Protein Monitoring and Interpretation of Spectroscopic Signals using Data-Driven and Physically Informed Modeling

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. Jakob Heyer-Müller
aus Friedrichshafen

Erstgutachter:	Prof. Dr. Jürgen Hubbuch
Zweitgutachterin:	apl. Prof. Dr. Gisela Guthausen
Tag der mündlichen Prüfung:	30.10.2026



Chapter 3 is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0):
https://creativecommons.org/licenses/by-nc-nd/4.0/deed.de



Chapters 4 and 5 are licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0):
https://creativecommons.org/licenses/by/4.0/deed.de

Acknowledgements

Im Laufe meiner Promotion durfte ich auf die Unterstützung, den fachlichen Austausch und die Ermutigung vieler Menschen zählen. Dafür möchte ich mich an dieser Stelle herzlich bedanken.

Prof. Dr. Jürgen Hubbuch danke ich herzlich für die Möglichkeit zur Promotion am Institut für Molekulare Aufarbeitung von Bioprodukten (MAB). Das Zusammenspiel aus Eigenverantwortung, fachlichem Freiraum und unterstützender Begleitung hat meine Entwicklung in dieser Zeit wesentlich gefördert.

Mein Dank gilt apl. Prof. Dr. Gisela Guthausen für die Übernahme des Zweitgutachtens.

Mein besonderer Dank gilt Robin Schiemer für die gemeinsame wissenschaftliche Arbeit in verschiedenen Projekten, aus der zwei gemeinsame Veröffentlichungen hervorgegangen sind. Die Zusammenarbeit war geprägt von offenem Austausch, kritischen Diskussionen und vielen wertvollen Impulsen, die meine Forschung deutlich bereichert haben.

Mein besonderer Dank gilt Dr. Lars Robbel und Michael Schmitt seitens der CSL Innovation GmbH. Die Zusammenarbeit war stets sehr angenehm, vertrauensvoll und konstruktiv. Besonders geschätzt habe ich den inhaltlichen Freiraum innerhalb des Projekts sowie das Gefühl, in den zahlreichen fachlichen Diskussionen gehört und mit den eigenen Perspektiven ernst genommen zu werden.

Mein Dank gilt Matthias Lopinski, Alina Käthe Stein, Caty Wang und Franka Willems für ihre engagierte Unterstützung in den gemeinsamen Projekten. Neben ihrer sehr sorgfältigen und zuverlässigen Arbeit haben insbesondere ihre eigenen Ideen und Perspektiven wesentlich zum Gelingen der Arbeiten beigetragen.

Außerdem danke ich allen aktuellen und ehemaligen MABlern für die durchweg angenehme Arbeitsatmosphäre und das besondere kollegiale Miteinander. Den fachlichen wie persönlichen Austausch habe ich in dieser Zeit sehr geschätzt.

Abstract

The manufacturing of biopharmaceutical proteins differs fundamentally from the synthesis of small-molecule drugs, as the macromolecular structure and its associated higher-order structural levels play a decisive role in therapeutic efficacy and safety. Within this context, product quality is not a static property but rather a dynamic variable that continuously responds to the physicochemical influences of the processing environment. A critical indicator of process-induced stress is protein aggregation, which not only reduces yield but also poses significant immunogenic risks. Under real-world downstream processing (DSP) conditions, however, aggregate formation rarely occurs as an isolated phenomenon; rather, it is intimately linked to conformational changes, varying protein concentrations, and modifications of the surrounding buffer matrix. This molecular complexity presents a dual challenge for process analytics: on the one hand, structural alterations must be captured in real time, while on the other, the spectral signatures of the various components within the process stream overlap, thereby complicating direct mathematical isolation.

While classical offline reference methods provide precise quantitative data, their suitability for proactive process control is limited by an inherent time delay and the burden of sampling. Inline and online spectroscopic sensors, particularly Raman spectroscopy, offer high-resolution access to structural protein features directly within the process stream. Consequently, the technological challenge shifts from the physical measurement itself toward the mathematical deconvolution of complex, superimposed spectra. To address this task, different modeling approaches exist, which were systematically investigated and compared within the framework of this work. Linear chemometric methods, most notably partial least squares (PLS) regression, are considered a robust and widely established industry standard, provided that calibration and subsequent application conditions are tightly aligned. In contrast, non-linear, purely data-driven approaches such as convolutional neural networks (CNNs) promise a more flexible description of highly complex, non-linear relationships within spectral data; however, they typically require extensive datasets and reduce mathematical transparency due to their inherent complexity. A fundamentally complementary path is pursued by physically informed hard-modeling methods, which utilize explicit assumptions regarding peak shapes and signal composition to resolve overlapping bands, thereby offering direct mechanistic interpretability of spectral alterations. Against this background, the present work is dedicated to the targeted advancement of spectroscopic PAT (Process Analytical Technology) strategies by combining Raman measurements with these complementary modeling approaches under realistic DSP conditions.

The methodology of this thesis is structured into three consecutive phases that describe the transition from purely quantitative concentration monitoring to mechanistic structural elucidation, with chapter 3 forming the initial starting point. To minimize the high calibration effort typically associated with spectroscopic models, an automated sampling and measurement setup was developed. By coupling a liquid chromatography system, including an autosampler, with a Raman spectrometer, representative synthetic mixtures of lysozyme and varying phosphate and citrate buffer systems could be automatically generated and measured. The partial least squares models trained on this dataset were successfully transferred to a dynamic ultrafiltration/diafiltration (UF/DF) process, enabling the inline real-time monitoring of protein and buffer concentrations. Consequently, this chapter provided the technical proof of concept for a cohesive analytical strategy spanning from automated data acquisition to process-adjacent application.

Building upon these foundations, chapter 4 addresses the selective differentiation and quantification of native proteins and their aggregates within a chromatographic environment. Using bovine serum albumin (BSA) as a model system, controlled thermal stress studies and statistically designed mixing series were conducted to identify spectral markers for the aggregation state independent of the total protein concentration. These models were subsequently validated in dynamic bind-and-elute anion-exchange chromatography (AEX) runs. A direct comparison between linear PLS models and non-linear CNN architectures revealed that high predictive performance within controlled training data does not necessarily correlate with robustness against real-world process fluctuations. The chapter demonstrates that data-driven approaches enable reliable real-time prediction of monomer and aggregate profiles in preparative chromatography, while simultaneously underscoring the necessity of careful model selection regarding data availability and transferability.

To bridge the remaining gap between purely predictive models and physical process understanding, a complementary hard-modeling approach was established in chapter 5. The focus was placed on the amide I region of the Raman spectra, which is highly sensitive to secondary structure alterations but is concurrently severely disrupted by salt-induced background drifts during chromatographic elutions. A constrained pseudo-Voigt model was developed that utilizes an integrated background model to mathematically separate salt-dependent influences from protein-specific signal contributions. The resulting model-based mathematical descriptors, such as the shift in the center of gravity of the amide I band, were successfully validated against quantitative offline reference data for monomers and aggregates. This approach illustrates that hard modeling bridges a mechanistic gap, delivering physically interpretable structural features directly from the active process without relying exclusively on correlative patterns.

In its entirety, this dissertation thus links three complementary levels of modern process analytics: Automated, reproducible data generation (chapter 3), robust, data-driven prediction of critical quality attributes under realistic process conditions (chapter 4), and physically informed resolution of complex spectral signals for mechanistic structural interpretation (chapter 5). The work demonstrates that Raman-based process analytics for therapeutic proteins must go far beyond the mere optimization of predictive accuracy alone. Only the deliberate interplay of process knowledge, tailored preprocessing, and the conscious evaluation of the trade-off between data-driven flexibility and physical transparency enables deeper process understanding. Consequently, this work provides an essential contribution to the implementation of intelligent, interpretable, and robust PAT strategies in biopharmaceutical downstream processing.

Zusammenfassung

Die Herstellung biopharmazeutischer Proteine unterscheidet sich grundlegend von der Synthese niedermolekularer Wirkstoffe, da die makromolekulare Struktur und die damit verbundenen höheren Strukturebenen eine entscheidende Rolle für die therapeutische Wirksamkeit und Sicherheit spielen. In diesem Kontext ist die Produktqualität keine statische Eigenschaft, sondern eine dynamische Variable, die kontinuierlich auf die physikochemischen Einflüsse der Prozessumgebung reagiert. Ein kritischer Indikator für prozessbedingten Stress ist die Proteinaggregation, die nicht nur die Ausbeute mindert, sondern auch erhebliche immunogene Risiken birgt. Unter realen Bedingungen des Downstream-Processings (DSP) tritt die Aggregatbildung jedoch selten als isoliertes Phänomen auf; vielmehr ist sie eng mit Konformationsänderungen, variierenden Proteinkonzentrationen und Modifikationen der umgebenden Puffermatrix verknüpft. Diese molekulare Komplexität stellt eine doppelte Herausforderung für die Prozessanalytik dar: Einerseits müssen strukturelle Veränderungen in Echtzeit erfasst werden, andererseits überlagern sich die spektralen Signaturen der verschiedenen Komponenten im Prozessstrom, was die direkte mathematische Isolierung erschwert.

Während klassische Offline-Referenzmethoden präzise quantitative Daten liefern, ist ihre Eignung für eine proaktive Prozesskontrolle durch eine inhärente Zeitverzögerung und den Aufwand der Probenahme begrenzt. Inline- und Online-Spektroskopiesensoren, insbesondere die Raman-Spektroskopie, bieten einen hochauflösenden Zugang zu strukturellen Proteinmerkmalen direkt im Prozessstrom. Folglich verschiebt sich die technologische Herausforderung von der physikalischen Messung selbst hin zur mathematischen Dekonvolution komplexer, überlagerter Spektren. Zur Bewältigung dieser Aufgabe existieren verschiedene Modellierungsansätze, die im Rahmen dieser Arbeit systematisch untersucht und verglichen wurden. Lineare chemometrische Methoden, allen voran die Partial-Least-Squares-Regression (PLS), gelten als robuster und weit verbreiteter Industriestandard, sofern die Kalibrierungs- und nachfolgenden Anwendungsbedingungen eng aufeinander abgestimmt sind. Im Gegensatz dazu versprechen nichtlineare, rein datengetriebene Ansätze wie Convolutional Neural Networks (CNNs) eine flexiblere Beschreibung hochkomplexer, nichtlinearer Zusammenhänge innerhalb spektraler Daten; sie erfordern jedoch typischerweise umfangreiche Datensätze und verringern aufgrund ihrer inhärenten Komplexität die mathematische Transparenz. Ein grundlegend komplementärer Weg wird durch physikalisch informierte Hard-Modeling-Methoden beschritten, die explizite Annahmen über Peakformen und Signalzusammensetzungen nutzen, um überlagerte Banden aufzulösen, und somit eine direkte mechanistische Interpretierbarkeit spektraler Veränderungen bieten. Vor diesem Hintergrund widmet sich die vorliegende Arbeit der gezielten Weiterentwicklung

spektroskopischer PAT-Strategien (Process Analytical Technology) durch die Kombination von Raman-Messungen mit diesen komplementären Modellierungsansätzen unter realistischen DSP-Bedingungen.

Die Methodik dieser Arbeit gliedert sich in drei aufeinanderfolgende Phasen, die den Übergang von einer rein quantitativen Konzentrationsüberwachung zur mechanistischen Strukturaufklärung beschreiben, wobei Kapitel 3 den Ausgangspunkt bildet. Um den mit spektroskopischen Modellen typischerweise verbundenen hohen Kalibrieraufwand zu minimieren, wurde ein automatisiertes Probenahme- und Messsystem entwickelt. Durch die Kopplung eines Flüssigkeitschromatographie-Systems inklusive Autosampler mit einem Raman-Spektrometer konnten repräsentative synthetische Mischungen aus Lysozym und variierenden Phosphat- sowie Citratpuffersystemen automatisiert generiert und gemessen werden. Die auf diesem Datensatz trainierten Partial-Least-Squares-Modelle wurden erfolgreich auf einen dynamischen Ultrafiltrations-/Diafiltrationsprozess (UF/DF) übertragen, was die Inline-Echtzeitüberwachung von Protein- und Pufferkonzentrationen ermöglichte. Folglich lieferte dieses Kapitel den technischen Machbarkeitsnachweis für eine geschlossene Analysestrategie, die von der automatisierten Datenerfassung bis zur prozessnahen Anwendung reicht.

Auf diesen Grundlagen aufbauend befasst sich Kapitel 4 mit der selektiven Differenzierung und Quantifizierung von nativen Proteinen und deren Aggregaten in einer chromatographischen Umgebung. Unter Verwendung von Rinderserumalbumin (BSA) als Modellsystem wurden kontrollierte thermische Stressstudien und statistisch geplante Mischungsreihen durchgeführt, um vom Proteingehalt unabhängige spektrale Marker für den Aggregationszustand zu identifizieren. Diese Modelle wurden anschließend in dynamischen Bind-and-Elute-Anionenaustauschchromatographie-Läufen (AEX) validiert. Ein direkter Vergleich zwischen linearen PLS-Modellen und nichtlinearen CNN-Architekturen zeigte, dass eine hohe Vorhersageleistung innerhalb kontrollierter Trainingsdaten nicht zwangsläufig mit der Robustheit gegenüber realen Prozessfluktuationen korreliert. Das Kapitel demonstriert, dass datengetriebene Ansätze eine zuverlässige Echtzeit-Vorhersage von Monomer- und Aggregatprofilen in der präparativen Chromatographie ermöglichen, unterstreicht jedoch gleichzeitig die Notwendigkeit einer sorgfältigen Modellauswahl im Hinblick auf Datenverfügbarkeit und Transferierbarkeit.

Um die verbleibende Lücke zwischen rein prädiktiven Modellen und physikalischem Prozessverständnis zu schließen, wurde in Kapitel 5 ein komplementärer Hard-Modeling-Ansatz etabliert. Der Fokus wurde auf die Amid-I-Region der Raman-Spektren gelegt, die hochsensitiv gegenüber Sekundärstrukturveränderungen ist, jedoch gleichzeitig durch salzinduzierte Hintergrunddrifts während chromatographischer Elutionen stark gestört wird. Es wurde ein restringiertes Pseudo-Voigt-Modell entwickelt, das ein integriertes Hintergrundmodell nutzt, um salzabhängige Einflüsse mathematisch von proteinspezifischen Signalbeiträgen zu trennen. Die resultierenden modellbasierten mathematischen Deskriptoren, wie die Verschiebung des Linienschwerpunkts der Amid-I-Bande, wurden erfolgreich gegen quantitative Offline-Referenzdaten für Monomere und Aggregate validiert. Dieser Ansatz veranschaulicht, dass das Hard-Modeling eine mechanistische Lücke schließt, indem es physikalisch interpretierbare Strukturmerkmale direkt aus dem active Prozess liefert, ohne sich ausschließlich auf korrelative Muster zu verlassen.

In ihrer Gesamtheit verknüpft diese Dissertation somit drei komplementäre Ebenen der modernen Prozessanalytik: Eine automatisierte, reproduzierbare Datengenerierung (Kapitel 3), die robuste, datengetriebene Vorhersage kritischer Qualitätsattribute unter realistischen Prozessbedingungen (Kapitel 4) und die physikalisch informierte Auflösung komplexer Spektralsignale zur mechanisti-

schen Strukturinterpretation (Kapitel 5). Die Arbeit zeigt, dass Raman-basierte Prozessanalytik für therapeutische Proteine weit über die bloße Optimierung der Vorhersagegenauigkeit hinausgehen muss. Erst das gezielte Zusammenspiel aus Prozesswissen, maßgeschneiderter Vorverarbeitung und der bewussten Abwägung des Trade-offs zwischen datengetriebener Flexibilität und physikalischer Transparenz ermöglicht ein tieferes Prozessverständnis. Folglich leistet diese Arbeit einen wesentlichen Beitrag zur Implementierung intelligenter, interpretierbarer und robuster PAT-Strategien im biopharmazeutischen Downstream-Processing.

Contents

Acknowledgements **i**

Abstract **iii**

Zusammenfassung **v**

Contents **ix**

1 Introduction **1**
1.1 Therapeutic Proteins and Process-Related Quality Attributes 3
1.1.1 Therapeutic Proteins and Structural Complexity 3
1.2 Protein Aggregation in Biopharmaceutical Processing 3
1.2.1 Protein Structure and Aggregation Mechanisms 3
1.2.2 Aggregation under Process Conditions 3
1.3 Analytical Strategies for Aggregation Characterization 4
1.3.1 Established Offline Analytical Methods 4
1.3.2 Limitations of Conventional Approaches 4
1.4 Process Analytical Technology 4
1.4.1 Evolution of Analytical Methods toward Spectroscopic Techniques 4
1.4.2 Raman Spectroscopy 5
1.4.3 Chemometric Analysis of Raman Spectra 5
1.4.4 Data-Driven and Physically-Informed Spectral Modeling 5

2 Thesis Outline **13**
2.1 Research Proposal 13
2.2 Comprehensive Overview of Publications 16

3 A novel Raman-chromatography assembly for automated calibration and in-line monitoring in bioprocessing **19**
Jakob Heyer-Müller, Robin Schiemer, Matthias Lopinski, Caty Wang, Franka Willems, Lars Robbel, Michael Schmitt and Jürgen Hubbuch
3.1 Introduction 20

3.2 Material and Methods . . . 21
3.2.1 Experimental setup . . . 21
3.2.2 Raman spectroscopy . . . 22
3.2.3 Case study: Sample preparation . . . 22
3.2.4 Data preprocessing and chemometric modeling . . . 22
3.2.5 Ultrafiltration/Diafiltration . . . 23
3.3 Results and Discussion . . . 23
3.3.1 Automating Raman spectroscopy measurements by combining an ÄKTA chromatography system with a Raman spectrometer . . . 23
3.3.2 Calibrating chemometric models off-line by synthetic mixtures . . . 24
3.3.3 Applying chemometric models to quantify protein and buffer concentrations in ultrafiltration/diafiltration processes . . . 25
3.4 Conclusions . . . 26
3.5 Figures . . . 28
3.6 Tables . . . 35

4 Development of Raman spectroscopy and machine learning methods for protein aggregate quantification: Application to BSA in chromatographic processes. 39

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

4.1 Introduction . . . 40
4.2 Material and Methods . . . 42
4.2.1 Experimental . . . 42
4.2.2 Data analysis . . . 44
4.3 Results and Discussion . . . 47
4.3.1 Forced degradation studies . . . 47
4.3.2 Mixing studies . . . 48
4.3.3 Bind-elute anion-exchange chromatography experiments . . . 50
4.4 Conclusions . . . 53
4.5 Figures . . . 55
4.6 Tables . . . 60

5 Mechanistic Deconvolution of BSA Size Variants by Constrained Raman Pseudo-Voigt Hard Modeling during Anion-Exchange Chromatography 67

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

5.1 Introduction . . . 68
5.2 Materials and Methods . . . 69
5.2.1 Experimental . . . 69
5.2.2 Data analysis . . . 70
5.3 Results and Discussion . . . 72
5.3.1 Hard modeling framework . . . 72
5.3.2 Dataset Coverage and Hard-Model Execution . . . 76
5.3.3 Hard-Model Spectral Reconstruction Quality . . . 77
5.3.4 Resolved Protein Components and Peak Constraints . . . 77
5.3.5 Quantitative Validation of Hard-Model-Derived Spectral Features . . . 78

5.4 Conclusion 80
5.5 Tables 82
5.6 Figures 83

6 General Discussion and Conclusion **89**

References **95**

List of Figures **105**

List of Tables **109**

A3 Supplementary Material of Chapter 3 **113**
A3.1 Figures 114

A4 Supplementary Material of Chapter 4 **117**
A4.1 Figures 118
A4.2 Tables 126

A5 Authorship Contributions **129**

1

Introduction

Biopharmaceutical processes for therapeutic proteins are characterized by a strong coupling between molecular structure and process conditions [1, 2]. In contrast to small-molecule systems, relevant product properties cannot be described solely by composition, but depend on the preservation and evolution of higher-order structure during manufacturing [3, 4]. Consequently, product quality emerges as a dynamic property, shaped by the interplay of physicochemical conditions, unit operations, and molecular stability [5, 6].

Within this context, protein aggregation represents a central and practically relevant manifestation of structural variability [7]. In the literature, aggregation has been extensively studied with respect to its molecular mechanisms, classification, and impact on product quality [7]. Established frameworks describe aggregation as a consequence of structural perturbations and protein–protein interactions, occurring across a broad size range and influenced by process conditions [8]. However, aggregation is typically treated as an isolated attribute, whereas in real process environments it occurs simultaneously with changes in conformation, concentration, and matrix composition [9]. As a result, process-derived samples reflect superimposed contributions from multiple sources rather than a single well-defined variable [10]. The coupled nature of these effects remains only partially resolved [11].

From an analytical perspective, considerable effort has been devoted to the development of robust methods for aggregate quantification [12, 13]. In particular, separation-based offline techniques are well established and provide reliable reference measurements under controlled conditions [14, 15]. However, these approaches are inherently limited in dynamic process environments [6, 16]. Beyond the constraints of sampling and time delay, a key challenge lies in the interpretation of measurements obtained under process conditions [15]. In such systems, analytical signals are not uniquely linked to the target variable, but are influenced by concurrent variations in background composition, optical properties, and measurement conditions [13]. While these effects are recognized, their systematic

impact on measurement interpretation and model development remains insufficiently addressed [10, 17].

Spectroscopic techniques have been proposed as a means to overcome these limitations by enabling direct, process-near measurements [18, 19]. In the literature, their application has demonstrated the potential to capture molecular and structural information without physical separation [13] . At the same time, it is well established that spectroscopic data represent composite signals, integrating overlapping contributions from multiple components [10]. Consequently, rather than simplifying the analytical problem, spectroscopy shifts the challenge from measurement to data interpretation [10]. The question of how to reliably extract relevant information from such complex signals under process conditions remains open.

Chemometric modeling provides the primary toolset to address this challenge [10, 20]. A large body of work has focused on the application of multivariate regression methods to relate spectral data to target variables [13]. These approaches have demonstrated strong predictive performance in controlled systems, particularly when calibration and application conditions are closely matched. However, their robustness under varying process conditions is often limited [10, 21].

In particular, models may capture correlations that are specific to the calibration dataset but do not generalize when background composition, measurement conditions, or process dynamics change [10]. The extent to which such models reflect causal relationships rather than coincidental correlations remains a critical open question.

Different modeling strategies have been explored to address these limitations, yet each introduces its own set of challenges [22]. Data-driven approaches emphasize predictive accuracy but may lack interpretability and robustness. More flexible nonlinear models have been proposed to capture complex dependencies, but their applicability is often constrained by data availability and limited transparency. In contrast, physically informed approaches aim to improve interpretability by incorporating prior knowledge, yet require strong assumptions and may be sensitive to model specification. While these approaches are well described individually, their comparative behavior under realistic process conditions and their suitability for different types of variability remain insufficiently understood.

Overall, the literature provides a solid foundation for both analytical techniques and modeling approaches, but a gap remains in their integration under process-relevant conditions [16]. In particular, the combined effects of spectral interference, process-induced variability, and model selection on predictive performance and interpretability have not been systematically addressed. This is especially critical for applications involving dynamic processes, where data distributions evolve over time and multiple sources of variability interact.

The following sections introduce the theoretical foundations required to address these challenges. Starting from the structural characteristics of therapeutic proteins (Section 1.1.1) and the mechanisms of aggregation (Section 1.2.1), the discussion proceeds to established analytical approaches (Section 1.3.1) and their limitations (Section 1.3.2). Building on this basis, spectroscopic methods (Section 1.4.2) and chemometric modeling strategies (Section 1.4.3, Section 1.4.4) are presented, providing the framework for the methodological developments addressed in this work.

1.1 Therapeutic Proteins and Process-Related Quality Attributes

1.1.1 Therapeutic Proteins and Structural Complexity

Therapeutic proteins represent a major class of biopharmaceuticals with high clinical relevance, driven by their specificity and ability to target complex biological pathways [23]. In contrast to small-molecule drugs, these biomolecules exhibit a high degree of structural complexity, encompassing multiple hierarchical levels ranging from primary sequence to higher-order conformations [3]. The biological function of proteins critically depends on the preservation of their higher-order structure, as even subtle conformational changes can affect activity, stability, and immunogenicity [3–5]. During manufacturing, proteins inherently exhibit a certain degree of product-related heterogeneity, which may include variants arising from post-translational modifications or conformational flexibility [7, 8]. While some forms of heterogeneity are intrinsic and acceptable, undesired structural changes such as aggregation or misfolding can compromise product quality and safety [4, 8, 24]. Consequently, product quality must be understood as a function of both molecular structure and the applied manufacturing process, as process conditions can directly influence structural integrity and stability [1, 4, 5]. This interplay between structure and process leads to the formation of structural variants during biopharmaceutical production, which are classified as critical quality attributes and must be carefully monitored and controlled using appropriate analytical strategies [1, 14, 17].

1.2 Protein Aggregation in Biopharmaceutical Processing

1.2.1 Protein Structure and Aggregation Mechanisms

Protein aggregation is closely linked to the hierarchical structure of proteins , comprising primary, secondary, tertiary, and quaternary levels [3]. Disruption of higher-order structure can induce unfolding or conformational rearrangements, exposing hydrophobic regions that promote intermolecular association [4, 5]. Aggregation may thus proceed via interactions between native or non-native species, often driven by protein–protein interactions and hydrophobic effects [5, 25, 26]. Aggregation can be classified as reversible or irreversible, depending on the strength and nature of the underlying interactions [4]. Reversible aggregates are typically stabilized by weak non-covalent forces, whereas irreversible aggregation is associated with structural changes or covalent modifications [4]. The resulting species span a broad size range, from soluble oligomers to subvisible and visible particles [24]. Aggregation phenomena can, for instance, be induced under process-relevant stress conditions such as freeze–thaw operations. In biopharmaceutical processing, such size variants are classified as critical quality attributes due to their impact on product efficacy, stability, and immunogenicity [27]. While conventional methods primarily quantify aggregates based on size, aggregation is inherently a structural phenomenon. Spectroscopic techniques therefore provide complementary information by probing underlying conformational changes.

1.2.2 Aggregation under Process Conditions

Protein aggregation can be induced or promoted during manufacturing and downstream processing, as proteins are exposed to a range of physicochemical stresses [4, 8]. Key process variables include protein concentration, temperature, pH shifts, ionic strength, as well as interfaces, shear forces, and residence time, all of which influence protein stability and intermolecular interactions [4, 8]. In particular, chromatographic environments and formulation-relevant conditions may alter not only

aggregation propensity but also the spectral background due to changing buffer composition and co-eluting species [11]. As a result, process conditions affect not only aggregate content, but can simultaneously induce conformational changes and modify the surrounding matrix [9]. This coupling complicates the direct interpretation of process-derived samples and motivates the application of advanced analytical strategies.

1.3 Analytical Strategies for Aggregation Characterization

1.3.1 Established Offline Analytical Methods

Established analytical strategies for aggregate characterization rely on a set of orthogonal methods to capture different size regimes and structural properties [7, 8]. Size-exclusion chromatography (SEC) represents the standard approach for the quantification of soluble size variants, enabling separation of monomeric and high-molecular-weight species based on hydrodynamic size [8, 12, 28]. Complementary techniques, including light scattering, analytical ultracentrifugation, and particle-based methods, extend the accessible size range and provide additional structural or morphological information [8, 29, 30]. While SEC provides robust and quantitative information on monomer and aggregate fractions, it remains limited to soluble species and does not capture subvisible or particulate aggregates [12, 28]. Despite these limitations, offline analytical methods remain indispensable as reference techniques [1, 14]. In particular, they provide the quantitative ground truth required for the development and assessment of predictive models. Consequently, offline analytics form the basis for model development and benchmarking [1].

1.3.2 Limitations of Conventional Approaches

Conventional offline analytical methods require manual sampling and thus interrupt process continuity [1, 14]. In addition, a time delay between sampling and result generation is inherent to these approaches [16]. Sampling, dilution, or storage steps may further perturb fragile aggregation equilibria, potentially altering the original sample state [4, 8]. Moreover, size-based quantification alone does not provide direct information on molecular-level structural changes underlying aggregation [4, 8]. Consequently, conventional reference analytics are limited in their ability to enable real-time process understanding and feedback control [1, 14]. These limitations motivate the implementation of process analytical technologies for in-line or on-line monitoring [14, 16].

1.4 Process Analytical Technology

1.4.1 Evolution of Analytical Methods toward Spectroscopic Techniques

Process analytical technology (PAT) provides a framework for improving process understanding and ensuring consistent product quality through enhanced process monitoring and control [14]. Analytical measurements can be implemented at-line, on-line, or in-line, depending on the degree of integration into the process stream [14, 16]. Process-near analytics are particularly desirable, as they enable timely data acquisition, reduce product loss, and increase data density compared to conventional offline methods [14, 16]. This development reflects a shift from separation-based endpoint analysis toward direct spectroscopic interrogation of process samples. Spectroscopic techniques are especially advantageous in systems where multiple compositional and structural

changes occur simultaneously [11]. Among these, Raman spectroscopy represents a promising approach for non-destructive monitoring of aggregation-related structural variation [17, 31–33].

1.4.2 Raman Spectroscopy

Raman spectroscopy is a vibrational spectroscopic technique based on inelastic light scattering, providing molecular-specific information through characteristic spectral fingerprints [31]. In this process, energy is exchanged between incident photons and molecular vibrations, giving rise to Stokes and anti-Stokes scattering, corresponding to transitions to higher or lower vibrational energy states, respectively [31]. The resulting energy differences are expressed as Raman shifts, and characteristic regions of enhanced intensity are referred to as Raman bands [31]. For proteins, Raman spectra arise from vibrational modes of both the peptide backbone and amino acid side chains, enabling access to structural information across multiple levels [17]. In particular, amide bands reflect backbone conformations, while aromatic residues such as phenylalanine, tyrosine, and tryptophan contribute distinct and sensitive spectral features [11, 17]. Raman spectroscopy offers several advantages for the analysis of aqueous protein systems, including weak interference from water, sensitivity to structural changes, and compatibility with in-line or on-line process environments [11, 17]. However, the technique is limited by inherently weak signal intensities, susceptibility to fluorescence and background interference, and sensitivity to acquisition conditions [17, 31]. As a result, Raman spectra are information-rich but not directly interpretable, necessitating the application of chemometric and model-based analysis approaches [11].

1.4.3 Chemometric Analysis of Raman Spectra

Chemometrics refers to the application of statistical and mathematical methods to extract relevant information from complex spectral data [11, 20]. Raman spectra are affected by multiple sources of unwanted variation, including noise, baseline drift, fluorescence and background contributions, as well as intensity variability and matrix effects [11, 31]. To address these challenges, preprocessing steps such as spectral cropping, normalization, smoothing, derivative calculation, and baseline correction are routinely applied [11, 34]. The selection and parametrization of these preprocessing operations strongly influence both spectral interpretability and downstream model performance [11, 20].

Spectral information can be represented through various features, including peak intensities, intensity ratios, band positions, centers of mass, and derivative-based descriptors [33]. Multivariate regression methods provide a general framework for linking such spectral features to target variables, such as concentration or aggregate fraction [22, 33]. Among these, partial least squares (PLS) regression is widely used due to its ability to handle collinearity and high-dimensional predictor spaces [11, 35]. While chemometric models often yield strong predictive performance, their interpretability depends on the underlying model structure and may be limited for more abstract representations [22]. This trade-off motivates the comparison of purely data-driven and physically informed modeling approaches.

1.4.4 Data-Driven and Physically-Informed Spectral Modeling

Different modeling philosophies exist for extracting aggregation-relevant information from Raman spectra, reflecting distinct strategies for handling spectral complexity and underlying physicochemical

processes. These approaches differ in their underlying assumptions, interpretability, flexibility, and robustness, leading to trade-offs between predictive performance and physical insight.[22, 35, 36] Chemometrics provides a comprehensive set of statistical and mathematical tools to extract relevant information from complex spectral data [11, 20]. Raw spectroscopic measurements typically contain various sources of variability, including instrumental noise, baseline distortions, fluorescence and background contributions, as well as matrix effects, which can obscure the underlying chemical information [11]. To mitigate these effects, preprocessing strategies such as normalization, baseline correction, background correction, smoothing, and derivative filtering are routinely applied [37–39]. In addition, variable selection methods can be employed to reduce dimensionality and focus on the most informative spectral regions [40]. However, the choice, parametrization, and sequence of preprocessing steps must be carefully considered to avoid the loss of relevant information and to ensure robust model performance [19]. Preprocessed data are commonly further standardized by centering and scaling prior to multivariate modeling.

1.4.4.1 Partial Least Squares Regression

Among statistical methods for quantitative prediction, partial least squares (PLS) regression represents the most widely applied approach in spectroscopic analysis [20, 22]. PLS extends classical linear regression by incorporating multiple correlated predictors, while simultaneously reducing dimensionality through latent variable decomposition [22, 41]. In contrast to principal component regression, which focuses on explaining variance in the predictor space, PLS identifies latent variables that maximize the covariance between predictor and response variables [22, 42].

In biopharmaceutical applications, spectral data are typically organized in a predictor matrix $\mathbf{X}$ containing spectral intensities, while the corresponding concentrations are stored in the response matrix $\mathbf{Y}$. The relationship between both is commonly described by a linear regression model of the form

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \tag{1.1}$$

where $\mathbf{B}$ denotes the regression coefficient matrix and $\mathbf{E}$ the residual error. Different regression approaches primarily differ in how $\mathbf{B}$ is estimated. While multiple linear regression directly minimizes the residual error, principal component regression operates on variance-maximizing components, and PLS extracts latent variables that maximize covariance between $\mathbf{X}$ and $\mathbf{Y}$. [22, 41]

$$\mathbf{X} = \mathbf{TP}^\top + \mathbf{E}_X \tag{1.2}$$

$$\mathbf{Y} = \mathbf{UQ}^\top + \mathbf{E}_Y \tag{1.3}$$

Here, $\mathbf{T}$ and $\mathbf{U}$ denote the score matrices representing the projections of $\mathbf{X}$ and $\mathbf{Y}$ onto latent variables, $\mathbf{P}$ and $\mathbf{Q}$ are the corresponding loading matrices describing the contribution of the original variables, and $\mathbf{E}_X$ and $\mathbf{E}_Y$ represent the residual matrices capturing unexplained variance.

