Smart PAT concepts for the downstream process of biologics

Application of multimodal soft sensors, machine learning, and data fusion

zur Erlangung des akademischen Grades eines
Doktors der Ingenieurswissenschaften (Dr.-Ing.)

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

Dissertation

von
Laura Maria Rolinger, M.Sc.

Tag der mündlichen Prüfung: 07.12.2021

Erstgutachter: Prof. Dr. Jürgen Hubbuch
Zweitgutachterin: Prof. Dr. Gisela Guthausen

Abstract

The rapidly increasing number of originator biopharmaceuticals and their biosimilars on the List of Essential Medicine (EML), published by the World Health Organization (WHO), emphasizes the importance of biopharmaceutical drugs for global health, but also the importance of affordable medicine. Biopharmaceuticals improve survival rates for a rising number of patients with previously incurable or untreatable indications. Regardless of the revolution of treatment for unmet indications, biopharmaceuticals come at a major socioeconomic cost due to on average 20 times higher treatment costs in comparison to chemically produced drugs.

One cause for the higher costs of biopharmaceuticals are higher expenses for the purification development and manufacturing. Biopharmaceuticals are therapeutic proteins, which cannot be chemically synthesized due to their large and complex chemical structure. Therefore, biopharmaceuticals are produced by animal, yeast, or bacteria cells. Cells represent a challenging manufacturing system, as they produce not only the desired product, but also variations of the product or impurities in various amounts, which need to be depleted to consistent and safe levels before the drug can be administered to patients. Due to the complex structure, chemical similarity to contaminants, and instability of biopharmaceuticals, production processes can be more complex and harder to monitor and control compared to chemical syntheses.

One approach for monitoring the purification process of biopharmaceuticals are Process Analytical Technologies (PATs). The goal of a PAT method is to measure in a timely manner critical quality and performance attributes of the process to control the process in the long run. Thereby, a PAT method not only guarantees the quality of the product, but also allows for process optimization leading to a reduction in production costs.

However, the close chemical relation between biologics and their impurities leads to high demands for the selectivity of PAT methods for the quantification of those critical attributes. Often one PAT method is not selective enough to distinguish between the product and a specific impurity.

Therefore, the combination of different PAT methods can be necessary for adequate control of the purification process.

The goal of this thesis was to develop smart sensor concepts as PAT tools for the downstream process of biologics. In order to fulfill this objective, the measurement capabilities of different sensors for proteins in aqueous solutions need to be understood. Prerequisites for the applicability to real-time monitoring are the operation in a flow-through setup, either in the process stream itself (in-line) or in a bypass (on-line), and the general observability of protein-related features. Especially optical spectroscopy checks these requirements. Chapter 3, therefore, evaluates the sensitivity and selectivity of optical spectroscopic techniques toward the quantification of proteins in the downstream process. Due to the close chemical relation of the desired product to the contaminants, also the specificity for measuring different protein structure attributes is important for the applicability of the spectroscopic techniques. Therefore, a focus of Chapter 3 lies on the measurability of different structure levels of proteins and thereby differentiation between proteins by different spectroscopic methods. Additionally, a guidance for Partial-Least Squares (PLS) model calibration and validation is given to avoid common pitfalls in analysis of spectroscopic data. From this review, Ultraviolet (UV) spectroscopy was identified as the most sensitive spectroscopic technique for the measurement of proteins in aqueous solutions due to the high absorption coefficients of proteins and low absorption of water. However, UV spectroscopy lacks a high selectivity to differentiate between different proteins due to broad and overlapping bands of different structure elements. Here, Raman spectroscopy seems to be promising due to availability of information on the primary, secondary, and tertiary protein structure in the spectra. The only spectroscopic technique, which can measure the most common form of aggregated proteins with main changes in quaternary structure level is light scattering. Unfortunately, there is no universal sensor, which can measure all Critical Quality Attributes (CQAs) in the downstream process. This makes sensor combinations necessary. Therefore Chapter 3 includes a summary of data fusion techniques to cope with the multi-block data from different sensors.

As a first example, the implementation of UV spectroscopy combined with PLS modeling is shown as a proof-of-concept for the real-time monitoring and control of the Protein A load phase in Chapter 4. It was demonstrated that PLS models based on UV absorption spectra can be applied to quantify the monoclonal antibody (mAb) concentration in the column effluent during the load phase despite the influence of many protein and non-protein-based impurities on the UV spectra, which is referred to as background in this thesis. Based on the quantification, the load phase was automatically terminated,

when a previously specified mAb concentration was reached. Consequently, the proposed method has potential for the monitoring and control of capture steps, like Protein A chromatography, at large-scale production for both batch and continuous processes. In batch chromatography, the loading volume can be determined dynamically with the proposed method, which allows for increased resin capacity utilization while keeping the product loss small. Additionally, the time-consuming off-line determination of the mAb titer in the Harvested Cell Culture Fluid (HCCF) prior to the start of the load phase could be eliminated. For continuous chromatography, the proposed method may also be interesting for controlling the column switch. A drawback of the study is that only a variation in mAb titer in the upstream was included into the study. Other variations, like the contaminant content in the HCCF or media component changes, were not investigated.

In a next step, the developed method with UV spectroscopy combined with PLS modeling was applied in a second study (see Chapter 5) to a revised design space, which included large process variation due to the use of different feedstocks with different mAb compositions, to test the robustness of the method and applicability to different products. The study showed, that the error of the method is increased due to the large design space. To overcome this challenge, a dynamic UV background subtraction based on the leveling out of the conductivity signal during the load phase was implemented to increase the prediction ability of the PLS model. It was demonstrated that by subtracting the background spectrum during the breakthrough, the prediction of the mAb concentration is facilitated and improved compared to models using the raw spectra. The conductivity-based background subtraction in combination with PLS modeling on UV spectra offers a robust quantification of the product breakthrough regardless of the large variability in the cell culture fluid. Additionally, it was shown, that by using the conductivity-based background subtraction, the use of a single absorption wavelength instead of a multivariate spectrum becomes feasible for the mAb quantification. This smart sensor concept shows great potential for application to production processes as the required univariate sensors are already implemented in most processes.

Even though, the results of the study presented in Chapter 5 are promising, it also shows, that UV spectroscopy lacks the selectivity to distinguish between the mAb and contaminants, which makes the background subtraction necessary. Other spectroscopy methods, especially Raman spectroscopy, have proven to be more selective. Therefore, Raman spectroscopy is frequently used in upstream processing to differentiate between various cell culture components and the product. A drawback of Raman spectroscopy is the long measurement time as the Raman effect is very weak compared

to absorption phenomena. In recent years, the application of Raman spectroscopy to the downstream process became feasible due to an increased measurement speed by instrumentation improvements.

In Chapter 6, the application of both Raman and UV spectroscopy for monitoring the Protein A load phase was presented to compare both methods and to evaluate the benefit of a combination of both methods by data fusion. As data fusion techniques, hierarchical PLS modeling, and Convolutional Neural Networks (CNNs) were tested. If no preprocessing was applied to the spectra, it was shown that UV spectroscopy has a slightly better prediction accuracy in comparison to Raman spectroscopy. However, when the dynamic background subtraction (developed in Chapter 5) is applied, the prediction accuracy of the UV-based models improves 20-fold. For Raman spectroscopy, the background subtraction does not improve the prediction ability, probably due to the increased noise. It seems, that Raman spectroscopy is more selective than UV spectroscopy, which might make a background subtraction not helpful. The main drawback of Raman spectroscopy was the observed non-linearities in the spectra, which led to an increased number of Latent Variables (LVs) of the PLS models and a lower model prediction in comparison to the UV-based model. PLS models as linear regression techniques are only able to fit non-linearities to a certain extend. The larger the design space, the worse the linear approximation. CNNs were applied for non-linear regression to overcome this problem. CNNs can improve the prediction ability slightly in comparison to PLS-based methods, but the training of CNNs is challenging, requires a larger amount of data and might converge to different solutions. Besides the evaluation of PLS and CNN models, data fusion algorithms were tested to potentially improve the prediction accuracy by combining the sensitivity of UV spectroscopy with the selectivity of Raman spectroscopy. However, no improvement was observed in comparison to the solely UV-based models. Even though, the combination of the high signal-to-noise ratio of the UV measurements with the selectivity of the Raman measurements seems promising, for the purpose of quantifying only the mAb concentration, UV-based methods, especially in combination with a background subtraction, seem to be the best option. Nevertheless, UV spectroscopy cannot monitor other attributes of interest, like the buffer composition, aggregate content or disulfide bridges which makes other sensor concepts necessary.

An example for a process step, where multiple attributes need to monitored to facilitate process development and assure consistent quality in production processes, is the the combined process step Ultrafiltration/Diafiltration (UF/DF). For UF/DF processes not only the monitoring of the product concentration, but also of the buffer composition and aggregate

content is important. In the study presented in Chapter 7, a lab-scale Cross-Flow Filtration (CFF) device was equipped with a Variable Pathlength (VP) Ultraviolet/Visible (UV/Vis) spectrometer, a light scattering photometer, and a micro Liquid Density Sensor (microLDS). The protein concentration was measured by VP UV/Vis spectrometer. Due to the large concentration range of the UF/DF step, the use of the VP technology was necessary to avoid detector saturation. The buffer exchange was monitored by density measurements of the microLDS. To calculate the apparent molecular weight, both the protein concentration determined by the VP UV/Vis spectrometer and the Static Light Scattering (SLS) signal measured by the light scattering photometer were necessary. The average hydrodynamic radius was calculated by the Dynamic Light Scattering (DLS) signal of the light scattering photometer, which was corrected by the viscosity determined by the microLDS. The setup was tested in three case studies to show the full potential of this setup. Off-line and on-line measurements were always in good agreement, if no protein precipitation occurred. The protein concentration could be monitored in-line and in a large concentration range. The buffer-dependent increase in apparent molecular weight of the mAb could be shown during diafiltration, giving valuable information for process development and stability assessment. The developed sensor concept was shown to be a powerful tool for monitoring protein concentration, buffer exchange, apparent molecular weight and hydrodynamic radius. The in Chapter 7 presented case study highlights the need for smart sensor concepts to measure all quality attributes of interest, which was not possible with a single sensor.

While aggregates and other product species may form during the UF/DF process, providing real-time information on their content is mainly interesting during process development. Instead, for the final UF/DF step in production, it is essential to achieve other quality metrics such as a sufficient buffer exchange and a compliant product concentration. Monitoring the protein and buffer components concentrations enables process automation of the UF/DF process by switching to the next process phase, when either the desired protein concentration or buffer component concentration are reached. Chapter 8 builds on the process monitoring foundations presented in Chapter 7 to enable more process automation. A Raman analyzer was implemented in the setup, because Raman spectroscopy is capable of measuring the protein concentration and Raman-active buffer components simultaneously. As the protein concentrations observed in the UF/DF process are significantly higher than during the Protein A load phase (presented in Chapter 6), Raman showed comparable results to UV spectroscopy for quantification of the protein concentration. However, the noise level in both the buffer signal of the Raman spectra and density were too large to allow for a process automation without data preprocessing. Therefore, an Extended Kalman Filter (EKF) was implemented to combine mechanistic process knowledge with the data to estimate the state of the process more accurately and thereby allow for process automation.

In summary, the potential of different spectroscopic methods to monitor the downstream process was evaluated in this thesis. Commonly implemented univariate sensors were evaluated to close the gaps of spectroscopic techniques or facilitate the implementation of the PAT methods. Smart sensor concepts for the Protein A capture step and the UF/DF step were introduced. Additionally, data fusion techniques and new concepts in machine learning, especially CNNs and an EKF, were evaluated for their ability to improve the prediction ability of spectroscopic methods. While CNNs can automate the preprocessing optimization in the convolutional layers and apply non-linear regression techniques in the fully connected layers, the performance in the tested case study did not justify the computational effort in comparison to PLS models. The implementation of an EKF on the other hand showed promising results, as mechanistic process knowledge is combined with the spectroscopic data, which allows for a more accurate prediction of the process state. As a result of the carried-out scientific studies, this thesis facilitates the implementation of PAT methods in the downstream process of biologics, because solutions to specific monitoring needs of the capture and UF/DF step are presented. As the capture and UF/DF step are drives of purification process costs due to the high cost of the capture resin and the high value of the purified product for the UF/DF step, a contribution to make critical biopharmaceutical drugs more affordable was made. The presented smart sensor concepts show potential to improve the monitoring and automation of productions processes resulting in more efficient and robust processes.

Zusammenfassung

Die rasant steigende Anzahl an Biopharmazeutika und deren Nachahmerprodukten, sogenannte *Biosimilars*, auf der Liste der essentiellen Medikamente der Weltgesundheitsorganisation, unterstreicht die Bedeutung von biopharmazeutischen Arzneimitteln für die globale Gesundheitsversorgung. Biopharmazeutika verbessern die Überlebensraten für eine steigende Anzahl an Patienten mit bisher unheilbaren oder nicht behandelbaren Indikationen. Unabhängig von den Therapieerfolgen sind Biopharmazeutika mit hohen sozioökonomischen Kosten verbunden, da die Behandlungskosten im Vergleich zu chemisch hergestellten Medikamenten durchschnittlich zwanzigmal höher sind. Daher steigt der Anteil der Biopharmazeutika an den gesamten Arzneimittelausgaben stetig an.

Eine Ursache für die höheren Kosten von Biopharmazeutika sind die schwierigere Aufreinigungsentwicklung und Herstellung. Biopharmazeutika sind therapeutische Proteine, die aufgrund ihrer großen und komplexen chemischen Struktur nicht chemisch synthetisiert werden können. Daher werden Biopharmazeutika in Tier-, Hefe- oder Bakterienzellen hergestellt. Zellen stellen ein anspruchsvolles Herstellungssystem dar, da sie nicht nur das gewünschte Produkt, sondern auch Variationen des Produkts oder Verunreinigungen in unterschiedlichen Mengen produzieren, die auf ein konsistentes und sicheres Niveau abgereichert werden müssen. Wegen der komplexen Struktur, chemischen Ähnlichkeit zu Kontaminanten und Instabilität der Biopharmazeutika,sind die Produktionsprozesse, im Vergleich zur chemischen Synthese, komplexer und schwieriger zu überwachen und zu steuern.

Ein Ansatz zur Überwachung des Aufreinigungsprozesses von Biopharmazeutika sind prozessanalytische Technologien (engl.: Process Analytical Technology *PAT*). Das Ziel von *PAT* ist es, zeitnah kritische Qualitäts- und Leistungsattribute des Prozesses zu messen, um den Prozess langfristig zu steuern. Dadurch garantieren *PATs* nicht nur die Qualität des Produktes, sondern ermöglicht auch eine Prozessoptimierung, die zu einer Reduzierung der Produktionskosten von Biopharmazeutika führt.

Die enge chemische Verwandtschaft zwischen Biopharmazeutika und ihren Verunreinigungen führt jedoch zu hohen Anforderungen an die Selektivität von *PATs* zur Quantifizierung dieser kritischen Eigenschaften. Oft ist eine *PAT* nicht selektiv genug, um zwischen Produkt und spezifischer Verunreinigung zu unterscheiden. Daher kann die Kombination verschiedener PATs für eine adäquate Kontrolle des Aufreinigungsprozesses notwendig sein.

Das Ziel dieser Arbeit war es, intelligente Sensorkonzepte als *PAT*-Werkzeuge für den Aufreinigungsprozess von Biopharmazeutika zu entwickeln. Dafür musste zunächst die Messbarbeit von Proteineigenschaften in wässrigen Lösungen durch verschiedene Sensoren verstanden werden. Voraussetzungen für die Anwendbarkeit zur Echtzeitüberwachung sind der Betrieb in einer Durchflussanordnung, entweder im Prozessstrom selbst oder in einem Bypass, und die generelle Beobachtbarkeit von proteinbezogenen Merkmalen. Insbesondere optische Spektroskopie erfüllt diese Anforderungen. Daher wird in Kapitel 3 die Sensitivität und Selektivität von optischer Spektroskopie zur Quantifizierung von Proteinen im Aufreinigungsrozess bewertet. Aufgrund der engen chemischen Verwandtschaft des gewünschten Produkts mit den Verunreinigungen ist auch die Selektivität für die Messung verschiedener Proteinstruktureigenschaften wichtig für die Anwendbarkeit der spektroskopischen Techniken. Daher liegt ein Schwerpunkt des Kapitels 3 auf der Messbarkeit verschiedener Strukturniveaus von Proteinen und damit der Differenzierung zwischen Proteinen durch verschiedene spektroskopische Methoden. Zusätzlich wird eine Anleitung zur Kalibrierung und Validierung von *Partial Least Squares* (*PLS*)-Modellen gegeben, um häufige Fehler bei der Datenanalyse zu vermeiden. Durch diese Evaluierung wurde die UV-Spektroskopie aufgrund der hohen Absorptionskoeffizienten von Proteinen und der geringen Absorption von Wasser als die empfindlichste spektroskopische Technik zur Messung von Proteinen in wässriger Lösung identifiziert. Allerdings fehlt der UV-Spektroskopie eine hohe Selektivität, um zwischen verschiedenen Proteinen zu unterscheiden, da die Banden verschiedener Strukturelemente breit und überlappend sind. Hier scheint die Raman-Spektroskopie aufgrund der Verfügbarkeit der primären, sekundären und tertiären Proteinstruktur in den Spektren vielversprechend zu sein. Die einzige spektroskopische Technik, die die häufigste Form von aggregierten Proteinen mit Hauptänderungen im quartären Strukturniveau messen kann, ist die Lichtstreuung. Leider gibt es keinen Sensor, der alle kritischen Qualitätsattribute im Aufreinigungsprozess messen kann. Dafür sind Kombinationen von verschiedenen Sensoren notwendig. Daher enthält Kapitel 3 eine Zusammenfassung von Datenfusionstechniken zur Auswertung der Multiblockdaten der verschiedenen Sensoren.

Als erstes Beispiel wird in Kapitel 4 die Implementierung der UV-Spektroskopie in Kombination mit der *PLS*-Modellierung als konzeptioneller Beweis für die Echtzeitüberwachung und -steuerung der Protein A-Beladungsphase gezeigt. Es wurde gezeigt, dass die *PLS*-Modellierung auf UV-Absorptionsspektren angewendet werden kann, um die Konzentration des monoklonalen Antikörper (engl.: *mAb*) am Säulenausgang während der Beladephase in Anwesenheit vieler protein- und nicht-proteinbasierter Verunreinigungen zu quantifizieren. Basierend auf der Quantifizierung wurde die Beladephase automatisch beendet, wenn eine vorher festgelegte *mAb*-Konzentration erreicht wurde. Folglich hat die vorgeschlagene Methode Potenzial für die Überwachung und Steuerung von Affinitätsschritten in der industriellen Produktion. Durch die gezeigte Methode kann das Beladungsvolumen mit der vorgeschlagenen Methode dynamisch bestimmt werden, was eine höhere Auslastung der Harzkapazitäten ermöglicht und gleichzeitig den Produktverlust gering hält. Außerdem kann auf die zeitaufwändige Bestimmung des *mAb*-Titers in der Fermentationsbrühe verzichtet werden. Für die kontinuierliche Chromatographie könnte die vorgeschlagene Methode auch für die Steuerung des Säulenwechsels bei Affinitässchritten interessant sein. Ein Nachteil der Studie ist, dass nur die erwartete Variation des *mAb*-Titers berücksichtigt wurde. Andere Variationen, wie z.B. der Kontaminationsgehalt oder Änderungen der Medienkomponenten, wurden nicht untersucht.

Deshalb wurde die Methode mit UV-Spektroskopie in Kombination mit *PLS*-Modellierung in Kapitel 5 auf einen überarbeiteten Prozessraum angewendet, der große Prozessvariationen aufgrund der Verwendung verschiedener Ausgangsstoffe mit unterschiedlichen *mAb*-Zusammensetzungen enthielt. Damit wurde die Robustheit der Methode und die Anwendbarkeit auf verschiedene Produkte getestet. Die Studie zeigte, dass der Fehler der Methode aufgrund des großen Prozessraums erhöht ist im Vergleich zur Machbarkeitsstudie aus Kapitel 4. Um die Präzision der Methode zu erhöhen, wurde eine dynamische UV-Hintergrundsubtraktion implementiert, die auf dem Leitfähigkeitssignals während der Beladung basiert. Durch die Subtraktion des Hintergrundspektrums während des Durchbruchs wurde die Vorhersage der *mAb*-Konzentration im Vergleich zu Modellen, die die Rohspektren verwenden, verbessert. Die leitfähigkeitsbasierte Hintergrundsubtraktion in Kombination mit der UV-Spektren basierten *PLS*-Modellierung bietet eine robuste Quantifizierung des Produktdurchbruchs unabhängig von großen Variabilitäten in der Zellkulturflüssigkeit. Zusätzlich wurde gezeigt, dass durch die leitfähigkeitsbasierte Hintergrundsubtraktion die Verwendung einer einzelnen Absorptionswellenlänge anstelle eines multivariaten Spektrums für die *mAb*-Quantifizierung machbar wird. Dieses intelligente Sensorkonzept

zeigt großes Potenzial für die Anwendung in Produktionsprozessen, da die erforderlichen Sensoren in den meisten Prozessen bereits implementiert sind.

Obwohl die Ergebnisse aus Kapitel 5 vielversprechend sind, zeigen sie auch, dass der UV-Spektroskopie die Empfindlichkeit fehlt, um zwischen *mAb* und Verunreinigungen zu unterscheiden, was die Hintergrundsubtraktion notwendig macht. Andere Spektroskopiemethoden, insbesondere die Raman-Spektroskopie, haben sich als selektiver erwiesen. Daher wird die Raman-Spektroskopie häufig während der Fermentation eingesetzt, um zwischen verschiedenen Zellkulturkomponenten und dem Produkt zu unterscheiden. Ein Nachteil der Raman-Spektroskopie sind die benötigten Messzeiten, da der Raman-Effekt im Vergleich zu Absorptionsphänomenen sehr schwach ist. In den letzten Jahren wurde die Anwendung der Raman-Spektroskopie während des Aufreinigungsprozess aufgrund einer verbesserten Messgeschwindigkeit durch technische Fortschritte im Spektrometer möglich.

In Kapitel 6 wurde die Anwendung sowohl der Raman- als auch der UV-Spektroskopie zur Überwachung des Protein-A-Schrittes vorgestellt, um beide Methoden zu vergleichen und den Nutzen einer Kombination beider Methoden zu bewerten. Als Datenfusionstechniken wurden die hierarchische *PLS*-Modellierung und *Convolutional Neural Networks* (*CNNs*) getestet. Ohne Spektrenvorverarbeitung wies die UV-Spektroskopie im Vergleich zur Raman-Spektroskopie eine bessere Vorhersagegenauigkeit auf. Wenn jedoch die dynamische Hintergrundsubtraktion (entwickelt in Kapitel 5) angewendet wird, verbessert sich die Vorhersagegenauigkeit der UV-basierten Modelle um rund das 20-fache. Bei der Raman-Spektroskopie verbessert die Hintergrundsubtraktion die Modellgenauigkeit nicht, was wahrscheinlich auf die Erhöhung des Signal-zu-Rausch- Verhältnisses zurückzuführen ist. Es scheint, dass die Raman-Spektroskopie selektiver ist als die UV-Spektroskopie, wodurch eine Hintergrundsubtraktion nicht hilfreich sein könnte. Ein Hauptnachteil der Raman-Spektroskopie waren die beobachteten Nichtlinearitäten in den Spektren, die zu einer erhöhten Anzahl von latenten Variablen der *PLS*-Modelle und einer geringeren Modellvorhersage im Vergleich zum UV-basierten Modell führten. *PLS*-Modelle sind als lineare Regressionsverfahren nur in der Lage, Nichtlinearitäten als lineare Approximationen anzupassen. Je größer der Prozessraum ist, desto schlechter ist die lineare Approximation. Um dieses Problem zu überwinden, wurden *CNNs* für die nichtlineare Regression eingesetzt. *CNNs* kann die Vorhersagefähigkeit im Vergleich zu *PLS*-basierten Methoden leicht verbessern, aber das Training von CNNs ist anspruchsvoll, erfordert eine größere Menge an Daten und kann zu unterschiedlichen Lösungen konvergieren. Neben der Auswertung von *PLS*- und *CNN*-Modellen wurden Datenfusionsalgorithmen getestet, um die Vorhersagegenauigkeit durch die Kombination der Empfindlichkeit der UV-

Spektroskopie mit der Selektivität der Raman-Spektroskopie möglicherweise zu verbessern. Es wurde jedoch keine Verbesserung im Vergleich zu den ausschließlich UV-basierten Modellen beobachtet. Auch wenn die Kombination des hohen Signal-zu-Rausch-Verhältnisses der UV-Messungen mit der Selektivität der Raman-Messungen vielversprechend erscheint, scheinen für den Zweck, nur die *mAb*-Konzentration zu quantifizieren, UV-basierte Methoden, insbesondere in Kombination mit einer Hintergrundsubtraktion, die beste Option zu sein. Auch wenn die UV-Spektroskopie am besten geeignet scheint, um die Beladungsphase des Protein-A-Schrittes zu überwachen, bedeutet dies nicht, dass sie auch für andere Prozessschritte im Aufreinigungsprozess geeignet ist. So kann die UV-Spektroskopie nicht für Prozesschritte eingesetzt werden, die der Überwachung der Pufferzusammensetzung, des Aggregatgehalts oder Disulfidbrückenformation bedürfen.

Ein Prozessschritt, bei dem nicht nur die Überwachung der Proteinkonzentration, sondern auch der Pufferzusammensetzung und des Aggregatgehalts wichtig ist, ist die Ultrafiltration/Diafiltration (UF/DF). Die Überwachung der genannten Attribute könnte nicht nur die Prozessentwicklung erleichtern, sondern auch eine gleichbleibende Qualität in Produktionsprozessen gewährleisten. In der in Kapitel 7 vorgestellten Studie wurde ein Gerät zur tangentialen Flussfiltration im Labormaßstab mit einem variablen Pfadlängen (VP) UV/Vis-Spektrometer, einem Lichtstreuungsphotometer und einem microLDS ausgestattet. Die Proteinkonzentration wurde durch VP UV/Vis-Spektroskopie gemessen, um den großen Konzentrationsbereich der Ultrafiltration (UF) abzudecken. Der Pufferaustausch wurde durch Dichtemessungen des *microLDS* überwacht. Zur Berechnung des scheinbaren Molekulargewichts sind sowohl die durch das UV/Vis-Spektrometer gemessene Proteinkonzentration als auch das vom Lichtstreuphotometer gemessene statische Lichtstreuungssignal erforderlich. Der mittlere hydrodynamische Radius wurde aus dem dynamischen Lichtstreuungssignal berechnet, das um die mit dem microLDS bestimmte Viskosität korrigiert wurde. Der Aufbau wurde in drei Fallstudien getestet, um das volle Potenzial dieses Aufbaus zu zeigen. Offline- und Online-Messungen waren immer in guter Übereinstimmung mit Korrelationskoeffizienten von über $0,92$, wenn keine Proteinausfällung auftrat. Zudem konnte die Proteinkonzentration in einem großen Bereich überwacht werden. Der pufferabhängige Anstieg des scheinbaren Molekulargewichts des *mAb* konnte während der Diafiltration gezeigt werden, was wertvolle Informationen für die Prozessentwicklung und Stabilitätsbewertung liefert. Das entwickelte Sensorkonzept hat sich als leistungsfähiges Werkzeug zur Überwachung von Proteinkonzentration, Pufferaustausch, scheinbarem Molekulargewicht und hydrodynamischem Radius erwiesen. Die in Kapitel 7 vorgestellte Studie zeigt, dass es oft nicht

möglich ist, alle wichtigen Qualitätsmerkmale mit nur einem Sensor zu messen. Daher sind intelligente Sensorkonzepte notwendig, um möglichst viele kritische Qualitätsmerkmale mit einer möglichst geringen Anzahl an Sensoren zu messen.

Zwar können sich während des UF/DF-Prozesses Aggregate und andere Produktvarianten bilden, doch ist die Bereitstellung von Echtzeitinformationen über deren Gehalt hauptsächlich während der Prozessentwicklung interessant. Für den abschließenden UF/DF-Schritt in der Produktion ist es stattdessen von entscheidender Bedeutung, andere Qualitätskennzahlen, wie einen vollständigen Pufferaustausch und eine konforme Endproduktkonzentration zu erreichen. Die Überwachung der Konzentration von Proteinen und Pufferkomponenten ermöglicht daher die Automatisierung des UF/DF-Prozesses durch Umschalten auf die nächste Prozessphase, wenn entweder die gewünschte Proteinkonzentration oder vollständige Pufferaustausch erreicht ist. Kapitel 8 baut auf den in Kapitel 7 vorgestellten Grundlagen der Prozessüberwachung auf, jedoch wurden Erweiterungen getroffen, die eine Prozessautomatisierung ermöglichen. Zusätzlich wurde ein Ramanspektrometer in den Aufbau implementiert, da die Ramanspektroskopie in der Lage ist, die Proteinkonzentration und eine Vielzahl von Raman-aktiven Pufferkomponenten gleichzeitig zu messen. Da die während des UF/DF Prozesses beobachteten Proteinkonzentrationen deutlich höher sind als während der Protein-A Beladungsphase (siehe Kapitel 6), zeigte die Ramanspektroskopie zur Quantifizierung der Proteinkonzentration vergleichbare Ergebnisse wie die UV-Spektroskopie, auch wenn die Quantifizierung wiederum auf dem Anstieg des Hintergrundspektrums/der Basislinie beruhte. Der Rauschpegel sowohl im Puffersignal der Raman-Spektren als auch in der Dichte war jedoch zu groß, um eine Prozessautomatisierung ohne Datenvorverarbeitung zu ermöglichen. Daher wurde ein *EKF* implementiert, um mechanistisches Prozesswissen mit den Daten zu kombinieren, um den Zustand des Prozesses genauer abzuschätzen und dadurch eine Prozessautomatisierung zu ermöglichen.

Zusammenfassend wird in dieser Arbeit das Potenzial verschiedener spektroskopischer Methoden zur Überwachung des Aufreinigungsprozesses bewertet. Gängige, univariate Sensoren wurden evaluiert, um die Lücken der spektroskopischen Verfahren zu schließen oder die Implementierung der *PAT*-Methoden zu erleichtern. Intelligente Sensorkonzepte für den Protein A-Schritt und den UF/DF-Schritt wurden vorgestellt. Zusätzlich wurden Datenfusionstechniken und neue Konzepte des maschinellen Lernens, insbesondere *CNNs* und einen *EKF*, als Regressionsmethoden untersucht. Während *CNNs* die Vorverarbeitungsoptimierung in den Faltungsschichten automatisieren und nichtlineare Regressionstechniken in den vollverknüpften

Schichten anwenden kann, rechtfertigte die Leistung in dem getesteten Anwendungsfall den Rechenaufwand im Vergleich zu *PLS*-Modellen nicht. Die Implementierung eines EKF hingegen zeigte vielversprechende Ergebnisse, da mechanistisches Prozesswissen mit den spektroskopischen Daten kombiniert wird, was eine genauere Vorhersage des Prozesszustandes ermöglichte. Als Ergebnis der durchgeführten wissenschaftlichen Studien erleichtert diese Arbeit die Implementierung von *PAT* im Aufreinigungsprozess von Biologika, da Lösungen für spezifische Überwachungsbedürfnisse des Protein A-Schritt- und UF/DF-Schrittes vorgestellt werden. Somit wird durch die vorgestellten intelligenten Sensorkonzepte ein Beitrag geleistet, kritische biopharmazeutische Medikamente erschwinglicher zu machen, um Produktionsprozesse zu verbessern.

Contents

Laura Rolinger, Matthias Rüdt, Jürgen Hubbuch

Matthias Rüdt[1], Nina Brestrich[1], Laura Rolinger, Jürgen Hubbuch ([1] contributed equally)

1 Introduction

In 2019, the WHO recognized the therapeutical equivalency of biosimilars to the originator biopharmaceuticals and added biosimilars of the essential medicines rituximab and trastuzumab to the EML [1, 2]. As the purpose of the EML is to summarize the medicines for the most important health care needs [3], the addition of an increasing number of biosimilars not only emphasizes the importance of biopharmaceuticals for the global health, but also the importance of affordable medicine. Biopharmaceuticals improve survival rates for a rising number of patients with previously incurable or untreatable indications, like immunologic diseases, or advanced melanoma [4, 5] or, HER2-overexpressed metastatic breast cancer [6]. Regardless of the treatment revolution for unmet indications, biopharmaceuticals come at major socioeconomic costs [7]. Biopharmaceutical treatment costs regularly between 20 k€ to 200 k€ per patient per year [8], which is 20 times higher compared to traditional chemically produced, so called small molecules or synthesized molecules, treatments [9]. Already in 2005, biopharmaceuticals accounted for 18 % of the total drug expenditures in the United States of America (USA) and the costs are rising [10].

Even though, the overall costs for the approval of drugs are comparable between small molecules and biopharmaceuticals [11], biopharmaceuticals face timeline delays for the supply of the first clinical phase and higher costs for purification development and manufacturing [11, 12]. These effects are likely related to the more challenging production of biopharmaceuticals. Biopharmaceuticals are large therapeutic proteins, which cannot be chemically synthesized due to their large and complex chemical structure. Therefore, biopharmaceuticals are produced by animal, yeast, or bacteria cells [13].

