to ensure equal contribution of both targets despite absolute concentration differences. The learning rate was initially set to 0.001 and a learning rate scheduler was configured to halve the learning rate if the validation loss was not reduced for 5 consecutive epochs. The non-augmented CNN used the original training data as input data. To be able to compute the validation loss at every epoch of the training cycle, the training data was randomly divided into calibration and validation set, reserving 25% of the original training data points for calculating the validation loss. In case of the augmented CNN, data augmentation was employed to generate a set of 10000 synthetic data points which served for calibration. The validation loss was calculated using 100% of the original training data. For the details of the data augmentation method, we refer to Sec. 7.2.1.3.

Data Augmentation Data augmentation was used to enhance model robustness, utilizing the local subset augmentation (LSA) technique introduced in Schiemer, Rüdts, and Hubbuch 2024. In short, the LSA technique extracts the theoretical pure component spectra from a subset of the training data and generates synthetic spectra by recombining the pure component spectra. A local subset size of 39 was used in all instances. The standardized Euclidean distance as given by $(d(x, y) = \sqrt{\sum_{i=1}^n \frac{(x_i - y_i)^2}{S_i^2}})$ was used as a metric to select the local subset. Random white noise with a standard deviation of 0.0001 intensity counts and normally-distributed shifts along the wavenumber axis with a standard deviation of 0.25 cm^{-1} were further introduced to the synthetic spectra.

Error metrics Model performance was assessed using the R^2 and root mean squared error (RMSE) metrics. In all cases, these metrics are reported separately for the monomer and aggregate species. Additionally, R^2 and RMSE are provided for the aggregate fraction as determined by $v_{\text{agg}} = c_{\text{agg}} / (c_{\text{agg}} + c_{\text{mono}})$, where c_{mono} and c_{agg} denote the absolute concentration of monomer and aggregate species given in g/L, respectively. To avoid division by small values, v_{agg} was calculated only for samples where the denominator summed to values greater than 0.1 g/L . The metrics for cross-validation were calculated based on the predicted values for the iterations when the respective experiments were assigned to the validation set.

Augmentation Test To assess the selectivity of the evaluated models, an augmentation test was conducted using the LSA technique described in Sec. 7.2.1.3 to evaluate whether the models correctly predict monomer and aggregate combinations which were not covered in the experimental data. To provide a challenging data set, the synthetic spectra were generated based on the experimental data originating from the test set. The LSA algorithm was used in the same configuration as laid out in Sec. 7.2.1.3. The monomer and aggregate combinations covered in the augmentation test were sampled from a uniform distribution in the ranges $[0, 20]$ and $[0, 5]$ for the monomer and aggregate species, respectively. In total, 1000 random samples were drawn for which synthetic spectra were generated. The calibrated models were finally used to predict the monomer and aggregate concentrations from the synthetic spectra and the model accuracy was evaluated based on the sampled values as the ground truth. The distribution of the sampled values is visualized in Fig. E.1. This methodology makes use of the test set data by diversifying the obtained data through synthetic amplification. Further, it enables a more systematic evaluation of the model selectivity for the individual species.

Test set rotation test To further compare the robustness and accuracy of the different model variants, we conducted a test set rotation based on a Leave-P-Batches-Out scheme with $P=4$. With 7 experiments in total, this yielded 35 different train-test splits. We conducted the rotation test with the best performing PLS model and the two different CNN models, one without and one with data augmentation. In each rotation, the respective model was calibrated using the selected training batches. For both model types, the methodology introduced in Sec. 7.2.1.3 and 7.2.1.3 was used. For the PLS model, the number of components and the derivative order were fixed to optimal setting found for the training-test split used in the rest of this study, i.e. no individual optimization was performed for each rotation. The optimal number of components was 2 and 14 and the order of derivative was 0 and 1 for the monomer and aggregate species, respectively, for the Crop-SG-PLS1 model.

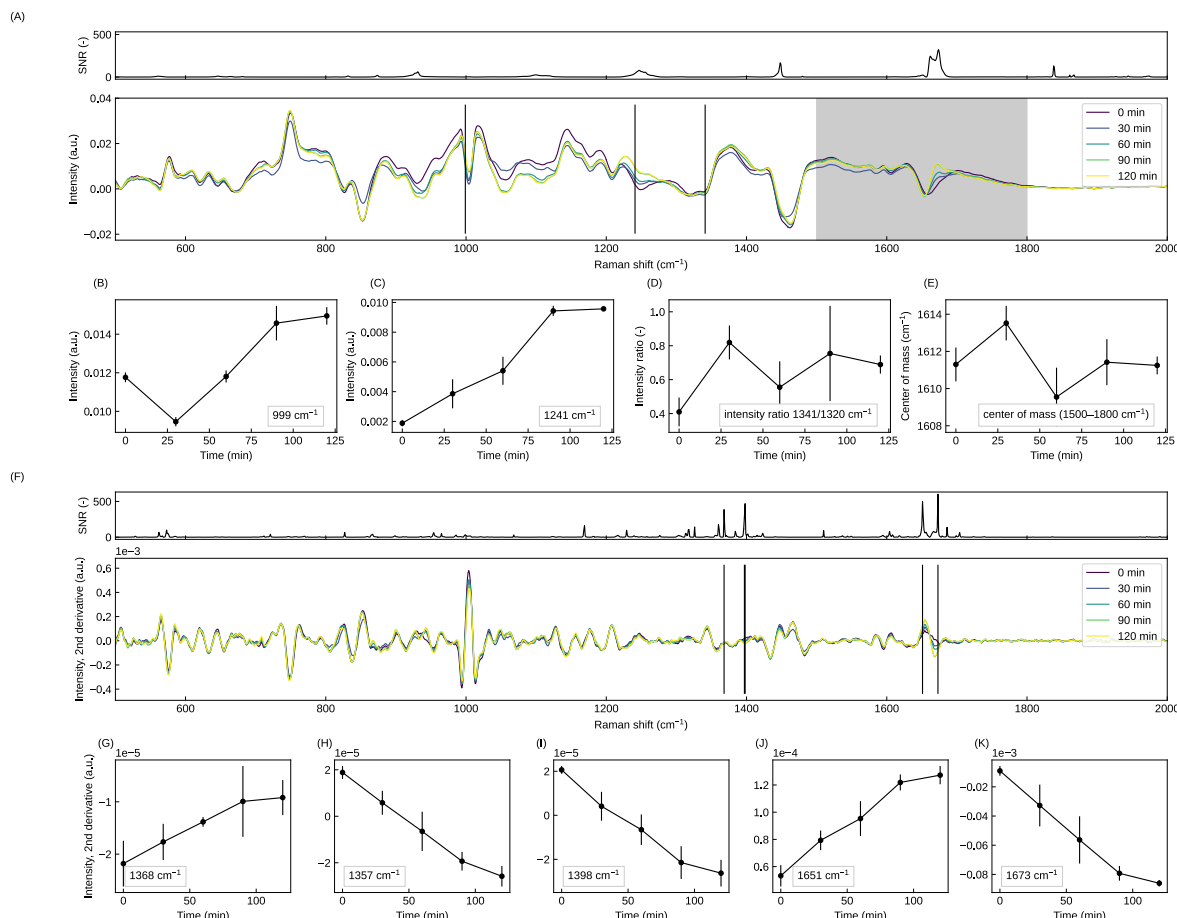


Figure 7.3 Results of the forced degradation experiment. Off-line recorded Raman spectra and SNR for each wavenumber (A). Spectra are shown after preprocessing as described in Sec. 7.2.1.1. The color gradient reflects the increasing duration of heating. Intensities at 999, 1241, the ratio 1341/1320 cm^{-1} and the center of mass of the amide I region (1500-1800 cm^{-1}) were extracted from the preprocessed Raman spectra and highlight the evolution over time (B-E). Second-derivative Raman spectra and corresponding SNR (F). Temporal evolution of selected wavenumbers (1368, 1397, 1398, 1651 and 1673 cm^{-1}) corresponding to the top 5 SNR values (G-K).

7.3 Results and Discussion

7.3.1 Forced degradation studies

The primary objective of forced degradation studies was to identify Raman spectral features associated with aggregation. Therefore, samples with a concentration of 30 g/L BSA were exposed to thermal stress at 65°C and analyzed for their aggregate content via UHP-SEC and Raman spectroscopy. Fig. 7.1(A) schematically presents the methodology and the aggregate content over time. The UHP-SEC results illustrated in Fig. 7.1, demonstrate a time-dependent formation of aggregated species which saturates after roughly 90 min. Comparative retention time analysis showed that heat-induced aggregates elute earlier than native aggregates, suggesting the formation of higher-order multimers (data not shown). Notably, the data indicate that with increasing ionic

strength, heat-induced aggregate formation decreased. Forced degradation has been commonly used in Raman spectroscopy studies for aggregate detection (Boulet-Audet, Byrne, and Kazarian 2014; Barnett et al. 2015; Zhou et al. 2015), as batch to batch variations do not cover the experimentally interesting range to study spectral effects of protein aggregation. However, the aggregates produced by e.g. heat stress are not necessarily structurally coherent with naturally occurring aggregates. Nevertheless, forced degradation is a helpful tool to efficiently produce aggregated proteins for further investigation. The authors considered the use of non-native aggregates sufficient for initially investigating spectral features affected by the structural changes induced by aggregation, but acknowledge the need for validation in native aggregate species.

To further investigate the structural changes affected by aggregation, offline-recorded Raman spectra of the thermally stressed samples were evaluated. Fig. 7.3 (A) and (F) show background-corrected and second derivative Raman spectra for the different sampling time points, respectively. The upper panels in (B) and (G) depict the SNR between the single spectral features and the aggregate content. The top correlative features together with manually selected (manual selection was based on visual inspection of all wavenumber ranges, in addition to those previously identified in literature) ones are shown in the lower panels (B-E) and (G-K) over time with their respective standard deviations. In the background-corrected spectra (cf. Fig. 7.3(A)), the top correlative features appear in the region of the phenylalanine band between 990 and 1010 cm^{-1} . Spectral features in the amide regions (amide III at 1241 cm^{-1} and amide I between 1550 and 1700 cm^{-1} as expressed by the center of mass) show minor correlations with higher standard deviations. The center of mass of the amide I band was calculated according to Eq. (7.1). The amide III ratio (I_{1340}/I_{1321}) was not affected. When additionally treated with the second derivative SG filter, the SNR display high degree of correlation in the fingerprint region, mainly between 1200 and 1400 cm^{-1} as well as in the amide I between 1550 and 1700 cm^{-1} with values of up to 300 and 500, respectively.

The amide I region has previously been identified as reliable marker for aggregation in Raman spectroscopy studies (Zhou et al. 2015; Shivu et al. 2013). Especially for proteins with predominantly alpha-helical structures such as BSA, the amide I band is considered a reliable structural marker. In particular, the formation of intermolecular beta-sheets is known to cause a shift in the amide I band with decreasing intensity around 1650 cm^{-1} and increasing intensity around 1670 cm^{-1} . This shift may either be observed in second-derivative spectra thanks to the enhanced selectivity by this treatment or on a macroscopic scale by a shift in center of mass (Militello et al. 2004). Moreover, the amide I is the most promising structural marker in Raman spectroscopy of proteins due to the high intensity compared to amide II and III vibration modes or those of tyrosine and tryptophane side chains (Zhou et al. 2015). However, the overlapping water band, which was removed in this analysis, may pose a challenge in determining accurate amide I patterns as the water band may be affected by the solution environment, such as salt concentrations (Fontana, Ben Mabrouk, and Kauffmann 2013).

7.3.2 Mixing studies

Following the identification of Raman spectral features associated with aggregation, samples were prepared from an unstressed and a stressed BSA stock solution, where protein concentration and aggregate content were systematically varied (cf. Sec. 7.2.0.2) to evaluate the robustness of the spectral effects found in Sec. 7.3.1. LHS was employed as a statistical approach to efficiently cover the possible design space with a given number of samples. For each sample, Raman spectroscopy

CHAPTER 7. MONITORING PROTEIN AGGREGATES IN DSP

measurements were subsequently performed off-line using the ARMP. Fig. 7.1 (B) schematically depicts the described workflow and Fig. 7.2 (A) - (F) present the strongest correlations of individual spectral features with overall protein concentration and aggregate content.

The spectral intensities at 999 cm^{-1} , representing the phenylalanine band, closely follows protein concentration in the samples with standard deviations below 5 %. A slight non-linear trend towards higher concentrations, along with a few outliers at high concentrations, can be observed. The non-linearity highlights the complexity of Raman spectroscopy data, as the raw signals undergo multiple preprocessing steps which may influence the extent of non-linearity observed. Amide I related spectral features such as the 1673 cm^{-1} bands in the second derivative spectra, as well as the center of mass of the amide I linearly follow the aggregate content. In all instances, outliers can be observed. These outliers mainly originate from samples with low protein concentrations ($< 3\text{ g/L}$), where analysis becomes more challenging due to the overall low signal intensity. These results corroborate findings from forced degradation experiments and demonstrate the potential of the identified spectral features as reliable markers for quantifying protein aggregation.

The selection of the 1673 and 1651 cm^{-1} features in Figure 7.2 reflects the outcomes of the forced degradation study (Figure 7.3), where the Amide I region exhibited the strongest and most consistent correlation with aggregation. These wavenumber regions correspond to opposing components of the Amide I band associated with shifts from native α -helical structure (approximately 1651 cm^{-1}) toward intermolecular β -sheet formation (approximately 1673 cm^{-1}). Their inclusion in the mixing-study analysis therefore provides a mechanistic link to the structural transitions identified earlier and enables an evaluation of whether these aggregation-related spectral changes remain robust when protein concentration and aggregate content are varied independently.

Beyond amide I-related markers, various Raman intensity ratios involving protein-specific bands were investigated as potential indicators of protein aggregation. Specifically, the amide III ratio (I_{1340}/I_{1321}) and the peak maximum within the $920\text{--}950\text{ cm}^{-1}$ range were assessed (Ashton and E. W. Blanch 2010; Militello et al. 2004; C. H. Li and T. Li 2009; Zhou et al. 2015; Alizadeh-Pasdar, Nakai, and Li-Chan 2002). However, both showed only weak correlations with aggregate levels, accompanied by large standard deviations ($>20\%$) and the presence of significant outliers, limiting their reliability. Although the $1341/1320\text{ cm}^{-1}$ ratio is included for completeness because it has been reported as a potential aggregation-sensitive marker in prior Raman studies, our results demonstrate that it is not consistently responsive under the conditions tested here. We therefore treat it as an auxiliary rather than a primary structural marker, and do not rely on it for quantitative interpretation.

These findings highlight that while spectral features beyond the Amide I band can provide supplementary structural insights into aggregation processes, their robustness and reproducibility for aggregate quantification remain limited. The phenylalanine band as well as the amide I region remain the most reliable spectral markers for accurately monitoring aggregation dynamics. In contrast, other markers showed high variability and limited statistical significance, reducing their suitability for quantitative analysis.

7.3.3 Bind-elute anion-exchange chromatography experiments

7.3.3.1 Qualitative spectral analysis

Building on insights from the mixing studies, bind-elute anion-exchange chromatography experiments were conducted to assess in-line Raman detection capabilities for naturally occurring BSA aggregates under operational conditions. Chromatographic conditions comprised gradient elutions with varying loading densities from two BSA batches and one step elution (cf. Tab. 7.1). Continuous Raman spectra acquisition during the elution phases facilitated direct comparison with UHP-SEC analyses of collected fractions. Compared to the heat-induced aggregation system, where aggregates with higher-order structures formed rapidly at elevated temperatures and low ionic strength, chromatography load material was solely comprised of monomeric and dimeric BSA species.

In Fig. 7.4, fraction analytics from all AEX experiments are overlaid with selected spectral markers, namely the amide III ratio I_{1341}/I_{1320} and the center of mass of the amide I band, obtained from in-line acquired Raman spectra. The raw spectra were averaged over the collection interval of each fraction and preprocessed the same way as in the forced degradation and mixing experiments. Due to the buffer system using in the chromatographic separations, the spectral background correction was performed using the first acquired spectrum in the elution phase. In case the previously defined spectral markers prove selective for aggregate detection, we expect the following behavior during AEX elutions: Due the timely separation of monomeric and dimeric species, the spectral markers would initially remain constant until the co-elution of the dimeric species, which is more strongly retained by the chromatography resin. When the co-elution begins, the spectral marker is expected to increase as the relative content of the dimeric species is successively increasing until the elution peak ends. For non-separating conditions, the spectral markers are expected to remain constant over the entire elution phase. The absolute value depends on the aggregate content in the elution pool.

In Fig. 7.4, both the amide III ratio and the center of mass follow the described behavior for all tested chromatographic conditions despite some non-idealities. For runs from batch 1 (A-C), the loading densities ranged from 15 to 22.5 g/L resin. For the longer gradients of 10 and 15 CVs (B, C), this led to an eluate pool dilution with maximum monomer concentrations of <15 g/L. In those cases, the spectral markers appear unstable with a high degree of scattering. As a result, it is vaguely possible to deduct the onset of the co-elution of the dimeric species. For experiments from batch 2 (D-F), for which loading densities ranged from 22.5 to 40 g/L, the spectral markers show a lower degree of scattering and fewer outliers were observed. While for the 10 and 15 CV runs (E, F), the actual onset of co-elution after approx. 52 and 62 min, respectively, is accurately represented by the spectral markers, the spectral markers in the 5 CV run (C), do not match the fraction analytics. Notably, the step elution experiment (G) fulfills the expectations and the spectral markers remain constant of the width of the elution profile. As the elution occurred within approx. 2 min, the number of collected fractions and corresponding Raman spectra is comparably low. In all experiments, outliers can be observed which are mostly corresponding to fractions at the front and end of the elution peak with low concentrations (<3 g/L).

The depicted selection of spectral markers for AEX experiments is not exhaustive. However, they showed the best qualitative behavior among the studied markers. Notably, previously identified amide I related markers such as the 1651 and 1673 cm^{-1} intensities in the second derivative spectra were not found to be selective for aggregate content as they were affected by the total protein concentration and therefore followed the overall elution profile. Hypothetically, the higher-order

structure aggregates generated during the heat-treatment may form more intermolecular β -sheets than natively occurring dimers and hence show a pronounced effect in the amide I region, which upholds variations in protein concentration (Barnett et al. 2015; Zhou et al. 2015). Inversely, the amide III ratio, which showed no significant effect in forced degradation or mixing experiments, served as a valid indicator in AEX experiments. This is likely because pH and salt gradients in AEX can alter hydrogen bonding and polarity around tyrosine residues, enhancing spectral differences between monomeric and aggregated forms (Antosiewicz and Shugar 2016). Despite the ineffectiveness of the amide I related single-variable intensities, the cumulative measure of the center of mass proved suitable across all experimental modes. Across the studied literature, the center of mass appears to be the most reliable qualitative spectral marker for protein aggregation. However, it is difficult to use this marker for direct quantification of aggregate content, due to possible non-linearities or background effects. Especially when used in biopharmaceutical processes, such as chromatography, changing buffer environments or excipients are common. Even in AEX gradient elutions, the salt concentration is constantly changing which has an effect on the water band (Furić, Ciglenečki, and Čosović 2000) and therefore affects background correction steps in Raman spectroscopy. In summary, these results present the first direct illustration of how protein aggregation affects Raman spectra in chromatographic separation processes and shows applicability of Raman spectroscopy to monitoring biopharmaceutical downstream operations. Due to concentration variations and spectral noise, the use of Raman spectroscopy for direct quantification necessitates the calibration of statistical models.

7.3.3.2 Statistical model building

To enable quantitative prediction of protein size variants from Raman spectra, multiple regression models were calibrated targeting the concentrations of monomeric and dimeric BSA species. Model training incorporated spectra from three representative bind-elute AEX experiments, while testing was performed on four independent datasets to assess generalizability across varying chromatographic conditions (cf. Tab. 7.1). Multiple different statistical model variants were compared using different performance criteria, namely their prediction error on training, cross-validation, augmentation test and external test sets. For PLS models, every preprocessing variants was tested using a single-output (PLS1) and a multi-output version (PLS2). The studied CNNs were exclusively multi-output. Here, we compared solely two variants, namely a non-augmented and a augmented CNN.

Fig. E.2 presents RMSE values for monomer and aggregate components in absolute terms (A and B), as well as RMSE for the aggregate fraction calculated from the predicted monomer and aggregate concentrations (C). This systematic comparison reveals distinct performance patterns both within and between PLS and CNN model families across different evaluation scenarios. For monomer prediction (A), CNN models demonstrate lower RMSE during training, cross-validation, and augmentation test phases. However, this advantage largely disappears in the external test, where PLS and CNN models achieve comparable performance. This convergence suggests that the apparent superiority of CNN models in earlier phases may reflect overfitting rather than genuine generalization capability for monomer prediction. The aggregate prediction results (B) show a more consistent pattern, with CNN models maintaining lower RMSE across all evaluation splits. While individual PLS variants occasionally match CNN performance for specific datasets, no single PLS model achieves comparable accuracy across all scenarios, indicating more robust performance by the CNN approach for this target. Fig. E.2 (C) confirms this trend for aggregate fraction prediction, with selected PLS models approaching but not consistently matching CNN-type model performance. Notably, data augmentation consistently increases RMSE relative to non-augmented variants across

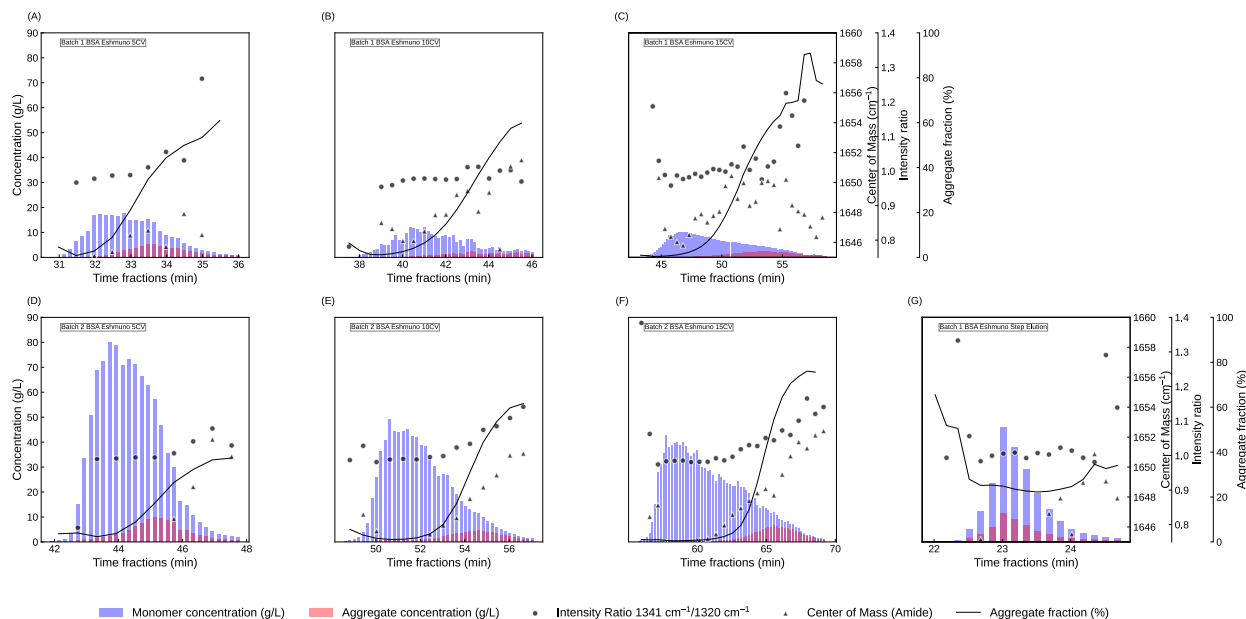


Figure 7.4 Elution profiles and structural markers of BSA under varying process conditions. Protein separation from two BSA batches was performed using different gradient lengths: (A, D) 5 CV; (B, E) 10 CV; (C, F) 15 CV; and (G) step elution. Blue and red bars represent monomer and aggregate concentrations (g/L), respectively. Black lines indicate aggregate fractions across elution time. Black circles and black triangles denote Raman spectral markers: intensity ratio ($1341\text{ cm}^{-1}/1320\text{ cm}^{-1}$) and Amide I band center of mass (cm^{-1}), respectively. To improve legibility for panels (A–F), Raman marker traces were smoothed by plotting the mean of three consecutive spectra at the corresponding mean elution time; panel (G) shows all measured points.

all model types and evaluation scenarios, suggesting that the augmentation strategy employed does not improve generalization and may introduce artifacts that degrade prediction quality.

To better visualize the accuracy of the different model variants, parity plots of observed versus predicted HMWC (%) for all calibrated PLS models are shown in Figs. E.4 to E.13, where panels (A)–(D) represent the model performance on the training, cross-validation, augmentation test and test datasets, respectively.

To resolve the ambiguity regarding data augmentation effectiveness, we conducted a comprehensive leave-4-groups-out rotation test (Fig. E.3), evaluating all possible training/test splits for each model type. The results demonstrate that the base CNN performs slightly better than the augmented CNN variant for both monomer and aggregate predictions. Both CNN approaches achieve substantially lower RMSE on average compared to the best-performing PLS model (Crop-SG-PLS1), particularly for aggregate prediction where the median RMSE is reduced by approximately 50%. The more compact distributions and lower median values for CNN models confirm their superior generalization capability when evaluated across diverse data splits.

Fig. 7.5 presents the model predictions for the Crop-SG-PLS1 and the CNN models in dashed and solid lines, respectively, for training (A), cross-validation (B) and test (C) sets. For the AEX runs in the training set, both models fit the UHP-SEC data accurately with minor deviations. For the

CHAPTER 7. MONITORING PROTEIN AGGREGATES IN DSP

Table 2(A). Coefficient of determination (R^2) for prediction of monomer and aggregate concentrations across all data splits and models.

Model	Analyte	Train	CV	Augmentation test	External test
PLS1	Monomer (g/L)	0.9761	0.9417	-0.1489	0.8474
	Aggregate (g/L)	0.9765	-0.5345	0.6823	0.6210
PLS2	Monomer (g/L)	0.9911	0.8710	-0.1621	0.8389
	Aggregate (g/L)	0.9878	-0.7969	0.5123	0.5756
SG-PLS1	Monomer (g/L)	0.9709	0.9422	0.1993	0.9030
	Aggregate (g/L)	0.6014	-0.3389	-0.2459	0.0365
SG-PLS2	Monomer (g/L)	0.9670	0.9313	0.5027	0.9323
	Aggregate (g/L)	0.5195	-0.2045	-0.4069	0.0371
Crop-SG-PLS1	Monomer (g/L)	0.9706	0.9529	-0.2536	0.7954
	Aggregate (g/L)	0.9642	0.2039	-2.9519	-0.2070
Crop-SG-PLS2	Monomer (g/L)	0.9772	0.9555	0.3686	0.9108
	Aggregate (g/L)	0.7694	-1.4017	0.8421	0.6350
PLS1-augmented	Monomer (g/L)	0.9632	0.9464	0.9152	0.8779
	Aggregate (g/L)	-0.0479	-0.1968	0.7373	-1.2896
PLS2-augmented	Monomer (g/L)	0.9625	0.9438	0.8955	0.8841
	Aggregate (g/L)	-0.0915	-0.2479	0.7803	-1.2446
CNN	Monomer (g/L)	0.9850	0.9752	0.9126	0.8274
	Aggregate (g/L)	0.9525	0.7804	0.6874	0.6894
CNN-Augmented	Monomer (g/L)	0.6775	0.8584	0.7112	0.5756
	Aggregate (g/L)	0.4114	0.3097	-0.0137	0.3051

Batch 1-5CV run, the PLS model closely follows the sharp dimer elution profile, while the CNN model predicts a more flattened, round profile. For the Batch-1-15CV, both models exhibit scattered predictions with the CNN being slightly less affected for both species. In the Batch 2-10CV run, both models match the UHP-SEC data accurately with the concentrations being the highest among the training runs. In the cross-validation, the accuracy and robustness of the CNN predictions are superior to the PLS predictions. In addition to the higher degree of scattering of the PLS model, the initial phases where only monomer is eluting are wrongly predicted by the PLS model and the selectivity for the monomeric species during co-elution is compromised (cf. Batch-2-10CV). The predictions for the test runs corroborate these findings and suggest a superiority of the CNN model. For all gradient elutions (Batch-2-5CV, Batch-1-10CV, Batch-2-15CV), the CNN correctly predicts the onset of co-elution and the absolute dimer content, while the PLS predictions are highly scattered and predict early co-elution onsets. Finally, the step elution confirms that both models are able to differentiate between monomeric and dimeric species as a negative control under non-separating conditions. Both models accurately predict a simultaneous elution of monomers and dimers despite an absolute offset in the monomer concentration compared to UHP-SEC data.

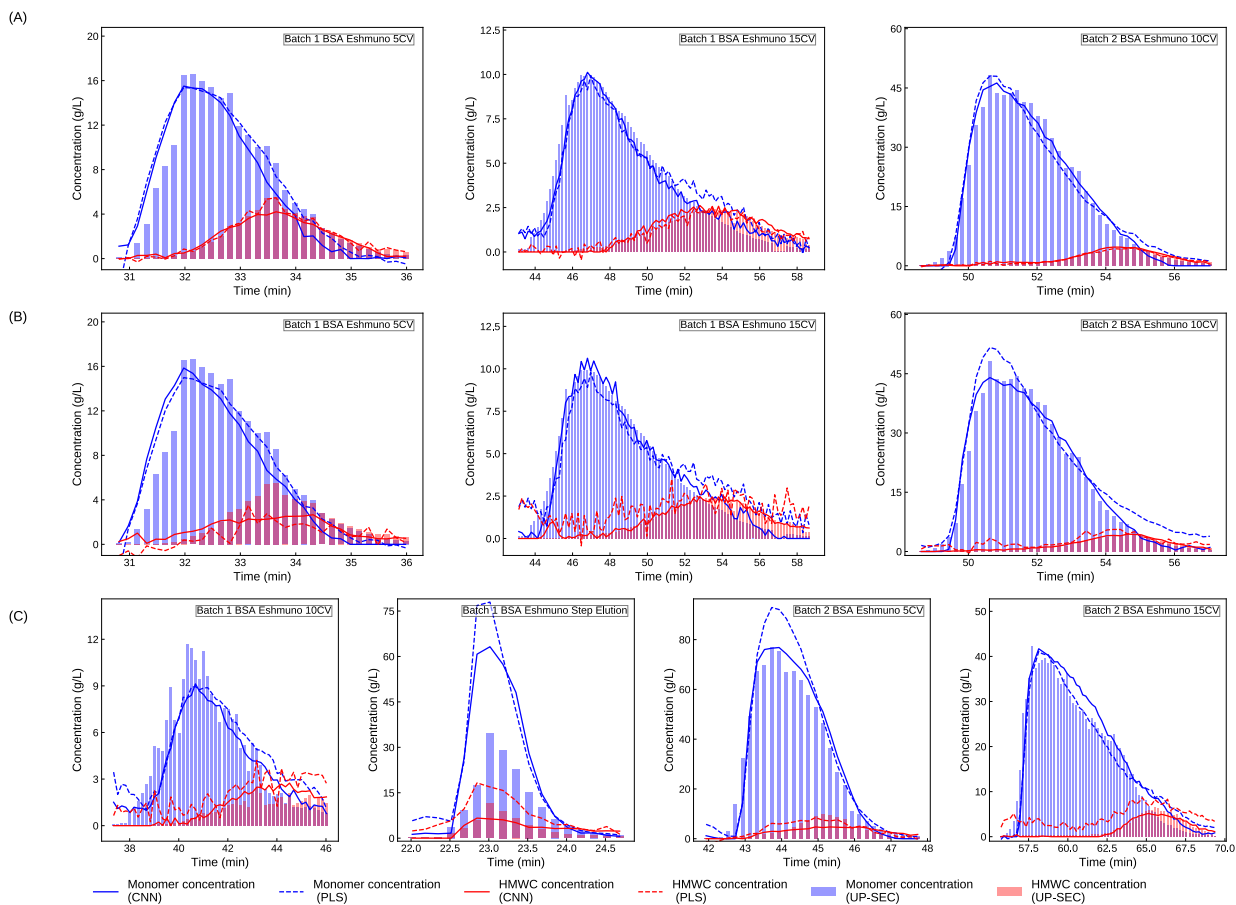


Figure 7.5 Predictions of Crop-SG-PLS1 (dashed lines) and CNN (solid lines) regression models for training (A), cross-validation (B) and test (C). Blue and red bars represent monomer and dimer concentrations as determined by UHP-SEC.

Considering all emphasized points, the CNN model is better than the studied PLS-based models, as it combines high accuracy with robustness across cross-validation, test and augmentation test for both target species. These findings confirm the work presented in (Schiemer, Rdt, and Hubbuch 2024) where the CNN models have been established for quantification of different proteins and monoclonal antibody size variants from ultraviolet-visible absorption spectra. In contrast to (Schiemer, Rdt, and Hubbuch 2024), the augmented CNN model used within this study achieved lower performance than the non-augmented variants across all data sets and test scenarios. In direct comparison with the PLS model, the augmented CNN still was found to be superior. Although data augmentation was not found to be beneficial for this particular case, it is notable that the transfer of the LSA algorithm to Raman spectra was successful. Another recent example of data augmentation for Raman spectra showed a beneficial effect for chemometric modeling (Lange et al. 2024).

Compared to established spectroscopy-based PAT applications in biopharma, such as glucose monitoring in fermentation processes (Cavinato et al. 1990; D. H. Mller, Koss, et al. 2024; Pontius et al. 2020) or UFDF monitoring (Rolinger, Rdt, and Hubbuch 2020a; Rolinger, Hubbuch, and Rdt 2023; Milewska et al. 2023), spectroscopic methods for aggregate quantification remain limited

CHAPTER 7. MONITORING PROTEIN AGGREGATES IN DSP

Table 2(B). Root-mean-square error (RMSE, g/L) for prediction of monomer and aggregate concentrations across all data splits and models.