PLS achieves dimensionality reduction by projecting the predictor space onto latent variables, which are constructed to retain the most relevant information for prediction. This enables robust handling of multicollinearity and high-dimensional data, which are characteristic of spectroscopic measurements. However, as a linear method, PLS may be limited in capturing non-linear relationships present in

complex spectral data, motivating the use of motivating the use of more flexible modeling approaches when required. [20, 41]

1.4.4.2 Convolutional Neural Networks

Convolutional neural network (CNN)s represent a class of nonlinear, data-driven models capable of learning hierarchical feature representations directly from raw input data [43, 44]. In contrast to classical chemometric approaches, CNNs do not rely on predefined feature extraction, but instead automatically identify relevant patterns through a sequence of convolutional operations and nonlinear transformations [43, 44]. This hierarchical learning enables the extraction of increasingly abstract spectral features across multiple representation levels [44, 45]. In the context of spectroscopic data, one-dimensional CNN architectures are typically employed, where convolutional filters are applied along the spectral axis [44, 45]. For an input spectrum $\mathbf{x}$, the convolutional operation can be expressed as

$$\mathbf{h}_k = \sigma(\mathbf{x} * \mathbf{w}_k + b_k) \tag{1.4}$$

where $*$ denotes the convolution operator, $\mathbf{w}_k$ represents the k-th filter kernel, b_k is a bias term, $\sigma(\cdot)$ is a nonlinear activation function, and $\mathbf{h}_k$ denotes the resulting feature map [43]. By stacking multiple convolutional layers, CNNs construct increasingly abstract feature representations, enabling nonlinear representation learning and the extraction of complex relationships within spectral data [44, 45]. CNNs are particularly attractive for Raman spectroscopy, as convolutional filters inherently capture local spectral patterns, such as peak shapes and relative intensity distributions [44, 46]. This local feature extraction provides robustness against small spectral shifts and peak broadening effects, which frequently occur under varying process conditions [44, 47]. In addition, the nonlinear structure of CNNs allows modeling of complex interactions between spectral features that cannot be captured by linear methods [44, 45]. These properties can lead to improved predictive performance, especially in cases where the relationship between spectral data and target variables is highly nonlinear or influenced by overlapping spectral contributions [44, 46]. However, the application of CNNs is associated with several limitations. In particular, CNNs typically require larger datasets for robust training, are more prone to overfitting, and exhibit reduced interpretability compared to linear chemometric models [43, 47]. Furthermore, their transferability across different instruments or process conditions may be limited without additional strategies such as domain adaptation or spectral standardization [44, 47].

Consequently, CNNs should be regarded as complementary to classical approaches such as PLS regression, rather than as direct replacements [35, 44]. While CNNs offer increased flexibility and predictive power, linear models retain advantages in terms of interpretability, robustness in small-data regimes, and ease of deployment in process analytical technology frameworks [20, 35, 47].

1.4.4.3 Physically-Informed Hard Modeling of Raman Spectra

In contrast to purely data-driven approaches, physically-informed hard modeling represents a parameterized spectral decomposition strategy based on explicit assumptions about peak shapes and component contributions [17, 20]. In this framework, measured spectra are described as superpositions of individual spectral components with predefined functional forms, enabling a direct link between spectral features and underlying physicochemical properties [17, 48]. This approach is particularly relevant for Raman spectroscopy of proteins, where the amide I region (approximately 1600–1700 cm^{-1}) contains overlapping contributions from different secondary structural elements

as well as background signals [49, 50]. The superposition of these contributions complicates direct interpretation and motivates the use of constrained decomposition strategies to resolve individual spectral components [11, 49].

The spectral signal $\mathbf{x}(\tilde{\nu})$ can be expressed as a sum of K parameterized peak functions and a background term

$$\mathbf{x}(\tilde{\nu}) = \sum_{k=1}^{K} a_k \, f_k(\tilde{\nu}; \theta_k) + b(\tilde{\nu}) + \varepsilon(\tilde{\nu}), \tag{1.5}$$

where a_k denotes the amplitude, f_k the peak shape function parameterized by θ_k (e.g., position, width, and shape), $b(\tilde{\nu})$ the background contribution, and $\varepsilon(\tilde{\nu})$ the residual error [17].

In Raman applications, pseudo-Voigt functions are commonly employed to describe individual bands, combining Gaussian and Lorentzian contributions to account for both instrumental and intrinsic broadening effects

$$f_k(\tilde{\nu}) = \eta_k \, L(\tilde{\nu}) + (1 - \eta_k) \, G(\tilde{\nu}), \tag{1.6}$$

where η_k denotes the mixing parameter between Lorentzian (L) and Gaussian (G) line shapes [51].

To ensure physically meaningful solutions, constraints on peak positions, widths, and relative contributions are typically introduced based on prior knowledge of protein structure and spectroscopic behavior [49, 50]. This constrained optimization enables the decomposition of highly overlapping spectral regions, such as the amide I band, into contributions associated with distinct structural states [17, 37]. In contrast to latent-variable methods, hard modeling explicitly aims at interpretable descriptors, such as band positions, widths, and relative intensities, which can be directly related to molecular structure [20, 37]. This provides mechanistic insight into structural changes and allows a more transparent interpretation of aggregation-related spectral variations [50]. The main advantages of this approach include its mechanistic interpretability, the explicit treatment of overlapping spectral contributions, and its direct structural relevance [17]. However, these benefits come at the cost of increased model complexity, sensitivity to initial parameter estimates, and dependence on the validity of underlying model assumptions [11, 20]. In addition, the fitting procedure may become computationally demanding and requires careful regularization to ensure stable solutions [20].

Overall, hard modeling represents a complementary strategy to data-driven approaches. While PLS regression provides linear latent-variable prediction and CNNs enable nonlinear feature learning, hard modeling performs an explicit, physically constrained decomposition of spectral signals, thereby bridging predictive modeling and mechanistic interpretation [35, 44].

Chapter References

[1] A. S. Rathore and H. Winkle. „Quality by design for biopharmaceuticals“. In: *Nature Biotechnology* 27.1 (2009), pp. 26–34. DOI: 10.1038/nbt0109-26.

[2] P. Diederich, M. Hoffmann, and J. Hubbuch. „High-throughput process development of purification alternatives for the protein avidin“. In: *Biotechnology Progress* 31.4 (2015), pp. 957–973. DOI: 10.1002/btpr.2104.

[3] C. M. Dobson. „Protein folding and misfolding“. In: *Nature* 426 (2003), pp. 884–890. DOI: 10.1038/nature02261.

[4] W. Wang. „Protein aggregation and its inhibition in biopharmaceutics“. In: *International Journal of Pharmaceutics* 289 (2005), pp. 1–30. DOI: 10.1016/j.ijpharm.2004.11.014.

[5] E. Y. Chi, S. Krishnan, T. W. Randolph, and J. F. Carpenter. „Physical stability of proteins in aqueous solution: mechanism and driving forces in nonnative protein aggregation“. In: *Pharmaceutical Research* 20 (2003), pp. 1325–1336. DOI: https://doi.org/10.1023/A:1025771421906.

[6] A. Sanden, S. Haas, and J. Hubbuch. „Modifying an ÄKTApurifier system for the automated acquisition of samples for kinetic modeling of batch reactions“. In: *SLAS Technology* (2019). DOI: 10.1177/2472630319891976.

[7] M. C. Manning, D. K. Chou, B. M. Murphy, R. W. Payne, and D. S. Katayama. „Stability of protein pharmaceuticals: An update“. In: *Pharmaceutical Research* 27.4 (2010), pp. 544–575. DOI: 10.1007/s11095-009-0045-6.

[8] H. C. Mahler, W. Friess, U. Grauschopf, and S. Kiese. „Protein aggregation: pathways, induction factors and analysis“. In: *Journal of Pharmaceutical Sciences* 98 (2009), pp. 2909–2934. DOI: 10.1002/jps.21566.

[9] C. Zhou, W. Qi, E. N. Lewis, and J. F. Carpenter. „Concomitant Raman spectroscopy and dynamic light scattering for characterization of therapeutic proteins at high concentrations“. In: *Analytical Biochemistry* 472 (2015), pp. 7–20. DOI: 10.1016/j.ab.2014.11.016.

[10] M. Rüdt, S. Andris, R. Schiemer, and J. Hubbuch. „Factorization of preparative protein chromatograms with hard-constraint multivariate curve resolution and second-derivative pretreatment“. In: *Journal of Chromatography A* 1585.5 (2019), pp. 152–160. DOI: 10.1016/j.chroma.2018.11.065.

[11] Å. Rinnan, F. van den Berg, and S. B. Engelsen. „Review of the most common pre-processing techniques for near-infrared spectra“. In: *TrAC Trends in Analytical Chemistry* 28 (2009), pp. 1201–1222. DOI: 10.1016/j.trac.2009.07.007.

[12] „A review size-exclusion chromatography for the analysis of protein biotherapeutics and their aggregates“. In: *Journal of Liquid Chromatography Related Technologies* 35.20 (2012), pp. 2923–2950. DOI: 10.1080/10826076.2012.743724.

[13] S. K. Hansen, E. Skibsted, A. Staby, and J. Hubbuch. „A label-free methodology for selective protein quantification by means of absorption measurements“. In: *Biotechnology and Bioengineering* 108.11 (2011), pp. 2661–2669. DOI: 10.1002/bit.23229.

[14] Food and Drug Administration. *PAT — A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance*. 2004.

[15] M. Rüdt, N. Brestrich, L. Rolinger, and J. Hubbuch. „Real-time monitoring and control of the load phase of a protein A capture step“. In: *Biotechnology and Bioengineering* 114.2 (2017), pp. 368–373. DOI: 10.1002/bit.26078.

[16] J. Glassey et al. „Process analytical technology (PAT) for biopharmaceuticals“. In: *Biotechnology Journal* 6.4 (2011), pp. 369–377. DOI: 10.1002/biot.201000356.

[17] T. D. Beer et al. „Raman spectroscopy as a process analytical technology“. In: *TrAC Trends in Analytical Chemistry* 30 (2011), pp. 1073–1086. DOI: 10.1016/j.trac.2011.06.012.

[18] K. A. Esmonde-White, M. Cuellar, C. Uerpmann, B. Lenain, and I. R. Lewis. „Raman spectroscopy as a process analytical technology for pharmaceutical manufacturing and bioprocessing“. In: *Analytical and Bioanalytical Chemistry* 409.3 (2017), pp. 637–649. DOI: 10.1007/s00216-016-9824-1.

[19] L. Rolinger et al. „Multi-attribute PAT for UF/DF of proteins—monitoring concentration and buffer exchange by Raman spectroscopy“. In: *Analytical and Bioanalytical Chemistry* 412 (2020), pp. 2123–2134. DOI: 10.1007/s00216-019-02318-8.

[20] T. Næs, T. Isaksson, T. Fearn, and T. Davies. *A User-Friendly Guide to Multivariate Calibration and Classification*. NIR Publications, 2002. DOI: 10.1255/978-1-906715-25-0.

[21] L. e. a. Rolinger. „Multivariate data analysis in downstream processing“. In: *Analytical and Bioanalytical Chemistry* 412 (2020), pp. 2045–2059. DOI: 10.1007/s00216-020-02407-z.

[22] H. Martens and T. Næs. *Multivariate Calibration*. Wiley, 1989.

[23] G. Walsh. „Biopharmaceutical benchmarks 2018“. In: *Nature Biotechnology* 36 (2018), pp. 1136–1145. DOI: 10.1038/nbt.4305.

[24] J. F. Carpenter, T. W. Randolph, W. Jiskoot, et al. „Overlooking subvisible particles in therapeutic protein products“. In: *Journal of Pharmaceutical Sciences* 98 (2009), pp. 1201–1205. DOI: 10.1002/jps.21530.

[25] N. Räkel, K. Schleining, F. Dismer, and J. Hubbuch. „Prediction of protein retention in ion-exchange chromatography based on protein–protein interactions“. In: *Journal of Chromatography A* 1277 (2013), pp. 60–68. DOI: 10.1016/j.chroma.2013.03.077.

[26] P. Vormittag, T. Klamp, and J. Hubbuch. „Optimization of protein chromatography processes“. In: *Frontiers in Bioengineering and Biotechnology* 8 (2020), p. 402. DOI: 10.3389/fbioe.2020.0088.

[27] A. S. Rosenberg. „Effects of protein aggregates: An immunologic perspective“. In: *The AAPS Journal* 8.3 (2006), p. 59. DOI: 10.1208/aapsj080359.

[28] K. K. Arthur, J. P. Gabrielson, B. S. Kendrick, and M. R. Stoner. „Detection of protein aggregates by sedimentation velocity analytical ultracentrifugation (SV-AUC): sources of variability and their relative importance“. In: *Journal of Pharmaceutical Sciences* 98.10 (2009), pp. 3522–3539. DOI: 10.1002/jps.21654.

[29] J. Stetefeld, S. A. McKenna, and T. R. Patel. „Dynamic light scattering: a practical guide and applications in biomedical sciences“. In: *Biophysical Reviews* 8.4 (2016), pp. 409–427. DOI: 10.1007/s12551-016-0218-6.

[30] J. Lebowitz, M. S. Lewis, and P. Schuck. „Modern analytical ultracentrifugation in protein science: a tutorial review“. In: *Protein Science* 11.9 (2002), pp. 2067–2079. DOI: 10.1110/ps.0207702.

[31] R. L. McCreery. „Raman Spectroscopy for Chemical Analysis“. In: *Wiley* 1 (2000), pp. 1–420. DOI: 10.1002/0471238961.

[32] A. Dietrich, R. Schiemer, J. Kurmann, S. Zhang, and J. Hubbuch. „Raman-based PAT for VLP precipitation: systematic data diversification and preprocessing pipeline identification“. In: *Frontiers in Bioengineering and Biotechnology* 12 (2024). DOI: 10.3389/fbioe.2024.1399938.

[33] J. Heyer-Müller, R. Schiemer, L. Robbel, M. Schmitt, and J. Hubbuch. „Development of Raman spectroscopy and machine learning methods for protein aggregate quantification: Application to BSA in chromatographic processes“. en. In: *Biotechnology and Bioengineering* 123.5 (2026), e70163. DOI: 10.1002/bit.70163.

[34] A. Savitzky and M. J. E. Golay. „Smoothing and Differentiation of Data by Simplified Least Squares Procedures“. In: *Analytical Chemistry* 36.8 (1964), pp. 1627–1639. DOI: 10.1021/ac60214a047.

[35] S. Wold, M. Sjöström, and L. Eriksson. „PLS-regression: a basic tool of chemometrics“. In: *Chemometrics and Intelligent Laboratory Systems* 58.2 (2001), pp. 109–130. DOI: 10.1016/S0169-7439(01)00155-1.

[36] K. A. Bakeev, ed. *Process Analytical Technology: Spectroscopic Tools and Implementation Strategies for the Chemical and Pharmaceutical Industries*. Wiley, 2010.

[37] Å. Rinnan, F. van den Berg, and S. B. Engelsen. „Review of the most common pre-processing techniques for near-infrared spectra“. In: *TrAC Trends in Analytical Chemistry* 28.10 (2009), pp. 1201–1222. DOI: 10.1016/j.trac.2009.07.007.

[38] A. Savitzky and M. J. E. Golay. „Smoothing and Differentiation of Data by Simplified Least Squares Procedures“. In: *Journal of Analytical Chemistry* 36.8 (1964), pp. 1627–1639.

[39] R. Gautam, S. S. G. Vanga, F. Ariese, and S. Umapathy. „Review of multidimensional data processing approaches for Raman and infrared spectroscopy“. In: *EPJ Techniques and Instrumentation* 2.1 (2015), p. 8. DOI: 10.1140/epjti/s40485-015-0018-6.

[40] T. Mehmood, K. H. Liland, L. Snipen, and S. Sæbø. „A review of variable selection methods in Partial Least Squares Regression“. In: *Chemometrics and Intelligent Laboratory Systems* 118 (2012), pp. 62–69. DOI: 10.1016/j.chemolab.2012.07.010.

[41] S. Wold, M. Sjöström, and L. Eriksson. „PLS-regression: a basic tool of chemometrics“. In: *Chemometrics and Intelligent Laboratory Systems* 58.2 (2001), pp. 109–130. DOI: 10.1016/S0169-7439(01)00155-1.

[42] S. Wold, M. Sjöström, and L. Eriksson. „PLS-regression: a basic tool of chemometrics“. In: *Chemometrics and Intelligent Laboratory Systems* 58.2 (2001), pp. 109–130. DOI: 10.1016/S0169-7439(01)00155-1.

[43] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[44] F. Lussier, V. Thibault, B. Charron, G. Q. Wallace, and J.-F. Masson. „Deep learning and artificial intelligence methods for Raman and surface-enhanced Raman scattering“. In: *Trends in Analytical Chemistry* 124 (2020), p. 115796. DOI: 10.1016/j.trac.2019.115796.

[45] Y. LeCun, Y. Bengio, and G. Hinton. „Deep learning“. In: *Nature* 521.7553 (2015), pp. 436–444. DOI: 10.1038/nature14539.

[46] C.-S. Ho et al. „Rapid identification of pathogenic bacteria using Raman spectroscopy and deep learning“. In: *Nature Communications* 10.1 (2019), p. 4927. DOI: 10.1038/s41467-019-12898-9.

[47] J. Acquarelli et al. „Convolutional neural networks for vibrational spectroscopic data analysis“. In: *Analytica Chimica Acta* 954 (2017), pp. 22–31. DOI: 10.1016/j.aca.2016.12.010.

[48] R. Tauler. „Multivariate curve resolution applied to second order data“. In: *Chemometrics and Intelligent Laboratory Systems* 30.1 (1995), pp. 133–146. DOI: 10.1016/0169-7439(95)00047-X.

[49] A. Barth. „Infrared spectroscopy of proteins“. In: *Biochimica et Biophysica Acta (BBA) - Bioenergetics* 1767.9 (2007), pp. 1073–1101. DOI: 10.1016/j.bbabio.2007.06.004.

[50] Z. Movasaghi, S. Rehman, and I. U. Rehman. „Raman spectroscopy of biological tissues“. In: *Applied Spectroscopy Reviews* 42.5 (2007), pp. 493–541. DOI: 10.1080/05704920701551530.

[51] T. Ida, M. Ando, and H. Toraya. „Extended pseudo-Voigt function for approximating the Voigt profile“. In: *Journal of Applied Crystallography* 33 (2000), pp. 1311–1316. DOI: 10.1107/S0021889800010219.

2

Thesis Outline

2.1 Research Proposal

Biopharmaceuticals have transformed modern medicine by enabling the targeted treatment of complex diseases through highly specific and structurally sophisticated molecules. In contrast to conventional small-molecule drugs, therapeutic proteins derive their functionality not merely from chemical composition, but from the preservation of higher-order structure across production, purification, and formulation. Product quality is therefore inseparably linked to process conditions, and process understanding becomes a prerequisite for ensuring consistent efficacy, stability, and safety. Within this context, protein aggregation represents one of the most relevant manifestations of structural variability. Aggregates may arise throughout biopharmaceutical processing and are widely recognized as critical quality attributes (CQAs) due to their impact on product performance and immunogenicity. However, in real process environments, aggregation does not occur in isolation. Instead, it develops alongside changes in protein conformation, concentration, and matrix composition, which complicates both analytical measurement and the interpretation of process-derived data. Conventional offline methods remain indispensable as quantitative reference techniques, yet their reliance on sampling and delayed analysis limits their value for real-time process understanding in dynamic downstream environments. These limitations motivate the development of process-near analytical strategies that can capture molecularly meaningful information directly under process conditions. Spectroscopic techniques offer a promising route to address this need because they enable direct, non-destructive interrogation of process samples without prior physical separation. Among them, Raman spectroscopy is particularly attractive for protein systems, as it provides access to structurally relevant vibrational information while being compatible with aqueous environments and in-line or on-line process integration. At the same time, Raman spectra are inherently composite signals. They integrate overlapping contributions from the protein itself, the surrounding matrix, and measurement-related background effects. As a result, the analytical challenge shifts from signal acquisition to signal interpretation.

Chemometric and model-based approaches provide the principal tools for this task, but important limitations remain. Linear multivariate methods such as partial least squares regression are well established and often robust in controlled settings, yet they may fail when relevant nonlinearities or process-induced shifts are present. More flexible nonlinear models such as convolutional neural networks can potentially capture more complex dependencies, but they generally require larger datasets and often lack transparency. Physically informed hard-modeling approaches promise a more interpretable decomposition of overlapping spectral features, but they depend on explicit assumptions and careful model specification. Although these different strategies are well described individually, their comparative behavior under realistic downstream processing conditions and their suitability for extracting aggregation-relevant information remain insufficiently understood. Against this background, the objective of this thesis is to advance process-relevant analytical strategies for therapeutic protein systems by integrating Raman spectroscopy with complementary modeling approaches under realistic downstream processing conditions. Particular emphasis is placed on understanding how process-induced variability, spectral background contributions, and model choice influence predictive performance and interpretability. Rather than treating Raman spectroscopy merely as an alternative process sensor, the present work aims to establish a coherent framework for extracting structurally meaningful and process-relevant information from complex Raman data. To this end, the thesis combines the development of an automated Raman-based measurement platform with the evaluation of both data-driven and physically informed approaches for quantifying and interpreting protein size variants during downstream processing.

A first study (Chapter 3) will address the practical challenge of generating representative Raman data for model calibration and validation under conditions that resemble the intended process application. Although Raman spectroscopy is widely discussed as a promising process analytical technology, calibration remains labor-intensive because reliable chemometric models require datasets that are both representative and reproducible. To address this limitation, a novel assembly combining a commercially available chromatography system, an autosampler, and a Raman spectrometer will be established for automated spectral acquisition. Using synthetic mixtures of phosphate buffer, citrate buffer, and lysozyme as a model protein, representative calibration and validation datasets will be generated and used to construct chemometric models based on partial least squares regression. These models will then be transferred to in-line ultra-/diafiltration experiments to evaluate whether off-line calibrated Raman models can provide reliable real-time monitoring of protein and buffer concentrations under process conditions. This study is expected to establish an analytical and technical foundation for subsequent work by demonstrating that automated data generation and process-near Raman implementation can be integrated into a single, practically deployable concept for bioprocess monitoring. Building on this foundation, the second study (Chapter 4) will focus on the selective quantification of protein aggregates in chromatographic environments. While Raman spectroscopy has substantial potential for monitoring aggregation-related structural variation, one of the major unresolved challenges lies in distinguishing aggregate-sensitive information from the influence of total protein concentration and changing matrix conditions. To address this problem, bovine serum albumin will be employed as a model system in controlled forced-degradation experiments and statistically designed mixing studies. These experiments will be used to identify and validate aggregation-sensitive spectral markers under conditions in which aggregate content and total protein concentration can be decoupled. The resulting markers will then be assessed during bind-elute anion-exchange chromatography experiments in order to test their applicability under realistic process conditions.

In parallel, multiple chemometric models will be established and compared, including linear latent-variable approaches based on PLS regression and nonlinear data-driven approaches based on convolutional neural networks. This comparative design will make it possible to assess not only predictive performance, but also the broader balance between robustness, data requirements, and interpretability. This study is expected to demonstrate to what extent Raman spectroscopy, combined with systematically developed machine learning models, can enable reliable real-time prediction of monomer and aggregate concentrations during chromatographic processing. At the same time, it will clarify how strongly model utility depends on whether the selected modeling strategy remains meaningful under changing process conditions and overlapping spectral variability. Whereas the second study will emphasize predictive quantification through data-driven modeling, the third study (Chapter 5) will address the complementary question of mechanistic interpretability. In many process-analytical settings, highly predictive models provide limited physical insight into the structural basis of the measured signal, particularly when protein-derived Raman features overlap with gradient-induced background effects. This problem is especially pronounced in the amide I region, which contains rich structural information but is also susceptible to matrix-driven drift during chromatography. To address this limitation, a constrained pseudo-Voigt hard-modeling framework will be developed for the mechanistic deconvolution of in-line Raman spectra acquired during anion-exchange chromatography of bovine serum albumin monomers and aggregates. A dedicated background model will be introduced to capture salt-induced spectral drift and thereby separate process-related matrix contributions from protein-specific structural features. The constrained hard model will subsequently be applied to chromatographic datasets in order to extract physically interpretable components and aggregation-sensitive descriptors, in particular the center of mass of the amide I region. These descriptors will then be quantitatively evaluated against chromatographic reference analytics for the prediction of monomeric and high-molecular-weight species. This study is expected to show that physically informed hard modeling can bridge predictive analysis and structural interpretation by explicitly resolving protein-related spectral contributions under process-relevant conditions. Thereby, it will establish a mechanistically transparent and comparatively calibration-light alternative or complement to purely data-driven models.

Taken together, the proposed studies address three closely linked methodological levels: analytical implementation, predictive modeling, and mechanistic interpretation. The first study establishes the technical basis for reproducible Raman-based process analytics. The second study extends this basis toward the selective prediction of protein size variants under chromatographic conditions and directly compares different data-driven modeling strategies. The third study then addresses the limits of purely predictive models by introducing a physically informed framework for structural interpretation of overlapping Raman signals. In combination, these studies are expected to provide a coherent contribution to the development of Raman-based process analytical strategies for therapeutic proteins. More broadly, the thesis aims to deepen the understanding of how spectroscopic sensor data, chemometric modeling, and mechanistic interpretation can be integrated under realistic downstream processing conditions. In this way, the work is intended to support future process analytical technology concepts that are not only predictive, but also interpretable, structurally meaningful, and robust in the presence of evolving process variability.

2.2 Comprehensive Overview of Publications

This section provides an overview of the three research articles included in this thesis. The following chapters contain the original publications, with minor formatting and layout adjustments to ensure consistency throughout the dissertation. As two of the articles resulted from close collaborative work, shared first authorship is indicated where applicable.

The included publications address a common methodological objective: the development of Raman-based process analytics for therapeutic protein downstream processing under realistic process conditions. In this context, Chapter 3 establishes an automated Raman-based platform for systematic spectral acquisition, calibration, and process-near monitoring. Building on this analytical foundation, Chapter 4 develops data-driven methods for the selective quantification of protein aggregates in chromatographic processes using BSA as a model system. Chapter 5 extends this framework toward mechanistic interpretation by introducing a physically informed hard-modeling approach for the deconvolution of overlapping Raman signals acquired during anion-exchange chromatography. Taken together, the three publications represent a coherent progression from analytical implementation to predictive modeling and finally to mechanistic spectral interpretation.

Chapter 3

A novel Raman-chromatography assembly for automated calibration and in-line monitoring in bioprocessing. 19

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

published in *Engineering in Life Sciences*, Volume 25(9), 2025
https://doi.org/10.1002%2Felsc.70044

Chapter 3 presents a novel assembly combining an ÄKTA chromatography system, an autosampler, and a Raman spectrometer for automated spectral acquisition and straightforward chemometric model calibration. Synthetic mixtures of phosphate buffer, citrate buffer, and lysozyme are prepared using two statistical designs and measured off-line to calibrate and validate partial least squares (PLS) regression models. By applying the calibrated models to in-line Raman spectra from ultra-/diafiltration (UF/DF) experiments, the monitoring of protein and buffer concentrations in real time is demonstrated. Further, the study evaluates the transferability of off-line calibrated models to in-line process data and assesses the suitability of the automated setup for bioprocess monitoring.

Chapter 4

Development of Raman spectroscopy and machine learning methods for protein aggregate quantification: Application to BSA in chromatographic processes. 39

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

published in *Biotechnology and Bioengineering*, 2026
https://doi.org/10.1002/bit.70163

Chapter 4 investigates the development of Raman spectroscopy- and machine learning-based methods for the selective quantification of protein aggregates in chromatographic processes using bovine serum albumin (BSA) as a model protein. Controlled forced degradation and statistically designed mixing studies were used to identify and validate aggregation-sensitive spectral markers while decoupling aggregate content from total protein concentration. These markers were subsequently evaluated during in-line anion-exchange chromatography (AEX) experiments to assess their applicability under process-relevant conditions. In addition, multiple chemometric models, including PLS regression and convolutional neural networks (CNNs), were established and compared for real-time prediction of monomer and aggregate concentrations, demonstrating the potential of Raman spectroscopy combined with advanced data-driven modeling for monitoring protein size variants in biopharmaceutical downstream processing.

Chapter 5

Mechanistic Deconvolution of BSA Size Variants by Constrained Raman Pseudo-Voigt Hard Modeling during Anion-Exchange Chromatography 67

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

published in *Biotechnology Progress*, 2026
https://doi.org/10.1002/btpr.88534

Chapter 5 establishes a constrained pseudo-Voigt hard-modeling framework for the mechanistic deconvolution of in-line Raman spectra acquired during AEX separation of BSA monomers and aggregates. A dedicated background model was introduced to capture salt-induced spectral drift in the Amide I region and thereby separate gradient-driven matrix effects from protein-specific structural contributions. The constrained hard model was subsequently applied to fraction-resolved chromatographic datasets to extract physically interpretable spectral components and aggregation-sensitive markers, in particular the Amide I center of mass. In addition, the model-derived features were quantitatively evaluated against ultrahigh performance - size exclusion chromatography (UHP-SEC) reference analytics for the prediction of monomer and high-molecular-weight species, demonstrating the potential of mechanistic Raman hard modeling as an interpretable and calibration-light alternative for monitoring protein size variants during biopharmaceutical downstream processing.

3

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

Jakob Heyer-Müller[1,*], Robin Schiemer[1,*], Matthias Lopinski[1], Caty Wang[1], Franka Willems[1], Lars Robbel[2], Michael Schmitt[2] and Jürgen Hubbuch[1]

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

[2] CSL Innovation GmbH, Marburg, 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 (R2 and Q2), 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.

3.1 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 [52], filtration [53], or precipitation [32] 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 [32, 54], artificial spiking with the target analytes or synthetic mixtures [55–57]. The use of synthetic mixtures, when available, facilitates an off-line model calibration [58] or can be used to supplement process-derived data sets [57].

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 [59], 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 [60]. 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 [61].

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 [58, 62, 63]. Monitoring the solution composition in UF/DF processes including protein and excipient concentrations have been realized through combinations of multiple sensors [19], near-infrared (NIR) spectroscopy [58] or more recently Raman spectroscopy [53]. Thakur et al. [58] have shown the suitability of off-line calibrated chemometric models to monitor UF/DF processes in real-time. [53] 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 µL. An 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.

3.2 Material and Methods

3.2.1 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. 3.1 (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. 3.1 (B). To enable automated equilibration, loading, injection, measurement and washing, the valves and tubing connections were chosen as visualized in Fig. 3.1 (C)-(G). For all injections in this study, an injection volume of 490 µl 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 29 s totaling to an acquisition of 24 Raman spectra (see Sec. 3.2.2 for more information on the Raman spectroscopy measurements). Afterwards, the sample was 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.

3.2.2 Raman spectroscopy

A Hyperflux PRO Plus (Tornado Spectral Systems, Toronto, Canada) equipped with a 785 nm laser, a Bioreactor BallProbe™ 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.

3.2.3 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 latin hypercube sampling (LHS) and a central composite design (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. 3.4 (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).

3.2.4 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 et al. [64]. 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 et al. [65]. The direct standardization problem was formulated as

$$\left\|\mathbf{X}_\mathrm{T}\mathbf{W} - \mathbf{X}_\mathrm{S} + \lambda_\mathrm{reg} * \mathbf{D}^T\mathbf{D}\right\| \tag{3.1}$$

, where $\mathbf{X}_\mathrm{T}$ and $\mathbf{X}_\mathrm{S}$ are the target and source spectra, respectively. The equation is further regularized using the second-order difference matrix $\mathbf{D}$ and a regularization factor $\lambda_\mathrm{reg} = 0.001$. 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 $\mathbf{X}_\mathrm{T}^{'} = \mathbf{X}_\mathrm{T}\mathbf{W}$ 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 [66]. 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 Wold et al. [67]. 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.

3.2.5 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. 3.1.

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 cm^2 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. 3.2.

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.

3.3 Results and Discussion

3.3.1 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. 3.3 (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. 3.2.4. 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. 3.3 (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 a 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 [68] 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. 3.3 (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 acquire data with in-process data.

3.3.2 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 [69]. 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. 3.4 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. 3.5 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 [58], 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 et al. [58], 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 [70].

3.3.3 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. 3.2. 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 3.1. Fig. 3.6 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 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 [71]. Under these conditions, the Donnan effect causes macroscopic pH shifts and retention of buffer substances [72]. 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.

3.4 Conclusions

This study demonstrates the feasibility of automating Raman spectroscopy measurements using a chromatography system with a minimal sample volume of 490 µL. 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.

Author contribution

Conceptualization, J.H.-M. and R.S.; methodology, J.H.-M. and R.S.; software, M.L.; validation, J.H.-M. and R.S.; formal analysis, J.H.-M. and R.S.; investigation, M.L., F.W., and C.W.; resources, L.R. and M.S.; data curation, J.H.-M. and R.S.; writing—original draft preparation, J.H.-M. and R.S.; writing—review and editing, all authors; visualization, J.H.-M. and R.S.; supervision, J.H.-M.,

R.S., L.R., M.S., and J.H.; project administration, J.H.; funding acquisition, J.H. All authors have read and agreed to the published version of the manuscript.

3.5 Figures

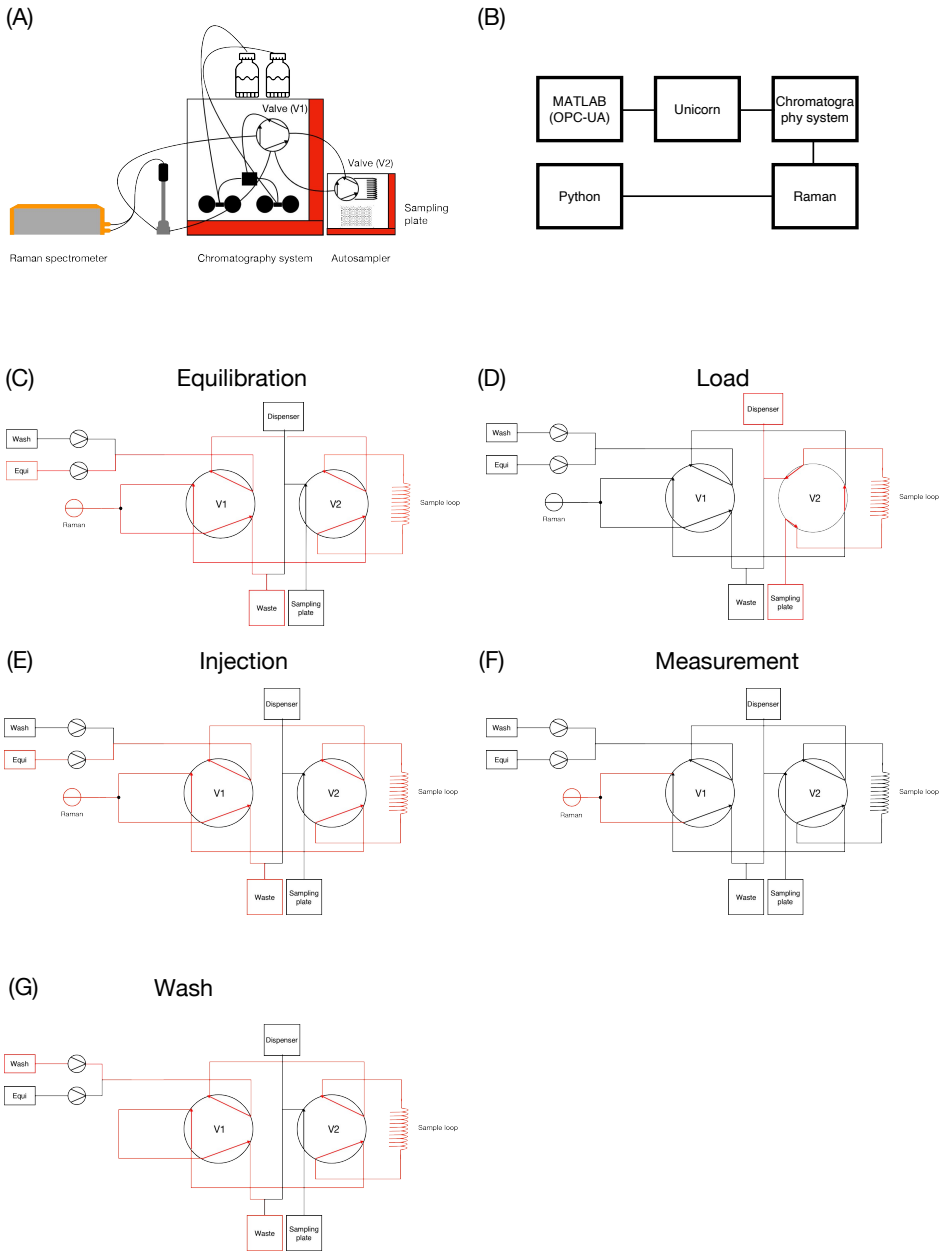


Figure 3.1 Experimental setup connecting ÄKTA purifier, A905 autosampler and Tornado Hyperflux PRO Plus. Schematic illustration of the experimental setup (A). Schematic depiction of software interfaces (B). Flow paths for different phases: equilibration (C), sample load (D), sample injection (E), measurement (F) and wash (G). Red and black lines indicate active and non-active flow paths, respectively.