Cells represent a challenging manufacturing system, as they produce not only the desired product, but also variations of the product or impurities in various amounts, which need to be depleted to consistent and safe levels before the drug can be administered to patients [14]. As a consequence, biopharmaceutical production processes can be more complex, and harder to monitor and control compared to chemical synthesis.

One approach to monitor the purification process of biopharmaceuticals are PAT methods [15, 16]. The goal of PAT is to measure in a timely manner critical quality and performance attributes of the process [17] to control the process in the long run. Thereby, PAT methods not only guarantee the quality of the product, but also allow for process optimization leading to a reduction in the production costs of biopharmaceuticals.

However, the close chemical relation between biologics and their impurities leads to high demands for the selectivity of the PAT method for the quantification of those critical attributes. Often one PAT method is not selective enough to distinguish between the product and a specific impurity. Therefore, the combination of different PAT methods can be necessary for adequate control of the purification process.

Early applications of PAT methods for biopharmaceuticals used on-line analytical chromatography for process monitoring and control [18–22], which combines automated sampling with a standard analytical method usually used for off-line quality attribute monitoring. The on-line application of analytical chromatography via High Performance Liquid Chromatography (HPLC) provides high resolution separation and quantification of different species. As HPLC relies on the separation of different protein species by chromatography, it produces a time-delay between sample drawing and analysis result. Depending on the decision time of a unit operation, this may lead to a late notice of process deviations or even completely prevent real-time monitoring. Additionally, chromatographic methods have high maintenance efforts due the buffer consumption and column aging.

An alternative can be the implementation of in-line sensors. Usually, the basic steps in the purification process are monitored by fairly simple in-line sensors, like pH, conductivity, or density. However if a differentiation between different protein is necessary, more advanced techniques, like spectroscopy, are required.

The following sections will give an overview of the production process of monoclonal antibodies (mAbs), PAT and a detailed introduction of optical spectroscopy as PAT method will be presented. The final section addresses the necessary data analysis for multivariate analysis.

1.1 Monoclonal Antibodies

Antibodies, also known as Immunoglobulins (Igs), are immunoreactive proteins that combine the ability to recognize invading pathogens by binding to surface antigens and trigger potent effector mechanisms for pathogen elimination [23]. mAbs are secreted by identical plasma cells cloned from a single parent cell [24]. Therefore, mAbs are highly similar and usually directed to a single epitope on an antigen surface. Due to their high specificity, mAbs have become one of the most important classes of biopharmaceutical products against cancer or chronic diseases [25], which makes mAbs also highly profitable due to the large quantities required for therapy [26]. The mass production of mAbs was enabled in 1975 by Köhler and Milstein, when they fused myeloma cell lines with B-cells resulting in identical immortalized cells, so called hybridoma cells [27]. Hybridoma cells combine the ability of the B-cell to produce antibodies with the longevity and reproductive fertility of the myeloma cells [28]. Due to the size and complex chemical structure of mAbs, like most proteins, cannot be produced by chemical synthesis, but require cells as expression systems.

1.1.1 Antibody Structure

Generally, antibodies, like all proteins, consist of a chain of amino acids. Amino acids are composed of a carboxyl group ($-\mathrm{COOH}^-$) and an amino group ($-\mathrm{NH}_2$) bonded to the same carbon atom along with a side chain. The side chain is specific for every amino acid. The amino acids tryptophan, tyrosine, and phenylalanine have aromatic residues, which can be measured by spectroscopy and used for protein quantification and identification [29] as explained in detail in Section 1.4. Amino acids can be linked by the removal of water from the carboxyl group and the amino group from the next amino acid. This reaction is called condensation or dehydration and the resulting covalent linkage is called peptide bond. When many amino acids are linked by peptide bonds, it is called a polypeptide. A Polypeptide with a molecular mass higher than 10 000 Da is referred to as protein. [30]

The linear sequence of amino acids is referred to as the primary structure. Due to the high Gibbs free energy, the linear chain is unstable and in order to reach a stable point, proteins are folded in the native form. Pauling and Corey [31] shaped the understanding of this spatial conformation, which was separated by them into secondary, tertiary, and quaternary structure. The secondary structure is the local conformation of protein segments as a result of hydrogen bonds between the C=O and the N-H groups of the peptide bond. The most prominent regular folding patterns of the polypeptide backbone are the α helix and β sheet conformations. [30]

The α helix is a right-handed helical structure, such that the polypeptide backbone is tightly wound around a longitudinal axis through the center of the helix. The side chains of the amino acids protrude outward of the helical backbone. Hydrogen bonds are formed between the amine hydrogen atom of the amino acid n and the carbonyl group of the amino acid n+3. [30]

Not all polypeptides can form a stable α helix due to interactions between amino acid side chains. Therefore a second repetitive structure was predicted by Pauling and Corey, called β sheet [31]. In this spatial arrangement, the polypeptide backbone is extended to a linear zigzag structure and hydrogen bonds are formed between vicinal chains. These chains can be linked in a parallel direction, which means that all amino acid chains have the same amino-to-carbonyl direction, or antiparallel direction, such that the amino acid chain direction is vice versa [30]. If an α or β conformation is not possible due to the arrangement of the amino acid side chains, the secondary structure is statistically distributed, called random coil [32]. As the hydrogen bond length between the carbonyl and amino group differs depending on the sort of secondary structure, spectroscopic techniques, like Fourier-Transform Infrared (FTIR) and Raman spectroscopy, can be used for structure identification of proteins [33].

The tertiary structure refers to a long-range arrangement of distant amino acid sequences, which reside due to different secondary structures. The secondary structure segments are linked by weak molecular interactions or sometimes by covalent disulfide bonds [30]. Changes in the tertiary structure can result in a change in the exposure of amino acid residues to the protein surface. If, for example, an aromatic amino acid of a protein in an aqueous solution moves to the surface of the protein, the change in hydrophobicity of the environment induces a spectral shift, which can be observed by Raman or UV/Vis spectroscopy [34].

The arrangement of these tertiary structure subunits in the three-dimensional space constitutes the quaternary structure [30]. The quaternary structure describes the size and form of the protein, which can be measured by light scattering. The principles to measure the molecular weight and hydrodynamic by light scattering are explained and Section 1.4 and the use of this measurement in the downstream process is shown in Chapter 7. The quaternary structure gives proteins some flexibility, which improves the ability to bind to other molecules.

All antibodies have a resembling basic structure composed of two identical heavy chains and light chains, respectively. There are five classes, in which antibodies are divided due to their structure, function, and distribution in the body: IgM, IgD, IgG, IgE, and, IgA [35]. IgG is the main antibody class in the bloodstream and has a crucial role in eliminating invading pathogens. As an example of an antibody, an IgG is depicted in Figure 1.1. Every light chain is linked by an inter-chain disulfide bond to a heavy chain. A light chain has a molecular weight of 23 kDa and consists of one Variable Region V_L (yellow) and one Constant Region C_L(light green). All regions are held together by an intra-chain disulfide bond. Heavy chains have a molecular weight of 50-70 kDa and consist of one Variable Region V_H and three Constant Regions $C_{H1} - C_{H3}$. This means, that all IgGs have similar biophysical and biochemical properties, because only amino acid sequences in two regions are variable. [36, 37]

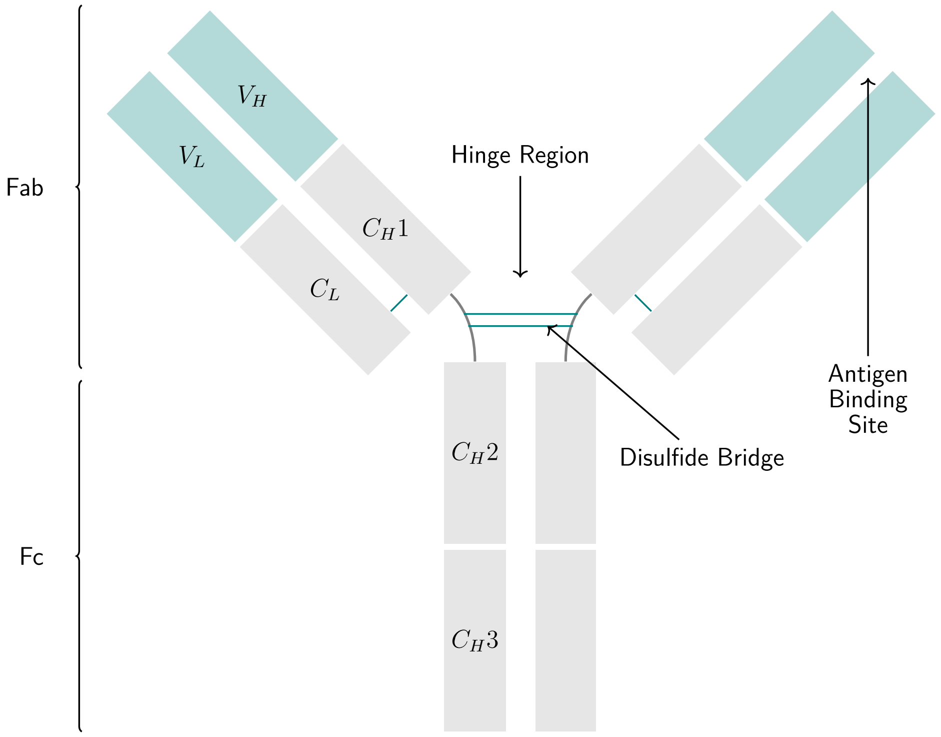

Figure 1.1: Typical IgG structure composed of two identical heavy chains and light chains, which are connected by disulfide bonds. The heavy chain comprises three constant regions (C_H) and one variable region (V_H). The light chain comprises also one constant region (C_L) and one variable region (V_L) with an antigen binding site at both ends. Adapted from [38] .

The antibody can be divided into three functional regions. The hinge region is located in the middle of the antibody, where both arms form a Y. This region allows some flexibility, which is needed due to varying spaces between epitopes on antigen, where the antigen binding sides, called paratopes, can bind [35]. The Fragment crystallized (Fc) region can bind to macrophages, which can engulf and digest aggregates of antigens and pathogens [37]. This cleaning process also triggers an immune reaction to cope with further pathogens. The arms of the antibody are called Fragment antigen binding (Fab) region, because they have an antigen binding site at the end of each variable region [35]. While traditional mAbs are monospecific, which means they have the same antigen binding side at each variable region, a bispecific antibody (bsAb) recognizes two different epitopes either on the same or on different antigens [39].

1.1.2 Bispecific Antibodies

bsAbs have gained increasing interest over recent years with their wide range of applications including diagnosis, imaging, prophylaxis, and therapy [40]. Due to their dual specificity, bsAbs can, for example, bind to the target cells using one antigen-binding site and bring other cells or molecules in close proximity by binding to the second antigen-binding site. Therefore the initial therapeutic application of bsAb focused mainly on redirecting different effector cell to a cancer cell which cannot be simultaneously recruited to tumor cells by normal antibodies [39].

Initially, bsAbs were produced by chemical conjugation of two different mAbs or by fusing two hybridomas resulting in a quadroma cell line producing [41, 42]. Recombinant Deoxyribonucleic Acid (DNA) engineering of the cells enabled direct expression of engineered bsAbs and resulted in a range of recombinant bispecific antibody formats, with over 50 different formats now available [39]. In Figure 1.2 a bsAb is depicted with the expressed mispaired antibody variants. In this example, the same light chain is used in both parent antibodies to minimize the number of mispaired antibody variants [43].

1.2 Production Process of mAbs

The size and complexity of mAbs requires the production in mammalian host cell lines with Chinese Hamster Ovary (CHO) cells being the predominant host used to produce about 70% of recombinant proteins [44]. First, an aliquot from the working cell bank is defrosted and subsequently cultivated in vessels with increasing volume until enough cells are produced for transfer in the production bioreactor [45]. Once a certain viable cell density is reached, a temperature shift prompts the cells to secrete more mAb [45].

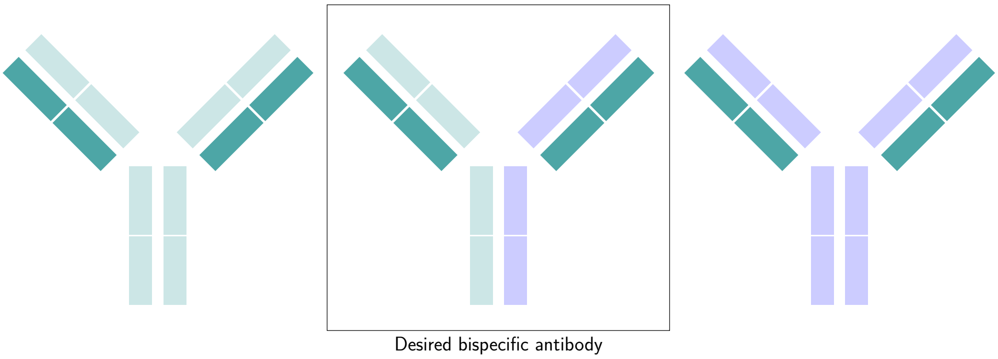

Figure 1.2: Structure of a bispecific antibody from two parent antibodies with identical light chain. Adapted from [43].

Cells not only produce the desired product, but also vital proteins and other by-products, which need to be depleted before administration to a patient [46]. The contaminants in the cell culture medium are differentiated into process-related impurities and product-related impurities [47]. Product-related impurities include protein variants such as aggregates, fragments, and, heterogeneities [26]. The most important process related impurities are Host Cell Proteins (HCPs), DNA, viruses, endotoxins, and leached Protein A [46]. In contrast to product-related impurities, process-related impurities are caused by the purification process and are not associated with the product.

Due to the very similar physicochemical properties, as mentioned in section 1.1.1, different mAbs can be purified by a process concept, called platform process, with minimal alterations in process parameters [48]. The platform process, as depicted in Figure 1.3, depletes product- and process-related contaminants to a level that the administration to the patient is considered safe. The first unit operation in a downstream process is the removal of cells and cell debris by centrifugation and depth filtration. This is called primary recovery or harvest and the supernatant is called HCCF. The primary recovery is followed by a capture step, in most cases Protein A chromatography, which is the "gold-standard" due to the high product purity of more than 98% [49]. Chromatography columns are blocked by cell debris, which makes complete removal in the primary recovery step essential [48]. At the fermentation pH of 6 to 8, the antibody binds to Protein A while the process-related impurities, such as HCPs, flows through. Mainly HCPs associated to the mAb by protein-protein interaction persist. These interactions can be disrupted by wash step with chaotropic agents at pH of 8 or higher, at which mAb-Protein A interactions are still strong [50]. The elution of the product is done at pH 2.5 to 4 with a following virus inactivation at this pH [46]. The capture step is followed by additional polishing steps, like cation exchange or anion exchange chromatography to further minimize contaminants. Then, the product is viral filtrated to provide a second orthogonal virus clearance step, which translates to approximately 12 to 18 log_{10} clearance of endogenous retroviruses during the whole process. Finally, the product is brought into formulation buffer by UF/DF. [46]

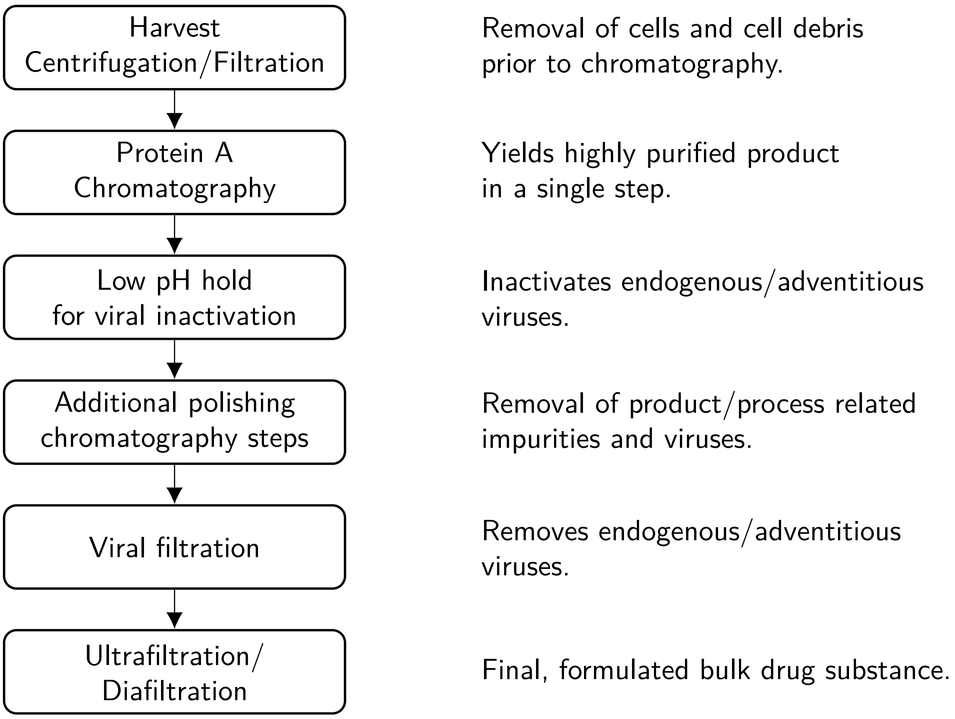

Figure 1.3: A typical platform process as downstream process of mAbs. [48]

Chromatography and UF/DF contribute most to the costs of the downstream process [51], probably due to the number of chromatographic steps and long process times of UF/DF. Protein A resin is also the most expensive resin in the platform process [52, 53]. Therefore, the implementation of PAT methods for the Protein A and UF/DF step are the economically most promising applications.

1.2.1 Protein A Chromatography

Protein A is derived from *Staphylococcus aureus* due to its ability to bind immunoglobulins [54]. In protein A chromatography, the Fc domain of mAbs

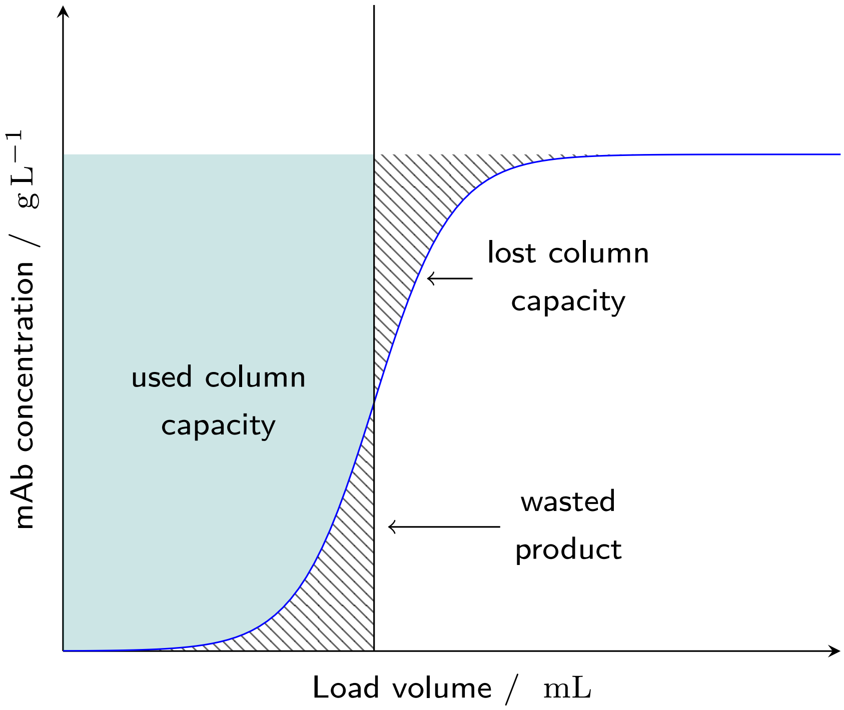

Figure 1.4: Dynamic breakthrough curve of a chromatography column. As the column is saturating with the loaded volume, the mAb concentration in the effluent increases. Adapted from [55].

binds to Protein A, while other molecules without Fc domain flow through resulting in an extensive removal of process-related impurities [50]. Due to the high costs of Protein A resin, an increase in used resin capacity by exhausting the binding places will directly result in fewer cycle times for the resin. A dynamic breakthrough curve is depicted in Figure 1.4 to illustrate the used column capacity. The feedstock, which is loaded onto the column has a certain mAb concentration. As the column reaches saturation, mAb starts to break through the column due to the mass transfer resistance of the resin beads [55]. The concentration in the effluent of the column increases further until all binding places are used and the mAb concentration in the column effluent reaches the mAb concentration in the feedstock. Depending on when the loading of the column is terminated, either more column capacity is left or more mAb is lost in the effluent of the column. Therefore, the ideal point for batch chromatography is to determine the point with as little as possible mAb in the breakthrough of the column.

For continuous processing modes, like for Periodic Counter Current Chromatography (PCCC), where the effluent of a first column is loaded onto a second column, a higher mAb concentration in the effluent is acceptable before the column positions are switched. In commercial manufacturing, ten cycles or more are used to process the HCCF on the protein A column to allow for smaller column volumes. The quantification of the mAb concentration in the column effluent would allow to use the resin capacity more efficiently by reduction of the cycle numbers. As the column lifetime is limited by the cleaning cycles, this would allow to process more mAb on one column and reduce costs. Karst et al . monitored the mAb concentration by at-line analytical chromatography for a two column continuous protein A chromatography process [56]. A drawback of the method is the time delay between sampling and analytical result, which leads to an increased wasted mAb mass in comparison to faster analytical techniques.

1.2.2 Ultrafiltration/Diafiltration

The purpose of the UF/DF step is to concentrate (UF) and buffer exchange (DF) the product. Therefore, UF/DF is often applied at the end of a process to prepare the product for formulation [57]. The two main filtration methods are dead-end filtration, in which the flow direction is perpendicular to the membrane surface, and CFF, in which the flow direction is parallel to the membrane [58]. CFF is the main filtration method used for UF/DF of mAbs, because the parallel flow can reduce the filtration cake formation on the membrane and thereby the flux is higher compared to dead-end filtration [59].

In Figure 1.5 is a UF/DF setup depicted. The feedstock is circulated by the pump over the membrane. The applied Transmembrane Pressure (TMP) forces molecules smaller than the membrane pore size to move through the membrane. The mAb remains in the UF/DF loop as it is significantly larger than the membrane pore size. The further the process runs, the more buffer molecules will leave the feedstock and, therefore, a concentration of the mAb takes places. When a diluent is added the buffer molecules of the diluent replace the original buffer system. Because no purification step follows the UF/DF step, it is important to minimize the formation of aggregates due to shear or water-air interfaces [60]. PAT applications for UF/DF can help to monitor the aggregate formation. Additionally, the buffer exchange progress can be monitored, which allows optimizing the process run time or enables the observation of exchange delaying effects, like the Donnan effect due to electrostatic interaction between proteins and excipients [61].

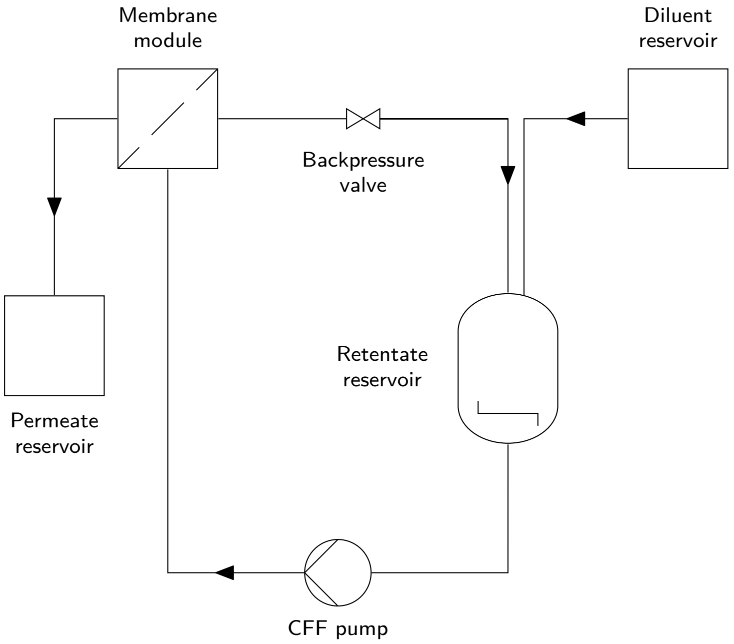

Figure 1.5: UF/DF setup: The feedstock is circulated by the pump over the membrane. The applied transmembrane pressure causes the feedstock to partly permeate through the membrane depending on the molecule size in the feedstock. The diluent is only added during diafiltration.

1.3 Process Analytical Technologies

PAT methods were historically part of process analytics or process analytical chemistry, which have been used in the chemical industry since the 1930s [62, 63]. The term PAT in its current sense has been coined by the United States Food and Drug Administration (FDA) as a system for ”designing, analysing, and controlling manufacturing through timely measurements of critical quality and performance attributes” [17] by the final guidance for implementing PAT in September 2004. PAT aims to improve product quality and, therefore, global health [17] by increasing process understanding and controlling of the manufacturing process. Furthermore, PAT methods support innovation in the manufacturing process and supplies regulatory strategies to accommodate innovation. The five pillars for the realization of these goals are process understanding, process analyzers, chemometrics, process control, and knowledge management [64].

As the first pillar is process understanding, it is important to set a Quality Target Product Profile (QTPP) describing the product criteria, like quality, safety, and efficacy, in order to find the CQAs and link these to the most Critical Process Parameters (CPPs). From these CPPs, a Design Space can be developed, in which a variation of process parameters is safe and does not affect product quality. Then, if the CQAs and CPPs can be monitored by a selected process analyzer and adequate data analysis, a dynamic control over the process can be achieved [65, 66] in order to reduce waste, production costs, and improve efficiency [64].

The appropriate selection of the process analyzer and data analysis strategy are the most problematic steps, because not every CQA can be monitored directly or in real-time. Standard process analyzers, like pH or conductivity sensors, can be implemented in the process stream, so called in-line, and the data interpretation is simple due to the univariate nature of the signal. If the required CQA is more challenging to monitor, like quantifying the main product and related impurities in the process, usually the use of on-line analytical assays or spectroscopy is required. Early approaches for PAT methods for biopharmaceuticals mainly used analytical chromatography with automated sampling, so called on-line analytical chromatography, for process monitoring and control [67]. Analytical chromatography can provide a high selectivity for the quantification of different species, but it cannot provide real-time measurements, because the sample needs to be removed from the process stream, separated by chromatography and analyzed. Depending on the process time of unit operations, a large runtime of the chromatography cannot lead to sufficient process control as attributes may change faster as chromatography can measure them.

Spectroscopy has proven to be a powerful alternative to analytical chromatography for process monitoring biopharmaceuticals [67]. Spectroscopic equipment has similar investment costs compared to on-line analytical chromatography [67], but usually the required consumable costs are lower and less laborious. In the following section, more insight into the quantification of CQAs by spectroscopy will be given.

1.4 Spectroscopic Methods

The physical principle behind the concentration quantification by spectroscopy is the linear dependence between the concentration and the observed signal intensity. For absorption spectroscopy, the Beer-Lambert law correlates the concentration of an analyte c to the absorbance A by Equation 1.1. The absorbance A is the logarithmic fraction of the intensity of the incoming light I_0 divided by the intensity of the outcoming light I. ϵ is the molar absorption coefficient, which depends on the wavelengths of the utilized radiation λ and on the measured protein. d is the path length or the thickness of the sample. [68]

$$A_{ab} = \log \frac{I_0}{I} = c \cdot d \cdot \epsilon(\lambda) \tag{1.1}$$

While this equation holds true for absorption spectroscopy, like UV and Infrared (IR) spectroscopy, for scattering based spectroscopic techniques, there also exists a linear relation between the intensity of scattered light and the concentration of the scattering molecule [69], see Equation 1.2

$$I_R = \sigma \cdot L \cdot c \cdot I_o \cdot k \tag{1.2}$$

where I_R is the observed intensity, σ is the apparent scatter-cross section in dependence on the species, environment, and excitation wavelength, L is the interrogated volume, c is the species concentration , I_o is the laser intensity and k is the instrument throughput.

Both Equation 1.1 and Equation 1.2 are used for a univariate signal, so the absorption at one wavelength or the scattering at one wavenumber to calculate the concentration of a single component sample. In a multi component mixture, those equations, with further analysis, still can be used, if the spectra of the analytes are not overlapping [68]. The ability of a spectroscopic method to resolve a mixture into unique signal ranges for every analyte is referred to as selectivity. The selectivity describes how well analytes can be differentiated by the analyzer from other components present in the sample. As the most challenging task is to differentiate between different components with a similar structure, the observability of different changes in the protein structure are of high relevance. Figure 1.6 gives an overview of the different levels of protein structure and useful spectroscopic methods for monitoring these levels.

As summary, UV and Raman spectroscopy can measure features in the primary structure, so the sequence of amino acids, through mainly the aromatic nature. The folding of the primary structure elements, so called secondary structure, can be observed with IR and Raman spectroscopy due to the changes in the bonds of the folded protein. Tertiary structure elements, so the placement of secondary structure elements to each other in a spatial arrangement, can be observed by fluorescence, UV, or Raman spectroscopy mostly due to environmentally induced changes in the spectra of aromatic acids. The total size or quaternary structure of proteins can be analyzed by light scattering.

From this overview, it becomes clear, that there is no spectroscopic techniques, that can measure all interesting features or CQAs. A combination

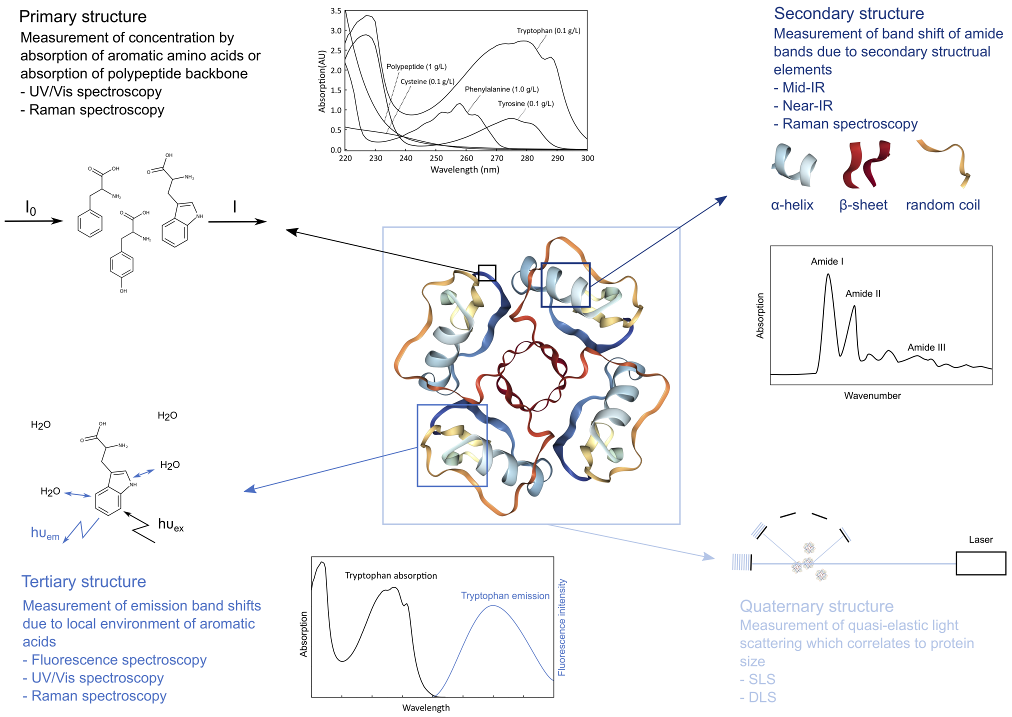

Figure 1.6: The four levels of protein structure and the respective spectroscopy for deriving information on this level are depicted. [70]

of different techniques is sometimes necessary to monitor every CQA for a certain unit operation. In the following sections, the different spectroscopic techniques used in this thesis will be explained more in detail.

1.4.1 UV/Vis spectroscopy

The main originators of the UV absorption of protein are the aromatic residues of the amino acids tryptophan, tyrosine, and phenylalanine, as well as the protein backbone [29, 34]. The three aromatic amino acids mainly absorb in the mid UV region from 220 nm to 300 nm due to the delocalized π electrons of the aromatic residues. Additionally, disulfide bridges and the polypeptide backbone absorb weakly in this region as well, as depicted in Figure 1.7 [29].

The selective quantification of protein mixtures relies on the spectral differences between these structure elements. Depending on the amount and ratio of the aromatic amino acids to each other and to the number of peptide bonds and disulfide bridges, the spectra for different proteins can vary in the position of the local absorption maxima, molar absorption coefficient, and shape of the spectrum. The difference in the spectrum can then be used for quantification by means of multivariate data analysis.