Model	Analyte	Train	CV	Augmentation test	External test
PLS1	Monomer	1.6568	2.5848	6.0818	6.8610
	Aggregate	0.1945	1.5711	0.8141	1.4377
PLS2	Monomer	1.0121	3.8464	6.1169	7.0501
	Aggregate	0.1399	1.7001	1.0086	1.5213
SG-PLS1	Monomer	1.8253	2.5742	5.0772	5.4720
	Aggregate	0.8007	1.4675	1.6122	2.2922
SG-PLS2	Monomer	1.9438	2.8073	4.0014	4.5705
	Aggregate	0.8792	1.3919	1.7132	2.2914
Crop-SG-PLS1	Monomer	1.8360	2.3248	6.3529	7.9450
	Aggregate	0.2398	1.1316	2.8712	2.5655
Crop-SG-PLS2	Monomer	1.6181	2.2585	4.5088	5.2476
	Aggregate	0.6091	1.9655	0.5739	1.4108
PLS1-augmented	Monomer	2.0550	2.4796	1.6523	6.1382
	Aggregate	1.2983	1.3875	0.7403	3.5335
PLS2-augmented	Monomer	2.0742	2.5386	1.8343	5.9808
	Aggregate	1.3250	1.4168	0.6769	3.4986
CNN	Monomer	1.3103	1.6861	1.6777	7.2968
	Aggregate	0.2765	0.5944	0.8075	1.3015
CNN-Augmented	Monomer	6.0814	4.0293	3.0495	11.4429
	Aggregate	0.9731	1.0537	1.4542	1.9467

(Wei et al. 2022; Brestrich, Rdt, et al. 2018). Brestrich et al. (Brestrich, Rdt, et al. 2018) relied on only four experimental runs, with just one reserved for validation, whereas Wei et al. (Wei et al. 2022) utilized aggregates derived from forced degradation experiments, while the cross-correlation structure in the provided data is unclear. Furthermore, reported results were predominantly acquired offline, allowing longer spectral acquisition times and eliminating the challenges associated with real-time process variations.

Although BSA serves as a practical and well-characterized surrogate for method development, we acknowledge that its structural simplicity and spectral characteristics do not fully capture the complexity of industrially-relevant biotherapeutics, like monoclonal antibodies. Extending the approach to mAbs requires consideration of their distinct secondary-structure compositions, particularly their substantially higher β -sheet content relative to the predominantly α -helical structure of BSA. This is expected to dampen the spectral difference and hence making the distinction of monomeric from dimeric species more difficult. Additional factors include the higher concentration ranges typically encountered for therapeutic antibodies in downstream processing (20–60 g/L), the aggregate levels relevant to industrial polishing steps (reduction from roughly 3 % to $\leq 0.5\%$), and detection-limit requirements of drug product formulations (0.1–1 % aggregates).

Ongoing follow-up studies with therapeutic antibodies aim to evaluate these aspects and assess the transferability and robustness of the method across biopharmaceutical modalities.

In summary, the use of CNN models demonstrated suitability by generating selective regression models capable of accurately quantifying monomer and dimer contents in real-time chromatographic processes. However, the development of robust chemometric models requires extensive evaluation of their selectivity and predictive performance through larger experimental datasets and systematic algorithmic validation.

7.4 Conclusions

This study demonstrates the feasibility of real-time aggregate quantification in biopharmaceutical downstream processing using Raman spectroscopy and structural spectral markers. By applying controlled aggregation stress and a Latin hypercube sampling design, distinct Raman features — particularly in the amide I region — were identified as reliable indicators of protein aggregation across varying concentrations and conditions.

These spectral markers were validated in in-line measurements during bind-elute anion-exchange chromatography, where they successfully tracked monomer–aggregate separation dynamics. The combination of Raman spectroscopy with multivariate modeling approaches, including partial least squares and a convolutional neural network enhanced through synthetic data augmentation, enabled robust and accurate quantification of size variants under process-relevant conditions. While the CNN model outperformed traditional approaches in terms of predictive performance and selectivity, limitations arose from low signal intensities at low protein concentrations and environmental influences on background correction.

Despite these limitations, this work highlights the potential of Raman-based structural markers and machine learning models for real-time monitoring of protein aggregation in downstream unit operations.

Future work should also aim to expand the experimental space and enhance spectral preprocessing to improve model robustness and transferability across different process steps and environmental conditions.

Conflict of interest

The authors declare no conflicts of interests.

Author contributions

Jakob Heyer-Müller and Robin Schiemer both evolved the concepts and methods presented in this manuscript, supervised experimental work, analyzed and interpreted the data, drafted the figures and wrote the final manuscript. Lars Robbel and Michael Schmitt, and Jürgen Hubbuch supervised the work. Jürgen Hubbuch acquired funding and initiated the work. All authors read and approved the final manuscript.

Acknowledgments

Open Access funding enabled and organized by Projekt DEAL.

Data availability statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Chapter references

- Alizadeh-Pasdar, Nooshin, Shuryo Nakai, and Eunice C. Y. Li-Chan (10/2002). „Principal Component Similarity Analysis of Raman Spectra To Study the Effects of pH, Heating, and κ -Carrageenan on Whey Protein Structure“. In: *Journal of Agricultural and Food Chemistry* 50.21, pp. 6042–6052.
- Antosiewicz, Jan M. and David Shugar (05/2016). „UV–Vis spectroscopy of tyrosine side-groups in studies of protein structure. Part 2: selected applications“. en. In: *Biophysical Reviews* 8. Open access review article, pp. 163–177.
- Ashton, Lorna and Ewan W. Blanch (06/2010). „pH-induced conformational transitions in α -lactalbumin investigated with two-dimensional Raman correlation variance plots and moving windows“. In: *Journal of Molecular Structure*. PROGRESS IN TWO-DIMENSIONAL CORRELATION SPECTROSCOPY 974.1, pp. 132–138.
- Barnett, Gregory V. et al. (12/2015). „Structural Changes and Aggregation Mechanisms for Anti-Streptavidin IgG1 at Elevated Concentration“. eng. In: *The Journal of Physical Chemistry. B* 119.49, pp. 15150–15163.
- Boulet-Audet, Maxime, Bernadette Byrne, and Sergei G. Kazarian (10/2014). „High-Throughput Thermal Stability Analysis of a Monoclonal Antibody by Attenuated Total Reflection FT-IR Spectroscopic Imaging“. In: *Analytical Chemistry* 86.19, pp. 9786–9793.
- Brestrich, Nina, Matthias Rüdte, et al. (02/2018). „Selective protein quantification for preparative chromatography using variable pathlength UV/Vis spectroscopy and partial least squares regression“. In: *Chemical Engineering Science* 176.C, pp. 157–164.
- Brestrich, Nina, Adrian Sanden, et al. (07/2015). „Advances in inline quantification of co-eluting proteins in chromatography: Process-data-based model calibration and application towards real-life separation issues“. eng. In: *Biotechnology and Bioengineering* 112.7, pp. 1406–1416.
- Capito, Florian, Romas Skudas, Harald Kolmar, et al. (2015). „At-line mid infrared spectroscopy for monitoring downstream processing unit operations“. In: *Process Biochemistry* 50.6, pp. 997–1005.
- Cavinato, Anna G. et al. (09/1990). „Noninvasive method for monitoring ethanol in fermentation processes using fiber-optic near-infrared spectroscopy“. In: *Analytical Chemistry* 62.18, pp. 2058–2063.
- Chen, Jingyi et al. (03/2024). „Application of Raman spectroscopy during pharmaceutical process development for determination of critical quality attributes in Protein A chromatography“. In: *Journal of Chromatography A* 1718, p. 464721.
- Dietrich, Annabelle et al. (05/2024). „Raman-based PAT for VLP precipitation: systematic data diversification and preprocessing pipeline identification“. en. In: *Frontiers in Bioengineering and Biotechnology* 12, p. 1399938.
- Eisenberg, Azaria Solomon and Laura J. Juszczak (07/2012). „Correlation of TrpGly and GlyTrp Rotamer Structure with W7 and W10 UV Resonance Raman Modes and Fluorescence Emission Shifts“. In: *Journal of Amino Acids* 2012, pp. 1–10.
- Erb, Donald (n.d.). *pybaselines: A Python library of algorithms for the baseline correction of experimental data*.
- Ettah, Ilokugbe and Lorna Ashton (07/2018). „Engaging with Raman Spectroscopy to Investigate Antibody Aggregation“. en. In: *Antibodies* 7.3, pp. 24–24.
- FDA (2004). „Guidance for Industry: PAT—a framework for innovative pharmaceutical development, manufacturing, and quality assurance“. In.
- Feidl et al. (10/2019). „Combining Mechanistic Modeling and Raman Spectroscopy for Monitoring Antibody Chromatographic Purification“. In: *Processes* 7.10, pp. 683–683.
- Fontana, Marc D., Kawther Ben Mabrouk, and Thomas H. Kauffmann (2013). „Raman spectroscopic sensors for inorganic salts“. In: *Spectroscopic Properties of Inorganic and Organometallic Compounds*. Ed. by Jack Yarwood, Richard Douthwaite, and Simon Duckett. Vol. 44. RSC Publishing, pp. 40–67.
- Furić, K, I Ciglenečki, and B Čosović (09/2000). „Raman spectroscopic study of sodium chloride water solutions“. In: *Journal of Molecular Structure* 550-551, pp. 225–234.
- Glasse, Jarka et al. (04/2011). „Process analytical technology (PAT) for biopharmaceuticals“. In: *Biotechnology Journal* 6.4, pp. 369–377.
- Gómez de la Cuesta, Ramón, Royston Goodacre, and Lorna Ashton (11/2014). „Monitoring Antibody Aggregation in Early Drug Development Using Raman Spectroscopy and Perturbation-Correlation Moving Windows“. In: *Analytical Chemistry* 86.22, pp. 11133–11140.
- Harada, Issei, Takashi Miura, and Hideo Takeuchi (01/1986). „Origin of the doublet at 1360 and 1340 cm⁻¹ in the Raman spectra of tryptophan and related compounds“. en. In: *Spectrochimica Acta Part A: Molecular Spectroscopy* 42.2-3, pp. 307–312.
- Heyer-Müller, Jakob et al. (2025). „A novel Raman-chromatography assembly for automated calibration and in-line monitoring in bioprocessing“. In: *Engineering in Life Sciences*.
- Lange, Christoph et al. (2024). „Data Augmentation Scheme for Raman Spectra with Highly Correlated Annotations“. In: *34th European Symposium on Computer Aided Process Engineering / 15th International Symposium on Process Systems Engineering*. Ed. by Flavio Manenti and Gintaras V. Reklaitis. Vol. 53. Computer Aided Chemical Engineering. Elsevier, pp. 3055–3060.
- Lewis, E. Neil et al. (12/2014). „Combined Dynamic Light Scattering and Raman Spectroscopy Approach for Characterizing the Aggregation of Therapeutic Proteins“. en. In: *Molecules* 19.12, pp. 20888–20905.
- Li, Cynthia H. and Tiansheng Li (2009). „Application of Vibrational Spectroscopy to the Structural Characterization of Monoclonal Antibody and its Aggregate“. en. In: *Current Pharmaceutical Biotechnology* 10.4, pp. 391–399.
- Lippert, J. L., D. Tyminski, and P. J. Desmeules (10/1976). „Determination of the secondary structure of proteins by laser Raman spectroscopy“. en. In: *Journal of the American Chemical Society* 98.22, pp. 7075–7080.
- Makki, Alaa A. et al. (02/2021). „Understanding the discrimination and quantification of monoclonal antibodies preparations using Raman spectroscopy“. In: *Journal of Pharmaceutical and Biomedical Analysis* 194, p. 113734.
- Milewska, Adrianna et al. (12/2023). „In-line monitoring of protein concentration with MIR spectroscopy during UFDF“. en. In: *Engineering in Life Sciences* 23.2, e2200050.

CHAPTER 7. MONITORING PROTEIN AGGREGATES IN DSP

- Militello, Valeria et al. (02/2004). „Aggregation kinetics of bovine serum albumin studied by FTIR spectroscopy and light scattering“. In: *Biophysical Chemistry* 107.2, pp. 175–187.
- Moussa, Ehab M. et al. (02/2016). „Immunogenicity of Therapeutic Protein Aggregates“. English. In: *Journal of Pharmaceutical Sciences* 105.2, pp. 417–430.
- Müller, David Heinrich, Hans-Jürgen Koss, et al. (02/2024). „Bioprocess in-line monitoring and control using Raman spectroscopy and Indirect Hard Modeling (IHM)“. en. In: *Biotechnology and Bioengineering* n/a.n/a.
- Ota, Chikashi et al. (04/2016). „Assessment of the Protein–Protein Interactions in a Highly Concentrated Antibody Solution by Using Raman Spectroscopy“. en. In: *Pharmaceutical Research* 33.4, pp. 956–969.
- Pham, Ngoc B. and Wilson S. Meng (07/2020). „Protein aggregation and immunogenicity of biotherapeutics“. en. In: *International Journal of Pharmaceutics* 585, p. 119523.
- Pontius, Katrin et al. (06/2020). „Monitoring yeast fermentations by nonlinear infrared technology and chemometrics – understanding process correlations and indirect predictions“. In: *Applied Microbiology and Biotechnology* 104.12, pp. 5315–5335.
- Roberts, Christopher J (12/2014). „Protein aggregation and its impact on product quality“. en. In: *Current Opinion in Biotechnology* 30, pp. 211–217.
- Rolinger, Laura, Jürgen Hubbuch, and Matthias Rüdts (02/2023). „Monitoring of ultra- and diafiltration processes by Kalman-filtered Raman measurements“. In: *Analytical and Bioanalytical Chemistry* 415.5, pp. 841–854.
- Rolinger, Laura, Matthias Rüdts, and Jürgen Hubbuch (04/2020a). „A critical review of recent trends, and a future perspective of optical spectroscopy as PAT in biopharmaceutical downstream processing“. In: *Analytical and Bioanalytical Chemistry* 412.9, pp. 2047–2064.
- Rolinger, Laura, Matthias Rüdts, and Jürgen Hubbuch (11/2021). „Comparison of UV- and Raman-based monitoring of the Protein A load phase and evaluation of data fusion by PLS models and CNNs“. In: *Biotechnology and Bioengineering* 118.11 (11), pp. 4255–4268.
- Rosenberg, Amy S. (09/2006). „Effects of protein aggregates: An immunologic perspective“. en. In: *The AAPS Journal* 8.3, p. 59.
- Rüdts, Matthias, Nina Brestrich, et al. (02/2017). „Real-time monitoring and control of the load phase of a protein A capture step“. In: *Biotechnology and Bioengineering* 114.2 (2), pp. 368–373.
- Rüdts, Matthias, Philipp Vormittag, et al. (06/2019). „Process monitoring of virus-like particle reassembly by diafiltration with UV/Vis spectroscopy and light scattering“. In: *Biotechnology and Bioengineering* 116.6, pp. 1366–1379.
- Rygula, A. et al. (2013). „Raman spectroscopy of proteins: A review“. In: *Journal of Raman Spectroscopy* 44.8, pp. 1061–1076.
- Sato, Yusui et al. (2023). „Raman Spectroscopic Analysis of Highly-Concentrated Antibodies under the Acid-Treated Conditions“. In: *Pharmaceutical Research* 40.7, pp. 1853–1864.
- Schiemer, Robin, Matthias Rüdts, and Jürgen Hubbuch (2024). „Generative data augmentation and automated optimization of convolutional neural networks for process monitoring“. In: *Frontiers in Bioengineering and Biotechnology* 12.
- Schiemer, Robin, Jan Tobias Weggen, et al. (05/2023). „An adaptive soft-sensor for advanced real-time monitoring of an antibody-drug conjugation reaction“. en. In: *Biotechnology and Bioengineering* 120.7, pp. 1914–1928.
- Shivu, Bhavana et al. (08/2013). „Distinct β -Sheet Structure in Protein Aggregates Determined by ATR–FTIR Spectroscopy“. In: *Biochemistry* 52.31, pp. 5176–5183.
- Soch, Joram and Carsten Allefeld (08/2018). „MACS – a new SPM toolbox for model assessment, comparison and selection“. In: *Journal of Neuroscience Methods* 306, pp. 19–31.
- Vázquez-Rey, María and Dietmar A. Lang (07/2011). „Aggregates in monoclonal antibody manufacturing processes“. en. In: *Biotechnology and Bioengineering* 108.7, pp. 1494–1508.
- Wang, Jiarui et al. (12/2023b). „In-line product quality monitoring during biopharmaceutical manufacturing using computational Raman spectroscopy“. In: *mAbs* 15.1, p. 2220149.
- Wei, Bingchuan et al. (12/2022). „Multi-attribute Raman spectroscopy (MARS) for monitoring product quality attributes in formulated monoclonal antibody therapeutics“. In: *mAbs* 14.1 (1).
- Williams, Robert W. (1986). „[14] Protein secondary structure analysis using Raman amide I and amide III spectra“. en. In: *Methods in Enzymology*. Vol. 130. Elsevier, pp. 311–331.
- Zhang, Chi et al. (2019). „Rapid, quantitative determination of aggregation and particle formation for antibody drug conjugate therapeutics with label-free Raman spectroscopy“. In: *Analytica Chimica Acta* 1081, pp. 138–145.
- Zhou, Chen et al. (03/2015). „Concomitant Raman spectroscopy and dynamic light scattering for characterization of therapeutic proteins at high concentrations“. en. In: *Analytical Biochemistry* 472, pp. 7–20.

General Discussion and Conclusion

The objective of this thesis was to advance chemometric model development for spectroscopic process monitoring in biopharmaceutical manufacturing through modern machine learning methods that address fundamental selectivity challenges and systematic frameworks for preprocessing optimization, automated calibration, and model validation. An adaptive soft-sensor framework was developed to overcome spectral selectivity limitations in antibody-drug conjugate (ADC) monitoring by fusing ultraviolet/visible (UV/Vis) spectroscopy with mechanistic process knowledge (Ch. 3). A novel data augmentation method was introduced to enable training of CNNs with limited experimental data, combined with automated hyperparameter optimization (HPO) (Ch. 4). Raman spectroscopy was systematically evaluated for monitoring virus-like particle (VLP) precipitation, establishing preprocessing pipelines that eliminate environmental interferences in particulate systems (Ch. 5). An automated platform integrating a chromatography system with Raman spectroscopy was developed for high-throughput data generation and model transfer (Ch. 6). Finally, Raman-based monitoring of protein aggregation during chromatographic separation was demonstrated using validated structural markers and deep learning approaches (Ch. 7). Overall, the integration of machine learning (ML) with systematic experimental and computational frameworks provided robust tools for spectroscopic PAT across diverse biopharmaceutical applications.

Real-time monitoring of ADC conjugation reactions presents fundamental challenges because multiple species with similar spectral properties coexist simultaneously. UV/Vis spectroscopy cannot directly resolve individual conjugation states, as the absorption spectra of unconjugated antibody and conjugated species are too similar for direct chemometric decomposition. To overcome this selectivity limitation, a hybrid soft-sensor was developed in Ch. 3 that combines spectroscopic measurements with mechanistic process knowledge through an extended Kalman filter (EKF). The soft-sensor fuses predictions from a Gaussian process regression (GPR) model trained on UV/Vis spectra with forecasts from a kinetic reaction model describing the site-directed conjugation mechanism. For the first time, time-variant uncertainty quantification was incorporated into both the sensor and process models, enabling dynamic balancing between spectroscopic observations and

mechanistic predictions as model accuracies evolve throughout the reaction. The GPR model with automated variable selection achieved 18% better accuracy than PLS regression for conjugated drug quantification, while the EKF framework reduced overall prediction errors by 8.7% during training and 17.2% during validation compared to the kinetic model alone. Importantly, robustness testing demonstrated that the adaptive soft-sensor compensates effectively for noisy sensor signals, incorrect model initialization, and transfer from batch to fed-batch operation modes. The time-variant process noise estimation through Monte Carlo simulation and direct propagation of GPR predictive variance as sensor noise proved crucial for achieving stable state estimation across diverse operating conditions. Overall, Ch. 3 shows that mechanistic knowledge integration enables monitoring of chemically indistinguishable species beyond the selectivity limits of spectroscopy alone, providing a framework for model-based process control in ADC manufacturing.

CNNs offer potential advantages over linear chemometric methods through their hierarchical feature extraction capabilities, yet their application to spectroscopic calibration faces fundamental obstacles: experimental datasets rarely provide sufficient samples for robust training, and configuring network architectures is laborious and computationally expensive. Ch. 4 addressed these challenges through a novel data augmentation method that generates synthetic spectra preserving physicochemical validity. Unlike existing approaches, the method accounts for concentration-dependent effects by estimating pure component spectra locally within concentration neighborhoods, then combining these estimates to generate realistic synthetic mixtures across the experimental space. Bayesian optimization was employed to automate the combined search of data augmentation parameters and network architecture which identified CNNs with three convolutional layers, with dropout and regularization as optimal. The resulting CNN models demonstrated up to 50% reduction in RMSE compared to PLS regression for quantification of monoclonal antibody (mAb) size variants, while simultaneously reducing preprocessing requirements. Importantly, model-agnostic interpretability assessment using Shapley additive explanations (SHAP) values revealed that optimized CNNs exhibit increased robustness against spectral noise and wavelength shifts compared to PLS models. The optimized CNN architecture transferred successfully from the original UV/Vis dataset to additional UV/Vis and infrared (IR) applications without re-optimization, demonstrating generalizability beyond specific datasets. To conclude, Ch. 4 provides an in-depth analysis for using deep learning in spectroscopic applications where data scarcity is a common limitation. The introduced data augmentation method and automated optimization workflow enables practical implementation in chemometrics.

The transition from UV/Vis to Raman spectroscopy introduces additional challenges related to environmental interference in complex bioprocess matrices. VLP precipitation from crude *E. coli* lysates represents a particularly challenging application where particulate suspensions, precipitating agents, and feedstock components dominate observed spectra over VLP-related features. Unlike relatively clean UV/Vis measurements, Raman spectra exhibit varying baseline shapes and scatter behaviors that continuously change with sample composition. In Ch. 5, a Raman-based PAT for monitoring VLP precipitation through systematic data diversification and preprocessing pipeline optimization was developed. A diverse data set was generated by conditioning clarified lysate through spiking with purified VLPs, host cell proteins (HCPs), and varying dilution factors prior to batch and fed-batch experiments, resulting in varying precipitation backgrounds and dynamics across eight batch and three fed-batch experiments. A systematic screening of preprocessing combinations identified an optimal pipeline comprising turbidity correction through normalization at 3299 cm^{-1} , asymmetric least squares baseline correction, orthogonal projection to latent structures (OPLS)-based background removal, difference spectra calculation, SG second derivative filtering, and spectral

cropping to the 800–1800 cm^{-1} region. For the first time, this systematic approach demonstrated effective elimination of process-related interferences while preserving chemical information relevant to VLP and precipitant quantification in particulate systems. PLS models achieved R^2 values of 0.98 and 0.97 for ammonium sulfate as the precipitant in batch and fed-batch modes respectively, with R^2 of 0.74 and 0.64 for VLP concentration. Importantly, spectral analysis of amide band ratios (1341/1660 cm^{-1}) and nucleic acid/protein ratios (781/1341 cm^{-1}) provided qualitative confirmation of selective VLP precipitation, revealing structural differences between VLPs and co-precipitated impurities. Variable importance in projection (VIP) analysis showed that models rely on residual precipitant information alongside protein-associated bands, explaining the non-linear relationship between spectral features and VLP concentration that limits resolution of fine differences. Overall, Ch. 5 demonstrates the suitability of a Raman-based PAT for protein quantification in particulate systems, with the preprocessing optimization showcasing a systematic comparison of different modeling approaches.

Traditional calibration procedures create an experimental bottleneck limiting widespread PAT implementation, as manual preparation and measurement of calibration samples is labor-intensive and error-prone. Furthermore, systematic differences between off-line calibration and in-line process measurements can degrade model performance when transferred to real-time applications. Ch. 6 introduced an automated platform integrating a commercial chromatography system with a Raman spectrometer to enable high-throughput data generation with minimal sample volumes and operator involvement. Statistical experimental designs were applied to efficiently sample multi-dimensional concentration spaces for lysozyme, phosphate, and citrate mixtures. An open platform communication - unified architecture (OPC-UA) control interface enabled fully automated execution of calibration experiments, reducing the manual effort required for chemometric model development. PLS models achieved excellent performance with RMSE of prediction of 1.88 g/L for protein and 2–3 mM for buffer components. Importantly, a direct standardization approach was employed to transfer off-line calibrated models to in-line applications by modeling systematic spectral differences between measurement configurations using synthetic spectra generated through the data augmentation method introduced in Ch. 4. Validation during ultra-/diafiltration (UF/DF) demonstrated that the transferred models correctly track concentration trends, with lysozyme increasing from 5 to 20 g/L, phosphate depleting from 40 to 0 mM, and citrate accumulating from 0 to 90 mM over the course of the UF/DF phases. However, negative phosphate predictions and citrate overestimation pointed out limitations in the deployed models indicating that direct standardization requires refinement for complex multi-component matrices. To conclude, Ch. 6 establishes a practical workflow for accelerated PAT implementation, reducing calibration effort while enabling model transfer from automated off-line platforms to continuous in-line monitoring. Monitoring protein aggregation during chromatographic separation presents a crucial lever for real-time process control, yet the distinction between monomers and aggregates is challenging due to subtle spectral differences between conformational states. While Raman spectroscopy provides molecular structural information sensitive to protein secondary structure, previous studies typically assessed aggregation qualitatively or under idealized conditions, lacking comprehensive validation that spectral features respond specifically to conformational changes rather than confounding factors. Ch. 7 systematically assessed Raman-based aggregate monitoring through controlled stress experiments, orthogonal mixing studies, and in-line detection during anion-exchange chromatography. Forced degradation of BSA at 65 °C identified the amide I region (1550–1700 cm^{-1}) as primary aggregation markers reflecting secondary structure changes. Mixing studies with 48 conditions spanning 0–15 g/L protein and 25–50% aggregate content validated that these markers respond

specifically to aggregation state independent of total protein concentration. The amide I center of mass was introduced as a robust structural marker tracking aggregate co-elution during preparative chromatography across seven experiments with varying loading densities (15–45 g/L resin) and gradient lengths (5–15 column volumes). A CNN model with the same architecture as previously identified in Ch. 4 was trained, achieving superior performance compared to PLS variants. Importantly, the CNN showed robust performance in gradient and step elutions with varying process parameters. In contrast to the findings in Ch. 4, data augmentation-based training did not yield superior CNNs compared to non-augmented versions. Accuracy limitations were observed at low concentrations (<3 g/L) where signal-to-noise ratios caused outliers. Overall, Ch. 7 establishes validated structural markers for protein aggregation monitoring and demonstrates that deep learning enables real-time aggregate quantification during biopharmaceutical downstream processing (DSP). In conclusion, this thesis presents methodological advances for chemometric model development in biopharmaceutical PAT through integration of modern ML methods and comprehensive frameworks for data preprocessing, model calibration, and model validation. The adaptive ADC soft-sensor demonstrated that hybrid approaches combining spectroscopy with mechanistic knowledge overcome fundamental selectivity limitations, enabling monitoring of species beyond direct spectroscopic resolution (Ch. 3). The novel data augmentation method addressed the critical bottleneck of data scarcity for deep learning, enabling the use of CNNs in spectroscopic applications (Ch. 4). The systematic screening of preprocessing pipelines for turbid VLP precipitation systems (Ch. 5) provided a solution for handling environmental interferences in complex bioprocess matrices. The automated calibration platform reduced implementation barriers for PAT adoption while establishing model transfer protocols for transitioning from off-line development to in-line deployment (Ch. 6). Finally, validated structural markers and deep learning approaches for protein aggregation monitoring (Ch. 7) demonstrated the capability for real-time quality assessment during chromatographic separation. Beyond individual contributions, several cross-cutting themes emerged: the importance of robustness of chemometric model estimates for robust decision-making (Chs. 3, 4, 7), the value of systematic experimental design for data diversification (Chs. 5, 6, 7), and the role of interpretability methods for ensuring models capture relevant chemical relationships (Chs. 4, 5, 7). While the research focused on biopharmaceutical applications including ADCs, VLPs, mAbs, and protein aggregates, the developed methods are broadly applicable to small molecule pharmaceutical manufacturing, chemical process industries, and other domains requiring robust spectroscopic monitoring. By building on proposed frameworks, the immediate detection of process deviations and implementation of corrective actions before product quality is compromised may become reality, while process development activities may be accelerated through reduced reliance on time-consuming off-line analytics. In essence, these developments contribute to enhanced process understanding and pave the way toward robust, interpretable, and automated spectroscopic PAT for advanced biopharmaceutical manufacturing.

References

- Abdallah, Fatma F. et al. (2019). „Orthogonal projection to latent structures and first derivative for manipulation of PLSR and SVR chemometric models’ prediction: A case study“. In: *PLoS ONE* 14.9, pp. 1–13.
- Abel, Jeff et al. (2018). „A Small-Scale Process for Predicting Donnan and Volume Exclusion Effects During Ultrafiltration/Diafiltration Process Development“. In: *Journal of Pharmaceutical Sciences* 107.5, pp. 1296–1303.
- Abu-Absi, Nicholas R. et al. (05/2011). „Real time monitoring of multiple parameters in mammalian cell culture bioreactors using an in-line Raman spectroscopy probe“. In: *Biotechnology and Bioengineering* 108.5, pp. 1215–1221.
- Acquarelli, Jacopo et al. (02/2017). „Convolutional neural networks for vibrational spectroscopic data analysis“. In: *Analytica Chimica Acta* 954. June 2018 (June 2018), pp. 22–31.
- Afseth, Nils Kristian and Achim Kohler (2012). „Extended multiplicative signal correction in vibrational spectroscopy, a tutorial“. In: *Chemometrics and Intelligent Laboratory Systems* 117, pp. 92–99.
- Agarwal, Paresh and Carolyn R Bertozzi (2015). „Site-specific antibody-drug conjugates: The nexus of bioorthogonal chemistry, protein engineering, and drug development“. In: *Bioconjugate Chemistry* 26.2, pp. 176–192.
- Akiba, Takuya et al. (2019). *Optuna: A Next-generation Hyperparameter Optimization Framework*.
- Alizadeh-Pasdar, Nooshin, Shuryo Nakai, and Eunice C. Y. Li-Chan (10/2002). „Principal Component Similarity Analysis of Raman Spectra To Study the Effects of pH, Heating, and κ -Carrageenan on Whey Protein Structure“. In: *Journal of Agricultural and Food Chemistry* 50.21, pp. 6042–6052.
- Alsmeyer, Frank, Hans Jürgen Koß, and Wolfgang Marquardt (2004). „Indirect spectral hard modeling for the analysis of reactive and interacting mixtures“. In: *Applied Spectroscopy* 58.8, pp. 975–985.
- Andersen, C. M. and R. Bro (11/2010). „Variable selection in regression-a tutorial“. In: *Journal of Chemometrics* 24.11-12, pp. 728–737.
- Anderson, N.T. et al. (10/2020). „Achieving robustness across season, location and cultivar for a NIRS model for intact mango fruit dry matter content“. en. In: *Postharvest Biology and Technology* 168, p. 111202.
- Andersson, Martin (2009). „A comparison of nine PLS1 algorithms“. In: *Journal of Chemometrics: A Journal of the Chemometrics Society* 23.10, pp. 518–529.

- Andris, Sebastian and Jürgen Hubbuch (2020). „Modeling of hydrophobic interaction chromatography for the separation of antibody-drug conjugates and its application towards quality by design“. In: *Journal of Biotechnology* 317.April, pp. 48–58.
- Andris, Sebastian, Matthias Rüdte, et al. (12/2018). „Monitoring of antibody-drug conjugation reactions with UV/Vis spectroscopy“. In: *Journal of Biotechnology* 288.October, pp. 15–22.
- Andris, Sebastian, Jonathan Seidel, and Jürgen Hubbuch (2019). „Kinetic reaction modeling for antibody-drug conjugate process development“. In: *Journal of Biotechnology* 306.September, pp. 71–80.
- Antosiewicz, Jan M. and David Shugar (05/2016). „UV–Vis spectroscopy of tyrosine side-groups in studies of protein structure. Part 2: selected applications“. en. In: *Biophysical Reviews* 8. Open access review article, pp. 163–177.
- Ashton, Lorna and Ewan W. Blanch (06/2010). „pH-induced conformational transitions in α -lactalbumin investigated with two-dimensional Raman correlation variance plots and moving windows“. In: *Journal of Molecular Structure. PROGRESS IN TWO-DIMENSIONAL CORRELATION SPECTROSCOPY* 974.1, pp. 132–138.
- Atkins, Peter and Julio de Paula (2006). *Physical Chemistry*. Oxford, GB: Oxford University Press.
- Bakeev, Katherine A (2005). *Process Analytical Technology*. Oxford, GB: Blackwell Publishing.
- Balabin, Roman M. and Sergey V. Smirnov (04/2011). „Variable selection in near-infrared spectroscopy: Benchmarking of feature selection methods on biodiesel data“. In: *Analytica Chimica Acta* 692.1-2, pp. 63–72.
- Banner, Matthew et al. (12/2021). „A decade in review: use of data analytics within the biopharmaceutical sector“. In: *Current Opinion in Chemical Engineering* 34, p. 100758.
- Barman, Ishan et al. (02/2010). „Development of Robust Calibration Models Using Support Vector Machines for Spectroscopic Monitoring of Blood Glucose“. In: *Analytical Chemistry* 82.23, pp. 9719–9726.
- Barnett, Gregory V. et al. (12/2015). „Structural Changes and Aggregation Mechanisms for Anti-Streptavidin IgG1 at Elevated Concentration“. eng. In: *The Journal of Physical Chemistry. B* 119.49, pp. 15150–15163.
- Barth, Andreas (2007). „Infrared Spectroscopy of Proteins“. In: *Biochimica et Biophysica Acta - Bioenergetics* 1767.9, pp. 1073–1101.
- Barth, Andreas and Christian Zscherp (11/2002). „What Vibrations Tell about Proteins“. In: *Quarterly Reviews of Biophysics* 35.4, pp. 369–430.
- Basalla, Andrew J. and Brent S. Kendrick (12/2023). „Correcting Ultraviolet-Visible Spectra for Baseline Artifacts“. In: *Journal of Pharmaceutical Sciences* 112.12, pp. 3240–3247.
- Bazylewski, Paul, Ranjith Divigalpitiya, and Giovanni Fanchini (01/2017). „In situ Raman spectroscopy distinguishes between reversible and irreversible thiol modifications in L-cysteine“. en. In: *RSC Advances* 7.5, pp. 2964–2970.
- Beer’s Law* (02/2023). [Online; accessed 2025-12-24].
- Belle, Vaishak and Ioannis Papantonis (07/2021). „Principles and Practice of Explainable Machine Learning“. In: *Frontiers in Big Data* 4.July, pp. 1–25.
- Ben-David, Avishai and Hsuan Ren (2005). „Comparison between orthogonal subspace projection and background subtraction techniques applied to remote-sensing data“. In: *Applied Optics* 44.18, pp. 3846–3855.
- Bergstra, J., D. Yamins, and D. D. Cox (2012). *Making a Science of Model Search*.