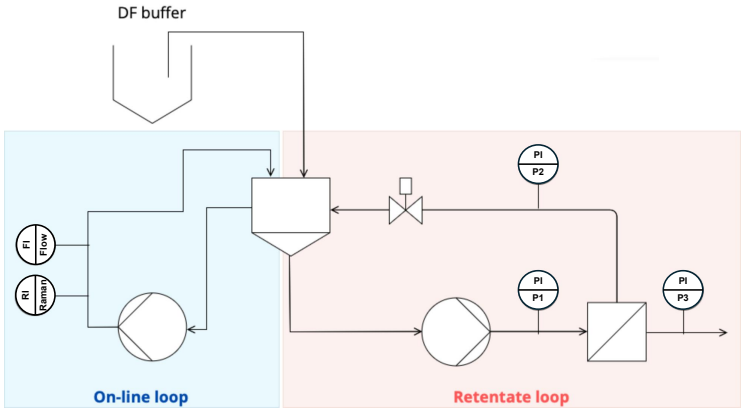


Figure 3.2 Process flow diagram of the CFF system with an external loop for on-line Raman measurements.

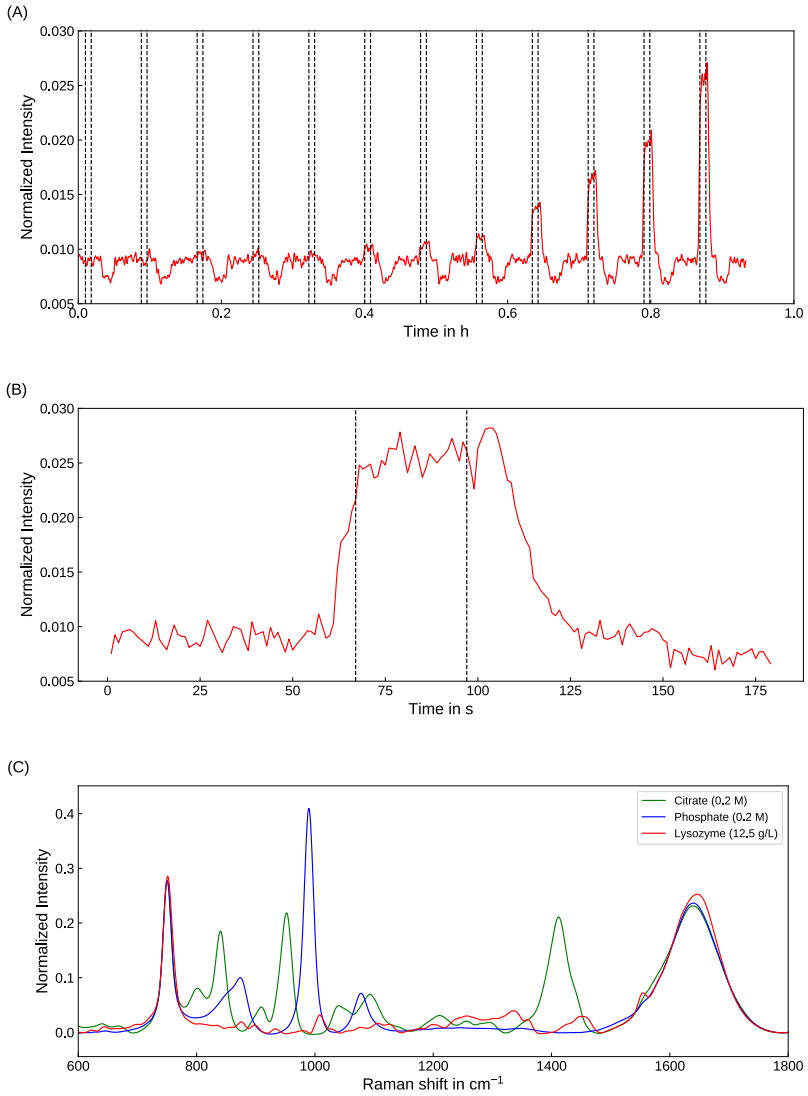


Figure 3.3 Injection sequence of selected samples with increasing lysozyme concentration depicted by the normalized Raman intensity using the $1004\,cm^{-1}$ band after baseline correction as an indicator (A) Dashed vertical lines indicate the time intervals for the static measurements. Zoomed Raman 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).

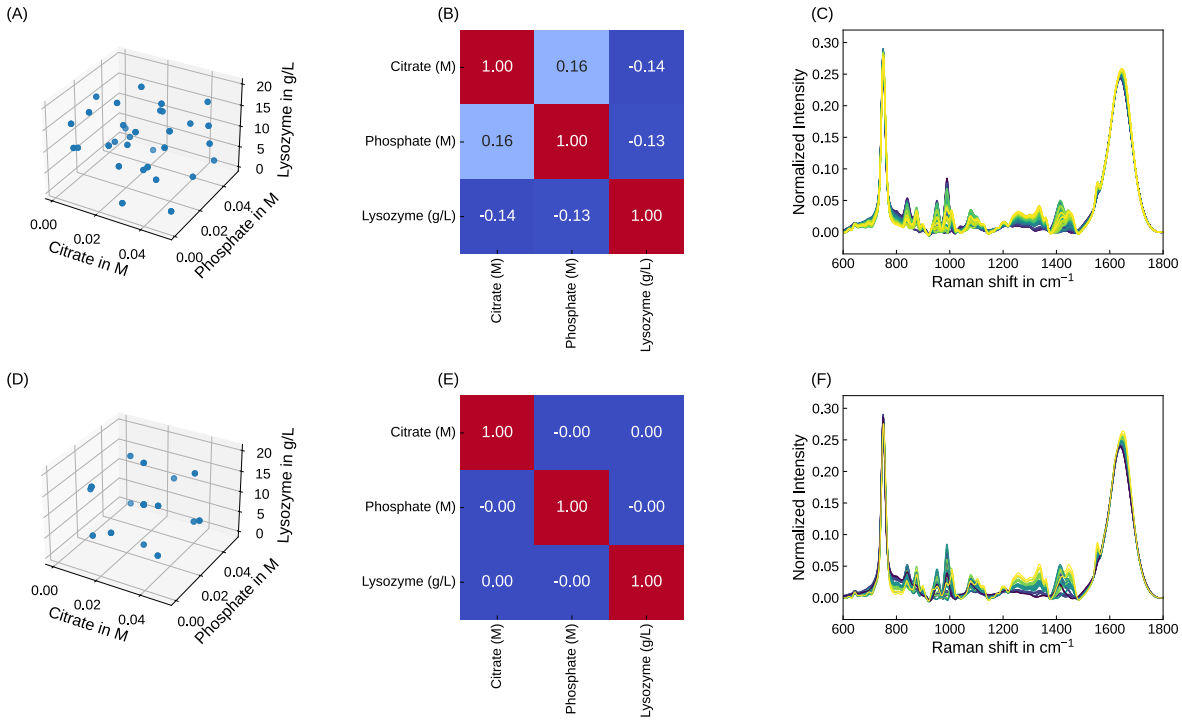


Figure 3.4 Sample distribution (A),(D), correlation coefficients (B),(E), and preprocessed Raman spectra (C), (F) for LHS and CCD datasets in the upper and lower row, respectively.

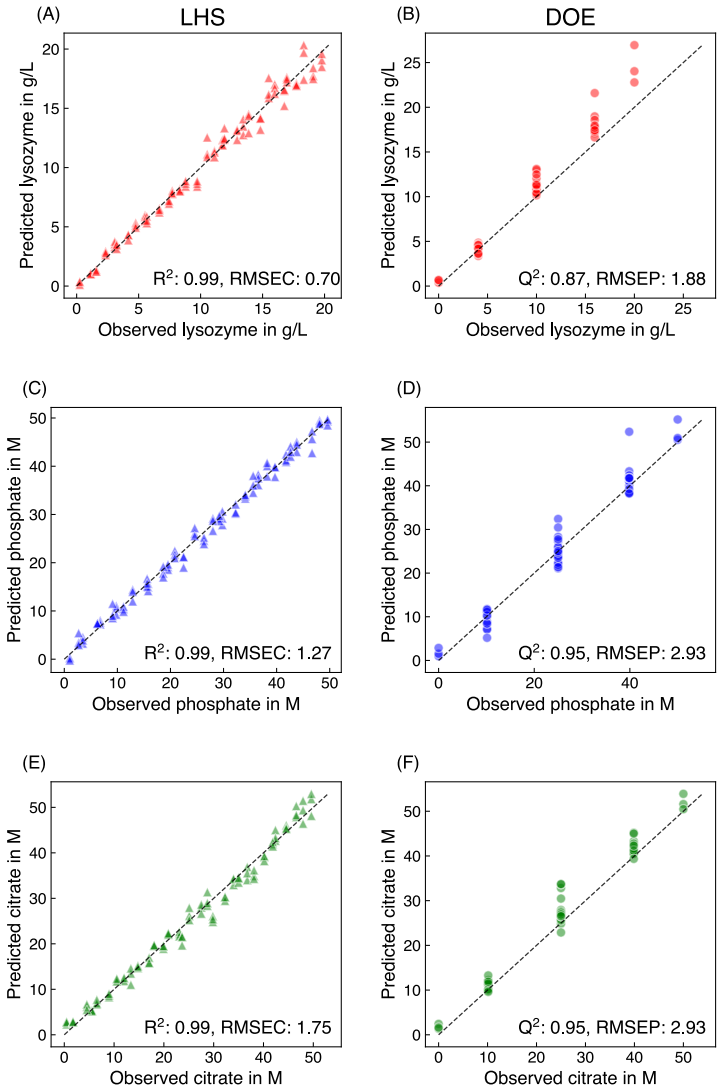


Figure 3.5 Regression errors for calibration datasets LHS and CCD. The errors are displayed for lysozyme, phosphate and citrate from top to bottom. R^2 and $RMSEC$ values refer to the error metrics obtained the data assigned to training, while Q^2 and $RMSEP$ refer to the data assigned to test.

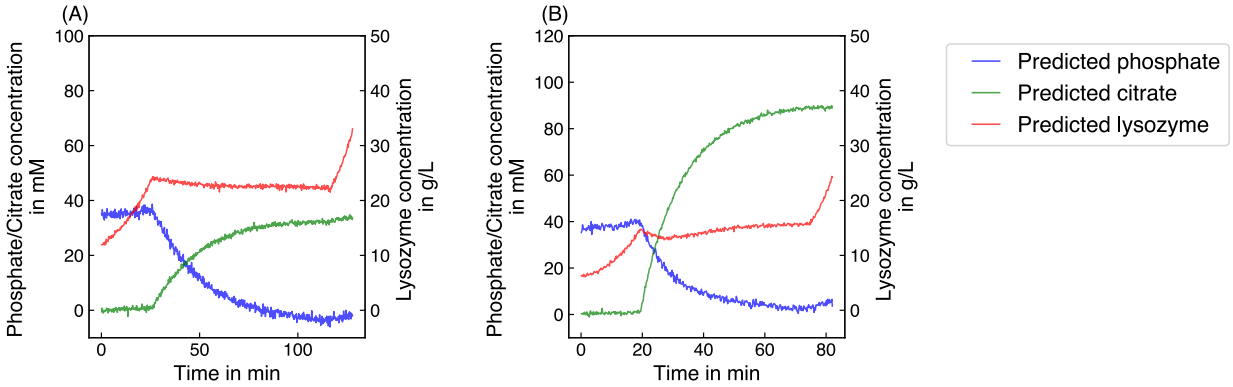


Figure 3.6 Model predictions for in-line data from UF/DF experiments.

3.6 Tables

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

Table 3.1 Experimental conditions for different runs.

Appendix

Appendix **A3** contains the Supporting Figures associated with this chapter.

Chapter References

[19] L. Rolinger et al. „Multi-attribute PAT for UF/DF of proteins—monitoring concentration and buffer exchange by Raman spectroscopy“. In: *Analytical and Bioanalytical Chemistry* 412 (2020), pp. 2123–2134. DOI: 10.1007/s00216-019-02318-8.

[32] A. Dietrich, R. Schiemer, J. Kurmann, S. Zhang, and J. Hubbuch. „Raman-based PAT for VLP precipitation: systematic data diversification and preprocessing pipeline identification“. In: *Frontiers in Bioengineering and Biotechnology* 12 (2024). DOI: 10.3389/fbioe.2024.1399938.

[52] Feidl et al. „Combining Mechanistic Modeling and Raman Spectroscopy for Monitoring Antibody Chromatographic Purification“. In: *Processes* 7.10 (2019), p. 683. DOI: 10.3390/pr7100683.

[53] L. Rolinger et al. „Monitoring of ultra- and diafiltration processes by Raman spectroscopy combined with Kalman filtering“. In: *Biotechnology and Bioengineering* 120 (2023), e28378. DOI: 10.1002/bit.28378.

[54] N. Brestrich, M. Rüdt, D. Büchler, and J. Hubbuch. „Selective protein quantification for preparative chromatography using variable pathlength UV/Vis spectroscopy and partial least squares regression“. In: *Chemical Engineering Science* 176 (2018), pp. 157–164. DOI: 10.1016/j.ces.2017.10.030.

[55] R. M. Santos, J.-M. Kessler, P. Salou, J. C. Menezes, and A. Peinado. „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 (2018), pp. 659–670. DOI: 10.1002/btpr.2635.

[56] P. Romann, J. Kolar, D. Tobler, C. Herwig, J.-M. Bielser, and T. K. Villiger. „Advancing Raman model calibration for perfusion bioprocesses using spiked harvest libraries“. In: *Biotechnology Journal* 17.11 (2022), e2200184. DOI: 10.1002/biot.202200184.

[57] M. Klaverdijk, M. Ottens, and M. E. Klijn. „Single compound data supplementation to enhance transferability of fermentation specific Raman spectroscopy models“. In: *Analytical and Bioanalytical Chemistry* (2025), pp. 1–12. DOI: 10.1007/s00216-025-05768-5.

[58] G. Thakur, V. Hebbi, and A. S. Rathore. „Near Infrared Spectroscopy as a PAT tool for monitoring and control of protein and excipient concentration in ultrafiltration of highly concentrated antibody formulations“. In: *International Journal of Pharmaceutics* 600 (2021), p. 120456. DOI: 10.1016/j.ijpharm.2021.120456.

[59] S. K. Hansen, B. Jamali, and J. Hubbuch. „Selective high throughput protein quantification based on UV absorption spectra“. In: *Biotechnology and Bioengineering* 110.2 (2013), pp. 448–460. DOI: 10.1002/bit.24712.

[60] J. Wang, J. Chen, J. Studts, and G. Wang. „Automated calibration and in-line measurement of product quality during therapeutic monoclonal antibody purification using Raman spectroscopy“. In: *Biotechnology and Bioengineering* (2023). DOI: 10.1002/bit.28514.

[61] J. Chen, J. Wang, R. Hess, G. Wang, J. Studts, and M. Franzreb. „Application of Raman spectroscopy during pharmaceutical process development for determination of critical quality attributes in Protein A chromatography“. In: *Journal of Chromatography A* (2024), p. 464721. DOI: 10.1016/j.chroma.2024.464721.

[62] J. Abel et al. „A small-scale process for predicting Donnan and volume exclusion effects during ultrafiltration/diafiltration process development“. In: *Journal of Pharmaceutical Sciences* 107 (2018), pp. 1600–1609. DOI: 10.1016/j.xphs.2018.01.010.

[63] J. M. West et al. „Process analytical technology for on-line monitoring of biologics processes“. In: *Biotechnology and Bioengineering* 118 (2021), pp. 1234–1245. DOI: 10.1002/bit.27741.

[64] R. Nikzad-Langerodi and E. Andries. „A chemometrician's guide to transfer learning“. In: *Journal of Chemometrics* 35.11 (2021), e3373. DOI: 10.1002/cem.3373.

[65] R. Schiemer, M. Rüdt, and J. Hubbuch. „Generative data augmentation and automated optimization of convolutional neural networks for process monitoring“. In: *Frontiers in Bioengineering and Biotechnology* 12 (2024).

[66] V. I. Korepanov. „Asymmetric least-squares baseline algorithm with peak screening for automatic processing of the Raman spectra“. In: *Journal of Raman Spectroscopy* 51 (2020), pp. 2061–2065. DOI: 10.1002/jrs.5952.

[67] S. Wold, M. Sjöström, and L. Eriksson. „PLS-regression: a basic tool of chemometrics“. In: *Chemometrics and Intelligent Laboratory Systems* 58.2 (2001), pp. 109–130. DOI: 10.1016/S0169-7439(01)00155-1.

[68] J. M. Smulko, N. C. Dingari, J. S. Soares, and I. Barman. „Anatomy of Noise in Quantitative Biological Raman Spectroscopy“. In: *Bioanalysis* 6.3 (2014), pp. 411–421. DOI: 10.4155/bio.13.337.

[69] L. Rolinger, M. Rüdt, and J. Hubbuch. „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 (2020), pp. 2047–2064. DOI: 10.1007/s00216-020-02407-z.

[70] D. H. Müller, M. Börger, J. Thien, and H.-J. Koß. „The Good pH probe: non-invasive pH in-line monitoring using Good buffers and Raman spectroscopy“. In: *Analytical and Bioanalytical Chemistry* 415.29 (2023), pp. 7247–7258. DOI: 10.1007/s00216-023-04993-0.

[71] F. Rischawy et al. „Modeling of biopharmaceutical UF/DF from laboratory to manufacturing scale“. In: *Computers and Chemical Engineering* 177 (2023), p. 108337. DOI: 10.1016/j.compchemeng.2023.108337.

[72] T. Briskot et al. „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 (2022), p. 120333. DOI: 10.1016/j.memsci.2022.120333.

4

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 Robbel[2], Michael Schmitt[2] and Jürgen Hubbuch[1]

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

[2] CSL Innovation GmbH, Marburg, 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.

4.1 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 [73–75] - potentially compromising both the safety and efficacy of therapeutic proteins [27, 76]. 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 [14]. 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 [77] to enable real-time impurity quantification [78–81], protein structure assessment [82], reaction monitoring [83] or precipitation behavior [32]. These innovations support a shift toward Quality by Design (QbD) approaches in

biopharmaceutical production [16].

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 [84, 85]. 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 [86]. Later studies reported similar spectral changes in the amide bands for different natural occuring proteins as well as for common therapeutic proteins such as immunoglobulin Gs (IgGs) [9, 87–89]. 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 solvatization during protein aggregation [9, 89]. However, the vibrational spectra are not only affected by protein aggregation but also by other environmental factors such as pH, protein concentration or excipients [90–92] as well as intrinsic fluorescence and Rayleigh scattering of the protein analytes requiring targeted preprocessing of the collected Raman spectra [32, 77].

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 [80, 87, 93, 94]. 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 [32], chromatography [15], or filtration [69]. Raman spectroscopy has previously been used in bioseparations, but has been shown to have limited accuracy due to the inherent variation in spectral signals [52, 69, 95]. Previous studies have shown that these limitations can be circumvented when combining chemometric models with mechanistic process models [83, 95] or using machine learning models such as convolutional neural networks (CNNs) [65].

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.

4.2 Material and Methods

4.2.1 Experimental

4.2.1.1 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. 4.2.1.4). Residual samples were diluted 1:100 and analyzed by ultrahigh performance - size exclusion chromatography (UHP-SEC) for reference analytics.

4.2.1.2 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 % aggregates. Each condition was prepared in duplicate at a total sample volume of 800 µL, 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 4.1.

4.2.1.3 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 µL, 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 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. 5.1.

4.2.1.4 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 µL (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 [96].

4.2.1.5 Raman spectroscopy - acqusition 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.

4.2.1.6 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 µm particle size, 4.6 × 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).

4.2.2 Data analysis

All data analysis and computations were performed in Python 3.12.7.

4.2.2.1 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 500–3250 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 [97] and was hence used for normalization. Normalization was done to compensate for small variations in total intensity potentially induced by turbidity [32]. 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 $d = 2$ [98]. 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.

4.2.2.2 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 1500-1800 cm^{-1} and 1200-1400 cm^{-1}, respectively [99–101]. 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} [101] and to provide a informative marker of local solvent environment [102, 103]. 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 (1500-1800 cm^{-1}),

$$\bar{\nu} = \frac{\sum_i I_i \nu_i}{\sum_i I_i} \tag{4.1}$$

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

$$SNR = \frac{\mathrm{Var}(\mathbf{x}_\lambda \hat{\boldsymbol{\beta}})}{\hat{\sigma}^2} \tag{4.2}$$

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 ε and $\hat{\sigma}^2$ being the residual variance of linear model for wavenumber λ according to [104]. To calculate SNR for the Raman spectroscopy data using Eq. (4.2), 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. 4.3), error bars represent the standard deviation of two independent replicates, each derived from an averaged spectrum comprising 20 acquisitions.

4.2.2.3 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. 5.1. Raman spectra were preprocessed as laid out in Sec. 4.2.2.1 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 [32, 69].

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. 4.2.2.3.

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 [65]. 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. 4.2.2.3.

Data Augmentation Data augmentation was used to enhance model robustness, utilizing the local subset augmentation (LSA) technique introduced in [65]. 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. 4.2.2.3 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. 4.2.2.3. 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. S1. 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. 4.2.2.3 and 4.2.2.3 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 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.

4.3 Results and Discussion

4.3.1 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. 4.1(A) schematically presents the methodology and the aggregate content over time. The UHP-SEC results illustrated in Fig. 4.1, 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 [9, 89, 105], 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 occuring 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. 4.2 (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. 4.2(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. (4.1). 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 [9, 106]. Especially for proteins with predominantly alpha-helical structures such as BSA, the amide I band is considered a realiable 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 [85]. 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 [9]. 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 [107].

4.3.2 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. 4.2.1.2) to evaluate the robustness of the spectral effects found in Sec. 4.3.1. 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 measurements were subsequently performed off-line using the ARMP. Fig. 4.1 (B) schematically depicts the described workflow and Fig. 4.3 (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 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 4.3 reflects the outcomes of the forced degradation study (Figure 4.2), 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 920–950 cm^{-1} range were assessed [9, 84, 85, 108, 109]. 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 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.

4.3.3 Bind-elute anion-exchange chromatography experiments

4.3.3.1 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. 5.1). 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. 4.4, 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. 4.4, 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 occured 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 occuring dimers and hence show a pronounced effect in the amide I region, which upholds variations in protein concentration [9, 89]. 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 [110]. 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 [97] 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.

4.3.3.2 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. 5.1. 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. S2 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. S2 (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 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. S4 to S13, 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. S3), 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. 4.5 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 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.

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 [65] 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 [65], 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 [111].
Compared to established spectroscopy-based PAT applications in biopharma, such as glucose monitoring in fermentation processes [112–114] or UFDF monitoring [69, 95, 115], spectroscopic methods for aggregate quantification remain limited [80, 116]. Brestrich et al. [116] relied on only four experimental runs, with just one reserved for validation, whereas Wei et al. [80] 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.

4.4 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.

Author contribution

Conceptualization, J.H.-M. and R.S.; methodology, J.H.-M. and R.S.; validation, J.H.-M. and R.S.; formal analysis, J.H.-M. and R.S.; investigation, J.H.-M. and R.S.; resources, L.R. and M.S.; data curation, J.H.-M. and R.S.; writing—original draft preparation, J.H.-M. and R.S.; writing—review and editing, all authors; visualization, J.H.-M. and R.S.; supervision, L.R., M.S., and J.H.; project administration, J.H.; funding acquisition, J.H. All authors have read and agreed to the published version of the manuscript.

4.5 Figures

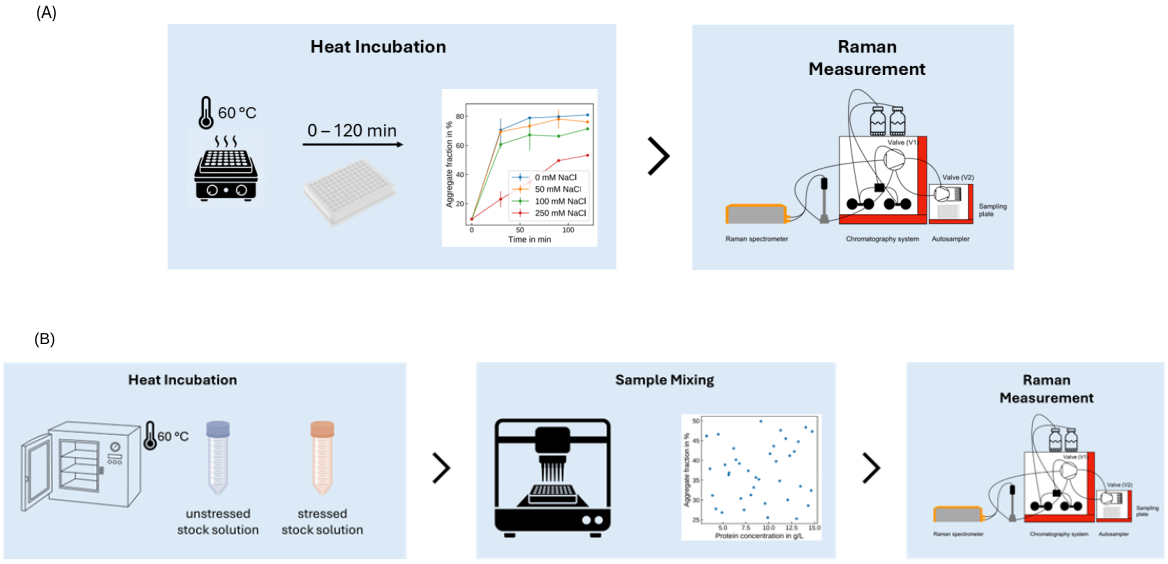


Figure 4.1 Schematic representation of the forced degradation (A) and mixing study workflows (B). (A) Heat-induced BSA aggregation was generated by incubating samples at 60 °C for 0-120 min at defined NaCl concentrations, followed by offline Raman spectroscopy and UHP-SEC reference analytics. (B) To 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.

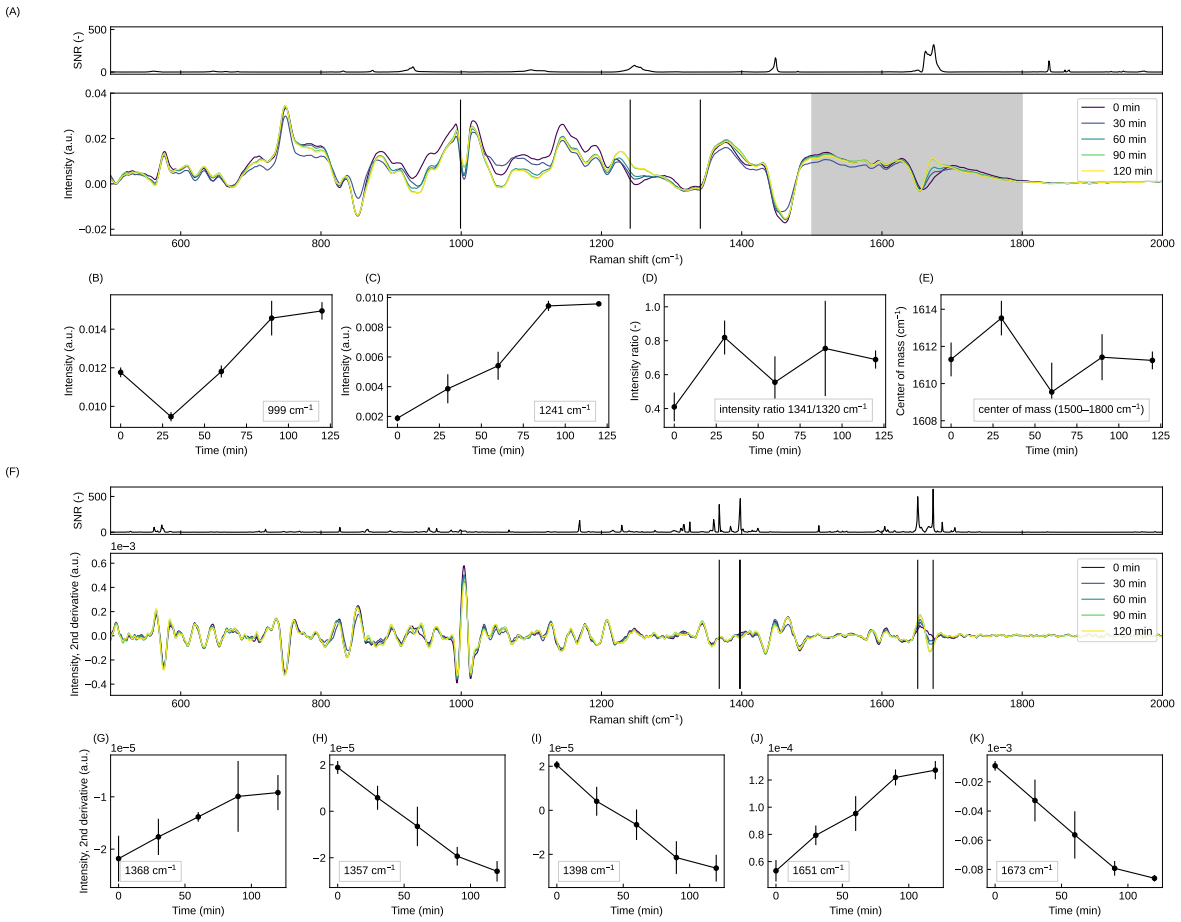


Figure 4.2 Results of the forced degradation experiment. Off-line recorded Raman spectra and SNR for each wavenumber (A). Spectra are shown after preprocessing as described in Sec. 4.2.2.1. The color gradient reflects the increasing duration of heating. Intensities at 999, 1241, the ratio 1341/1320 cm^{-1} 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).

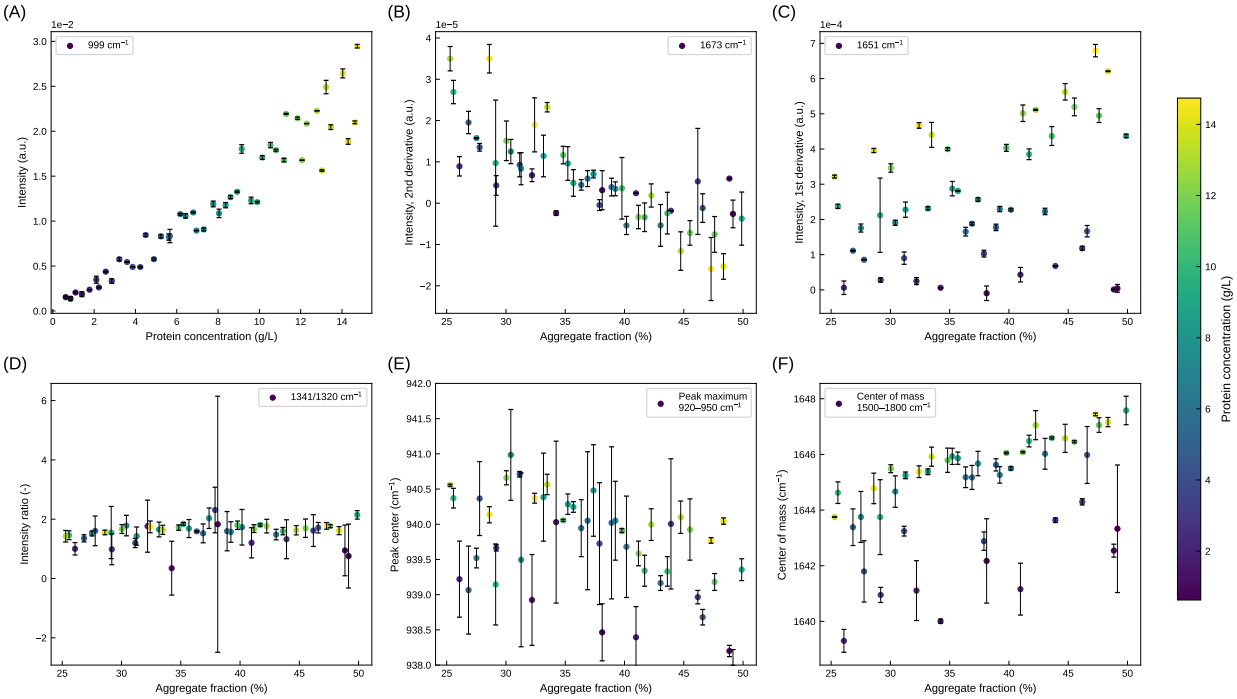


Figure 4.3 Spectral features observed in offline-recorded Raman spectra of BSA obtained during mixing studies, preprocessed as described in Section 4.2.2.1. Raman intensities, peak shifts, and band ratios are plotted as functions of total protein concentration (g/L), aggregate fraction (%), and high 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).

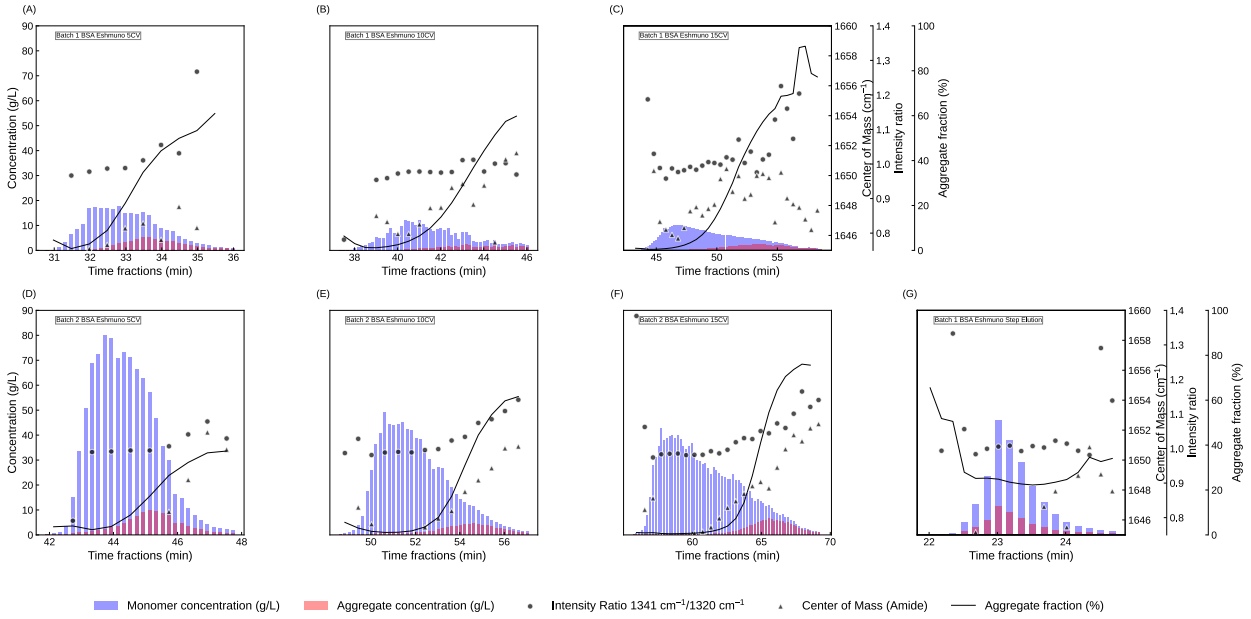


Figure 4.4 Elution profiles and structural markers of BSA under varying process conditions. Protein separation from two BSA batches was performed using different gradient lengths: (A, D) 5 CV; (B, E) 10 CV; (C, F) 15 CV; and (G) step elution. Blue and red bars represent monomer and aggregate concentrations (g/L), respectively. Black lines indicate aggregate fractions across elution time. Black circles and black triangles denote Raman spectral markers: intensity ratio (1341 cm^{-1}/1320 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.