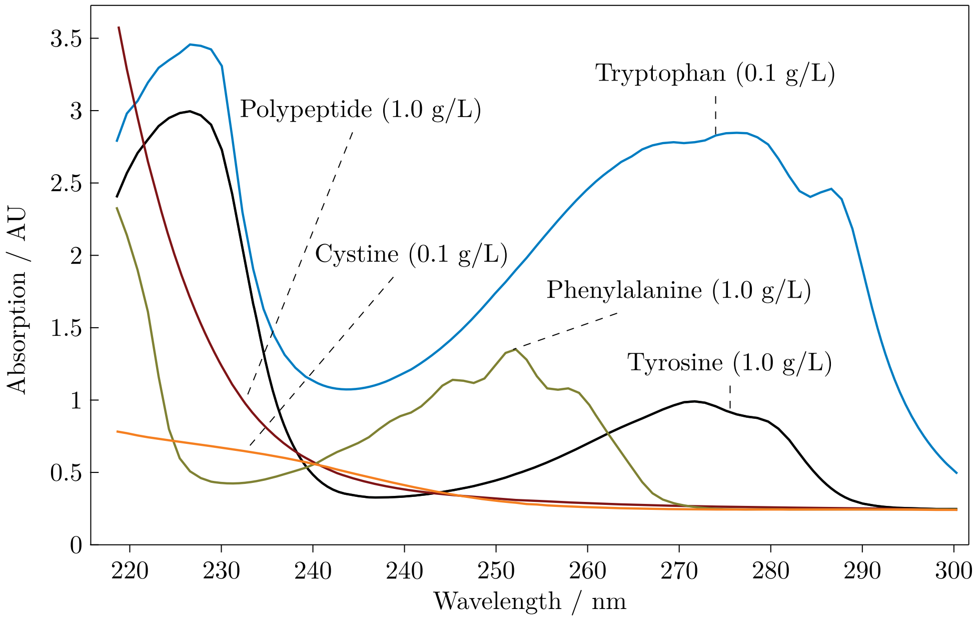

Figure 1.7: Spectra of the main contributors to the mid-UV spectrum of proteins. Adapted from [29]

UV spectroscopy combined with Multivariate Data Analysis (MVDA) has been successfully applied to process monitoring and control in Downstream Processing (DSP). For example, pooling decisions were made by UV [71], selective in-line quantification of co-eluting proteins in chromatography [72], and quantification of aggregate and fragment levels during a cation exchange in a mAb purification was achieved [73].

Also changes in the tertiary structure can be seen in the UV spectrum, if the environment around a tryptophan amino acid changes. Even though features of the secondary structure elements can be observed in the far UV region from 180 nm to 220 nm in theory, in practice the absorption of common solution components, like inorganic ions and dissolved oxygen, in this region makes the application difficult [74].

1.4.2 Fourier Transform Infrared spectroscopy

In contrast to UV, IR spectroscopy does not observe the electronic absorption of the primary structure elements of proteins, instead, IR spectroscopy is based on the vibration of the amide bonds [75]. Atoms within a molecule oscillate around an equilibrium position, which leads to a change in bond length, also referred to as stretching, or a change in the bond angle between atoms, also referred to as bending. As the frequency of these changes lay within the IR range, IR radiation can excite vibrational motions. This causes absorption of the irradiation [34]. For practical and historical reasons, instead of the wavelength, the reciprocal of the wavelength in centimeters is used for the notation of the absorption spectra.

The stretching and bending of free, planar (hypothetical) peptide groups of amide bonds give rise to nine characteristic IR absorption bands, namely in order of decreasing wavelength, amide A, B, and I-VII. The amide I (≈1650/cm C=O stretching) and amide II (1550/cm, C-N stretching, N-H bending) bands are the most prominent vibrational bands of the protein backbone. As the frequencies of the stretching and bending are influenced by the strength of any hydrogen bond involved in the amide bonds, the location of those bonds to each other influences the vibrational bands. For example, the intensity of the C=O stretching depends on the strength of the hydrogen bridge bonds to the C=O group. A strong hydrogen bridge weakens the double bond of the carbonyl group and therefore decreases the vibration energy or frequency. Hence, the secondary structure has a huge influence on the intensity of the carbonyl group vibration, because the folding is a result of hydrogen bonds between carbonyl and amide groups [76].

The nine amid bands and the influence of the secondary structure on the position and shape of the amide I band is shown in Figure 1.8. As the position and shape of the amide I and II bands have the highest intensities, these bands are used to probe changes in the secondary structure of proteins [77, 78]. The solvent has a huge influence on the conformation of proteins and hence influences the IR spectra [34].

Since the secondary structure of proteins has a significant influence on the vibrations of a protein, FT-IR has been used to analyze protein refolding and is also applied to quantify protein mixtures. It was shown that mAb, HCP levels, and aggregate levels could successfully be monitored [79].

1.4.3 Raman spectroscopy

While both FTIR spectroscopy and UV/Vis rely on the absorption of photons, Raman spectroscopy relies on the emitted photons by the induced change in energy due a change in the dipole moment or polarization by the interaction with the incident photons. This makes the Raman scattering theory more

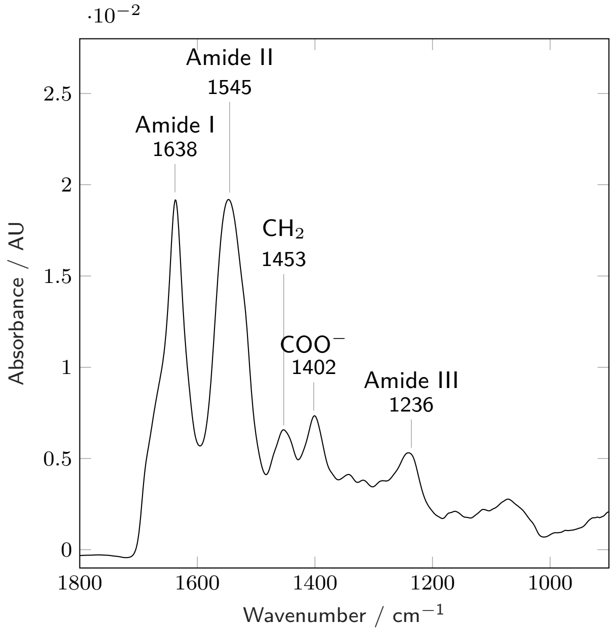

Figure 1.8: FTIR spectrum of a mAb with amide I-III, COO^- and, CH_2 peaks

complex as two photons are involved in the scattering process, compared to one photon for absorption effects [69].

In Figure 1.9, the working principle of Raman spectroscopy and differences to IR and UV/Vis spectroscopy are described with an electronic-vibrionic energy diagram. The absorption of light in the IR range induces an excitation from the ground electronic state g_0 to the first vibrational level of the ground electronic state g_1. In Raman scattering, the same state change of the molecules is achieved by a two photon process. First, a photon must be absorbed to transit the molecule to a virtual state, which has an extremely short lifetime. Subsequently, a photon with a lower energy than the absorbed photon in the first step is emitted from the molecule to relax to a vibrational excited state. This is termed Stokes transition.

The excitation from the ground state to the excited state cannot be induced by absorption from the energy of the environment. If a photon emitted by the laser hits a molecule, which is already in an excited vibrational level, the molecule can be excited to a second virtual state, from which it can release an Anti-Stokes photon to relax into the ground state. Since the populations of molecules in the excited vibrational state are always smaller compared to the populations in the ground state (Boltzmann distribution

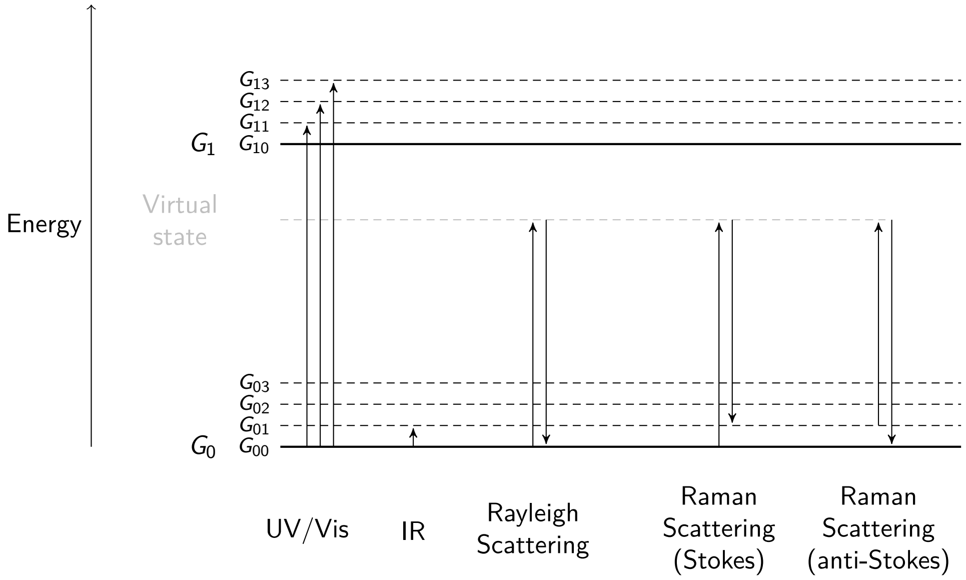

Figure 1.9: Energy level diagram comparing the different light–matter interactions. In this schematic, the length of the straight arrows are proportional to the energy of the photon involved in the process. [80]

law), the anti-Stokes Raman scattering is even weaker, so occurs less often, compared to the Stokes-scattering. If the incident photon energy approaches an excited electronic state, a resonance effect can enhance the efficiency of the Raman scattering, so called resonance Raman scattering.

Compared to IR spectroscopy, only the amid I and III bands are strong in Raman spectroscopy. As those bands are sensitive to differences in the secondary structure, Raman spectroscopy can be used similarly to IR spectroscopy for differentiating proteins with different secondary structures. [81] Additionally, the aromatic amino acids and disulfide bridges can be observed by Raman spectroscopy as well due to the large amount of π electrons, which cause a large polarizability [34]. Raman scattering describes the inelastic scattering of photons as energy is exchanged between the photon and the scatterer [69]. A far more likely event is elastic scattering, where no energy is exchanged and as a result no wavelength or directional changes of the scattered light are observed [82], which will be discussed in the next section.

1.4.4 Elastic Light Scattering

Elastic light scattering is mainly used for particle size characterization. If the particle is smaller than 5% of the incident wavelength, all regions in the particle show a similar electric field and the waves are in phase, which is referred to as Rayleigh scattering [83]. As biopharmaceutical proteins and small aggregates are often not exceeding 30 nm, the symmetrical character of the scattered photons simplifies the particle size determination [34]. For the measurement of proteins, SLS and DLS are the most often used techniques. SLS techniques measure the averaged intensity of the scattered light, while DLS techniques measure the intensity fluctuations due to the inference between moving particles [34]. SLS and DLS allow the measurement without sample preparation, which can make concentration dependent or buffer composition dependent aggregation observable [84].

Static Light Scattering

SLS is the most widely used method to determine the molecular weight of dissolved macromolecules or the detection of large aggregates [34]. The angle-dependent Rayleigh Ratio R_θ is the normalized intensity of the scattered light, which is defined as

$$R_\theta = \frac{I_\theta}{I_0} = K \cdot M_w \cdot c \tag{1.3}$$

where I_0 is the incident light intensity, I_θ is the measured intensity at a distance r from the scattering volume, c is the concentration, M_W is the overall molecular weight and K is the instrument constant [34]. In case of a polarized incident light, the dependence of the scattered light on the observation angle θ becomes 1 and K can be therefore expressed as

$$K = \frac{4\pi^2}{\lambda N_A} n_{\text{Solvent}}^2 \left(\frac{\partial n}{\partial c}\right)^2 \tag{1.4}$$

where n_o is the refractive index of the solvent, $\frac{\partial n}{\partial c}$ is the refractive index increment, N_A is the Avogadro number [34]. Equations 1.3 and 1.4 are only valid for diluted, monodisperse solutions. For undiluted, polydisperse solution the Zimm equation (1.5) represents a connection between the Rayleigh Ratio R_θ, the protein concentration c, the instrument constant K and the overall molecular weight M_W. A_2 and A_3 are the second and third virial coefficient and describe the interactions between the dissolved particles [34].

$$\frac{Kc}{R} = \frac{1}{M_W} + 2A_2c + 3A_3c^2 \tag{1.5}$$

Thereby the Zimm equation (1.5) can be used to determine the molecular weight M_W and the virial coefficients through the measurement of dilution series [85].

Dynamic Light Scattering

The intensity measured by SLS is usually a time average, because the scattered light intensity underlies fluctuations due to the Brownian motion of the particles, which depend on the size of the particles and the interaction between particles. The fluctuations themselves can be used for the calculation of the hydrodynamic radius of the particles by DLS. As small solutes move quicker than larger solutes, the intensity fluctuations are hence faster [34].

The intensity fluctuations are described mathematically by the autocorrelation function $G(\tau)$ in Equation 1.6. $G(\tau)$ can be expressed as an integral over the product of intensities of scattered light $I(t)$ at time t and delayed time $(t + \tau)$.

$$G(\tau) = \langle I(t)I(t + \tau)\rangle \tag{1.6}$$

It can be shown that when a large number of monodisperse solute particles are moving randomly without interaction, $G(\tau)$ shows a single exponential decay over time [86].

$$G(\tau) = A + B \cdot exp(-2\Gamma\tau) \tag{1.7}$$

A is often called the baseline, B is an empirical, experimental coefficient [87] and both A and B depend on the instrument. The decay rate Γ is the product of the translational diffusion coefficient D_τ and the square of the scattering vector q [34].

$$\Gamma = q^2 D_\tau \tag{1.8}$$

$$q = \frac{4\pi n_0}{\lambda_0} \sin\left(\frac{\theta}{2}\right) \tag{1.9}$$

Using the Stokes-Einstein equation for non-interacting monodisperse particles 1.10, the diffusion coefficient D can be converted into the hydrodynamic radius r_h of an equivalent spherical particle [88]. For this, the Boltzmann's constant k_B, the absolute temperature T in Kelvin and the dynamic viscosity η of the solvent are required.

$$D_\tau = \frac{k_B T}{6\pi\eta r_h} \tag{1.10}$$

For polydisperse samples, the decay rate Γ is a sum of the individual decay rates of the different species. In Equation 1.7, Equation 1.11 must be inserted [89].

$$exp(-\Gamma\tau) = \int_0^\infty G(\Gamma)\exp(-\Gamma\tau)\mathrm{d}\Gamma \tag{1.11}$$

The resulting equation has the form of a Laplace transformation and can be solved for the size distribution $G(\Gamma)$ of the particles, for example, with the method of the cumulants [90]. This yields the following solution:

$$\Gamma = \left[\Gamma\tau + \frac{\mu_2\tau^2}{2!} - \frac{\mu_3\tau^2}{3!} + \cdots\right] \tag{1.12}$$

Here Γ is the first moment of the particle distribution and in this case the intensity weighted average of the diffusion coefficients and μ_2, μ_3 are the second and third moment respectively. Thus also the calculated hydrodynamic radius, is an intensity-weighted average of the diffusion coefficients of all species, called the z-average [89].

Due to the fact that large particles scatter light more than small particles, large particles have a significantly greater influence on the calculated z-average value.

1.5 Multivariate Data Analysis

As the last section focused on the measurability of certain attributes by spectroscopy, this section will explain the information extraction by MVDA from multivariate spectra. Usually, the variables in high-resolution spectra are highly correlated to each other and carry similar information. Therefore not all variables are needed to describe the information contained in the spectra. Principal Component Analysis (PCA) is a method to reduce the dimensions of a data set by condensing variables with similar information into so called Principal Components (PCs) [91]. The PCs can be used for the correlation between the spectra and quality attributes. As PC regression bears a resemblance to PLS regression, the next section will explain PCA in detail to lay the groundwork to understand PLS as linear regression tools. The last section of this chapter will explain the background behind Artificial Neural Networks (ANNs) as non-linear regression tools. PCA, PLS, and ANNs are also referred to as machine learning tools [92].

1.5.1 Principal Component Analysis

PCA was first introduced to data analysis in economics and social science studies in the 1940s [93, 94], but became popular in chemistry as well simultaneously to the increase in computational power.

PCA extracts a set of orthogonal factors from the original data, called PCs, consisting of linear combinations of the original data. The main idea is, that high dimensional data matrices are not of full rank, meaning that some variables contain similar or identical information which can be condensed to PCs. The dimension of the matrix **X** is reduced by projecting the data into a subspace formed by the PCs. [91, 95].

The basis for PCA is the mean-centered data matrix $\boldsymbol{X}$, which comprises the M variables of N samples. As outlined by Equation 1.13, matrix $\boldsymbol{X}$ is factorized into the loadings matrix $\boldsymbol{P}$ and the scores matrix $\boldsymbol{T}$.

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^{\intercal} + \boldsymbol{E} \tag{1.13}$$

The number of columns in the scores matrix $\boldsymbol{T}$ represents the number of PCs, on which the PCA is based. The residual matrix $\boldsymbol{E}$ carries the information from $\boldsymbol{X}$, which is not described by the matrix multiplication $\boldsymbol{T}\boldsymbol{P}^{\intercal}$. For a good PCA model, $\boldsymbol{E}$ carries mostly measure errors and noise. [95] The loadings are the new dimensions of space and the scores are the new coordinates in this space. The transformation and visualization of high-dimensional data as low-dimensional planes is an important application of PCA. [91]

In Figure 1.10, a geometrical visualization of PCA is depicted. Basically, the PCA is a transformation of the main axes. Therefore, each observation of the mean-centered $\boldsymbol{X}$-matrix is positioned in the M-dimensional variable space. Mean-centered means, that the mean value of the variables is subtracted from the data. The mean value of the variables can be displayed as a vector and hence as a point in the M-dimensional space, namely the origin (red). The mean-centering corresponds to a re-positing in the coordinate system. PCA now finds lines, planes and hyperplanes in the M-dimensional space approximating the data in the least square sense. The first PCA line goes through the average point. Hereby, the first PC, PC1, reflects the direction of the highest variance. Each observation is then projected into the line in order to get the co-ordinate value, known as score. The variance of the scores in PC1 direction is maximized and the residual variance is minimized. The second PC is represented in the space as orthogonal line to PC1 through the average point [96]. PC1 and PC2 vectors define the plane the projected configuration is known as score plot. The further away from the origin the variable lies, the stronger is the impact on the model. [96]

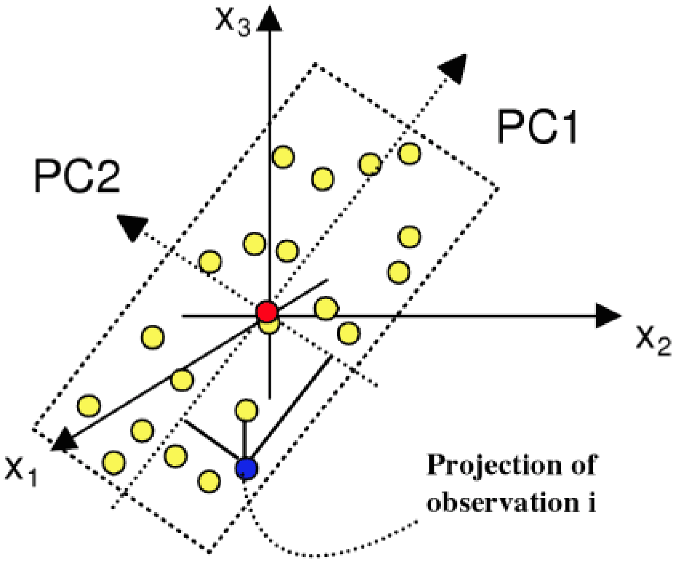

Figure 1.10: A geometrical visualization of PCA in the multidimensional space. PC1 reflects the direction with the highest variation. PC2 reflects the direction with the second highest variation and is orthogonal to PC1. PC1 and PC2 form a window into the multidimensional plan. [96]

PCA can be seen as a dimensionality reduction of the data by separating the information in the data from the measurement noise $\boldsymbol{E}$, which can increase the interpretability of the data. If quantitative information is needed, the scores of the PCA can be used for regression to a $\boldsymbol{Y}$-variables (Principal Component Regression (PCR)). Also desired quantitative information in the $\boldsymbol{Y}$-variables can be used to influence the scores. Then only the variance important to predict $\boldsymbol{Y}$ is captured in the scores, which is referred to as PLS.

1.5.2 Partial Least Square (PLS) Regression

PLS regression was first introduced to data analysis in social science studies, but became popular in chemometrics [62]. The goal of PLS regression is to predict $\boldsymbol{Y}$ from a data matrix $\boldsymbol{X}$ and describe the relationship between both. In most chemometric cases, $\mathbf{Y}$ are the concentrations of an analyte and $\boldsymbol{X}$ are the recorded spectra. When the $\boldsymbol{X}$-variables are numerous, noisy and correlated, as usual for spectra, ordinary multiple linear regression (MLR) according to Equation 1.14 is no longer feasible. [91, 97]

$$y = b_o + b_1x_1 + b_2x_2 + ... + b_nx_n \tag{1.14}$$

MLR needs more samples than $\boldsymbol{X}$-variables for determining the unique regression coefficients, which is rarely the case for spectroscopic data sets. If variables x_i and x_j are correlated (linear dependent), MLR leads to unstable prediction of the regression coefficients due to the corruption by measurement noise [98]. The solution to the multicollinearity problem is to do a matrix factorization and, hence, reduce the $\boldsymbol{X}$-variables to a smaller space, where multilinear regression is again possible, also referred to as PCR. The main difference between a PCR and PLS is, that PCR recovers the information of the most dominant variables for describing the $\boldsymbol{X}$ matrix. In contrast, PLS recovers the information of the $\boldsymbol{X}$ matrix, which are most important for the description of the $\boldsymbol{Y}$ matrix. [95]

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^{\mathsf{T}} + \boldsymbol{E} \tag{1.15}$$

$$\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{C}^{\mathsf{T}} + \boldsymbol{F} \tag{1.16}$$

For the PLS regression, the data matrix $\boldsymbol{X}$ and $\boldsymbol{Y}$ are both factorized (Equation 1.15 and 1.16). Similar to PCA, $\boldsymbol{X}$ is factorized into the loadings matrix $\boldsymbol{P}$ and the scores matrix $\boldsymbol{T}$. The residual matrix $\boldsymbol{E}$ contains errors and noise (Equation 1.15). $\boldsymbol{Y}$ is factorized into the loadings matrix $\boldsymbol{U}$ and the scores matrix $\boldsymbol{C}$. The residual matrix is called $\boldsymbol{F}$ (Equation 1.16). Both spaces from $\boldsymbol{X}$ and $\boldsymbol{Y}$ are linked, because the highest $\boldsymbol{U}$ score vector is the basis for $\boldsymbol{T}$ score calculation. This means, that the first estimated t_1 vector is actually as well the u_1 vector, hence $\boldsymbol{Y}$ data space influences the $\boldsymbol{X}$ via $\boldsymbol{U}$ scores matrix and vice versa. In the context of PLS, the ”PCs” in the scores matrices are called LVs or PLS components, because they are not the same components as for a PCA.

The actual calculation of the different matrices of PLS can be done by an Nonlinear Iterative Partial Least Squares (NIPALS)-algorithm. NIPALS is a simple algorithm developed by Herman Wold to estimate the parameters of a PLS regression [99]. The regression matrix $\boldsymbol{B}$ can finally be calculated with Equation 1.17. $\boldsymbol{B}$ is then used to calculate y from an x for an independent data set.

$$\boldsymbol{B} = \frac{\boldsymbol{U}^{\mathsf{T}}\boldsymbol{T}}{(\boldsymbol{T}^{\mathsf{T}}\boldsymbol{T})} \tag{1.17}$$

PLS models are linear regression models. Therefore, non-linearities in the data can only be fitted as a linear approximation. Even though, the correlation between an analyte concentration y and spectral measurement intensity is expected to be linear, there are effects, like non-linearity of the detector due to the very high or low number of measured photons, concentration-dependent effects in the spectra or effects of the background matrix, which lead to a deviation from the expected linear correlation. Therefore, those effects need to be removed from the spectra by preprocessing methods, if possible, for calibration of an accurate model. More information on preprocessing, calibration, and validation of PLS models will be given in Chapter 3.

Sometimes, the removal of the non-linear effects from the spectra is difficult or the removal introduces more noise into the spectra (e.g. by applying derivates). An alternative can be the use of non-linear regression techniques, like ANN.

1.5.3 Artificial Neural Networks

ANNs originated as an attempt for a mathematical representation of information processing in nervous systems according to the laws of theoretical neurophysiology [92, 100]. Simplified, nervous systems consist of a net of so called neurons or nodes, which are interconnected. Figure 1.11 depicts the architecture of a simple fully connected ANN. ANNs consist of an input layer, a variable number of hidden layers, and an output layer. The nodes in the input layer receive the data and forward it to the nodes in the hidden layer. The number of nodes in the input layer is defined by the amount of inputs $x_1, ..., x_n$. The nodes in the hidden layers and output layer are responsible for the calculation of the weighted sum of received signals and the biases, as well as its processing with transfer functions. Eventually, the output signal is sent to all receiving nodes in the next layer. [92]

The mathematical function of a neuron can be explained as well with an analogy to the nervous system. On the occurrence of an electrical impulse, the neuron has a threshold, which the excitation must exceed to allow the signal to pass through the neuron and to be processed [100]. For ANNs, the input values $x_1, ..., x_n$ are given to a neuron in the next layer of the ANN as the activation a_j, which resembles the electrical impulse in a physical neuron. The activation a_j is calculated as a linear combination of the input values, weighted with parameters w_{ji} and a bias parameter w_{j0} according to Equation 1.18. The superscript indicates to which layer in the model the parameters belong.

$$a_j = \sum_{i=0}^{n} {w_{ji}}^{(1)} x_i + {w_{j0}}^{(1)} \tag{1.18}$$

The activation a_j is then transformed by the neuron with an activation function $f(\cdot)$ to the output of the neuron h_j according to Equation 1.19.

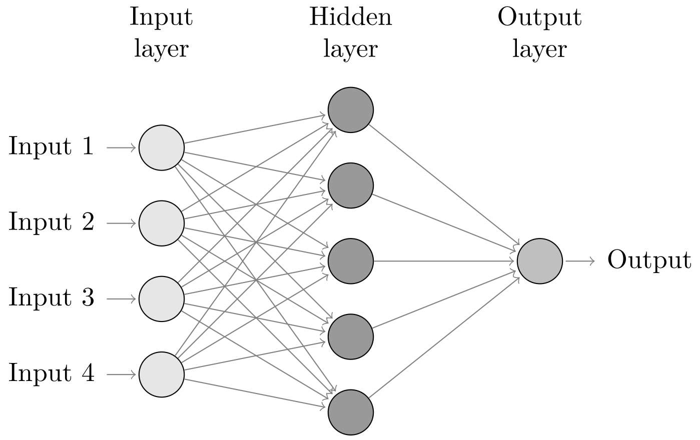

Figure 1.11: Schematic illustration of a multilayer perceptron with x neurons forming the input layer, one hidden layer of y neurons, and an output layer consisting of z neurons.

$$h_j = f(a_j) \tag{1.19}$$

The choice of activation function is determined by the nature of the data and the assumed distribution of target variables. For binary classification problems, generally logistic sigmoid functions (Equation 1.20 are used as output unit activation functions [92] due to the binary limits of the function in a positive and negative direction.

$$\sigma(a) = \frac{1}{1 + e^{-a}} \text{ in } [0, 1] \tag{1.20}$$

For regression problems, where the output of the ANN is proportional to the input vector, rectified linear unit (ReLu) (Equation 1.21) is often used. Similar to the function of a physical neuron, the ReLu activation function is the positive part of its argument. Therefore, the return value is zero below a certain activation value, in this case zero), and above an activation of zero, the ReLu function returns the activation.

$$R(a) = \max(o, a) \text{ in } [0, 1] \tag{1.21}$$

The different layers of the ANN can be combined to an overall network function that, for an ANN with three layers (input, hidden, and output layer) takes the form

$$y(x, w) = f\left(\sum_{j=0}^{m} {w_{kj}}^{(2)} f\left(\sum_{i=0}^{n} {w_{ji}}^{(1)} x_i + {w_{j0}}^{(1)}\right) + {w_{j0}}^{(2)}\right). \tag{1.22}$$

Thus, an ANN is a just a parametric function, which calculates an output y from an input vector of variables $\boldsymbol{x}$ and a matrix of weight parameters $\boldsymbol{w}$ [92]. Depending on the number of layers and activation function, the ANN can fit any function [101]. The higher flexibility of ANN comes at the cost of more parameters in comparison to a PLS model, which makes a large number of samples necessary for adequate training of the ANN.

2

Thesis Outline

2.1 Research Proposal

Digital transformation has been identified as a key pillar for improved value generation in different industries in the past. As the biopharmaceutical industry struggles to provide cheap and widely available drugs to the world population, also the conservative pharmaceutical sector is catching up on the possibilities for real-time monitoring and control of production processes for cost reduction. The implementation of PAT methods for monitoring of CQAs and CPPs builds the foundation for an effective process control in order to move to continuous and robust processes by reduction of process variability.

Biopharmaceutical production processes have high requirements for PAT methods as proteins are larger and more complex compared to synthesized drugs. Additionally, biopharmaceuticals are closely related to some contaminants, like HCPs or product-variants, due to the proteinous nature of both. The close chemical relation impedes the selective monitoring of each species by PAT tools, which often makes the use of advanced multivariate measurements necessary. The ability of multivariate spectroscopy in combination with PLS modeling has been proven to selectively quantify protein concentration in complex mixtures of model proteins. However, the applicability to real downstream processing units needs to be proven.

The economically most interesting use cases for PAT methods are the capture step at the beginning of the downstream process due to the high costs of Protein A resin and the UF/DF step as last step of the downstream process due to the high costs of the purified start material. From a scientific point of view, the capture step and the UF/DF step are very contrary as the first is a chromatographic step to purify and concentrate a product from a crude mixture and the latter is a filtration step to prepare the pure product for formulation. Along with the different purposes of both steps sides different monitoring needs and sensor requirements.

While the strengths and weaknesses of different sensors vary, there is no universal sensor to measure all CQAs or CPPs, which require monitoring during the whole process. The implementation of sensor arrays combining several methods may also not be desirable due to the high investment costs and increased probability of equipment failure. Instead, a conscious selection of different sensors or sensor combinations could be a viable solution for specific challenges.

The correlation of the measured feature by one multivariate sensor to the quality attribute of interest is often done by PLS modeling. Due to the increased availability of computational power, also more advanced and flexible models, like ANNs, have been proven to solve complex classification and regression problems. As the relation between measured feature and calculated quality attributes does not always follow a linear relation, preprocessing of the data is often required to make the relationship more linear. Here, the application of more advanced machine learning techniques offer great benefits. However, the applicability to biopharmaceutical processes has to be proven as more complex models generally require more data than there is available during development to train to increased number of parameters in comparison to simpler, linear models.

In case a combination of different sensors is required to measure the quality attributes of interest, data fusion methods have to be developed and applied. In case of a physical relation between the sensor signals and the quality attribute of interest, adequate filter and error compensation techniques need to be implemented for use before calculation. In case the same information is measured by different sensors, multi-block data fusion techniques need to be evaluated.

The objective of this research project is to meet unsatisfied needs in the real-time monitoring of biologics by leveraging commonly implemented process sensors and by applying smart sensor concepts and data analysis strategies. In order to fulfill this objective, the measurement capabilities of different sensors for proteins in aqueous solutions need to be understood. Prerequisites for the applicability to real-time monitoring are the operation in a flow-through setup, either in the process stream itself (in-line) or in a bypass (on-line), and the general observability of protein-related features. Especially optical spectroscopy checks these requirements. Therefore, a review of optical spectroscopic sensors will stand at the beginning of this thesis.

As multivariate UV spectroscopy is the most used tool to selectively quantify proteins in aqueous solutions, the second study will focus on the application of UV spectroscopy combined with PLS to the Protein A capture step. Here, the method will be used to quantify the mAb concentration in the effluent of a Protein A capture and terminate the loading when a defined concentration is reached. The aim of this second study is to prove the applicability of multivariate UV spectroscopy to a real process for a specific product. First, a small design space will be investigated, covering only the mainly observed process variation, which is the mAb concentration in the load material. However, the study will not reflect real process variation, as the concentration of impurities can change during the fermentation as well. As the calibration of models in general demands multiple runs and a new model for every product and process due to the changes in the spectra, the application of the method will be labor intensive.

To overcome these implementation hurdles, a consecutive study with a broad design space covering multiple products will be evaluated, if the initial study proves to be successful. As the variation in the spectra grows due to increased variability in the load material, a method to remove the background variation from the spectra could be necessary to improve the prediction accuracy. The background subtraction could be realized by subtraction of a UV spectrum, when the initial break through of impurities through the column is finished. The determination of this time point might require a second sensor, which is not sensitive to changes in the protein concentration, but rather the correlating effects like the buffer composition. If the background variation are removed, the developed model could be applicable to any process and to different products with a similar UV spectrum. This would decrease the development efforts significantly. An additional benefit of the background subtraction could be the use of univariate UV sensors for mAb concentration, which are usually implemented. The implementation of the background subtraction for the UV spectra could improve the limited selectivity of UV spectroscopy.