- Berry, Brandon et al. (2015). „Cross-scale predictive modeling of CHO cell culture growth and metabolites using Raman spectroscopy and multivariate analysis“. In: *Biotechnology Progress* 31.2, pp. 566–577.
- Bjerrum, Esben Jannik, Mads Glahder, and Thomas Skov (10/2017). „Data Augmentation of Spectral Data for Convolutional Neural Network (CNN) Based Deep Chemometrics“. In: pp. 1–10.
- Blazhko, Uladzislau et al. (08/2021). „Comparison of augmentation and pre-processing for deep learning and chemometric classification of infrared spectra“. en. In: *Chemometrics and Intelligent Laboratory Systems* 215, p. 104367.
- Bocklitz, Thomas et al. (2011). „How to pre-process Raman spectra for reliable and stable models?“ In: *Analytica Chimica Acta* 704.1-2, pp. 47–56.
- Boelens, Hans F M et al. (2004). „New background correction method for liquid chromatography with diode array detection, infrared spectroscopic detection and Raman spectroscopic detection“. In: *Journal of Chromatography A* 1057.1-2, pp. 21–30.
- Boulet-Audet, Maxime, Bernadette Byrne, and Sergei G. Kazarian (10/2014). „High-Throughput Thermal Stability Analysis of a Monoclonal Antibody by Attenuated Total Reflection FT-IR Spectroscopic Imaging“. In: *Analytical Chemistry* 86.19, pp. 9786–9793.
- Brestrich, Nina, Tobias Hahn, and Jürgen Hubbuch (03/2016). „Application of spectral deconvolution and inverse mechanistic modelling as a tool for root cause investigation in protein chromatography“. In: *Journal of Chromatography A* 1437, pp. 158–167.
- Brestrich, Nina, Matthias Rüdert, et al. (02/2018). „Selective protein quantification for preparative chromatography using variable pathlength UV/Vis spectroscopy and partial least squares regression“. In: *Chemical Engineering Science* 176.C, pp. 157–164.
- Brestrich, Nina, Adrian Sanden, et al. (07/2015). „Advances in inline quantification of co-eluting proteins in chromatography: Process-data-based model calibration and application towards real-life separation issues“. eng. In: *Biotechnology and Bioengineering* 112.7, pp. 1406–1416.
- Brewster, Victoria L., Lorna Ashton, and Royston Goodacre (08/2011). „Monitoring the Glycosylation Status of Proteins Using Raman Spectroscopy“. In: *Analytical Chemistry* 83.15, pp. 6074–6081.
- Briskot, Till et al. (04/2022). „Modeling the Gibbs–Donnan effect during ultrafiltration and diafiltration processes using the Poisson–Boltzmann theory in combination with a basic Stern model“. In: *Journal of Membrane Science* 648, p. 120333.
- Brunel, Benjamin, Fatima Alsamad, and Olivier Piot (01/2021). „Toward automated machine learning in vibrational spectroscopy - use and settings of genetic algorithms for pre-processing and regression optimization“. In: *Chemometrics and Intelligent Laboratory Systems* 219.104444, pp. 1–10.
- Buckley, Kevin and Alan G. Ryder (2017). „Applications of Raman Spectroscopy in Biopharmaceutical Manufacturing: A Short Review“. In: *Applied Spectroscopy* 71.6, pp. 1085–1116.
- Bulcha, Jote T. et al. (02/08/2021). „Viral Vector Platforms within the Gene Therapy Landscape“. In: *Signal Transduction and Targeted Therapy* 6.1, p. 53.
- Burkart, Nadia and Marco F. Huber (01/2021). „A Survey on the Explainability of Supervised Machine Learning“. In: *Journal of Artificial Intelligence Research* 70, pp. 245–317.
- Butler, Holly J. et al. (2016). „Using Raman Spectroscopy to Characterize Biological Materials“. In: *Nature Protocols* 11.4, pp. 664–687.

- Cabaneros Lopez, Pau et al. (2020). „Towards a digital twin: a hybrid data-driven and mechanistic digital shadow to forecast the evolution of lignocellulosic fermentation“. In: *Biofuels, Bioproducts and Biorefining* 14.5, pp. 1046–1060.
- Cai, Wensheng, Yankun Li, and Xueguang Shao (2008). „A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra“. In: *Chemometrics and Intelligent Laboratory Systems* 90.2, pp. 188–194.
- Cao, Mingyan et al. (2019). „Site-specific antibody-drug conjugate heterogeneity characterization and heterogeneity root cause analysis“. In: *mAbs* 11.6, pp. 1064–1076.
- Capito, Florian, Romas Skudas, Harald Kolmar, et al. (2015). „At-line mid infrared spectroscopy for monitoring downstream processing unit operations“. In: *Process Biochemistry* 50.6, pp. 997–1005.
- Capito, Florian, Romas Skudas, Bernd Stanislawski, et al. (01/2013). „Matrix effects during monitoring of antibody and host cell proteins using attenuated total reflection spectroscopy“. In: *Biotechnology Progress* 29.1, pp. 265–274.
- Carta, Giorgio and Alois Jungbauer (2010). *Protein Chromatography - Process Development and Scale-Up*. Weinheim, Germany: Wiley-VCH Verlag GmbH und Co. KGaA.
- Castelli, María Sofía, Paul McGonigle, and Pamela J. Hornby (12/2019). „The Pharmacology and Therapeutic Applications of Monoclonal Antibodies“. In: *Pharmacology Research & Perspectives* 7.6, e00535.
- Cavinato, Anna G. et al. (09/1990). „Noninvasive method for monitoring ethanol in fermentation processes using fiber-optic near-infrared spectroscopy“. In: *Analytical Chemistry* 62.18, pp. 2058–2063.
- Chackerian, Bryce (2007). „Virus-like particles: Flexible platforms for vaccine development“. In: *Expert Review of Vaccines* 6.3, pp. 381–390.
- Chen, Hongshu, Bhavik R. Bakshi, and Prem K. Goel (2007). „Toward Bayesian chemometrics-A tutorial on some recent advances“. In: *Analytica Chimica Acta* 602.1, pp. 1–16.
- Chen, Jingyi et al. (03/2024). „Application of Raman spectroscopy during pharmaceutical process development for determination of critical quality attributes in Protein A chromatography“. In: *Journal of Chromatography A* 1718, p. 464721.
- Chen, Tao, Julian Morris, and Elaine Martin (2007). „Gaussian process regression for multivariate spectroscopic calibration“. In: *Chemometrics and Intelligent Laboratory Systems* 87.1, pp. 59–71.
- Chen, Tao and Bo Wang (2010). „Bayesian variable selection for Gaussian process regression: Application to chemometric calibration of spectrometers“. In: *Neurocomputing* 73.13-15, pp. 2718–2726.
- Chen, Yuan-yuan and Zhi-bin Wang (08/2019). „Feature selection based convolutional neural network pruning and its application in calibration modeling for NIR spectroscopy“. In: *Chemometrics and Intelligent Laboratory Systems* 191, pp. 103–108.
- Christler, Anna et al. (10/2021). „Technology transfer of a monitoring system to predict product concentration and purity of biopharmaceuticals in real-time during chromatographic separation“. In: *Biotechnology and Bioengineering* 118.10, pp. 3941–3952.
- Cooper, Arik and Yosef Shaul (02/2005). „Recombinant viral capsids as an efficient vehicle of oligonucleotide delivery into cells“. In: *Biochemical and Biophysical Research Communications* 327.4, pp. 1094–1099.
- Coumans, Ruud G.E. et al. (2020). „A Platform for the Generation of Site-Specific Antibody-Drug Conjugates That Allows for Selective Reduction of Engineered Cysteines“. In: *Bioconjugate Chemistry* 31.9, pp. 2136–2146.

- Covert, Ian, Scott Lundberg, and Su-In Lee (11/2020a). „Explaining by Removing: A Unified Framework for Model Explanation“. In: *Journal of Machine Learning Research* 22, pp. 1–90.
- Covert, Ian, Scott Lundberg, and Su-In Lee (04/2020b). „Understanding Global Feature Contributions With Additive Importance Measures“. In: *Advances in Neural Information Processing Systems* 2020-Decem.NeurIPS.
- Cui, Chenhao and Tom Fearn (02/2017). „Comparison of partial least squares regression, least squares support vector machines, and Gaussian process regression for a near infrared calibration“. en. In: *Journal of Near Infrared Spectroscopy* 25.1 (1), pp. 5–14.
- Cui, Chenhao and Tom Fearn (11/2018). „Modern practical convolutional neural networks for multivariate regression: Applications to NIR calibration“. In: *Chemometrics and Intelligent Laboratory Systems* 182, pp. 9–20.
- Curtis, R. A., J. M. Prausnitz, and H. W. Blanch (01/1998). „Protein-protein and protein-salt interactions in aqueous protein solutions containing concentrated electrolytes“. In: *Biotechnology and Bioengineering* 57.1, pp. 11–21.
- Dietrich, Annabelle et al. (05/2024). „Raman-based PAT for VLP precipitation: systematic data diversification and preprocessing pipeline identification“. en. In: *Frontiers in Bioengineering and Biotechnology* 12, p. 1399938.
- Effio, Christopher Ladd and Jürgen Hubbuch (05/2015). „Next generation vaccines and vectors: Designing downstream processes for recombinant protein-based virus-like particles“. In: *Biotechnology Journal* 10.5, pp. 715–727.
- Eilers, Paul H. C. (2003). „A Perfect Smoother“. In: *Analytical Chemistry* 75.14, pp. 3631–3636.
- Eisenberg, Azaria Solomon and Laura J. Juszczak (07/2012). „Correlation of TrpGly and GlyTrp Rotamer Structure with W7 and W10 UV Resonance Raman Modes and Fluorescence Emission Shifts“. In: *Journal of Amino Acids* 2012, pp. 1–10.
- Elsheikh, Mohamed et al. (2021). „A Comparative Review of Multi-Rate Moving Horizon Estimation Schemes for Bioprocess Applications“. In: *Computers & Chemical Engineering* 146, p. 107219.
- EMA (2014). „Guideline on the Use of near Infrared Spectroscopy by the Pharmaceutical Industry and the Data Requirements for New Submissions and Variations“. In: *EMA/CHMP/CVMP/QWP/17760/2009 Rev1*.
- Erb, Donald (n.d.). *pybaselines: A Python library of algorithms for the baseline correction of experimental data*.
- Ettah, Ilokugbe and Lorna Ashton (07/2018). „Engaging with Raman Spectroscopy to Investigate Antibody Aggregation“. en. In: *Antibodies* 7.3, pp. 24–24.
- Falkenberg, Heiner et al. (05/2019). „Mass spectrometric evaluation of upstream and downstream process influences on host cell protein patterns in biopharmaceutical products“. In: *Biotechnology Progress* 35.3.
- FDA (2004). „Guidance for Industry: PAT—a framework for innovative pharmaceutical development, manufacturing, and quality assurance“. In.
- FDA (2015). „Development and Submission of Near Infrared Analytical Procedures Guidance for Industry“. In: March, p. 24.
- FDA (2023). *Artificial Intelligence in Drug Manufacturing*.
- Feidl et al. (10/2019). „Combining Mechanistic Modeling and Raman Spectroscopy for Monitoring Antibody Chromatographic Purification“. In: *Processes* 7.10, pp. 683–683.
- Feng, Steven Y. et al. (2021). „A Survey of Data Augmentation Approaches for NLP“. In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 968–988.

- Feurer, Matthias and Frank Hutter (2019). „Hyperparameter Optimization“. In: *Automated Machine Learning: Methods, Systems, Challenges*. Ed. by Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren. Cham: Springer International Publishing, pp. 3–33.
- Flatman, Stephen et al. (03/2007). „Process analytics for purification of monoclonal antibodies“. In: *Journal of Chromatography B* 848.1, pp. 79–87.
- Fontana, Marc D., Kawther Ben Mabrouk, and Thomas H. Kauffmann (2013). „Raman spectroscopic sensors for inorganic salts“. In: *Spectroscopic Properties of Inorganic and Organometallic Compounds*. Ed. by Jack Yarwood, Richard Douthwaite, and Simon Duckett. Vol. 44. RSC Publishing, pp. 40–67.
- Fu, Zhiwen et al. (03/22/2022). „Antibody Drug Conjugate: The “Biological Missile” for Targeted Cancer Therapy“. In: *Signal Transduction and Targeted Therapy* 7.1, p. 93.
- Furić, K, I Ciglencečki, and B Čosović (09/2000). „Raman spectroscopic study of sodium chloride water solutions“. In: *Journal of Molecular Structure* 550-551, pp. 225–234.
- Gargalo, Carina L. et al. (2021). „Towards the Development of Digital Twins for the Bio-manufacturing Industry“. In: *Digital Twins: Tools and Concepts for Smart Biomanufacturing*. Ed. by Christoph Herwig, Ralf Pörtner, and Johannes Möller. Cham: Springer International Publishing, pp. 1–34.
- Gasteiger, Elisabeth et al. (2005). „Protein Identification and Analysis Tools on the ExPASy Server“. In: *The Proteomics Protocols Handbook*, pp. 571–607.
- Gaiglitz, Günter and David S. Moore (2014). *Handbook of Spectroscopy*. Ed. by Gaiglitz Günter and David S. Moore. Vol. 4. Weinheim: Wiley-VCH.
- Gautam, Rekha et al. (2015). „Review of multidimensional data processing approaches for Raman and infrared spectroscopy“. In: *EPJ Techniques and Instrumentation* 2.1.
- Genova, E. et al. (2018). „SERS of cells: What can we learn from cell lysates?“ In: *Analytica Chimica Acta* 1005, pp. 93–100.
- Gerretzen, Jan et al. (09/2015). „Simple and Effective Way for Data Preprocessing Selection Based on Design of Experiments“. In: *Analytical Chemistry* 87.24, pp. 12096–12103.
- Gillespie, Christopher et al. (2022). „Systematic assessment of process analytical technologies for biologics“. en. In: *Biotechnology and Bioengineering* 119.2, pp. 423–434.
- Glassey, Jarka et al. (04/2011). „Process analytical technology (PAT) for biopharmaceuticals“. In: *Biotechnology Journal* 6.4, pp. 369–377.
- Golabgir, Aydin and Christoph Herwig (2016). „Combining Mechanistic Modeling and Raman Spectroscopy for Real-Time Monitoring of Fed-Batch Penicillin Production“. In: *Chemie-Ingenieur-Technik* 88.6, pp. 764–776.
- Goldrick, Stephen, David Lovett, et al. (2018). „Influence of incident wavelength and detector material selection on fluorescence in the application of raman spectroscopy to a fungal fermentation process“. In: *Bioengineering* 5.4, pp. 1–17.
- Goldrick, Stephen, Alexandra Umprecht, et al. (09/2020). „High-Throughput Raman Spectroscopy Combined with Innovate Data Analysis Workflow to Enhance Biopharmaceutical Process Development“. In: *Processes* 8.9 (9), pp. 1179–1179.
- Gómez de la Cuesta, Ramón, Royston Goodacre, and Lorna Ashton (11/2014). „Monitoring Antibody Aggregation in Early Drug Development Using Raman Spectroscopy and Perturbation-Correlation Moving Windows“. In: *Analytical Chemistry* 86.22, pp. 11133–11140.
- Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. MIT Press.
- Goodfellow, Ian J., Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. Cambridge, MA, USA: MIT Press.

- Grön, Heidi et al. (2003). „Dynamic in-process examination of particle size and crystallographic form under defined conditions of reactant supersaturation as associated with the batch crystallization of monosodium glutamate from aqueous solution“. In: *Industrial and Engineering Chemistry Research* 42.20, pp. 4888–4898.
- Großhans, Steffen, Susanna Suhm, and Jürgen Hubbuch (2019). „Precipitation of complex antibody solutions: influence of contaminant composition and cell culture medium on the precipitation behavior“. In: *Bioprocess and Biosystems Engineering*.
- Großhans, Steffen et al. (2018). „In-line Fourier-transform infrared spectroscopy as a versatile process analytical technology for preparative protein chromatography“. In: *Journal of Chromatography A* 1547, pp. 37–44.
- Gu, Jiuxiang et al. (05/2018). „Recent advances in convolutional neural networks“. In: *Pattern Recognition* 77, pp. 354–377.
- Guardalini, Luis Giovanni de Oliveira, Vinícius Aragão Tejo Dias, et al. (01/2023). „Comparison of chemometric models using Raman spectroscopy for offline biochemical monitoring throughout the VLP-making upstream process“. In: *Biochemical Engineering Journal* 198, pp. 109013–109013.
- Guardalini, Luis Giovanni de Oliveira, Rafaela Morais Rangel, et al. (12/2023). „Monitoring by Raman spectroscopy of rabies virus-like particles production since the initial development stages“. en. In: *Journal of Chemical Technology & Biotechnology* n/a.n/a.
- Guardalini, Luis Giovanni de Oliveira, Paulo Eduardo da Silva Cavalcante, et al. (2023). „Biochemical monitoring throughout all stages of rabies virus-like particles production by Raman spectroscopy using global models“. eng. In: *Journal of Biotechnology* 363, pp. 19–31.
- Guindo, Mahamed Lamine et al. (01/2023). „Chemometric Approach Based on Explainable AI for Rapid Assessment of Macronutrients in Different Organic Fertilizers Using Fusion Spectra“. In: *Molecules* 28.2.
- Gunter Jagschies et al. (01/2018). *Biopharmaceutical Processing : Development, Design, and Implementation of Manufacturing Processes*. Amsterdam, Netherlands: Elsevier.
- Guo, Fan, Ruimin Xie, and Biao Huang (02/2020). „A deep learning just-in-time modeling approach for soft sensor based on variational autoencoder“. In: *Chemometrics and Intelligent Laboratory Systems* 197, p. 103922.
- Guo, Shuxia, Thomas Bocklitz, and Jürgen Popp (2016). „Optimization of Raman-spectrum baseline correction in biological application“. In: *Analyst* 141 (8), pp. 2396–2404.
- Haghi, R. K., E. Pérez-Fernández, and A. H.J. Robertson (08/2021). „Prediction of various soil properties for a national spatial dataset of Scottish soils based on four different chemometric approaches: A comparison of near infrared and mid-infrared spectroscopy“. In: *Geoderma* 396.
- Hämmerling, Frank et al. (01/2017). „Investigation and prediction of protein precipitation by polyethylene glycol using quantitative structure–activity relationship models“. In: *Journal of Biotechnology* 241, pp. 87–97.
- Hansen, Sigrid K., Babak Jamali, and Jürgen Hubbuch (02/2013). „Selective high throughput protein quantification based on UV absorption spectra“. eng. In: *Biotechnology and Bioengineering* 110.2, pp. 448–460.
- Hansen, Sigrid K., Erik Skibsted, et al. (11/2011). „A label-free methodology for selective protein quantification by means of absorption measurements“. eng. In: *Biotechnology and Bioengineering* 108.11, pp. 2661–2669.
- Harada, Issei, Takashi Miura, and Hideo Takeuchi (01/1986). „Origin of the doublet at 1360 and 1340 cm⁻¹ in the Raman spectra of tryptophan and related compounds“. en. In: *Spectrochimica Acta Part A: Molecular Spectroscopy* 42.2-3, pp. 307–312.

- Harner, Richard S. et al. (2009). „Use of a fiber-optic turbidity probe to monitor and control commercial-scale unseeded batch crystallizations“. In: *Organic Process Research and Development* 13.1, pp. 114–124.
- Harris, Charles R. et al. (09/2020). „Array programming with NumPy“. In: *Nature* 585.7825, pp. 357–362.
- Haseltine, Eric L. and James B. Rawlings (2005). „Critical evaluation of extended Kalman filtering and moving-horizon estimation“. In: *Industrial and Engineering Chemistry Research* 44.8, pp. 2451–2460.
- Hassebroek, Anna M. et al. (02/2023). „A hepatitis B virus core antigen-based virus-like particle vaccine expressing SARS-CoV-2 B and T cell epitopes induces epitope-specific humoral and cell-mediated immune responses but confers limited protection against SARS-CoV-2 infection“. In: *Journal of Medical Virology* 95.2.
- Hassoun, Mohamed et al. (2018). „A droplet-based microfluidic chip as a platform for leukemia cell lysate identification using surface-enhanced Raman scattering“. In: *Analytical and Bioanalytical Chemistry* 410.3, pp. 999–1006.
- He, Hao et al. (01/2024). „Noise learning of instruments for high-contrast, high-resolution and fast hyperspectral microscopy and nanoscopy“. In: *Nature Communications* 15.1, p. 754.
- He, Shixuan et al. (2014). „Baseline correction for Raman spectra using an improved asymmetric least squares method“. In: *Anal. Methods* 6.12, pp. 4402–4407.
- Helleckes, Laura M. et al. (06/2023). „Machine Learning in Bioprocess Development: From Promise to Practice“. In: *Trends in Biotechnology* 41.6, pp. 817–835.
- Heyer-Müller, Jakob et al. (2025). „A novel Raman-chromatography assembly for automated calibration and in-line monitoring in bioprocessing“. In: *Engineering in Life Sciences*.
- Hillebrandt, Nils and Jürgen Hubbuch (05/2023). „Size-selective downstream processing of virus particles and non-enveloped virus-like particles“. In: *Frontiers in Bioengineering and Biotechnology* 11.
- Hillebrandt, Nils, Philipp Vormittag, Nicolai Bluthardt, et al. (2020). „Integrated Process for Capture and Purification of Virus-Like Particles: Enhancing Process Performance by Cross-Flow Filtration“. In: *Frontiers in Bioengineering and Biotechnology* 8.May.
- Hillebrandt, Nils, Philipp Vormittag, Annabelle Dietrich, and Jürgen Hubbuch (06/2022). „Process monitoring framework for cross-flow diafiltration-based virus-like particle disassembly: Tracing product properties and filtration performance“. In: *Biotechnology and Bioengineering* 119.6, pp. 1522–1538.
- Hillebrandt, Nils, Philipp Vormittag, Annabelle Dietrich, Christina H. Wegner, et al. (2021). „Process development for cross-flow diafiltration-based VLP disassembly: A novel high-throughput screening approach“. In: *Biotechnology and Bioengineering* 118.10, pp. 3926–3940.
- Hollas, J. Michael (2010). *Modern Spectroscopy*. 4. ed., reprinted. Chichester: Wiley.
- Hong, Moo Sun et al. (2018). „Challenges and opportunities in biopharmaceutical manufacturing control“. In: *Computers & Chemical Engineering* 110, pp. 106–114.
- Hu, Xi et al. (2017). „Development of Commercial-Ready Processes for Antibody Drug Conjugates“. In: *Organic Process Research and Development* 21.4, pp. 601–610.
- Huang, Hongxia and Haibin Qu (2018). „In-situ monitoring of saccharides removal of alcohol precipitation using near-infrared spectroscopy“. In: *Journal of Innovative Optical Health Sciences* 11.5, pp. 1–12.

- Hutter, Frank, Lars Kotthoff, and Joaquin Vanschoren, eds. (2019). *Automated Machine Learning: Methods, Systems, Challenges*. The Springer Series on Challenges in Machine Learning. Cham: Springer International Publishing.
- ICH (2009a). *ICH guideline Q8 (R2) pharmaceutical development*.
- ICH (2009b). *Q8(R2) Pharmaceutical Development*. Tech. rep. June.
- Ioffe, Sergey and Christian Szegedy (2015). „Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift“. In: *32nd International Conference on Machine Learning, ICML 2015* 1, pp. 448–456.
- Iverius, P.H. and T.C. Laurent (02/1967). „Precipitation of some plasma proteins by the addition of dextran or polyethylene glycol“. In: *Biochimica et Biophysica Acta (BBA) - Protein Structure* 133.2, pp. 371–373.
- James, G. et al. (2013). *An Introduction to Statistical Learning*. 244–251.
- Jiang, Mo et al. (11/2017). „Opportunities and challenges of real-time release testing in biopharmaceutical manufacturing“. In: *Biotechnology and Bioengineering* 114 (11), pp. 2445–2456.
- Joubert, Nicolas et al. (2020). „Antibody–Drug Conjugates: The Last Decade“. In: *Pharmaceuticals* 13.9.
- Kadlec, Petr, Bogdan Gabrys, and Sibylle Strandt (04/2009). „Data-Driven Soft Sensors in the Process Industry“. In: *Computers and Chemical Engineering* 33.4, pp. 795–814.
- Kazaks, Andris et al. (2017). „Production and purification of chimeric HBc virus-like particles carrying influenza virus LAH domain as vaccine candidates“. In: *BMC Biotechnology* 17.1, pp. 1–11.
- Kazemzadeh, Mohammadrahim et al. (05/2022). „Deep convolutional neural networks as a unified solution for Raman spectroscopy-based classification in biomedical applications“. In: *Optics Communications* 510, p. 127977.
- Keller, William R. and Michaela Wendeler (2021). „Using multimodal chromatography for post-conjugation antibody-drug conjugate purification: A methodology from high throughput screening to in-silico process development“. In: *Journal of Chromatography A* 1653, p. 462378.
- Kennard, R. W. and L. A. Stone (1969). „Computer Aided Design of Experiments“. In: *Technometrics* 11.1, pp. 137–148.
- Kessler, Waltraud (2007). *Multivariate Datenanalyse für die Pharma-, Bio- und Prozessanalytik - Ein Lehrbuch*. Wiley VCH Verlag GmbH & Co. KGaA.
- Kheirvari, Milad, Hong Liu, and Ebenezer Tumban (05/02/2023). „Virus-like Particle Vaccines and Platforms for Vaccine Development“. In: *Viruses* 15.5, p. 1109.
- Kim, Hyoung Jin et al. (2010). „One-step chromatographic purification of human papillomavirus type 16 L1 protein from *Saccharomyces cerevisiae*“. In: *Protein Expression and Purification* 70.1, pp. 68–74.
- Kingma, Diederik P. and Jimmy Lei Ba (2015). „Adam: A method for stochastic optimization“. In: *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pp. 1–15.
- Kjeldahl, Karin and Rasmus Bro (07/2010). „Some common misunderstandings in chemometrics“. en. In: *Journal of Chemometrics* 24.7–8, pp. 558–564.
- Klamp, Thorsten et al. (01/2011). „Highly Specific Auto-Antibodies against Claudin-18 Isoform 2 Induced by a Chimeric HBcAg Virus-Like Particle Vaccine Kill Tumor Cells and Inhibit the Growth of Lung Metastases“. In: *Cancer Research* 71.2, pp. 516–527.

- Klaverdijk, Maarten, Marcel Ottens, and Marieke E. Klijn (2025). „Single compound data supplementation to enhance transferability of fermentation specific Raman spectroscopy models“. In: *Analytical and Bioanalytical Chemistry*, pp. 1–12.
- Kohler, A. et al. (03/2008). „Estimating and Correcting Mie Scattering in Synchrotron-Based Microscopic Fourier Transform Infrared Spectra by Extended Multiplicative Signal Correction“. In: *Applied Spectroscopy* 62.3, pp. 259–266.
- Koho, Tiia et al. (2012). „Purification of norovirus-like particles (VLPs) by ion exchange chromatography“. In: *Journal of Virological Methods* 181.1, pp. 6–11.
- Kolås, S., B. A. Foss, and T. S. Schei (2009). „Constrained nonlinear state estimation based on the UKF approach“. In: *Computers and Chemical Engineering* 33.8, pp. 1386–1401.
- Korepanov, Vitaly I. (10/2020). „Asymmetric least-squares baseline algorithm with peak screening for automatic processing of the Raman spectra“. en. In: *Journal of Raman Spectroscopy* 51.10, pp. 2061–2065.
- Krause, Julius et al. (04/2021). „New active learning algorithms for near-infrared spectroscopy in agricultural applications“. In: *at - Automatisierungstechnik* 69.4, pp. 297–306.
- Kriesten, E. et al. (2008). „Fully automated indirect hard modeling of mixture spectra“. In: *Chemometrics and Intelligent Laboratory Systems* 91.2, pp. 181–193.
- Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton (05/2017). „ImageNet classification with deep convolutional neural networks“. en. In: *Communications of the ACM* 60.6, pp. 84–90.
- Kushnir, Natasha, Stephen J. Streatfield, and Vidadi Yusibov (12/2012). „Virus-like particles as a highly efficient vaccine platform: Diversity of targets and production systems and advances in clinical development“. In: *Vaccine* 31.1, pp. 58–83.
- Kuzmin, V. V. et al. (10/2020). „Raman spectra of polyethylene glycols: Comparative experimental and DFT study“. In: *Journal of Molecular Structure* 1217, p. 128331.
- Lakowicz, J.R. (2006). *Principles of Fluorescence Spectroscopy*. Springer US.
- Lange, Christoph et al. (2024). „Data Augmentation Scheme for Raman Spectra with Highly Correlated Annotations“. In: *34th European Symposium on Computer Aided Process Engineering / 15th International Symposium on Process Systems Engineering*. Ed. by Flavio Manenti and Gintaras V. Reklaitis. Vol. 53. Computer Aided Chemical Engineering. Elsevier, pp. 3055–3060.
- Lewis, E. Neil et al. (12/2014). „Combined Dynamic Light Scattering and Raman Spectroscopy Approach for Characterizing the Aggregation of Therapeutic Proteins“. en. In: *Molecules* 19.12, pp. 20888–20905.
- Li, Boyan et al. (2010). „Rapid characterization and quality control of complex cell culture media solutions using Raman spectroscopy and chemometrics“. In: *Biotechnology and Bioengineering* 107.2, pp. 290–301.
- Li, Cynthia H. and Tiansheng Li (2009). „Application of Vibrational Spectroscopy to the Structural Characterization of Monoclonal Antibody and its Aggregate“. en. In: *Current Pharmaceutical Biotechnology* 10.4, pp. 391–399.
- Li, Lian et al. (2014). „The relevance study of effective information between near infrared spectroscopy and chondroitin sulfate in ethanol precipitation process“. In: *Journal of Innovative Optical Health Sciences* 7.6, pp. 1–7.
- Li, Meng-Yao et al. (03/2018). „Real-time monitoring of antibody glycosylation site occupancy by in situ Raman spectroscopy during bioreactor CHO cell cultures“. In: *Biotechnology Progress* 34.2 (2), pp. 486–493.

- Liland, Kristian Hovde, Achim Kohler, and Nils Kristian Afseth (06/2016). „Model-based pre-processing in Raman spectroscopy of biological samples“. en. In: *Journal of Raman Spectroscopy* 47.6, pp. 643–650.
- Lippert, J. L., D. Tyminski, and P. J. Desmeules (10/1976). „Determination of the secondary structure of proteins by laser Raman spectroscopy“. en. In: *Journal of the American Chemical Society* 98.22, pp. 7075–7080.
- Liu, Jinchao et al. (2017). „Deep convolutional neural networks for Raman spectrum recognition: A unified solution“. In: *Analyst* 142.21 (21), pp. 4067–4074.
- Liu, Yi, Markus Pietzsch, and Joachim Ulrich (2014). „Determination of the phase diagram for the crystallization of L-asparaginase II by a turbidity technique“. In: *Crystal Research and Technology* 49.4, pp. 262–268.
- Lohmann, Lara Julia and Jochen Strube (2021). „Process analytical technology for precipitation process integration into biologics manufacturing towards autonomous operation–mab case study“. In: *Processes* 9.3.
- Long, James R., Vasilis G. Gregoriou, and Paul J. Gemperline (09/1990). „Spectroscopic calibration and quantitation using artificial neural networks“. In: *Analytical Chemistry* 62.17 (17), pp. 1791–1797.
- Lorber, Avraham (1986). „Error Propagation and Figures of Merit for Quantification by Solving Matrix Equations“. In: *Analytical Chemistry* 58.6, pp. 1167–1172.
- Loving, Galen S., Matthieu Sainlos, and Barbara Imperiali (02/2010). „Monitoring Protein Interactions and Dynamics with Solvatochromic Fluorophores“. In: *Trends in Biotechnology* 28.2, pp. 73–83.
- Lundberg, Scott M, Paul G Allen, and Su-In Lee (2017). „A Unified Approach to Interpreting Model Predictions“. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30, pp. 4765–4774.
- Luttmann, Reiner et al. (2012). „Soft sensors in bioprocessing: A status report and recommendations“. In: *Biotechnology Journal* 7.8, pp. 1040–1048.
- Mahynski, Nathan A. et al. (10/2022). „Building Interpretable Machine Learning Models to Identify Chemometric Trends in Seabirds of the North Pacific Ocean“. In: *Environmental Science and Technology* 56.20, pp. 14361–14374.
- Maiti, Nakul C. et al. (03/2004). „Raman Spectroscopic Characterization of Secondary Structure in Natively Unfolded Proteins: α -Synuclein“. In: *Journal of the American Chemical Society* 126.8, pp. 2399–2408.
- Makki, Alaa A. et al. (02/2021). „Understanding the discrimination and quantification of monoclonal antibodies preparations using Raman spectroscopy“. In: *Journal of Pharmaceutical and Biomedical Analysis* 194, p. 113734.
- Malek, Salim, Farid Melgani, and Yakoub Bazi (2018). „One-dimensional convolutional neural networks for spectroscopic signal regression“. In: *Journal of Chemometrics* 32.5 (5), pp. 1–17.
- Markl, Daniel et al. (05/2020). „Review of real-time release testing of pharmaceutical tablets: State-of-the art, challenges and future perspective“. In: *International Journal of Pharmaceutics* 582, p. 119353.
- Martens, Harald, Jesper Pram Nielsen, and Seren Bailing Engelsen (2003). „Light scattering and light absorbance separated by extended multiplicative signal correction. Application to near-infrared transmission analysis of powder mixtures“. In: *Analytical Chemistry* 75.3, pp. 394–404.