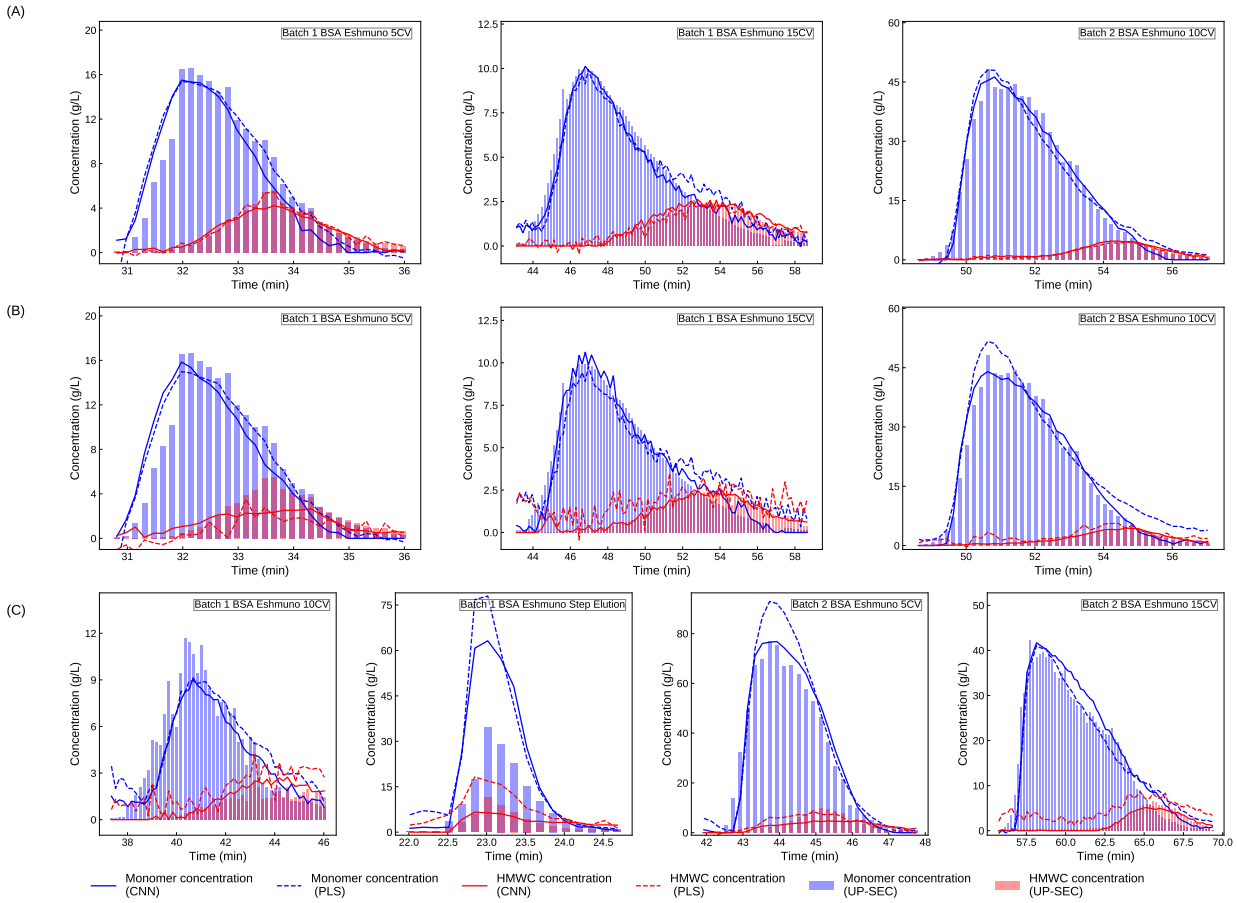


Figure 4.5 Predictions of Crop-SG-PLS1 (dashed lines) and CNN (solid lines) regression models for training (A), cross-validation (B) and test (C). Blue and red bars represent monomer and dimer concentrations as determined by UHP-SEC.

4.6 Tables

Table 4.1 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

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

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

Appendix

Appendix **A4** contains the Supporting Figures associated with this chapter.

Chapter References

[9] C. Zhou, W. Qi, E. N. Lewis, and J. F. Carpenter. „Concomitant Raman spectroscopy and dynamic light scattering for characterization of therapeutic proteins at high concentrations“. In: *Analytical Biochemistry* 472 (2015), pp. 7–20. DOI: 10.1016/j.ab.2014.11.016.

[14] Food and Drug Administration. *PAT — A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance*. 2004.

[15] M. Rüdt, N. Brestrich, L. Rolinger, and J. Hubbuch. „Real-time monitoring and control of the load phase of a protein A capture step“. In: *Biotechnology and Bioengineering* 114.2 (2017), pp. 368–373. DOI: 10.1002/bit.26078.

[16] J. Glassey et al. „Process analytical technology (PAT) for biopharmaceuticals“. In: *Biotechnology Journal* 6.4 (2011), pp. 369–377. DOI: 10.1002/biot.201000356.

[27] A. S. Rosenberg. „Effects of protein aggregates: An immunologic perspective“. In: *The AAPS Journal* 8.3 (2006), p. 59. DOI: 10.1208/aapsj080359.

[32] A. Dietrich, R. Schiemer, J. Kurmann, S. Zhang, and J. Hubbuch. „Raman-based PAT for VLP precipitation: systematic data diversification and preprocessing pipeline identification“. In: *Frontiers in Bioengineering and Biotechnology* 12 (2024). DOI: 10.3389/fbioe.2024.1399938.

[52] Feidl et al. „Combining Mechanistic Modeling and Raman Spectroscopy for Monitoring Antibody Chromatographic Purification“. In: *Processes* 7.10 (2019), p. 683. DOI: 10.3390/pr7100683.

[65] R. Schiemer, M. Rüdt, and J. Hubbuch. „Generative data augmentation and automated optimization of convolutional neural networks for process monitoring“. In: *Frontiers in Bioengineering and Biotechnology* 12 (2024).

[69] L. Rolinger, M. Rüdt, and J. Hubbuch. „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 (2020), pp. 2047–2064. DOI: 10.1007/s00216-020-02407-z.

[73] M. Vázquez-Rey and D. A. Lang. „Aggregates in monoclonal antibody manufacturing processes“. In: *Biotechnology and Bioengineering* 108.7 (2011), pp. 1494–1508. DOI: 10.1002/bit.23155.

[74] C. J. Roberts. „Protein aggregation and its impact on product quality“. In: *Current Opinion in Biotechnology* 30 (2014), pp. 211–217. DOI: 10.1016/j.copbio.2014.08.001.

[75] N. B. Pham and W. S. Meng. „Protein aggregation and immunogenicity of biotherapeutics“. In: *International Journal of Pharmaceutics* 585 (2020), p. 119523. DOI: 10.1016/j.ijpharm.2020.119523.

[76] E. M. Moussa et al. „Immunogenicity of Therapeutic Protein Aggregates“. In: *Journal of Pharmaceutical Sciences* 105.2 (2016), pp. 417–430. DOI: 10.1016/j.xphs.2015.11.002.

[77] L. Rolinger et al. „Multi-attribute PAT for UF/DF of Proteins—Monitoring Concentration, particle sizes, and Buffer Exchange“. In: *Analytical and Bioanalytical Chemistry* 412.9 (2020), pp. 2123–2136. DOI: 10.1007/s00216-019-02318-8.

[78] N. Brestrich, A. Sanden, A. Kraft, K. McCann, J. Bertolini, and J. Hubbuch. „Advances in inline quantification of co-eluting proteins in chromatography: Process-data-based model calibration and application towards real-life separation issues“. In: *Biotechnology and Bioengineering* 112.7 (2015), pp. 1406–1416. DOI: 10.1002/bit.25546.

[79] F. Capito, R. Skudas, H. Kolmar, and C. Hunzinger. „At-line mid infrared spectroscopy for monitoring downstream processing unit operations“. In: *Process Biochemistry* 50.6 (2015), pp. 997–1005. DOI: 10.1016/j.procbio.2015.03.005.

[80] B. Wei et al. „Multi-attribute Raman spectroscopy (MARS) for monitoring product quality attributes in formulated monoclonal antibody therapeutics“. In: *mAbs* 14.1 (2022). DOI: 10.1080/19420862.2021.2007564.

[81] J. Chen, J. Wang, R. Hess, G. Wang, J. Studts, and M. Franzreb. „Application of Raman spectroscopy during pharmaceutical process development for determination of critical quality attributes in Protein A chromatography“. In: *Journal of Chromatography A* 1718 (2024), p. 464721. DOI: 10.1016/j.chroma.2024.464721.

[82] M. Rüdt, P. Vormittag, N. Hillebrandt, and J. Hubbuch. „Process monitoring of virus-like particle reassembly by diafiltration with UVVis spectroscopy and light scattering“. In: *Biotechnology and Bioengineering* 116 (2019), pp. 1366–1379. DOI: 10.1002/bit.26935.

[83] R. Schiemer, J. T. Weggen, K. M. Schmitt, and J. Hubbuch. „An adaptive soft-sensor for advanced real-time monitoring of an antibody-drug conjugation reaction“. In: *Biotechnology and Bioengineering* 120.7 (2023), pp. 1914–1928. DOI: 10.1002/bit.28428.

[84] N. Alizadeh-Pasdar, S. Nakai, and E. C. Y. Li-Chan. „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 (2002), pp. 6042–6052. DOI: 10.1021/jf0116956.

[85] V. Militello, C. Casarino, A. Emanuele, A. Giostra, F. Pullara, and M. Leone. „Aggregation kinetics of bovine serum albumin studied by FTIR spectroscopy and light scattering“. In: *Biophysical Chemistry* 107.2 (2004), pp. 175–187. DOI: 10.1016/j.bpc.2003.09.004.

[86] I. Ettah and L. Ashton. „Engaging with Raman Spectroscopy to Investigate Antibody Aggregation“. In: *Antibodies* 7.3 (2018), p. 24. DOI: 10.3390/antib7030024.

[87] R. G. de la Cuesta, R. Goodacre, and L. Ashton. „Monitoring Antibody Aggregation in Early Drug Development Using Raman Spectroscopy and Perturbation-Correlation Moving Windows“. In: *Analytical Chemistry* 86.22 (2014), pp. 11133–11140. DOI: 10.1021/ac5038329.

[88] E. N. Lewis, W. Qi, L. H. Kidder, S. Amin, S. M. Kenyon, and S. Blake. „Combined Dynamic Light Scattering and Raman Spectroscopy Approach for Characterizing the Aggregation of Therapeutic Proteins“. In: *Molecules* 19.12 (2014), pp. 20888–20905. DOI: 10.3390/molecules191220888.

[89] G. V. Barnett et al. „Structural Changes and Aggregation Mechanisms for Anti-Streptavidin IgG1 at Elevated Concentration“. In: *The Journal of Physical Chemistry B* 119.49 (2015), pp. 15150–15163. DOI: 10.1021/acs.jpcb.5b08748.

[90] C. Ota, S. Noguchi, S. Nagatoishi, and K. Tsumoto. „Assessment of the Protein–Protein Interactions in a Highly Concentrated Antibody Solution by Using Raman Spectroscopy“. In: *Pharmaceutical Research* 33.4 (2016), pp. 956–969. DOI: 10.1007/s11095-015-1842-8.

[91] Y. Sato, S. Nagatoishi, S. Noguchi, and K. Tsumoto. „Raman Spectroscopic Analysis of Highly-Concentrated Antibodies under the Acid-Treated Conditions“. In: *Pharmaceutical Research* 40.7 (2023), pp. 1853–1864. DOI: 10.1007/s11095-023-03526-9.

[92] A. A. Makki et al. „Understanding the discrimination and quantification of monoclonal antibodies preparations using Raman spectroscopy“. In: *Journal of Pharmaceutical and Biomedical Analysis* 194 (2021), p. 113734. DOI: 10.1016/j.jpba.2020.113734.

[93] C. Zhang, J. S. Springall, X. Wang, and I. Barman. „Rapid, quantitative determination of aggregation and particle formation for antibody drug conjugate therapeutics with label-free Raman spectroscopy“. In: *Analytica Chimica Acta* 1081 (2019), pp. 138–145. DOI: 10.1016/j.aca.2019.07.007.

[94] J. Wang, J. Chen, J. Studts, and G. Wang. „In-line product quality monitoring during biopharmaceutical manufacturing using computational Raman spectroscopy“. In: *mAbs* 15.1 (2023), p. 2220149. DOI: 10.1080/19420862.2023.2220149.

[95] L. Rolinger, J. Hubbuch, and M. Rüdt. „Monitoring of ultra- and diafiltration processes by Kalman-filtered Raman measurements“. In: *Analytical and Bioanalytical Chemistry* 415.5 (2023), pp. 841–854. DOI: 10.1007/s00216-022-04477-7.

[96] J. Heyer-Müller et al. „A novel Raman-chromatography assembly for automated calibration and in-line monitoring in bioprocessing“. In: *Engineering in Life Sciences* (2025). DOI: 10.1002/elsc.70044.

[97] K. Furić, I. Ciglenečki, and B. Ćosović. „Raman spectroscopic study of sodium chloride water solutions“. In: *Journal of Molecular Structure* 550–551 (2000), pp. 225–234. DOI: 10.1016/S0022-2860(00)00388-4.

[98] D. Erb. *pybaselines: A Python library of algorithms for the baseline correction of experimental data*. DOI: 10.5281/zenodo.5608581.

[99] J. L. Lippert, D. Tyminski, and P. J. Desmeules. „Determination of the secondary structure of proteins by laser Raman spectroscopy“. In: *Journal of the American Chemical Society* 98.22 (1976), pp. 7075–7080. DOI: 10.1021/ja00438a057.

[100] R. W. Williams. „[14] Protein secondary structure analysis using Raman amide I and amide III spectra“. In: *Methods in Enzymology*. Vol. 130. Elsevier, 1986, pp. 311–331. DOI: 10.1016/0076-6879(86)30016-8.

[101] A. Rygula, K. Majzner, K. M. Marzec, A. Kaczor, M. Pilarczyk, and M. Baranska. „Raman spectroscopy of proteins: A review“. In: *Journal of Raman Spectroscopy* 44.8 (2013), pp. 1061–1076. DOI: 10.1002/jrs.4335.

[102] I. Harada, T. Miura, and H. Takeuchi. „Origin of the doublet at 1360 and 1340 cm−1 in the Raman spectra of tryptophan and related compounds“. In: *Spectrochimica Acta Part A: Molecular Spectroscopy* 42.2–3 (1986), pp. 307–312. DOI: 10.1016/0584-8539(86)80193-3.

[103] A. S. Eisenberg and L. J. Juszczak. „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 (2012), pp. 1–10. DOI: 10.1155/2012/735076.

[104] J. Soch and C. Allefeld. „MACS – a new SPM toolbox for model assessment, comparison and selection“. In: *Journal of Neuroscience Methods* 306 (2018), pp. 19–31. DOI: 10.1016/j.jneumeth.2018.05.017.

[105] M. Boulet-Audet, B. Byrne, and S. G. Kazarian. „High-Throughput Thermal Stability Analysis of a Monoclonal Antibody by Attenuated Total Reflection FT-IR Spectroscopic Imaging“. In: *Analytical Chemistry* 86.19 (2014), pp. 9786–9793. DOI: 10.1021/ac502529q.

[106] B. Shivu, S. Seshadri, J. Li, K. A. Oberg, V. N. Uversky, and A. L. Fink. „Distinct -Sheet Structure in Protein Aggregates Determined by ATR–FTIR Spectroscopy“. In: *Biochemistry* 52.31 (2013), pp. 5176–5183. DOI: 10.1021/bi400625v.

[107] M. D. Fontana, K. B. Mabrouk, and T. H. Kauffmann. „Raman spectroscopic sensors for inorganic salts“. In: *Spectroscopic Properties of Inorganic and Organometallic Compounds*. Ed. by J. Yarwood, R. Douthwaite, and S. Duckett. Vol. 44. RSC Publishing, 2013, pp. 40–67. DOI: 10.1039/9781849737791-00040.

[108] L. Ashton and E. W. Blanch. „pH-induced conformational transitions in α-lactalbumin investigated with two-dimensional Raman correlation variance plots and moving windows“. In: *Journal of Molecular Structure* 974.1 (2010), pp. 132–138. DOI: 10.1016/j.molstruc.2010.03.005.

[109] C. H. Li and T. Li. „Application of Vibrational Spectroscopy to the Structural Characterization of Monoclonal Antibody and its Aggregate“. In: *Current Pharmaceutical Biotechnology* 10.4 (2009), pp. 391–399. DOI: 10.2174/138920109788488950.

[110] J. M. Antosiewicz and D. Shugar. „UV–Vis spectroscopy of tyrosine side-groups in studies of protein structure. Part 2: selected applications“. In: *Biophysical Reviews* 8 (2016), pp. 163–177. DOI: 10.1007/s12551-016-0197-7.

[111] C. Lange et al. „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 F. Manenti and G. V. Reklaitis. Vol. 53. Computer Aided Chemical Engineering. Elsevier, 2024, pp. 3055–3060. DOI: 10.1016/B978-0-443-28824-1.50510-X.

[112] A. G. Cavinato, D. M. Mayes, Z. Ge, J. B. Callis, et al. „Noninvasive method for monitoring ethanol in fermentation processes using fiber-optic near-infrared spectroscopy“. In: *Analytical Chemistry* 62.18 (1990), pp. 2058–2063. DOI: 10.1021/ac00217a015.

[113] D. H. Müller, H.-J. Koss, et al. „Bioprocess in-line monitoring and control using Raman spectroscopy and Indirect Hard Modeling (IHM)“. In: *Biotechnology and Bioengineering* (2024). DOI: 10.1002/bit.28724.

[114] K. Pontius, H. Junicke, K. V. Gernaey, and M. Bevilacqua. „Monitoring yeast fermentations by nonlinear infrared technology and chemometrics – understanding process correlations and indirect predictions“. In: *Applied Microbiology and Biotechnology* 104.12 (2020), pp. 5315–5335. DOI: 10.1007/s00253-020-10604-0.

[115] A. Milewska, G. Baekelandt, S. Boutaieb, V. Mozin, and A. Falconbridge. „In-line monitoring of protein concentration with mid-infrared spectroscopy during ultrafiltration/diafiltration“. In: *Engineering in Life Sciences* 23.2 (2023), e2200050. DOI: 10.1002/elsc.202200050.

[116] N. Brestrich, M. Rüdt, D. Büchler, and J. Hubbuch. „Selective protein quantification for preparative chromatography using variable pathlength UV/Vis spectroscopy and partial least squares regression“. In: *Chemical Engineering Science* 176.C (2018), pp. 157–164. DOI: 10.1016/j.ces.2017.10.030.

5

Mechanistic Deconvolution of BSA Size Variants by Constrained Raman Pseudo-Voigt Hard Modeling during Anion-Exchange Chromatography

Jakob Heyer-Müller[1], Robin Schiemer[1], Lars Robbel[2], Michael Schmitt[2] and Jürgen Hubbuch[1]

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

[2] CSL Innovation GmbH, Marburg, Germany

Abstract

In biopharmaceutical manufacturing, protein aggregation is a critical quality attribute, necessitating rapid and reliable analytical strategies during downstream processes like anion- exchange chromatography (AEX). While Raman spectroscopy enables continuous monitoring of protein secondary structure, standard data-driven regression models struggle to decouple intrinsic structural changes from gradient-induced solvent and buffer drifts under dynamic chromatographic conditions. Addressing this methodological gap, this study establishes a constrained pseudo-Voigt hard modeling framework for the mechanistic deconvolution of bovine serum albumin (BSA) size variants during in-line Raman monitoring of AEX processes. By explicitly defining a parametric background model to capture salt-induced spectral drift, the methodology effectively isolates matrix variations from genuine protein-specific signals. The constrained hard model was applied to 285 in-line spectra across diverse chromatographic conditions, achieving reconstruction fidelity while maintaining stable, physically interpretable component identities. The mechanistically derived Amide I center of mass emerged as a robust, aggregation-sensitive descriptor that preserves structural information despite strong concentration dynamics. Furthermore, the extracted spectral features demonstrated strong predictive performance for monomer concentration and acceptable accuracy for high molecular weight components. Collectively, these results demonstrate that constrained spectral hard modeling provides a highly interpretable, robust, and calibration-light alternative to classical partial least squares approaches for the real-time monitoring of protein size variants.

5.1 Introduction

Protein aggregation is a critical quality attribute in biopharmaceutical manufacturing, necessitating robust analytical strategies for rapid detection during downstream separation processes like anion-exchange chromatography (AEX) [73–75]. While Raman spectroscopy enables continuous monitoring of protein secondary structure via the structurally sensitive Amide I vibrational band, standard data-driven methods, such as partial least squares (PLS) regression, struggle under dynamic chromatographic conditions [33, 78–81], protein structure assessment [82], reaction monitoring [83]. Gradient elution introduces pronounced solvent and buffer-related spectral drifts that PLS cannot reliably decouple from intrinsic structural changes, resulting in entangled, calibration-heavy models that lack physical interpretability. To circumvent these limitations, spectral hard modeling integrates mechanistic prior knowledge by representing overlapping bands as a linear superposition of defined mathematical functions (e.g., pseudo-Voigt profiles) [117–119]. By enforcing physically motivated constraints, hard modeling preserves component identity and isolates background interference from genuine analyte variance. However, robustly resolving the highly overlapping contributions of solvent, monomeric proteins, and aggregated species within the narrow Amide I window remains a fundamental challenge for dynamic bioprocesses [9, 86–89]. Addressing this methodological gap requires a systematic evaluation of spectral deconvolution under realistic operational conditions, guided by three core questions: How can mechanistic prior knowledge effectively isolate gradient-induced matrix effects from protein-specific structural signals? How accurately can individual spectral contributions corresponding to background, monomeric, and aggregated species be resolved within the dynamic Amide I envelope? To what extent do mechanistically derived spectral features improve physical interpretability without sacrificing the predictive performance required for quantitative process monitoring? To answer these questions, this work establishes a constrained pseudo-Voigt

hard modeling framework for the mechanistic deconvolution of bovine serum albumin (BSA) size variants during in-line Raman monitoring of AEX processes. By explicitly defining a parametric background model to capture salt-induced spectral drift, this methodology isolates solvent variations and extracts physically interpretable parameters for the robust, real-time resolution of protein monomers and aggregates.

5.2 Materials and Methods

5.2.1 Experimental

5.2.1.1 Bind–elute anion-exchange chromatography

Bind–elute anion-exchange chromatography (AEX) chromatography experiments with in-line Raman spectroscopy were conducted to generate process data for the separation of bovine serum albumin (BSA) monomers and aggregates under process-relevant conditions with dynamically changing buffer composition. Experiments were performed using a 5 mL Eshmuno Q column (Merck KGaA, Darmstadt, Germany) on an Äkta Pure system (Cytiva, Uppsala, Sweden). Unstressed BSA feed solutions with a total protein concentration of 15 g/L were prepared using a single commercial lot of lyophilized powder (Sigma-Aldrich, St. Louis, MO, USA). To evaluate the impact of feedstock preparation and reconstitution reproducibility, two independent batches were prepared separately by dissolving the identical starting material in 20 mM Tris at pH 8. These separate dissolution procedures resulted in native baseline high molecular weight component (HMWC) contents of 25.33% for Batch 1 and 15.67% for Batch 2, as characterized via SE-UPLC reference analytics.

It is important to emphasize that no artificial stress mechanisms (such as heat, pH excursions, or chemical denaturation) were applied to induce aggregation. The HMWC fraction represents the inherent baseline aggregation of the commercial lyophilized powder, consisting primarily of non-covalently associated oligomers and native disulfide-linked dimers formed during industrial manufacturing, lyophilization, and subsequent storage of the purchased material prior to reconstitution. Utilizing this unstressed material ensures that the Raman spectra reflect realistic, process-relevant aggregation pathways rather than artificially denatured conformations. While an initial aggregate content of 15 to 25% is unusually high for a commercial-grade protein, this elevated baseline is highly advantageous for the present study. The primary objective of this work is to rigorously evaluate the spectral hard modeling workflow and its mathematical decoupling capabilities across a broad dynamic range of size variants, making the absolute purity of the starting material secondary to its structural diversity.

Prior to sample loading, the column was equilibrated for 5 column volumes (CVs) using 20 mM Tris (pH 8). Following sample loading, the column was washed for 5 CVs with equilibration buffer to remove unbound components and stabilize the baseline prior to elution. Gradient elution was subsequently performed using linear NaCl gradients ranging from 0 to 500 mM NaCl at different gradient lengths. After the gradient phase, a strip step at 1 M NaCl was applied for 5 CVs to remove strongly retained species. The column was then regenerated using 1 M NaOH for 5 CVs and re-equilibrated with equilibration buffer to restore the initial conditions. All elution phases were operated at a constant flow rate of 1 mL/min. During the elution phase, fractions were collected at intervals of 200 µL. Raman spectra were recorded in-line at an exposure time of 500 ms and a laser power of 495 mW, resulting in 24 individual acquisitions per fraction. To preserve

temporal resolution during the chromatographic process, single 500 ms acquisitions were used. These acquisitions were subsequently averaged over the time intervals corresponding to each ultrahigh performance - size exclusion chromatography (UHP-SEC)-analyzed fraction to enable quantitative comparison with fraction-wise analytics. The Raman measurement point was incorporated into the flow path between the outlet valve and the fractionator, ensuring that the recorded spectra correspond to the effluent immediately upstream of fractionation. In total, six AEX runs were performed using different gradient lengths, loading densities, and feedstocks derived from two independent BSA batches. Loading densities were intentionally scaled in direct proportion to the gradient lengths across the experiments. This experimental design was deployed to break any systematic, co-linear mathematical correlation between absolute protein elution concentration profiles and local salt kinetics, thereby strictly validating the decoupling capability of the hard-modeling framework. A summary of the experimental conditions, including the assignment to training and test sets, is provided in Table 5.1. Despite the variance in the absolute initial aggregate percentages induced by the separate reconstitution events, SE-UPLC analytics confirmed that both prepared feedstocks exhibited a highly consistent distribution profile of dimeric and higher-order oligomeric species. This structural consistency justifies the treatment of the HMWC fraction as a functionally homogeneous class during spectral feature extraction. Throughout this manuscript, these chromatographic datasets serve as the basis for evaluating the robustness of the hard modeling workflow.

5.2.1.2 Raman acquisition settings

Raman spectroscopy measurements were performed using a Raman BioReactor BallProbe inserted into a Raman flow cell with a dead volume of 240 µL (both MarqMetrix, Seattle, WA, USA) connected to a HyperFlux™ PRO Plus 785 spectrometer operated using SpectralSoft 3.2.600.1 software (Tornado Spectral Systems, Mississauga, Ontario, Canada). All Raman measurements were performed at the maximum laser power of 495 mW. To preserve temporal resolution during continuous process measurements, the exposure time was set to 500 ms per acquisition.

5.2.1.3 Ultrahigh-performance size-exclusion chromatography

To quantify the distribution of protein size variants and provide fraction-wise reference analytics for the Raman measurements, samples were analyzed using UHP-SEC. A TSKgel SuperSW mAb HTP column (4 µm particle size, 4.6×150 mm) was operated at a flow rate of 0.3 mL/min. The mobile phase consisted of 15 mM sodium phosphate buffer at pH 6.2. Prior to injection, samples were diluted to protein concentrations within the range of 0–1 mg/mL to ensure operation within the linear response regime and to prevent column overloading. Analyses were performed using a Vanquish UHPLC system controlled by Chromeleon software (version 7.2) (Thermo Fisher Scientific, Waltham, MA, USA). For each fraction, the absolute concentrations of monomeric and aggregated protein species were determined by peak integration. These concentrations were subsequently used for (i) qualitative comparison with Raman-derived spectral markers and (ii) quantitative evaluation of model-derived trajectories.

5.2.2 Data analysis

All data analysis and computations were performed in Python 3.12.7. The analysis workflow consisted of multiple steps, including (i) spectral preprocessing, (ii) definition and fitting of a

constrained background model for gradient-induced matrix effects, (iii) constrained pseudo-Voigt hard modeling of the protein Amide I band, (iv) extraction of physically interpretable parameters and derived spectral markers and (v) quantitative evaluation of model-derived trajectories. All processing steps were applied consistently across chromatographic runs to ensure comparability of the fitted parameters and marker trajectories.

5.2.2.1 Spectral preprocessing

Raman spectral processing consisted of multiple steps, including truncation, normalization, background subtraction, baseline correction, smoothing, feature selection, and derivative computation. Each operation was chosen to suppress non-chemical variability while retaining aggregation-relevant structural information in the Amide I region.

Truncation All spectra were first truncated to 500–3250 cm^{-1}. The lower bound of 500 cm^{-1} was chosen as smaller wavenumbers solely contain baseline contributions and bands stemming from the sapphire window built into the probe head. The upper bound of 3250 cm^{-1} corresponds to the maximum of the water stretching band within the accessible spectral range and was retained to ensure consistent normalization [97].

Normalization Spectra were subsequently normalized using index normalization to the local baseline minimum adjacent to the C–H stretching region at 2750 cm^{-1}. While the general methodological concept of utilizing the C–H stretching envelope as a stable scaling reference under aqueous, protein-containing conditions is derived from Dietrich et al. [32], the specific wavenumber of 2750 cm^{-1} was selected empirically in this study. This frequency is established as a reliable intensity anchor point due to the absence of intrinsic protein or salt vibrational modes in the immediate region preceding the aliphatic C–H stretching profile [120]. Index-based normalization was preferred over global area normalization to avoid confounding changes in broadband intensity contributions from water bands[cite: 75]. Unlike Standard Normal Variate (SNV) normalization, which scales based on global spectral variance and can distort peak ratios when solvent backgrounds change dynamically, index-based scaling to this local baseline anchor remains unaffected by salt-induced drifts in water intensity.

Background subtraction Subsequently, background subtraction was performed to reduce solvent- and buffer-related contributions prior to baseline correction. A representative protein-free buffer spectrum originating from the start of the respective gradient elution phase, where no protein signal was detectable, was selected for each chromatographic experiment. This buffer spectrum was subtracted from all subsequently acquired spectra on an index-wise basis. Background subtraction was performed to improve visualization of protein-related spectral features and stabilize downstream baseline correction, particularly in the Amide I region, where water and buffer contributions can overlap with protein vibrations. As the NaCl concentration changes continuously during gradient elution, background subtraction alone is insufficient to fully remove salt-induced spectral variation. Residual salt-dependent changes were therefore explicitly addressed in the hard modeling framework by parametric background modeling (Section 5.3.1.1).

Baseline correction Furthermore, baseline correction was performed using a Whittaker smoothing approach through the derivative peak-screening asymmetric least squares algorithm (DERPSALSA)

as implemented in the *pybaselines* library (v. 1.1.0), with parameters $\lambda = 10^5$, $k = 0.02$, and $d = 2$ [98]. These parameters were selected to suppress broad baseline components (e.g., fluorescence, instrument drift, residual scattering) while preserving peak shapes and relative intensities in the Amide I window.

Smoothing and derivatives Finally, high-frequency noise reduction was performed using a Savitzky Golay (SG) filter as implemented in *scipy* (v. 1.14.1), with a second-order polynomial and a window size of 21 points. Smoothed spectra were used for visualization and as inputs for peak initialization. The SG filter was also employed for second-derivative computation using the same filter configuration to enhance subtle spectral variations and support robust peak localization in overlapping band regions (Figure 5.1). The second derivative was explicitly preferred over the first derivative as it simultaneously eliminates linear baseline slopes and converts convoluted inflection shoulders into distinct, sharp downward minima. In the present work, derivative spectra were used exclusively for peak localization and initialization; all final fitting and quantitative marker extraction were performed on the corresponding smoothed (non-derivative) spectra to preserve physically meaningful intensity information.

Feature selection All downstream analyses were restricted to the Amide I region (1500–1801 cm^{-1}), which is sensitive to protein secondary structure and aggregation-related conformational changes. Restricting the spectral window was performed to reduce the number of fitted degrees of freedom, improve the numerical stability of constrained fitting, and focus the hard model on the most structurally informative region. For the remainder of this manuscript, "Amide I" refers to this window.

5.3 Results and Discussion

5.3.1 Hard modeling framework

To enable a mechanistic and physically interpretable decomposition of Amide I Raman spectra into contributions from solvent/buffer and protein structure, a hard modeling framework was developed. In contrast to purely data-driven multivariate approaches, the hard model enforces component identity across spectra by constraining peak centers and widths, enabling direct comparison of fitted parameters between fractions and experiments. The hard model was designed to disentangle (i) intrinsic structural changes associated with aggregation from (ii) extrinsic matrix-induced spectral variation under gradient elution, thereby supporting robust analysis of size variant behavior under process conditions.

5.3.1.1 Buffer background modeling and salt-induced effects

During bind–elute AEX chromatography, gradient elution induces systematic changes in buffer composition, primarily through increasing NaCl concentrations. These changes affect the Raman spectra through solvent-related contributions, including the H–O–H bending vibration and additional oxygen-related bands overlapping with the Amide I region [33]. To prevent these effects from being absorbed into protein-associated components during spectral deconvolution, background contributions were explicitly modeled as a low-dimensional parametric hard model prior to protein analysis. A protein-free buffer reference spectrum was obtained from the first spectrum of the 5 CV

gradient elution segment, assuming negligible protein contribution. To ensure numerical stability and consistent parameterization, the reference spectrum was interpolated to a uniform wavenumber grid ranging from 1500 to 1800 cm^{-1} with a spacing of 1 cm^{-1}. This interpolation ensures that peak localization and pseudo-Voigt evaluation operate on an identical spectral grid across all chromatographic runs. Candidate background peak positions were subsequently identified using a continuous wavelet transform (CWT) [121] with a Mexican hat (Ricker) wavelet [122]. Wavelet coefficients were computed across multiple scales and summed over low-to-intermediate scales to emphasize peak-like structures in the presence of residual noise. The resulting peak candidates were refined by restricting the search to the range 1550–1700 cm^{-1}, thereby focusing on solvent-related contributions overlapping with the Amide I region while excluding regions dominated by protein bands. The background contribution in the Amide I region was represented as the superposition of two pseudo-Voigt components (visualized in Figure 5.2A),

$$y_{\mathrm{bg}}(\nu) = \sum_{k=1}^{2} A_k^{\mathrm{bg}} \cdot PV(\nu; \nu_k^{\mathrm{bg}}, \sigma_k^{\mathrm{bg}}, \eta_k^{\mathrm{bg}}), \tag{5.1}$$

with initial center estimates located near 1550 and 1640 cm^{-1}. Model parameters were estimated using nonlinear least-squares optimization and stored as a reference background model for subsequent analysis. To capture the dominant solvent- and buffer-related features within the Amide I window while avoiding overparameterization, the two-component model was selected as a parsimonious representation.