Even though the quantification of mAb in real process steps by UV spectroscopy seems promising, the comparison of UV spectroscopy to more selective spectroscopic techniques is useful. Raman spectroscopy has been applied successfully in upstream processes, among other things, for the quantification of mAb in the cell culture fluid. Therefore, a comparison of both spectroscopy techniques will be the core of a next study. Additionally, as Raman spectroscopy and UV spectroscopy have different advantages, like the higher selectivity of Raman or the higher measurement speed of UV spectroscopy, multi-block data fusion methods will be tested to evaluate the combinability of the advantages. As Raman spectra are more complex than UV spectra, advanced preprocessing is generally done to improve the predictions of models. As machine learning has shown impressive results for the fit of complex problems, like speech recognition, CNNs will be applied as regression models and compared to PLS models. CNNs combine filter layers (convolutional layers) and densely connected layers of a common ANN to preprocess features and fit non-linear models. As preprocessing generally tries to remove non-linearities from the spectra, CNNs automate the preprocessing of spectra and improve the predictions due to the non-linear fit. Additionally, CNNs could be used as data fusion methods, when different spectra are chosen as input.

A different approach to multi-block data fusion is the sensor combination based on physical principles. In the last study, UV spectroscopy, light scattering and density/viscosity measurements will be combined to monitor protein concentration, buffer exchange, apparent molecular weight and hydrodynamic radius during a UF/DF step. The protein concentration can be determined by VP UV spectroscopy. Due to the large concentration range, a variable path length will be necessary to avoid a detector saturation at high concentrations, but allow for an accurate quantification at low concentrations. The buffer exchange could be monitored by density measurements as different buffers generally have different densities and the change in density corresponds to a change in buffer composition. For the monitoring of the apparent molecular weight, the protein concentration determined by VP UV spectroscopy will be combined with the static light scattering measurement, as this measurement is influenced by the concentration of scattering particles and the size of the particles. The by dynamic light scattering observed hydrodynamic radius is influenced by the surrounding viscosity of the medium. Therefore, a correction for the changing viscosity due to the change in buffer or protein concentration will be done by measurements of the density/viscosity sensor. With the smart combination of UV spectroscopy with light scattering and density/viscosity measurements, typical monitoring needs of the UF/DF process could be fulfilled to enable process automation. In a next step, the UF/DF setup will be updated for process automation and the necessary sensors to control the protein concentration and buffer exchange progress will be implemented. Additionally, an EKF will be implemented to support the prediction of the buffer exchange progress by mechanistic process knowledge.

3

A Critical Review of Recent Trends, and a Future Perspective of Optical Spectroscopy as PAT in Biopharmaceutical Downstream Processing

Laura Rolinger[1], Matthias Rüdt[1], Jürgen Hubbuch[1]

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

Abstract

As competition in the biopharmaceutical market gets keener due to the market entry of biosimilars, PATs play an important role for process automation and cost reduction. This article will give a general overview and address the recent innovations and applications of spectroscopic methods as PAT tools in the downstream processing of biologics. As data analysis strategies are a crucial part of PAT, the review discusses frequently used data analysis techniques and addresses data fusion methodologies as the combination of several sensors is moving forward in the field. The last chapter will give an outlook on the application of spectroscopic methods in combination

with chemometrics and Model Predictive Control (MPC) for downstream processes.

3.1 Introduction

The biopharmaceutical industry currently faces major changes because of increasing competition in the field due to the market entry of biosimilars and increasing costs in research and development (R&D) of new drugs [102]. Since 1950, the number of approved drugs per billion US dollars spent for R&D has halved approximately every 9 years. This behavior is termed 'Eroom's Law' as it describes the opposite of 'Moore's Law' [103]. Not only are the costs per approved drug increasing, but the sales of off-patent blockbuster drugs are slowing down due to price competition from a variety of biosimilar products [104]. More companies seek to capitalize on the rapidly growing biologics market, which creates a competitive climate driving innovations for cheaper production, faster development, and improved quality of the biologics in order to gain a competitive edge [104, 105].
Digital transformation has already proven to drive the performance of companies in other industry sectors and has started to be adapted by the rather conservative biopharmaceutical industry as key strategy for production improvements as well [106, 107]. Part of the digital transformation of production processes are the implementation of appropriate measurement sensors and data analytics, i.e. PAT, as information input for process control algorithms [107]. The achieved process control allows for optimal production runs and improves process robustness. The product quality may be improved by coping with process variability. Process robustness also shortens the development-to-market times, e.g. by facilitating scale-up, resulting in a competitive advantage [108].

While PAT has been successfully implemented as a pillar of process control for numerous small-molecule pharmaceuticals [109, 110], the high complexity of biopharmaceutical proteins and the close chemical similarity of contaminants impose a challenge for finding suitable PAT methods [111]. Ideally, a PAT method would be able to differentiate between product, process-related contaminants, and product-related contaminants in real-time. However, some product-related contaminants (such as subtle structural differences in oxidation or deamidation of single amino acids to the product) are detected by time-consuming analytical methods [112] e.g. analytical HPLC methods which typically take 30 minutes or more [25]. Larger structural differences (e.g. aggregation, misfolds, or pegylated species) can be detected by on-line HPLC within 4 min to 6 min [67, 113], or by in-line spectroscopic methods in real-time [114, 115]. Here, spectroscopic methods offer several advantages over on-line PAT methods, such as rapid and automated detection with no sample preparation, conditioning, or destruction at comparable equipment costs [62]. However, one optical spectroscopic method alone offers a limited selectivity for the structural integrity of proteins [67], but optical spectroscopic methods can be easily combined with other spectroscopic or non-spectroscopic sensors to measure a large variety of attributes [116, 117]. Therefore, improved measurability and accuracy can be achieved by multiple sensors as compared to a single sensor [118, 119].

As the data complexity increases through the combination of multiple, possibly multivariate, spectroscopic sensors, advanced data analysis is required to extract information from the multivariate data about critical process parameters or critical quality attributes [17]. Data analysis from chemical data itself is also referred to as chemometrics [120]. Even though chemometrics generally covers the basic analysis from multiple data sources, data fusion methodologies are applied to chemical data for classification and prediction improvement [121]. As data analysis is often performed by software, the combination of sensors and data analysis for attribute estimation is often referred to as soft sensor [122].

Following this line of arguments, the section below will give a general overview and address the recent innovations and applications of optical spectroscopic methods as PAT tools in the downstream processing of biologics. This is meant as an addition to the comprehensive review by Rüdt et al. [67] in 2017. This review will focus only on optical spectroscopy, because other tools have been review in full elsewhere [123, 124]. As data analysis strategies are a crucial part of PAT especially for the interpretation of spectroscopic data, the third section will give a review about frequently used data analysis techniques and address data fusion methodologies as the combination of several sensors is moving forward in the field. The last section will give an outlook on the application of soft sensors (spectroscopic methods in combination with chemometrics) and model predictive control for downstream processes.

3.2 Improvements in Spectroscopy and Applications

3.2.1 Spectroscopic Methods and their Applicability to Protein Monitoring

The selection of appropriate techniques consisting of a spectroscopic method as well as a measurement setup is a key element in PAT [125]. The most important selection criteria are sensitivity and selectivity to evaluate the feasibility of the application. Other factors, like costs or complexity of the instrument, have to be evaluated for a successful process implementation in industry [62, 125]. In downstream processing of biologics, the dynamic range and measurement speed are important factors for the technology selection as well, because the concentration ranges are generally the largest in production and the feasible measurement times are the shortest.

The measurement environment (bulk solvent, temperature, pressure, etc.) greatly influences the sensitivity and selectivity of different methods. As the solvent often contributes the majority of molecules to the sample, it needs special consideration [125]. For biopharmaceutical processes, the solvent is in most cases water. Thus, high water signals are a typical problem in protein measurements. In Figure 3.1, the bulk water absorption coefficients are depicted with reference wavelength regions for various spectroscopy types. UV spectroscopy, intrinsic fluorescence, and often also Raman spectroscopy take place in regions of the electromagnetic spectrum with low water absorptivities. Even though Near-Infrared (NIR) and Major Immunodominant Region (MIR) measurements are generally thought of as selective and relatively sensitive, when it comes to measuring in water, these methods are impaired by the high water absorptivity caused by the OH band. In the NIR and MIR region, the water absorption spectrum dominates over the protein absorption (cf. Table 3.1). Additionally, the temperature sensitivity of the OH bands is a severe drawback for measuring aqueous solution in NIR and MIR, which makes tempered sample holders necessary [126].

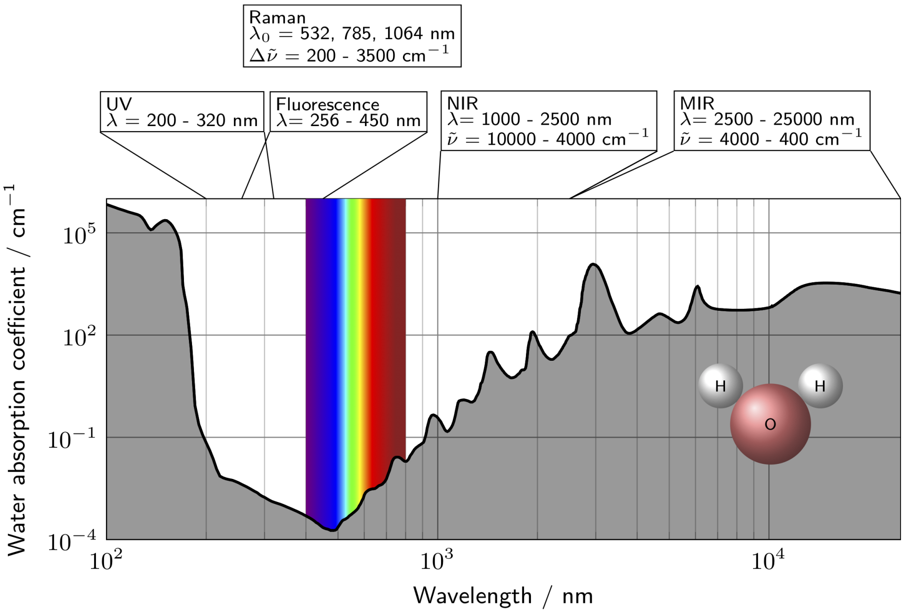

Figure 3.1: Typical wavelength ranges of UV, fluorescence, NIR, MIR, and Raman spectroscopy for the analysis of proteins are depicted. Additionally, the bulk water absorption coefficient is plotted over the wavelength to emphasize the effect of water on the different techniques. The visible spectrum is indicated for orientation. The data for the bulk absorption coefficient was taken from Segelstein [127].

Table 3.1: Molecular cross-sections and extinction coefficients (if applicable) of IgG measured with different spectroscopic techniques.

Spectroscopic method	cross-section σ as $-\log\left(\sigma/(\mathrm{cm}^2/\mathrm{molecule})\right)$	absorption/emission coefficient /(L/(g cm))	extinction coefficient water /1/cm
UV (280 nm)	16	1.2 to 1.5 [128]	$2.6 \cdot 10^{-3}$
Fluorescence	17	0.16 to 0.2 [129]	$1.3 \cdot 10^{-3}$
NIR	16	1.2 [130]	25.6
MIR	15	12	1400
Raman (532 nm)	27	-	$4.2 \cdot 10^{-4}$
Resonance Raman (229 nm)	25 [131]	-	$6 \cdot 10^{-3}$
Rayleigh (633 nm)	18-19	-	$3 \cdot 10^{-3}$

To compare different spectroscopic methods based on their sensitivity to proteins in water, the molecular cross sections, extinction coefficients, and the water absorption coefficients are listed in Table 3.1 for the different methods. The listed protein values are representative of an Immunoglobulin G (IgG). Further information on the calculations are given in Supplementary Material 3.5. [id=2nd]Table 3.1 gives an overview on the sensitivity of the different spectroscopic methods by comparing the different scatter cross-sections. However, it is important to consider the surrouding solvent water. It is benefical to achieve a high ratio of protein scatter cross-section to water absorption. Table 3.2 gives an overview which protein structural elements are measurable with different spectroscopic methods. [id=2nd]Table 3.2 helps to evaluate, whether the protein structure feature of interest can measured with the selected spectroscopic method. Table 3.1 provides a lead on the measurability of a certain protein concentration in water with the selected spectroscopic method. Generally, it is important to look at the protein and water absorption in the wavelength range of a spectroscopic method to draw the right conclusions.

Table 3.2: Structural elements of proteins observed with different spectroscopic methods. The information was compiled from [34] and [132].

Spectroscopic method	Relevant structural elements
UV	Aromatic amino acids, peptide bonds, disulfide bridges size (light scattering)
Fluorescence	Aromatic amino acids
NIR and MIR	Peptide bonds
Raman	Aromatic amino acids, peptide bonds, disulfide bridges
Resonance Raman	Excitation $\leq 220\,\mathrm{nm}$: peptide bonds Excitation $\geq 229\,\mathrm{nm}$: aromatic amino acids
Rayleigh	Protein weight and shape

In the NIR and MIR regions, proteins show high absorption coefficients compared to the other methods due to the strong absorption of the C=O bond [133]. However, since water absorption in this region can be a 100-fold higher for dilute concentration, NIR and MIR are not well suited for quantifications down to 1 g/L [134], which means that the quantification of contaminants in the process will be challenging due to the low concentrations. In contrast, UV and intrinsic fluorescence spectroscopy show little water interference, but absorption and emission coefficients comparable to those in the NIR and MIR regions. Therefore, quantification of proteins in the mg/L range is possible with UV and fluorescence spectroscopy [135]. Rarely, there are deviations from the Beer-Lambert law due to, e.g. adsorption to the measurement cell walls, which can impair the quantification limits [136]. Even though intrinsic fluorescence spectroscopy can quantify proteins to the mg/L range, it behaves only linearly at low concentrations (absorbance below 0.05) due to the so-called inner filtering effect. The inner filtering effect is caused by light absorption in the sample and results in distorted emission intensities and spectra, which cause a nonlinearity between fluorescence intensity and protein concentration [34, 137]. Consequently, UV spectroscopy typically offers a greater linear range than fluorescence spectroscopy [138].

Like UV and intrinsic fluorescence spectroscopy, Raman spectroscopy usually has very low water interference as well [126] but, due to very small protein scattering cross-sections (cf. Table 3.1), the water bands are dominant for dilute protein solutions. Therefore, protein structure studies often utilize the resonance enhancement effect in the UV range [139] to increase the intensity of protein bands and take advantage of the low water absorptivity in the UV. The resonant effect of the Raman scattering in the UV region, referred to as UV Resonance Raman (UVRR), is caused by the absorption of aromatic amino acids or the polypeptide backbone of proteins. The Raman cross-section of the modes coupled to these resonant electronic transitions can increase by a magnitude of five [131]. Besides the enhancement advantages of UVRR, there are some drawbacks like photodamage due to exposure to UV light or a loss of linearity between the signal intensity and the concentration of protein due to the reabsorption of photons [134]. This effect is comparable to the inner filter effect observed in fluorescence measurements [134].

Not only does the broad concentration range during purification of biologics impose a challenge on the linear range and sensitivity of analytical methods but the complexity and chemical similarity of contaminants to the respective product call for a high level of selectivity for quantification as well [25, 46]. The International Union of Pure and Applied Chemistry (IUPAC) defines selectivity as “the quantitative characterization of a systematic error in the measure of a signal caused by the presence of concomitants in a sample” [140]. In other words, it is the accuracy of quantifying an analyte in a mixture [141]. For spectroscopy, this implies that the signal/bands of interferent and analyte need to be distinguishable for a high selectivity [142]. UV spectroscopy observes the electronic state transitions. The most prevalent chromophores in proteins are the peptide backbone, the aromatic amino acids (tryptophan, tyrosine, and phenylalanine), and disulfide bridges formed by oxidation of two cysteine residues to cystine [34, 143]. Furthermore, UV spectra contain information on protein folding (via wavelength shifts of the involved chromophores) to aggregation (via light scattering), even though these different energy states overlap to the broad electronic absorption spectra usually observed in solution [34]. This information can be used in combination with multivariate data analysis tools, like PLS models, to deconvolute several species, which has been shown in several case studies [72, 73, 114, 143, 144].

In MIR, up to nine characteristic bands can be observed for proteins, namely and in order of decreasing wavenumber amide A, B, and I to VII [145]. The amide I band (1610/cm to 1700/cm, mostly C=O stretching) and amide II (1480/cm to 1575/cm N-H bending and C-N stretching) are most pronounced. These bonds are influenced by the hydrogen bonds around them, formed by the folding of secondary structure elements [146]. Aromatic amino acids absorb as well, but mainly in the spectral region of the amide I band from 1610/cm to 1700/cm [147]. Due to the overlapping absorptions, highly convoluted and similar spectra are observed for proteins. However, MIR spectroscopy can be used to distinguish between proteins and other substances used by the biopharmaceutical industry, like Polyethylene Glycole (PEG) or Triton-X [115, 148]. These measurements were carried out with FTIR, which is not entirely suitable for processes due to moving parts and vibrational sensitivity [149].

NIR spectroscopy has the advantage of having no moving parts. However, the selectivity is generally low, due to the superposition of different overtones and combination bands in the NIR region [150].

As a complementary vibrational spectroscopic method to MIR, Raman spectroscopy provides similar information on the secondary structure of proteins. Similar to MIR, the amide bands (especially amid I and III) are strong in Raman spectra [34]. Additionally, Raman offers more structural details on aromatic amino acids and disulfide bonds that reflect the protein tertiary structure. These information can be observed because some molecular groups in the protein side chains, such as C=C, C-C, S-S, C-S, S-H groups, have large polarizabilities which results in large Raman activities [62]. In contrast to MIR, these bands generally overlap less with the amide bands [151] and, therefore, the selectivity of Raman for proteins is generally higher. Furthermore, as discussed above, the impact of the bulk water is smaller for Raman spectroscopy.

The selectivity can be improved by chemometric methods, also referred to as computational selectivity [141], which will be further addressed in Section 3.3. The initial selectivity of a sensor is, however, an important driver of the computational selectivity [152]. This might be the reason why UV spectroscopy in combination with chemometric methods has successfully been applied to a wide variety of problems in the last decade [67] as a result of its strong sensitivity and decent selectivity. Raman spectroscopy is frequently applied in upstream processing in research and industry due to its high selectivity and low water interference [153] despite the relatively long measurement times. Instrumental innovations shorten measurement times and make Raman spectroscopy more amendable for downstream processing as well. New applications of UV, fluorescence, Raman, and multimodal spectroscopy as PAT tools for downstream processing will be addressed in the following subsections in detail.

3.2.2 UV Spectroscopy

A challenge of UV spectroscopy is the limited linear range of the instruments [67]. The application of VP UV spectroscopy allows for concentration measurements in an extended dynamic range. The necessary equipment has been commercialized and is available under the brand names FlowVPE and SoloVPE [154, 155]. Recent applications of VP UV spectroscopy showed the applicability to a mAb chromatography step from 0 g/L to 80 g/L [114] and to an UF/DF process with a range from 2.8 g/L to 120 g/L [156]. For most flow rates, the FlowVPE can be used in-line. Due to the used monochromator, the FlowVPE takes a significant amount of time (typically $\geq$30 s) [114] to collect a full spectrum. Replacing the monochromator with a polychromator and a diode array detector could improve measurement time in the future and reduce the number of moving parts in the VP spectroscopy system.

Alternatively, the use of Attenuated Total Reflection (ATR) flow cells could be of interest for measuring UV spectra in high concentration protein solutions. However, to the best of our knowledge, no studies with a focus on biologics have been published using UV ATR flow cells.

3.2.3 Fluorescence Spectroscopy

Pathak et al. demonstrated that the fouling of Protein A resin can be observed by diffuse transmission fluorescence spectroscopy [157]. While it is interesting that the fluorescence increases due to protein fouling on the resin, a direct correlation is difficult. Due to the setup path length of 1 cm, the study is not directly applicable for industrial scale. Higher path lengths might result in a more pronounced inner filter effect and nonlinearities. Additionally, Zhang et al. [158] showed, that the resin fouling is not homogeneous over the column, which makes multiple measurements necessary to provide a holistic picture over the column.

3.2.4 Raman Spectroscopy

In general, Raman scattering is a weak effect because only about 1 in 10^{10} photons undergoes Raman scattering in aqueous protein solutions [159]. To set this into perspective with absorption experiments where a mAb (ϵ= 14 L/(g cm)) will absorb around 90% of the incident photons over 1 cm cuvette at a concentration of 0.7 g/L [159]. The low scattering cross-section explains why the first Raman scattering measurements took days [160]. Due to the development of compact and high power lasers, charge-coupled devices, fiber-optics probes, and further optical component enhancements, measurements can be realized in minutes today because of the increased photon output and collection efficiency [69, 161]. With standard Raman analyzers, measurement times of 12.5 min (785 nm excitation, 75 s collection time with 10 exposures) [162, 163] are frequently applied to upstream processes. As upstream processes can take a couple of weeks [164], a measurement time of 12.5 min is sufficient. But for downstream process units with operation times of a few hours [164], measurements need to be significantly faster. Usually, 30 s are considered near real-time in downstream processing [114].

There are several factors influencing the strength of the Raman signal and hence the measurement speed, but all of them rely either on increasing the amount of scattered photons or converting more scattered photons to a signal. The Raman efficiency increases by a fourth-order function as the laser frequency is decreased. Hence, the shorter the laser wavelength, the more intense is the Raman signal [69]. Unfortunately, a shorter wavelength does not always result in a better Raman spectrum because fluorescence can overshadow the Raman signal. At the very least, a laser excitation wavelength and according Raman scattering range outside the intrinsic fluorescence range of proteins from 257 nm to 450 nm [129] should be chosen for the downstream process to avoid fluorescence overpowering the Raman signal. This is assuming, that other potential fluorophores, like phenol red from the cell culture medium [165], which fluorescence outside the intrinsic protein fluorescence range, are not present. At a laser excitation wavelength below the intrinsic protein fluorescence range, e.g. 254 nm, there is no interference from fluorescence. While it might be difficult to apply standard laser emission wavelengths, like 532 nm or even 785 nm, to upstream processes due to fluorophores in cell culture media, these wavelengths can usually be utilized for downstream processing.

Besides lowering the excitation wavelength, the Raman signal intensity can be enhanced by increasing the laser power, increasing the interaction length between the laser and the sample by multiple-pass arrangements [166], or

increasing the collected light through sample optics with reduced photon losses in the spectrometer [167].
Feidl et al. [168] made a multi-pass flow cell by using a concave mirror behind a cuvette to increase the signal to monitor the breakthrough of a Protein-A column. Even though this is the first application of Raman spectroscopy to downstream processing, the publication shows that advanced chemometrics and a significant computational effort were necessary to reach a model that is comparable to UV spectroscopy combined with a basic PLS model [169]. It is worth noting that the obtained Raman spectra were dominated by water. Therefore, it might be possible that the displacement of water due to an overall increase in protein concentration may be important for the underlying correlation.

3.2.5 Multimodal Spectroscopy

As outlined by Rüdt et al. [67], one sensor alone will not be able to measure every product quality attribute during production. Even for measuring one quality attribute, the combination of multiple sensors might be necessary. For example, the real-time monitoring of the mean molecular weight during a flow-through Hydrophobic-Interaction Chromatography (HIC) step for a mAb has been realized by static light scattering and concentration measurements by UV spectroscopy [170]. Because the scattered-light intensity is not only influenced by the molecular weight but by the concentration as well, a concentration measurement is necessary to calculate the molecular weight. Based on the calculated mean molecular weight signal, the flow-through step was terminated after a 1.5 % dimer breakthrough. It should be mentioned that this setup is limited to near isocratic buffer conditions. For, e.g. Cation-Exchange (CEX) with high and low salt conditions and therefore a changing refractive index, additional sensors, like a refractometer, might be necessary for accurate quantification.

Another application of light scattering is the downstream process of Virus-Like Particles (VLPs). Rüdt et al. monitored the diafiltration reassembly steps of three different VLP constructs at different conditions with UV spectroscopy and light scattering [144]. The scattered-light intensity was correlated to the assembly progress and UV spectroscopy provided information on the concentration of the VLPs as well as the rate of the assembly due to changes in the local environment of tyrosine residues.

Another approach, besides calculating the attributes of interest from different sensors by physically founded equations, is to fuse all data for statistical model building. This approach was applied by Walch et al. [116], where fusing data from seven sensors lead to a total of 15,725 input variables. These input variables were then used for PLS model building to predict antibody concentration, High Moleculare Weight Variant (HMW), DNA, HCP, and monomer content by PLS regression. It is important to note that such an approach can lead to physically unrealistic results. In the study, the pH was used in a PLS model to predict the mAb concentration. PLS modelling is a linear regression approach, which can only handle nonlinearities to a point, where a linear approximation of a nonlinear problem is feasible. A logarithmic pH value might not be a meaningful input for a linear regression model without a variable transformation. Similarly, ratios, like HMW, DNA, or HCP content, as output values should be handled with care as they are not linearly related to unscaled spectroscopic data. In a small range, where the relationship between the ratio and the spectral data can be linearly approximated, the use of PLS models is feasible [171, 172]. For strong nonlinearities, nonlinear methods, like nonlinear PLS models [173] or ANNs [174], should be considered. In our opinion, for the prediction of ratios with values covering several orders of magnitude (i.e. DNA content, and HCP content) nonlinear methods should be used. Based on the data published by Walch et al., it cannot be precluded either that the PLS models rather correlate the DNA and HCP content to the inverse of the protein concentration than being based on an actual causal relationship. Therefore, these PLS models might only work in a limited design space, where every run has the same trends and the DNA and HCP concentration in the eluate is constant. Then, the DNA and HCP concentrations per part of mAb are only influenced by the mAb concentration and could be well predicted to unrealistic concentration limits for optical spectroscopy. Additionally, if a large number of input variables and only a small number of samples is available, spurious correlations between two data sets are likely to occur when variable selection is done even while using Cross-Validation (CV) [175].

Sauer et al. [117] used the same experimental setup as Walch et al. [116] but chose to use the statistical framework of STructured Additive Regression (STAR), which provides means to include a wide range of nonlinear effects into model building, e.g. by including bivariate interaction terms [176]. However, the authors chose to exclude bivariate interaction terms for all spectroscopy sensors due to the required computational power. Therefore, it remains unclear how the model structure reflects the nonlinear response of, e.g. the DNA and HCP to mAb concentration. The additional degrees of freedom do not only affect the computational demand during calibration. During validation, it also becomes far more challenging to assert that the model does not overfit compared to purely linear models.

When using multiple sensors in a process stream, it is important to account for dispersion between the detectors. Especially for lab-scale chromatographic setups, the peak will change its shape as the detectors are passed and time alignment alone might not be sufficient to overlay the signal of the different sensors. Here, proper data treatment and analysis are important to draw the right conclusions which will be discussed in the next chapter.

3.3 Advanced Data Analysis and Machine Learning

Machine learning refers to different algorithms to develop models for pattern recognition, classification, and prediction derived from existing data [177]. PLS models and its variations are the most frequently used machine learning methods for MVDA of spectral data in bioprocesses [178, 179]. In Figure 3.2, a general workflow for model building is depicted with illustrations from Raman spectral data for concentration determination as example. Generally, model building starts by choosing the design space for the model and recording spectral data. Subsequently, spectra are preprocessed, outliers are removed, and the data is pretreated to improve data quality. Model building may include CV and model optimization until the optimal model is found. Before productive use, it is compulsory to evaluate the model performance with an external data set as it has been shown, that internal validation is not sufficient [180]. All necessary steps to obtain a valid model are discussed in more detail in the following.

3.3.1 Sample Selection

Generally, it is advisable to choose samples, which are representative of the purpose of the model [181]. Therefore, known process variations should be included into the model. This could be done, for example, by recording different runs with variations in the normal operating ranges, like different batches, upper and lower limits for buffer composition, and load density of chromatography columns, etc. If there are no restrictions on the compositions of the samples, the use of a D-optimal design for a Design of Experiment (DoE) approach is applicable to the distribution of samples in the design space [182]. Regarding the minimal sample size required for PLS calibration, rough heuristic rules advocate at least five or ten samples per adaptive parameter, i.e latent variables [92, 183, 184].. Generally, it is not possible to choose more latent variables than calibration samples

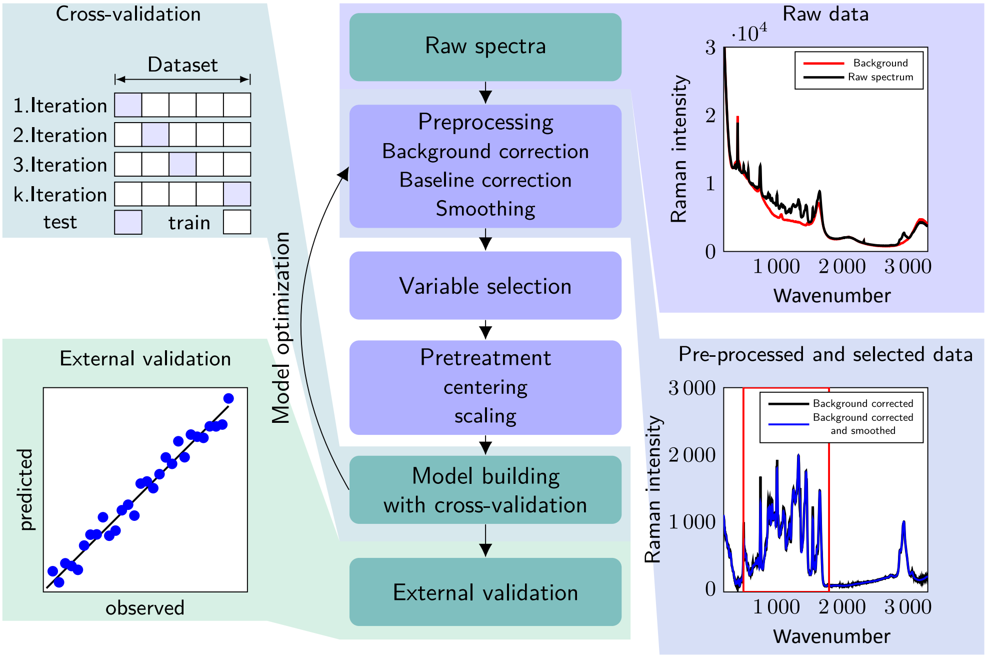

Figure 3.2: General workflow for PLS model building. More information on the different steps of the workflow are provided in Section 3.3.

as this is a restriction of the algorithm. PLS models with as many latent variables as samples will be without doubt over-fitted. Depending on the data complexity, PLS models for spectroscopic data can even have around 10 latent variables without over-fitting [96, 185]. The data set is split into calibration and external validation test set at a ratio of 2/3 to 3/4 in terms of calibration samples to the sample size of the data set [182]. The exact ratio depends on the sample size of the data set [182]. For smaller data sets with fewer samples, a higher ratio of calibration samples to available samples is chosen. To ensure a uniform distribution of calibration and validation samples over the design space, a supervised sample selection such as the Kennard-Stone algorithm, is preferred compared to random sampling [186].

3.3.2 Preprocessing

The objective of the preprocessing of spectral data is to remove extraneous variance, such that the data adheres closer to the Beer-Lambert law [187]. Depending on the spectroscopy method, different preprocessing steps are

required to reach this objective [188]. A review on preprocessing for Raman and FTIR is given by Gautam et al. [188]. For UV, 2D fluorescence, and light scattering usually no extensive preprocessing, except for the background correction, is necessary.
Often, the spectrometer software and correct calibration of the instrument remove instrument- or method-specific effects, such as detector nonlinearities, wavelength shifts, or interfering signals. Especially for Raman spectrometers, instrument calibration is necessary due to possible shifts in the laser excitation wavelength. Therefore, Raman spectrometers are generally calibrated with external light sources and reference substances to calibrate x- and y-axis and the laser wavelength [189]. Usually, cosmic rays are already removed before preprocessing begins.
The most common preprocessing steps for UV, NIR, MIR, fluorescence, and Raman spectroscopy include smoothing as well as baseline, background, and scatter correction [190]. Background correction procedures minimize the effect of a varying background caused by fluorescence, if applicable, of the sample or thermal fluctuations on the detector [191] and the buffer contribution to the spectrum for dissolved samples. Usually, if the background correction corrects for drifts of the spectrometer, no additional baseline correction is necessary. However, if a baseline correction is necessary, detrending, Alternating Least-Squares (ALS), or derivations [190] could be used. De-trending relies on fitting a polynomial to the spectrum and subtracting it from the spectrum while ALS involves an inert estimation of the background by an asymmetric least-squares fit. First-order derivatives eliminate a constant offset while second-order derivatives remove a constant offset and slope. Because derivatives make high-frequency noise more pronounced, Savitzky-Golay filters are often used to smooth and derive [187, 190]. However, Savitzky-Golay derivations are also prone to high-frequency noise, depending on the window width. High-frequency noise can influence the model and cause overfitting [187]. Therefore, (extended) Multiplicative Signal Correction (MSC) is generally recommended as preprocessing technique [187, 190, 192]. In practice, derivatives are still frequently used due to their simplicity and ease of use. For solely smoothing data, Savitzky-Golay filters are still the most used smoothers due to their superior preservation of peak shapes compared to e.g. the moving average filter [95].