- Martens, Harald and Edward Stark (01/1991a). „Extended multiplicative signal correction and spectral interference subtraction: New preprocessing methods for near infrared spectroscopy“. In: *Journal of Pharmaceutical and Biomedical Analysis* 9.8, pp. 625–635.
- Martens, Harald and Edward Stark (01/1991b). „Extended multiplicative signal correction and spectral interference subtraction: New preprocessing methods for near infrared spectroscopy“. In: *Journal of Pharmaceutical and Biomedical Analysis* 9.8, pp. 625–635.
- Mayerhöfer, Thomas G., Susanne Pahlow, and Jürgen Popp (09/2020). „The Bouguer-Beer-Lambert Law: Shining Light on the Obscure“. In: *ChemPhysChem* 21.18, pp. 2029–2046.
- McHardy, Rose G. et al. (08/2023). „Augmentation of FTIR spectral datasets using Wasserstein generative adversarial networks for cancer liquid biopsies“. en. In: *Analyst* 148.16, pp. 3860–3869.
- Mehmood, Tahir et al. (08/2012). „A review of variable selection methods in Partial Least Squares Regression“. In: *Chemometrics and Intelligent Laboratory Systems* 118, pp. 62–69.
- Meyer-Kirschner, Julian et al. (2016). „In-line Monitoring of Monomer and Polymer Content during Microgel Synthesis Using Precipitation Polymerization via Raman Spectroscopy and Indirect Hard Modeling“. In: *Applied Spectroscopy* 70.3, pp. 416–426.
- Milewska, Adrianna et al. (12/2023). „In-line monitoring of protein concentration with MIR spectroscopy during UFDF“. en. In: *Engineering in Life Sciences* 23.2, e2200050.
- Militello, Valeria et al. (02/2004). „Aggregation kinetics of bovine serum albumin studied by FTIR spectroscopy and light scattering“. In: *Biophysical Chemistry* 107.2, pp. 175–187.
- Miller, Charles E. (2007). „Chemometrics in Process Analytical Chemistry“. In: *Process Analytical Technology*. Blackwell Publishing, pp. 226–328.
- Mishra, Puneet and Ittai Herrmann (08/2021). „GAN meets chemometrics: Segmenting spectral images with pixel2pixel image translation with conditional generative adversarial networks“. In: *Chemometrics and Intelligent Laboratory Systems* 215, p. 104362.
- Mishra, Puneet and Dário Passos (05/2021a). „A synergistic use of chemometrics and deep learning improved the predictive performance of near-infrared spectroscopy models for dry matter prediction in mango fruit“. In: *Chemometrics and Intelligent Laboratory Systems* 212.March (March), p. 104287.
- Mishra, Puneet and Dário Passos (2021b). „Deep chemometrics: Validation and transfer of a global deep near-infrared fruit model to use it on a new portable instrument“. en. In: *Journal of Chemometrics* 35.10, e3367.
- Mishra, Puneet and Dário Passos (05/2021c). „Realizing transfer learning for updating deep learning models of spectral data to be used in new scenarios“. In: *Chemometrics and Intelligent Laboratory Systems* 212, p. 104283.
- Mobini, Saeed et al. (08/2020). „Computational Design of a Novel VLP-Based Vaccine for Hepatitis B Virus“. In: *Frontiers in Immunology* 11.
- Mohsen, Mona O. and Martin F. Bachmann (08/12/2022). „Virus-like Particle Vaccinology, from Bench to Bedside“. In: *Cellular & Molecular Immunology* 19.9, pp. 993–1011.
- Moleirinho, Mafalda G. et al. (05/2020). „Current challenges in biotherapeutic particles manufacturing“. In: *Expert Opinion on Biological Therapy* 20.5, pp. 451–465.
- Moradi Vahdat, Maryam et al. (03/2021). „Hepatitis B core-based virus-like particles: A platform for vaccine development in plants“. In: *Biotechnology Reports* 29, e00605.
- Moreno, A. et al. (2000). „Turbidity as a useful optical parameter to predict protein crystallization by dynamic light scattering“. In: *Journal of Molecular Structure* 519.1-3, pp. 243–256.
- Moussa, Ehab M. et al. (02/2016). „Immunogenicity of Therapeutic Protein Aggregates“. English. In: *Journal of Pharmaceutical Sciences* 105.2, pp. 417–430.

- Müller, David Heinrich, Marieke Börger, et al. (12/2023). „The Good pH probe: non-invasive pH in-line monitoring using Good buffers and Raman spectroscopy“. en. In: *Analytical and Bioanalytical Chemistry* 415.29, pp. 7247–7258.
- Müller, David Heinrich, Hans-Jürgen Koss, et al. (02/2024). „Bioprocess in-line monitoring and control using Raman spectroscopy and Indirect Hard Modeling (IHM)“. en. In: *Biotechnology and Bioengineering* n/a.n/a.
- Narayanan, Harini et al. (2020). „Hybrid-EKF: Hybrid model coupled with extended Kalman filter for real-time monitoring and control of mammalian cell culture“. In: *Biotechnology and Bioengineering* 117.9, pp. 2703–2714.
- Ng, Andrew Y (2004). „L1 and L2 Regularisation Comparisation“. In: *Proceedings of the 21 st International Conference on Machine Learning*.
- Nikzad-Langerodi, Ramin and Erik Andries (2021). „A chemometrician’s guide to transfer learning“. en. In: *Journal of Chemometrics* 35.11, e3373.
- Nikzad-Langerodi, Ramin, Edwin Lughofer, et al. (2018). „Calibration model maintenance in melamine resin production: Integrating drift detection, smart sample selection and model adaptation“. In: *Analytica Chimica Acta* 1013, pp. 1–12.
- Nooraei, Saghi et al. (02/2021). „Virus-like particles: preparation, immunogenicity and their roles as nanovaccines and drug nanocarriers“. In: *Journal of Nanobiotechnology* 19.1, p. 59.
- Nørgaard, L. et al. (03/2000). „Interval Partial Least-Squares Regression (*i* PLS): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy“. en. In: *Applied Spectroscopy* 54.3, pp. 413–419.
- Ohadi, Kaveh, Raymond L. Legge, and Hector M. Budman (2015). „Development of a soft-sensor based on multi-wavelength fluorescence spectroscopy and a dynamic metabolic model for monitoring mammalian cell cultures“. In: *Biotechnology and Bioengineering* 112.1, pp. 197–208.
- Ota, Chikashi et al. (04/2016). „Assessment of the Protein–Protein Interactions in a Highly Concentrated Antibody Solution by Using Raman Spectroscopy“. en. In: *Pharmaceutical Research* 33.4, pp. 956–969.
- Otto, Matthias (2017). *Chemometrics - Statistics and Computer Application in Analytical Chemistry*. Wiley VCH Verlag GmbH & Co. KGaA.
- Ouyang, Jun (2013). „Drug-to-antibody ratio (DAR) and drug load distribution by hydrophobic interaction chromatography and reversed phase high-performance liquid chromatography“. In: *Antibody-Drug Conjugates*. Springer, pp. 275–283.
- Paananen, Topi et al. (12/2019). „Variable selection for Gaussian processes via sensitivity analysis of the posterior predictive distribution“. In: *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)* PLMR.89, pp. 1743–1752.
- Panowski, Siler et al. (2014). „Site-specific antibody drug conjugates for cancer therapy“. In: *mAbs* 6.1, pp. 34–45.
- Passos, Cláudia P. et al. (2010). „Application of Fourier transform infrared spectroscopy and orthogonal projections to latent structures/partial least squares regression for estimation of procyanidins average degree of polymerisation“. In: *Analytica Chimica Acta* 661.2, pp. 143–149.
- Passos, Dário and Puneet Mishra (08/2021). „An automated deep learning pipeline based on advanced optimisations for leveraging spectral classification modelling“. In: *Chemometrics and Intelligent Laboratory Systems* 215.May (May), p. 104354.
- Passos, Dário and Puneet Mishra (04/2022). „A tutorial on automatic hyperparameter tuning of deep spectral modelling for regression and classification tasks“. In: *Chemometrics and Intelligent Laboratory Systems* 223, p. 104520.

- Pedregosa, F. et al. (2011). „Scikit-learn: Machine Learning in Python“. In: *Journal of Machine Learning Research* 12, pp. 2825–2830.
- Peticolas, Warner L. (1995). „Raman spectroscopy of DNA and proteins“. In: *Methods in Enzymology*. Vol. 246. Elsevier Inc., pp. 389–416.
- Petrovskis, Ivars et al. (01/2021). „Production of the HBc Protein from Different HBV Genotypes in E. coli. Use of Reassociated HBc VLPs for Packaging of ss- and dsRNA“. In: *Microorganisms* 9.2, p. 283.
- Pham, Ngoc B. and Wilson S. Meng (07/2020). „Protein aggregation and immunogenicity of biotherapeutics“. en. In: *International Journal of Pharmaceutics* 585, p. 119523.
- Poison, A. (01/1977). „A Theory for the Displacement of Proteins and Viruses with Polyethylene Glycol“. In: *Preparative Biochemistry* 7.2, pp. 129–154.
- Pons Royo, Maria del Carme et al. (2022). „Mode and dosage time in polyethylene glycol precipitation process influences protein precipitate size and filterability“. In: *Process Biochemistry* 114, pp. 77–85.
- Pontius, Katrin et al. (06/2020). „Monitoring yeast fermentations by nonlinear infrared technology and chemometrics – understanding process correlations and indirect predictions“. In: *Applied Microbiology and Biotechnology* 104.12, pp. 5315–5335.
- Porterfield, J. Zachary and Adam Zlotnick (11/2010). „A simple and general method for determining the protein and nucleic acid content of viruses by UV absorbance“. In: *Virology* 407.2, pp. 281–288.
- Preston, Caroline M. and W. A. Adams (04/1979). „A Laser Raman Spectroscopic Study of Aqueous Orthophosphate Salts“. In: *The Journal of Physical Chemistry* 83.7, pp. 814–821.
- Prince, Simon J. D. (2023). *Understanding Deep Learning*. Cambridge, Massachusetts London, England: The MIT Press.
- Qian, Ciyang et al. (03/2020). „Recent Progress on the Versatility of Virus-Like Particles“. In: *Vaccines* 8.1, p. 139.
- Raschka, Sebastian (11/2018). „Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning“. In.
- Rasmussen, Carl Edward and Christopher K. I. Williams (2006). *Gaussian processes for machine learning*. Ed. by Thomas Dietterich. Vol. 38. Cambridge, MA, USA: MIT Press.
- Rathore, A. S., R. Bhambure, and V. Ghare (09/2010). „Process analytical technology (PAT) for biopharmaceutical products“. In: *Analytical and Bioanalytical Chemistry* 398.1 (1), pp. 137–154.
- Read, E. K. et al. (2010). „Process Analytical Technology (PAT) for biopharmaceutical products: Part II. Concepts and applications“. In: *Biotechnology and Bioengineering* 105.2, pp. 285–295.
- Read, E.K. et al. (02/2010). „Process analytical technology (PAT) for biopharmaceutical products: Part I. concepts and applications“. In: *Biotechnology and Bioengineering* 105 (2), pp. 276–284.
- Rinnan, Åsmund et al. (2009). „Data Pre-processing“. In: *Infrared Spectroscopy for Food Quality Analysis and Control*. Vol. 3. Academic Press, pp. 29–50.
- Rischawy, Federico et al. (09/2023). „Modeling of biopharmaceutical UF/DF from laboratory to manufacturing scale“. In: *Computers & Chemical Engineering* 177, p. 108337.
- Roberts, Christopher J (12/2014). „Protein aggregation and its impact on product quality“. en. In: *Current Opinion in Biotechnology* 30, pp. 211–217.
- Rodríguez-Rodríguez, José A. et al. (2021). „The Impact of Linear Motion Blur on the Object Recognition Efficiency of Deep Convolutional Neural Networks“. en. In: *Pattern Recognition. ICPR International Workshops and Challenges*. Ed. by Alberto Del Bimbo et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 611–622.

- Roessler, Ulrich et al. (2014). „In situ protein secondary structure determination in ice: Raman spectroscopy-based process analytical tool for frozen storage of biopharmaceuticals“. In: *Journal of Pharmaceutical Sciences* 103.8, pp. 2287–2295.
- Rolinger, Laura, Jürgen Hubbuch, and Matthias Rüdts (02/2023). „Monitoring of ultra- and diafiltration processes by Kalman-filtered Raman measurements“. In: *Analytical and Bioanalytical Chemistry* 415.5, pp. 841–854.
- Rolinger, Laura, Matthias Rüdts, Juliane Diehm, et al. (02/2020). „Multi-attribute PAT for UF/DF of Proteins–Monitoring Concentration, particle sizes, and Buffer Exchange“. In: *Analytical and Bioanalytical Chemistry* 412.9, pp. 2123–2136.
- Rolinger, Laura, Matthias Rüdts, and Jürgen Hubbuch (04/2020a). „A critical review of recent trends, and a future perspective of optical spectroscopy as PAT in biopharmaceutical downstream processing“. In: *Analytical and Bioanalytical Chemistry* 412.9, pp. 2047–2064.
- Rolinger, Laura, Matthias Rüdts, and Jürgen Hubbuch (2020b). „A multi-sensor approach for improved protein A load phase monitoring by conductivity-based background subtraction of UV spectra“. In: *Biotechnology and Bioengineering* October.
- Rolinger, Laura, Matthias Rüdts, and Jürgen Hubbuch (11/2021). „Comparison of UV- and Raman-based monitoring of the Protein A load phase and evaluation of data fusion by PLS models and CNNs“. In: *Biotechnology and Bioengineering* 118.11 (11), pp. 4255–4268.
- Romann, Patrick et al. (11/2022). „Advancing Raman model calibration for perfusion bioprocesses using spiked harvest libraries“. In: *Biotechnology Journal* 17.11, e2200184.
- Rosebrock, A. (2018). *Deep Learning for Computer Vision with Python*. pyimagesearch.
- Rosenberg, Amy S. (09/2006). „Effects of protein aggregates: An immunologic perspective“. en. In: *The AAPS Journal* 8.3, p. 59.
- Rüdts, Matthias, Nina Brestrich, et al. (02/2017). „Real-time monitoring and control of the load phase of a protein A capture step“. In: *Biotechnology and Bioengineering* 114.2 (2), pp. 368–373.
- Rüdts, Matthias, Till Briskot, and Jürgen Hubbuch (03/2017). „Advances in downstream processing of biologics – Spectroscopy: An emerging process analytical technology“. In: *Journal of Chromatography A* 1490, pp. 2–9.
- Rüdts, Matthias, Philipp Vormittag, et al. (06/2019). „Process monitoring of virus-like particle reassembly by diafiltration with UV/Vis spectroscopy and light scattering“. In: *Biotechnology and Bioengineering* 116.6, pp. 1366–1379.
- Rygula, A. et al. (2013). „Raman spectroscopy of proteins: A review“. In: *Journal of Raman Spectroscopy* 44.8, pp. 1061–1076.
- S. De Jong (1993). „SIMPLS: an alternative approach to partial least squares regression“. In: *Chemometrics and intelligent laboratory systems* 18, pp. 251–263.
- Sadiki, Amissi et al. (12/2020). „Site-Specific Conjugation of Native Antibody“. In: *Antibody Therapeutics* 3.4, pp. 271–284.
- Saggu, Miguel, Jun Liu, and Ankit Patel (09/2015). „Identification of Subvisible Particles in Biopharmaceutical Formulations Using Raman Spectroscopy Provides Insight into Polysorbate 20 Degradation Pathway“. In: *Pharmaceutical Research* 32.9, pp. 2877–2888.
- Saleh, David et al. (2020). „Straightforward Method for Calibration of Mechanistic Cation Exchange Chromatography Models for Industrial Applications“. In: *Biotechnology Progress* 800.
- Sanden, Adrian et al. (12/2019). „Fourier-transform infrared spectroscopy as a process analytical technology for near real time in-line estimation of the degree of PEGylation in chromatography“. In: *Journal of Chromatography A* 1608, p. 460410.

- Santos, Rafael M. et al. (2018). „Monitoring mAb cultivations with in-situ raman spectroscopy: The influence of spectral selectivity on calibration models and industrial use as reliable PAT tool“. In: *Biotechnology Progress* 34.3, pp. 659–670.
- Santos, Vianney O. et al. (08/2005). „A comparative study of diesel analysis by FTIR, FTNIR and FT-Raman spectroscopy using PLS and artificial neural network analysis“. In: *Analytica Chimica Acta* 547.2 (2), pp. 188–196.
- Sato, Yusui et al. (2023). „Raman Spectroscopic Analysis of Highly-Concentrated Antibodies under the Acid-Treated Conditions“. In: *Pharmaceutical Research* 40.7, pp. 1853–1864.
- Sauer, Dominik Georg et al. (08/2019). „Real-time monitoring and model-based prediction of purity and quantity during a chromatographic capture of fibroblast growth factor 2“. In: *Biotechnology and Bioengineering* 116.8 (8), pp. 1999–2009.
- Savitzky, Abraham. and M. J. E. Golay (07/1964). „Smoothing and Differentiation of Data by Simplified Least Squares Procedures“. In: *Analytical Chemistry* 36.8, pp. 1627–1639.
- Schiemer, Robin, Matthias Rüdert, and Jürgen Hubbuch (2024). „Generative data augmentation and automated optimization of convolutional neural networks for process monitoring“. In: *Frontiers in Bioengineering and Biotechnology* 12.
- Schiemer, Robin, Jan Tobias Weggen, et al. (05/2023). „An adaptive soft-sensor for advanced real-time monitoring of an antibody-drug conjugation reaction“. en. In: *Biotechnology and Bioengineering* 120.7, pp. 1914–1928.
- Schmid, Franz-Xaver (2001). „Biological Macromolecules: UV-visible Spectrophotometry“. In: Schneider, René and Christos Georgakis (2013). „How to NOT make the extended kalman filter fail“. In: *Industrial and Engineering Chemistry Research* 52.9, pp. 3354–3362.
- Schuhmacher, Alexander, Oliver Gassmann, and Markus Hinder (02/2016). „A Review of the Pharmaceutical R&D Efficiency: Costs, Timelines, and Probabilities“. In: *Value Creation in the Pharmaceutical Industry*, pp. 60–79.
- Selvaraju, Ramprasaath R. et al. (2020). „Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization“. In: *International Journal of Computer Vision* 128 (2), pp. 336–359.
- Shivu, Bhavana et al. (08/2013). „Distinct β -Sheet Structure in Protein Aggregates Determined by ATR-FTIR Spectroscopy“. In: *Biochemistry* 52.31, pp. 5176–5183.
- Shorten, Connor and Taghi M. Khoshgoftaar (2019). „A survey on Image Data Augmentation for Deep Learning“. In: *Journal of Big Data* 6.1.
- Shukla, Abhinav A. et al. (03/2007a). „Downstream Processing of Monoclonal Antibodies—Application of Platform Approaches“. In: *Polyclonal and Monoclonal Antibody Production, Purification, Process and Product Analytics* 848.1, pp. 28–39.
- Shukla, Abhinav A. et al. (2007b). „Downstream processing of monoclonal antibodies—Application of platform approaches“. In: *Journal of Chromatography B* 848.1, pp. 28–39.
- Simon, Dan (2006). „Nonlinear Kalman filtering“. In: *Optimal State Estimation*. John Wiley & Sons, Ltd. Chap. 13, pp. 393–431.
- Simutis, Rimvydas and Andreas Lübbert (03/2017). „Hybrid approach to state estimation for bioprocess control“. In: *Bioengineering* 4 (1).
- Sinfield, Joseph V. and Chike K. Monwuba (2014). „Assessment and correction of turbidity effects on Raman observations of chemicals in aqueous solutions“. In: *Applied Spectroscopy* 68.12, pp. 1381–1392.
- Sinner, Peter et al. (2021). „Usage of Digital Twins Along a Typical Process Development Cycle“. In: *Digital Twins: Tools and Concepts for Smart Biomanufacturing*. Ed. by Christoph Herwig, Ralf Pörtner, and Johannes Möller. Cham: Springer International Publishing, pp. 71–96.

- Smulko, Janusz M et al. (02/2014). „Anatomy of Noise in Quantitative Biological Raman Spectroscopy“. en. In: *Bioanalysis* 6.3, pp. 411–421.
- Soch, Joram and Carsten Allefeld (08/2018). „MACS – a new SPM toolbox for model assessment, comparison and selection“. In: *Journal of Neuroscience Methods* 306, pp. 19–31.
- Socrates, G. (2004). *Infrared and Raman Characteristic Group Frequencies: Tables and Charts*. Wiley.
- Spann, Robert et al. (2018). „A probabilistic model-based soft sensor to monitor lactic acid bacteria fermentations“. In: *Biochemical Engineering Journal* 135, pp. 49–60.
- Spinner, E. (05/2003). „Raman-spectral depolarisation ratios of ions in concentrated aqueous solution. The next-to-negligible effect of highly asymmetric ion surroundings on the symmetry properties of polarisability changes during vibrations of symmetric ions.“ In: *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 59.7, pp. 1441–1456.
- Srivastava, Nitish et al. (01/2014). „Dropout: a simple way to prevent neural networks from overfitting“. In: *J. Mach. Learn. Res.* 15.1, pp. 1929–1958.
- Stelzer, Ines V., Julian Kager, and Christoph Herwig (2017). „Comparison of Particle Filter and Extended Kalman Filter Algorithms for Monitoring of Bioprocesses“. In: *Proceedings of the 27th European Symposium on Computer Aided Process Engineering – ESCAPE 27*. Vol. 40. Elsevier Masson SAS, pp. 1483–1488.
- Štrumbelj, Erik and Igor Kononenko (12/2014a). „Explaining Prediction Models and Individual Predictions with Feature Contributions“. In: *Knowledge and Information Systems* 41.3, pp. 647–665.
- Štrumbelj, Erik and Igor Kononenko (12/2014b). „Explaining prediction models and individual predictions with feature contributions“. In: *Knowledge and Information Systems* 41.3 (3), pp. 647–665.
- Sun, Zhongyu, Jiajin Fan, et al. (2020). „Assessment of the human albumin in acid precipitation process using NIRS and multi-variable selection methods combined with SPA“. In: *Journal of Molecular Structure* 1199, p. 126942.
- Sun, Zhongyu, Jiayue Wang, et al. (2018). „Calibration transfer of near infrared spectrometers for the assessment of plasma ethanol precipitation process“. In: *Chemometrics and Intelligent Laboratory Systems* 181. August, pp. 64–71.
- Szegedy, Christian et al. (2015). „Going deeper with convolutions“. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9.
- team, The pandas development (02/2020). *pandas-dev/pandas: Pandas*. Version latest.
- Thakur, Garima, Vishwanath Hebhi, and Anurag S. Rathore (2020). „An NIR-based PAT approach for real-time control of loading in Protein A chromatography in continuous manufacturing of monoclonal antibodies“. In: *Biotechnology and Bioengineering* 117.3, pp. 673–686.
- Thakur, Garima, Vishwanath Hebhi, and Anurag S. Rathore (05/2021). „Near Infrared Spectroscopy as a PAT tool for monitoring and control of protein and excipient concentration in ultrafiltration of highly concentrated antibody formulations“. en. In: *International Journal of Pharmaceutics* 600. December 2020, p. 120456.
- Thakur, Garima, Swapnil Thori, and Anurag S. Rathore (11/2020). „Implementing PAT for single-pass tangential flow ultrafiltration for continuous manufacturing of monoclonal antibodies“. en. In: *Journal of Membrane Science* 613. July, p. 118492.
- Thissen, U et al. (2004). „Comparing support vector machines to PLS for spectral regression applications“. In: *Chemometrics and Intelligent Laboratory Systems* 73.2, pp. 169–179.

- Thomann, Marco et al. (2016). „Fc-galactosylation modulates antibody-dependent cellular cytotoxicity of therapeutic antibodies“. In: *Molecular Immunology* 73, pp. 69–75.
- Thomas, Anish, Beverly A. Teicher, and Raffit Hassan (2016). „Antibody–drug conjugates for cancer therapy“. In: *The Lancet Oncology* 17.6, e254–e262.
- Tiernan, Hannah, Bernadette Byrne, and Sergei G. Kazarian (11/2020). „ATR-FTIR Spectroscopy and Spectroscopic Imaging for the Analysis of Biopharmaceuticals“. In: *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 241, p. 118636.
- Top Pharma Products by Global Sales 2024* (2025). Statista. URL: <https://www.statista.com/statistics/258022/top-10-pharmaceutical-products-by-global-sales-2011/> (visited on 09/12/2025).
- Trampuž, Marko, Dušan Teslić, and Blaž Likozar (2020). „Process analytical technology-based (PAT) model simulations of a combined cooling, seeded and antisolvent crystallization of an active pharmaceutical ingredient (API)“. In: *Powder Technology* 366, pp. 873–890.
- Trygg, Johan and Svante Wold (2002). „Orthogonal projections to latent structures (O-PLS)“. In: *Journal of Chemometrics* 16.3, pp. 119–128.
- Tsoka, Sophia et al. (01/2000). „Selective Flocculation and Precipitation for the Improvement of Virus-Like Particle Recovery from Yeast Homogenate“. In: *Biotechnology Progress* 16.4, pp. 661–667.
- Tulsyan, Aditya et al. (2021). „Spectroscopic models for real-time monitoring of cell culture processes using spatiotemporal just-in-time Gaussian processes“. In: *AIChE Journal* 67.5 (5).
- Tuveri, Andrea et al. (2021). „Sensor fusion based on Extended and Unscented Kalman Filter for bioprocess monitoring“. In: *Journal of Process Control* 106, pp. 195–207.
- Tyther, Raymond and Nigel Jenkins (2011). „Quality Issues Arising from Post-translational Modification of Recombinant Antibodies“. In: *Antibody Expression and Production*. Ed. by Mohamed Al-Rubeai. Dordrecht: Springer Netherlands, pp. 293–303.
- Ündey, Cenk et al. (2010). „Applied advanced process analytics in biopharmaceutical manufacturing: Challenges and prospects in real-time monitoring and control“. In: *Journal of Process Control* 20.9, pp. 1009–1018.
- Valappil, Jaleel and Christos Georgakis (2000). „Systematic estimation of state noise statistics for extended Kalman filters“. In: *AIChE Journal* 46.2, pp. 292–308.
- Valensise, C. M. et al. (06/2020). „Removing non-resonant background from CARS spectra via deep learning“. In: *APL Photonics* 5.6, p. 061305.
- Valentic, Angela and Jürgen Hubbuch (2024). „Effective removal of host cell-derived nucleic acids bound to hepatitis B core antigen virus-like particles by heparin chromatography“. In: *Frontiers in Bioengineering and Biotechnology* Volume 12 - 2024.
- Valentic, Angela, Jakob Müller, and Jürgen Hubbuch (07/2022). „Effects of Different Lengths of a Nucleic Acid Binding Region and Bound Nucleic Acids on the Phase Behavior and Purification Process of HBcAg Virus-Like Particles“. In: *Frontiers in Bioengineering and Biotechnology* 10.
- Van Brink, Mark Den, Michel Pepers, and Alex M. Van Herk (2002). „Raman spectroscopy of polymer latexes“. In: *Journal of Raman Spectroscopy* 33.4, pp. 264–272.
- Vanderlaan, Martin et al. (2018). „Experience with host cell protein impurities in biopharmaceuticals“. In: *Biotechnology Progress* 34.4, pp. 828–837.
- Vázquez-Rey, María and Dietmar A. Lang (07/2011). „Aggregates in monoclonal antibody manufacturing processes“. en. In: *Biotechnology and Bioengineering* 108.7, pp. 1494–1508.

- Venton, J. et al. (12/2021). „Robustness of convolutional neural networks to physiological electrocardiogram noise“. en. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 379.2212, p. 20200262.
- Walsh, Gary and Eithne Walsh (12/2022). „Biopharmaceutical benchmarks 2022“. en. In: *Nature Biotechnology* 40.12, pp. 1722–1760.
- Wang, Bryan et al. (2022). „Process analytical technologies in cell therapy manufacturing: State-of-the-art and future directions“. In: *Journal of Advanced Manufacturing and Processing* 4.1, pp. 1–17.
- Wang, Jiarui et al. (2023a). „Automated calibration and in-line measurement of product quality during therapeutic monoclonal antibody purification using Raman spectroscopy“. In: *Biotechnology and Bioengineering*.
- Wang, Jiarui et al. (12/2023b). „In-line product quality monitoring during biopharmaceutical manufacturing using computational Raman spectroscopy“. In: *mAbs* 15.1, p. 2220149.
- Watson, G. H., W. B. Daniels, and C. S. Wang (02/1981). „Measurements of Raman intensities and pressure dependence of phonon frequencies in sapphire“. In: *Journal of Applied Physics* 52.2, pp. 956–958.
- Weber, Alfons et al. (01/1967). „High-Resolution Raman Spectroscopy of Gases with cw-Laser Excitation*“. In: *JOSA* 57.1, pp. 19–28.
- Weber, Dennis and Jürgen Hubbuch (2021). „Raman spectroscopy as a process analytical technology to investigate biopharmaceutical freeze concentration processes“. en. In: *Biotechnology and Bioengineering* 118.12, pp. 4708–4719.
- Wegner, Christina H., Ines Zimmermann, and Jürgen Hubbuch (02/02/2022). „Rapid Analysis for Multicomponent High-Throughput Crystallization Screening: Combination of UV–Vis Spectroscopy and Chemometrics“. In: *Crystal Growth & Design* 22.2, pp. 1054–1065.
- Wei, Bingchuan et al. (12/2022). „Multi-attribute Raman spectroscopy (MARS) for monitoring product quality attributes in formulated monoclonal antibody therapeutics“. In: *mAbs* 14.1 (1).
- Wen, Zai-Qing (11/2007). „Raman spectroscopy of protein pharmaceuticals“. In: *Journal of Pharmaceutical Sciences* 96.11, pp. 2861–2878.
- West, Jay M. et al. (2021). „Process analytical technology for on-line monitoring of quality attributes during single-use ultrafiltration/diafiltration“. In: *Biotechnology and Bioengineering* 118.6, pp. 2293–2300.
- Wiercigroch, Ewelina et al. (10/2017). „Raman and Infrared Spectroscopy of Carbohydrates: A Review“. In: *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 185, pp. 317–335.
- Wilcox, R. R (2005). *Introduction to Robust Estimation and Hypothesis Testing*. Netherlands: Elsevier Science, p. 100.
- Williams, Robert W. (1986). „[14] Protein secondary structure analysis using Raman amide I and amide III spectra“. en. In: *Methods in Enzymology*. Vol. 130. Elsevier, pp. 311–331.
- Witjes, H. et al. (2000). „Automatic correction of peak shifts in Raman spectra before PLS regression“. In: *Chemometrics and Intelligent Laboratory Systems* 52.1, pp. 105–116.
- Wold, Svante, Henrik Antti, et al. (1998). „Orthogonal signal correction of near-infrared spectra“. In: *Chemometrics and Intelligent Laboratory Systems* 44.1-2, pp. 175–185.
- Wold, Svante and Michael Sjöström (06/01/1977). „SIMCA: A Method for Analyzing Chemical Data in Terms of Similarity and Analogy“. In: *Chemometrics: Theory and Application*. Vol. 52. ACS Symposium Series 52. AMERICAN CHEMICAL SOCIETY, pp. 243–282.