To quantify gradual salt-induced spectral changes during gradient elution, the reference background model was fitted to buffer spectra acquired across the NaCl gradient using three constraint strategies. First, a global scaling approach was applied in which the background model was multiplied by a non-negative constant factor $c \geq 0$ calculated independently per spectrum to dynamically compensate for frame-wise linear fluctuations in absolute baseline intensity caused by path length variations while keeping all pseudo-Voigt parameters fixed. Second, a fully flexible refit was evaluated that allowed all background parameters (A_k^{bg}, ν_k^{bg}, σ_k^{bg}, η_k^{bg}) to vary independently for each spectrum. Third, a partially flexible refit was implemented in which the model was kept fixed except for selected parameters of the second background component, namely the center position and amplitude, which were allowed to vary to capture salt-driven spectral shifts and intensity changes while maintaining model parsimony (Figure 5.2B). The resulting parameter trajectories served two purposes: (i) providing empirical evidence of the magnitude and direction of salt-driven drift in the Amide I background and (ii) defining the constrained background treatment employed during protein hard modeling to prevent confounding between background and protein components.

A dedicated parametric background model was fitted to protein-free buffer spectra and subsequently utilized as a constrained background contribution during protein deconvolution within the Amide I window (1500–1800 cm^{-1}). The reference background fit yielded excellent agreement with the measured buffer spectrum, with a coefficient of determination of $R^2 = 0.99954$ and an root mean squared error (RMSE) of 1.225×10^{-3} in model intensity units, confirming that the selected low-dimensional pseudo-Voigt representation adequately captures the dominant solvent-related contributions in this spectral region. The reference background model comprised two pseudo-Voigt components centered at approximately 1555 cm^{-1} and 1641.5 cm^{-1}. These peak positions are consistent with solvent-dominated contributions overlapping the Amide I region and align with the expected spectral signatures of water-related bending modes and associated background

features under aqueous buffer conditions. To assess salt-induced drift during gradient elution, the background model was fitted to the first and last buffer spectra of each elution experiment. Across all experiments, a systematic drift of the dominant water/background component ($p1$) was observed. The center position exhibited shifts of $\Delta\nu_{p1} = +0.12$ to $+0.89$ cm^{-1}, while the corresponding amplitude increased by $\Delta A_{p1} = +0.18$ to $+1.55$ in model intensity units. Despite these gradual changes, the background-only fits remained highly stable throughout the elution, with $R^2_{\text{start}} = 0.99934$–$0.99959$ at the beginning and $R^2_{\text{end}} = 0.99932$–$0.99961$ at the end of the gradient. These results demonstrate that gradient elution introduces systematic matrix drift even in protein-free spectra, confirming that explicit background treatment is required for mechanistic interpretation in the Amide I region. By isolating solvent- and buffer-related variation in dedicated components ($p0$–$p1$), the protein-associated components can be fitted under stable constraints without absorbing gradient-driven covariance. This approach directly addresses a known limitation of purely data-driven models, which often entangle concentration effects and matrix variation and therefore require extensive recalibration when process conditions change.

5.3.1.2 Model definition

The measured Amide I spectrum $y(\nu)$ was modeled as a linear superposition of pseudo-Voigt components and a residual term,

$$y(\nu) = \sum_{i=1}^{N} A_i \cdot PV(\nu; \nu_i, \sigma_i, \eta_i) + \varepsilon(\nu), \tag{5.2}$$

where A_i denotes the component amplitude, ν_i the center position, σ_i the width parameter, and η_i the Lorentzian fraction of component i. The pseudo-Voigt function was selected to flexibly capture intermediate peak shapes between the Gaussian and Lorentzian limits while maintaining interpretable parameters. Solvent-related contributions were addressed by the dedicated background model and were not absorbed into the protein component parameters during protein deconvolution.

5.3.1.3 Parameter initialization and constraints

Constrained optimization necessitates stable initialization and physically motivated bounds to ensure consistent component identity across spectra and to prevent non-physical solutions. Initial peak center estimates were systematically and automatically obtained from a Mexican-hat wavelet analysis applied to smoothed second-derivative spectra, thereby eliminating manual trial-and-error initialization. The Mexican hat wavelet was selected due to its sensitivity to localized curvature changes and its analytical relationship to the derivatives of Gaussian functions, which enables robust peak localization in overlapping band regions. The wavelet-derived centers were used exclusively as initialization points and to define narrow center boundaries; wavelet-transformed spectra were not subjected to fitting or quantitative interpretation. Peak center positions ν_i were constrained within component-specific bounds aligned with known vibrational assignments and empirically observed peak neighborhoods in the Amide I region. These constraints prevent peak swapping and enforce consistent component identity across fractions and experiments, which is essential for interpreting parameter trajectories as mechanistic indicators. All amplitudes were constrained to non-negative values ($A_i \geq 0$) to enforce physically meaningful contributions. No upper bounds were imposed, allowing the amplitudes to reflect concentration-driven changes during chromatographic elution. The width parameters σ_i were constrained to physically plausible ranges to prevent overfitting of

noise through artificially narrow peaks or compensatory broadening that could mask the underlying spectral structure. These width constraints were kept constant across all spectra to maintain comparability of the fitted components and to limit the effective degrees of freedom.

5.3.1.4 Optimization strategy

Parameter estimation was performed hierarchically to decouple solvent-driven background variability from protein-related contributions and to stabilize the spectral deconvolution under process conditions. First, a parametric background model was established as described in Section 5.3.1.1. For protein-containing spectra, the background parameters were not jointly optimized with the protein parameters. Instead, the background contribution was applied as a fixed or partially constrained term depending on the selected salt-effect strategy, thereby preventing matrix-driven variance from being absorbed into the protein components. Subsequently, spectra from early elution fractions containing exclusively monomeric protein species were used to calibrate a reference protein model (cf. Figure 5.3A). During this step, the background contribution was accounted for using the background model, while the protein peak centers and widths were optimized and subsequently fixed. Only the amplitudes of the protein components were allowed to vary across spectra. This calibration establishes a reference monomer state that anchors the component interpretation and minimizes parameter drift across the chromatographic run. For spectra containing aggregated species, the calibrated monomer parameters were retained to preserve structural consistency and ensure that amplitude changes remain interpretable as concentration effects. Additional pseudo-Voigt components associated with aggregation-related conformational changes were introduced while maintaining the constrained background treatment (cf. Figure 5.3B). This approach separates concentration-driven amplitude variation from aggregation-specific changes in band shape and shoulder contributions. Finally, all model parameters were estimated on a per-spectrum basis using constrained nonlinear least-squares optimization. Bounds and fixed-parameter settings were enforced consistently across all spectra to ensure that the fitted parameters and derived markers remain comparable across time, fractions, and experiments. This strategy reduces parameter cross-correlation and limits compensatory solutions between the background and protein components.

5.3.1.5 Model validation and predictive evaluation

Model performance was evaluated using complementary criteria that assess both numerical fit quality and the predictive relevance of the extracted hard-model parameters. First, the quality of the spectral fits was examined using several diagnostic metrics. The coefficient of determination (R^2) was used as the primary measure of explained variance within the fitted spectral window, while residual spectra were inspected to identify systematic deviations such as unmodeled shoulders or baseline curvature that would indicate insufficient model complexity or inappropriate parameter constraints. In addition, reduced χ^2 values were used to evaluate fit quality relative to the expected noise level and the number of fitted parameters, providing a complementary criterion to R^2 for detecting potential overfitting or underfitting. Reduced χ^2 values close to unity indicate that the residuals are consistent with the expected noise magnitude. To balance fit accuracy and model parsimony, model complexity was further evaluated using the Akaike information criterion (AIC). The AIC was used to compare candidate model configurations and to ensure that additional components were justified by improved explanatory power rather than by overparameterization. Because AEX elution introduces pronounced concentration dynamics across the chromatographic profile, these fit diagnostics were

evaluated across the entire elution trajectory, including low-signal front and tail fractions. This approach provides a realistic estimate of model behavior under process analytical technology (PAT)-relevant conditions, where early detection of co-elution or tailing effects can be more critical than the accurate fitting of high-intensity apex fractions. Beyond spectral fit quality, the predictive relevance of the parameters extracted from the hard modeling framework was evaluated using a grouped cross-validation strategy. This analysis aimed to determine whether hard-model-derived features contain transferable information regarding protein concentration and aggregation-related critical quality attributes across independent chromatographic runs. For this purpose, experiment-wise hard-model outputs were loaded from stored result dictionaries containing the fitted parameter tables, residual spectra, and reconstructed Amide I signals. Derived spectral descriptors, including the Amide I center of mass and the $1654/1610\,\mathrm{cm}^{-1}$ intensity ratio, were used together with the amplitudes of all fitted pseudo-Voigt components as candidate predictor variables. Corresponding reference values for monomer, aggregate, and total protein concentrations were obtained from the associated result structures to compute the aggregate fraction. Samples with total protein concentrations below 0.01 g/L were excluded from the regression analysis to ensure a meaningful protein signal contribution. Predictive performance was assessed using a leave-one-experiment-out cross-validation scheme. This grouping strategy prevents overly optimistic performance estimates that could arise from randomly splitting highly correlated neighboring fractions originating from the same chromatographic run. Consequently, the resulting performance reflects the ability of hard-model-derived features to generalize to previously unseen experiments. Within each cross-validation fold, the selected features were standardized using a z-score transformation fitted exclusively on the training data. Regression models were trained using ridge regression with internal selection of the regularization strength, providing a transparent linear baseline capable of handling correlated predictor variables while reducing coefficient instability. The evaluated feature set consisted of the amplitudes of all fitted pseudo-Voigt components and the Amide I center of mass. Separate regression models were calibrated for the prediction of monomer concentration, aggregate concentration, and aggregate fraction. Model performance was quantified for each fold using the coefficient of determination and the root-mean-square error on the corresponding test set. This evaluation provides a run-wise estimate of how effectively mechanistically derived hard-model parameters support quantitative prediction of protein size variants under previously unseen chromatographic conditions.

5.3.2 Dataset Coverage and Hard-Model Execution

The constrained hard modeling workflow was applied to all in-line Raman spectra acquired during the AEX elution phases and mapped to fraction-wise UHP-SEC labels. In total, 285 fraction-averaged spectra were evaluated across six AEX experiments, including 5 CVs, 10 CVs, and 15 CVs linear gradients, as well as step elution experiments originating from two independent batches. As determined by UHP-SEC reference analytics, the total protein concentration spanned a range of 0–79.40 mg/mL, thereby encompassing the full dynamic elution window from baseline fractions to peak maxima (Figure 5.4). The aggregate fraction ranged from 0 to 100 %, with extreme values predominantly observed in low-signal tail fractions. This broad coverage of concentration and aggregate concentrations ensures that model behavior is evaluated under realistic chromatographic conditions rather than only in apex fractions with favorable signal-to-noise ratios. Furthermore, the inclusion of low-signal and tail fractions is particularly relevant for PAT applications, as these regions are often critical for pooling and process decisions. The hard model was structured as a superposition of constrained pseudo-Voigt components within the Amide I window. Specifically, two

components ($p0$–$p1$) were assigned to solvent and background contributions, while five components ($p2$–$p6$) were utilized to describe the protein-associated Amide I envelope. This component structure was kept constant across all spectra, which is essential for maintaining parameter comparability and for interpreting amplitude changes mechanistically rather than as model reconfiguration artifacts.

5.3.3 Hard-Model Spectral Reconstruction Quality

The constrained pseudo-Voigt hard model captured the Amide I band shape with high fidelity across the complete chromatographic elution profiles. Across all evaluated spectra ($n = 285$), a global goodness-of-fit of $R^2 = 0.99890 \pm 0.00118$ (mean $\pm$ SD) was achieved, spanning a range from $R^2_{\text{min}} = 0.99548$ to $R^2_{\text{max}} = 0.99975$. Within the modeled 1500–1800 cm^{-1} window, the corresponding global reconstruction error amounted to RMSE $= 1.98 \times 10^{-3} \pm 1.17 \times 10^{-3}$. This uniformly high reconstruction fidelity demonstrates that the constrained hard model is sufficiently flexible to describe the Amide I envelope without requiring unconstrained overparameterization. Furthermore, a concentration-dependent trend in fit stability was observed. The lowest goodness-of-fit values occurred in very low-signal tail fractions, where the total protein concentration approached baseline levels and the Amide I contribution became comparable to residual noise. This slight decrease in fit quality in low-signal fractions is expected and reflects signal-to-noise limitations rather than systematic model failure. Restricting the evaluation to fractions with increased total protein concentration reduced the influence of near-baseline spectra and yielded consistently high fit quality across experiments, with experiment-wise R^2_{min} values ranging from 0.9959 to 0.9996. Inspection of the residual spectra revealed no persistent band-shaped deviations. Instead, the residuals were dominated by high-frequency noise without systematic structure. This absence of structured residuals indicates that the selected number of components and the imposed constraints were sufficient to describe the Amide I envelope under dynamically changing gradient conditions, thereby capturing genuine spectral information rather than compensating for missing features through non-physical parameter drift.

5.3.4 Resolved Protein Components and Peak Constraints

Across all experiments, the five protein-associated pseudo-Voigt components remained confined to narrow and physically plausible center ranges, thereby ensuring consistent component identity and comparability between fractions and batches. Upon pooling all spectra, the mean center positions and observed ranges were as follows: $p2$ at 1562.20 cm^{-1} (1559.00–1569.00 cm^{-1}), $p3$ at 1605.16 cm^{-1} (1599.00–1609.00 cm^{-1}), $p4$ at 1623.55 cm^{-1} (1615.67–1625.00 cm^{-1}), $p5$ at 1654.05 cm^{-1} (1651.00–1661.00 cm^{-1}), and $p6$ at 1679.51 cm^{-1} (1672.00–1682.00 cm^{-1}). Such stable center positions across experiments are a central prerequisite for interpreting the fitted components as physically meaningful and comparable descriptors rather than purely numerical basis functions. Furthermore, no systematic peak swapping or boundary violations were observed across all chromatographic runs. For all spectra, the fitted center positions remained within the predefined constraint intervals, and inter-experimental variability was limited to small shifts within these bounds. Consequently, the resulting parameter distributions were unimodal for each component, showing no evidence of bimodal behavior or experiment-specific clustering (Figure 5.6). The absence of peak swapping or boundary accumulation confirms that the chosen initialization and constraints are sufficiently informative to stabilize the optimization. Together with the high fit quality, these findings support the use of amplitude trajectories and derived markers as mechanistically

interpretable outputs of the hard model. Given the strictly linear nature of the NaCl gradients (0 to 500 mM NaCl), the elution buffer composition can be tracked along the adjusted time axis. Monomer elution occurred predominantly between 150 and 280 mM NaCl, while the HMWC fraction enriched in the tail regions up to 400 mM NaCl before entering the high-salt column strip.

5.3.5 Quantitative Validation of Hard-Model-Derived Spectral Features

To evaluate aggregation sensitivity across the full chromatographic concentration range, including low-signal front fractions, apex fractions, and aggregate-enriched tail fractions, spectral markers were extracted from the reconstructed protein-only signal following the removal of background contributions. Across the complete dataset ($n = 285$), aggregation-dependent trends were detectable throughout the elution profile. However, marker variance increased in low-signal fractions where absolute Raman intensities approached the detector noise level. In these regions, the Amide I contribution becomes comparable to residual noise, affecting the numerical stability of intensity-based descriptors and increasing the dispersion of center-of-mass (CoM) estimates. Despite this increased variability, systematic aggregation-related trends remained observable across the entire dataset.

Across all fractions, the CoM exhibited a positive monotonic relationship with increasing aggregate fraction. For fractions with $v_{\mathrm{agg}} < 0.05$, the CoM amounted to 1643.32 ± 2.31 cm^{-1} ($n = 85$). In the intermediate aggregation range of $0.20 \leq v_{\mathrm{agg}} < 0.40$, the CoM increased to 1647.56 ± 3.96 cm^{-1} ($n = 43$). For higher aggregate fractions of $0.40 \leq v_{\mathrm{agg}} < 0.80$, the CoM reached 1651.29 ± 5.45 cm^{-1} ($n = 49$). These systematic shifts indicate that aggregation-related structural changes within the Amide I envelope are captured by the hard-model decomposition across the entire chromatographic concentration regime. Consequently, the CoM emerged as the most robust aggregation-sensitive descriptor because it integrates the full Amide I envelope and is therefore less susceptible to local noise contributions than single-point or ratio-based markers.

In addition to the CoM, a ratiometric marker defined as I_{1654}/I_{1610} was calculated from the reconstructed protein signal. Across the complete dataset, this ratio exhibited larger variance than the CoM due to numerical instability when the denominator intensity approached zero in low-signal fractions. In particular, when the intensity at 1610 cm^{-1} approached the noise floor, inflated ratio values and heavy-tailed distributions were observed. The instability of the I_{1654}/I_{1610} ratio in low-signal fractions highlights an important limitation of pointwise markers for online process decisions, especially when denominator intensities approach the noise floor. Nevertheless, the ratiometric marker also showed a positive association with aggregate fraction, increasing on average from 3.65 for $v_{\mathrm{agg}} < 0.05$ to 8.40 for $0.40 \leq v_{\mathrm{agg}} < 0.80$. Compared to the CoM, the ratio remained more sensitive to noise contributions and component cross-correlation effects, particularly in intermediate concentration fractions.

To assess whether the parameters extracted from the hard-model decomposition contain predictive information regarding protein concentration and aggregation-related attributes, the amplitudes of all fitted pseudo-Voigt components together with the Amide I CoM were used as features in a multivariate regression analysis. Using a ridge regression model trained on the reconstructed spectral features, the apparent predictive performance across all gradient experiments yielded a coefficient of determination of $R^2 = 0.979$ for monomer concentration and $R^2 = 0.557$ for HMWC concentration (Figure 5.5). Absolute concentration units (g/L) are intentionally retained throughout the analysis to ensure mathematical stability across the dynamic elution profile. In front and tail fractions where

total protein concentration approaches baseline levels, relative percentage metrics (%) become highly sensitive to minor noise floor fluctuations, whereas absolute mass concentration (g/L) provides an un-skewed representation of the actual mass balance. The stronger apparent predictive performance for monomer concentration than for HMWC concentration is consistent with the dominance of the monomer contribution in the chromatographic dataset and the lower abundance of aggregate-rich fractions. To contextualize these quantitative outcomes, a rigorous comparison with the current literature status quo for Raman-based high molecular weight component (HMWC) and protein purity tracking is warranted. The absolute cross-validation test error of 1.587 g/L achieved in this study translates to a relative error range of 2.0% to 4.5% across the peak fractions, where total protein concentrations reach up to 79.40 mg/mL. While conventional partial least squares (PLS) models optimized for stable, matrix-matched monoclonal antibody (mAb) formulations or steady-state bioprocesses can achieve lower cross-validation errors—with values frequently reported in narrow bounds such as 0.046–0.801% [**massei2026use**], 0.037–0.089% [**pan2024pioneering**], 0.15–0.52% [80], or 0.12–0.65% [81]—these purely data-driven baselines are structurally constrained to unvarying chemical backgrounds. In such steady-state literature studies, the solvent matrix does not undergo pronounced dynamic shifts that could mimic or confound genuine structural protein vibrations. In contrast, the presented hard-modeling framework successfully handles massive gradient matrix variations across a 0 to 500 mM NaCl elution window without requiring product-specific, extensive calibration sets or large labeled training libraries.

Consequently, the multivariate regression model for HMWC serves primarily as a qualitative screening tool, whereas precise tracking of aggregation kinetics is achieved via the more robust Center of Mass (CoM) descriptor.

A systematic under-prediction of HMWC is observed primarily in the late tail fractions where total protein concentrations reach approximately 10 g/L and the aggregate ratio equals or exceeds 50%. Because the baseline protein hard model was mathematically calibrated using early fractions containing purely monomeric species, rapid non-linear multi-component shape deformations within these extreme aggregate-rich zones induce minor cross-correlation errors between adjacent protein components ($p4, p5,$ and $p6$).

Overall, the analysis reveals a concentration-dependent hierarchy in marker robustness. The CoM represents a comparatively stable descriptor of aggregation-related spectral changes across the chromatographic profile, while the ratiometric marker exhibits higher sensitivity to noise-driven denominator fluctuations. Collectively, these findings indicate that hard-model-derived parameters consistently reflect aggregation-induced spectral shifts within the reconstructed protein signal, demonstrating their capacity to serve both as mechanistically interpretable markers and as structured feature sets for downstream quantitative models. For practical bioprocess implementations, such as automated peak-cutting operations, several caveats must be considered. Due to the high test uncertainty associated with the HMWC concentration regression model ($R^2_{\text{test}} = 0.285$), direct utilization of the absolute predicted mass values (g/L) for fraction pooling decisions is not recommended. Instead, industrial implementation must rely on tracking the integrated Amide I Center of Mass (CoM) descriptor. This parameter operates as a robust, noise-insensitive indicator of structural shape deformation. Fraction cutting should be executed by triggering pooling boundaries when the continuous CoM trajectory exceeds an empirical baseline threshold established during the pure monomer elution phase.

5.4 Conclusion

The implementation of the constrained pseudo-Voigt hard modeling framework successfully integrates mechanistic prior knowledge, enabling a highly interpretable spectral decomposition that effectively separates solvent and background contributions from protein-associated Amide I components during gradient elution. By relying on physically constrained spectral components rather than instrument-specific latent variables, the workflow offers a robust, calibration-light alternative to classical partial least squares (PLS) models, thereby reducing the dependence on large labeled datasets and facilitating transferability across different process conditions. Furthermore, explicit background modeling accurately captured systematic salt-induced spectral drift, ensuring that gradient-driven matrix covariance was successfully isolated and prevented from being erroneously absorbed into the protein-associated components. The mechanistically derived spectral features demonstrated exceptional stability and strong apparent predictive performance for monomer concentration ($R^2 = 0.979$), confirming that the structured feature sets capture the dominant variance of the dataset.

High molecular weight components (HMWC) were captured with a training R^2 of 0.557 and a leave-one-out cross-validation test R^2 of 0.285. While this validation score is low for strict quantitative concentration regression, the performance is evaluated as acceptable and passable exclusively from a process analytical technology (PAT) screening perspective, where continuous trend tracking is prioritized over absolute mass-balance precision.

Finally, the Amide I center of mass, extracted from the reconstructed protein signal, emerged as a highly robust, aggregation-sensitive descriptor that preserves structural information despite strong concentration dynamics. This circumvents the noise-driven instabilities of pointwise markers and serves as a promising candidate for online pooling decisions. While the current study focuses on BSA as a model protein, the framework is methodologically transferable to other complex systems, such as monoclonal antibodies, by adjusting the initial peak constraints to the respective native secondary structure.

Author contribution

Conceptualization, J.H.-M.; methodology, J.H.-M. and R.S.; validation, J.H.-M.; formal analysis, J.H.-M.; investigation, J.H.-M.; resources, L.R., M.S., and J.H.; data curation, J.H.-M.; writing—original draft preparation, J.H.-M.; writing—review and editing, all authors; visualization, J.H.-M.; supervision, L.R., M.S., and J.H.; project administration, J.H.; funding acquisition, J.H. All authors have read and approved the final manuscript.

5.5 Tables

Table 5.1 Experimental conditions for bind-elute AEX chromatography experiment used in in-line Raman monitoring of BSA monomer and aggregate separation.

Run ID	Gradient length *(CVs)*	Load density *(mg/mL resin)*	Total feed concentration *(mg/mL)*	HMWC content *(%)*
Batch 1 5CV	5	15	15	25.33
Batch 1 15CV	15	22.5	15	25.33
Batch 1 10CV	10	17.5	15	25.33
Batch 2 5CV	5	22.5	15	15.67
Batch 2 10CV	10	30	15	15.67
Batch 2 15CV	15	45	15	15.67

5.6 Figures

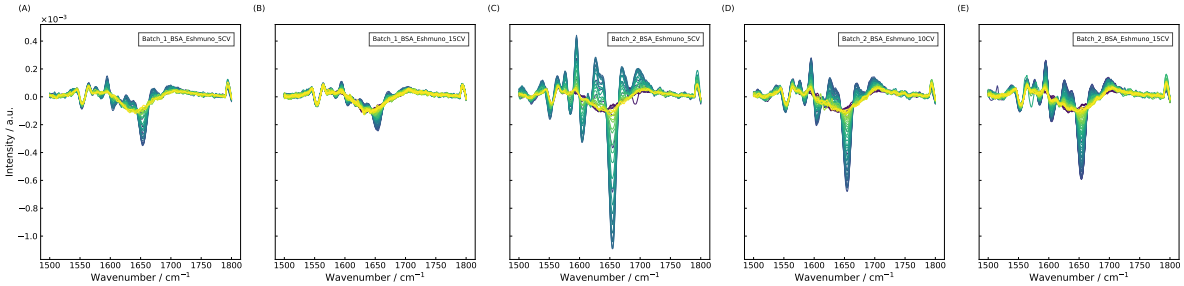


Figure 5.1 Comparison of second-derivative Raman spectra in the Amide I region across all chromatographic experiments. Subplots show the spectral variations corresponding to all chromatographic runs in a wavenumber range of 1500 to 1800 cm^{-1}.

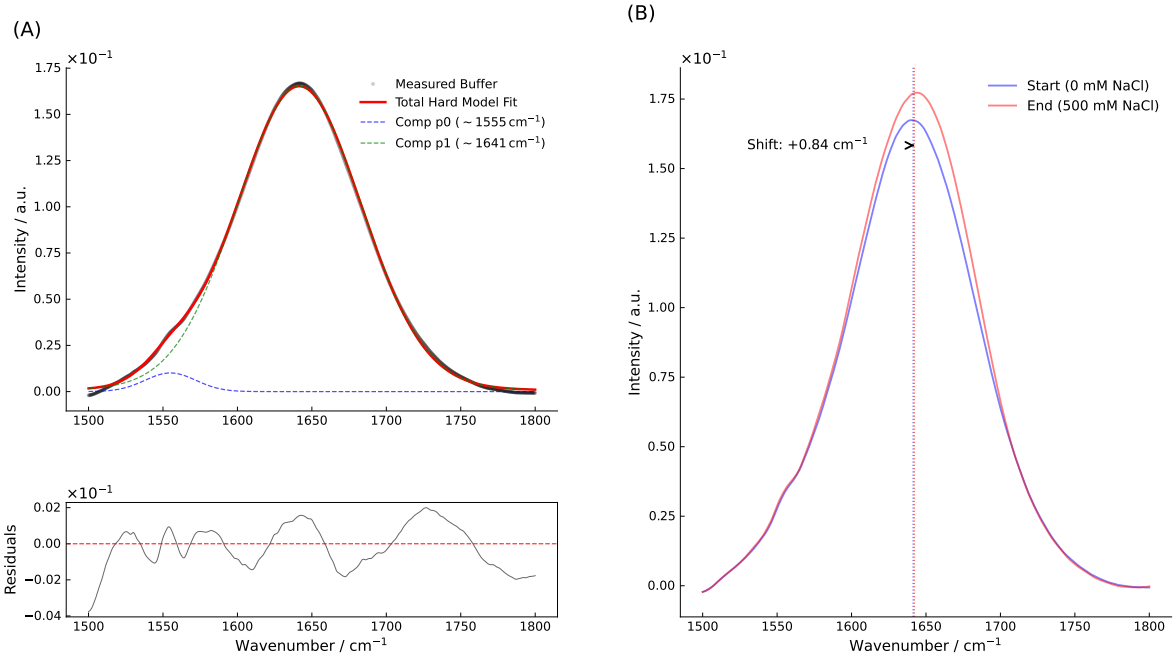


Figure 5.2 Background modeling and salt-induced spectral drift during gradient elution in the Amide I region. (A) Measured buffer spectrum overlaid with total hard model fit, demonstrating underlying components p0 (1555 cm^{-1}) and p1 (1641 cm^{-1}), alongside fitting residuals. (B) Spectral drift occurring between start (0 mM NaCl) and the end condition (500 mM NaCl).

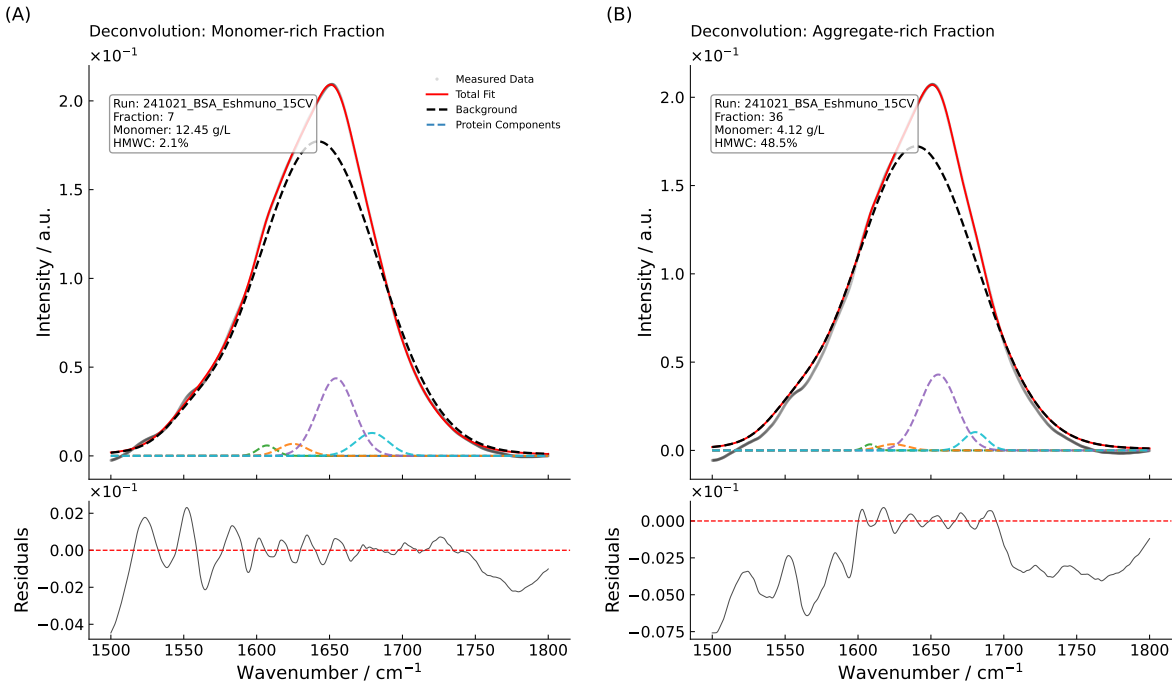


Figure 5.3 Representative constrained pseudo-Voigt deconvolution of Amide I spectra. (A) Deconvolution of a monomer-rich fraction yielding 12.45 g/L monomer and 2.1% HMWC. (B) Deconvolution of an aggregate-rich fraction yielding 4.12 g/L monomer and 48.5% HMWC. Both panels plot the measured data, total fit, background, individual protein components, and model residuals.

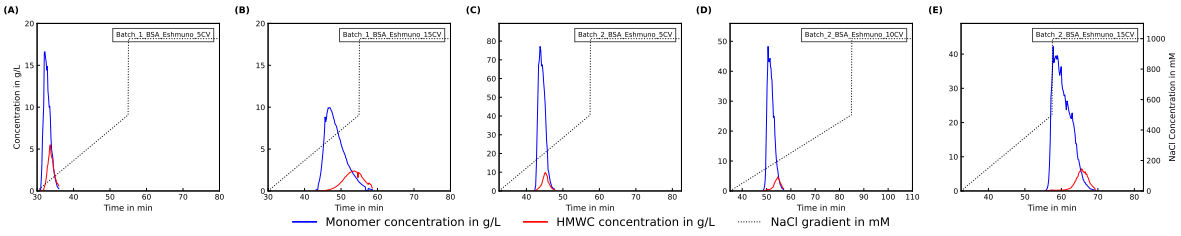


Figure 5.4 Comparison of monomer and HMWC elution profiles across the investigated AEX experiments. Blue traces represent monomer concentration, red traces represent HMWC concentration in g/L over time. Panels depict different gradient lengths and batch conditions

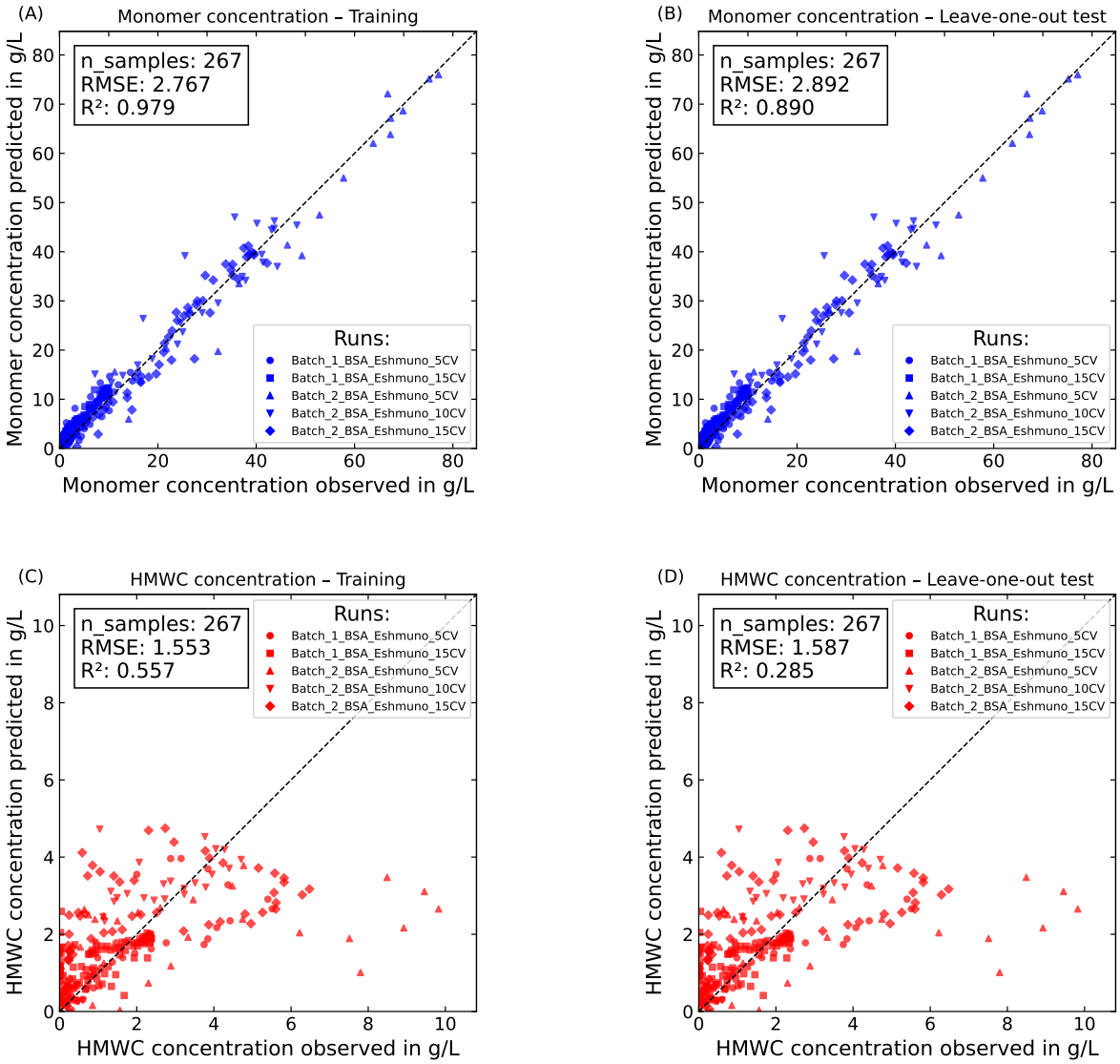


Figure 5.5 Model validation for monomer and HMWC prediction based on hard-model-derived spectral features across n=267 samples. (A) Monomer concentration prediction on the training set, (B) Monomer concentration prediction using a leave-one-out test set, (C) HMWC concentration prediction on the training set, (D) HMWC concentration prediction using a leave-one-out test set.