For scatter correction, the MSC algorithm was developed by Martens et al. [193]. MSC uses a blank spectrum as reference, if available, or a mean of all recorded spectra to estimate correction coefficients for the spectra. Later on, the MSC algorithm was expanded to include the wavelength dependency of the scattering intensity and corrections for known spectra, referred to as Extended Multiplicative Signal Correction (EMSC). This caused the development of other de-trending techniques, like Orthogonal Signal Correction (OSC), Orthogonal PLS (O-PLS) [187]. The use of MSC or related techniques can reduce the number of latent variables in a PLS model and enhance the chemical information in the spectra to facilitate interpretation [95]. Additionally, the EMSC can normalize the spectra. However, normalization of spectra removes absolute concentration information and is therefore not recommended for concentration-dependent applications.

Generally, it is worth to keep in mind that preprocessing may also remove useful information (e.g. fine structures in the spectra, informative scattering effects) [192]. Therefore, it is sometimes beneficial to preprocess data less in order to preserve most information.

3.3.3 Outlier Detection

Proper handling of outliers is essential for data analysis because outliers introduce large variance to the model which can disturb the model [175]. PCA is a useful tool to look at the variance of the data to evaluate whether it is an unusual variance in the model plane or outside of the model plane. A common way to remove outliers within the model plane is to look whether samples lie outside of the 95% confidence limit of the *Hotelling's* T^2 ellipse in the PCA t_i vs. $t_{j \neq i}$ score plots for each score to another [194]. The ellipse shows the distance from the origin in the model plane with the chosen confidence. Additionally, outliers outside the model plane can be evaluated by calculating the distance of an observation in the training set to the model hyperplane [96] or by calculating the residuals of the observations [62].

As PCA reflects the main variations in the X-data, the results of a PCA-based outlier detection might be misleading if the main variations in the data is not correlated to the Y-variables [181]. As the purpose of preprocessing is to remove variance outside of the Beer-Lambert law, the main variance in the X-data should be correlated to the Y-variables. Outliers due to erroneous measurement should be removed before variable selection. Outliers with a large variance in the model should be either removed during sample selection due to the irrelevance to the model or be included as important process variance. However, outlier detection was not included in the general workflow for PLS model building depicted in Figure 3.2, because it can be part of sample selection with manual inspection of the spectra for erroneous measurements or take place before model optimization.

If in doubt whether to remove an outlier or not, it is useful to compare the models before and after removal. If the model changes dramatically, e.g. in the amount of latent variables, scores, etc., the outlier removal is important. Otherwise, the sample can be included [194].

Generally, outlier detection and removal can be automized, but it is important to point out the risk of automatic outlier removal. Outliers may carry valuable information about the system and process. For instance, the ozone hole could have been detected earlier, if it had not been for automatic outlier detection methods [195]. In context with optical spectroscopy in processes, outliers indicate unusual disturbances of the spectrum. Here, outliers could be used to detect process failures, e.g. equipment failures or air entrapment.

A more extensive overview of outlier removal is given by Hadi et. al. [196].

3.3.4 Variable Selection

PLS models and the corresponding conclusions can be highly dependent on the included X-variables [181]. Even though weighting of the X-variables according to the information content for the prediction of a univariate y-variable is an inherent property of the PLS algorithms, the inclusion of irrelevant and noisy variables can increase the prediction error of the PLS models [197]. Therefore, areas in the spectrum with high variance, but little to no correlation to the chemical properties of the sample, and areas containing only noise should be left out of the model to improve the prediction ability [175]. Further exclusion of X-variables can still improve the prediction ability of the model but the model robustness can decrease due to the increased risk of over-fitting by choosing less causal X-variables but with a higher correlation to y [175, 181]. Andersen et al. [175] showed that variable selection can lead to a statistical significant correlation of random X-data to a y-variable for more X-variables than samples even when using CV. Therefore, a comparison between selected variables and variables known for containing the desired information on the chemical or physical behavior of the system is important to prevent over-fitting and can give more insight into the data. A review of variable selection techniques would go beyond the scope of this manuscript. However, reviews about various variable selection methods for spectral data are given by Anderson et al. [175] or Mehmoood et al. [197].

3.3.5 Pretreatment

Data pretreatment strategies focus on the relation between different samples in one variables (i.e. column vectors), in contrast to preprocessing, which focuses on the different variables from one sample. Sometimes, pretreatment techniques are also referred to as preprocessing. In our opinion, distinct terms should be used to emphasize the underlying differences. Next to the already mentioned difference regarding to which matrix dimension the methods are applied (i.e. applied variable/block-wise versus in the spectral direction), it is also worth noting, that data pretreatment is not limited to the X-data but can also be applied to the Y-data. Importantly, the pretreated values will change, when samples are removed from the calibration set, while the preprocessed values stay the same.

Centering, scaling, or variable transformations are used as most common pretreatment techniques [198]. Mean-centering is often applied to data that is obtained with a single instrument, as all variables are defined with the same unit [96]. Centering may improve the numerical stability and interpretability of the results, as the model is focused on explaining data variance rather than data magnitude [99, 198].

Scaling methods divide each column vector by a different factor, e.g. to give each column vector a unit variance [96]. The goal of scaling is to reduce the influence of large numeric values in order to focus on correlating the X- to the Y-variables. Pretreatment is especially important if variables are measured by different sensors, as this may result in variables with different scales. Models, such as PLS and PCA, often try to explain the largest covariance in data, which is bias to variables with the largest numerical values [96]. There are a plethora of different scaling techniques to account for different effects [198] which is important for handling multiple differently scaled variables. This topic will be discussed further in Section 3.3.8.

Transformations are necessary if the numeric values of X-variables are not linearly correlated to the Y-variables for linear modeling. This can be important to e.g. diffuse reflectance intensities or pH values.

3.3.6 Model Building and Model Optimization

An important point during model building is to select the correct model type, when having multiple Y-variables. For spectral data where the Y-data (e.g. concentrations of multiple components) are not correlated, it is useful to make a PLS model for each component, also referred to as PLS1-models [95, 99].

During model building, it is essential to determine the correct number of latent variables for the PLS model, also referred to as model complexity. Due to numerous and collinear X-variables, there is a substantial risk of overfitting the model. Overfitting occurs, when added latent variables only fit random noise, which results in a loss of the predictive power. CV has proven to be a useful tool to determine the influence of latent variables on model performance and reduce the possibility of random correlations [99, 199].

To perform CV, the data set is divided into multiple subsets (between five to nine[200]), and PLS models are formed for a given number of latent variables until every subset has been left out once. Subsequently, the sum of squared differences between experimental and predicted Y-values is calculated for the left-out data for all computed models to estimate the predictive ability, or goodness of prediction Q^2, of the model. The number of latent variables is set to the lowest number where adding another variable does not significantly increase the predictive ability [96, 99].

Besides the number of latent variables, data preprocessing and variable selection are other approaches, that can be optimized in order to obtain an improved PLS model [201]. Preprocessing and variable selection usually rely on experience and manual inspection of the samples, where a certain preprocessing algorithm and windows of the spectra are selected. While this improves the performance of the PLS model, it is often not intuitive to find the best combination of all optimizable parameters [175]. Therefore, the use of a parallel Genetic Algorithm (GA) can be useful to find the optimal PLS model [202] to optimize the preprocessing and variable selection in one algorithm. However, since GA are prone to overfitting, it is important to use multiple GA runs and set the optimization parameters, e.g. window size, properly [203]. A comprehensible review on variable selection techniques was published by Andersen et al. [175].

A different approach for model optimization is used by Feidl et al. [168] and Narayanan et el. [204], where all useful combinations of preprocessing, pretreatment, outlier removal, smoothing, and variable selection were calculated and the best preprocessing and pretreatment method was chosen judged by the decrease in Root Mean Square Error of Cross Validation (RMSECV) and Root Mean Square Error of Prediction (RMSEP).

In this case the RMSECV and RMSEP indicated the same optimized preprocessing an pretreatment method. Therefore the model optimization was not infleunced by the RMSEP. Nevertheless, it is important to note, that models must not be optimized by use of the RMSEP. It is counterproductive to use the same key figure for optimization and evaluation of the model, because the model is then optimized to give the lowest RMSEP and not to find an actual correlation.

3.3.7 Model Validation

The goal of model validation is to ensure the quality of the prediction in terms of a causal and robust correlation [62]. There are several key figures to evaluate models [95, 96]. The Root Mean Square Error (RMSE) is the Predicted Residual Error Sum of Squares (PRESS) divided by the sample size n, see equations 3.2. For the calculation of the PRESS with Equation 3.1, y_i is the measured value and $\hat{y}_i$ is the predicted value. The difference between RMSECV and RMSEP is the used data to calculate the error. In case of the RMSECV, it is the RMSE of the samples, that were left out in the CV step, also known as internal validation. In case of the RMSEP, the samples from an external validation sets are used.

$$PRESS = \sum_{n=1}^{N} (y_i - \hat{y}_i)^2, \tag{3.1}$$

$$RMSE = \sqrt{\frac{PRESS}{n}} = \sqrt{\frac{\sum_{n=1}^{N} (y_i - \hat{y}_i)^2}{n}}. \tag{3.2}$$

Especially for small data sets, the RMSECV and RMSEP depend heavily on the used samples. Therefore when comparing different PLS models with the same data set, the same samples should be used for calibration and validation, respectivly. For comparison of different PLS models with different data sets, it is useful to evaluate the model by the coefficient of determination for the calibration R^2 after Equation 3.3, where $\bar{y}$ is the mean of y. The coefficient of determination for the CV Q^2 is calculated after Equation 3.3 as well for the left-out samples during CV. It should be noted, that the difference between R^2 and Q^2 are the samples used for calculation. R^2, also referred to as R^2Y is the variation of the Y-variables explained by the model. Q^2, also referred to as Q^2Y, is the variation of the Y-variables predicted by the model. It should be noted, that as a replacement for the RMSEP the Q^2_{ext} calculated with the external validation set used for the RMSEP calculation can be used as well to give a more representative key figure for the prediction ability on an external validation set [96].

$$R^2 = 1 - \frac{PRESS}{\sum_{n=1}^{N} (y_i - \bar{y})^2} = \frac{\sum_{n=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{n=1}^{N} (y_i - \bar{y})^2}. \tag{3.3}$$

While statistic methods try to establish a correlation between X- and Y-variables, it is important to emphasize that this correlation might not necessarily be a causal relation [95, 179, 181]. Even if model building was successful, a spurious correlation or an indirect correlation possibly may have been found. Indirect correlations can sometimes be used to quantify a component A, if, e.g. actually component B is measured, but is converted into component A at a fixed ratio [205]. Even in this case, it is useful to be aware of this indirect correlation to draw the right conclusions from the model. Indirect and spurious correlation have been widely discussed for Quantitative Structure-Activity Relationship (QSAR) models, because QSAR models can be prone to these kinds of correlation due to the vast amount of X-variables which make it possible to almost always find some kind of correlation. For verification of meaningful correlations, Wold et al. [200] published a method consisting of originally four tools for model validation of QSAR models that can be adapted for spectral data resulting in three different tools.

Tool 1 is the permutation test (also referred to as significance test or randomization test). The main idea is to repetitively randomize a certain amount of the Y-variables in the training set while the X-data stays intact. In each cycle, the full data analysis is carried out on these scrambled data and the R^2 and Q^2 values are recorded. If, in each case, the scrambled data give much lower R^2 and Q^2 values than the original data, it is likely, that a real correlation was found.

Tool 2 is CV as explained above. It is a frequently applied and useful approach to model validation. However, CV results may also be misleading. If the validation groups during CV are too small, the model selection is biased. For example, if the number of groups is equal to the sample size, also referred to as leave-one-out, the permutation during the CV is too small and the resulting Q^2 values will approach the R^2 value [206]. In practice, five to nine subsets are recommended [200]. Additionally, CV might not work for variable selection, because only the variables with correlation to the Y-data are chosen and this might lead to the selection of X-variables with spurious correlations to Y [207].

Tool 3 is related to appropriate sample selection and in particular the external validation set. Ideally, an external validation data set should span across the complete design space in an evenly distributed manner. The validation set can also include samples outside the calibrated range for the Y-values to improve the confidence in the built model.

We recommend the use all of these tools for model validation to avoid spurious correlations, especially tool 1. When looking at the data published by Walch et al. [116], tools 2 and 3 have been applied, but not tool 1. A permutation test and inclusion of the mAb concentration as X-variable could reveal in this example if the concentrations of DNA, HCP, and HMW were predicted from the mAb concentration. For increasing mAb concentrations, decreasing impurity levels were calculated and vice versa. This may have little to do with actual concentration measurements of these components, because the amount of impurities per mAb concentration is not constant for every sample and batch. Especially when a large number of X-variables from different sensors are available, extensive variable selection can lead to spurious correlations [175].

3.3.8 Data Fusion

When multiple or multimodal sensors are involved in a measurement, different data fusion strategies can be utilized for model building [208]. Data fusion is generally categorized into low-level, mid-level and high-level data fusion [121, 209, 210]. A general overview is given in Figure 3.3. Here, each sensor provides a block of data which needs to be fused to all the other blocks for analysis. Low-level data fusion concatenates the different raw or preprocessed data blocks and applies an appropriate block-wise pretreatment before model building. This is important, because the variables in the blocks typically have different scales. Variables with a higher numeric value would otherwise contribute more to the model. To overcome this problem, unit variance scaling could be performed. Block scaling can be used to multiply the block with an additionally scaling weight to account for the importance of these variables for the prediction of the Y-variable [96].

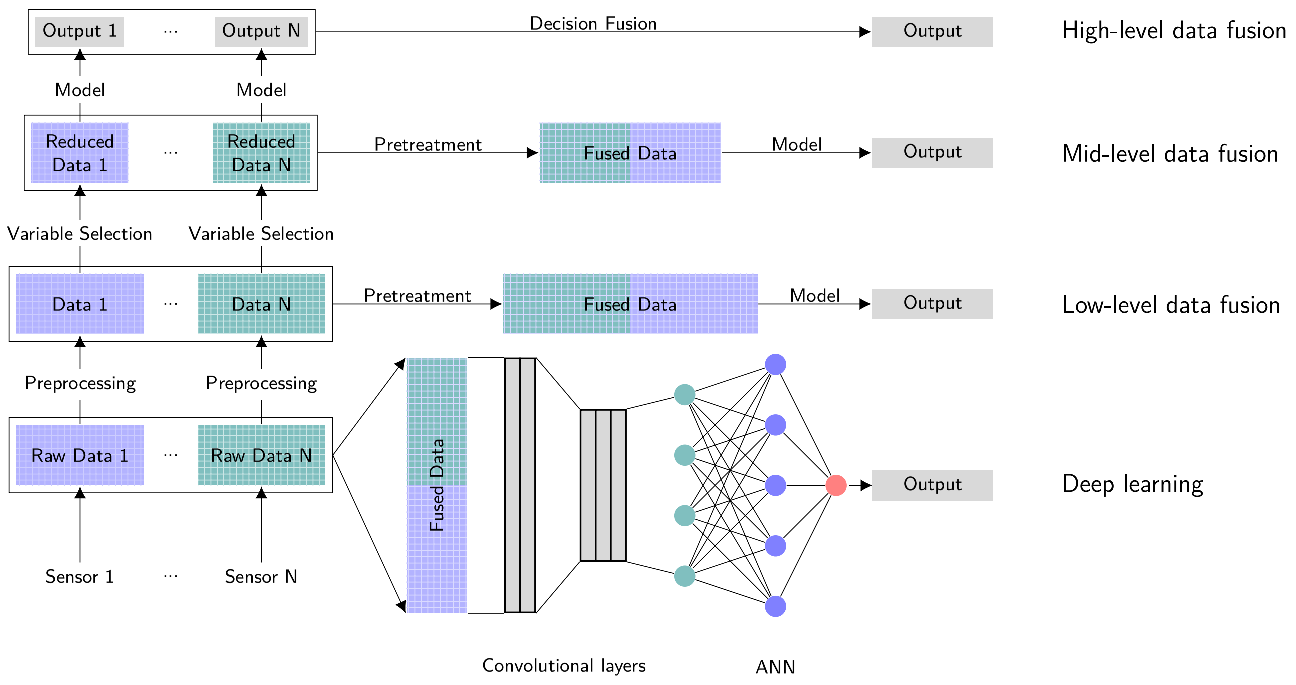

Figure 3.3: Methodology for model building in low-level, mid-level and high-level data fusion and, additionally, deep learning. Adapted after [121].

Mid-level data fusion applies variable selection before concatenating the different data blocks to reduce the influence of a large amount of unimportant variables. This can be done by variable selection for the data blocks or by hierarchical multiblock PLS. Hierarchical multiblock PLS is based on the decomposition of the blocks into scores and latent variables. The obtained block scores are subsequently used for PLS model building on the upper level [211]. This increases the interpretability of the model, because the relations between the blocks are emphasized due to the upper data level from which the model is built. An additional benefit of hierarchical multi-block PLS is the improved prediction of the block models as they are less sensitive to mild scaling inaccuracies [211].

High-level data fusion is a fusion of the outcome of a model. Therefore, it may rather be termed decision fusion than data fusion [118]. This means that block-scaling is unnecessary and the models can be separately optimized. Methods for decision fusion include different techniques like weighted decision methods, Bayesian inference, Dempster-Shafer inference or fuzzy logic theory [212]. Additionally, if a time dependency is available, state estimation methods like Kalman-filters can be used.

Recently, CNNs have gained momentum in spectral analysis [213–215]. Originally, CNNs were designed to cope with shift and distortion variances for image recognition [216] or speech recognition [217], which is desirable for spectral analysis as well. CNNs are a variant of feed-forward ANNs with additionally convolutional layers to filter the data by weighting the summation of the inputs in windows [218]. The kernels in the convolutional layers are sparsely connected and share weights. CNNs focus rather on local features, which makes them easier to train and interpret, and less prone to overfitting [214]. In higher structural data, pooling layers are used to pool similar features and bring the data in 1D form. For spectral data (already in 1D form), pooling layers are not always used [214].

CNNs are the oldest form of deep learning architectures [219] with multiple levels of nonlinear functions due to many hidden layers. This architecture of CNNs results in a filter ability. Therefore, CNNs can handle raw data, which can make human interference for preprocessing the data unnecessary [216]. However, it has been shown, that CNNs work better on preprocessed data similar to how PLS models behave [214]. CNNs are highly flexible and can fit highly nonlinear correlations. Nevertheless, for linear problems, usually linear methods perform better [220].

3.4 Perspectives for the Biopharmaceutical Downstream Process

This final section of the review is intended to give a more abstract view of the present and future of PAT in downstream processing of biopharmaceutical proteins. A special focus is set on different product- and process-related impurities and on how the current approaches could be further integrated towards holistic process monitoring.

In biopharmaceutical processes, relevant impurities and the product need to be monitored and controlled in a broad concentration range. Figure 3.4 illustrates this with the typical concentrations occurring during manufacturing of a mAb. Figure 3.4 also includes the typically maximum allowed impurity concentrations in the drug product. Information on the involved data analysis is provided in the Supplementary Data. Considering the lowest and highest relevant concentrations for both contaminants and mAb, downstream processing is spanning more than seven orders of magnitude of concentration values. Furthermore, each species is a diverse group of substances. For example, the term HCP refers to any protein produced by the host cells in addition to the target product. Thus, HCPs are a very diverse group of proteins which additionally complicates detection or concentration measurements of these contaminants [221, 222]. While the diversity for other species in biopharmaceutical production may not be as extreme as for HCPs, similar arguments hold for DNA, aggregates, fragments, or other product isoforms. The broad concentration ranges in combination with the diversity of the relevant species in downstream processing pose a major challenge for PAT.

In recent publications, implemented in-line soft sensors (spectroscopic methods in combination with chemometrics) achieved limits of detection for aggregate and fragment levels below the concentration limits set by the regulatory agencies for drug products [73, 114, 170]. On a lab-scale, the feasibility for measuring these important contaminants with the necessary accuracy was thus demonstrated. Future projects may work towards a closed-loop control of the process steps of interest. Product-related isoforms occur at similar concentrations as aggregates and fragments. Spectroscopic PAT methods are likely to achieve similar limits of detection as long as there is a measurable change in the spectroscopic properties of the isoforms. It seems likely that some processes may also use spectroscopic soft sensors for controlling isoform profiles in the future. However, there also remains a large fraction of isoforms which cannot be distinguished from the product

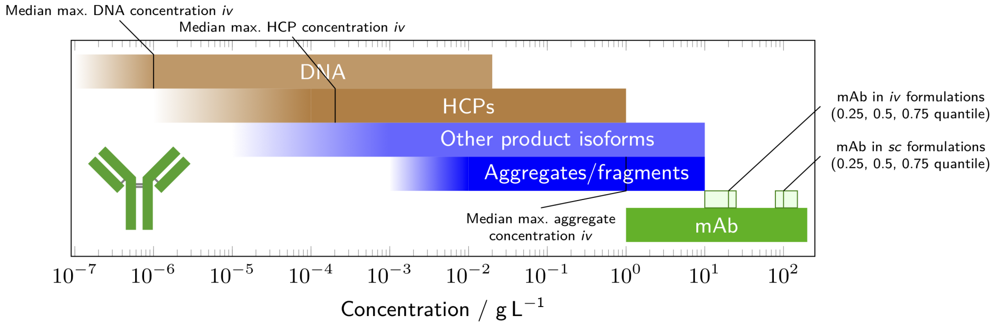

Figure 3.4: In biopharmaceutical processes, different species need to be monitored in a concentration range spanning many orders of magnitudes. This is illustrated here by the example of mAb processes. Each horizontal bar denotes concentration ranges for the major species covered in biopharmaceutical processes. In green, the mAb concentration is shown. The boxes in light green correspond to the monoclonal antibody concentrations of the marketed mAbs in US for *intravenous (iv)* and *subcutanous (sc)* administration. Product- and process-related impurities are shown in blue and brown, respectively. Impurity concentration limits as accepted by the regulatory agencies are marked by black lines in the corresponding concentration bars.

by optical spectroscopy. In such cases, other sensors or control strategies should be evaluated.

For the process-related impurities HCPs and DNA, in-line monitoring may be achievable for early steps in downstream processing, such as capture steps, where the process-related impurity concentrations are still high. During further polishing steps, process-related impurity concentrations are typically by a factor of 10^5 to 10^8 lower than the product concentration. To further complicate detection, HCPs are polypeptides and therefore chemically highly similar to the product. DNA is more distinct from the product, but typically also occurs at the lower end of the concentration scale. Based on regulatory guidelines, DNA must be depleted to concentrations approximately 10^7 times lower than the product concentration. The quantification of HCP and DNA by optical spectroscopic PAT methods towards the end of the downstream process seems very challenging and probably not achievable in the near future. Furthermore, at the current state of research, a purely measurement-driven approach does not seem practical for monitoring and controlling all CQA in downstream processing in real-time.

Fortunately, there are alternative approaches to monitoring and controlling production processes. For example, model-based predictions of CQAs from observed process parameters have reached an impressive accuracy in a number of studies [223–225]. These studies showed that statistical models can capture a significant amount of the hidden process dynamics and the effects on the CQA of the product while neglecting the actual time evolution of the system. In a next step, it would be interesting to also obtain time-dependent predictions of the process trajectory. Here, mechanistic, hybrid, or empirical models could be applied to predict the underlying system dynamics. As soon as a fast dynamic process model for different CQAs is available, the model could also be leveraged for process control.

While different approaches to process control exist, MPC is regarded as one of the most important tools in advanced process control [226, 227]. MPC is well established in various industries including refining, petro-chemical, and food applications [228]. MPC is founded on a mathematical model of the process dynamics, i.e. a model which describes the time evolution of the investigated system. To control the process, the model is leveraged by taking current and future process dynamics into account. Based on the model and an objective function, MPC aims to optimize the process performance over a given time frame into the future (the so-called receding horizon) by calculating a number of control actions. At each time step, an optimization is performed to find the optimal control actions. Then, the first calculated control action is applied to the system and the optimization is repeated with the receding horizon reaching one time step further into the future. This approach allows to neglect the future of the process beyond the receding horizon, thus simplifying the control problem. Among the benefits of the MPC framework is also its high flexibility. MPC provides means for accepting input variables, maintains an estimate of the current system state, and predicts the current and future plant outputs. Due to the model-based foundation of MPC, it is particularly well aligned with the motive of Quality by Design (QbD) of building the quality into the product through product and process understanding (see [229] for an extended discussion).

MPC was already investigated for a number of applications in biopharmaceutical manufacturing. For upstream processing, a number of different MPC schemes have been applied and reviewed [229, 230]. For downstream processing, research focused on the control of continuous chromatography. MPC for Multi-Column Solvent Gradient Purification (MCSGP) was developed and advanced in a variety of publications [231–233]. The application of MPC allowed for improved process performance and robust control of the purification processes as demonstrated by *in silico* studies. The need for reliable PAT was pointed out multiple times to provide feedback to the model. Initial research also exists towards coupling upstream and downstream unit operations *in silico* for an overall advanced process control [232].

Regarding process- and product-related impurities, MPC and its underlying model could build the basis for controlling CQAs based on inferred sensing of different species. In such a scenario, inferred state variables may track CQAs (e.g. HCP and DNA concentration) within the process which are not directly available from measurements [62, 227]. Based on an in-depth understanding ingrained into a model, MPC provides the ability to control impurities throughout the process, building a so called Digital Twin of the production. An additional key advantage of MPC is its capability to respect constraints. Thus, the objective function can be adjusted to fulfill the predefined quality metrics. Based on such an approach, manufacturing can be tailored towards Real-Time Release (RTR) [155].

3.5 Appendix: Calculations of Molecular Cross-sections and Absorption Coefficients

Equation 3.2 from Singh et al. [234] was used to convert molar absorption coefficients ϵ_{molar} in L/(mol cm) to molecular cross-sections σ in cm^2/Molecule.

$$\frac{\sigma}{\mathrm{cm}^2} = 3823 \cdot 10^{-24} \frac{\epsilon_{molar}}{\mathrm{Lmol}^{-1}\mathrm{cm}^{-1}} \tag{3.4}$$

The molar absorption coefficient ϵ_{molar} was calculated from the absorption coefficient ϵ in L/(g cm) and the molar mass M in g/mol according to Equation 3.5.

$$\epsilon_{molar} = \frac{\epsilon}{M} \tag{3.5}$$

3.5.1 Appendix: Fluorescence

Trypthophan is the most dominant aromatic amino acid in the UV spectrum regarding the absorption coefficient. It's quantum yield is 0.13 [129]. This information was used to convert the absorption coefficient at 280 nm to an emission coefficient.

3.5.2 Appendix: MIR

Typically, mAbs consist mainly of β-sheet secondary structure elements [235]. The extinction coefficient of C=O stretch in the amid I band at 1619/cm for β-sheet structures is 980 L/(mol cm) [236, 237]. For the calculations, it was assumed, that mAbs have roughly 1500 peptide bonds.

3.5.3 Appendix: NIR

NIR band intensities are much weaker than their corresponding MIR fundamentals by a factor of 10 to 100 depending on the order of the overtone [150].

3.5.4 Appendix: Raman

The Raman scatter cross-section was calculated from recorded data through comparision of the amid I band with the scattering area of water. The Raman scatter cross-section of water 5×10^{-30}/cm and a molar concentration of water of 55.5 mol/L were used for the calculation [238].

3.5.5 Appendix: Rayleigh scatter

11 nm was used as hydrodynamic diameter of a standard antibody [86, 239]. The Rayleigh scatter cross-section was calculated after Cox et al. [240].

4

Real-time Monitoring and Control of the Load Phase of a Protein A Capture Step

Matthias Rüdt*,1, Nina Brestrich*,1, Laura Rolinger1, Jürgen Hubbuch1

* Contributed equally

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

Abstract

The load phase in preparative Protein A capture steps is commonly not controlled in real-time. The load volume is generally based on an off-line quantification of the mAb prior to loading and on a conservative column capacity determined by resin-life time studies. While this results in a reduced productivity in batch mode, the bottleneck of suitable real-time analytics has to be overcome in order to enable continuous mAb purification. In this study, PLS modeling on UV/Vis absorption spectra was applied to quantify mAb in the effluent of a Protein A capture step during the load phase. A PLS model based on several breakthrough curves with variable mAb titers in the harvested cell culture fluid was successfully calibrated. The PLS model predicted the mAb concentrations in the effluent

of a validation experiment with a RMSE of 0.06 mg/ml. The information was applied to automatically terminate the load phase, when a product breakthrough of 1.5 mg/ml was reached. In a second part of the study, the sensitivity of the method was further increased by only considering small mAb concentrations in the calibration and by subtracting an impurity background signal. The resulting PLS model exhibited a RMSE of prediction of 0.01 mg/ml and was successfully applied to terminate the load phase, when a product breakthrough of 0.15 mg/ml was achieved. The proposed method has hence potential for the real-time monitoring and control of capture steps at large scale production. This might enhance the resin capacity utilization, eliminate time-consuming off-line analytics, and contribute to the realization of continuous processing.

4.1 Introduction

A capture step is the first unit operation in the protein purification process which is used to bind the target protein from crude HCCF. It increases product concentration as well as purity and prevents proteolytic degradation. Due to its high selectivity, Protein A capture is widely used in current mAb purification platform processes [46, 49, 50, 241, 242].

A difficulty in Protein A capture is a lack of real-time analytics for mAb quantification in the HCCF and in the column effluent during loading. As both the mAb and impurities contribute to the absorption at 280 nm (A_{280}), single wavelength measurements are not suitable as selective analytics [243]. To determine the mAb titer in the HCCF, elaborate off-line analytics is commonly performed [18, 19]. As mAb titers are influenced by variability in the cell culture, this off-line analytics has to be repeated for every lot in order to adapt the load volume onto the column [18]. While this results in a reduced productivity in batch mode, the bottleneck of suitable real-time analytics has to be overcome to enable continuous mAb purification.

In addition to the mAb titer in the HCCF, the optimal load volume onto the column is also influenced by the resin capacity. Due to leaching and degradation of the Protein A ligands as well as pore and ligand blocking by leftover impurities or product, the capacity of the resin decreases over cylce time [244]. In batch mode, a conservative loading is commonly applied to avoid breakthrough of the expensive product at the cost of productivity. In contrast to that, columns are overloaded in continuous mode to maximize productivity [245]. In this case, the determination of the the percentual product breakthrough is necessary for process control [246].

To perform (near) real-time process monitoring and control, several PAT tools have been developed to enable fast mAb quantification in the cell culture fluid and in the column effluent during loading. For instance, at-line mid-IR spectroscopy in combination with multivariate data analysis has been applied for secreted mAb quantification during a CHO cell culture process [247]. Selective mAb quantification in upstream processing was also successfully realized by at-line matrix-assisted laser desorption/ionization mass spectrometry [248]. For the control of the load phase of a two column continuous protein A chromatography process, which was connected to a CHO perfusion culture, at-line analytical chromatography was applied [56]. At-line monitoring however bears the risk of human errors resulting in contamination, time-delays, or missing data.

In order to minimize human impact, automated sampling can be applied. Automated analytical chromatography has been used in upstream processing to monitor the mAb titers [249–251]. In downstream processing, this technique was successfully used for mAb quantification in the column effluent during the load phase of Protein A chromatography. As soon as 1 % mAb breakthrough was detected, the load phase was automatically terminated [18]. Automated analytical chromatography is relatively easy to develop and equipment is commercially available. However, the equipment is expensive and the technique error-prone. Besides from the risk of contamination, the time delay between sampling and analytical results bears the risk of late reaction or requires a slow-down of the process.

PAT tools that operate in real-time, such as UV-based methods, overcome these limitations. In a patent application, a UV-based control method for determining binding capacities in Protein A capture was disclosed [252]. The method is based on the calculation of a difference signal between two detectors situated at the column in- and outlet. During the load phase, the post column signal is supposed to stabilize and is referred to as impurity baseline. As soon as the mAb breaks through, there is an increase in the post-column UV signal above the impurity baseline which corresponds to a breakthrough level of the product. Consequently, the method is very suitable for determining column switching times in continuous Protein A capture. It allows for an equal loading in terms of percentual breakthrough regardless of the mAb titer variability in the feed or decreasing column capacities. However, it requires two detectors posing a risk of unequal detector drifts. A further limitation might be displacement effects of contaminants that prevent a stabilized impurity baseline. The technique might also be limited to the equipment of the future patent holder.

Another recently published UV/Vis-based method for monitoring and control in protein chromatography applies UV/Vis absorption spectra instead of single wavelength measurements [72, 73]. Different protein species exhibit distinct variations in their UV absorption spectra. Consequently, PLS technique has been used to correlate absorption spectra with selective protein concentrations. The method was successfully applied for a selective in-line protein quantification and for product purity-based pooling decisions in real-time. However, no load control in Protein A chromatography has been performed so far using this technique.

In this study, PLS models correlating UV/Vis absorption spectra with mAb concentrations were applied for real-time monitoring and control of the load phase in Protein A chromatography. In contrast to previous publications in this field, this application requires the monitoring of one protein in the background of many protein and non protein-based contaminants. For the PLS model calibration, several breakthrough experiments were performed and the corresponding absorption spectra of the effluent were acquired. In order to generate variable mixing ratios of mAb and contaminants for a PLS model training data set, experiments with variable mAb titers in the feed were performed. The column effluent was collected in fractions and analyzed using analytical Protein A chromatography. The recorded absorption spectra were averaged according to the fraction time and correlated with the determined mAb concentrations using PLS technique. The PLS model was eventually applied for a real-time control of the load phase and terminated loading, when 5 % or 50 % product breakthrough was reached.