- Wold, Svante, Michael Sjöström, and Lennart Eriksson (10/2001). „PLS-regression: a basic tool of chemometrics“. en. In: *Chemometrics and Intelligent Laboratory Systems* 58.2 (2), pp. 109–130.
- Wu, Man et al. (12/2021). „Deep learning data augmentation for Raman spectroscopy cancer tissue classification“. en. In: *Scientific Reports* 11.1, p. 23842.
- Xiaobo, Zou et al. (05/2010). „Variables selection methods in near-infrared spectroscopy“. en. In: *Analytica Chimica Acta* 667.1-2, pp. 14–32.
- Yosinski, Jason et al. (06/2015). „Understanding Neural Networks Through Deep Visualization“. In: Yuanyuan, Chen and Wang Zhibin (10/2018). „Quantitative analysis modeling of infrared spectroscopy based on ensemble convolutional neural networks“. In: *Chemometrics and Intelligent Laboratory Systems* 181, pp. 1–10.
- Zahin, Maryam et al. (2016). „Scalable production of HPV16 L1 protein and VLPs from tobacco leaves“. In: *PLoS ONE* 11.8, pp. 1–16.
- Zavala-Ortiz, Daniel Arturo et al. (2020). „Support Vector and Locally Weighted regressions to monitor monoclonal antibody glycosylation during CHO cell culture processes, an enhanced alternative to Partial Least Squares regression“. In: *Biochemical Engineering Journal* 154. August 2019, p. 107457.
- Zeiler, Matthew D. and Rob Fergus (11/2013). „Visualizing and Understanding Convolutional Networks“. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 8689 LNCS (PART 1), pp. 818–833.
- Zelger, Monika et al. (2016). „Real-time monitoring of protein precipitation in a tubular reactor for continuous bioprocessing“. In: *Process Biochemistry* 51.10, pp. 1610–1621.
- Zeltins, Andris (01/2013). „Construction and Characterization of Virus-Like Particles: A Review“. In: *Molecular Biotechnology* 53.1, pp. 92–107.
- Zhang, Chi et al. (2019). „Rapid, quantitative determination of aggregation and particle formation for antibody drug conjugate therapeutics with label-free Raman spectroscopy“. In: *Analytica Chimica Acta* 1081, pp. 138–145.
- Zhang, Feng et al. (03/2020). „Baseline correction for infrared spectra using adaptive smoothness parameter penalized least squares method“. In: *Spectroscopy Letters* 53.3, pp. 222–233.
- Zhang, Pengfei et al. (02/2021). „A novel variable selection method based on combined moving window and intelligent optimization algorithm for variable selection in chemical modeling“. en. In: *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 246, p. 118986.
- Zhou, Chen et al. (03/2015). „Concomitant Raman spectroscopy and dynamic light scattering for characterization of therapeutic proteins at high concentrations“. en. In: *Analytical Biochemistry* 472, pp. 7–20.
- Zhu, Yiran et al. (07/23/2022). „RNA-based Therapeutics: An Overview and Prospectus“. In: *Cell Death & Disease* 13.7, p. 644.
- Zlotnick, A. et al. (01/1996). „Dimorphism of Hepatitis B Virus Capsids Is Strongly Influenced by the C-Terminus of the Capsid Protein“. In: *Biochemistry* 35.23, pp. 7412–7421.

List of Figures

1.1	Typical downstream processes for biopharmaceutical products: mAbs, VLPs and ADCs with possible PAT integration sites	4
1.2	Schematic illustration of different spectroscopic effects: Jablonski diagram, electromagnetic spectrum and carbodioxide vibrations.	8
1.3	Illustration of the effect of common preprocessing operations using an exemplary Raman spectrum of protein at a concentration of 25 g/L.	16
1.4	Schematic illustration of a GPR.	24
1.5	Schematic illustration of one-dimensional feed-forward (A) and convolutional neural networks (B).	26
1.6	Comparison of grid search and Bayesian optimization for HPO.	30
1.7	Visualization of a generalized ML workflow.	31
1.8	Visualization of different cross-validation schemes.	32
3.1	Schematic overview of the purposes of the data sets and the methodology of the designed model-based soft-sensor.	55
3.2	Predictions for the GPR-VS model.	59
3.3	EKF prediction error regarding the $nRMSE$ during k_Q tuning	61
3.4	Model predictions and calculated process and sensor noise covariance for run 3.	62
3.5	Comparison of the $RMSEP$ values of the GPR, the kinetic model and the EKF obtained from 100 simulations for each scenario	63
3.6	Model predictions for fed-batch run 11 for the concentration of conjugated drug and the ADC reaction species.	64
4.1	Illustration of the study design.	75
4.2	Visualization of the LSA data augmentation method.	77
4.3	Effect of parameter tuning of the LSA data augmentation method on the performance for data set 1 and data set 2.	85
4.4	Local estimations of the pure component profiles for ribonuclease A (Rib A), cytochrome C (Cyt C), lysozyme (Lys) for data set 1 and monomer and aggregate species for data set 2.	86
4.5	Influence of selected hyperparameters on the predictive cross-validation performance (RMSECV) for each component in data set 1.	87

4.6	Hyperparameter evolution profiles during optimization for data set 2.	89
4.7	Model performance for initial and optimized CNN models for data set 1 and data set 2.	90
4.8	Model-specific and model-agnostic importance measures for data set 2.	91
4.9	Evaluation of CNN robustness and transferability.	93
5.1	Illustration of pretreatment of the ultraviolet (UV) absorbance measurements.	115
5.2	Comparison of pretreated UV absorbance data and VLP precipitation curves.	115
5.3	Illustration of the preprocessing of the Raman spectra.	117
5.4	Comparison of collected Raman spectra in precipitant-containing buffer, supernatant and precipitate samples.	118
5.5	Raman spectra of the spiking solutions and comparison of selected wavenumber ratios.	119
5.6	Comparison of the effects of preprocessing operations on Raman spectra in precipitate samples for the wavenumber region $1020\text{-}1800\text{ cm}^{-1}$	121
5.7	Comparison of preprocessing pipelines and model types.	123
5.8	PLS model predictions with regard to precipitated VLP for batch and fed-batch experiments and the corresponding VIP scores for both models.	124
6.1	Experimental setup connecting ÄKTA purifier, A905 autosampler and Tornado Hyperflux PRO Plus.	140
6.2	Sample distribution, correlation coefficients, and preprocessed Raman spectra for LHS and central composite design (CCD) datasets in the upper and lower row, respectively.	141
6.3	Process flow diagram of the cross-flow filtration (CFF) system with an external loop for on-line Raman measurements.	143
6.4	Injection sequence of selected samples with increasing lysozyme concentration depicted by the normalized Raman intensity using the 1004 cm^{-1} band after baseline correction.	144
6.5	Regression errors for calibration datasets LHS and CCD.	146
6.6	Model predictions for in-line data from UF/DF experiments.	148
7.1	Schematic representation of the forced degradation and mixing study workflows.	155
7.2	Spectral features observed in offline-recorded Raman spectra of BSA obtained during mixing studies.	158
7.3	Results of the forced degradation experiment. Off-line recorded Raman spectra and SNR for each wavenumber.	161
7.4	Elution profiles and structural markers of BSA under varying process conditions.	166
7.5	Predictions of Crop-SG-PLS1 and CNN regression models for training, cross-validation and test.	168
A.1	Absorbance spectra and detected shift of the location of the peak maximum for all runs used in this study.	222
A.2	Raw absorbance spectra for the surrogate drug without addition of mAb solution.	223
A.3	Selected variables for GPR model in combination with variable selection algorithm.	223
A.4	Model and EKF predictions for all training runs.	224
A.5	EKF predictions and uncertainty estimates for all test runs derived from the state covariance matrix.	225
A.6	Model and EKF predictions for all test runs.	225
B.1	Hyperparameter evolution profiles during optimization for data set 1.	231

B.2	Visualization of the architecture of the optimized CNN models for data set 1 and data set 2.	232
B.3	Predictions of the optimized CNN models for data sets 1 and 2.	233
B.4	Exemplary raw UV/Vis spectra for data set 1 and data set 2.	234
B.5	Layer visualization for all convolutional layer for the optimized CNN model for data set 2.	235
B.6	Model-specific and model-agnostic importance measures for data set 1.	236
B.7	Layer visualization for all convolutional layer for the optimized CNN model for data set 1.	237
C.1	Comparison of the effects of preprocessing operations on Raman spectra in supernatant samples for the wavenumber region 1020-1800 cm^{-1}	241
C.2	Comparison of model types and preprocessing pipelines without turbidity correction. . .	242
C.3	PLS model predictions with regard to precipitated VLP concentration for batch experiments.	242
C.4	PLS model predictions with regard to the precipitated VLP concentration for batch experiments over ammonium sulfate (AMS) concentrations.	243
C.5	PLS model predictions with regard to precipitated VLP concentration for fed-batch experiments.	244
C.6	PLS model predictions with regard to AMS concentration for batch experiments. . . .	244
C.7	PLS model predictions with regard to AMS concentration for fed-batch experiments. . .	245
C.8	Visualization of preprocessing operations for fed-batch experiments F1-F3 from left to right.	246
D.1	Comparison of off-line and in-line spectra after different stages of preprocessing. . . .	248
E.1	Distribution of sampled values for augmentation test.	250
E.2	Performance comparison for all studied model variants for training, cross-validation, augmentation test and external test data sets.	251
E.3	Performance comparison of representative PLS and CNN models for rotation test. . . .	252
E.4	Parity plots for % HMWC prediction using the PLS1 model across training, cross-validation, and test datasets.	252
E.5	Parity plots for % HMWC prediction using the PLS2 model across training, cross-validation, and test datasets.	253
E.6	Parity plots for % HMWC prediction using the SG-PLS1 model across training, cross-validation, and test datasets.	253
E.7	Parity plots for % HMWC prediction using the SG-PLS2 model across training, cross-validation, and test datasets.	254
E.8	Parity plots for % HMWC prediction using the Crop-SG-PLS1 model across training, cross-validation, and test datasets.	254
E.9	Parity plots for % HMWC prediction using the Crop-SG-PLS2 model across training, cross-validation, and test datasets.	255
E.10	Parity plots for % HMWC prediction using the PLS1-Augmented model across training, cross-validation, and test datasets.	255
E.11	Parity plots for % HMWC prediction using the PLS2-Augmented model across training, cross-validation, and test datasets.	256
E.12	Parity plots for % HMWC prediction using the CNN model across training, cross-validation, and test datasets.	256

E.13 Parity plots for % HMWC prediction using the CNN-Augmented model across training, cross-validation, and test datasets.	257
--	-----

List of Tables

3.1	Overview of conducted conjugation runs using different mAb concentrations and NPM excesses.	52
3.2	Overview of predictive performance of selected chemometric models.	58
3.3	Comparison of the three studied combinations of Q and R regarding their accomplished $RMSE$ reduction compared to the kinetic model in training and test set.	60
4.1	Overview of hyperparameters used for data augmentation,CNN architecture and training for data sets 1 and 2.	80
5.1	Overview of the eight batch precipitation experiments B1-B8 comprising parameters for initial lysate conditioning and the final, curve fit-derived, apparent concentration of precipitated VLPs.	110
5.2	Overview of the three fed-batch precipitation experiments F1-F3 comprising parameters for initial lysate conditioning and fed-batch processing. Lysate batch 3 was used for all experiments.	110
6.1	Experimental conditions for different runs.	143
7.1	Experimental conditions for bind-elute AEX chromatography experiment used in in-line Raman monitoring of BSA monomer and aggregate separation.	156
B.1	Overview of experimental conditions for data set 1 and 2.	229
B.2	Optimized PLS hyperparameters for data sets 1-4.	229
B.3	Comparison of LSA and extended multiplicative scatter augmentation (EMSA) method for data set 2 and 3 as measured by the RMSE for all components.	230

Acronyms

E. coli

Escherichia coli

ADC

antibody-drug conjugate

AEX

anion-exchange chromatography

AMS

ammonium sulfate

ANN

artificial neural network

ARMP

automated Raman measurement platform

asPLS

adaptive smoothness penalized least squares

ATR

attenuated total reflectance

BSA

bovine serum albumin

CCD

central composite design

CFF

cross-flow filtration

CLS

classical least squares

CNN

convolutional neural network

CPP

critical process parameter

CQA

critical quality attribute

CV

column volume

Cyt C

cytochrome C

DAR

drug-to-antibody ratio

DERPSALSA

derivative peak-screening asymmetric least squares algorithm

DF

diafiltration

DSP

downstream processing

DV

diafiltration volume

EKF

extended Kalman filter

ELISA

enzyme-linked immunosorbent assay

ELU

exponential linear unit

EMA

European Medicines Agency

EMSA

extended multiplicative scatter augmentation

EMSC

extended multiplicative signal correction

FC

fully connected

FDA

United States Food and Drug Administration

FFNN

feed-forward neural network

GAN

generative adversarial network

GMP

Good manufacturing practice

GP

gaussian process

GPR

Gaussian process regression

GradCAM

gradient-weighted class activation map

HBcAg

Hepatitis B core Antigen

HC

host cell

HCCF

harvest cell culture material

HCP

host cell protein

HIC

hydrophobic interaction chromatography

HMW

high molecular weight species

HMWS

high molecular weight species

HPLC

high-performance liquid chromatography

HPO

hyperparameter optimization

IgG

immunoglobulin G

IHM

indirect hard modeling

iPLS

iterative partial-least squares

IR

infrared

KDE

kernel-density estimation

kNN

k-nearest neighbor

LHS

latin hypercube sampling

LMW

low molecular weight species

LMWS

low molecular weight species

LOBO

leave-one-batch-out

LSA

local subset augmentation

Lys

lysozyme

mAb

monoclonal antibody

MAE

mean absolute error

MAPE

mean absolute percentage error

MES

4-morpholinoethanesulfonic acid

ML

machine learning

MLR

multiple linear regression

MSE

mean squared error

MWCO

molecular weight cut-off

NIPALS

nonlinear iterative partial least squares

NIR

near-infrared

NNLS

non-negative least squares

OPC-UA

open platform communication - unified architecture

OPLS

orthogonal projection to latent structures

PAT

process analytical technology

PCA

principal component analysis

PCR

principal component regression

PEG

polyethylene glycol

Phe

phenylalanine

PLS

partial least squares

QbD

quality by design

RBF

radial basis function

ReLU

rectified linear unit

Rib A

ribonuclease A

RMSE

root mean squared error

SAGE

Shapley additive global importance

SEC

size exclusion chromatography

SG

Savitzky–Golay

SGF

Savitzky-Golay filter

SHAP

Shapley additive explanations

SIMPLS

straightforward implementation of a statistically inspired modification of partial least squares

SNR

signal-to-noise ratio

SS

scaled subtraction

SVM

support vector machine

tanh

hyperbolic tangent

TMP

transmembrane pressure

TPE

tree-structured parzen estimator

Tris

trishydroxymethylaminomethane

Trp

tryptophan

Tyr

tyrosine

UF

ultrafiltration

UF/DF

ultra-/diafiltration

UHP-SEC

ultrahigh performance - size exclusion chromatography

UHPLC

ultrahigh-performance liquid chromatography

USP

upstream processing

UV

ultraviolet

UV/Vis

ultraviolet/visible

VAE

variational autoencoder

VIP

variable importance in projection

VLP

virus-like particle



Appendix Chapter 1

Supplementary Material for Chapter 3

Gaussian process regression

Mathematical formulation

In Gaussian process regression for spectroscopic data, a functional relationship between an input spectrum $\mathbf{x}$ with M wavelengths with an output variable y is established. The functional relationship is created by a Gaussian process (GP) which can be defined as a collection of random variables such that any finite subset exhibits a joint Gaussian distribution (Rasmussen and C. K. I. Williams 2006; Cui and Fearn 2017). A GP is mathematically defined by its mean function m and covariance function C and can be written as

$$P(y|\mathbf{X}) = N(m(\mathbf{x}), C(\mathbf{x}, \mathbf{x})) \quad (\text{A.1})$$

assuming a Gaussian posterior $P(y|\mathbf{X})$ with $\mathbf{X}$ being the training data consisting of N samples of input spectra. Considering mean-centered input data, the mean function of the GP can be set to $m(\mathbf{x}) = 0$, leaving the covariance function C as the only configurable of the GP. The covariance

function $C(\mathbf{x}, \mathbf{x}^*)$ computes a distance measure of two arbitrary input spectra $\mathbf{x}$ and $\mathbf{x}^*$. The choice of C , hence, affects the computed relationships between different input spectra. In this study, a composite covariance function was used which is constructed from the sum of the so-called linear and radial-basis function (RBF) kernels, such that

$$C(\mathbf{x}, \mathbf{x}^*) = a_0 + \sigma_0^2 \sum_{i=1}^N \mathbf{x}^T \mathbf{x}^* + \sigma_1^2 \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}^*\|_2^2}{2l}\right) + \sigma_e^2 \quad (\text{A.2})$$

with a_0 , σ_0^2 , σ_1^2 and l being the constant offset, the variance of the linear kernel, the variance of the RBF kernel and the length-scale, respectively. The GP covariance function also assumes Gaussian additive noise with variance σ_e^2 . All parameters are considered hyperparameters and need to be inferred from the training data.

For parameter inference, the log-likelihood function is maximized with regard to the hyperparameters. According to (Rasmussen and C. K. I. Williams 2006), the log-likelihood can be calculated as

$$L = -\frac{1}{2}(\log \det \Sigma - \mathbf{y}^T \Sigma^{-1} \mathbf{y} - N \log 2\pi) \quad (\text{A.3})$$

with Σ being the covariance matrix as computed by Eq. A.2 using all spectra available in the training data and $\mathbf{y}$ the corresponding output values.

Using the estimated set of hyperparameters, the predictive posterior distribution is given by the mean y^* and variance σ^{*2} as

$$y^* = \mathbf{k}_{\text{GP}}^T(\mathbf{x}^*) \Sigma^{-1} \mathbf{y} \quad (\text{A.4})$$

$$\sigma^{*2} = C(\mathbf{x}, \mathbf{x}^*) - \mathbf{k}_{\text{GP}}^T(\mathbf{x}^*) \Sigma^{-1} \mathbf{k}_{\text{GP}}(\mathbf{x}^*), \quad (\text{A.5})$$

where $\mathbf{k}_{\text{GP}} = [C(\mathbf{x}_1, \mathbf{x}^*), \dots, C(\mathbf{x}_N, \mathbf{x}^*)]^T$.

Hyperparameter optimization

The kernel parameters were estimated by expectation maximization as implemented in (Pedregosa et al. 2011). The resulting hyperparameter values for the GPR-VS model were $a_0 = 1\text{e-}5$, $\sigma_0 = 143$, $\sigma_1 = 0.105$, $\sigma_e = 0.0133$ and $l = 0.0536$.

Kinetic modeling

The herein applied kinetic model assumes that the reactive cysteines of the oxidized mAb conjugate irreversibly with a drug in a two-stage consecutive reaction. The two reaction steps for the first and second conjugation are described by second-order rate laws having two kinetic rates k_1 and k_2 . The inactivation of free drug is described by a first-order kinetic with the rate constant k_3 . In theory, a free drug molecule is expected to conjugate with each of the two reactive cysteine residues which were inserted in a single mAb. In practice, a residual level of un- and mono-conjugated mAb are present after the conjugation reaction. To account for this, the model assumes an initial distribution of reactive cysteine residues for the oxidized mAb intermediate. This distribution is experimentally determined based on the final composition of conjugated species in runs with a NPM excess of 3x or

higher. The kinetic model for the consecutive conjugation reaction of a mixture of mAbs with zero, one or two binding sites with the surrogate payload NPM is described by a set of seven ordinary differential equations (ODEs):

$$\frac{dc_{\text{mAb}_{2c}}}{dt} = -k_1 \cdot c_{\text{mAb}_{2c}} \cdot c_{\text{NPM}} \quad (\text{A.6a})$$

$$\frac{dc_{\text{mAb}_{1c}}}{dt} = -k_1 \cdot c_{\text{mAb}_{1c}} \cdot c_{\text{NPM}} \quad (\text{A.6b})$$

$$\frac{dc_{\text{mAb}_{0c}}}{dt} = 0 \quad (\text{A.6c})$$

$$\frac{dc_{\text{Conj1}_{1c}}}{dt} = k_1 \cdot c_{\text{mAb}_{2c}} \cdot c_{\text{NPM}} - k_2 \cdot c_{\text{Conj1}_{1c}} \cdot c_{\text{NPM}} \quad (\text{A.6d})$$

$$\frac{dc_{\text{Conj1}_{0c}}}{dt} = k_1 \cdot c_{\text{mAb}_{1c}} \cdot c_{\text{NPM}} \quad (\text{A.6e})$$

$$\frac{dc_{\text{Conj2}}}{dt} = k_2 \cdot c_{\text{mAb}_{1c}} \cdot c_{\text{NPM}} \quad (\text{A.6f})$$

$$\frac{dc_{\text{NPM}}}{dt} = -k_1 \cdot c_{\text{mAb}_{2c}} \cdot c_{\text{NPM}} - k_1 \cdot c_{\text{mAb}_{1c}} \cdot c_{\text{NPM}} - k_2 \cdot c_{\text{Conj1}_{1c}} \cdot c_{\text{NPM}} - k_3 \cdot c_{\text{NPM}} \quad (\text{A.6g})$$

where $c_{\text{mAb}_{2c}}$, $c_{\text{mAb}_{1c}}$ and $c_{\text{mAb}_{0c}}$ denote the concentration of unconjugated mAbs with two, one and zero available binding sites, respectively. The free NPM concentration is given as c_{NPM} . Further, the concentrations $c_{\text{Conj1}_{1c}}$, $c_{\text{Conj1}_{0c}}$, c_{Conj2} define the concentration of Conj1 with one attached NPM molecules and one binding sites, concentration of Conj1 with one attached NPM molecule and no binding site, and the concentration of Conj2 with two attached NPM molecules, respectively.

Extended Kalman Filter

The working principle of the EKF can be described by a recursive execution of a prediction and correction step. During the prediction step, the state function is solved for a discrete time interval $\Delta t = t_k - t_{k-1}$ to propagate the state vector $\mathbf{x}$ to the next time step according to Eq. A.7. To describe the covariance in each state, the EKF uses the state covariance matrix $\mathbf{P}$ which is propagated to the next time step in parallel according to Eq. A.8

$$\mathbf{x}^-(t_k) = \mathbf{x}^+(t_{k-1}) + \int_{t_{k-1}}^{t_k} (f(\mathbf{x}, u(\tau), \boldsymbol{\theta}) d\tau \quad (\text{A.7})$$

$$\mathbf{P}^-(t_k) = \mathbf{P}^+(t_{k-1}) + \int_{t_{k-1}}^{t_k} (\mathbf{Z}(\tau)\mathbf{P}(\tau) + \mathbf{P}(\tau)\mathbf{Z}^T(\tau) + \mathbf{Q}(\tau)) d\tau \quad (\text{A.8})$$

where the superscripts of the state vector and state covariance matrix (+) and (-) indicate the corrected values at the previous time instance and the predicted values at the current time instance, respectively. The process noise covariance matrix is denoted as $\mathbf{Q}$. To estimate the change in the state covariance, the local derivatives $\mathbf{Z}(t)$ of the state function with regard to all model states are introduced as

$$\mathbf{Z}(t_k) = \left(\frac{\partial f}{\partial \mathbf{x}} \right)_{\mathbf{x}(t_k), \boldsymbol{\theta}}. \quad (\text{A.9})$$

For the first iteration, the reaction states and state covariance matrix require initialization. In this study, all reaction states and the corresponding covariance matrix were initialized as suggested in (Schneider and Georgakis 2013).

During the correction step, the predicted estimate is fused with the available measurement $y(t_k)$ using the Kalman Gain $\mathbf{K}(t_k)$. The Kalman Gain takes into account the state covariance $\mathbf{P}^-(t_k)$, the measurement noise $\mathbf{R}(t_k)$ and the local derivatives $\mathbf{C}(t)$ of the measurement function with regard to all model states and is computed according to

$$\mathbf{K}(t_k) = \mathbf{P}^-(t_k) \mathbf{C}(t_k)^T (\mathbf{C}(t_k) \mathbf{P}^-(t_k) \mathbf{C}(t_k)^T + \mathbf{R}(t_k))^{-1} \quad (\text{A.10})$$

where $\mathbf{C}(t_k)$ is defined as

$$\mathbf{C}(t_k) = \left(\frac{\partial h}{\partial \mathbf{x}} \right)_{\mathbf{x}(t_k), \boldsymbol{\theta}}. \quad (\text{A.11})$$

The corrected states vector and the corrected state covariance are then computed according to

$$\mathbf{x}^+(t_k) = \mathbf{x}^-(t_k) + \mathbf{K}(t_k) (y(t_k) - h(\mathbf{x}^-(t_k))) \quad (\text{A.12})$$

$$\mathbf{P}^+(t_k) = (\mathbf{I} - \mathbf{K}(t_k) \mathbf{C}(t_k)) \mathbf{P}^-(t_k). \quad (\text{A.13})$$

Error metrics

The root mean square error (*RMSE*) is defined as

$$RMSE = \sqrt{\frac{\sum_{n=1}^N (y_i - \hat{y}_i)^2}{n}}, \quad (\text{A.14})$$

where $\hat{y}_i$ denotes the predicted concentration values, y_i the reference data from RP-UPHLC and n the sample size. To evaluate model equality, the *RMSECV* and the *RMSEP* are calculated based on the Eq. A.14. To compute the *RMSECV*, the *RMSE* for the validation data of each iteration are averaged. For the *RMSEP*, Eq. A.14 is applied using only the test data. To optimize PLSR model using cross-validation, a scaled sum of squared errors SSE_{scaled} was introduced according to (S. Wold, M. Sjöström, and Eriksson 2001) which is given as

$$SSE_{\text{scaled}} = \frac{\sum_{n=1}^N (y_i - \hat{y}_i)^2}{n - n_{\text{PLS}} - 1} \quad (\text{A.15})$$

with n_{PLS} being the number of latent variables used in the PLSR model.

Chemometric model development

Spectroscopic data

This section compiles spectroscopic data used for the closer evaluation of the presented results in this study. Fig. A.1 presents an overview of all absorbance spectra used in this study. From the raw absorbance spectra, the location of the peak maxima of the two main peak bands of NPM were extracted and are displayed over time side-by-side with the raw spectra. Fig. A.2 displays the timely absorbance spectra for the surrogate drug without adding any mAb to the solution.

Variable Selection

The selected variables of the GPR-VS model which were optimized based on the cross-validation error are displayed in Fig. A.3. The selected variables are indicated by gray vertical lines which accumulate around the absorbance peaks between 250-280 nm and 320-350 nm. This is also the spectral area where the two peak shifts of the NPM molecule occur during the conjugation reaction as shown described above. The same bathochromic red shift of the NPM spectrum has previously been reported by (Andris, Rüdte, et al. 2018).

Soft-sensor development

This section compiles the individual soft-sensor predictions for all conjugation runs used in this study. Fig. A.4 and Fig. A.6 show the model predictions of the spectroscopic sensor model, the kinetic model and the EKF soft-sensor for the conjugated drug concentration and all individual species in the test and training subset.

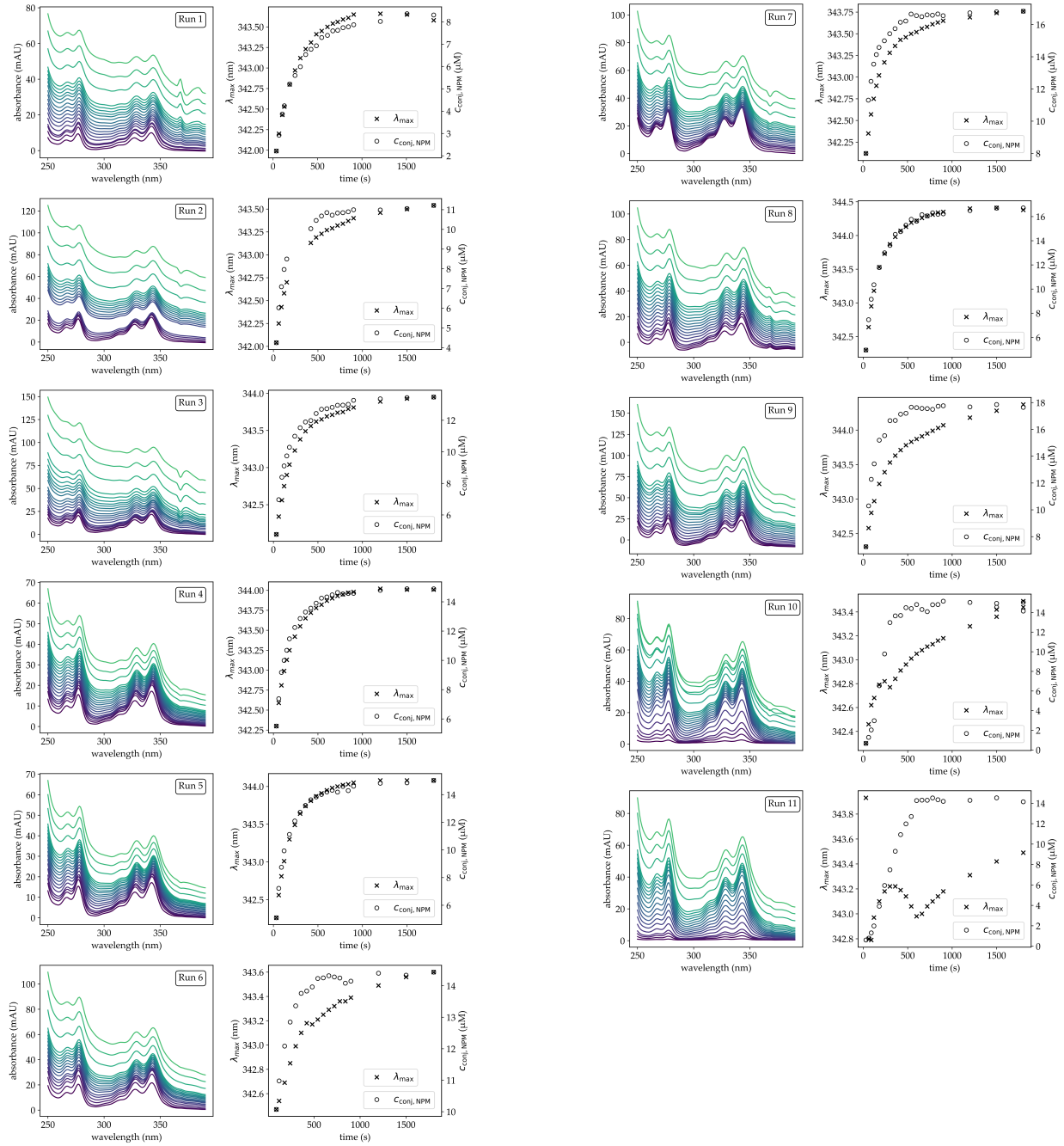


Figure A.1 Absorbance spectra and detected shift of the location of the peak maximum for all runs used in this study.

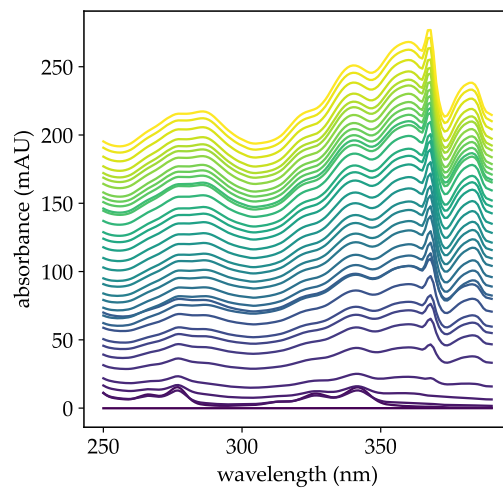


Figure A.2 Raw absorbance spectra for the surrogate drug without addition of mAb solution. The surrogate drug was continuously stirred and circulated through the on-line loop for 30 min. Only spectra for intervals of 25 s are shown.

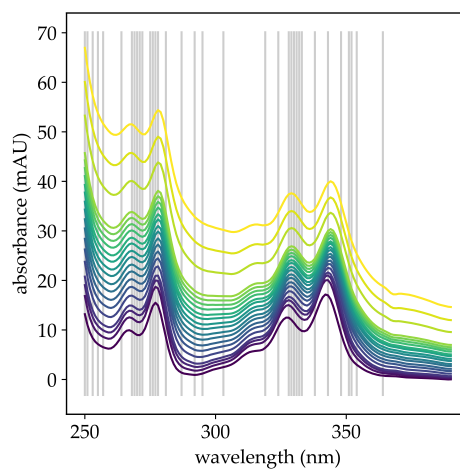


Figure A.3 Selected variables for GPR model in combination with variable selection algorithm. Selected variables are indicated by vertical gray lines. The absorbance spectra of run 5 are shown exemplarily and are colored by the reaction time from purple (0 min) to yellow (30 min).