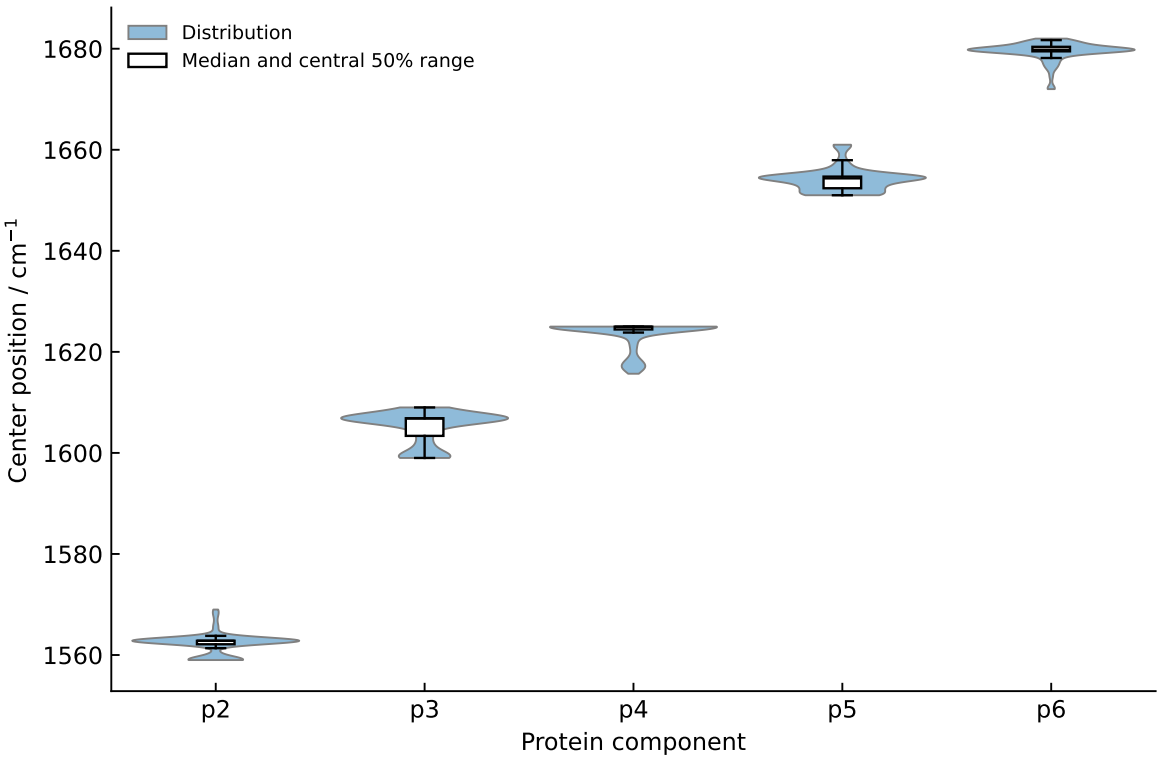


Figure 5.6 Distribution of fitted center positions (cm^{-1}) for the protein-associated pseudo-Voigt components across all spectra. The violin plots (blue) depict the underlying probability distribution for components p2 through p6, the inner box plots denote the median and central 50 % range.

Chapter References

[9] C. Zhou, W. Qi, E. N. Lewis, and J. F. Carpenter. „Concomitant Raman spectroscopy and dynamic light scattering for characterization of therapeutic proteins at high concentrations“. In: *Analytical Biochemistry* 472 (2015), pp. 7–20. DOI: 10.1016/j.ab.2014.11.016.

[32] A. Dietrich, R. Schiemer, J. Kurmann, S. Zhang, and J. Hubbuch. „Raman-based PAT for VLP precipitation: systematic data diversification and preprocessing pipeline identification“. In: *Frontiers in Bioengineering and Biotechnology* 12 (2024). DOI: 10.3389/fbioe.2024.1399938.

[33] J. Heyer-Müller, R. Schiemer, L. Robbel, M. Schmitt, and J. Hubbuch. „Development of Raman spectroscopy and machine learning methods for protein aggregate quantification: Application to BSA in chromatographic processes“. en. In: *Biotechnology and Bioengineering* 123.5 (2026), e70163. DOI: 10.1002/bit.70163.

[73] M. Vázquez-Rey and D. A. Lang. „Aggregates in monoclonal antibody manufacturing processes“. In: *Biotechnology and Bioengineering* 108.7 (2011), pp. 1494–1508. DOI: 10.1002/bit.23155.

[74] C. J. Roberts. „Protein aggregation and its impact on product quality“. In: *Current Opinion in Biotechnology* 30 (2014), pp. 211–217. DOI: 10.1016/j.copbio.2014.08.001.

[75] N. B. Pham and W. S. Meng. „Protein aggregation and immunogenicity of biotherapeutics“. In: *International Journal of Pharmaceutics* 585 (2020), p. 119523. DOI: 10.1016/j.ijpharm.2020.119523.

[78] N. Brestrich, A. Sanden, A. Kraft, K. McCann, J. Bertolini, and J. Hubbuch. „Advances in inline quantification of co-eluting proteins in chromatography: Process-data-based model calibration and application towards real-life separation issues“. In: *Biotechnology and Bioengineering* 112.7 (2015), pp. 1406–1416. DOI: 10.1002/bit.25546.

[79] F. Capito, R. Skudas, H. Kolmar, and C. Hunzinger. „At-line mid infrared spectroscopy for monitoring downstream processing unit operations“. In: *Process Biochemistry* 50.6 (2015), pp. 997–1005. DOI: 10.1016/j.procbio.2015.03.005.

[80] B. Wei et al. „Multi-attribute Raman spectroscopy (MARS) for monitoring product quality attributes in formulated monoclonal antibody therapeutics“. In: *mAbs* 14.1 (2022). DOI: 10.1080/19420862.2021.2007564.

[81] J. Chen, J. Wang, R. Hess, G. Wang, J. Studts, and M. Franzreb. „Application of Raman spectroscopy during pharmaceutical process development for determination of critical quality attributes in Protein A chromatography“. In: *Journal of Chromatography A* 1718 (2024), p. 464721. DOI: 10.1016/j.chroma.2024.464721.

[82] M. Rüdt, P. Vormittag, N. Hillebrandt, and J. Hubbuch. „Process monitoring of virus-like particle reassembly by diafiltration with UVVis spectroscopy and light scattering“. In: *Biotechnology and Bioengineering* 116 (2019), pp. 1366–1379. DOI: 10.1002/bit.26935.

[83] R. Schiemer, J. T. Weggen, K. M. Schmitt, and J. Hubbuch. „An adaptive soft-sensor for advanced real-time monitoring of an antibody-drug conjugation reaction“. In: *Biotechnology and Bioengineering* 120.7 (2023), pp. 1914–1928. DOI: 10.1002/bit.28428.

[86] I. Ettah and L. Ashton. „Engaging with Raman Spectroscopy to Investigate Antibody Aggregation“. In: *Antibodies* 7.3 (2018), p. 24. DOI: 10.3390/antib7030024.

[87] R. G. de la Cuesta, R. Goodacre, and L. Ashton. „Monitoring Antibody Aggregation in Early Drug Development Using Raman Spectroscopy and Perturbation-Correlation Moving Windows“. In: *Analytical Chemistry* 86.22 (2014), pp. 11133–11140. DOI: 10.1021/ac5038329.

[88] E. N. Lewis, W. Qi, L. H. Kidder, S. Amin, S. M. Kenyon, and S. Blake. „Combined Dynamic Light Scattering and Raman Spectroscopy Approach for Characterizing the Aggregation of Therapeutic Proteins“. In: *Molecules* 19.12 (2014), pp. 20888–20905. DOI: 10.3390/molecules191220888.

[89] G. V. Barnett et al. „Structural Changes and Aggregation Mechanisms for Anti-Streptavidin IgG1 at Elevated Concentration“. In: *The Journal of Physical Chemistry B* 119.49 (2015), pp. 15150–15163. DOI: 10.1021/acs.jpcb.5b08748.

[97] K. Furić, I. Ciglenečki, and B. Ćosović. „Raman spectroscopic study of sodium chloride water solutions“. In: *Journal of Molecular Structure* 550–551 (2000), pp. 225–234. DOI: 10.1016/S0022-2860(00)00388-4.

[98] D. Erb. *pybaselines: A Python library of algorithms for the baseline correction of experimental data*. DOI: 10.5281/zenodo.5608581.

[117] D. H. Müller, C. Flake, T. Brands, and H.-J. Koß. „Bioprocess in-line monitoring using Raman spectroscopy and Indirect Hard Modeling (IHM): A simple calibration yields a robust model“. en. In: *Biotechnology and Bioengineering* 120.7 (2023). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bit.28424, pp. 1857–1868. DOI: 10.1002/bit.28424.

[118] M. Melcher et al. „The potential of random forest and neural networks for biomass and recombinant protein modeling in Escherichia coli fed-batch fermentations“. In: *Biotechnology Journal* 10.11 (2015), pp. 1770–1782. DOI: https://doi.org/10.1002/biot.201400790.

[119] E. Kriesten, D. Mayer, F. Alsmeyer, C. B. Minnich, L. Greiner, and W. Marquardt. „Identification of unknown pure component spectra by indirect hard modeling“. In: *Chemometrics and Intelligent Laboratory Systems* 93.2 (2008), pp. 108–119. DOI: https://doi.org/10.1016/j.chemolab.2008.05.002.

[120] M. N. Siamwiza et al. „Interpretation of the doublet at 850 and 830 cm-1 in the Raman spectra of tyrosyl residues in proteins and certain model compounds“. In: *Biochemistry* 14.22 (1975), pp. 4870–4876. DOI: 10.1021/bi00693a014.

[121] R. Carmona, W.-L. Hwang, and B. Torresani. *Practical Time-Frequency Analysis: Gabor and wavelet transforms, with an implementation in S*. Vol. 9. Academic Press, 1998.

[122] P. Du, W. A. Kibbe, and S. M. Lin. „Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching“. In: *Bioinformatics* 22.17 (2006). _eprint: https://academic.oup.com/bioinformatics/article-pdf/22/17/2059/48840388/bioinformatics_22_17_2059.pdf, pp. 2059–2065. DOI: 10.1093/bioinformatics/btl355.

6

General Discussion and Conclusion

As outlined in the Introduction, biopharmaceutical process development for therapeutic proteins is governed by the interplay between molecular structure and process conditions. In contrast to small-molecule systems, relevant product properties cannot be described by composition alone, but depend on the preservation and evolution of higher-order structure during manufacturing. Within this context, protein aggregation represents a central source of structural variability. Under process conditions, however, aggregation does not occur in isolation, but coincides with changes in protein conformation, concentration, and matrix composition. Process-derived analytical signals therefore reflect multiple superimposed contributions rather than a single target variable. This coupling of structural variation, matrix effects, and process dynamics was identified as a central analytical challenge that remains only partly resolved by established approaches.

The thesis further started from the observation that conventional offline reference methods remain indispensable for aggregate quantification, but are inherently limited in dynamic process environments. Sampling, time delay, and possible perturbation of fragile aggregation equilibria restrict their value for real-time process understanding. Spectroscopic methods, and Raman spectroscopy in particular, therefore represent a promising route toward process-near monitoring. At the same time, spectroscopy does not remove analytical complexity, but shifts the challenge from measurement to data interpretation. Raman spectra are information-rich, yet they are shaped not only by the target analyte, but also by overlapping background contributions and measurement-dependent variability. Chemometric modeling was therefore treated as the central means of translating complex spectral data into process-relevant information. The key unresolved gap was defined accordingly: although spectral interference, process-induced variability, and model selection had each been addressed to some extent, their combined effects on predictive performance and interpretability under realistic process conditions remained insufficiently understood.

Against this background, the three studies formed a coherent methodological progression.

The first study addressed an enabling problem: representative Raman datasets for calibration and validation are difficult to generate in a systematic, efficient, and process-relevant manner. Without this basis, subsequent work on more demanding applications would remain methodologically weak. The second study then moved from concentration monitoring to the selective quantification of structurally relevant protein size variants under chromatographic conditions, where overlapping contributions from aggregate content, total concentration, and matrix composition constitute a major challenge. The third study addressed the remaining limitation that successful predictive models do not necessarily provide physical insight when spectral overlap becomes severe. The thesis thus progressed from analytical implementation to predictive modeling and finally to mechanistic interpretation. This sequence directly reflects the problem structure developed in the Introduction: first, the need for process-near spectral acquisition under relevant conditions; second, the need to extract aggregation-relevant information despite process-induced variability; and third, the need to move beyond prediction toward interpretable and structurally meaningful signal decomposition.

The red thread of the dissertation therefore lies in the stepwise reduction of one central uncertainty: how Raman spectroscopy can be developed from a promising but analytically demanding sensor into a robust and interpretable process-analytical framework for therapeutic protein systems. Each chapter addressed one layer of this problem. Chapter 3 established the measurement and calibration basis, Chapter 4 examined the extraction of aggregation-relevant information in a more demanding application, and Chapter 5 investigated how physically informed modeling can address the limitations of purely data-driven approaches when interpretability becomes critical. Rather than constituting separate contributions, the dissertation shows that robust process analytics require all three elements: representative data generation, suitable predictive modeling, and mechanistic understanding of the modeled signal.

A first unresolved issue identified at the outset of the thesis concerned the practical implementation of Raman spectroscopy as a process analytical technology. Although Raman-based process analytical technology (PAT) had been discussed extensively in the literature, routine application remained constrained by the need to generate representative calibration and validation datasets under conditions resembling the intended process environment. This limitation was not merely technical; it directly affected model robustness and transferability, since models trained under poorly matched conditions are unlikely to perform reliably in process-near applications. Chapter 3 addressed this gap by introducing a novel assembly that combined a chromatography system, an autosampler, and a Raman spectrometer for automated spectral acquisition and chemometric model calibration. Synthetic mixtures of phosphate buffer, citrate buffer, and lysozyme were generated through statistical designs and used to calibrate and validate partial least squares (PLS) models, which were subsequently transferred to in-line ultra-/diafiltration (UF/DF) data. The central conclusion of this study was that automated off-line generation of Raman calibration data can be aligned with in-line process measurement and can provide a practical route toward robust PAT implementation.

At the dissertation level, the significance of Chapter 3 lies in the fact that it solved an enabling methodological problem rather than the final analytical one. It did not yet address selective aggregate quantification, but established the analytical and technical basis for the subsequent studies. The chapter reduced uncertainty as to whether Raman-based process analytics can be implemented in a reproducible and practically deployable manner. This distinction is important, because the thesis did not begin with the most complex application case, but first established a controlled and transferable experimental foundation. The broader contribution of Chapter 3

therefore lies in demonstrating that process-near Raman implementation can be embedded in a structured workflow for systematic data generation, chemometric calibration, and model transfer. In this way, the chapter addressed one component of the gap identified in the Introduction: the need to move from promising spectroscopic concepts toward practically usable analytical frameworks under realistic bioprocess conditions.

Once the basic feasibility of automated Raman-based process analytics had been established, the dissertation moved to the central analytical challenge: the extraction of aggregation-relevant information from complex process spectra in which target signals are superimposed with concentration effects and matrix variability. Chapter 4 addressed this problem in chromatographic environments, which are particularly demanding because changing buffer composition, elution conditions, and protein concentration all affect spectral appearance. The unresolved issue was therefore not simply whether Raman spectroscopy is sensitive to aggregates, but whether aggregate-sensitive information can be separated from confounding factors under process-relevant conditions. To address this, bovine serum albumin (BSA) was used as a model system in forced degradation and statistically designed mixing studies that allowed aggregate content and total concentration to be decoupled. On this basis, aggregation-sensitive spectral markers were identified and then evaluated in bind-elute anion-exchange chromatography (AEX) experiments. In parallel, multiple chemometric models were established and compared, including PLS regression and convolutional neural networks (CNNs). The study concluded that Raman spectroscopy, when combined with systematically developed machine learning models, can support real-time prediction of monomer and aggregate concentrations in chromatographic processes. At the same time, it showed that successful process application depends not only on model accuracy in controlled datasets, but also on whether the selected modeling strategy remains meaningful under changing process conditions and overlapping spectral variability.

This conclusion is central for the dissertation because it moves the argument beyond feasibility toward selective analytical relevance. Chapter 4 showed that the key challenge is not simply to build a predictive model, but to ensure that the model captures aggregation-relevant variation rather than process-specific coincidence. In this sense, the chapter directly addressed the concern raised in the Introduction that models may learn correlations specific to the calibration dataset that do not generalize when background composition or process dynamics change. By comparing linear and nonlinear models, the chapter also clarified that model choice is not only a question of predictive flexibility, but of balancing data demands, robustness, and interpretability. What was resolved here was therefore not the general problem of aggregate monitoring in all settings, but a more specific and important one: the study demonstrated that selective, process-relevant prediction is possible, while also revealing the limits of using predictive performance alone as evidence of analytical validity. The chapter thus sharpened the overall thesis argument by showing that data-driven success must be interpreted in the context of process variability and model behavior.

A remaining limitation after Chapter 4 was that predictive performance did not necessarily provide physical insight into the structural basis of the measured Raman signal. This limitation had already been introduced in the Introduction, where different modeling philosophies were discussed in terms of their trade-offs between predictive performance and physical insight. Chapter 5 addressed this issue by developing a constrained pseudo-Voigt hard-modeling framework for the mechanistic deconvolution of in-line Raman spectra acquired during AEX of BSA monomers and aggregates. A dedicated background model was introduced to describe salt-induced drift in the amide I region and to separate gradient-driven matrix effects from protein-specific contributions. The model was subsequently

applied to fraction-resolved chromatographic datasets, from which physically interpretable spectral features, in particular the amide I center of mass, were extracted and quantitatively validated against ultrahigh performance - size exclusion chromatography (UHP-SEC) reference analytics. The chapter concluded that hard modeling can bridge predictive analysis and structural interpretation by explicitly resolving protein-related spectral components under process-relevant conditions, and that physically informed Raman models can serve as interpretable and calibration-light complements to purely data-driven approaches.

At the level of the dissertation as a whole, Chapter 5 addressed a problem that Chapter 3 and Chapter 4 had deliberately left open. If Chapter 3 asked how Raman data can be acquired and calibrated systematically, and Chapter 4 asked how aggregation-relevant quantities can be predicted, then Chapter 5 asked what those predictive relationships mean in spectral and structural terms. The chapter therefore represents more than a technical extension; it changes the analytical status of the framework. Instead of remaining dependent on latent-variable or nonlinear models whose internal structure is difficult to interpret, the work demonstrates that a physically constrained decomposition can provide descriptors that are both predictive and mechanistically meaningful. What was resolved here was not merely a fitting problem in the amide I region, but a more general limitation of process-near Raman analytics: the need to connect predictive output to interpretable spectral structure. This is relevant because it supports a stronger form of process understanding. The resulting insight is not only that a process state can be predicted, but also how structural and background-related spectral contributions are organized under dynamic chromatographic conditions.

Taken together, the three papers show that the central gaps identified were not due to a single missing method, but to a missing integration of experimental implementation, predictive modeling, and mechanistic interpretation under process-relevant conditions. The dissertation addressed this gap in a stepwise and coherent manner. First, it established that Raman-based PAT implementation can be embedded in a practical workflow for automated spectral acquisition and representative calibration. Second, it showed that aggregation-relevant information can be extracted from chromatographic Raman data, provided that model development is systematic and that predictive success is evaluated in light of process-induced variability. Third, it demonstrated that physically informed hard modeling can provide an interpretative layer that complements and extends purely data-driven methods. In this way, the dissertation solved three connected problems: it reduced the technical barrier to representative Raman calibration, demonstrated selective prediction of protein size variants under challenging process conditions, and provided a route toward structurally meaningful interpretation of overlapping spectral signals.

A further contribution of the work lies in how it reframed the role of modeling in process analytics. The dissertation does not treat PLS, CNNs, and hard modeling as competing methods from which a single winner must emerge. Instead, it shows that they address different analytical tasks. PLS provides a robust and established baseline for multivariate prediction when calibration and application conditions are sufficiently aligned. CNNs offer greater flexibility in capturing complex dependencies, but their deployment is constrained by data requirements and limited transparency. Hard modeling, in contrast, does not primarily compete on predictive flexibility, but on mechanistic interpretability and structural relevance. The dissertation therefore contributes a more differentiated view of model selection: the decisive question is not only which model predicts best, but which model is appropriate for the type of variability, the available data regime, and the intended analytical

purpose. This directly addresses the Introduction's concern that the comparative suitability of different modeling strategies under realistic process conditions remained insufficiently understood.

Despite these advances, the dissertation does not eliminate all limitations associated with Raman-based process analytics for therapeutic proteins. First, the work remains centered on model systems and selected process contexts, in particular lysozyme-buffer mixtures in UF/DF and BSA during AEX. These systems were chosen deliberately because they enable controlled investigation of methodological questions, but broader transferability to other therapeutic proteins, process formats, or formulation environments remains to be demonstrated. Proteins with different structural sensitivities, aggregation pathways, or spectral signatures may impose additional demands on calibration design and model structure. Second, although Chapter 4 demonstrated that data-driven models can support selective prediction of monomer and aggregate concentrations, it also showed that predictive success depends strongly on whether the model remains meaningful under changing process conditions. This leaves open the question of how far model transfer can be generalized across instruments, scales, and process variants without loss of analytical validity. Third, although Chapter 5 established a hard-modeling framework with mechanistic interpretability, the approach depends on explicit assumptions regarding peak structure, constraints, and background behavior. Its robustness under broader spectral regimes, more heterogeneous systems, or less controlled process backgrounds remains an important topic for future work.

Another limitation concerns the relationship between prediction and causal interpretation. The Introduction had emphasized that one critical open question in chemometric modeling is the extent to which models reflect causal relationships rather than coincidental correlations. The dissertation made substantial progress on this point by identifying selective spectral markers, comparing model classes under process-relevant variability, and introducing a physically informed decomposition framework. Nevertheless, the broader question of causality in spectroscopy-based PAT remains unresolved. Even interpretable descriptors such as the amide I center of mass remain embedded in a model context and therefore require careful validation against orthogonal reference analytics. Accordingly, future work should not only continue to improve prediction accuracy, but also strengthen the linkage between spectroscopic descriptors, underlying structural states, and process-induced perturbations. This will be particularly important if Raman-based PAT is to support not only monitoring, but eventually more direct process control and mechanistically informed decision-making.

The outlook emerging from this dissertation is therefore not one of closure, but of a strengthened methodological basis for future process analytics. One promising direction lies in extending the automated Raman calibration concepts developed in Chapter 3 toward broader classes of representative training datasets, including more diverse process states, formulation conditions, and protein systems. Such extensions would improve the robustness of process-near chemometric models and could reduce one of the major bottlenecks of Raman-based PAT implementation. A second important direction concerns the systematic development of transfer strategies across instruments and process variants. Since process-induced variability and spectral background shifts remain central limiting factors, future studies should examine calibration transfer, domain adaptation, and hybrid modeling approaches to make predictive models more robust across realistic manufacturing scenarios. A third direction follows directly from Chapter 5: physically informed spectral decomposition should be extended beyond the amide I region and beyond the current model system to test how far mechanistic hard modeling can be generalized as an interpretable complement to statistical learning approaches.

More broadly, the dissertation suggests that future PAT strategies for therapeutic proteins should not be designed around prediction alone. Instead, the most relevant developments are likely to arise from deliberate integration of process context, representative data generation, chemometric robustness, and mechanistic interpretability. In this sense, the work contributes not only individual methods, but also a conceptual perspective on how Raman-based analytics can be advanced in biopharmaceutical downstream processing. The central achievement of the dissertation lies in showing that meaningful process-near spectroscopy requires more than signal acquisition and model fitting. It requires a structured understanding of what is measured, under which process conditions it is measured, how the signal is transformed into prediction, and to what extent that prediction can be interpreted in structural terms. By addressing these questions across three connected studies, the work provides a coherent basis for future PAT concepts that are not only technically feasible and predictively useful, but also analytically credible and mechanistically informative under realistic bioprocess conditions.

References

[1] A. S. Rathore and H. Winkle. „Quality by design for biopharmaceuticals“. In: *Nature Biotechnology* 27.1 (2009), pp. 26–34. DOI: 10.1038/nbt0109-26.

[2] P. Diederich, M. Hoffmann, and J. Hubbuch. „High-throughput process development of purification alternatives for the protein avidin“. In: *Biotechnology Progress* 31.4 (2015), pp. 957–973. DOI: 10.1002/btpr.2104.

[3] C. M. Dobson. „Protein folding and misfolding“. In: *Nature* 426 (2003), pp. 884–890. DOI: 10.1038/nature02261.

[4] W. Wang. „Protein aggregation and its inhibition in biopharmaceutics“. In: *International Journal of Pharmaceutics* 289 (2005), pp. 1–30. DOI: 10.1016/j.ijpharm.2004.11.014.

[5] E. Y. Chi, S. Krishnan, T. W. Randolph, and J. F. Carpenter. „Physical stability of proteins in aqueous solution: mechanism and driving forces in nonnative protein aggregation“. In: *Pharmaceutical Research* 20 (2003), pp. 1325–1336. DOI: https://doi.org/10.1023/A:1025771421906.

[6] A. Sanden, S. Haas, and J. Hubbuch. „Modifying an ÄKTApurifier system for the automated acquisition of samples for kinetic modeling of batch reactions“. In: *SLAS Technology* (2019). DOI: 10.1177/2472630319891976.

[7] M. C. Manning, D. K. Chou, B. M. Murphy, R. W. Payne, and D. S. Katayama. „Stability of protein pharmaceuticals: An update“. In: *Pharmaceutical Research* 27.4 (2010), pp. 544–575. DOI: 10.1007/s11095-009-0045-6.

[8] H. C. Mahler, W. Friess, U. Grauschopf, and S. Kiese. „Protein aggregation: pathways, induction factors and analysis“. In: *Journal of Pharmaceutical Sciences* 98 (2009), pp. 2909–2934. DOI: 10.1002/jps.21566.

[9] C. Zhou, W. Qi, E. N. Lewis, and J. F. Carpenter. „Concomitant Raman spectroscopy and dynamic light scattering for characterization of therapeutic proteins at high concentrations“. In: *Analytical Biochemistry* 472 (2015), pp. 7–20. DOI: 10.1016/j.ab.2014.11.016.

[10] M. Rüdt, S. Andris, R. Schiemer, and J. Hubbuch. „Factorization of preparative protein chromatograms with hard-constraint multivariate curve resolution and second-derivative pretreatment“. In: *Journal of Chromatography A* 1585.5 (2019), pp. 152–160. DOI: 10.1016/j.chroma.2018.11.065.

[11] Å. Rinnan, F. van den Berg, and S. B. Engelsen. „Review of the most common pre-processing techniques for near-infrared spectra“. In: *TrAC Trends in Analytical Chemistry* 28 (2009), pp. 1201–1222. DOI: 10.1016/j.trac.2009.07.007.

[12] „A review size-exclusion chromatography for the analysis of protein biotherapeutics and their aggregates“. In: *Journal of Liquid Chromatography Related Technologies* 35.20 (2012), pp. 2923–2950. DOI: 10.1080/10826076.2012.743724.

[13] S. K. Hansen, E. Skibsted, A. Staby, and J. Hubbuch. „A label-free methodology for selective protein quantification by means of absorption measurements“. In: *Biotechnology and Bioengineering* 108.11 (2011), pp. 2661–2669. DOI: 10.1002/bit.23229.

[14] Food and Drug Administration. *PAT — A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance*. 2004.

[15] M. Rüdt, N. Brestrich, L. Rolinger, and J. Hubbuch. „Real-time monitoring and control of the load phase of a protein A capture step“. In: *Biotechnology and Bioengineering* 114.2 (2017), pp. 368–373. DOI: 10.1002/bit.26078.

[16] J. Glassey et al. „Process analytical technology (PAT) for biopharmaceuticals“. In: *Biotechnology Journal* 6.4 (2011), pp. 369–377. DOI: 10.1002/biot.201000356.

[17] T. D. Beer et al. „Raman spectroscopy as a process analytical technology“. In: *TrAC Trends in Analytical Chemistry* 30 (2011), pp. 1073–1086. DOI: 10.1016/j.trac.2011.06.012.

[18] K. A. Esmonde-White, M. Cuellar, C. Uerpmann, B. Lenain, and I. R. Lewis. „Raman spectroscopy as a process analytical technology for pharmaceutical manufacturing and bioprocessing“. In: *Analytical and Bioanalytical Chemistry* 409.3 (2017), pp. 637–649. DOI: 10.1007/s00216-016-9824-1.

[19] L. Rolinger et al. „Multi-attribute PAT for UF/DF of proteins—monitoring concentration and buffer exchange by Raman spectroscopy“. In: *Analytical and Bioanalytical Chemistry* 412 (2020), pp. 2123–2134. DOI: 10.1007/s00216-019-02318-8.

[20] T. Næs, T. Isaksson, T. Fearn, and T. Davies. *A User-Friendly Guide to Multivariate Calibration and Classification*. NIR Publications, 2002. DOI: 10.1255/978-1-906715-25-0.

[21] L. e. a. Rolinger. „Multivariate data analysis in downstream processing“. In: *Analytical and Bioanalytical Chemistry* 412 (2020), pp. 2045–2059. DOI: 10.1007/s00216-020-02407-z.

[22] H. Martens and T. Næs. *Multivariate Calibration*. Wiley, 1989.

[23] G. Walsh. „Biopharmaceutical benchmarks 2018“. In: *Nature Biotechnology* 36 (2018), pp. 1136–1145. DOI: 10.1038/nbt.4305.

[24] J. F. Carpenter, T. W. Randolph, W. Jiskoot, et al. „Overlooking subvisible particles in therapeutic protein products“. In: *Journal of Pharmaceutical Sciences* 98 (2009), pp. 1201–1205. DOI: 10.1002/jps.21530.

[25] N. Räkel, K. Schleining, F. Dismer, and J. Hubbuch. „Prediction of protein retention in ion-exchange chromatography based on protein–protein interactions“. In: *Journal of Chromatography A* 1277 (2013), pp. 60–68. DOI: 10.1016/j.chroma.2013.03.077.

[26] P. Vormittag, T. Klamp, and J. Hubbuch. „Optimization of protein chromatography processes“. In: *Frontiers in Bioengineering and Biotechnology* 8 (2020), p. 402. DOI: 10.3389/fbioe.2020.0088.

[27] A. S. Rosenberg. „Effects of protein aggregates: An immunologic perspective“. In: *The AAPS Journal* 8.3 (2006), p. 59. DOI: 10.1208/aapsj080359.

[28] K. K. Arthur, J. P. Gabrielson, B. S. Kendrick, and M. R. Stoner. „Detection of protein aggregates by sedimentation velocity analytical ultracentrifugation (SV-AUC): sources of variability and their relative importance“. In: *Journal of Pharmaceutical Sciences* 98.10 (2009), pp. 3522–3539. DOI: 10.1002/jps.21654.

[29] J. Stetefeld, S. A. McKenna, and T. R. Patel. „Dynamic light scattering: a practical guide and applications in biomedical sciences“. In: *Biophysical Reviews* 8.4 (2016), pp. 409–427. DOI: 10.1007/s12551-016-0218-6.

[30] J. Lebowitz, M. S. Lewis, and P. Schuck. „Modern analytical ultracentrifugation in protein science: a tutorial review“. In: *Protein Science* 11.9 (2002), pp. 2067–2079. DOI: 10.1110/ps.0207702.

[31] R. L. McCreery. „Raman Spectroscopy for Chemical Analysis“. In: *Wiley* 1 (2000), pp. 1–420. DOI: 10.1002/0471238961.

[32] A. Dietrich, R. Schiemer, J. Kurmann, S. Zhang, and J. Hubbuch. „Raman-based PAT for VLP precipitation: systematic data diversification and preprocessing pipeline identification“. In: *Frontiers in Bioengineering and Biotechnology* 12 (2024). DOI: 10.3389/fbioe.2024.1399938.

[33] J. Heyer-Müller, R. Schiemer, L. Robbel, M. Schmitt, and J. Hubbuch. „Development of Raman spectroscopy and machine learning methods for protein aggregate quantification: Application to BSA in chromatographic processes“. en. In: *Biotechnology and Bioengineering* 123.5 (2026), e70163. DOI: 10.1002/bit.70163.

[34] A. Savitzky and M. J. E. Golay. „Smoothing and Differentiation of Data by Simplified Least Squares Procedures“. In: *Analytical Chemistry* 36.8 (1964), pp. 1627–1639. DOI: 10.1021/ac60214a047.

[35] S. Wold, M. Sjöström, and L. Eriksson. „PLS-regression: a basic tool of chemometrics“. In: *Chemometrics and Intelligent Laboratory Systems* 58.2 (2001), pp. 109–130. DOI: 10.1016/S0169-7439(01)00155-1.

[36] K. A. Bakeev, ed. *Process Analytical Technology: Spectroscopic Tools and Implementation Strategies for the Chemical and Pharmaceutical Industries*. Wiley, 2010.

[37] Å. Rinnan, F. van den Berg, and S. B. Engelsen. „Review of the most common pre-processing techniques for near-infrared spectra“. In: *TrAC Trends in Analytical Chemistry* 28.10 (2009), pp. 1201–1222. DOI: 10.1016/j.trac.2009.07.007.

[38] A. Savitzky and M. J. E. Golay. „Smoothing and Differentiation of Data by Simplified Least Squares Procedures“. In: *Journal of Analytical Chemistry* 36.8 (1964), pp. 1627–1639.