4.2 Materials and Methods

4.2.1 Cell Culture Fluid and Buffers

HCCF and mock were obtained from Lek Pharmaceuticals d.d. (Mengeš, Slovenia) and stored at −80°C before experimentation. The HCCF and mock were filtered with a cellulose acetate filter with a pore size of 0.22 μm (Pall, Port Washington, NY, USA) before use. In order to achieve a variable mAb concentration in the feed, the HCCF was diluted with mock.

For all preparative runs, the following buffers were applied: Equilibration with 25 mM tris and 0.1 M sodium chloride at pH 7.4, wash with 1 M tris and 0.5 M potassium chloride at pH 7.4, elution with 20 mM citric acid at pH 3.6, sanitization with 50 mM sodium hydroxide and 1 M sodium chloride, and storage with 10 mM sodium phosphate, 130 mM sodium chloride, 20 % ethanol.

For analytical Protein A chromatography, column equilibration was carried out using a buffer with 10 mM phosphate (from sodium phosphate and potassium phosphate) with 0.65 M sodium ions (from sodium chloride and potassium chloride) at pH 7.1. Elution was performed with the same buffer, but titrated to pH 2.6 with hydrochloric acid. All buffer components were purchased from VWR, West Chester, USA. The buffers were prepared with Ultrapure Water (PURELAB Ultra, ELGA LabWater, Viola Water Technologies, Saint-Maurice, France), filtrated with a cellulose acetate filter with a pore size of 0.22 μm (Pall), and degassed by sonification.

4.2.2 Chromatographic Instrumentation

All preparative runs were realized with an Akta Pure 25 purification system controlled with Unicorn 6.4.1 (GE Healthcare, Chalfont St Giles, UK). The system was equipped with a sample pump S9, a fraction collector F9-C, a column valve kit (V9-C, for up to 5 columns), a UV-monitor U9-M (2 mm pathlength), a conductivity monitor C9, and an I/O-box E9. Additionally, an UltiMate 3000 Diode Array Detector (DAD) equipped with a semi-preparative flow cell (0.4 mm optical pathlength) and operated with Chromeleon 6.8 (Thermo Fisher Scientific, Waltham, USA) was connected to the Akta Pure. The DAD was positioned between the conductivity monitor and the fraction collector.

The communication between Unicorn and Chromeleon was implemented analogous to the protocol published in [72]. Shortly, Unicorn triggers the DAD data acquisition by sending a digital signal to a Matlab script (MathWorks, Natick, USA), which communicates with Chromeleon via a Visual Basics for Application Macro (Microsoft, Redmond, USA). If a certain condition such as a defined mAb concentration is fulfilled, the Matlab script sends a signal back to Unicorn to terminate a phase in the chromatographic method.

Reference analysis of collected fractions was performed using a Dionex UltiMate 3000 rapid separation liquid chromatography system (Thermo Fisher Scientific). The system was composed of a HPG-3400RS pump, a WPS-3000 analytical autosampler, a TCC-3000RS column thermostat, and a DAD-3000RS detector.

4.2.3 Chromatography Runs

In order to generate variable mixtures between mAb and impurities for the PLS model calibration and validation, breakthrough experiments with variable mAb titers in the feed were performed. The mAb titers in the different experiments were 2.7, 2.85, 3, 3.15, and 3.3 mg/ml. For each experiment, a Sartobind 2 ml Protein A membrane (Sartorius, Göttingen, Germany) was first equilibrated for 3 membrane volumes (MVs) and then loaded with 33.15 mg of mAb. At the beginning of the load phase, the DAD was triggered to record absorption spectra between 200-410 nm and the membrane flow-through was collected in 200 μl fractions. After a first wash with equilibration buffer for 4.5 MVs, the membrane was flushed with wash buffer for 5.5 MVs and with equilibration buffer for 4.5 MVs. Elution was carried out for 5 MVs followed by a re-equilibration of 1.5 MVs. Eventually, the column was sanitized for 5 MVs and, between the runs, kept in the storage buffer. The flow rate was 1 ml/min for all phases and experiments.

4.2.4 Analytical Chromatography

As displayed in Figure 4.1, the collected fractions of all runs were examined by analytical Protein A chromatography to obtain the mAb concentrations. For each sample, a 2.1x30 mm POROS prepacked Protein A column (Applied Biosystems, Foster City, USA) was equilibrated with 2.6 column volumes (CVs) of equilibration buffer, flowed by an injection of 20 µl sample. The column was then equilibrated with 0.8 CVs of equilibration buffer and eluted with 1.4 CVs of elution buffer. The flow rate was 2 ml/min for all phases and experiments.

4.2.5 Data Analysis

For the correlation of the absorption spectra with the mAb concentrations, PLS technique was applied using SIMCA (MKS Data Analytics Solutions, Umeå, Sweden). SIMCA applies the NIPALS-algorithm for PLS. Before performing PLS, all spectra were preprocessed by mean centering using SIMCA. PLS finds variation in the spectral data matrix, which is relevant for the correlation with the mAb concentrations and thereby separates information in the matrix from detector noise [96, 172, 253]. In order to achieve this separation, collinearity in the data is reduced by summarizing variables (here wavelengths) with similar information in LVs. This is done in a way such that the content of relevant information for the correlation included in each LV is highest for the first LV and decreases for the following ones. The number of applied LVs in a PLS model is hence a measure of data reduction and only a few LVs are required to obtain the correlation between absorption spectra and mAb concentrations.

The number of applied LVs has to be evaluated thoroughly to avoid under- or overfitting of a model. In order to determine a reasonable number of LVs, the root mean square error (RMSE) for the prediction of validation

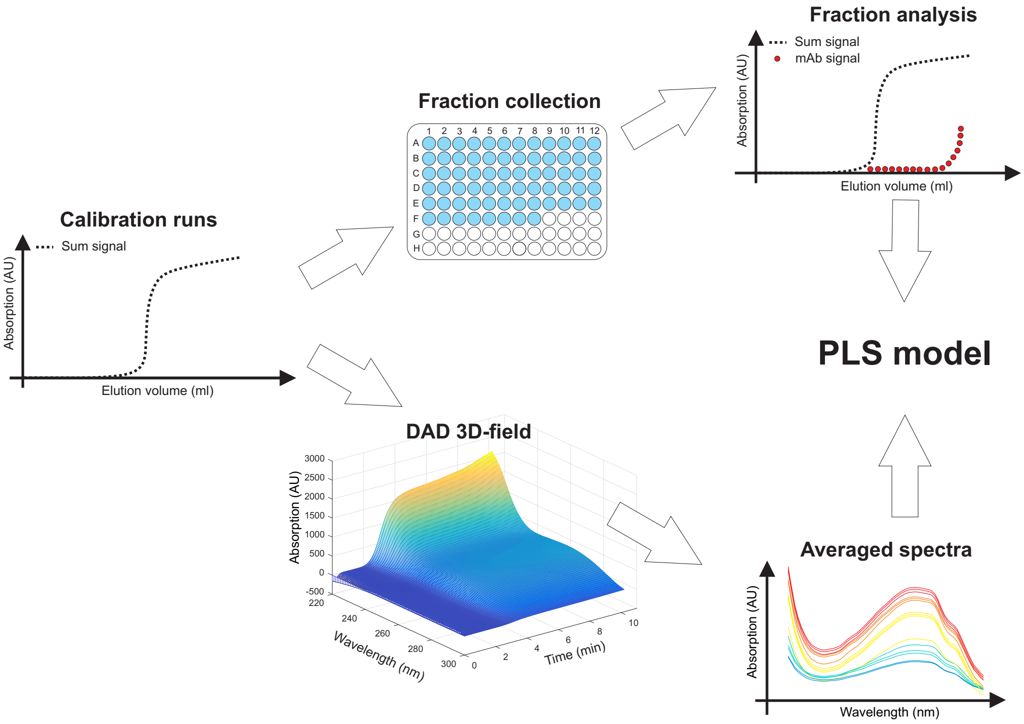

Figure 4.1: Experimental procedure for the PLS model calibration: For each calibration run, 200 μl fractions were collected and analyzed by analytical Protein A chromatography to obtain the mAb breakthrough curves. In addition, averaged spectra corresponding to the fraction size were calculated from the time, wavelength, and absorption 3D-field. Averaged spectra and mAb concentrations were eventually correlated using PLS technique.

samples is usually determined in dependence on the number of LVs applied in a PLS model. The minimum corresponds to the optimal number of LVs. In this study, cross validation was performed to determine an optimal number of LVs. Therefore, the calibration data was separated into seven groups. One group was then excluded during model calibration and the RMSE for theses samples was calculated subsequently. For every number of LVs, this procedure was performed until each group was excluded. Based on the so obtained number of LVs, completely independent runs were predicted to evaluate the final models.

A first PLS model calibration was based on the results of the runs with the following mAb titers in the feed: 2.7, 2.85, 3.15, and 3.3 mg/mL. The results of the corresponding spectral acquisitions are time, wavelength and absorption 3D-fields. The 3D-fields were averaged in time according to the fraction duration as displayed in Figure 4.1. The results of theses calculations were stored in an absorption matrix. Afterwards, PLS was carried out to correlate the mAb concentrations of the collected fractions with the the corresponding absorption matrix. For lower protein concentrations, a second PLS model was calibrated. Only samples with mAb concentrations below 0.5 mg/mL were considered in the model calibration. For those samples, a background subtraction was performed. As soon as the change in absorption signal after impurity breakthrough fell under a predefined threshold, an average absorption was calculated for every wavelength. This impurity background was subtracted from the absorption of all following data points.

4.2.6 Real-Time Monitoring and Control

The first calibrated PLS model was subsequently applied for a real-time monitoring of the mAb concentrations in a run with a mAb titer of 3 mg/mL in the feed. While the calibration of the PLS model was performed using averaged spectra, predictions were based on the 3D-fields. This means that the a spectrum at each time point was applied to predict the mAb concentrations. The absorption spectra of the effluent were recorded and translated into mAb concentrations in real-time by the calibrated PLS model. The calculation of the mAb concentrations was executed in Matlab. In a first run, a stop criterion of 1.5 mg/mL mAb concentration (50 % product breakthrough) was set in the Matlab evaluation script. As soon as the termination criterion was reached, a digital signal was send from Matlab to Unicorn and the load phase was terminated. In a second run, the stop criterion to terminate the load phase was set to a target concentration of 0.15 mg/mL (5 % product breakthrough). For this condition, the second PLS model was used.

4.3 Results and Discussion

As described above, the breakthrough of mAb was monitored in real-time by UV/Vis spectroscopy in combination with a PLS model. To calibrate the PLS model, 4 chromatographic runs at mAb concentrations of 2.7, 2.85, 3.15 and 3.3 mg/mL in the feed were performed and analyzed by off-line analytics. The model was eventually confirmed by performing a real-time control of two runs with a mAb titer of 3 mg/mL. The difference in the mAb titers in the feed ensured variable mixing ratios between product and contaminants. This was done to imitate variability in upstream processing and to span a calibrated design space for the PLS model.

4.3.1 PLS Model Calibration

The results of the model calibration are illustrated by Figure 4.2. It compares the A_{280} (recorded at a pathlength of 0.4 mm and displayed as dashed black line) to the concentrations measured by off-line analytics (blue bars) and the signal calculated by the calibrated PLS model (solid red lines). The number of LVs was set to 4 based on a minimal RMSE of 0.08 mg/mL in the cross validation. The calibrated PLS model was applied to evaluate all 3D-fields. In contrast to model calibration, where averaged spectra were used, the spectral raw data at each time point was translated into concentrations. The estimated concentrations by the PLS model closely follow the measured values by off-line analytics. It is worth noting that no clear plateau of the A_{280} is reached after the breakthrough of media components. Instead, the A_{280} continuous to increase. This may be caused by different impurities being retained differently on the membrane. Indeed, it has previously been shown, that major interactions between HCPs, the stationary phase and mAbs may occure [50, 254]. The advent of mAb breakthrough cannot be clearly distinguished from A_{280} alone. Based on the multivariate spectral data, the PLS model is able to predict protein concentrations, which allows for real-time monitoring and control.

4.3.2 Real-Time Monitoring and Control

For the confirmation of the obtained results, the calibrated PLS model was used to control the load phase of a Protein A capture step in real-time. In a first run, a target breakthrough concentration of 1.5 mg/mL was set, which corresponds to 50 % product breakthrough. Figure 4.3 A shows the A_{280} (dashed black line), the real-time prediction of mAb concentrations (solid red line) and the corresponding off-line analytics (blue bars). The model reached an RMSE for prediction of 0.06 mg/mL compared to the off-line analytics. This approach may be of interest for controlling a continuous chromatography system. In this context, the prediction of lower mAb concentrations is not so crucial.For a possible application in batch chromatography, the sensitivity of the model was further improved. A second PLS model was hence calibrated based on the calibration data set as described in the method section. The recalibration was performed to increase the sensitivity in the given concentration range. It was noticed, that it is difficult to accurately calibrate a PLS model for broad concentration ranges. By reducing the concentration calibration range, smaller RMSE values could be achieved. The model was used predict and stop a load phase in a second run at 0.15 mg/mL, which corresponds to 5 % product breakthrough. The results

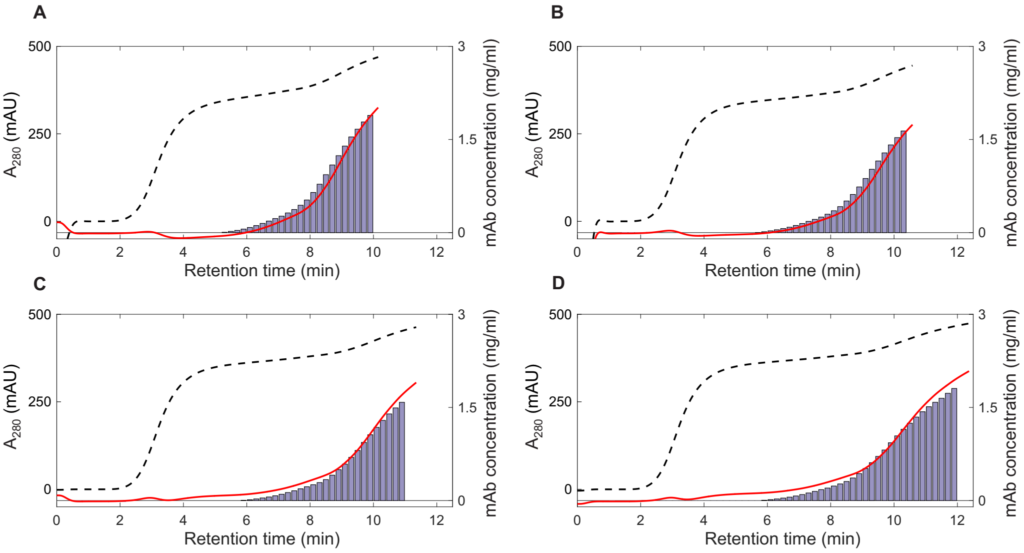

Figure 4.2: Results of the PLS model calibration. The A_{280} (measured at a pathlength of 0.4 mm and displayed as dashed black line) is compared with the results of the off-line analytics for mAb quantification (blue bars). The PLS model prediction is illustrated as red lines. The four runs exhibited variable mAb titers in the feed A: 3.3 mg/mL, B: 3.15 mg/mL, C: 2.85 mg/mL, D: 2.7 mg/mL.

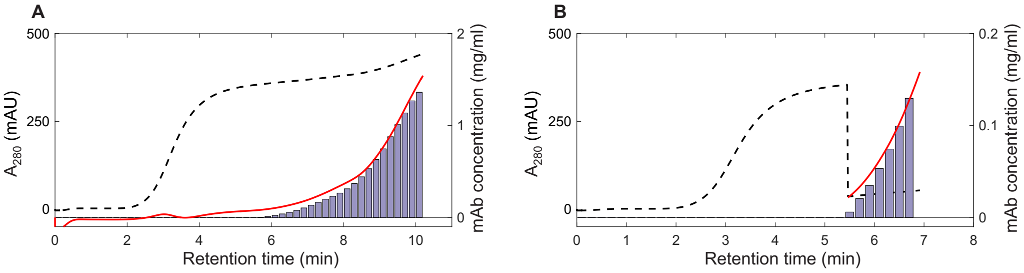

Figure 4.3: Results of the model evaluation by performing a real-time control of the load phase using a mAb titer of 3 mg/mL in the feed. The PLS model prediction (red lines) is compared with the results of the off-line analytics (blue bars) as well as the A_{280} (measured at a pathlength of 0.4 mm and displayed as dashed black line). The load phase was automatically terminated, when a mAb concentration in the effluent of A: 1.5 mg/mL or B: 0.15 mg/mL was reached. The sudden decrease in the A_{280} arises from the background subtraction.

of this second run are displayed in Figure 4.3 B. As an impurity background was subtracted to increase the sensitivity of the method, the A_{280} suddenly decreases. The second PLS model reached an RMSE for prediction of 0.01 mg/mL.

During both runs, the respective load phases were successfully terminated close to the intended breakpoints. In Table 4.1, a summary of intended and measured mAb concentrations in the last fraction of both confirmation runs is shown. The Matlab script sent a digital signal to Unicorn and terminated the load phase, when the targeted breakthrough concentration was reached. As the targeted breakthrough set points were concentrations at discrete time points, they are expected to be slightly higher than the concentrations of the last fraction determined by off-line analytics. This was observed for both confirmation runs (cf. Table 4.1). For an easier comparison between model and off-line analytics, a concentration based on an averaged absorption spectrum was calculated for the last fractions of both runs and compared with the corresponding off-line analytics. For the first run, the deviation between prediction and reference was 8.0 %, while for the second run a deviation of 2.3 % was found. This demonstrates that the described method can be successfully used to control the load phase in a Protein A capture step.

Table 4.1: Results of both confirmation runs: The targeted concentration to terminate loading is compared with the mAb concentration in the last fraction determined by off-line analytics. In addition, a PLS model prediction for the last fraction based on an averaged absorption spectrum is shown for comparison.

c_{target} [mg/mL]	$c_{\text{analytics}}$ [mg/mL]	$c_{\text{mean,PLS}}$ [mg/mL]
1.5	1.36	1.469
0.15	0.129	0.126

4.4 Conclusion and Outlook

A real-time monitoring and control of the load phase in a Protein A capture step was successfully realized in this study. It was demonstrated that PLS modelling on UV/Vis absorption spectra can be applied to quantify mAb in the effluent during the load phase despite of the background of many protein and non protein-based impurities. Based on the quantification, the load phase was automatically terminated, when a product breakthrough concentration of 1.5 mg/mL or 0.15 mg/mL was reached. Consequently, the proposed method has potential for the monitoring and control of capture steps at large scale production. In batch chromatography, the loading volume may be defined dynamically to allow for increased resin capacity utilization while still keeping the product loss small. Additionally, time-consuming off-line determination of the mAb titer in HCCF could be eliminated. The method may also be interesting for controlling column switching times in continuous chromatographic capture steps. Future challenges are especially related to the scale up and robustness of the method. Regarding the latter, especially upstream variations should be calibrated into the PLS model. Research will now focus on the migration of the method to the control of continuous capture steps.

Acknowledgment

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 635557. We kindly thank Lek Pharmaceuticals d.d. (Mengeš, Slovenia) for providing the HCCF and mock. This article is related to a patent application filed at the Euorpean Patent Office under the number EP15196599.3.

5

A multi-sensor approach for improved protein A load phase monitoring by conductivity-based background subtraction of UV spectra

Laura Rolinger[1], Matthias Rüdt[1], Jürgen Hubbuch[1]

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

Abstract

Real-time monitoring and control of protein A capture steps by PATs promises significant economic benefits due to the improved usage of the column's binding capacity, by eliminating time-consuming off-line analytics and costly resin lifetime studies, and enabling continuous production. The proposed PAT method in this study relies on UV spectroscopy with a dynamic background subtraction based on the leveling out of the conductivity signal. This point in time can be used to collect a reference spectrum for removing the majority of spectral contributions by process-related contaminants. The removal of the background spectrum facilitates chemometric model build-

ing and model accuracy. To demonstrate the benefits of this method, five different feedstocks from our industry partner were used to mix the load material for a case study. To our knowledge, such a large design space, which covers possible variations in upstream condition besides the product concentration, has not been disclosed yet. By applying the conductivity-based background subtraction, the RMSEP of the PLS model improved from 0.2080 g/L to 0.0131 g/L. Finally, the potential of the background subtraction method was further evaluated for single wavelength-based predictions to facilitate implementation in production processes. A RMSEP of 0.0890 g/L with univariate linear regression was achieved, showing that by subtraction of the background better prediction accuracy is achieved then without subtraction and a PLS model. In summary, the developed background subtraction method is versatile, enables accurate prediction results and is easily implemented into existing chromatography setups with typically already integrated sensors.

5.1 Introduction

The profitability of biopharmaceutical companies is decreasing [255] due to decreasing Research and Development (MCSGP) productivity and increased drug price competition from biosimilars [102]. Therefore, the sector is looking to reduce costs in MCSGP and production by automation of the production processes [104, 256]. The implementation of PAT is key for the digital transformation and automation of processes in order to gain a competitive edge over business rivals. As automation in the downstream process is economically most valuable for Protein A capture steps due to the high costs of protein A resin, this area has received a lot of attention [169], especially in the past year [168, 257, 258]. Rüdt et al. published an approach in 2017, where UV/Vis spectra were used to monitor the breakthrough of a protein A column and to control the load phase, if a certain concentration in the breakthrough was reached [169]. While the approach itself is interesting, little explanation was given in the article on the used PLS model and what spectral changes it leverages. Additionally, a background subtraction at a constant UV signal was necessary to improve the prediction for low concentrations as the change in HCP in different feeds influenced the model. This background subtraction at constant absorption is difficult, as a displacement of HCP species or highly concentrated feed stock can lead to insufficient fulfillment of UV criteria and thereby to the failure of the method.

Feidl at al. [168, 257] published an approach to monitor the breakthrough with Raman spectroscopy. Due to the low scatter efficiency of proteins, measurement times of 30 s per spectra were necessary [168, 257] and with an average of two spectra [168], resulting in a measurement time of 1 min. Measurement times of 1 min can be insufficient for process control, especially when looking at protein A membranes with high flow rates and short load times. Even though measurement times per spectra were quite high compared to UV/Vis, additional extensive data analysis was necessary to remove high noise and make accurate predictions possible.

A limitation of current publications is furthermore the comparably small change in HCCF composition due to the usage of only one or two feed stocks in each study. Rüdt et al. used HCCF and mixed it with mock from a different cultivation [169]. Feidl et al. used HCCF from a perfusion reactor with two different mAb concentration. Thakur et al. prepared flow-through and purified mAb from one batch of HCCF for a NIR-based control for continuous chromatography. In all three studies, the calibration space was thus spanned by only one or two HCCF batches. Since inter-batch variations can result in a significant impact on HCP composition and DNA content [259], the obtained models may be limited in their predictive power for an independent HCCF batch.

In order to tackle sensor complexity and model validity over upstream fluctuations in this study, a product containing HCCF was mixed with three different mock materials and purified bispecific mAb. This accounts for various changes in the cell line, cell culture medium, host cell profile and also for changes in the bispecific product profile due to the changes in the concentration of mispaired species relative to the product. Due to the increased and random variability compared to previous studies, a prediction of the mAb concentration in the breakthrough becomes more challenging. To compensate the increased variability in the background, a novel background subtraction method was developed in this study. Specifically, a background spectrum is subtracted when the conductivity reaches a stable point. This allows to determine the breakthrough of the flow-through as the protein concentration contributes very little to the overall conductivity of the HCCF. Finally, the usage of single wavelength absorption in combination with the conductivity-based background subtraction for product concentration prediction in the effluent is evaluated. The use of only one absorption wavelength and conductivity allows for an easy implementation of load control strategies in current manufacturing processes as those sensors are typically implemented in chromatographic equipment.

5.2 Materials and Methods

5.2.1 Biologic Material and Buffers

All biologic material was stored at 5°C before experimentation after delivery from our industry partner. In order to obtain a variable mAb concentration —in this study a bispecific mAb—, a variable mispaired species to product ratio, and a variable impurity profile in the load material, the product containing HCCF (Feedstock 1) with a product concentration of 2 g/L was mixed with purified product (Feedstock 2) and three different mock HCCFs solutions (Feedstock 3-5). One mock solution was cultivated with a non-producing cell line. The other two mock solutions were prepared as flow-through by preparative protein A chromatography. These two mock solutions were derived from HCCFs of two different cell lines, which produce two different mAbs, respectively. Prior to this study, it was ensured that the protein A flow-through did not contain antibodies in detectable concentrations (based on analytical protein A chromatography). For product spiking, the used bispecific mAb (Feedstock 2) was purified to the second polishing step by our industry partner and was concentrated up to 20 g/L to reduce dilution effects of the impurities by addition of the concentrated product.

In the product containing HCCF (Feedstock 1), different mispaired species were present, while the purified product (Feedstock 2) only contained the desired mAb. By mixing the product containing HCCF with the purified product, variation in the concentration of the different mAb species was introduced into the design space as well.

The product containing HCCF, purified mAb and the three mock HCCFs were filtered with a cellulose acetate filter with a pore size of 0.22 µm (Pall, Port Washington, NY, USA) before mixing. In Table 5.1, the used volumina of the different stock material for each run are shown. The composition of the mixtures between the three mock materials was determined by Latin Hypercube Sampling to provide a random multidimensional distribution.

For all preparative runs, the following buffers were applied: Equilibration with 25 mM Tris(hydroxymethyl)aminomethane (TRIS) and 0.1 M sodium chloride at pH 7.4, wash with 1 M TRIS and 0.5 M potassium chloride at pH 7.4, elution with 20 mM citric acid at pH 3.6, sanitization with 50 mM sodium hydroxide and 1 M sodium chloride, and storage with 10 mM sodium phosphate, 130 mM sodium chloride, 20 % ethanol.

For analytical protein A chromatography, column equilibration was carried out using a buffer with 10 mM phosphate (from sodium phosphate and potassium phosphate) with 0.65 M chloride ions (from sodium chloride and potassium chloride) at pH 7.1. Elution was performed with the same buffer, but titrated to pH 2.6 with hydrochloric acid. All buffer components were purchased from VWR, West Chester, USA. The buffers were prepared with Ultrapure Water (PURELAB Ultra, ELGA LabWater, Viola Water Technologies, Saint-Maurice, France), filtrated with a cellulose acetate filter with a pore size of 0.22 µm (Pall), and degassed by sonification.

Table 5.1: Sample composition for the calibration runs 1 to 4 and the validation run 5 with volumes of the product containing HCCF (Feedstock 1), purified mAb (Feedstock 2), mock HCCF (Feedstock 3), flow-through 1 (flow-t.1), and flow-through 2 (flow-t.2) (Feedstock 4 and 5).

Run number	data usage -	HCCF in mL	mAb in mL	flow-t.1 in mL	flow-t.2 in mL	mock HCCF in mL
Run 1	calibration	52.50	0.00	9.85	21.91	20.74
Run 2	calibration	35.00	1.75	14.82	1.36	17.08
Run 3	calibration	21.00	3.15	6.48	3.93	7.44
Run 4	calibration	17.50	3.50	6.57	6.13	1.30
Run 5	validation	26.25	2.63	2.01	12.65	8.96

5.2.2 Chromatographic Instrumentation

All preparative runs were realized with an Äkta Pure 25 purification system controlled with Unicorn 6.4.1 (GE Healthcare, Chicago, USA). The system was equipped with a sample pump S9, a fraction collector F9-C, a column valve kit (V9-C, for up to 5 columns), a UV-monitor U9-M (2 mm pathlength), a conductivity monitor C9, a pH valve kit (V9-pH) and an I/O-box E9. Additionally, an UltiMate 3000 DAD equipped with a semi-preparative flow cell (0.4 mm optical pathlength) and operated with Chromeleon 6.8 (Thermo Fisher Scientific, Waltham, USA) was connected to the Äkta Pure. The DAD was positioned between the conductivity monitor and the V9-pH valve. Additionally, a second sensor and flow cell were positioned before the DAD. The data was not used for this study.

Reference analysis of collected fractions was performed using a Vanquish Flex Binary HPLC system (Thermo Fisher Scientific, Wilminton, US) by analytical protein A chromatography. The system consisted of a Binary Pump F, Split Sampler FT, Column Compartment H and a Diode Array Detector HL. Chromeleon Version 7.2 SR4 (Thermo Fisher Scientific) was used to control the HPLC.

5.2.3 Chromatography Runs

In order to generate variable mixtures between the product bispecific mAb, mispaired species and, other impurities for the PLS model calibration and validation, breakthrough experiments with variable mAb titers in the feed were performed. The mAb titers in the different load materials were 1, 1.5, 2, 2.5, and 3 g/L. For each experiment, a prepacked 5×50 mm, MabSelect SuRe column (0.982 mL) (Repligen, Waltham, US) was first equilibrated for 5 Column Volume (CV) and then loaded with 100 mg of mAb. At the beginning of the load phase, the DAD equipped with a semi-preparative flow cell (optical pathlength 0.4 mm) was triggered to record absorption spectra between 200 nm to 800 nm and the column flow-through was collected in 200 µL fractions.

5.2.4 Analytical Chromatography

The collected fractions of all runs were examined by analytical protein A chromatography to obtain the mAb concentrations. For each sample, a 2.1×30 mm POROS prepacked protein A column (Applied Biosystems, Foster City, USA) was equilibrated with 2 CV of equilibration buffer, flowed by an injection of 20 µL of sample. The column was then equilibrated with 0.8 CV of equilibration buffer and eluted with 1.4 CV of elution buffer. The flow rate was 2 mL/min for all phases and experiments.

5.2.5 Data Analysis

The data analysis workflow is depicted in Figure 5.1. The recorded 3D-field, results from the analytical chromatography, and run data from the Äkta system were read in and pre-processed with MATLAB 2019R (The MathWorks, Inc., Natick, USA). From the conductivity data, the stable point of the conductivity was determined by smoothing the data with a moving mean filter with a window size of 5 s. If the conductivity did not change in the third decimal point for 10 s after the first CV, the conductivity was seen as stable. This point was used to subtract the background spectrum from the UV spectra, as depicted in Figure 5.2. The goal of this background subtraction is to remove signal originating from contaminants from the spectrum to improve product concentration predictions.

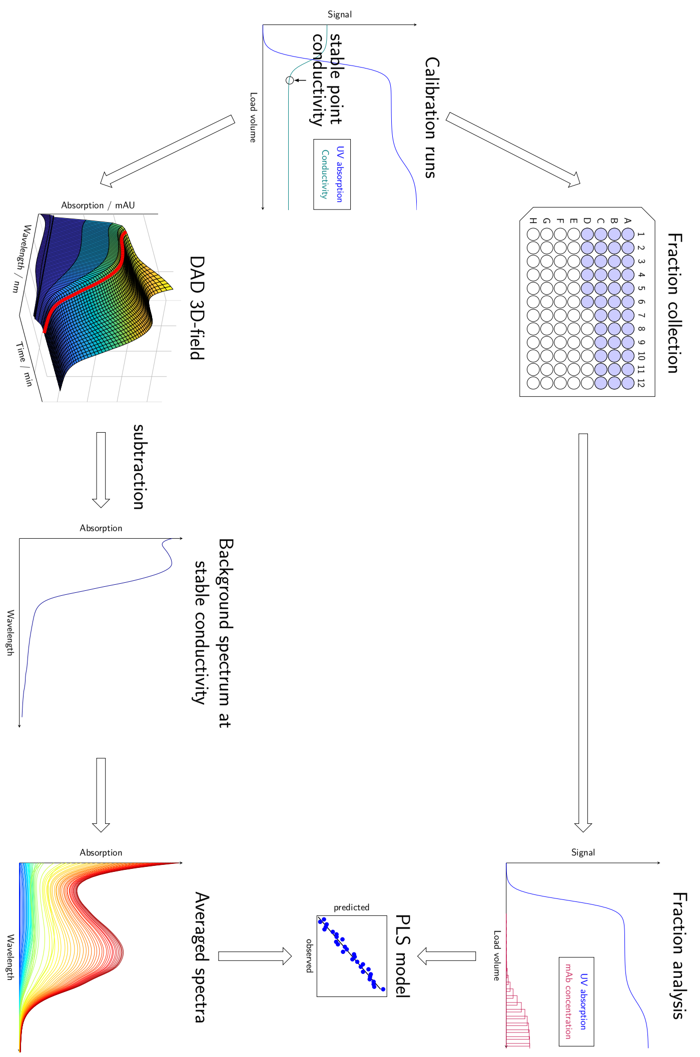

Figure 5.1: Experimental procedure for the PLS model calibration with background correction: For each calibration run, 200 µL fractions were collected and analyzed by analytical protein A chromatography to obtain the mAb breakthrough curves. During the breakthrough, 3D chromatograms and the conductivity were recorded. When the initial breakthrough of impurities was completed, determined by the stability of the conductivity signal, this background spectrum (highlighted red in 3D field) was subtracted from the 3D-field. Then the averaged spectra corresponding to the fraction size were calculated from the background corrected absorption 3D-field. Averaged spectra and mAb concentrations were correlated using PLS modeling.