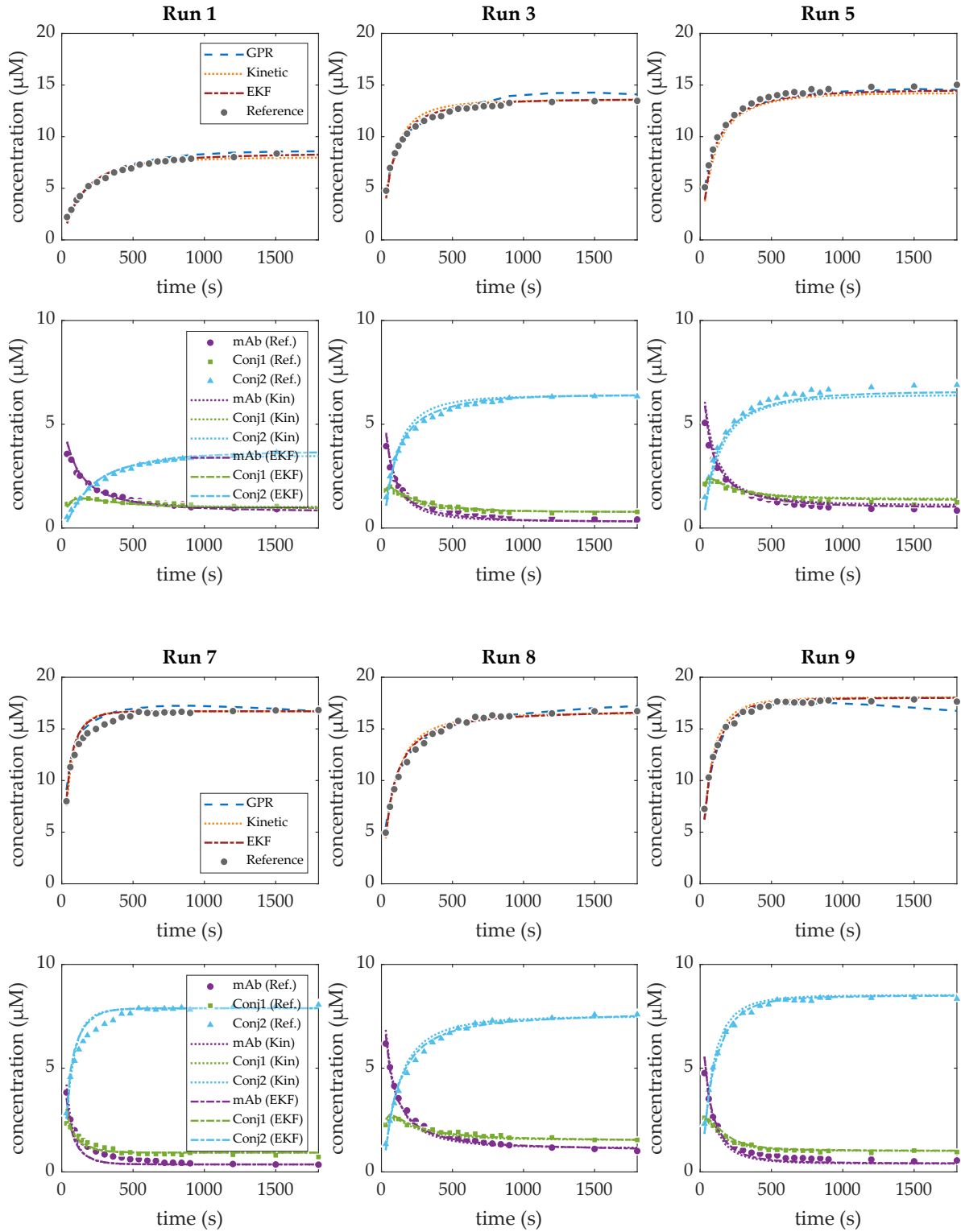


Figure A.4 Model and EKF predictions for all training runs. For each run, the first row shows the predictions of conjugated drug concentration and the second row the ADC species concentration.

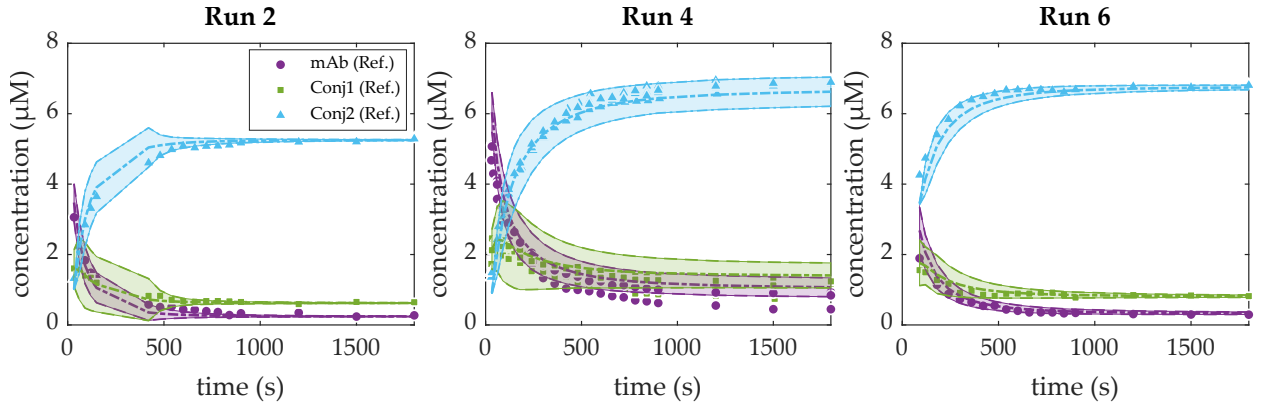


Figure A.5 EKF predictions and uncertainty estimates for all test runs derived from the state covariance matrix. The uncertainty bands were estimated from $\pm 2\sigma_{x(t)}$ for the corresponding states in the EKF state space. For Run 4, additional reference data from replicate runs, which were not used otherwise within this study, were added to compare the experimental and soft-sensor variance.

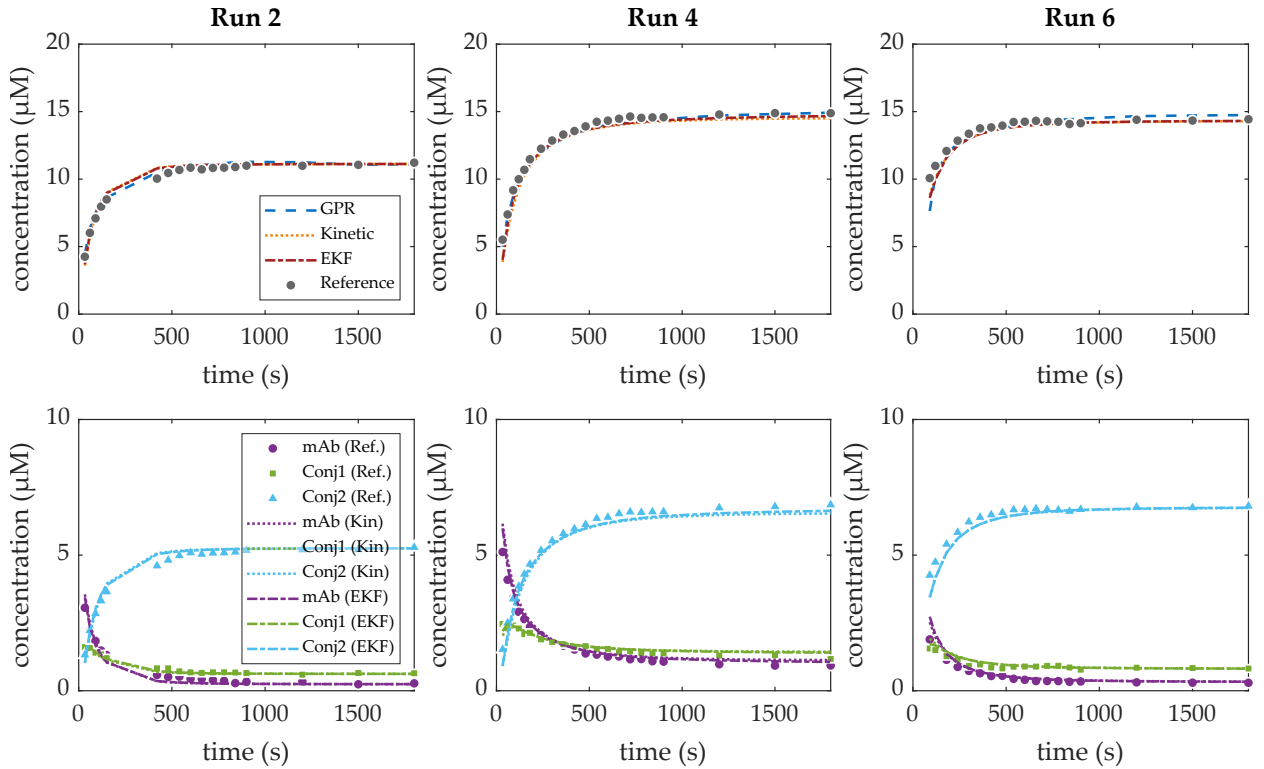


Figure A.6 Model and EKF predictions for all test runs. For each run, the first row shows the predictions of conjugated drug concentration and the second row the ADC species concentration.

B

Appendix Chapter 2 Supplementary Material for Chapter 4

Subset selection for data augmentation

This section provides additional information on the subset selection performed during data augmentation. For clarity, we provide a simple example which may directly be transferred to any case dealt with in this study. Let the data set contain 100 training samples at concentrations from 0 to 5 g L^{-1} for component 1 and from 0 to 10 g L^{-1} for component 2. The following is a stepwise explanation of the LSA method:

1. **Estimation of the concentration distributions** The kernel-density estimation (KDE) aims to approximate the concentration distribution in the experimental data set. This is done using Gaussian kernels which are summed to obtain a multimodal Gaussian distribution. Figuratively, the experimentally obtained concentration values are plotted as a histogram. The KDE tries to approximate the histogram by fitting multiple Gaussian distributions. We compared the KDE approach with a uniform and a unimodal Gaussian distribution by training convolutional neural networks on the augmented data sets generated with the three

distribution types. As the experimental data set only consists of discrete points with the most samples having low to medium protein concentration, the uniform and unimodal Gaussian approximation of the experimental data are thought to oversample regions where only few experimental data points are available. Analogously, the KDE approach is considered to only sample in similar concentration regions as provided in the experimental data.

2. **Sampling of a new concentration vector** From the estimated concentration distributions, a random sample is drawn. In our example, let this point have the concentrations of 1 and 2 g L^{-1} for components 1 and 2, respectively.
3. **Computation of the vector distance** The vector distances between the new concentration vector for which a new spectrum is to be generated and all concentration vectors (i.e. all samples in the training set) are computed. For an exemplary concentration vector from the training data with the concentrations 3 and 5 g L^{-1} for components 1 and 2, respectively, the respective vector distance is given by Equation (1) in the manuscript and yields $\sqrt{(1-3)^2 + (2-5)^2}$ if we are using the Euclidean distance. This is repeated for all 100 data points in the training data.
4. **Selection of a subset** The vector distances are sorted in ascending order and a subset is formed by the n_{LSA} data points with smallest distance, effectively grouping together data points with similar concentrations.
5. **Generation of the *in silico* spectra** From the selected subset, the pure component profiles are estimated by solving Equation (2) in the manuscript. Using Equation (3) in the manuscript, the pure component profiles are recombined into a new *in silico* spectrum for the sampled concentration vector. Steps 2 to 5 are repeated until the number of desired augmented data points is reached.

Practical remarks for hyperparameter optimization

Although HPO facilitates automated tuning of ML models, it is important to mention that several prerequisites should be fulfilled when starting the optimization: (1) The objective function should be representative of the modeling objective. The herein used sum of coefficients of determination ensure equal weighting of all components which could similarly be achieved by other forms of normalization. (2) The configuration of the optimization algorithm was found to be crucial in determining a globally optimal solution to not end up in premature, local minima of the objective function. The random sampling employed in the first 100 trials ensured that the search space was explored without any bias from initially sampled trials. Alternatively, the tree-structured parzen estimator (TPE) sampler can be adjusted to balance between exploration and exploitation of the search spaces to achieve this. (3) In accordance with (1), it should further be ensured that all monitored optimization metrics are representative of one another such that e.g. the cross-validation error portrays the training or validation error in order to be able to guide to optimization towards a model capable of generalization. Especially, when pruning is used, the pruning metric should be selected to resemble the actual optimization metric. And lastly (4), it is advisable to conduct a preceding parameter study for selecting critical hyperparameters to be optimized. Adding more hyperparameters to the optimization problem, considerably increases the search space and thus may require longer optimization periods.

Supplementary Tables and Figures

Tables

Table B.1 Overview of experimental conditions for data set 1 and 2.

Data set	Experiment	Elution Conditions	Number of samples	Subset
1	Run 1	250-400 mM NaCl (3 CV)	60	Training
1	Run 2	150-600 mM NaCl (3 CV)	56	Training
1	Run 3	230-460 mM NaCl (3 CV)	57	Training
1	Run 4	170-540 mM NaCl (3 CV)	60	Training
1	Run 5	200-500 mM NaCl (3 CV)	57	Test
2	Run 1	4 CV	101	Training
2	Run 2	5 CV	119	Test
2	Run 3	6 CV	139	Training
2	Run 4	7 CV	192	Training
3	Run 1	10 CV	72	Training
3	Run 2	20 CV	112	Test
3	Run 3	30 CV	164	Training
4	Randomly assigned	-	64	Training
4	Randomly assigned	-	16	Test

Table B.2 Optimized PLS hyperparameters for data sets 1-4. All data were preprocessed using a Savitzky-Golay filter (SGF) and mean-centering before being passed to the PLS models. For data sets 1-3, the number of components, the SGF window size and the order of derivative were optimized using cross-validated grid-based search in the ranges 2-15, 3-31 and 0-2 for the number of components, the SGF window and the order of derivative, respectively. For data set 4, solely the number of components was optimized in the range of 2-15.

Hyperparameter	Data set 1			Data set 2		Data set 3			HMWS2	Oil	Data set 4	
	Rib A	Cyt C	Lys	Monomer	Aggregate	low molecular weight species (LMWS)1	Monomer	high molecular weight species (HMWS)1			Protein	Starch
number of components	5	7	7	4	9	11	2	4	7	14	5	6
SGF window	17	3	9	21	3	9	3	7	11	19	19	19
order of derivative	2	1	2	0	0	1	2	2	2	2	2	2

Table B.3 Comparison of LSA and EMSA method for data set 2 and 3 as measured by the RMSE for all components.

		LSA				EMSA				PLS			
	Component	Training		Test		Training		Test		Training		Test	
		RMSE	R2	RMSE	R2	RMSE	R2	RMSE	R2	RMSE	R2	RMSE	R2
Data set 2	monomer	1.044	0.990	1.1811	0.991	1.727	0.973	2.487	0.958	1.099	0.989	1.223	0.99
	aggregate	0.175	0.978	0.1903	0.98	0.320	0.927	0.370	0.923	0.411	0.879	0.494	0.862
Data set 4	oil	0.079	0.805	0.070	0.808	0.017	0.991	0.108	0.546	0.036	0.960	0.098	0.630
	protein	0.143	0.915	0.138	0.927	0.044	0.992	0.171	0.890	0.135	0.924	0.121	0.945
	starch	0.355	0.821	0.453	0.532	0.072	0.992	0.291	0.806	0.301	0.871	0.290	0.807

Figures

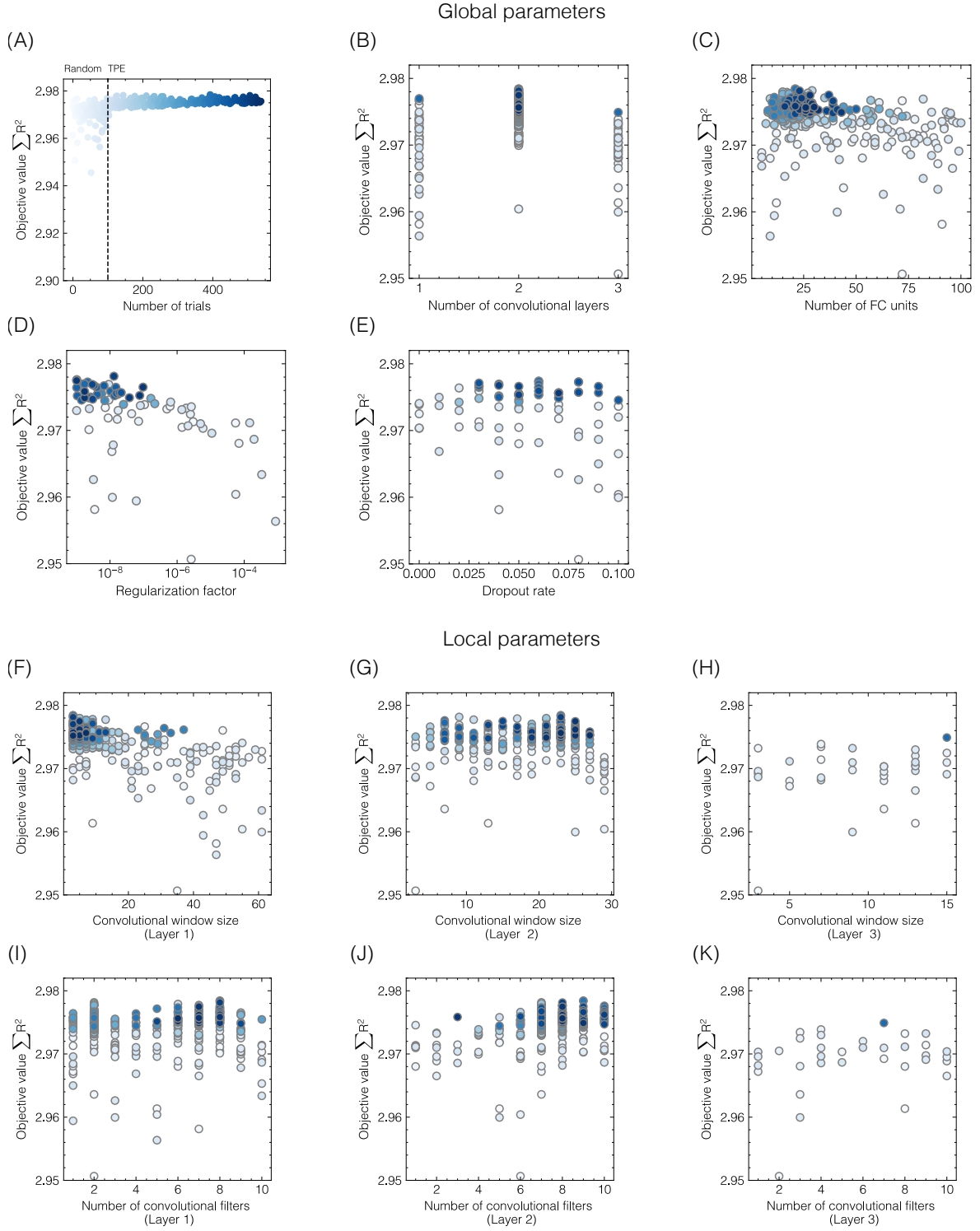
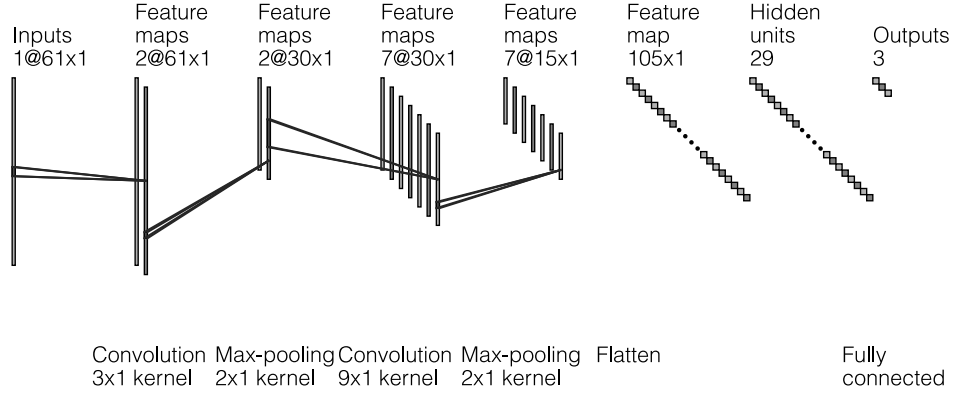


Figure B.1 Hyperparameter evolution profiles during optimization for data set 1. The objective values $\sum R^2$ are shown for all trials (A) and each hyperparameter individually (B – K). The colors of the circles indicate the number of HPO trials with darker shades of blue corresponding to later sampling points. Random sampling was performed for the first 100 trials after which median pruning and the TPE-based optimization were enabled.

(A)



(B)

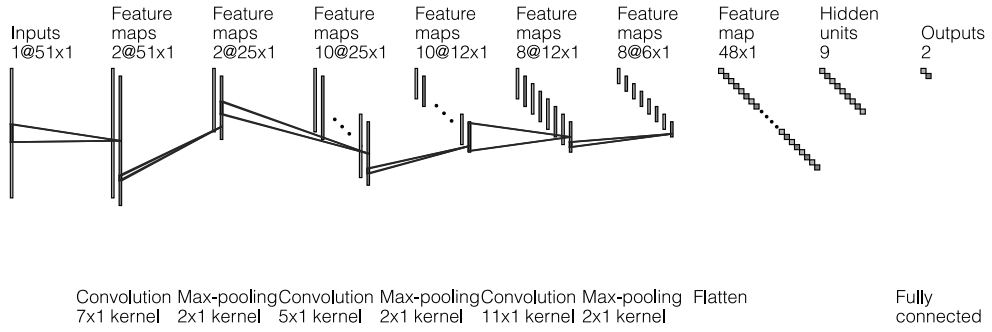


Figure B.2 Visualization of the architecture of the optimized CNN models for data set 1 (A) and data set 2 (B). Spectral data being passed through the convolutional and pooling layers are shown as rectangles. The convolutional window is illustrated by the connecting lines. The exact dimensions of the individual layers and their respective inputs and outputs are given as $n_{\text{filters}}@d_{\text{spectrum}}$ as part of the labels above. The dimension of the convolution and pooling operations are given as part of the labels below. For each convolution operation, a number of filters n_{filters} are used.

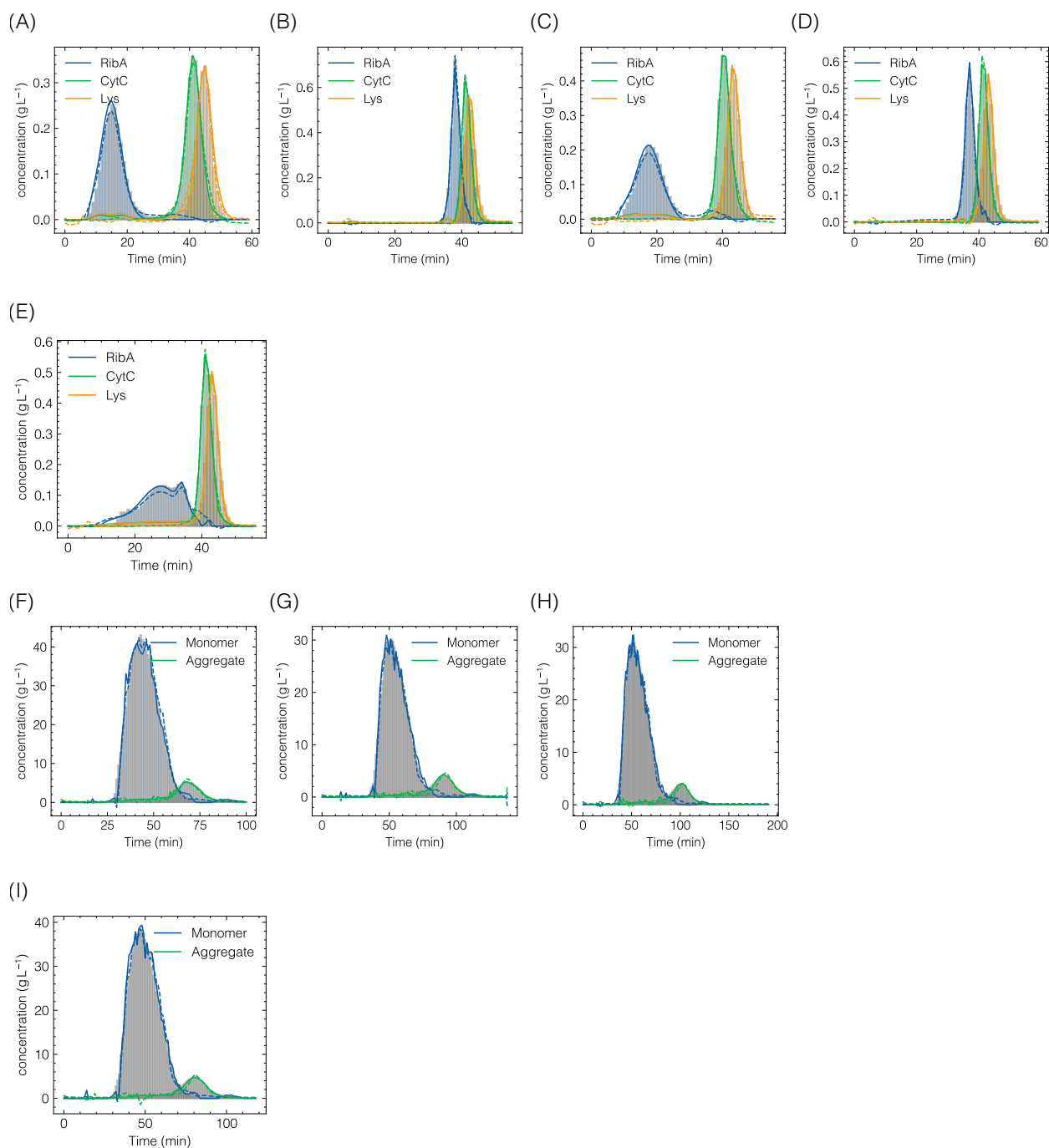


Figure B.3 Predictions of the optimized CNN models. For data set 1, the training runs are shown in (A-D) and the test run is shown in (E). For data set 2, the training runs are shown in (F-H) and the test run in (H). The reference data is overlaid with the predicted values for all species as indicated by the bars and the solid lines, respectively. The PLS predictions are shown as dashed lines.

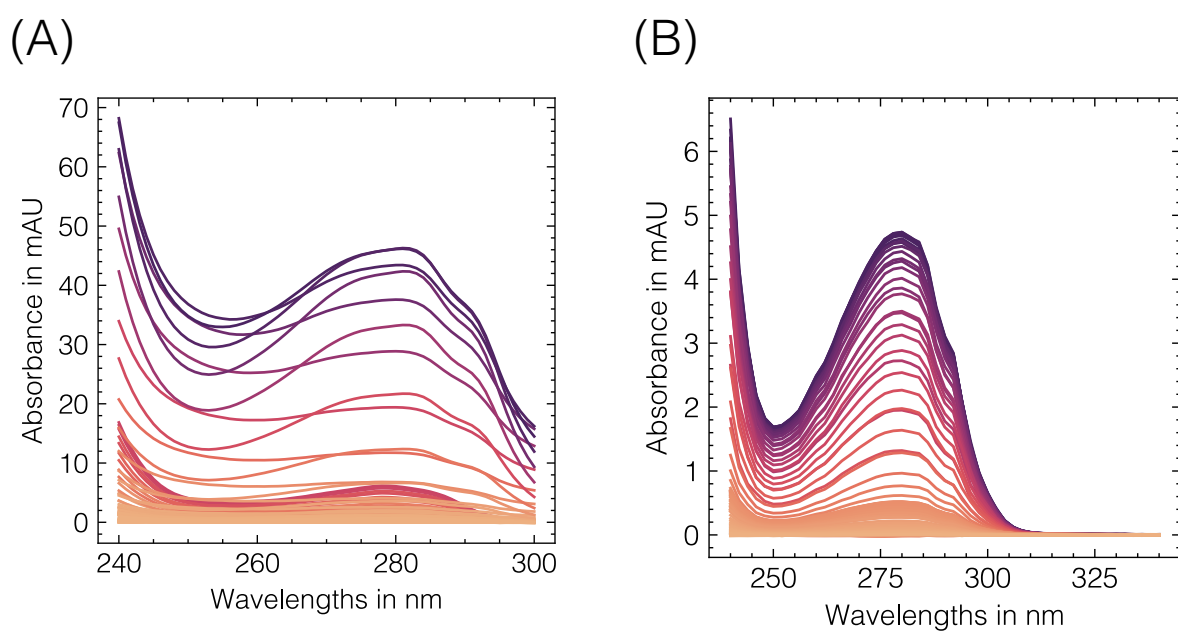


Figure B.4 Exemplary raw UV/Vis spectra for data set 1 (A) and data set 2 (B). Run 1 and Run 3 are exemplarily shown for data set 1 and 2, respectively. The spectra are colored according to the total protein concentration in each sample with darker colors corresponding to higher concentration.

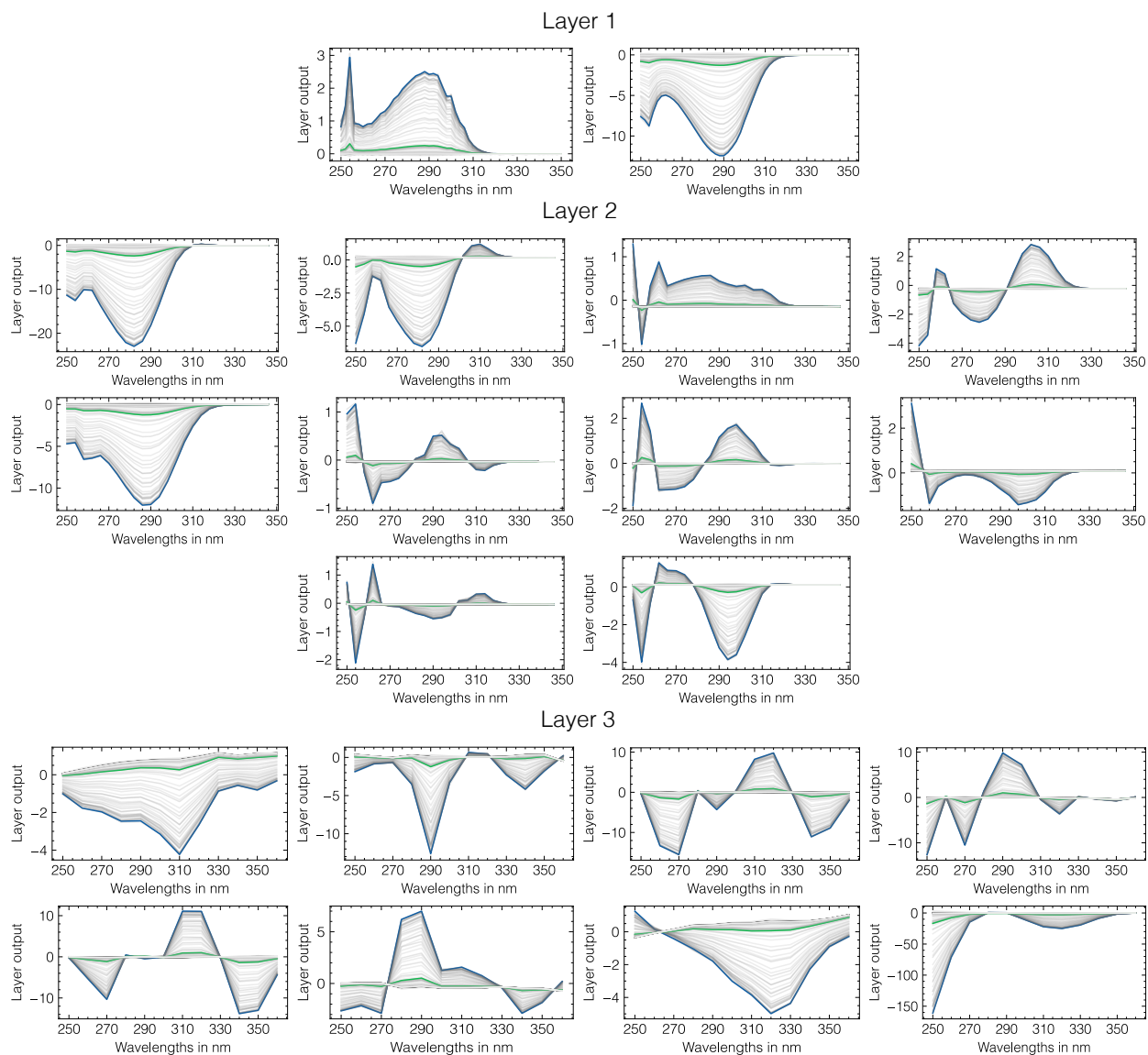


Figure B.5 Layer visualization for all convolutional layer for the optimized CNN model for data set 2. For each convolutional filter, the outputs for all available spectra in the test subset are shown as gray lines. The spectra with the highest reference concentration for monomer and aggregate species are marked in blue and green, respectively. The dimensions of the outputs are reduced by a factor of 2 due to interposed pooling layers causing the resolution of the displayed lines to decrease.

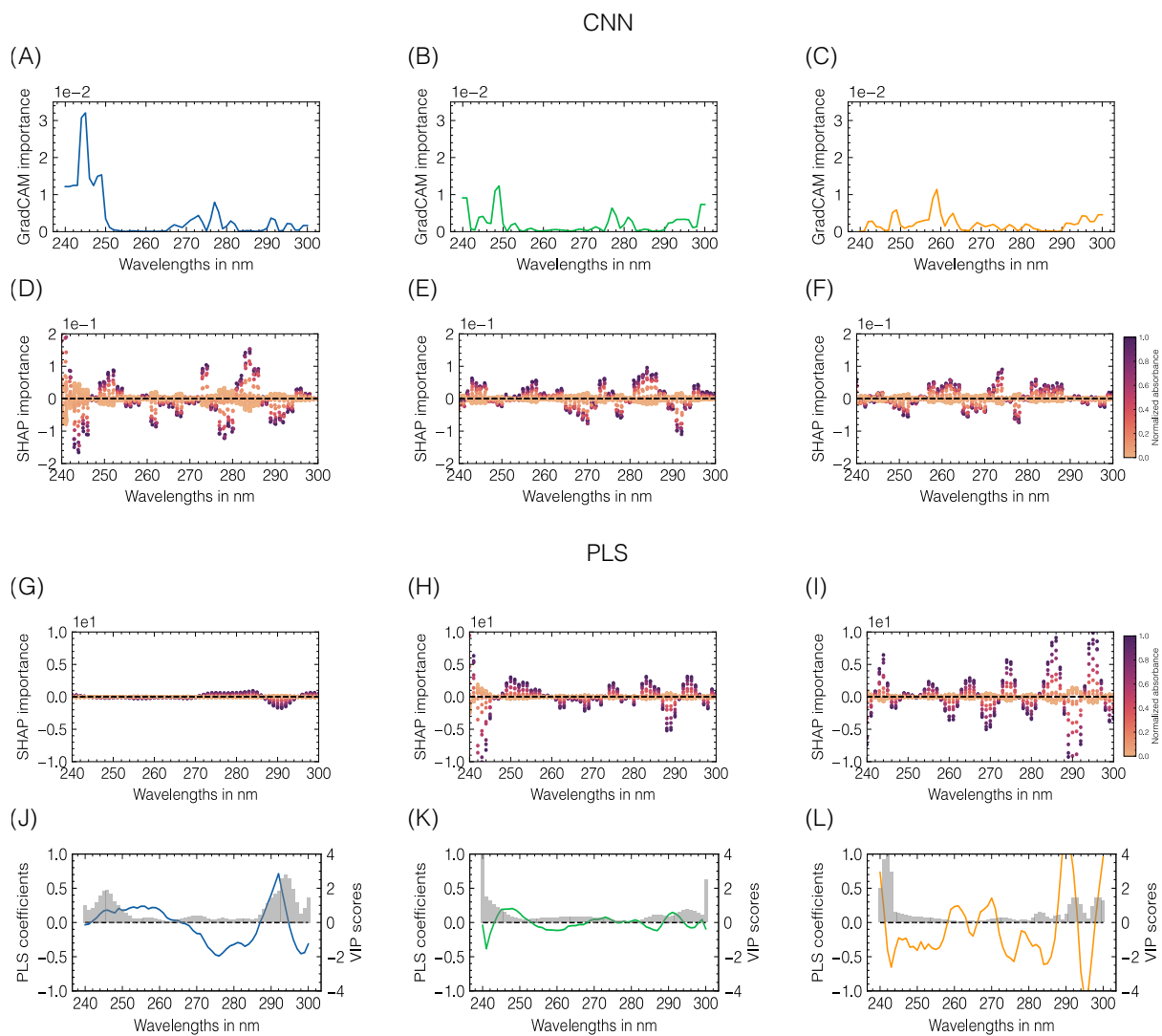


Figure B.6 Model-specific and model-agnostic importance measures for data set 1. CNN-specific gradient-weighted class activation map (GradCAM) importance values (A-C), SHAP values for CNN (D-F) and PLS (G-I) and PLS-specific regression coefficients and VIP scores (J-L). The regression coefficients and VIP scores are shown as solid lines and bars, respectively. The left, middle and right columns correspond to importance metrics for Rib A, Cyt C and Lys, respectively.