[39] R. Gautam, S. S. G. Vanga, F. Ariese, and S. Umapathy. „Review of multidimensional data processing approaches for Raman and infrared spectroscopy“. In: *EPJ Techniques and Instrumentation* 2.1 (2015), p. 8. DOI: 10.1140/epjti/s40485-015-0018-6.

[40] T. Mehmood, K. H. Liland, L. Snipen, and S. Sæbø. „A review of variable selection methods in Partial Least Squares Regression“. In: *Chemometrics and Intelligent Laboratory Systems* 118 (2012), pp. 62–69. DOI: 10.1016/j.chemolab.2012.07.010.

[41] S. Wold, M. Sjöström, and L. Eriksson. „PLS-regression: a basic tool of chemometrics“. In: *Chemometrics and Intelligent Laboratory Systems* 58.2 (2001), pp. 109–130. DOI: 10.1016/S0169-7439(01)00155-1.

[42] S. Wold, M. Sjöström, and L. Eriksson. „PLS-regression: a basic tool of chemometrics“. In: *Chemometrics and Intelligent Laboratory Systems* 58.2 (2001), pp. 109–130. DOI: 10.1016/S0169-7439(01)00155-1.

[43] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[44] F. Lussier, V. Thibault, B. Charron, G. Q. Wallace, and J.-F. Masson. „Deep learning and artificial intelligence methods for Raman and surface-enhanced Raman scattering“. In: *Trends in Analytical Chemistry* 124 (2020), p. 115796. DOI: 10.1016/j.trac.2019.115796.

[45] Y. LeCun, Y. Bengio, and G. Hinton. „Deep learning“. In: *Nature* 521.7553 (2015), pp. 436–444. DOI: 10.1038/nature14539.

[46] C.-S. Ho et al. „Rapid identification of pathogenic bacteria using Raman spectroscopy and deep learning“. In: *Nature Communications* 10.1 (2019), p. 4927. DOI: 10.1038/s41467-019-12898-9.

[47] J. Acquarelli et al. „Convolutional neural networks for vibrational spectroscopic data analysis“. In: *Analytica Chimica Acta* 954 (2017), pp. 22–31. DOI: 10.1016/j.aca.2016.12.010.

[48] R. Tauler. „Multivariate curve resolution applied to second order data“. In: *Chemometrics and Intelligent Laboratory Systems* 30.1 (1995), pp. 133–146. DOI: 10.1016/0169-7439(95)00047-X.

[49] A. Barth. „Infrared spectroscopy of proteins“. In: *Biochimica et Biophysica Acta (BBA) - Bioenergetics* 1767.9 (2007), pp. 1073–1101. DOI: 10.1016/j.bbabio.2007.06.004.

[50] Z. Movasaghi, S. Rehman, and I. U. Rehman. „Raman spectroscopy of biological tissues“. In: *Applied Spectroscopy Reviews* 42.5 (2007), pp. 493–541. DOI: 10.1080/05704920701551530.

[51] T. Ida, M. Ando, and H. Toraya. „Extended pseudo-Voigt function for approximating the Voigt profile“. In: *Journal of Applied Crystallography* 33 (2000), pp. 1311–1316. DOI: 10.1107/S0021889800010219.

[52] Feidl et al. „Combining Mechanistic Modeling and Raman Spectroscopy for Monitoring Antibody Chromatographic Purification“. In: *Processes* 7.10 (2019), p. 683. DOI: 10.3390/pr7100683.

[53] L. Rolinger et al. „Monitoring of ultra- and diafiltration processes by Raman spectroscopy combined with Kalman filtering“. In: *Biotechnology and Bioengineering* 120 (2023), e28378. DOI: 10.1002/bit.28378.

[54] N. Brestrich, M. Rüdt, D. Büchler, and J. Hubbuch. „Selective protein quantification for preparative chromatography using variable pathlength UV/Vis spectroscopy and partial least squares regression“. In: *Chemical Engineering Science* 176 (2018), pp. 157–164. DOI: 10.1016/j.ces.2017.10.030.

[55] R. M. Santos, J.-M. Kessler, P. Salou, J. C. Menezes, and A. Peinado. „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 (2018), pp. 659–670. DOI: 10.1002/btpr.2635.

[56] P. Romann, J. Kolar, D. Tobler, C. Herwig, J.-M. Bielser, and T. K. Villiger. „Advancing Raman model calibration for perfusion bioprocesses using spiked harvest libraries“. In: *Biotechnology Journal* 17.11 (2022), e2200184. DOI: 10.1002/biot.202200184.

[57] M. Klaverdijk, M. Ottens, and M. E. Klijn. „Single compound data supplementation to enhance transferability of fermentation specific Raman spectroscopy models“. In: *Analytical and Bioanalytical Chemistry* (2025), pp. 1–12. DOI: 10.1007/s00216-025-05768-5.

[58] G. Thakur, V. Hebbi, and A. S. Rathore. „Near Infrared Spectroscopy as a PAT tool for monitoring and control of protein and excipient concentration in ultrafiltration of highly concentrated antibody formulations“. In: *International Journal of Pharmaceutics* 600 (2021), p. 120456. DOI: 10.1016/j.ijpharm.2021.120456.

[59] S. K. Hansen, B. Jamali, and J. Hubbuch. „Selective high throughput protein quantification based on UV absorption spectra“. In: *Biotechnology and Bioengineering* 110.2 (2013), pp. 448–460. DOI: 10.1002/bit.24712.

[60] J. Wang, J. Chen, J. Studts, and G. Wang. „Automated calibration and in-line measurement of product quality during therapeutic monoclonal antibody purification using Raman spectroscopy“. In: *Biotechnology and Bioengineering* (2023). DOI: 10.1002/bit.28514.

[61] J. Chen, J. Wang, R. Hess, G. Wang, J. Studts, and M. Franzreb. „Application of Raman spectroscopy during pharmaceutical process development for determination of critical quality attributes in Protein A chromatography“. In: *Journal of Chromatography A* (2024), p. 464721. DOI: 10.1016/j.chroma.2024.464721.

[62] J. Abel et al. „A small-scale process for predicting Donnan and volume exclusion effects during ultrafiltration/diafiltration process development“. In: *Journal of Pharmaceutical Sciences* 107 (2018), pp. 1600–1609. DOI: 10.1016/j.xphs.2018.01.010.

[63] J. M. West et al. „Process analytical technology for on-line monitoring of biologics processes“. In: *Biotechnology and Bioengineering* 118 (2021), pp. 1234–1245. DOI: 10.1002/bit.27741.

[64] R. Nikzad-Langerodi and E. Andries. „A chemometrician's guide to transfer learning“. In: *Journal of Chemometrics* 35.11 (2021), e3373. DOI: 10.1002/cem.3373.

[65] R. Schiemer, M. Rüdt, and J. Hubbuch. „Generative data augmentation and automated optimization of convolutional neural networks for process monitoring“. In: *Frontiers in Bioengineering and Biotechnology* 12 (2024).

[66] V. I. Korepanov. „Asymmetric least-squares baseline algorithm with peak screening for automatic processing of the Raman spectra“. In: *Journal of Raman Spectroscopy* 51 (2020), pp. 2061–2065. DOI: 10.1002/jrs.5952.

[67] S. Wold, M. Sjöström, and L. Eriksson. „PLS-regression: a basic tool of chemometrics“. In: *Chemometrics and Intelligent Laboratory Systems* 58.2 (2001), pp. 109–130. DOI: 10.1016/S0169-7439(01)00155-1.

[68] J. M. Smulko, N. C. Dingari, J. S. Soares, and I. Barman. „Anatomy of Noise in Quantitative Biological Raman Spectroscopy“. In: *Bioanalysis* 6.3 (2014), pp. 411–421. DOI: 10.4155/bio.13.337.

[69] L. Rolinger, M. Rüdt, and J. Hubbuch. „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 (2020), pp. 2047–2064. DOI: 10.1007/s00216-020-02407-z.

[70] D. H. Müller, M. Börger, J. Thien, and H.-J. Koß. „The Good pH probe: non-invasive pH in-line monitoring using Good buffers and Raman spectroscopy“. In: *Analytical and Bioanalytical Chemistry* 415.29 (2023), pp. 7247–7258. DOI: 10.1007/s00216-023-04993-0.

[71] F. Rischawy et al. „Modeling of biopharmaceutical UF/DF from laboratory to manufacturing scale“. In: *Computers and Chemical Engineering* 177 (2023), p. 108337. DOI: 10.1016/j.compchemeng.2023.108337.

[72] T. Briskot et al. „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 (2022), p. 120333. DOI: 10.1016/j.memsci.2022.120333.

[73] M. Vázquez-Rey and D. A. Lang. „Aggregates in monoclonal antibody manufacturing processes“. In: *Biotechnology and Bioengineering* 108.7 (2011), pp. 1494–1508. DOI: 10.1002/bit.23155.

[74] C. J. Roberts. „Protein aggregation and its impact on product quality“. In: *Current Opinion in Biotechnology* 30 (2014), pp. 211–217. DOI: 10.1016/j.copbio.2014.08.001.

[75] N. B. Pham and W. S. Meng. „Protein aggregation and immunogenicity of biotherapeutics“. In: *International Journal of Pharmaceutics* 585 (2020), p. 119523. DOI: 10.1016/j.ijpharm.2020.119523.

[76] E. M. Moussa et al. „Immunogenicity of Therapeutic Protein Aggregates“. In: *Journal of Pharmaceutical Sciences* 105.2 (2016), pp. 417–430. DOI: 10.1016/j.xphs.2015.11.002.

[77] L. Rolinger et al. „Multi-attribute PAT for UF/DF of Proteins—Monitoring Concentration, particle sizes, and Buffer Exchange“. In: *Analytical and Bioanalytical Chemistry* 412.9 (2020), pp. 2123–2136. DOI: 10.1007/s00216-019-02318-8.

[78] N. Brestrich, A. Sanden, A. Kraft, K. McCann, J. Bertolini, and J. Hubbuch. „Advances in inline quantification of co-eluting proteins in chromatography: Process-data-based model calibration and application towards real-life separation issues“. In: *Biotechnology and Bioengineering* 112.7 (2015), pp. 1406–1416. DOI: 10.1002/bit.25546.

[79] F. Capito, R. Skudas, H. Kolmar, and C. Hunzinger. „At-line mid infrared spectroscopy for monitoring downstream processing unit operations“. In: *Process Biochemistry* 50.6 (2015), pp. 997–1005. DOI: 10.1016/j.procbio.2015.03.005.

[80] B. Wei et al. „Multi-attribute Raman spectroscopy (MARS) for monitoring product quality attributes in formulated monoclonal antibody therapeutics“. In: *mAbs* 14.1 (2022). DOI: 10.1080/19420862.2021.2007564.

[81] J. Chen, J. Wang, R. Hess, G. Wang, J. Studts, and M. Franzreb. „Application of Raman spectroscopy during pharmaceutical process development for determination of critical quality attributes in Protein A chromatography“. In: *Journal of Chromatography A* 1718 (2024), p. 464721. DOI: 10.1016/j.chroma.2024.464721.

[82] M. Rüdt, P. Vormittag, N. Hillebrandt, and J. Hubbuch. „Process monitoring of virus-like particle reassembly by diafiltration with UVVis spectroscopy and light scattering“. In: *Biotechnology and Bioengineering* 116 (2019), pp. 1366–1379. DOI: 10.1002/bit.26935.

[83] R. Schiemer, J. T. Weggen, K. M. Schmitt, and J. Hubbuch. „An adaptive soft-sensor for advanced real-time monitoring of an antibody-drug conjugation reaction“. In: *Biotechnology and Bioengineering* 120.7 (2023), pp. 1914–1928. DOI: 10.1002/bit.28428.

[84] N. Alizadeh-Pasdar, S. Nakai, and E. C. Y. Li-Chan. „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 (2002), pp. 6042–6052. DOI: 10.1021/jf0116956.

[85] V. Militello, C. Casarino, A. Emanuele, A. Giostra, F. Pullara, and M. Leone. „Aggregation kinetics of bovine serum albumin studied by FTIR spectroscopy and light scattering“. In: *Biophysical Chemistry* 107.2 (2004), pp. 175–187. DOI: 10.1016/j.bpc.2003.09.004.

[86] I. Ettah and L. Ashton. „Engaging with Raman Spectroscopy to Investigate Antibody Aggregation“. In: *Antibodies* 7.3 (2018), p. 24. DOI: 10.3390/antib7030024.

[87] R. G. de la Cuesta, R. Goodacre, and L. Ashton. „Monitoring Antibody Aggregation in Early Drug Development Using Raman Spectroscopy and Perturbation-Correlation Moving Windows“. In: *Analytical Chemistry* 86.22 (2014), pp. 11133–11140. DOI: 10.1021/ac5038329.

[88] E. N. Lewis, W. Qi, L. H. Kidder, S. Amin, S. M. Kenyon, and S. Blake. „Combined Dynamic Light Scattering and Raman Spectroscopy Approach for Characterizing the Aggregation of Therapeutic Proteins“. In: *Molecules* 19.12 (2014), pp. 20888–20905. DOI: 10.3390/molecules191220888.

[89] G. V. Barnett et al. „Structural Changes and Aggregation Mechanisms for Anti-Streptavidin IgG1 at Elevated Concentration“. In: *The Journal of Physical Chemistry B* 119.49 (2015), pp. 15150–15163. DOI: 10.1021/acs.jpcb.5b08748.

[90] C. Ota, S. Noguchi, S. Nagatoishi, and K. Tsumoto. „Assessment of the Protein–Protein Interactions in a Highly Concentrated Antibody Solution by Using Raman Spectroscopy“. In: *Pharmaceutical Research* 33.4 (2016), pp. 956–969. DOI: 10.1007/s11095-015-1842-8.

[91] Y. Sato, S. Nagatoishi, S. Noguchi, and K. Tsumoto. „Raman Spectroscopic Analysis of Highly-Concentrated Antibodies under the Acid-Treated Conditions“. In: *Pharmaceutical Research* 40.7 (2023), pp. 1853–1864. DOI: 10.1007/s11095-023-03526-9.

[92] A. A. Makki et al. „Understanding the discrimination and quantification of monoclonal antibodies preparations using Raman spectroscopy“. In: *Journal of Pharmaceutical and Biomedical Analysis* 194 (2021), p. 113734. DOI: 10.1016/j.jpba.2020.113734.

[93] C. Zhang, J. S. Springall, X. Wang, and I. Barman. „Rapid, quantitative determination of aggregation and particle formation for antibody drug conjugate therapeutics with label-free Raman spectroscopy“. In: *Analytica Chimica Acta* 1081 (2019), pp. 138–145. DOI: 10.1016/j.aca.2019.07.007.

[94] J. Wang, J. Chen, J. Studts, and G. Wang. „In-line product quality monitoring during biopharmaceutical manufacturing using computational Raman spectroscopy“. In: *mAbs* 15.1 (2023), p. 2220149. DOI: 10.1080/19420862.2023.2220149.

[95] L. Rolinger, J. Hubbuch, and M. Rüdt. „Monitoring of ultra- and diafiltration processes by Kalman-filtered Raman measurements“. In: *Analytical and Bioanalytical Chemistry* 415.5 (2023), pp. 841–854. DOI: 10.1007/s00216-022-04477-7.

[96] J. Heyer-Müller et al. „A novel Raman-chromatography assembly for automated calibration and in-line monitoring in bioprocessing“. In: *Engineering in Life Sciences* (2025). DOI: 10.1002/elsc.70044.

[97] K. Furić, I. Ciglenečki, and B. Ćosović. „Raman spectroscopic study of sodium chloride water solutions“. In: *Journal of Molecular Structure* 550–551 (2000), pp. 225–234. DOI: 10.1016/S0022-2860(00)00388-4.

[98] D. Erb. *pybaselines: A Python library of algorithms for the baseline correction of experimental data*. DOI: 10.5281/zenodo.5608581.

[99] J. L. Lippert, D. Tyminski, and P. J. Desmeules. „Determination of the secondary structure of proteins by laser Raman spectroscopy“. In: *Journal of the American Chemical Society* 98.22 (1976), pp. 7075–7080. DOI: 10.1021/ja00438a057.

[100] R. W. Williams. „[14] Protein secondary structure analysis using Raman amide I and amide III spectra“. In: *Methods in Enzymology*. Vol. 130. Elsevier, 1986, pp. 311–331. DOI: 10.1016/0076-6879(86)30016-8.

[101] A. Rygula, K. Majzner, K. M. Marzec, A. Kaczor, M. Pilarczyk, and M. Baranska. „Raman spectroscopy of proteins: A review“. In: *Journal of Raman Spectroscopy* 44.8 (2013), pp. 1061–1076. DOI: 10.1002/jrs.4335.

[102] I. Harada, T. Miura, and H. Takeuchi. „Origin of the doublet at 1360 and 1340 cm—1 in the Raman spectra of tryptophan and related compounds“. In: *Spectrochimica Acta Part A: Molecular Spectroscopy* 42.2–3 (1986), pp. 307–312. DOI: 10.1016/0584-8539(86)80193-3.

[103] A. S. Eisenberg and L. J. Juszczak. „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 (2012), pp. 1–10. DOI: 10.1155/2012/735076.

[104] J. Soch and C. Allefeld. „MACS – a new SPM toolbox for model assessment, comparison and selection“. In: *Journal of Neuroscience Methods* 306 (2018), pp. 19–31. DOI: 10.1016/j.jneumeth.2018.05.017.

[105] M. Boulet-Audet, B. Byrne, and S. G. Kazarian. „High-Throughput Thermal Stability Analysis of a Monoclonal Antibody by Attenuated Total Reflection FT-IR Spectroscopic Imaging“. In: *Analytical Chemistry* 86.19 (2014), pp. 9786–9793. DOI: 10.1021/ac502529q.

[106] B. Shivu, S. Seshadri, J. Li, K. A. Oberg, V. N. Uversky, and A. L. Fink. „Distinct -Sheet Structure in Protein Aggregates Determined by ATR–FTIR Spectroscopy“. In: *Biochemistry* 52.31 (2013), pp. 5176–5183. DOI: 10.1021/bi400625v.

[107] M. D. Fontana, K. B. Mabrouk, and T. H. Kauffmann. „Raman spectroscopic sensors for inorganic salts“. In: *Spectroscopic Properties of Inorganic and Organometallic Compounds*. Ed. by J. Yarwood, R. Douthwaite, and S. Duckett. Vol. 44. RSC Publishing, 2013, pp. 40–67. DOI: 10.1039/9781849737791-00040.

[108] L. Ashton and E. W. Blanch. „pH-induced conformational transitions in α-lactalbumin investigated with two-dimensional Raman correlation variance plots and moving windows“. In: *Journal of Molecular Structure* 974.1 (2010), pp. 132–138. DOI: 10.1016/j.molstruc.2010.03.005.

[109] C. H. Li and T. Li. „Application of Vibrational Spectroscopy to the Structural Characterization of Monoclonal Antibody and its Aggregate“. In: *Current Pharmaceutical Biotechnology* 10.4 (2009), pp. 391–399. DOI: 10.2174/138920109788488950.

[110] J. M. Antosiewicz and D. Shugar. „UV–Vis spectroscopy of tyrosine side-groups in studies of protein structure. Part 2: selected applications“. In: *Biophysical Reviews* 8 (2016), pp. 163–177. DOI: 10.1007/s12551-016-0197-7.

[111] C. Lange et al. „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 F. Manenti and G. V. Reklaitis. Vol. 53. Computer Aided Chemical Engineering. Elsevier, 2024, pp. 3055–3060. DOI: 10.1016/B978-0-443-28824-1.50510-X.

[112] A. G. Cavinato, D. M. Mayes, Z. Ge, J. B. Callis, et al. „Noninvasive method for monitoring ethanol in fermentation processes using fiber-optic near-infrared spectroscopy“. In: *Analytical Chemistry* 62.18 (1990), pp. 2058–2063. DOI: 10.1021/ac00217a015.

[113] D. H. Müller, H.-J. Koss, et al. „Bioprocess in-line monitoring and control using Raman spectroscopy and Indirect Hard Modeling (IHM)“. In: *Biotechnology and Bioengineering* (2024). DOI: 10.1002/bit.28724.

[114] K. Pontius, H. Junicke, K. V. Gernaey, and M. Bevilacqua. „Monitoring yeast fermentations by nonlinear infrared technology and chemometrics – understanding process correlations and indirect predictions“. In: *Applied Microbiology and Biotechnology* 104.12 (2020), pp. 5315–5335. DOI: 10.1007/s00253-020-10604-0.

[115] A. Milewska, G. Baekelandt, S. Boutaieb, V. Mozin, and A. Falconbridge. „In-line monitoring of protein concentration with mid-infrared spectroscopy during ultrafiltration/diafiltration“. In: *Engineering in Life Sciences* 23.2 (2023), e2200050. DOI: 10.1002/elsc.202200050.

[116] N. Brestrich, M. Rüdt, D. Büchler, and J. Hubbuch. „Selective protein quantification for preparative chromatography using variable pathlength UV/Vis spectroscopy and partial least squares regression“. In: *Chemical Engineering Science* 176.C (2018), pp. 157–164. DOI: 10.1016/j.ces.2017.10.030.

[117] D. H. Müller, C. Flake, T. Brands, and H.-J. Koß. „Bioprocess in-line monitoring using Raman spectroscopy and Indirect Hard Modeling (IHM): A simple calibration yields a robust model“. en. In: *Biotechnology and Bioengineering* 120.7 (2023). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bit.28424, pp. 1857–1868. DOI: 10.1002/bit.28424.

[118] M. Melcher et al. „The potential of random forest and neural networks for biomass and recombinant protein modeling in Escherichia coli fed-batch fermentations“. In: *Biotechnology Journal* 10.11 (2015), pp. 1770–1782. DOI: https://doi.org/10.1002/biot.201400790.

[119] E. Kriesten, D. Mayer, F. Alsmeyer, C. B. Minnich, L. Greiner, and W. Marquardt. „Identification of unknown pure component spectra by indirect hard modeling“. In: *Chemometrics and Intelligent Laboratory Systems* 93.2 (2008), pp. 108–119. DOI: https://doi.org/10.1016/j.chemolab.2008.05.002.

[120] M. N. Siamwiza et al. „Interpretation of the doublet at 850 and 830 cm-1 in the Raman spectra of tyrosyl residues in proteins and certain model compounds“. In: *Biochemistry* 14.22 (1975), pp. 4870–4876. DOI: 10.1021/bi00693a014.

[121] R. Carmona, W.-L. Hwang, and B. Torresani. *Practical Time-Frequency Analysis: Gabor and wavelet transforms, with an implementation in S*. Vol. 9. Academic Press, 1998.

[122] P. Du, W. A. Kibbe, and S. M. Lin. „Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching“. In: *Bioinformatics* 22.17 (2006). _eprint: https://academic.oup.com/bioinformatics/article-pdf/22/17/2059/48840388/bioinformatics_22_17_2059.pdf, pp. 2059–2065. DOI: 10.1093/bioinformatics/btl355.

List of Figures

3.1 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. 29

3.2 Process flow diagram of the cross-flow filtration (CFF) system with an external loop for on-line Raman measurements. 30

3.3 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). 31

3.4 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. 32

3.5 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. 33

3.6 Model predictions for in-line data from UF/DF experiments. 34

4.1 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. 55

4.2 Results of the forced degradation experiment. Off-line recorded Raman spectra and signal-to-noise ratio (SNR) for each wavenumber (A). Spectra are shown after preprocessing as described in Sec. 4.2.2.1. 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). 56

4.3 Spectral features observed in offline-recorded Raman spectra of BSA obtained during mixing studies, preprocessed as described in Section 4.2.2.1. 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). 57

4.4 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 cm^{-1}/1320 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. 58

4.5 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. 59

5.1 Comparison of second-derivative Raman spectra in the Amide I region across all chromatographic experiments. Subplots show the spectral variations corresponding to all chromatographic runs in a wavenumber range of 1500 to 1800 cm^{-1}. 83

5.2 Background modeling and salt-induced spectral drift during gradient elution in the Amide I region. (A) Measured buffer spectrum overlaid with total hard model fit, demonstrating underlying components p0 (1555 cm^{-1}) and p1 (1641 cm^{-1}), alongside fitting residuals. (B) Spectral drift occurring between start (0 mM NaCl) and the end condition (500 mM NaCl). 83

5.3 Representative constrained pseudo-Voigt deconvolution of Amide I spectra. (A) Deconvolution of a monomer-rich fraction yielding 12.45 g/L monomer and 2.1% HMWC. (B) Deconvolution of an aggregate-rich fraction yielding 4.12 g/L monomer and 48.5% HMWC. Both panels plot the measured data, total fit, background, individual protein components, and model residuals. 84

5.4 Comparison of monomer and HMWC elution profiles across the investigated AEX experiments. Blue traces represent monomer concentration, red traces represent HMWC concentration in g/L over time. Panels depict different gradient lengths and batch conditions 84

5.5 Model validation for monomer and HMWC prediction based on hard-model-derived spectral features across n=267 samples. (A) Monomer concentration prediction on the training set, (B) Monomer concentration prediction using a leave-one-out test set, (C) HMWC concentration prediction on the training set, (D) HMWC concentration prediction using a leave-one-out test set. 85

5.6 Distribution of fitted center positions (cm^{-1}) for the protein-associated pseudo-Voigt components across all spectra. The violin plots (blue) depict the underlying probability distribution for components p2 through p6, the inner box plots denote the median and central 50 % range. 86

A3.1 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. 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. 115

A4.1 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. 118

A4.2 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. 119

A4.3 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. 120

A4.4 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. 120

A4.5 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. 121

A4.6 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. 121

A4.7 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. 122

A4.8 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. 122

A4.9 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. 123
A4.10 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. 123
A4.11 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. 124
A4.12 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. 124
A4.13 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. 125

List of Tables

3.1 Experimental conditions for different runs. 35

4.1 Experimental conditions for bind-elute AEX chromatography experiment used in in-line Raman monitoring of BSA monomer and aggregate separation. 60

5.1 Experimental conditions for bind-elute AEX chromatography experiment used in in-line Raman monitoring of BSA monomer and aggregate separation. 82

A4.1 Summary of samples and preparation conditions. 127
A4.2 Regression model variants used for performance comparison. 128

Abbreviations

Acronym	Meaning
KIT	Karlsruher Institut für Technologie
LHS	latin hypercube sampling
LOBO	leave-one-batch-out
LSA	local subset augmentation
LMWS	low-molecular weight species
MAB	Institut für Molekulare Aufarbeitung von Bioprodukten
mAbs	monoclonal antibodies
MIR	mid-infrared
MLR	multiple linear regression
MSE	mean squared error
MWCO	molecular weight cut-off
NIR	near-infrared
OPC-UA	open platform communication unified architecture
OPLS	orthogonal projection to latent structures
PAT	process analytical technology
PCR	principal component regression
Phe	phenylalanine
PL	phospholipid
PLS	partial least squares
PLS1	single-output partial least squares
PLS2	multi-output partial least squares
QbD	Quality by Design
RMSE	root mean squared error
RMSEC	root mean squared error of calibration
RMSECV	root mean squared error of cross-validation
RMSEP	root mean squared error of prediction
SD	standard deviation

Continued on next page

Acronym	Meaning
SEC	size-exclusion chromatography
SG / SGF	Savitzky-Golay / Savitzky-Golay filter
SNR	signal-to-noise ratio
SNV	standard normal variate
TMP	transmembrane pressure
Trp	tryptophan
Tyr	tyrosine
UF	ultrafiltration
UF/DF	ultrafiltration/diafiltration
UHP-SEC	ultrahigh performance - size exclusion chromatography
UPSEC / UPLC	ultra performance size exclusion chromatography / ultra performance liquid chromatography
UV	ultraviolet
UV/Vis	ultraviolet-visible
VIP	variable importance in projection

A3

Appendix Chapter 3

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

Jakob Heyer-Müller[1,*], Robin Schiemer[1,*], Matthias Lopinski[1], Caty Wang[1], Franka Willems[1], Lars Robbel[2], Michael Schmitt[2] and Jürgen Hubbuch[1]

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

[2] CSL Innovation GmbH, Marburg, Germany

[*] Contributed equally.

A3.1 Figures

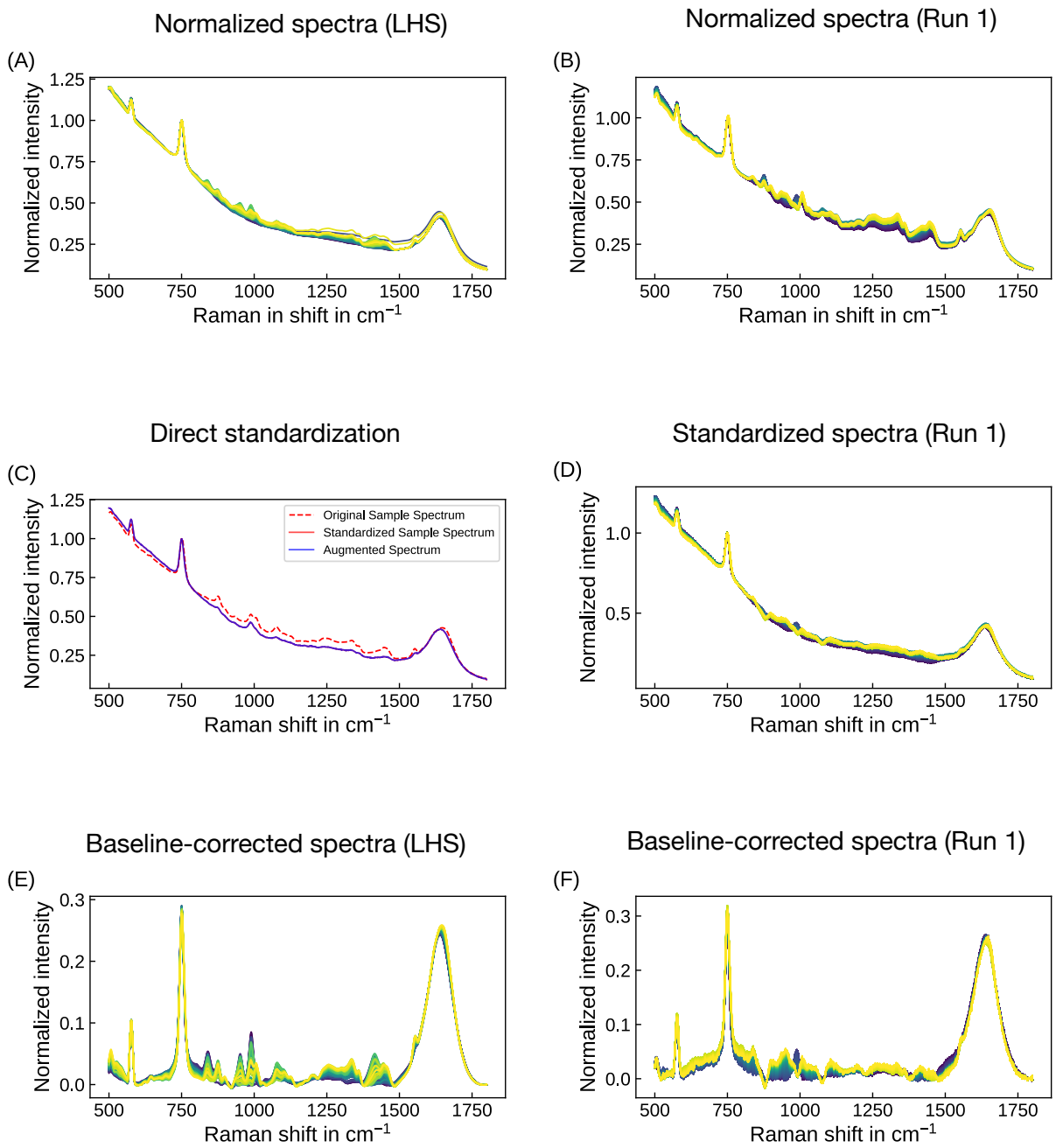


Figure A3.1 Comparison of off-line and in-line spectra after different stages of preprocessing. Spectra from the LHS dataset after smoothing, cropping and normalization at 750 cm^{-1} (A). Spectra from the UFDF Run 1 dataset after smoothing, cropping and normalization at 750 cm^{-1} (B). Illustration 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. 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.

A4

Appendix Chapter 4

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 Robbel[2], Michael Schmitt[2] and Jürgen Hubbuch[1]

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

[2] CSL Innovation GmbH, Marburg, Germany

[*] Contributed equally.

A4.1 Figures

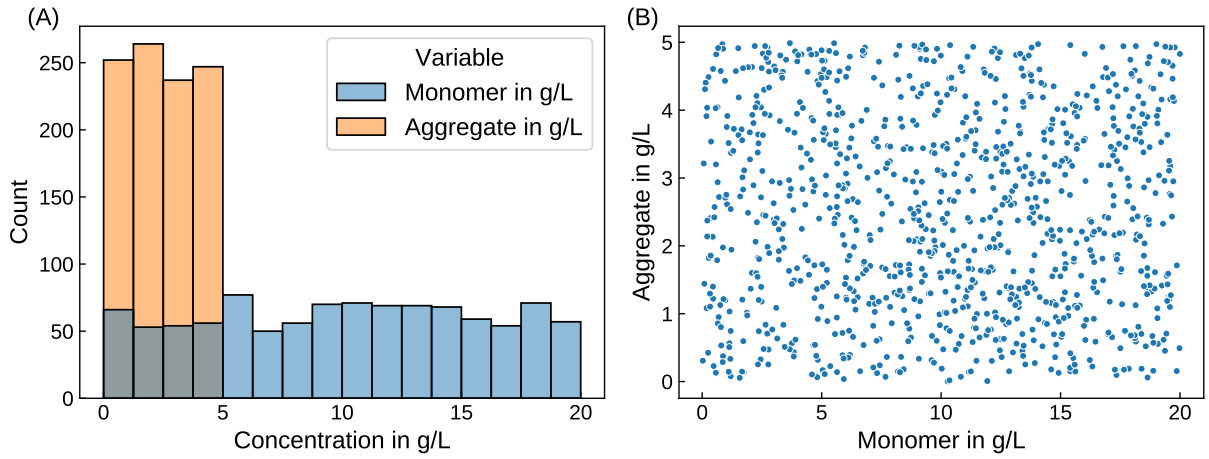


Figure A4.1 Distribution of sampled values for augmentation test. (A) Histogram of the sampled values for monomer and aggregate concentrations. (B) Two-dimensional scatter plot across the sampled value range.