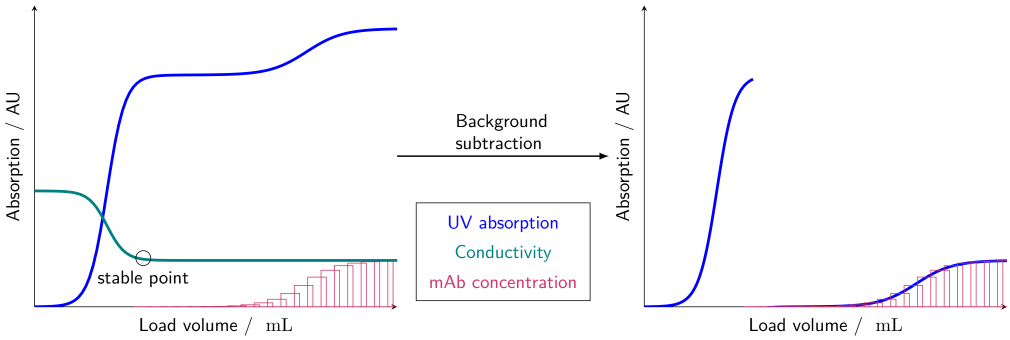

Figure 5.2: The goal of the background subtraction is to determine the complete breakthrough of the HCCF background by conductivity and to subtract the spectrum at complete background breakthrough. Through this most effects of the background are removed from the spectrum and estimation of the mAb concentration can be improved. Additionally, background effects in the HCCF due to changing conditions in the medium, HCP profile or DNA amount are excluded.

The background subtraction was performed by subtracting the measured UV spectrum closest to the stable point of the conductivity. The spectra were averaged according to the fraction size data from the Äkta. For the correlation of the averaged absorption spectra with the mAb concentrations, PLS models were calibrated using SIMCA 13.0.3 (Sartorius, Göttingen, Germany). SIMCA applies the NIPALS-algorithm for PLS model building [96]. Before the PLS model calibration, all spectra and the mAb concentration were pre-treated by mean-centering using SIMCA. For the calibration of the PLS model, Run 1, Run 2, Run 3 and Run 4 were used as calibration dataset. SIMCA applies a 7-fold cross validation as internal validation. The number of LV was determined by the autofit function of SIMCA. Run 5 was chosen as external validation.
The model complexity, in this case the number of LV, is important for the robustness of the model [96]. It is important to find the right compromise between fit and predictive ability of the model. While an increase in LVs increases the fit of the model, also noise in the data can be fitted, which reduces the prediction ability of the model for new data with unknown noise or other non-idealities [95].

5.3 Results and Discussion

In this study, the breakthrough of mAb during the protein A load phase was monitored by UV spectroscopy in combination with a PLS model. To calibrate the PLS model, four chromatographic runs (Run 1-4) at mAb concentrations of 1, 1.5, 2.5, and 3 g/L in the feed were performed and analyzed by off-line analytics. The actual concentration in the load material were slightly higher due to inaccuracies in the initial titer measurement of the HCCF and purified product. A validation run (Run 5) was performed at a mAb concentration of 2 g/L in the feed. Not only was the mAb concentration varied, but also the composition of mock mixture to dilute the HCCF. This was done to imitate possible variability in upstream processing, like changes in cell culture medium, different amounts of DNA through different harvest timepoints, and changes in the HCP profile. This variation generates a large design space for model application.

Figure 5.3 compares the absorption at 280 nm A_{280} recorded by the DAD to the conductivity recorded by the Äkta. The stability criterium of the conductivity is reached between 6.6 mL to 10.6 mL, depending on the remaining buffer volume in the sample pump due to incomplete purging. It can be seen, that while the conductivity is stable after this point, the absorption at 280 nm is still increasing due to the displacement of impurities. It has been shown, that DNA and certain HCP species interact with the mAb bound to the Protein A resin [254, 260–262]. This interaction can lead to a retention effect of the interacting impurities in comparison to non-interacting impurities, which could lead to a delayed breakthrough of the interacting impurities. The difference in interaction strength between the impurities and the bound mAb could also lead to a displacement of weakly interacting contaminants by stronger interacting HCP species with progression of the load. The increase in absorption due to the displacement, while no mAb breakthrough occurs, varies between runs as the impurity profile varies. Therefore, the conductivity-based criterium is more robust for the background subtraction than a UV-based criterium.

5.3.1 PLS Model Calibration and Validation

The results of the model calibration without background subtraction are depicted in Figure 5.4. It compares the absorption at 280 nm A_{280} to the concentrations measured by off-line analytics and the prediction calculated by the calibrated PLS model. It can be seen, that from the A_{280} alone, it is not possible to determine the breakthrough of mAb, because no clear plateau is visible. Likely HCPs are displaced during loading which is overlaying

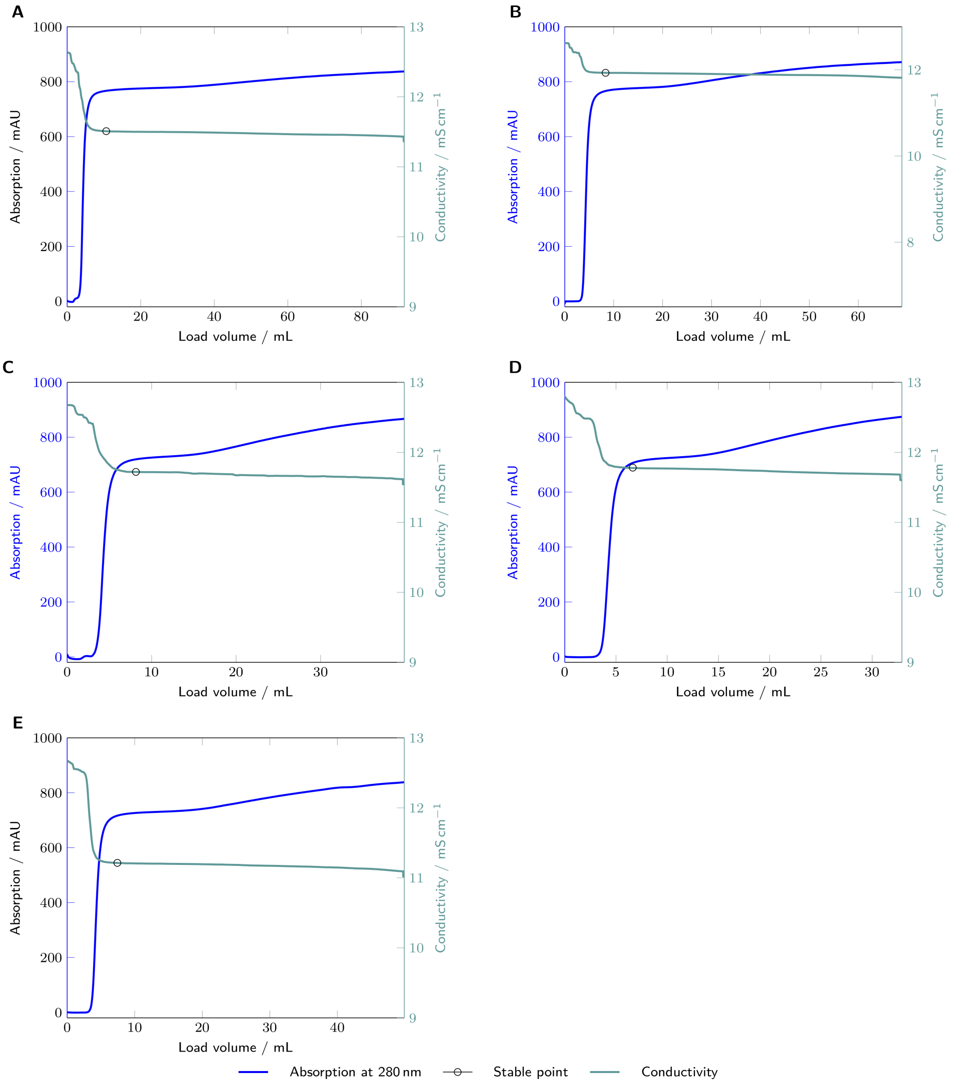

Figure 5.3: The absorption at 280 nm A_{280} recorded by the DAD (displayed as blue line) is compared with the conductivity recorded by the Äkta (teal line). The calculated stable point of the conductivity is indicated as black circle. All five runs exhibited variable mAb titers in the feed A: 1 g/L, B: 1.5 g/L, C: 2 g/L, D: 2.5 g/L and E: 3 g/L.

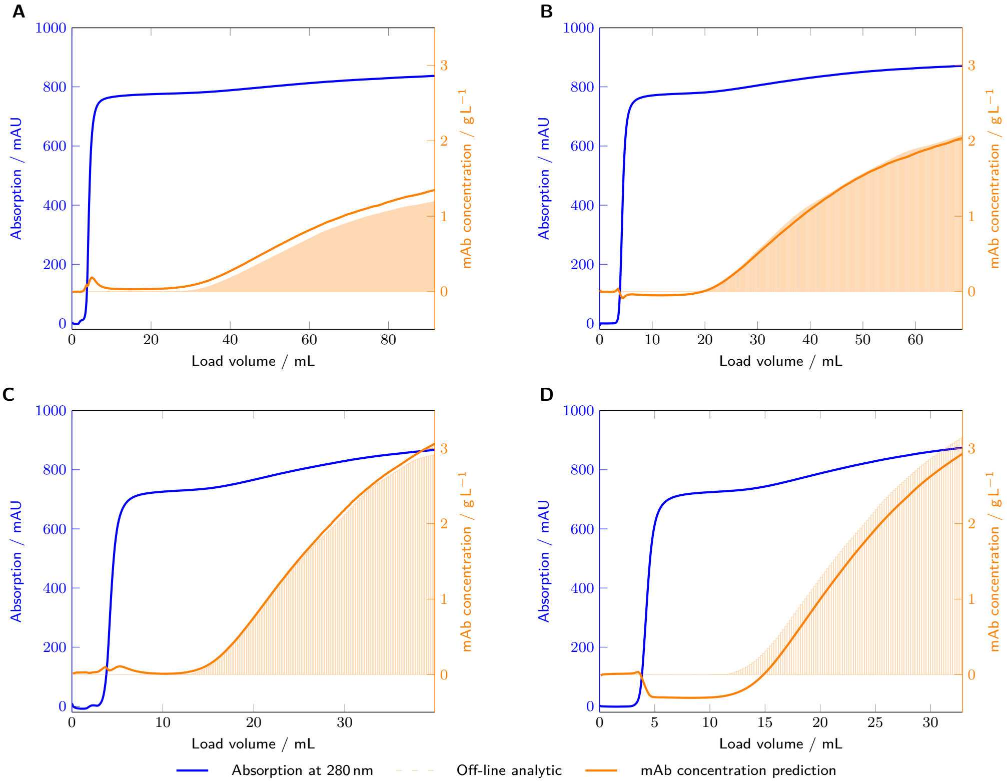

Figure 5.4: Results of the PLS model calibration without background subtraction. The absorption at 280 nm A_{280} recorded by the DAD (displayed as blue line) is compared with the results of the off-line analytics for mAb quantification (orange bars). The PLS model prediction is illustrated as orange lines. The four runs (Run 1-4) exhibited variable mAb titers in the feed A: 1 g/L, B: 1.5 g/L, C: 2.5 g/L, and D: 3 g/L.

with the breakthrough of the mAb [50, 254]. The data show, that with decreasing mAb concentration and increased background variation, the offset between the model prediction and actual concentration at low concentrations is increasing. In Table 5.2, the coefficient of determination R^2, the cross-validated coefficient of determination Q^2, the RMSECV and number of LVs are compared for the model with background subtraction and without. In general, the model with background subtraction has a higher R^2 and Q^2 with 0.999 compared to 0.980, respectively, for the model without background subtraction.

The results of the model calibration with background subtraction are depicted in Figure 5.5. The PLS model fits the actual concentration pro-

Table 5.2: R^2, Q^2, RMSECV, RMSEP and number of LVs for both PLS models.

Background subtraction	R^2	Q^2	RMSECV in g/L	RMSEP in g/L	number of LVs
No	0.980	0.980	0.1170	0.2080	3
Yes	0.999	0.999	0.0246	0.0131	2

file over the complete concentration range better than the model without subtraction. The corrected absorption at 280 nm $A_{280\mathrm{corr}}$ does not plateau during the load, showing that conductivity-based background subtraction is better suitable than a UV-based criterium. To further visualize the accuracy of both methods, an observed versus predicted mAb concentration plot is discussed in the Appendix 5.5.
Additionally, the PLS model with background subtraction has two LVs compared to three LVs of the PLS model without background subtraction. In general, models with less LVs are preferred as the chances of overfitting are smaller and therefore the robustness of the model can be better.

Figure 5.6 compares the spectra of the uncorrected and corrected data. Apparently, the background contributes most to the UV spectrum of the load material. Typically, the background contributes between 710 mAU to 768 mAU at 280 nm to the overall absorption, while the mAb and the displaced proteins contribute between 0 mAU to 162 mAU. This indicates, that other UV active components in the load material besides the product are the main contributors to the spectrum. The local spectral maxima shift for the not background subtracted spectra from 271 nm towards 272 nm or 273 nm, depending on the background spectrum. Probably different DNA concentrations in the load material cause difference in the local maximum at same concentration in different runs. DNA has a local absorption maximum around 260 nm, which could cause the spectrum of the load material to lay between 260 nm and 280 nm, depending on the DNA concentration. The varied concentration of UV-active components in the different load materials is probably the reason for the shift and the different total absorption values of the spectra without background subtraction by mAb concentration in Figure 5.6. As the mAb concentration increases the local maximum shifts in the direction of 280 nm, which is typically considered as the local maximum of proteins [263].

The local maxima in the background corrected spectra (Figure 5.6B) remain constant at 279 nm as soon as the mAb concentration starts to

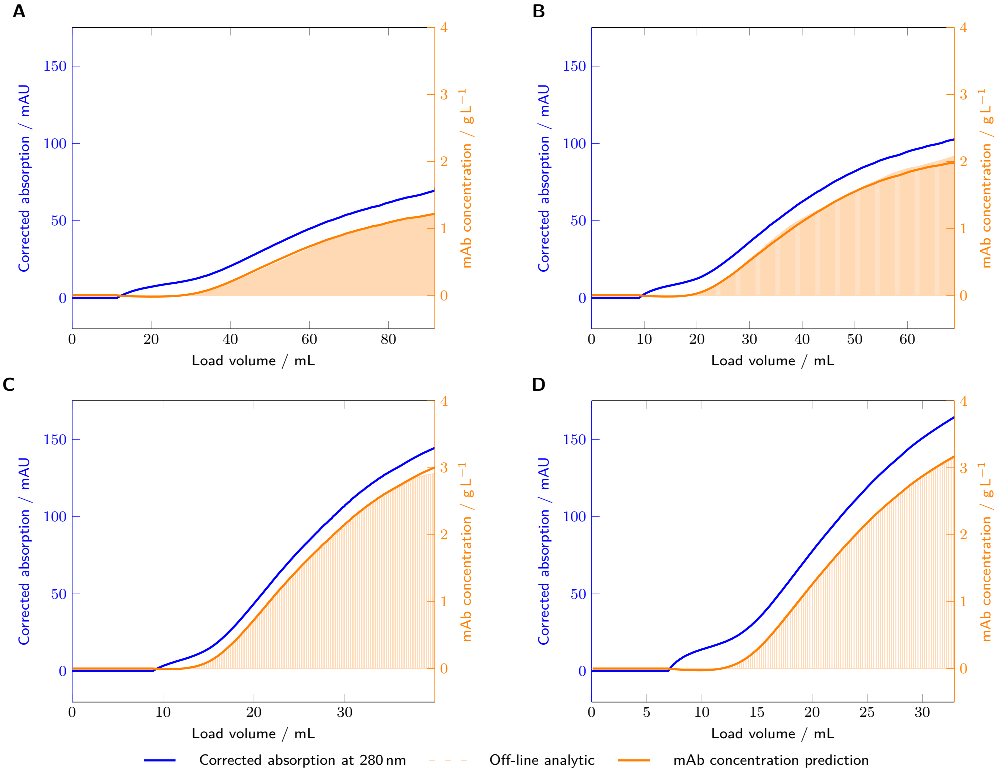

Figure 5.5: Results of the PLS model calibration with background subtraction. The absorption at 280 nm A_{280} recorded by the DAD (displayed as blue line) is compared with the results of the off-line analytics for mAb quantification (orange bars). The PLS model prediction is illustrated as orange line. The four runs (Run 1-4) exhibited variable mAb titers in the feed A: 1 g/L, B: 1.5 g/L, C: 2.5 g/L, D: 3 g/L.

increase. While the mAb concentration is still 0 g/L, the overall absorption still increases over time, maybe due to a baseline drift by the DAD. Additionally a small contribution of the background is visible. This may be caused by the displaced of impurities from the column due to the binding of the mAb. Both phenomena could explain, why the absorption does not stay at 0 mAU for all wavelengths, while no mAb is breaking through the column. To compare the background corrected spectra with the absorption spectrum of the product, the absorption spectrum during the elution of Run 1 is plotted in black in Figure 5.6B. The absorption spectrum during the elution of Run 1 was normalized to maximum absorption in Figure 5.6 and shifted up by 5 mAU to enhance readability. The elution spectrum has its local maximum at 279 nm like the background corrected spectra. The

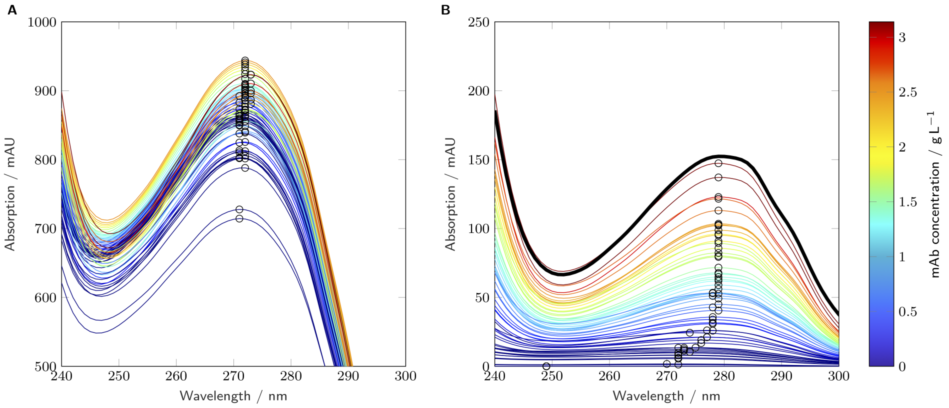

Figure 5.6: Comparison of the averaged spectra before (Figure 5A) and after background subtraction (Figure 5B). Every twentieth spectrum is plotted. The spectra are colored accordingly to the mAb concentration in the spectra from low concentration (blue) to high concentration (red). The local maxima of the spectra are highlighted with black circles. It is shown, that without the background subtraction, the positions of the local maximum shift with higher mAb concentration closer from 264 nm towards 270 nm. For the background subtracted spectra, the position of the local maxima stay consistent at 279 nm after the initial breakthrough. Highlighted in black is a normalized absorption spectrum during the elution of Run 1.

absorption in the elution spectrum around the local minimum at 252 nm is lower compared to the background corrected spectra. This could be caused by impurities contributing to the background corrected spectra. It seems more challenging for the PLS model to extract the mAb concentration from the spectra with the random variation in the background, because the PLS model without background subtraction needs more LVs to fitted the data. The spectra with background subtraction are ordered according to mAb concentration and the local maximum stays at 279 nm, indicating that the spectrum originates from a proteinous source.

Additionally, from Figure 5.5 it seems, that the product concentration does not follow the absorption at 280 nm entirely, because the difference between the absorption and product concentration grows bigger with increase in product concentration. The higher the mAb concentration in the breakthrough the more HCPs seem to be displaced from the column as the column saturates. The results of the model validation without background subtraction and with background subtraction are depicted in Figure 5.7. For the prediction without background subtraction, an offset between the actual mAb concentration at low protein concentration persists as in the

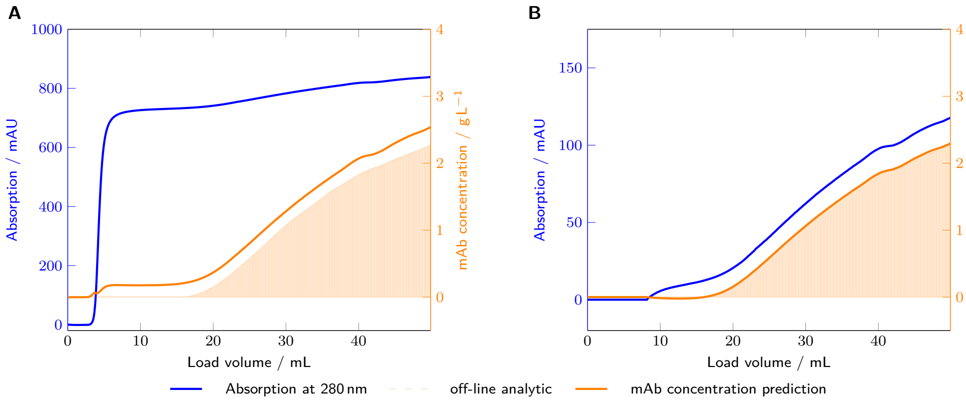

Figure 5.7: Results of the PLS model validation (Run 5). The absorption at 280 nm A_{280} recorded by the DAD and displayed as blue line) is compared with the results of the off-line analytics for mAb quantification (orange bars). The PLS model prediction is illustrated as orange lines. Figure 7A shows the model prediction without background subtraction and Figure 7B shows the model prediction with background subtraction at a feed concentration of 2 g/L.

calibration data. Again, the model with background subtraction fits the actual breakthrough at low concentration better. This is also represented in the RMSEP of 0.0131 g/L of the model with background subtraction compared with the RMSEP of 0.2080 g/L of the model without background subtraction (see Table 5.2).

Additionally, we provide the limit of detection (LOD) and the limit of quantitation (LOQ) of both models with and without background subtraction in the Appendix 5.7.

5.3.2 Comparison to Other Publications

To set the results of this study into perspective to recent publications, the results are compared to the obtained results by Thakur et al. [258] for the usage of NIR spectroscopy to monitor the breakthrough and to the results by Feidl et al. [168] for the usage of Raman spectroscopy. As these studies were carried out on different data set and different steps for model optimization were undertaken, a final conclusion cannot be drawn by solemnly comparison of the results. However, a comparison can give a general steer on which method might be the most suitable for the monitoring of the Protein A load phase.

Thakur et al. [258] published a RMSEP for the breakthrough experiments in their publication of 0.1540 g/L for NIR spectroscopy in combination with PLS models. This error is almost 10-fold higher than the RMSEP of 0.0186 g/L for the model with background subtraction from this article. As it is sometimes misleading to compare RMSEPs due to difference in involved sample concentration and sample distribution in the design space, it would be better to as well compare the goodness of fit R^2 and the goodness of prediction during cross-validation Q^2 values. However, the R^2 mentioned in the paper must not be mistaken for the R^2, as the coefficient of determination describes only the goodness of fit of the regression line on the observed versus predicted plot and not on the actual prediction. Another point to consider is, that no orthogonal off-line analytic was performed by Thakur et al. It remains unclear, how the actual mAb concentration was calculated, if it was not measured. Also, the chemometric side of the data analysis is not explained. This makes it difficult to evaluate, whether an effect caused by the actual difference in mAb concentration was measured or an effect due to the mixture of the feedstocks is correlated with the mAb concentration. Additionally, a more challenging design space presented in this study, as five different feedstocks were used compared to one in the case study by Thakur et al.

Therefore, it is difficult to compare the NIR-based model with UV-based model. With the presented evidence, however, we would conclude that UV-based models seem to have a lower prediction error compared to NIR-based methods. This is also in good agreement with literature, which generally concludes, that UV absorption spectroscopy has a higher accuracy due to the low impact of temperature and water background on the spectra [125, 138, 264]. The same can be said about Raman spectroscopy, which is also reported to a lower accuracy and higher limit of detection in comparison to UV absorption spectroscopy for proteins [125, 138, 264]. An average RMSEP of 0.12 g/L was published by Feidl et al. [168] for the breakthrough monitoring with Raman spectroscopy and PLS modelling in a concentration range, which is comparable to this study. This is again an almost 10-fold higher RMSEP as for the model with background subtraction presented in this study. Also extensive chemometric model optimization was used to achieve this RMSEP, whereas in this study no optimization was undertaken, because it was not necessary. In a next study, Feidl. et al. [257] investigated the usage of a lumped kinetic model and an extended Kalman filter to improve the PLS model prediction for low mAb concentrations. In the lower concentration range between 0 g/L to 0.42 g/L, a RMSEP of 0.055 g/L was achieved. The implementation of an extended Kalman filter improved the RMSEP to 0.026 g/L. Even a RMSEP of 0.026 g/L is still almost double as high as the best RMSEP of this study. Although the use of an extended

Kalman leads to a prediction improvement at first, the underlying model can change during the lifetime a Protein A column due to column fouling, which could make the predictions worse in the long run. Additionally, the Raman measurements were quite slow with a total measurement time of 1 min [168] in comparison to NIR or UV measurements, which can be carried out in less than a second. RMSEP of 0.12 g/L [168] obtained with Raman spectroscopy and a RMSEP of 0.026 g/L [257] of Raman spectroscopy with an extended Kalman filter and extensive chemometric processing, the RMSEP in this study is still lower even though the concentration range was 10 times larger. It seems, that in general the prediction obtained by Raman spectroscopy is more corrupted by measurement noise and the use of a signal filter is obligatory to derive a more reliable prediction compared to the raw prediction.

5.3.3 Application of Single Wavelength UV-measurements

The implementation of a DAD is not standard in most production processes. Therefore, the use of the absorption only at 280 nm was tested, with and without background subtraction. In Table 5.3 the R^2, the Q^2, the RMSECV and RMSEP are compared for the model with background subtraction and without. Without background subtraction, the model cannot fit the breakthrough of mAb. The R^2 and Q^2 are with 0.172 too low for spectroscopic models [96] and the RMSEP is with 0.7348 g/L too high for an effective control of a protein A load phase. With background subtraction, R^2 and Q^2 of 0.985 and an RMSEP of 0.0890 g/L are achieved. This shows, that the background subtraction eliminates most effects not caused by the increase in mAb concentration. In Appendix 5.8, a further discussion visualization of prediction capability of the single wavelength approach is given.The simple linearregression on a single wavelength with background subtraction allows the implementation in production processes with already available process sensor, i.e. conductivity and absorption at 280 nm. No advanced chemometric methods are necessary. Instead, the approach works almost out-of-the-box. As the accuracy of the sensors are crucial for the application, low-noise sensors are required in the process.

5.4 Conclusion and Outlook

In this study, a multi-sensor approach for real-time monitoring of the load phase in a protein A capture step was presented and compared to other

Table 5.3: R^2, Q^2, RMSECV, and RMSEP for both PLS models with only 279 nm as input.

Background subtraction	R^2 -	Q^2 -	RMSECV in g/L	RMSEP in g/L
No	0.172	0.172	0.7574	0.7348
Yes	0.985	0.985	0.101	0.0890

published approaches. The proposed method relies on a dynamic UV background subtraction based on the leveling out of the conductivity signal. The background corrected spectra can be used for product breakthrough predictions in combination with a PLS model or by single wavelength regression. In this study, a large design space with possible variations arising during fermentation was created by using five different feedstocks to mix the load material for the protein A step. The mixtures accounts for possible changes in contaminant profile and concentration, like buffer components, DNA, HCP and mispaired species of the bispecific mAb. It was demonstrated that by subtracting the background spectrum during the breakthrough, the prediction of the mAb concentration is facilitated and improved compared to models using the raw spectra. The proposed method offers a robust quantification of the product breakthrough regardless of large variability in the cell culture fluid.

We conclude that UV-based methods, especially with background subtraction, yield better prediction accuracies than NIR- or Raman-based methods judged by the RMSEPs published in other publications [168, 257, 258]. The application of the background subtraction to product concentration determination with only one absorption wavelength shows great potential for the application to production processes as the required sensors are already implemented in most processes.

Conflict of interest

The authors declare no conflict of interest.

5.5 Appendix: Linearity over Concentration Range for the PLS models

Figure 5.8 shows the predicted mAb concentration over the observed/measured mAb concentration for the PLS models with and without background subtraction. Predicted and observed mAb concentrations show for both models a linear relationship. Deviations from the linear relationship could be possibly caused by the off-line analytic due to carry-over between samples. However the not background corrected models shows different offsets depending on the individual run. The largest offset can be observed for Run 1. These offset could originate for the differences in load material. The PLS model with background subtraction shows no significant offsets, which seems to be a results of the removal of different spectral contributions from the different feed stock material.

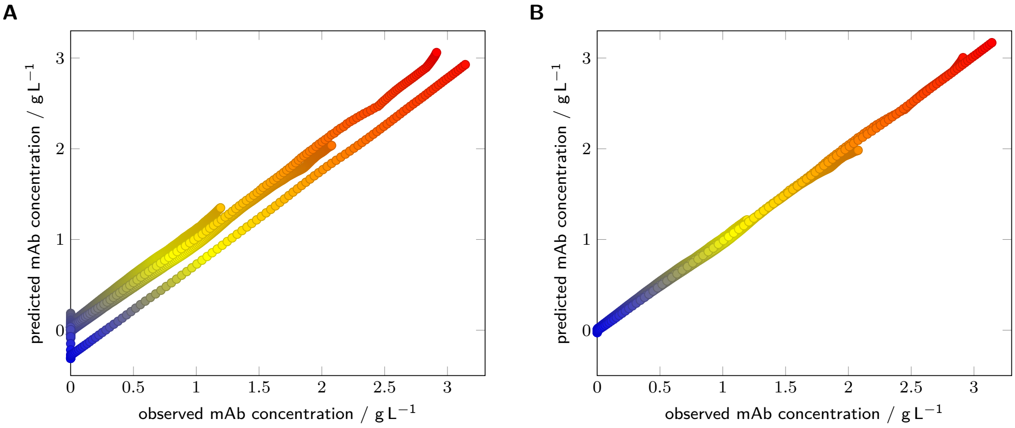

Figure 5.8: Predicted mAb concentration by the PLS model over the measured (observed) mAb concentration for (A) the not background corrected PLS model and (B) the background corrected PLS model.

5.6 Appendix: Background Composition

All feedstocks used in this study could be differentiated by the color of the HCCF. Figure 5.9 shows the different background spectra, which were subtracted. As contaminants, like DNA, HCP, some buffer componants and scattering molecules contribute to this background spectrum, the diversity in the feed stock can be spectrally assessed. Interestingly, the background spectra cluster into 2 groups. Even though Run 1 and Run 2 have a very different composition, the background spectra look similar with Run 2 possibly having a higher DNA concentration due to the increased absorption at 260 nm, but not at 280 nm. Also Run 3, Run 4 and Run 5 show similar background spectra with regards to the total amount of absorption, but also differ in the composition possibly due to different DNA, HCP and amount of large molecules, which cause light scattering.

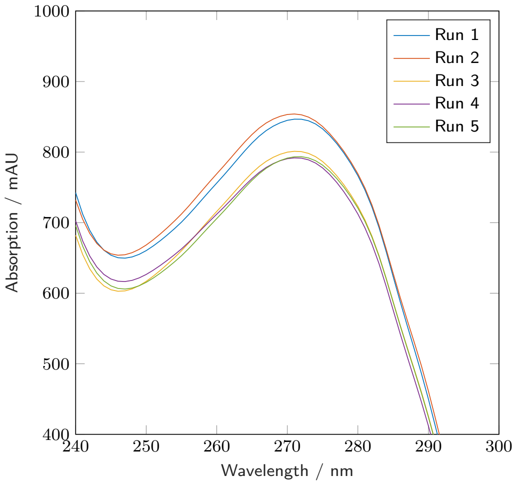

Figure 5.9: Comparison of the background spectra for the calibration runs (Run 1-4) and the validation run (Run 5).

5.7 Appendix: Limit of Detection

The LOD and LOQ interval for the dataset was calculated based on the MATLAB code provided by Allegrini [265]. The results are displayed in Table 5.4. If the background subtraction is done, both the LOD and LOQ interval are lower in comparison to without background subtraction. Additionally are the intervals itself smaller with background subtraction. The reduced spectral contribution of interfering components due to the background subtraction could explain these findings, allowing for better detection and quantification.

Table 5.4: LOD interval and LOQ interval for multivariate models with and without background subtraction.

Background subtraction	LOD interval in g/L	LOQ interval in g/L
Yes	0.0130-0.0144	0.039-0.043
No	0.0752-0.0940	0.226-0.282

5.8 Appendix: Single wavelength prediction

Figure 5.10 shows the predicted mAb concentration over the observed/measured mAb concentration for the linear regression models with and without background subtraction at 279 nm. The regression without background subtraction shows large offsets for the different runs, which seem to be driven by the contribution of the background spectra (see Figure 5.9). The regression model with background substraction shows little offsets. Only Run 4 seems to have a larger offset compared to the other runs, which could be explained by a comparable little earlier subtraction of the background as with the other runs. Interestingly, even though the offsets are minimized by the background subtraction, the predicted mAb concentration over the observed/measured mAb concentration show different slopes for the different runs. This could be caused by the different interacting species present in the load material, which are displaced at a different rate from the column between the different runs.