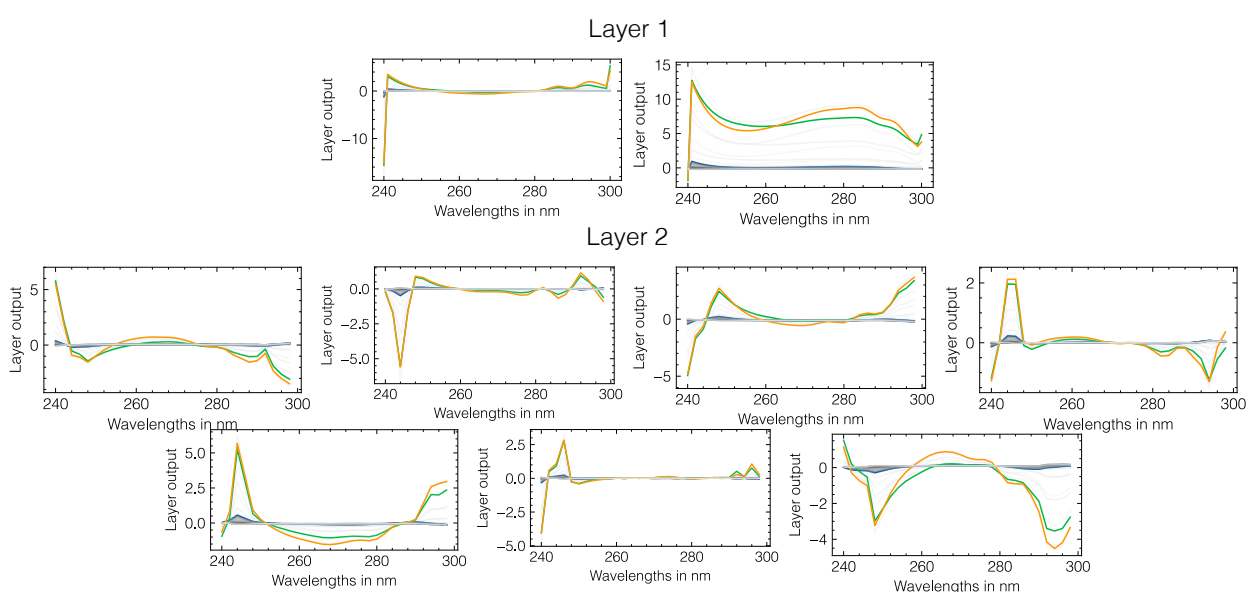


Figure B.7 Layer visualization for all convolutional layer for the optimized CNN model for data set 1. For each convolutional filter, the outputs for all available spectra in the test subset are shown as gray lines. The spectra with the highest reference concentration for Rib A, Cyt C and Lys are marked in blue, green and orange, respectively. The dimensions of the outputs are reduced by a factor of 2 due to interposed pooling layers causing the resolution of the displayed lines to decrease.



Appendix Chapter 3

Supplementary Material for Chapter 5

Buffers, solutions, and spiking materials

If not otherwise stated, chemicals (Merck KGaA, Darmstadt, DE) and ultrapure water (PURELAB Ultra, ELGA LabWater, Lane End, High Wycombe, UK) were used for buffer and stock solution preparation. Buffers were pH-adjusted with 32 % HCl using a SenTix62 pH electrode (WTW, Weilheim, DE) coupled to a HI 3220 pH meter (Hanna Instruments, Woonsocket, US). Buffers and stock solutions were filtered through a 0.2 μ m pore-size cellulose acetate filter (VWR International, Radnor, US). For all experiments, stock solutions used for polysorbate adjustment and precipitation experiments were 10% (v/v) polysorbate 20 and 4 M AMS, respectively. Lysis or high-salt lysis buffers consisted of 50 mM Tris, 100 mM or 1500 mM NaCl, 1 mM EDTA (AppliChem GmbH, Darmstadt, DE) at pH 8.0, respectively.

For the VLP-enriched spiking material, clarified lysate was further purified by precipitation and re-dissolution, diafiltration, and chromatography as described by Hillebrandt, Vormittag, Dietrich, Wegner, et al. (2021). The purified VLPs were further dialyzed into lysis buffer overnight using 10 kDa MWCO Slide-A-Lyzer G2 cassettes (Thermo Fisher Scientific Inc., Waltham, US) and concentrated

by ultrafiltration with Vivaspin 20 centrifugal filters (Sartorius Stedim Biotech GmbH, Göttingen, DE) in a 5810R centrifuge (Eppendorf, Hamburg, DE). VLP concentration was determined using a NanoDrop 2000c UV/Vis spectrometer (Thermo Fisher Scientific Inc.) and the theoretical extinction coefficient at 280 nm of $1.764 \text{ L g}^{-1} \text{ cm}^{-1}$ as provided by the ProtParam tool (Gasteiger et al. 2005). VLP-enriched spiking material concentrations were 6.95 g L^{-1} for batch precipitation experiment B5 and 3.85 g L^{-1} for batch and fed-batch precipitation experiments B6 and F2, respectively. For both VLP-enriched spiking materials, a A260/A280 ratio of 0.63 indicated almost pure protein. For the HCP-enriched spiking material, clarified lysate was adjusted to 0.25% (v/v) polysorbate 20 and precipitation was conducted at 1.1 M AMS for 30 min under stirred conditions. Precipitate solution was centrifuged at 12000 G and 4 °C for 30 min. Supernatant was collected and a similar procedure of dialysis and concentration was performed as for the VLP-enriched spiking material, but using 2 kDa MWCO devices. HCP-enriched spiking material characteristics were a A280 value of 57.3 mAU and a A260/A280 ratio of 1.88, both indicating host cell (HC) nucleic acids in the spike solution.

Supplementary Tables and Figures

Figures

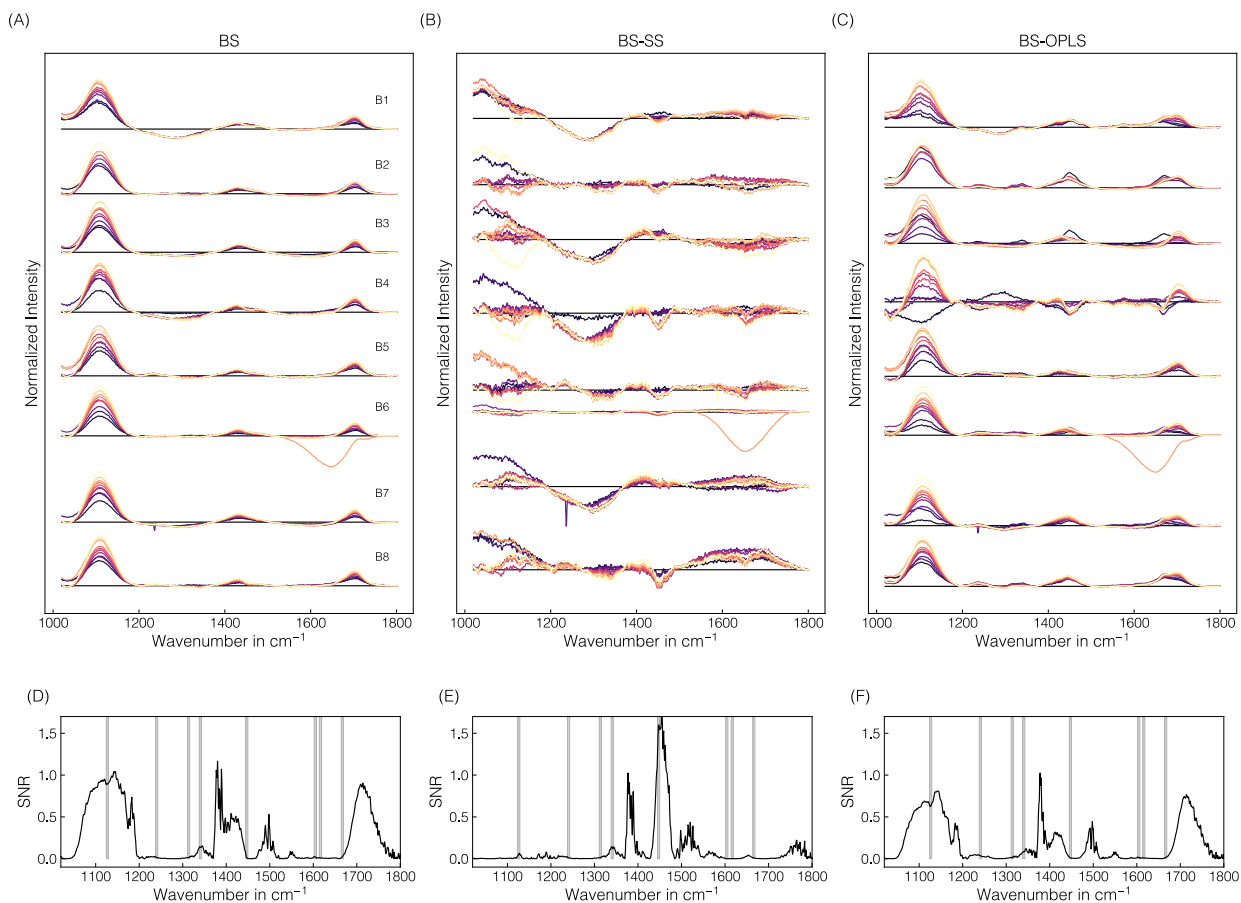


Figure C.1 Comparison of the effects of preprocessing operations on Raman spectra in supernatant samples for the wavenumber region $1020\text{-}1800\text{ cm}^{-1}$. In (A)-(C), difference spectra are shown for all batch experiments B1-B8 after turbidity and baseline correction, with incorporated scaled subtraction (SS)-background correction, or with incorporated OPLS-background correction, respectively. The spectra are colored according to the AMS concentration with brighter colors denote higher concentrations. In (D)-(F), the corresponding SNR are shown. Gray shaded areas indicate protein-related Raman regions according to literature (Maiti et al. 2004; Rygula et al. 2013).

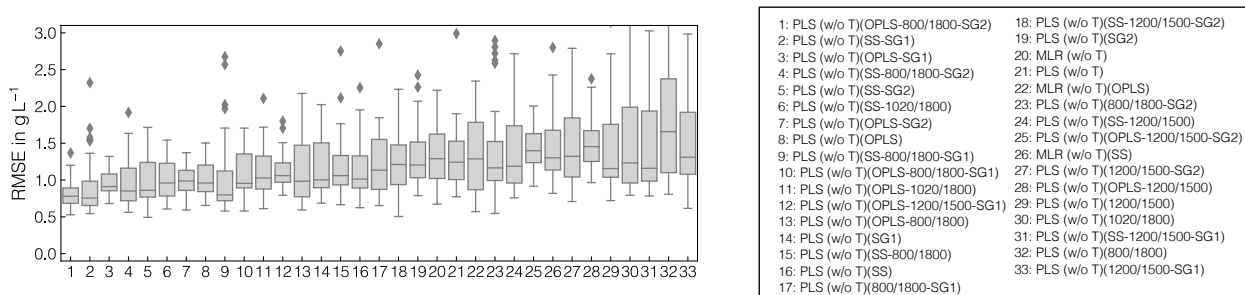


Figure C.2 Comparison of model types and preprocessing pipelines without turbidity correction. The distributions of RMSE of the outer cross-validation are presented and ranked by the mean of the RMSEs for all tested model configurations. All tested model configurations comprised baseline correction and difference spectra. The solid lines within the boxes represent the median of the obtained performances.

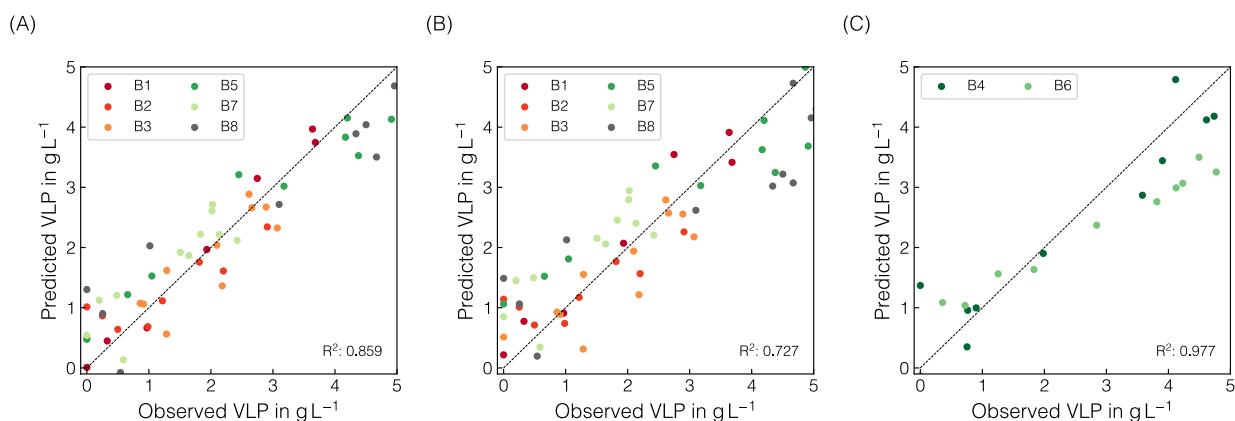


Figure C.3 PLS model predictions with regard to precipitated VLP concentration for batch experiments. The predicted and observed VLP concentrations for calibration (A), cross-validation (B) and test sets (C) are plotted against each other with the tie line indicating the optimal result.

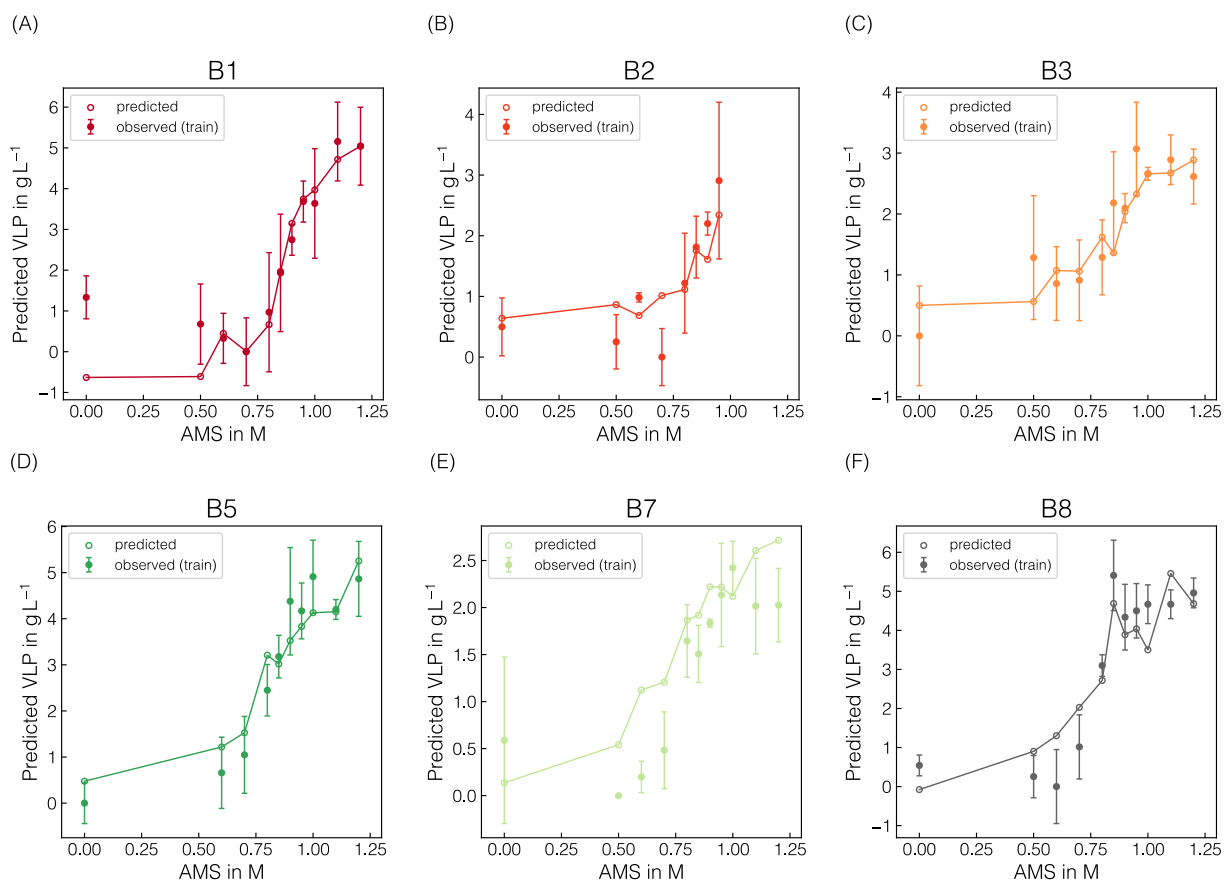


Figure C.4 PLS model predictions with regard to the precipitated VLP concentration for batch experiments over AMS concentrations. Batch experiments assigned to the training set (B1, B2, B3, B5, B7, B8) are shown in (A)-(F), respectively. The predicted and observed VLP concentrations for all calibration experiments are displayed as empty and full circles, respectively. For visual purposes, solid are shown to linearly connect the PLS model predictions. For batch experiment B2, the data points above 0.95 M were excluded due to defective Raman spectra.

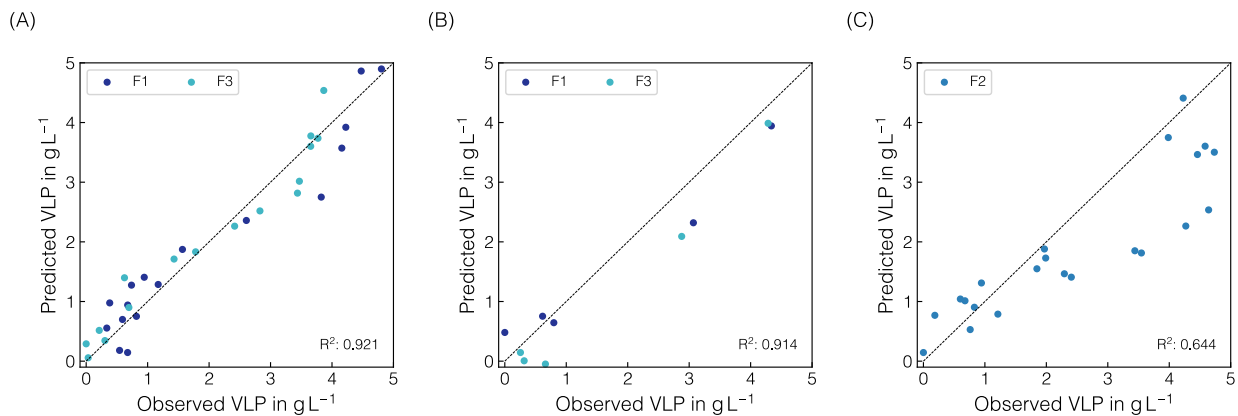


Figure C.5 PLS model predictions with regard to precipitated VLP concentration for fed-batch experiments. The predicted and observed VLP concentrations for calibration **(A)**, validation **(B)** and test sets **(C)** are plotted against each other with the tie line indicating the optimal result.

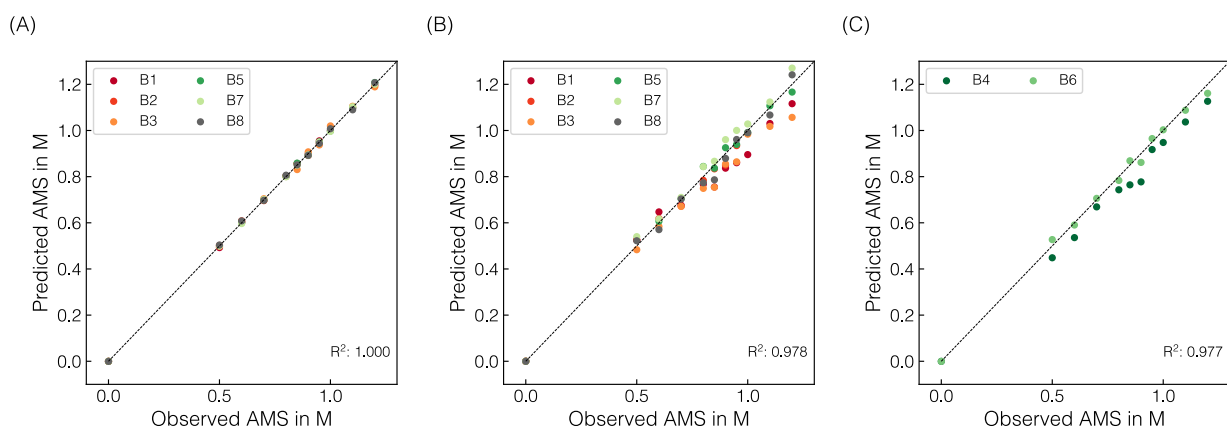


Figure C.6 PLS model predictions with regard to AMS concentration for batch experiments. The predicted and observed AMS concentrations for training **(A)**, cross-validation **(B)** and test sets **(C)** are plotted against each other with the tie line indicating the optimal result.

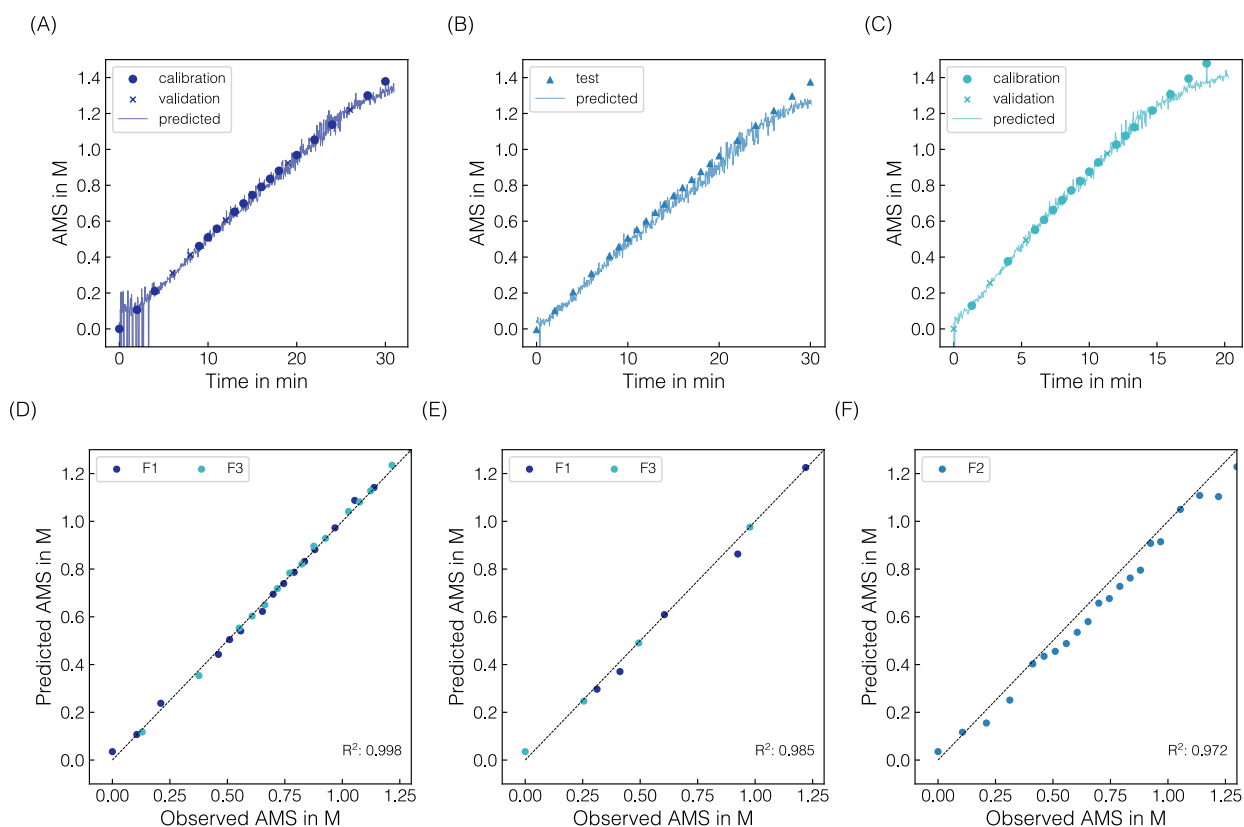


Figure C.7 PLS model predictions with regard to AMS concentration for fed-batch experiments. In (A)-(C), the timely predictions obtained from real-time Raman spectra are shown as solid lines and observed AMS concentrations are shown as scattered points. The data points belonging to the calibration, validation and test sets are marked with circles, crosses and triangles, respectively. The PLS predictions for the calibration, validation and test sets are plotted against the observed AMS concentrations in (D)-(F), respectively.

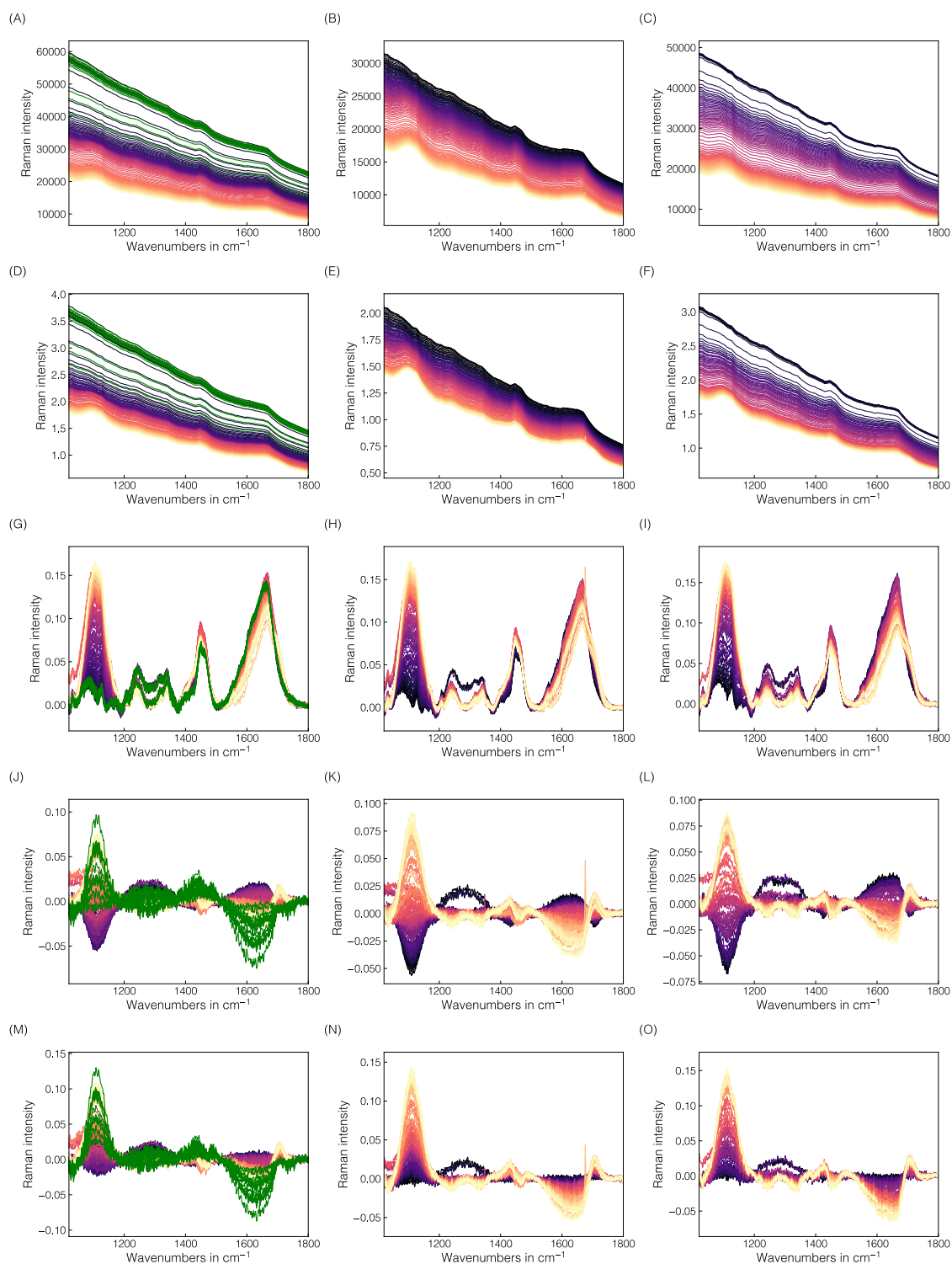


Figure C.8 Visualization of preprocessing operations for fed-batch experiments F1-F3 from left to right. The individual rows depict the Raman spectra at different stages of the preprocessing pipeline: Raw (A)-(C), turbidity correction (D)-(F), baseline (G)-(I), OPLS background (J)-(L) and difference spectra (M)-(O). For visual purposes, only every 5th spectra from each fed-batch experiment is shown. The Raman spectra affected by scattering in the PLS predictions are marked in green. The Raman spectra are colored according to the theoretical AMS concentration in the system with brighter colors denoting higher concentrations.

D

Appendix Chapter 4 Supplementary Material for Chapter 6

Supplementary Figures and Tables

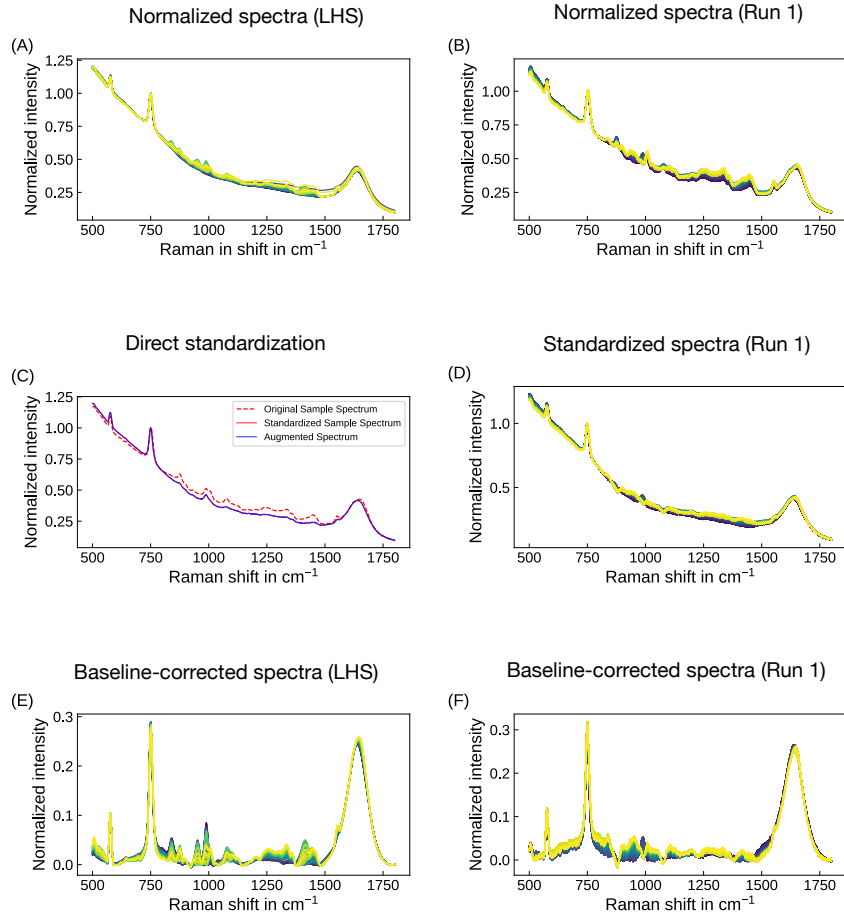


Figure D.1 Comparison of off-line and in-line spectra after different stages of preprocessing. Spectra from the LHS dataset after smoothing, cropping and normalization at 750 cm^{-1} (A). Spectra from the UFDF Run 1 dataset after smoothing, cropping and normalization at 750 cm^{-1} (B). Illustration of the direct standardization approach (C): An augmented spectrum (blue) is generated using the normalized spectra from the LHS dataset which matches the initial conditions of the UFDF Run 1 using the algorithm as described in (Schiemer, Rdt, and Hubbuch 2024). The initial spectrum from the UFDF Run 1 dataset (red dashed) is transformed using the direct standardization weights. The weights are subsequently used to transform all spectra in the dataset to produce the standardized spectra (D). Baseline-corrected spectra from the LHS (E). Baseline-corrected spectra from the UFDF Run 1.