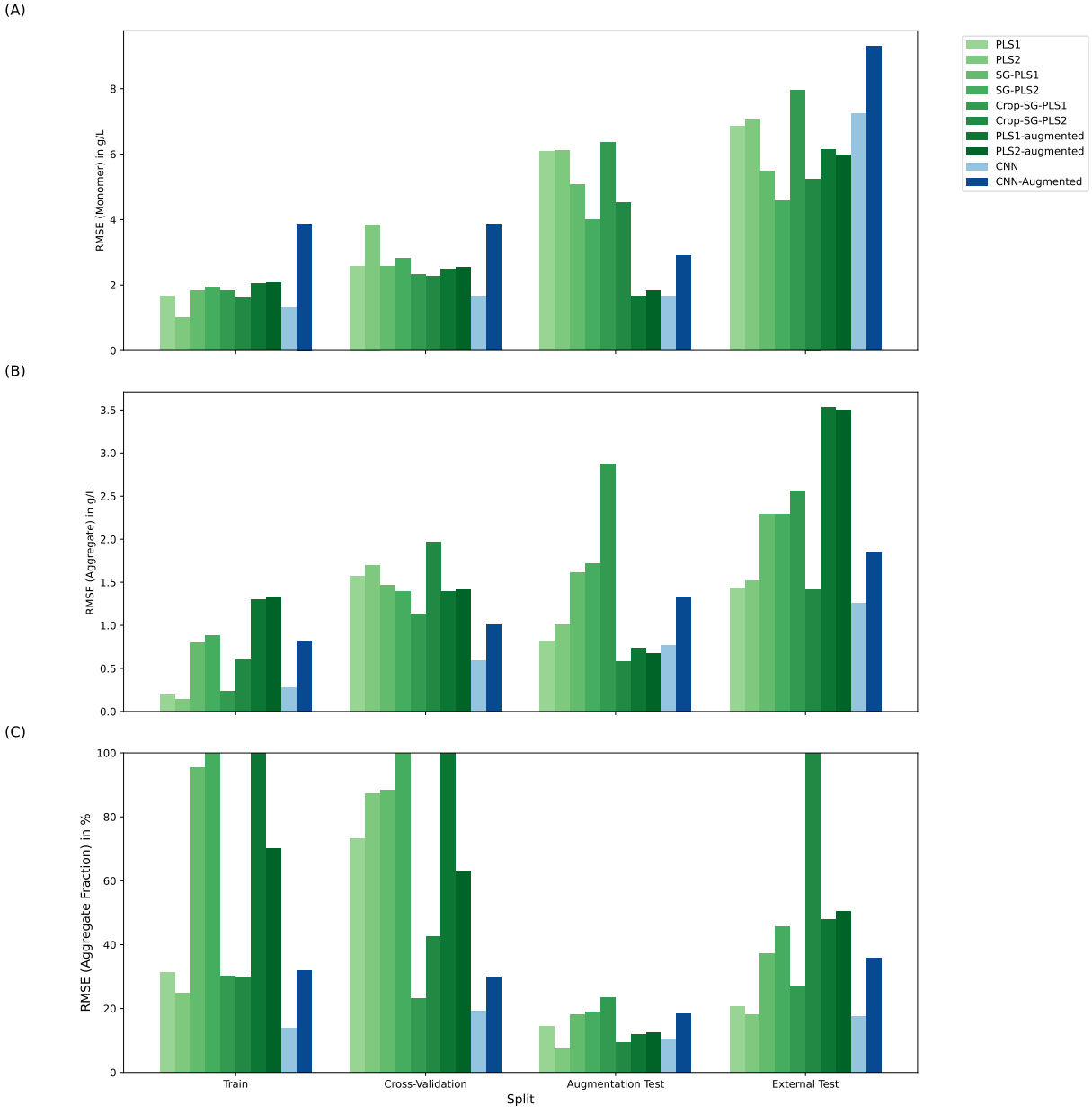


Figure A4.2 Performance comparison for all studied model variants for training, cross-validation, augmentation test and external test data sets. (A) RMSE for Monomer component, (B) RMSE for Aggregate component, (C) RMSE for Aggregate fraction.

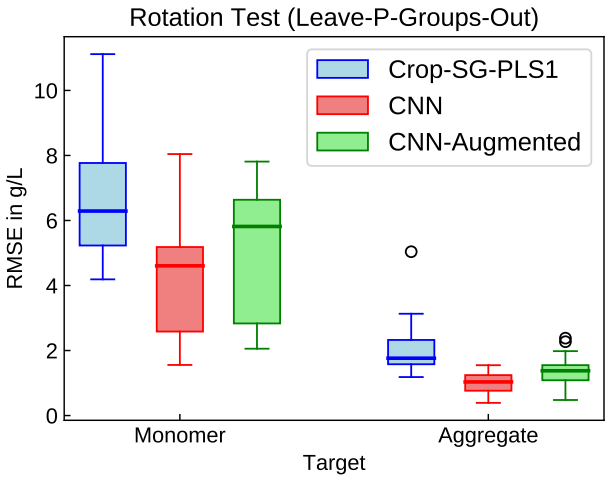


Figure A4.3 Performance comparison of representative PLS and CNN models for rotation test. Distribution of RMSE for the rotated test sets during a leave-4-batches-out cross-validation scheme using all available batches.

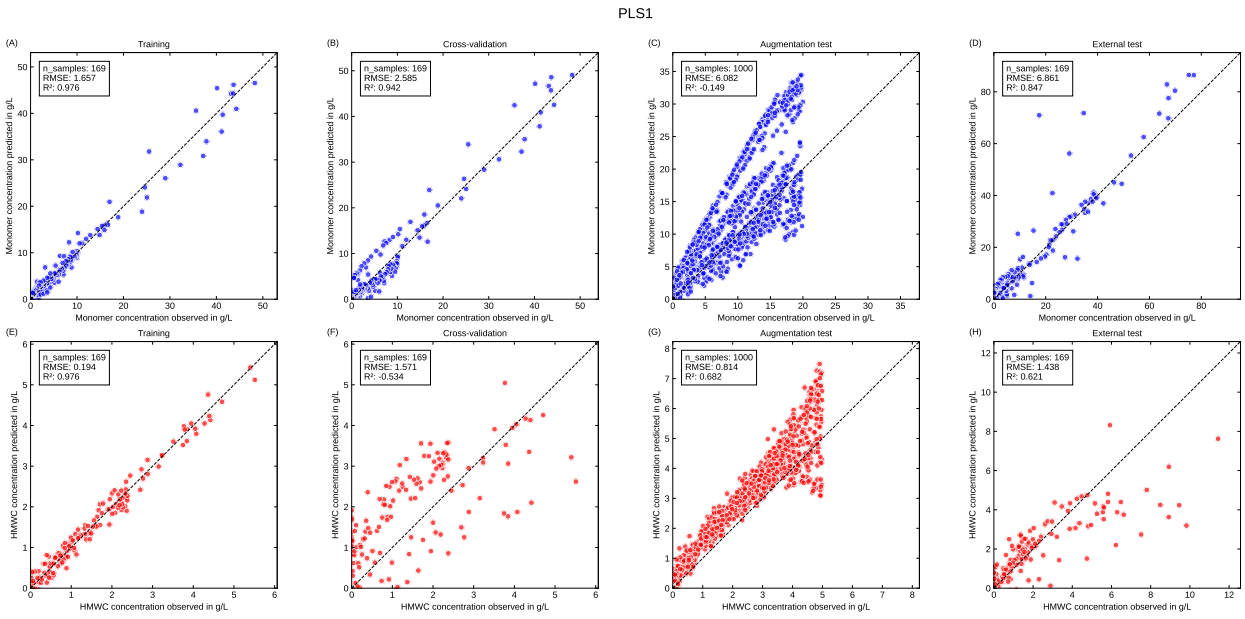


Figure A4.4 Parity plots for % HMWC prediction using the PLS1 model across training, cross-validation, and test datasets. The dashed line marks the ideal parity. R^2 and RMSE values are annotated in each panel.

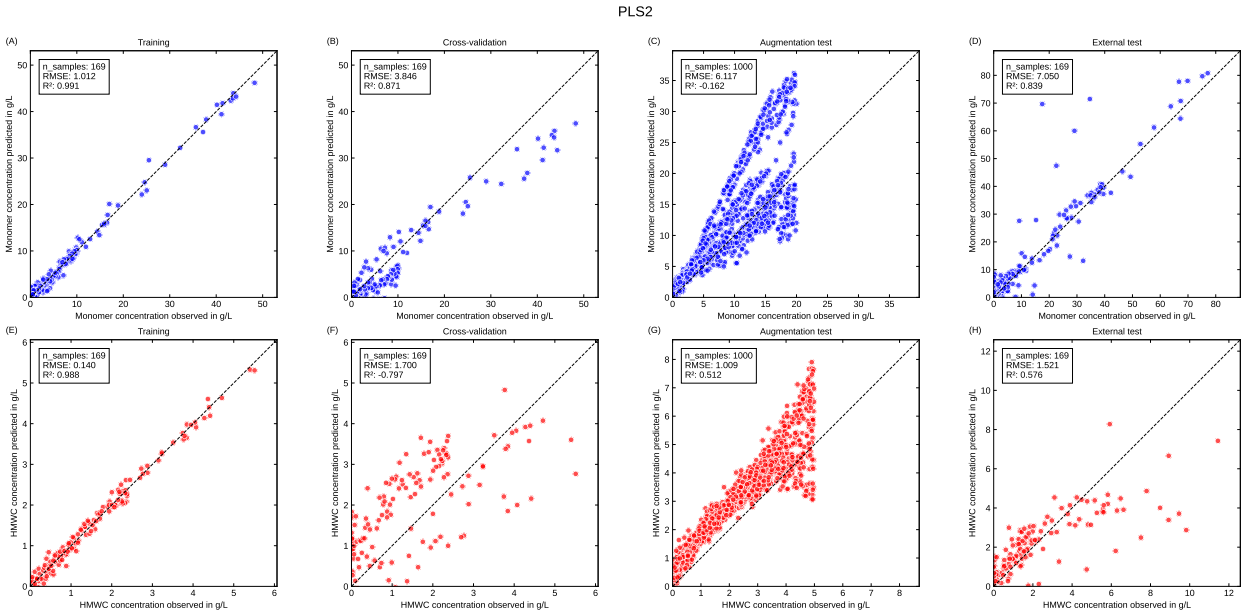


Figure A4.5 Parity plots for % HMWC prediction using the PLS2 model across training, cross-validation, and test datasets. The dashed line marks the ideal parity. R^2 and RMSE values are annotated in each panel.

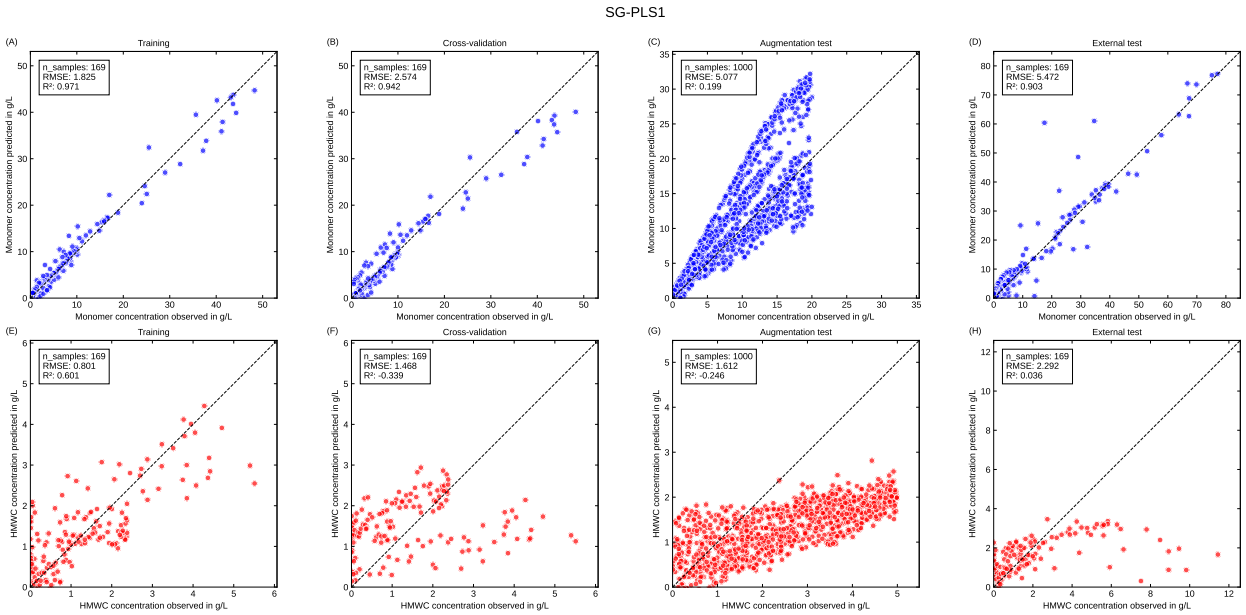


Figure A4.6 Parity plots for % HMWC prediction using the SG-PLS1 model across training, cross-validation, and test datasets. The dashed line marks the ideal parity. R^2 and RMSE values are annotated in each panel.

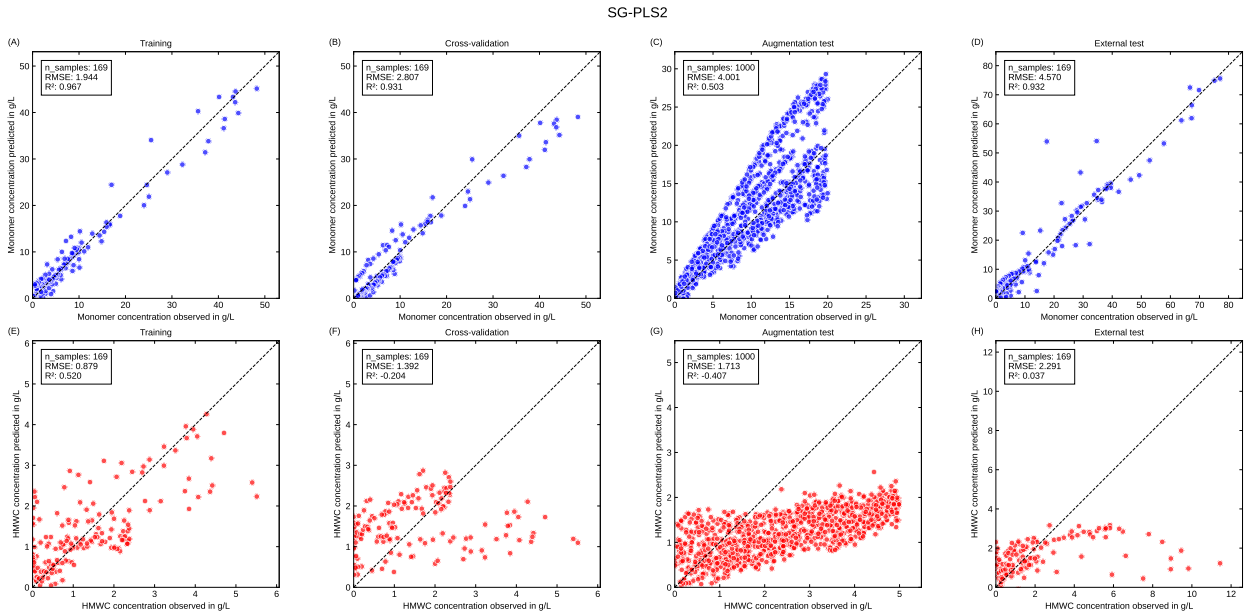


Figure A4.7 Parity plots for % HMWC prediction using the SG-PLS2 model across training, cross-validation, and test datasets. The dashed line marks the ideal parity. R^2 and RMSE values are annotated in each panel.

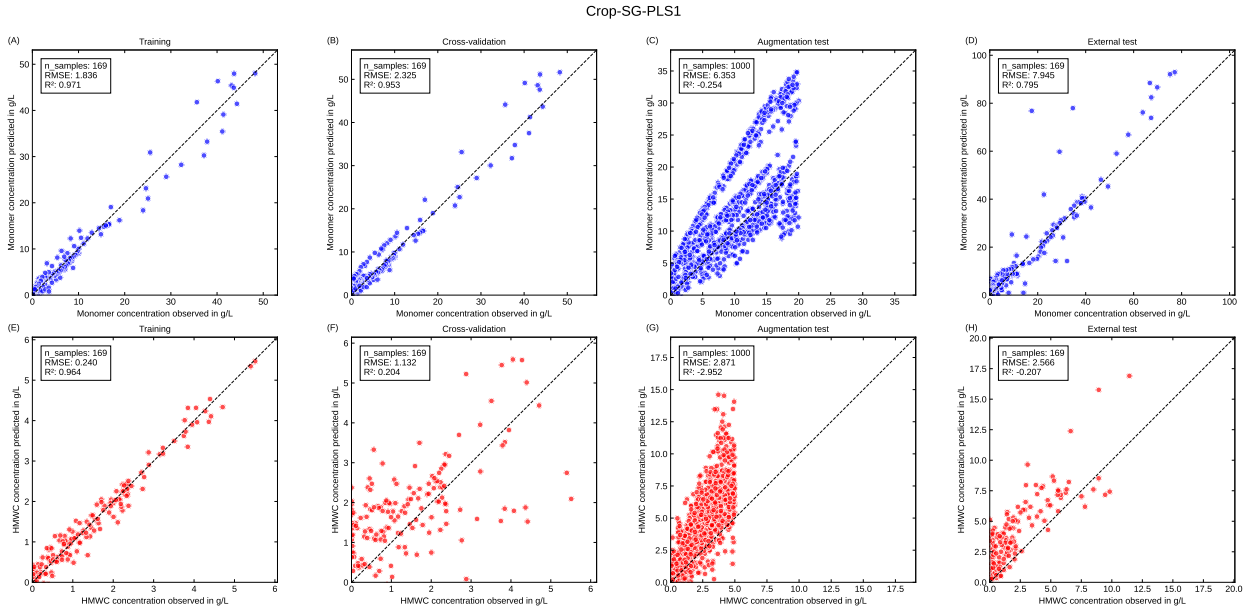


Figure A4.8 Parity plots for % HMWC prediction using the Crop-SG-PLS1 model across training, cross-validation, and test datasets. The dashed line marks the ideal parity. R^2 and RMSE values are annotated in each panel.

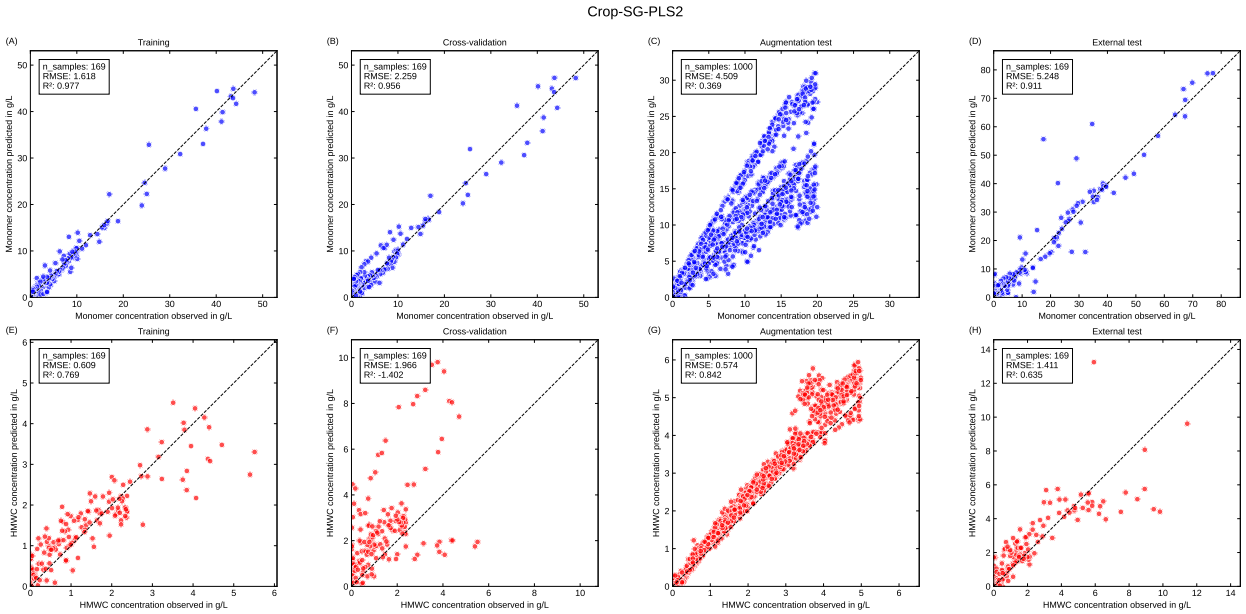


Figure A4.9 Parity plots for % HMWC prediction using the Crop-SG-PLS2 model across training, cross-validation, and test datasets. The dashed line marks the ideal parity. R^2 and RMSE values are annotated in each panel.

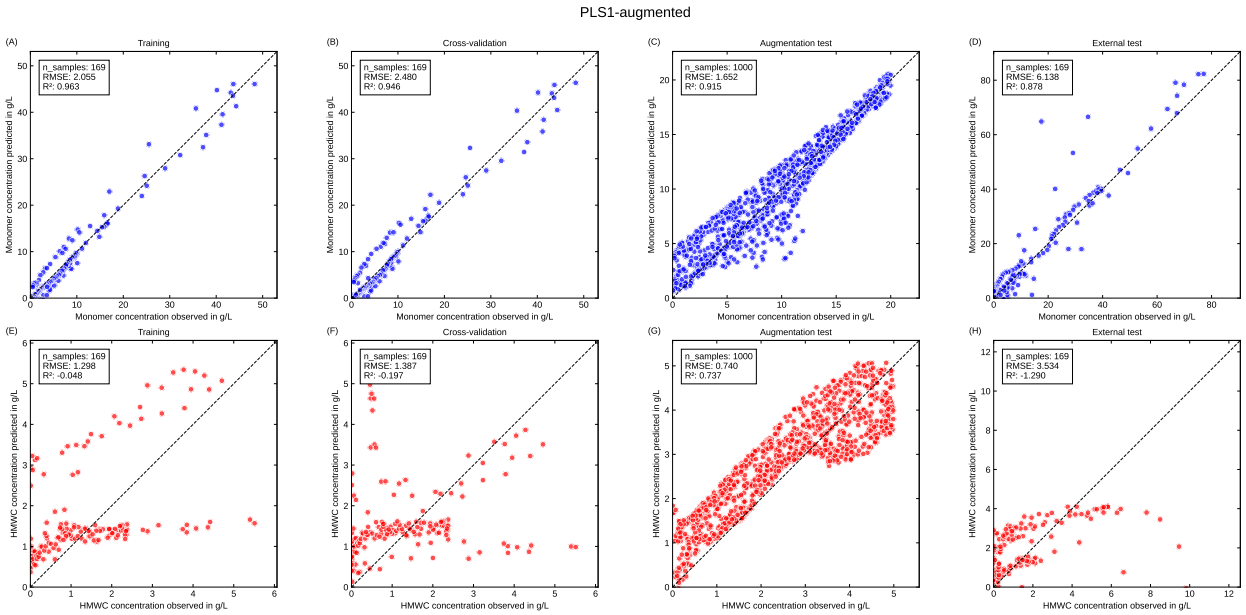


Figure A4.10 Parity plots for % HMWC prediction using the PLS1-Augmented model across training, cross-validation, and test datasets. The dashed line marks the ideal parity. R^2 and RMSE values are annotated in each panel.

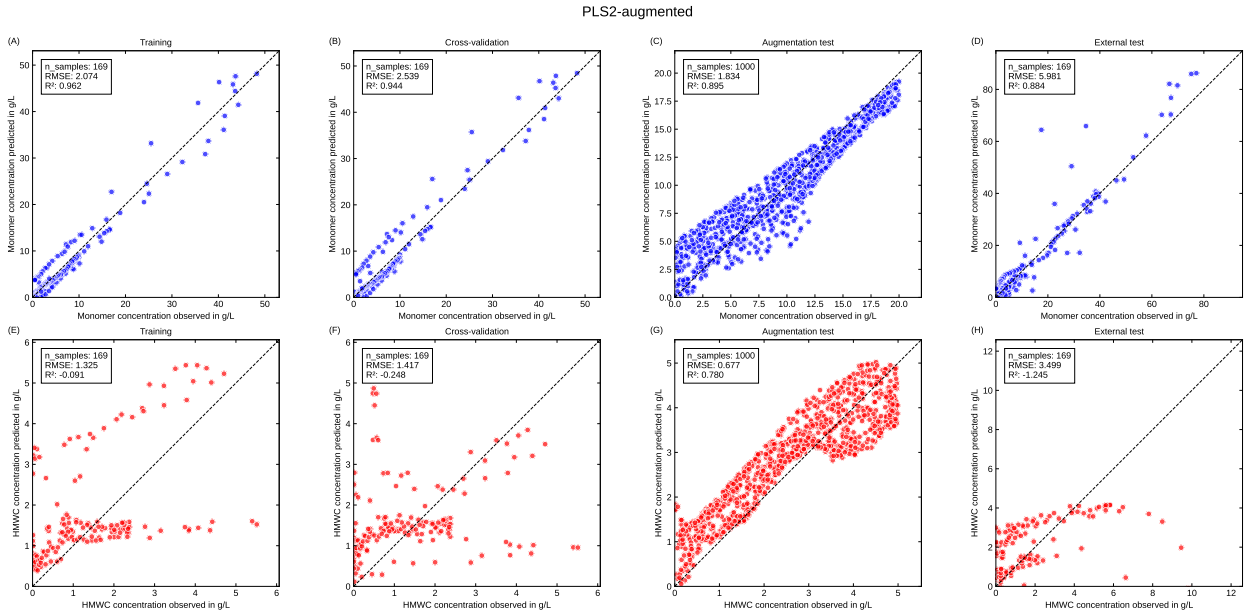


Figure A4.11 Parity plots for % HMWC prediction using the PLS2-Augmented model across training, cross-validation, and test datasets. The dashed line marks the ideal parity. R^2 and RMSE values are annotated in each panel.

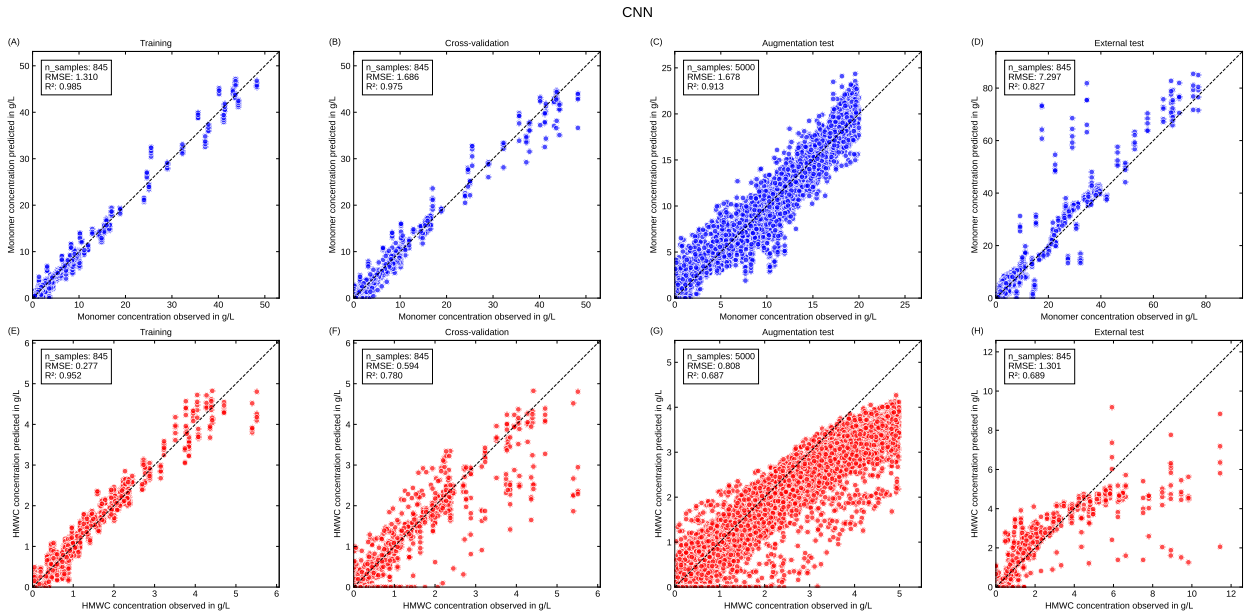


Figure A4.12 Parity plots for % HMWC prediction using the CNN model across training, cross-validation, and test datasets. The dashed line marks the ideal parity. R^2 and RMSE values are annotated in each panel.

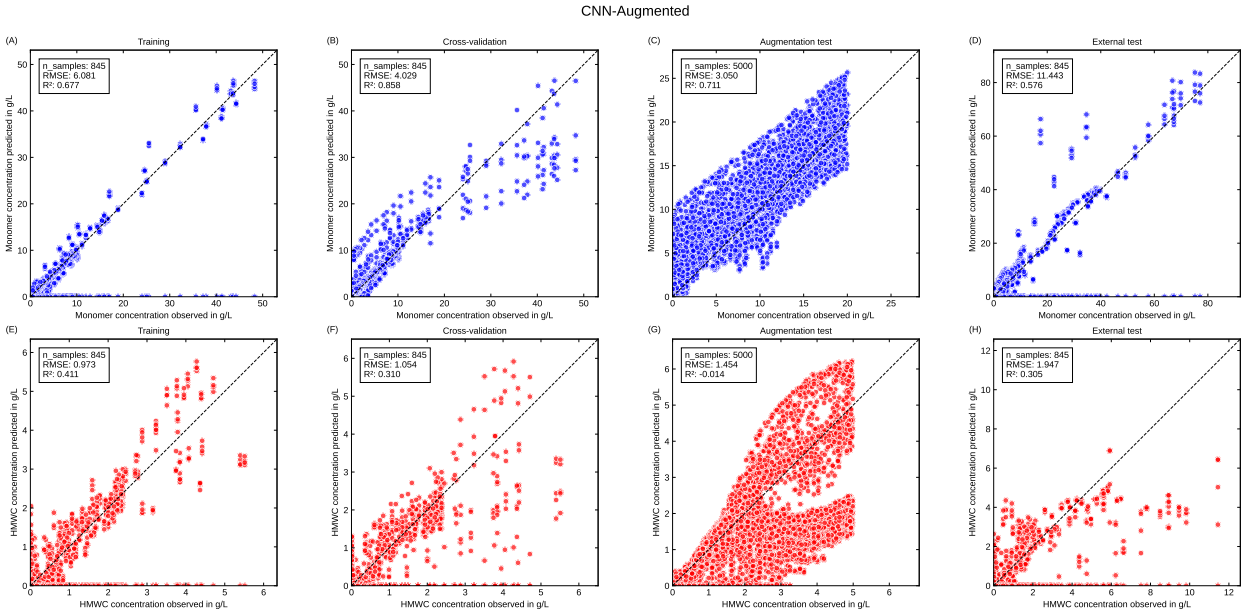


Figure A4.13 Parity plots for % HMWC prediction using the CNN-Augmented model across training, cross-validation, and test datasets. The dashed line marks the ideal parity. R^2 and RMSE values are annotated in each panel.

A4.2 Tables

Table A4.1 Summary of samples and preparation conditions.

Study	Total protein conc. (g/L)	NaCl conc. (mM)	Process conditions	Total samples	Replicates / Notes
Forced degradation	30	0, 50, 100, 250	65 °C, 0–120 min, no agitation	20	triplicate sample preparation; Raman spectra acquired on ARMP; aggregates quantified by UHP-SEC (1:100 dilution).
Mixing study	5–20	150 (fixed)	Unstressed (fresh) and stressed stocks (65 °C, 24 h) mixed	96 (48 LHS × 2)	duplicate sample preparation; Raman spectra recorded on ARMP; monomer and aggregate fractions quantified by UHP-SEC.
AEX chromatography	15 (feed)	0–500	5–15 CV linear gradient + step elution	7 runs	n/a (process runs); 5 mL Eshmuno Q column; fractions 200 µL; 24 Raman spectra averaged per fraction; monomer and aggregate concentrations determined by UHP-SEC.

Table A4.2 Regression model variants used for performance comparison.

Model ID	Pipeline	Multi-Output
PLS1	PLS Regression	no
PLS2	PLS Regression	yes
SG-PLS1	Savitzky-Golay, PLS Regression	no
SG-PLS2	Savitzky-Golay, PLS Regression	yes
Crop-SG-PLS1	Cropping (600–1800 cm^{-1}), Savitzky-Golay, PLS Regression	no
Crop-SG-PLS2	Cropping (600–1800 cm^{-1}), Savitzky-Golay, PLS Regression	yes
PLS1-Aug.	Local Subset Augmentation, PLS Regression	no
PLS2-Aug.	Local Subset Augmentation, PLS Regression	yes
CNN	Convolutional Neural Network	yes
CNN-Aug.	Local Subset Augmentation, Convolutional Neural Network	yes

A5

Appendix Chapter 5
Authorship Contributions

Title	A novel Raman-chromatography assembly for automated calibration and in-line monitoring in bioprocessing
Journal	Engineering in Life Science
Authors	Jakob Heyer-Müller[1], Robin Schiemer[1], Matthias Lopinski, Caty Wang, Franka Willems, Lars Robbel, Michael Schmitt, Jürgen Hubbuch[*] [1]contributed equally [*]corresponding author
Position	The content of this paper has been included in Chapter 3 of the dissertation.

Author	**Contribution**	
Heyer-Müller, Jakob	• conceptualization • methodology • data curation • writing – original draft • supervision	• formal analysis • validation • visualization • writing – review & editing
Schiemer, Robin	• conceptualization • methodology • data curation • writing – original draft • supervision	• formal analysis • validation • visualization • writing – review & editing
Lopinski, Matthias	• software • writing – review & editing	• investigation
Wang, Caty	• investigation	• writing – review & editing
Willems, Franka	• investigation	• writing – review & editing
Robbel, Lars	• resources • writing – review & editing	• supervision
Schmitt, Michael	• resources • writing – review & editing	• supervision
Hubbuch, Jürgen	• resources • project administration • writing – review & editing	• supervision • funding acquisition

Title	Development of Raman spectroscopy and machine learning methods for protein aggregate quantification: Application to BSA in chromatographic processes
Journal	Biotechnology and Bioengineering
Authors	Jakob Heyer-Müller[1], Robin Schiemer[1], Lars Robbel, Michael Schmitt, Jürgen Hubbuch[*] [1]contributed equally [*]corresponding author
Position	The content of this paper has been included in Chapter 4 of the dissertation.

Author	**Contribution**	
Heyer-Müller, Jakob	• conceptualization • methodology • investigation • visualization • writing – review & editing	• formal analysis • validation • data curation • writing – original draft
Schiemer, Robin	• conceptualization • methodology • investigation • visualization • writing – review & editing	• formal analysis • validation • data curation • writing – original draft
Robbel, Lars	• resources • writing – review & editing	• supervision
Schmitt, Michael	• resources • writing – review & editing	• supervision
Hubbuch, Jürgen	• resources • project administration • writing – review & editing	• supervision • funding acquisition

Title	Mechanistic Deconvolution of BSA Size Variants by Constrained Raman Pseudo-Voigt Hard Modeling during Anion-Exchange Chromatography
Journal	Biotechnology Progress
Authors	Jakob Heyer-Müller, Robin Schiemer, Lars Robbel, Michael Schmitt, Jürgen Hubbuch[*] [*]corresponding author
Position	The content of this paper has been included in Chapter 5 of the dissertation.

Author	**Contribution**	
Heyer-Müller, Jakob	• conceptualization • methodology • investigation • visualization • writing – review & editing	• formal analysis • validation • data curation • writing – original draft
Schiemer, Robin	• methodology	• writing – review & editing
Robbel, Lars	• resources • writing – review & editing	• supervision
Schmitt, Michael	• resources • writing – review & editing	• supervision
Hubbuch, Jürgen	• resources • project administration • writing – review & editing	• supervision • funding acquisition