Appendix Chapter 5
Supplementary Material for Chapter 7

Figures

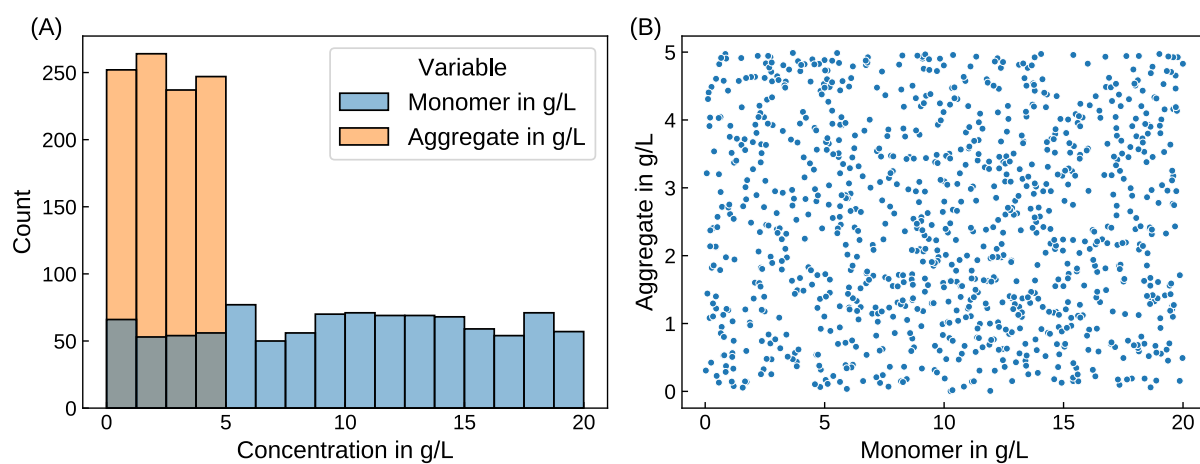


Figure E.1 Distribution of sampled values for augmentation test. (A) Histogram of the sampled values for monomer and aggregate concentrations. (B) Two-dimensional scatter plot across the sampled value range.

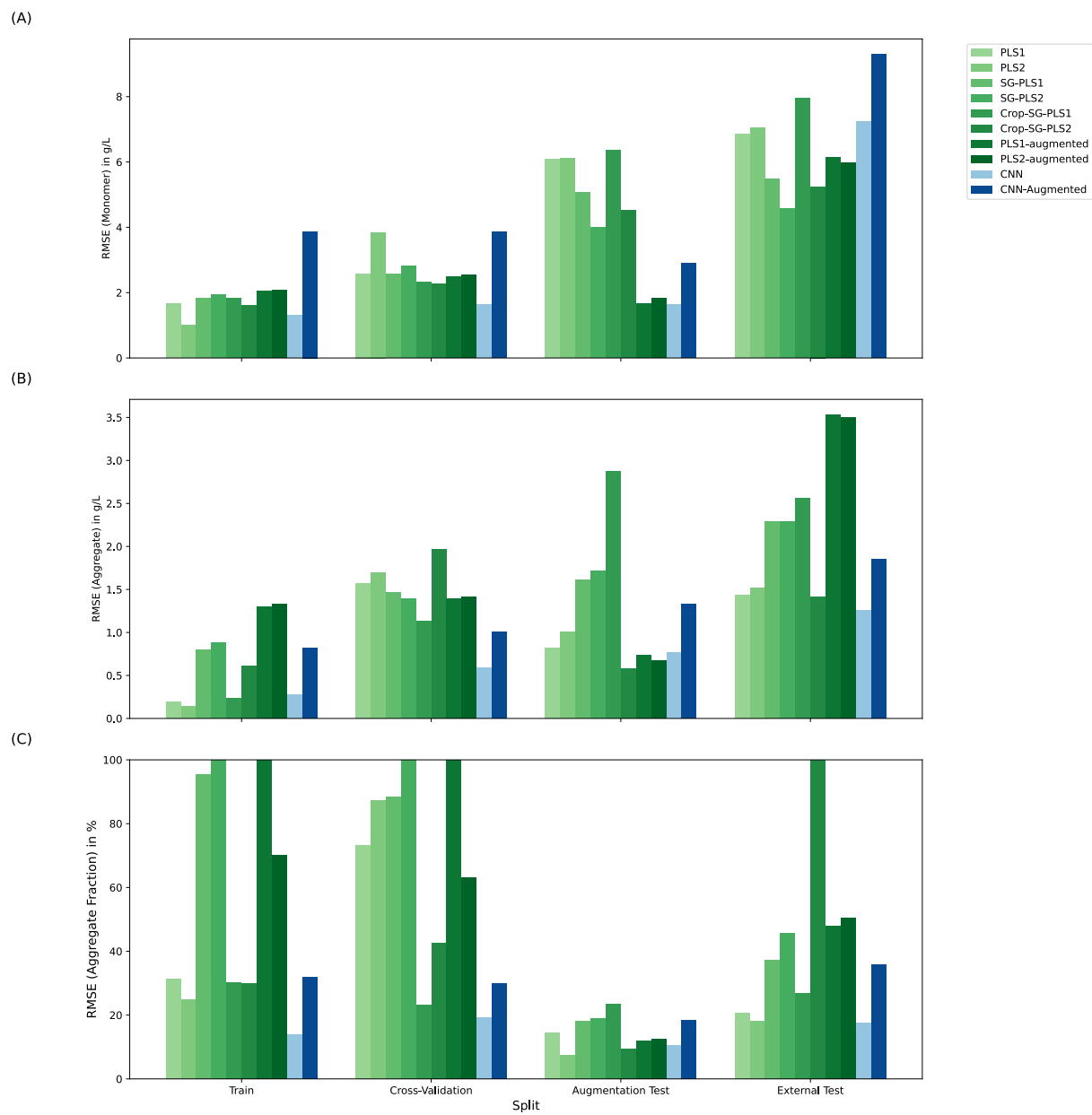


Figure E.2 Performance comparison for all studied model variants for training, cross-validation, augmentation test and external test data sets. (A) RMSE for Monomer component, (B) RMSE for Aggregate component, (C) RMSE for Aggregate fraction.

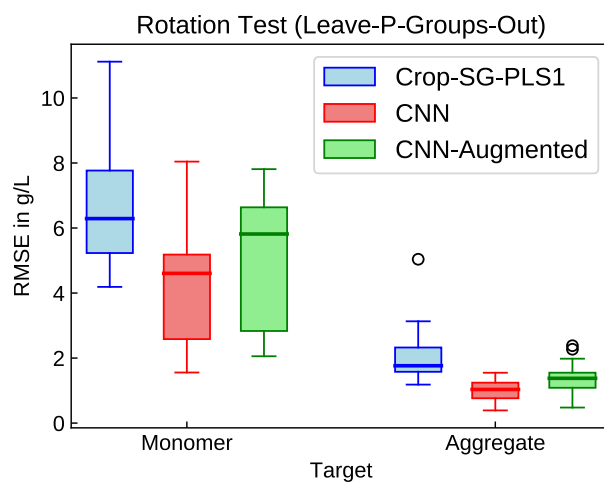


Figure E.3 Performance comparison of representative PLS and CNN models for rotation test. Distribution of RMSE for the rotated test sets during a leave-4-batches-out cross-validation scheme using all available batches.

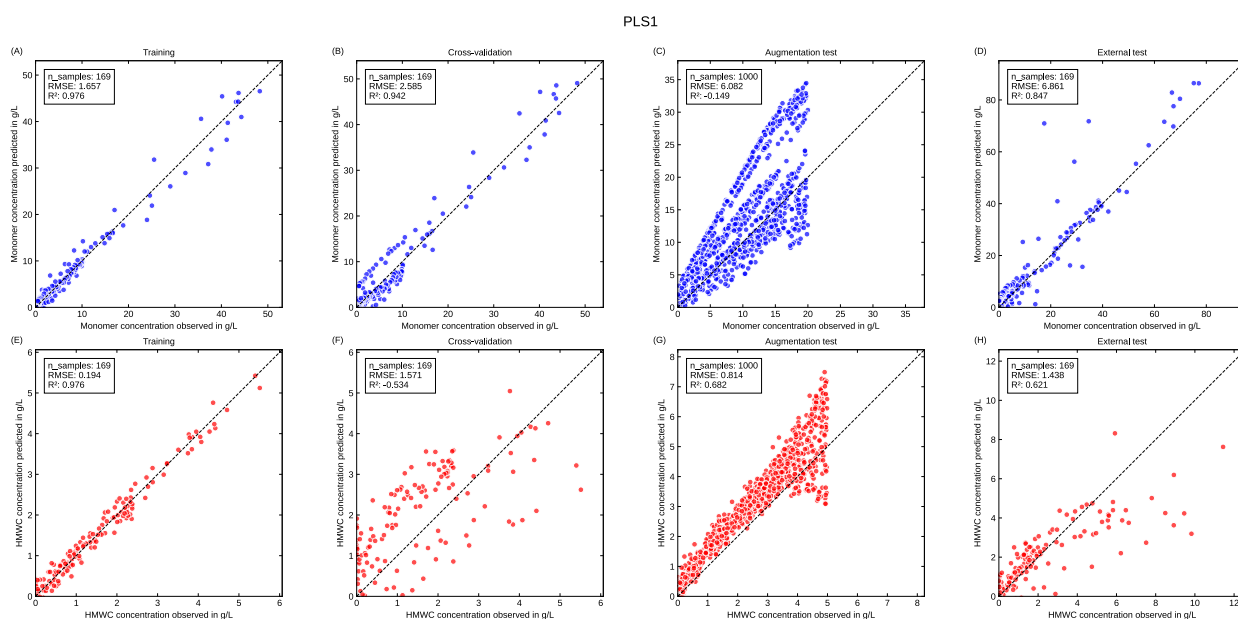


Figure E.4 Parity plots for % HMWC prediction using the PLS1 model across training, cross-validation, and test datasets. The dashed line marks the ideal parity. R^2 and RMSE values are annotated in each panel.

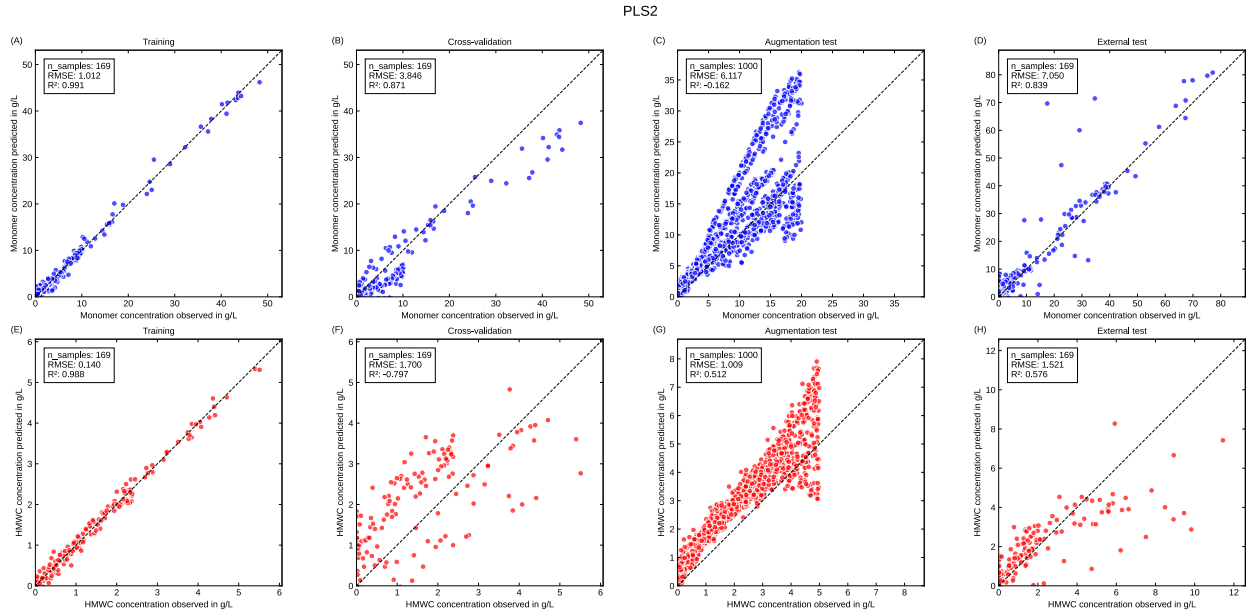


Figure E.5 Parity plots for % HMWC prediction using the PLS2 model across training, cross-validation, and test datasets. The dashed line marks the ideal parity. R^2 and RMSE values are annotated in each panel.

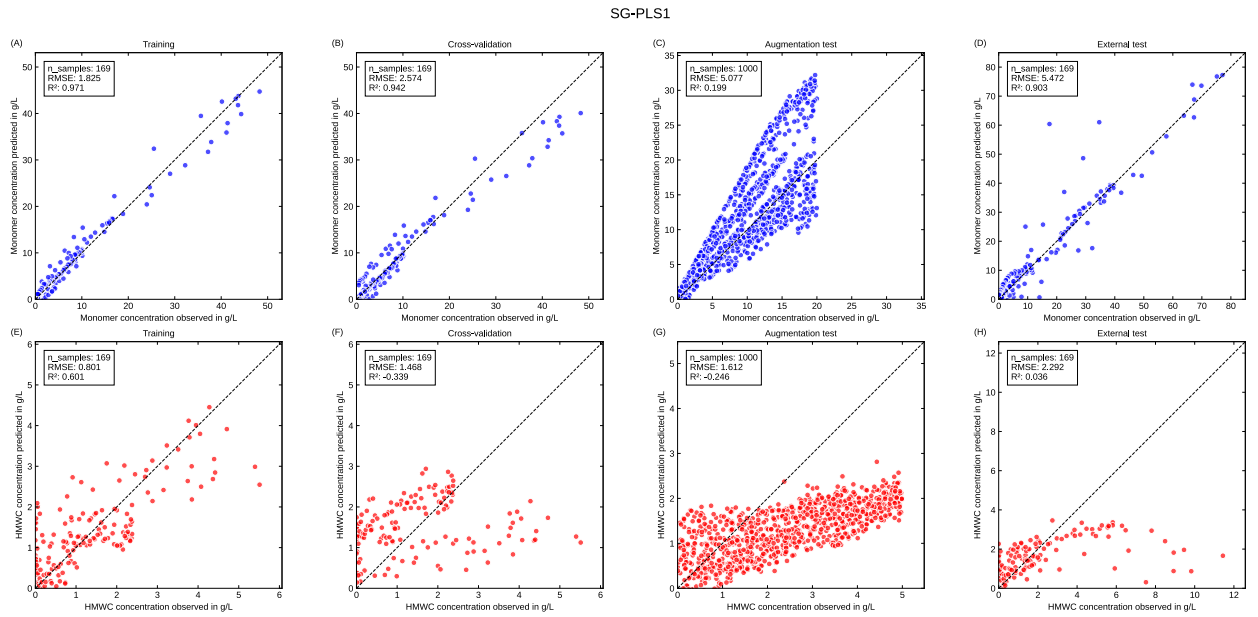


Figure E.6 Parity plots for % HMWC prediction using the SG-PLS1 model across training, cross-validation, and test datasets. The dashed line marks the ideal parity. R^2 and RMSE values are annotated in each panel.

SG-PLS2

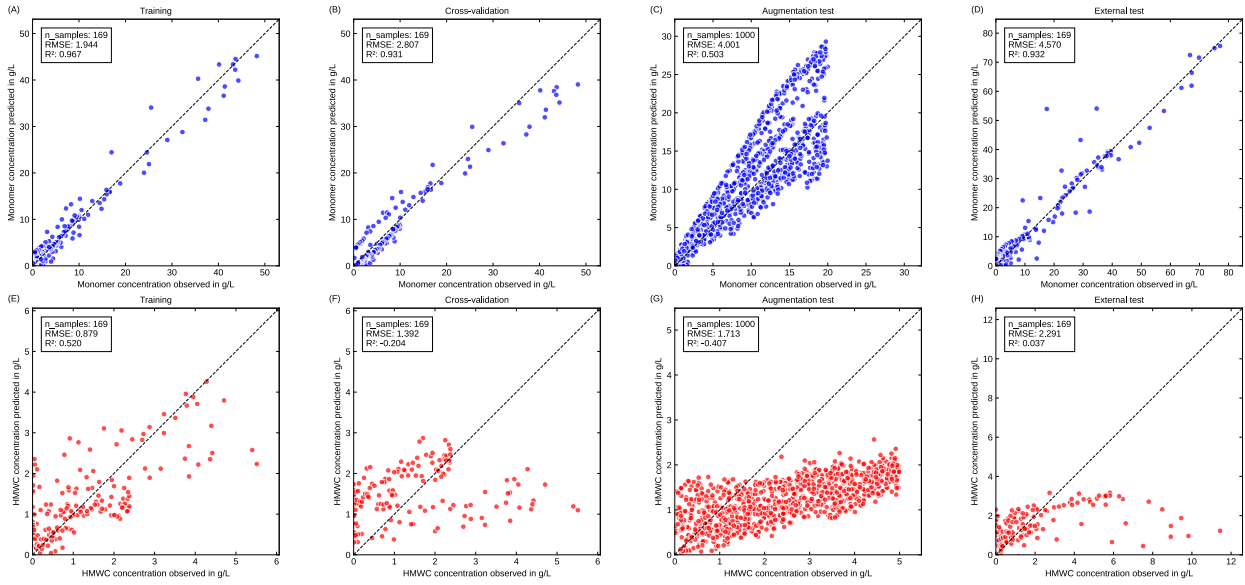


Figure E.7 Parity plots for % HMWC prediction using the SG-PLS2 model across training, cross-validation, and test datasets. The dashed line marks the ideal parity. R^2 and RMSE values are annotated in each panel.

Crop-SG-PLS1

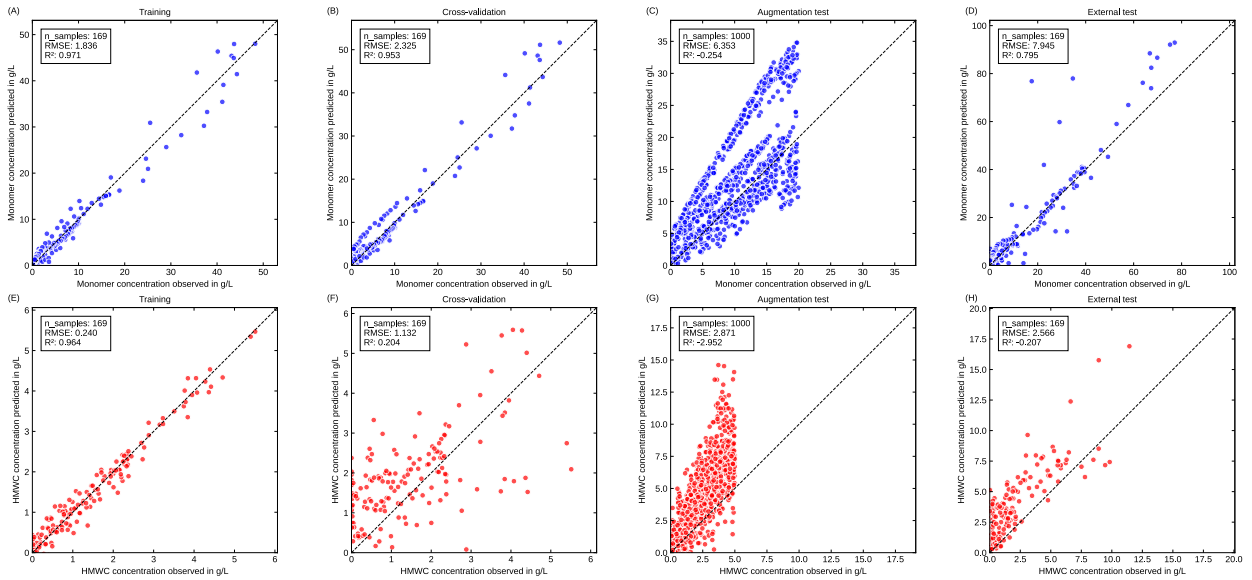


Figure E.8 Parity plots for % HMWC prediction using the Crop-SG-PLS1 model across training, cross-validation, and test datasets. The dashed line marks the ideal parity. R^2 and RMSE values are annotated in each panel.

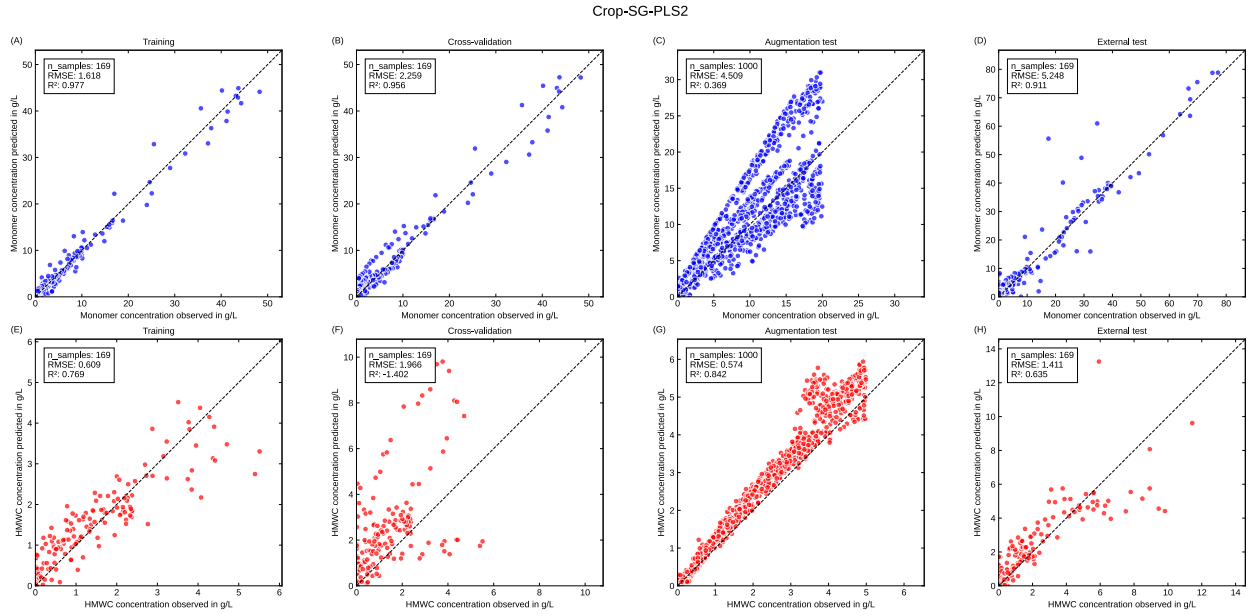


Figure E.9 Parity plots for % HMWC prediction using the Crop-SG-PLS2 model across training, cross-validation, and test datasets. The dashed line marks the ideal parity. R^2 and RMSE values are annotated in each panel.

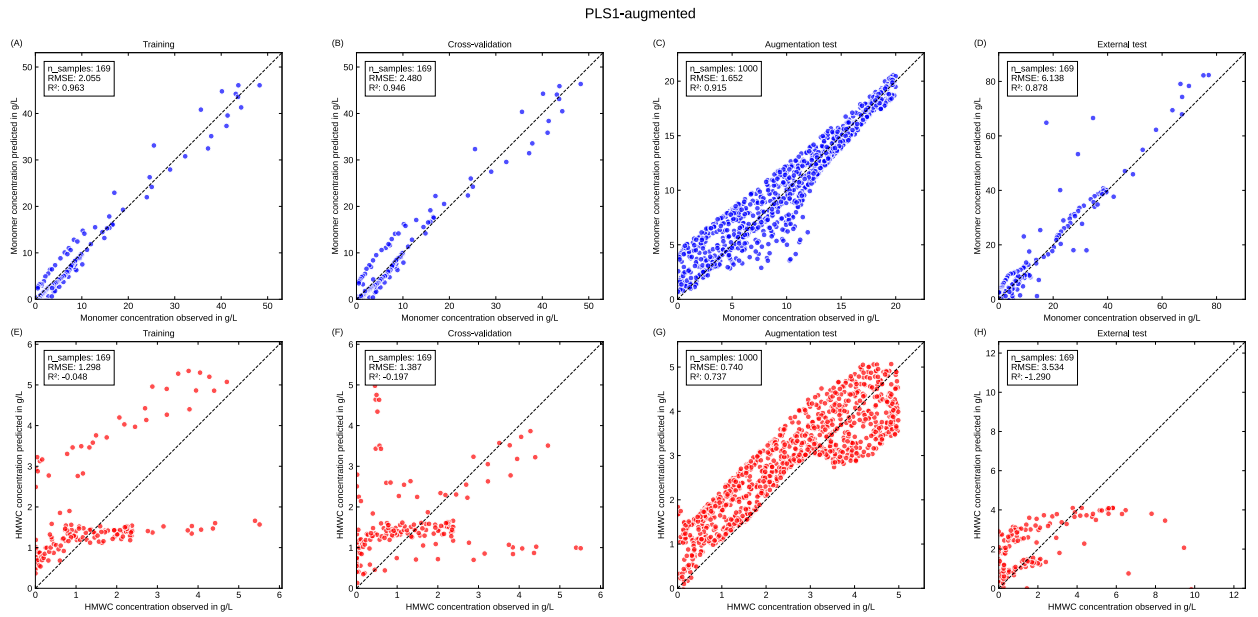


Figure E.10 Parity plots for % HMWC prediction using the PLS1-Augmented model across training, cross-validation, and test datasets. The dashed line marks the ideal parity. R^2 and RMSE values are annotated in each panel.

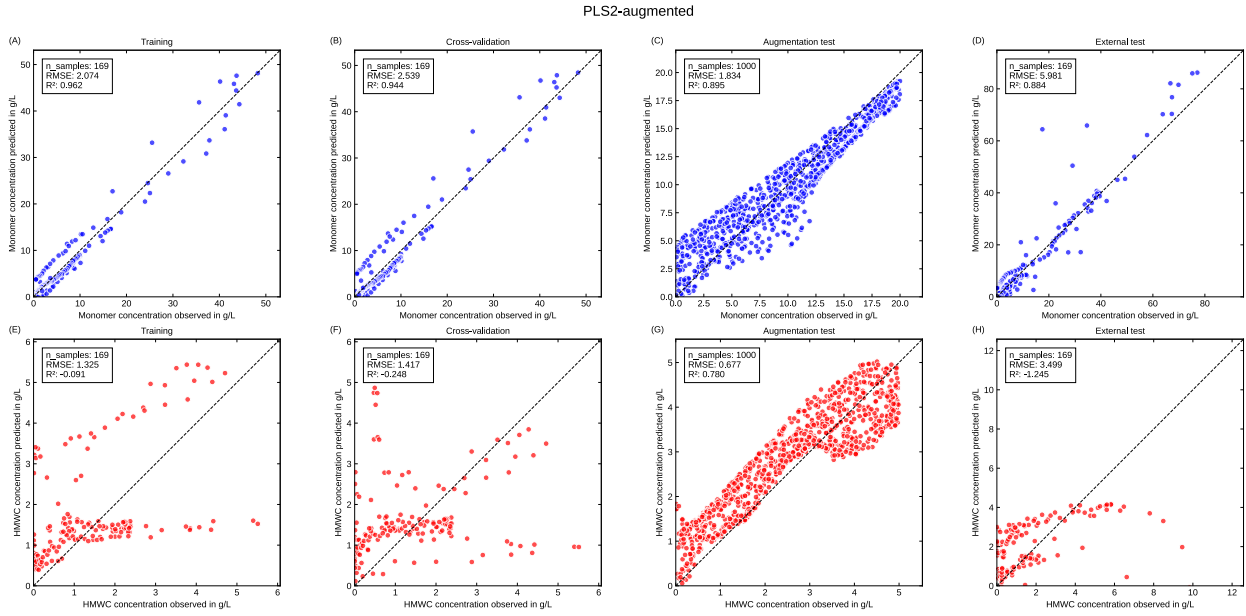


Figure E.11 Parity plots for % HMWC prediction using the PLS2-Augmented model across training, cross-validation, and test datasets. The dashed line marks the ideal parity. R^2 and RMSE values are annotated in each panel.

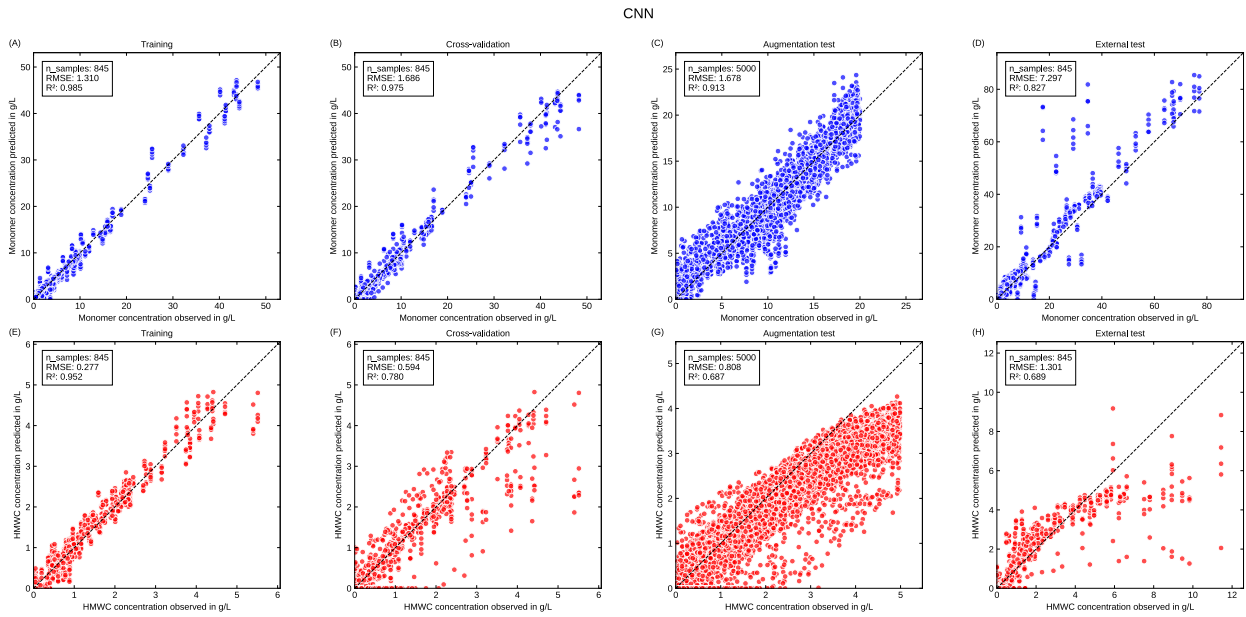


Figure E.12 Parity plots for % HMWC prediction using the CNN model across training, cross-validation, and test datasets. The dashed line marks the ideal parity. R^2 and RMSE values are annotated in each panel.

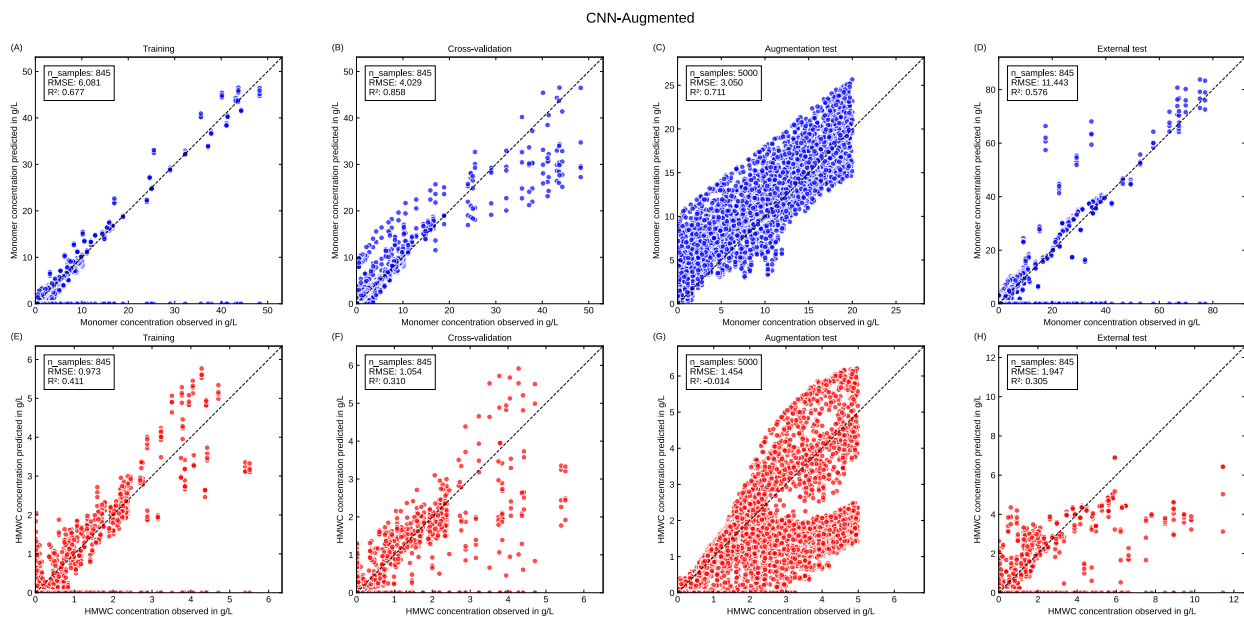


Figure E.13 Parity plots for % HMWC prediction using the CNN-Augmented model across training, cross-validation, and test datasets. The dashed line marks the ideal parity. R^2 and RMSE values are annotated in each panel.