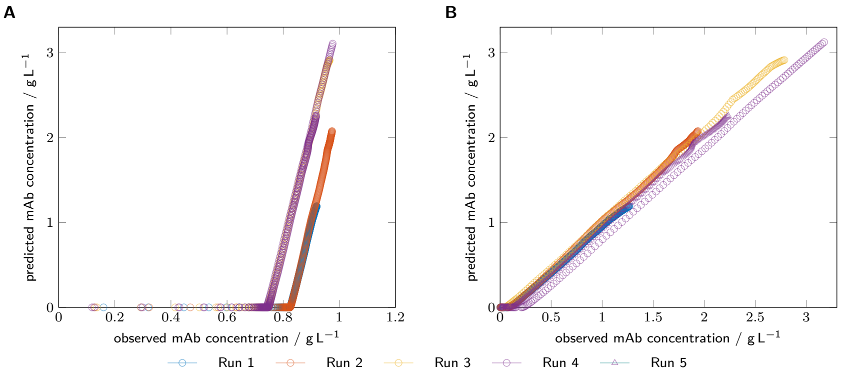

Figure 5.10: Predicted mAb concentration by the PLS model over the measured (observed) mAb concentration for (A) the not background corrected PLS model and (B) the background corrected linear regression model.

6

Comparison of UV- and Raman-based monitoring of the protein A load phase and evaluation of data fusion by PLS models and CNNs

Laura Rolinger[1], Matthias Rüdt[1], Jürgen Hubbuch[1]

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

Abstract

A promising application of PAT to the downstream process of mAbs is the monitoring of the Protein A load phase as its control promises economic benefits. Different spectroscopic techniques have been evaluated in literature with regard to the ability to quantify the mAb concentration in the column effluent. Raman and UV spectroscopy are among the most promising techniques. In this study, both were investigated in an in-line setup and directly compared. The data of each sensor were analyzed independently with PLS models and CNNs for regression. Furthermore, data fusion strategies were investigated by combining both sensors in hierarchical PLS models

or in CNNs. Among the tested options, UV spectroscopy alone allowed for the most precise and accurate prediction of the mAb concentration. A RMSEP of 0.013 g/L was reached with the UV-based PLS model. The Raman-based PLS model reached an RMSEP of 0.232 g/L. The different data fusion techniques did not improve the prediction accuracy above the prediction accuracy of the UV-based PLS model. Data fusion by PLS models seems meritless when combining a very accurate sensor with a less accurate signal. Furthermore, the application of CNNs for UV and Raman spectra did not yield significant improvements of the prediction quality. For the presented application, linear regression techniques seem to be better suited compared to advanced non-linear regression techniques, like CNNs. In summary, the results favor the application of UV spectroscopy and PLS modeling for future research and development activities aiming to implement spectroscopic real-time monitoring of the Protein A load phase.

6.1 Introduction

In biopharmaceutical downstream processing of mAbs, a focus of PAT research has been on the monitoring of the Protein A load phase [168, 169, 257, 258] as this application promises the most economic benefits due to the high costs of Protein A resin [266]. Economic improvements may be achieved due to multiple aspects. In conventional batch production, the Protein A column capacity is typically under-used. The acceptance range for the column loading density is set such that it can be kept constant during the resin lifetime. A dynamic termination of the load phase by detecting product breakthrough allows to use the optimal column capacity throughout resin life time. Furthermore, real-time PAT eliminates the need for completing at- or off-line titer measurements before starting the downstream process resulting in a more stream-lined production. As pharmaceutical companies move towards continuous processes, real-time monitoring of the Protein A load phase becomes more interesting to support robust process control. In continuous Protein A chromatography, the effluent of a first column is commonly loaded onto a second column, which allows to overload the columns without losing product. If a continuous load stream with a variable mAb titer is used, monitoring the product concentration in the breakthrough continuously reduces the dependence of the process on at- or off-line analytics and thus improves the process control.

Different spectroscopic sensors, like UV [169, 266], NIR [258], and Raman [168, 257], have been investigated for the purpose of quantifying the mAb concentration in the column effluent with varying success. Based on the literature data, UV spectroscopy and Raman spectroscopy seem to be the most promising techniques for the breakthrough monitoring of the Protein A load.

Raman spectroscopy has been successfully implemented to monitor various attributes during the upstream process of mAbs, including the mAb concentration in the complex cell culture fluid [162, 267–269]. A limiting factor for the application of Raman spectroscopy to the downstream process are the long acquisition times to derive a good signal-to-noise ratio. This is important, because process steps in the downstream take hours in comparison to days during the fermentation [264]. Therefore, Feidl et al. [168, 257] applied advanced preprocessing of the spectra and mechanistic modeling for the prediction of the mAb concentration to overcome the noise limitation of the Raman spectra due to short measurement times.

For monitoring the downstream process, the application of UV-based PAT methods was proven to be successful for selective mAb concentration determination in complex mixtures [73, 114, 169, 266, 270]. Raman spectroscopy has been proven to selectively quantify protein [131] and different buffer components [271], which can be interesting for UF/DF steps and formulation. In comparison to Raman-based techniques, UV spectroscopy offers a higher measurement speed and a better signal-to-noise ratio for quantification of proteins in aqueous solutions with the drawback of less selectivity for different protein features [264]. To compensate the lower selectivity and thereby improve the prediction of the UV-based PAT methods, dynamic background subtraction methods have been investigated to remove the influence of process-related impurities on the UV spectra [169, 266]. Another drawback of the UV spectroscopy in comparison to Raman spectroscopy is the detector saturation at high protein concentrations. To resolve this, a flow cell with adequate pathlength or with variable pathlength needs to be chosen. Raman spectroscopy has a larger working range due to more possibilities in laser and detector settings to avoid the saturation of the detector.

The comparison of both techniques with results from different studies remains difficult as different sample conditions and different methods for model optimization and model validation can influence the results dramatically. Therefore, a final conclusion can only be drawn, when using the different sensors on the same sample set and by applying the same model methodology. An application to the same sample set can be realized by serial in-line measurements with both sensors. This also enables the application of data fusion algorithms on the multimodal data set. Data fusion from multiple sensors promises advantages over data from a single source, like the statistical advantage of improving the number of measurements and the improved observability by combining multimodal measurement data [118]. The development and use of chemometric data fusion algorithms of multimodal spectroscopic sensors has been driven by food science [121, 272], but data fusion is starting to be used in biopharmaceutical production as well [264]. Up to the present, mostly low-level data fusion is used and a thorough investigation into the improved prediction by data fusion methods in comparison to single sensor models is missing.

In this study, Raman spectroscopy and UV spectroscopy are evaluated based on their ability to quantify the mAb concentration in the column effluent of the protein A column. It is discussed what molecular features the spectroscopic techniques measures in order to quantify the mAb concentration of complex mixtures. Additionally, data fusion techniques are applied to evaluate the benefit of two orthogonal sensors. First, traditional data fusion techniques, which are based on PLS modeling, are compared to the base PLS models of the individual sensors. Special emphasis is put on the considerations for variable and data block scaling, and on the comparison to the single sensor models. In a second step, the application of CNNs as non-linear regression techniques is evaluated for Raman and UV spectroscopy. Lastly, the potential of CNNs as a data fusion technique is explored and compared to the traditional PLS based data fusion techniques.

6.2 Materials and Methods

6.2.1 Biologic Material

All biologic material was stored at 5 °C before experimentation after delivery from our industry partner Sanofi-Aventis (Frankfurt, DE). In order to obtain a variable mAb concentration and a variable impurity profile in the load material, the product containing HCCF with a product concentration of 2 g/L (Feedstock 1) was mixed with purified product (Feedstock 2) and three different mock HCCF solutions (Feedstock 3 to 5). One mock solution was cultivated with a non-producing cell line. The other two mock solutions were prepared as flow-through by preparative Protein A chromatography. These two mock solutions were derived from HCCFs of two different cell lines which produce two different mAbs, respectively. Prior to this study, it was ensured that the Protein A flow-through did not contain antibodies in detectable concentrations (based on analytical Protein A chromatography). For product spiking, the used mAb (Feedstock 2) was purified to the second polishing step by our industry partner and was concentrated up to 20 g/L to reduce dilution effects of the impurities by addition of the concentrated product.

The product containing HCCF, purified mAb and mock HCCFs were filtered with a cellulose acetate filter with a pore size of 0.22 µm (Pall, Port Washington, NY, USA) before mixing. In Table 6.1, the used volumina of the different stock material for each run are shown. The composition of the mixtures between the three mock materials was determined by Latin Hypercube Sampling to provide a random multidimensional distribution.

Table 6.1: Sample composition for the calibration runs 1 to 4 and the validation run 5 with volumes of the product containing HCCF (Feedstock 1), purified mAb (Feedstock 2), mock HCCF (Feedstock 3), flow-through 1 (flow-t.1), and flow-through 2 (flow-t.2) (Feedstock 4 and 5).

Run	data usage	HCCF	mAb	flow-t.1	flow-t.2	mock HCCF
number	-	in mL	in mL	in mL	in mL	in mL
Run 1	calibration	52.50	0.00	9.85	21.91	20.74
Run 2	calibration	35.00	1.75	14.82	1.36	17.08
Run 3	calibration	21.00	3.15	6.48	3.93	7.44
Run 4	calibration	17.50	3.50	6.57	6.13	1.30
Run 5	validation	26.25	2.63	2.01	12.65	8.96

6.2.2 Chromatography Runs and Sensors

All preparative runs were realized with an Äkta Pure 25 purification system controlled by Unicorn 6.4.1 (Cytiva, Chicago, USA). The system was equipped with a sample pump S9, a fraction collector F9-C, a column valve kit (V9-C, for up to 5 columns), a UV-monitor U9-M (2 mm pathlength), a conductivity monitor C9, a pH valve kit (V9-pH) and an I/O-box E9. To monitor the breakthrough by Raman spectroscopy, a MarqMetrix BioReactor Ballprobe (MarqMetrix, Seattle, USA) was inserted into an in-house made flow cell. The probe was connected to a HyperFlux PRO Plus 785 Raman analyzer with Spectralsoft 2.8.0 (Tornado Spectral Systems, Toronto, Canada). The laser power during acquisition was set to 495 mW with an acquisition time of 800 ms and ten acquisitions per spectrum. The flow cell was placed after the conductivity monitor of the Äkta system. In Figure 6.1 the flow cell is displayed. X-, Y- and laser calibration were done before the experiment according to the manual. More information on the Raman measurement setup is given in the supplemental data 6.5.

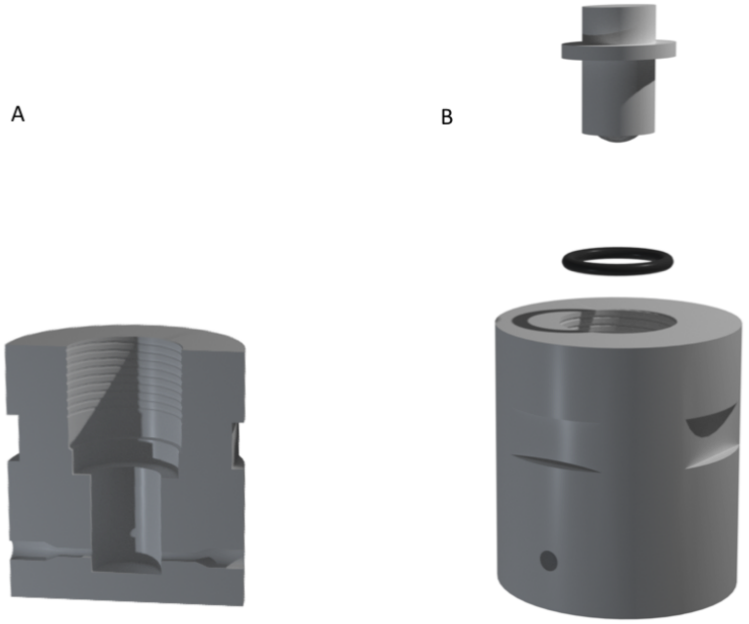

Figure 6.1: Cut of the (A) and exploded view of the in-house made flow cell, O-ring and MarqMetrix Ballprobe with welded flange (B). The flow cell consists of a block of stainless steel with a *Panzergewinde* (PG) 13.5-sized threaded borehole to insert the Ballprobe and two boreholes for 1/16 inch Äkta fingertight connectors.

Additionally, an UltiMate 3000 DAD equipped with a semi-preparative flow cell (0.4 mm optical pathlength) and operated with Chromeleon 6.8 (Thermo Fisher Scientific, Waltham, USA) was connected to the Äkta Pure. The DAD was positioned between the Raman flow cell and the V9-pH valve.

For the PLS model calibration and validation, breakthrough experiments with variable mAb titers in the feed were performed. The mAb titers in the different load materials were 1 g/L, 1.5 g/L, 2 g/L, 2.5 g/L, and 3 g/L. For each experiment, a prepacked 5 mm × 50 mm, MabSelect SuRe column (0.982 mL) (Repligen, Waltham, US) was first equilibrated for 5 CVs with a 25 mM TRIS and 0.1 mM sodium chloride buffer at pH 7.4, and then loaded with 100 mg of mAb. At the beginning of the load phase, the DAD equipped with a semi-preparative flow cell (optical pathlength 0.4 mm) was triggered to record absorption spectra between 200 nm to 800 nm and the column flow-through was collected in 200 µL fractions, as explained in more detail by Rüdt et al. [169]. An additional command was inserted into the MATLAB script (MATLAB version R2019b from the MathWorks, Inc., Natick, USA) to trigger the Raman measurements over Transmission Control Protocol/Internet Protocol (TCP/IP).

After the load phase, the column was washed for 4.5 CVs with equilibration buffer, before the mAb was eluted with 20 mM citric acid at pH 3.6. A sanitization was conducted with 50 mM sodium hydroxide and 1 mM sodium chloride for 5 CVs after each run.

6.2.3 Analytical Chromatography

Reference analysis of the collected fractions was performed using a Vanquish Flex Binary HPLC system (Thermo Fisher Scientific, Wilminton, US) by analytical Protein A chromatography. The system consisted of a Binary Pump F, Split Sampler FT, Column Compartment H and a Diode Array Detector HL. Chromeleon Version 7.2 SR4 (Thermo Fisher Scientific) was used to control the HPLC. The collected fractions of all runs were examined by analytical protein A chromatography to obtain the mAb concentrations. For each sample, a 2.1 mm × 30 mm POROS prepacked protein A column (Applied Biosystems, Foster City, USA) was equilibrated with 2 CVs of equilibration buffer, followed by an injection of 20 µL of sample. The column was then equilibrated with 0.8 CVs of equilibration buffer and eluted with 1.4 CVs of elution buffer. The flow rate was 2 mL/min for all phases and experiments.

Column equilibration was carried out using a buffer with 10 mM phosphate (from sodium phosphate and potassium phosphate) with 0.65 M chloride ions (from sodium chloride and potassium chloride) at pH 7.1. Elution was performed with the same buffer, but titrated to pH 2.6 with hydrochloric acid. All buffer components were purchased from VWR, West Chester, USA. The buffers were prepared with Ultrapure Water (PURELAB Ultra, ELGA LabWater, Viola Water Technologies, Saint-Maurice, France), filtrated with a cellulose acetate filter with a pore size of 0.22 µm (Pall), and degassed by sonification.

6.2.4 Data Analysis

Figure 6.2 shows an overview of the applied data analysis. First, the sensor signals were gathered and combined with the mAb concentration. For the UV and Raman spectra, various types of preprocessing were evaluated by two-block PLS modeling. Subsequently, the best preprocessing technique was applied to the raw data resulting in the data used for both data fusion by PLS modeling and CNN regression. These data were concatenated and pretreated for low-level data fusion by PLS modeling. Additionally the data were used to build the base PLS model for each spectroscopic technique. From the base models, the scores were concatenated and pretreated for mid-level data fusion by PLS modeling. Additionally, the predictions of the hierarchical models were taken for decision fusion PLS modeling for high level data fusion. Further details on the raw data analysis, PLS model calibration and evaluation, and CNN training is given below.

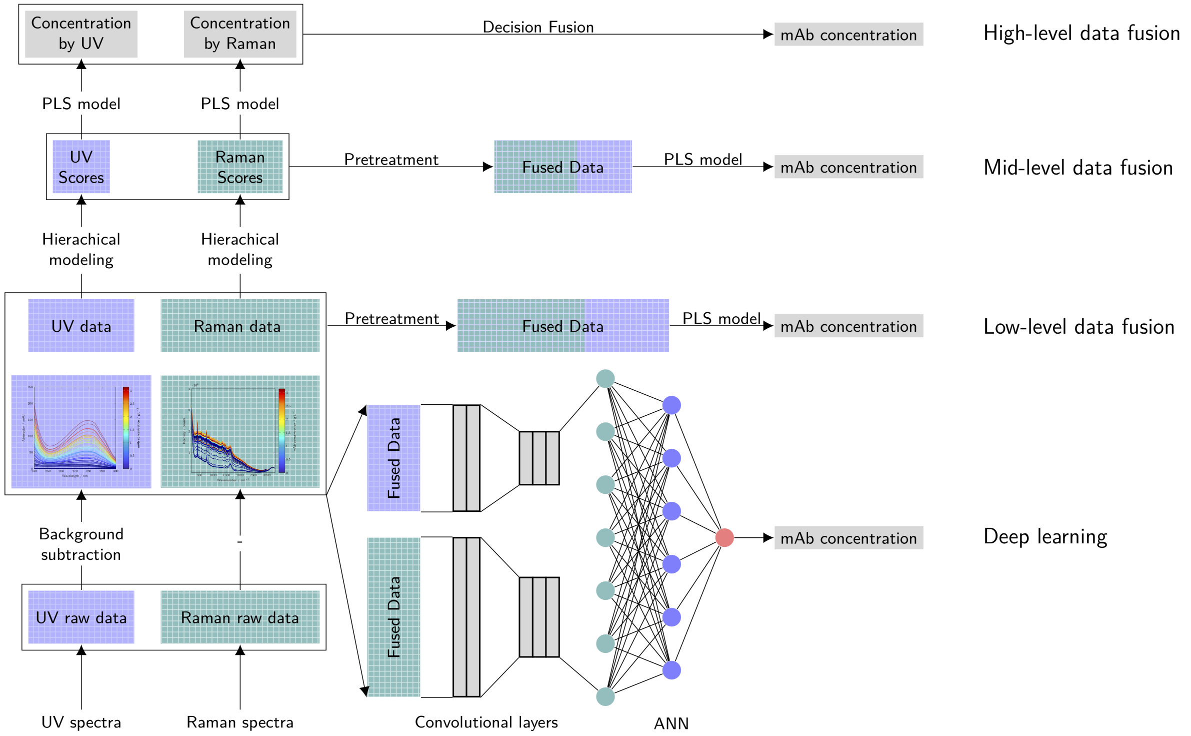

Figure 6.2: Methodology for the applied model building in low-level, mid-level, high-level data fusion, and deep learning.

Raw Data Analysis

The recorded Raman and UV spectra, the measured mAb concentration by analytical chromatography, and run data from the Äkta system were read in and processed with MATLAB R2019b (The MathWorks, Inc.). A background subtraction to remove the influence of contaminants on the spectra was evaluated for both spectra sets as described in Rolinger et al. [266]. After the background subtraction, the spectra were averaged according to the fraction size data from the Äkta. For the calibration/training of the different models, Run 1, Run 2, Run 4 and Run 5 were used as calibration dataset. Run 3 was always used as external validation, because it is the center point of the design space.

PLS Modeling

For the calibration of PLS models, SIMCA 13.0.3 (Sartorius, Göttingen, Germany) was used. SIMCA applies a 7-fold cross-validation as internal validation, by splitting the calibration data set in seven parts and leaving each part out of the calibration once. SIMCA applies the NIPALS-algorithm for PLS model building [96]. For the UV-based model, no spectral preprocessing was done except the previously explained subtraction of the background. All spectra and the mAb concentration were pretreated by mean-centering. The resulting model was chosen as base model for all PLS-based data fusion efforts.

For the Raman-based models, first, different spectral preprocessing steps were evaluated to improve the model prediction and linearity during calibration. This involved the use of an EMSC filter, first and second derivation, baseline removal and a background subtraction. Additionally, the different spectral preprocessing options were compared in Solo 8.9 (Eigenvector Research, Inc., Wenatchee, USA) with the optimization tool. After the evaluation of different preprocessing options, the best Raman model was chosen as base data along with the UV model for comparing the prediction quality and data fusion purposes.

Often data fusion is grouped into three different levels, namely low-level, mid-level and high-level data fusion [121, 212]. In this study, the results of the different fusion level will be compared to each other. Low-level data fusion is the concatenation of the preprocessed UV and Raman spectra. Mid-level data fusion refers to additional variable selection prior to the concatenation of the spectra. In this study, hierarchical PLS modeling will be used as main variable selection technique. With hierarchical PLS modeling, the score vectors of the base model are taken as input variables, also referred to as "super variables", for a new PLS model [211]. For high-level data fusion, an output fusion of the base PLS models was carried out by hierarchical PLS modeling.

The basis for successful data fusion is proper data alignment [118]. Here, both data sets were already aligned time-wise and averaged according to the collected fractions before preprocessing or concatenation. Due to the two-dimensional nature of the UV and Raman spectra, no dimension reduction before concatenation was necessary. However, the UV and Raman spectra differ in the number of variables and in the total value of the variables. To prevent the greater influence of one data set onto the model by either the total value of the variables or the number of variables in the data set, proper scaling is important [96].

The preprocessing methods used in this study are mean-centering, unit variance scaling and Pareto scaling. Mean-centering performs a subtraction of the mean value of a signal $\bar{x}_j$ (Equation 6.1) from the measured values x_{ij} with i being the sample number and j being the signal number. In case of unit variance scaling, the mean-centered value is divided by the standard deviation of the signal s_j (Equation 6.2) to account for any difference in the signal variance. Pareto scaling is an intermediate between mean-centering and unit variance scaling, as the mean-centered values are divided by the square root of the standard deviation s_j [198].

$$\bar{x}_j = \frac{\sum_{i=1}^{n} x_{ij}}{n} \tag{6.1}$$

$$s_j = \sqrt{\frac{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2}{n-1}} \tag{6.2}$$

$$\text{Center} \quad \hat{x}_{ij} = x_{ij} - \bar{x}_j \tag{6.3}$$

$$\text{Unit variance} \quad \hat{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \tag{6.4}$$

$$\text{Pareto} \quad \hat{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\sqrt{s_i}} \tag{6.5}$$

CNN

The neural networks were built in Python version 3.6 (Python Software Foundation, Wilmington, USA) using NumPy version 1.18.5 [273], pandas version 1.0.5 [274] and TensorFlow version 2.2.0 [275] as libraries. For all models, a hyperparameter optimization was done via Bayesian optimization (Keras Tuner, version 1.0.1 [276]).

The structure of the used CNNs may be broadly split into convolutional blocks and a fully connected block. Every convolutional block consisted of a convolutional layer, a pooling layer and a dropout layer. The number of such convolutional blocks was optimized in the range from 1 to 3 and from 1 to 2 for the Raman- and UV-based model, respectively. The window width of the first convolutional layer was allowed to change from 60 to 130 for the Raman-based model and from 4 to 30 for the UV-based model. To initialize the kernel of the first convolutional layer of the Raman model, a first and second derivative Gaussian wavelet was used. Thereafter, a dense layer with 1 to 52 neurons was optimized. Swish was used as activation function [277]. As beta was not specified, Swish is equivalent to a Sigmoid-weighted Linear Unit. The output layer was fixed with one densely-connected neuron with a leaky ReLu activation function (alpha of 0.1) and a bias. This was chosen due to the linearity of the ReLu function in the positive domain and the attenuation of negative values. The weights of the neurons were optimized with Adaptive Moment Estimation (Adam) [278]. The learning rate of Adam optimizer was a further hyperparameter varied by Bayesian optimization. As loss function Mean Square Error (MSE) was used.

For the combined Raman and UV-based CNN model, only a hyperparameter optimization of an additional dense layer on top of the individual dense layers way done to combine both models. Bayesian optimization was used again with a range between 12 and 64 neurons in the dense layer with the same conditions for the learning rate as for single sensor models.

6.3 Results and Discussion

This paper focuses on a comparison of UV- and Raman-based monitoring of the Protein A breakthrough as well as the evaluation of data fusion techniques for both sensor signals. UV data was preprocessed as described by Rolinger et al. [266], which leads to a significantly improved prediction as it suppresses absorption from interfering co-eluting species. For an analysis of the UV spectra during the load phase, a comparison to elution spectra and a detailed discussion on the effects of the preprocessing, we refer to Rolinger et al.. In the following, the focus is set towards an analysis of the Raman spectra and the comparison of the prediction quality based on UV- and Raman-based models. First, the observable features of Raman spectra will be analyzed followed by a discussion on the performance of the different PLS models for the Raman spectra and data fusion. Finally, the results from the CNN models are introduced and discussed for the individual sensors and the fused data.

6.3.1 Raman Spectra

Figure 6.3 shows the Raman raw spectra, the first, and second derivative colored according to mAb concentration. For further data analysis, only the raw spectra were used. The first and second derivative are plotted to show the influence of the background removal on the spectra. It is interesting to note that the raw spectra show an underlying baseline effect that increases with increasing run time. The intensity of this effect varies for every feed stock. The background spectra for each run are shown in the supplemental data 6.6. Therefore, when looking at multiple runs, the raw spectra are not primarily sorted by mAb concentration but rather by run-specific baseline effects. For every individual run, a trend of increasing baseline with increasing run time after the impurity breakthrough is apparent. Within each run, the baseline increase is visually the strongest effect over the run time in the spectra. The first derivative mostly removes the baseline effects except for the steep increase below 400 cm^{-1}. The second derivative removes the baseline effect completely. However, it also becomes obvious that very little change remains in the spectra after removal of the baseline by derivation. Additionally, the signal-to-noise ratio is decreased by the derivation.

In Figure 6.4 the Raman spectra over the course of run 2 are plotted to show the formation of the Raman bands over the process time. The most prominent effect, which also partly correlates with the mAb concentration, is the increase in background scattering. The spectrum with the lowest overall intensities is the first spectrum of the run, where only buffer is measured. The sapphire band at 418 cm^{-1} is the strongest band in the spectrum. No wavenumber-dependent intensity correction was performed. Otherwise the water bands around 3000 cm^{-1} would be more prominent as well. Proteins have low Raman scatter cross sections [264], which makes the contribution of water in the spectrum more prominent. The strongest protein bands seem to be caused by phenylalanine (1006 cm^{-1}), tryptophan (1360 cm^{-1}), C-H deformations (1421 cm^{-1}, 1468 cm^{-1})[151, 279] and C-H stretching at 2952 cm^{-1} [34]. Overall, with increasing run time there are more weak protein-based peaks present in the spectral range 500 cm^{-1} to 1700 cm^{-1}, which are corrupted by noise.

Jiskoot et al. estimates the limit of quantification for proteins in aqueous solutions to range between 1 % to 5 % [34] which corresponds to a concentration 10 g/L to 50 g/L. Wen et al. claim that therapeutical proteins can be quantified from 1 g/L due to significant instrument improvements [131]. From the shown spectra, it seems that a quantification to lower concentrations is possible with our setup. In general, the quantification does not seem to rest on features generated by the protein backbone, i.e. the amid

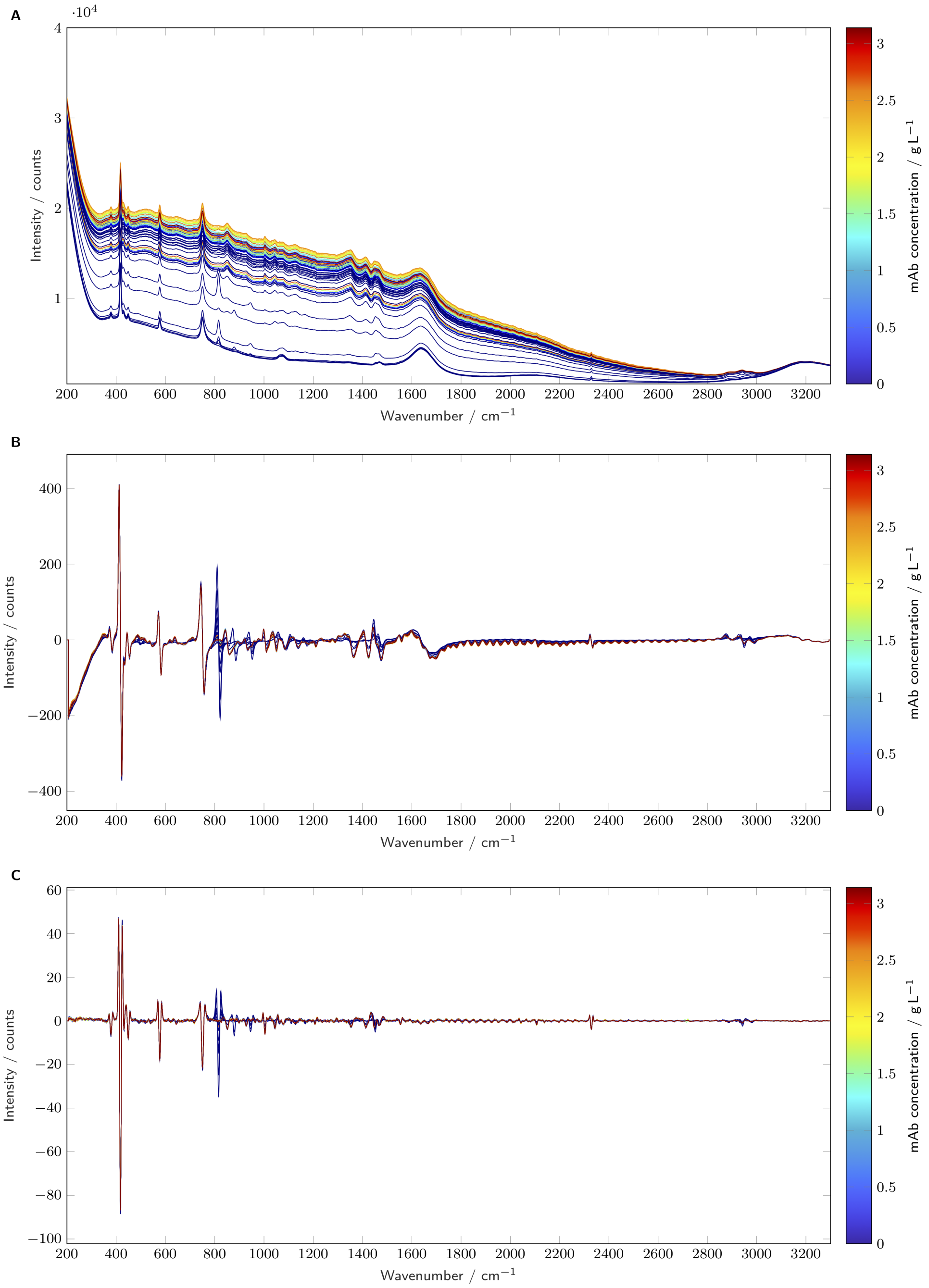

Figure 6.3: The raw (A), first derivative (B) and second derivative (C) spectra of the calibration runs. The spectra are colored by mAb concentration.

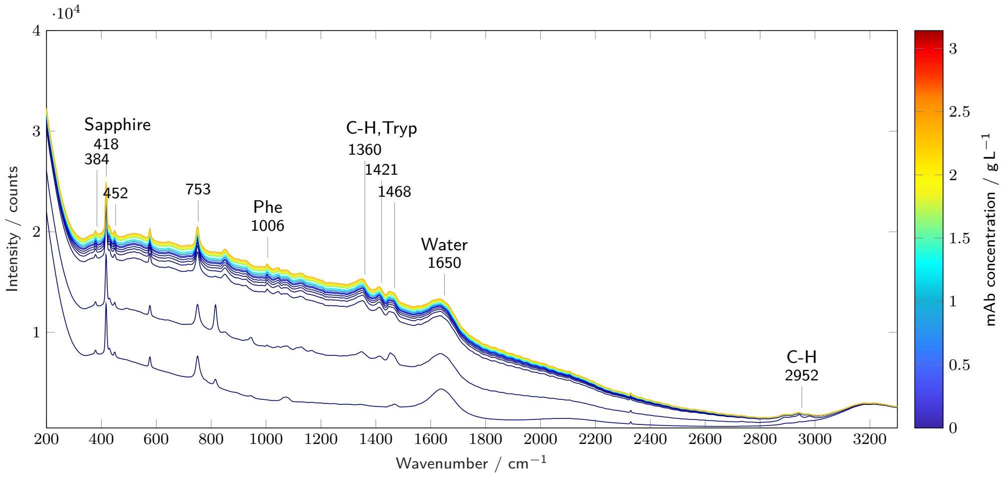

Figure 6.4: Every 10th Raman spectrum of run 2 is plotted and colored by the mAb concentration. The prominent bands in the spectra are assigned to the generating species sapphire glass, water, buffer and protein.

bands, but rather on bands related to aromatic groups and C-H vibrations. A selective quantification by Raman spectroscopy between different protein species, based on other protein structure elements than aromatic groups and C-H vibrations, in the investigated concentration range seems difficult due to the low signal-to-noise ratio of the amide bands.

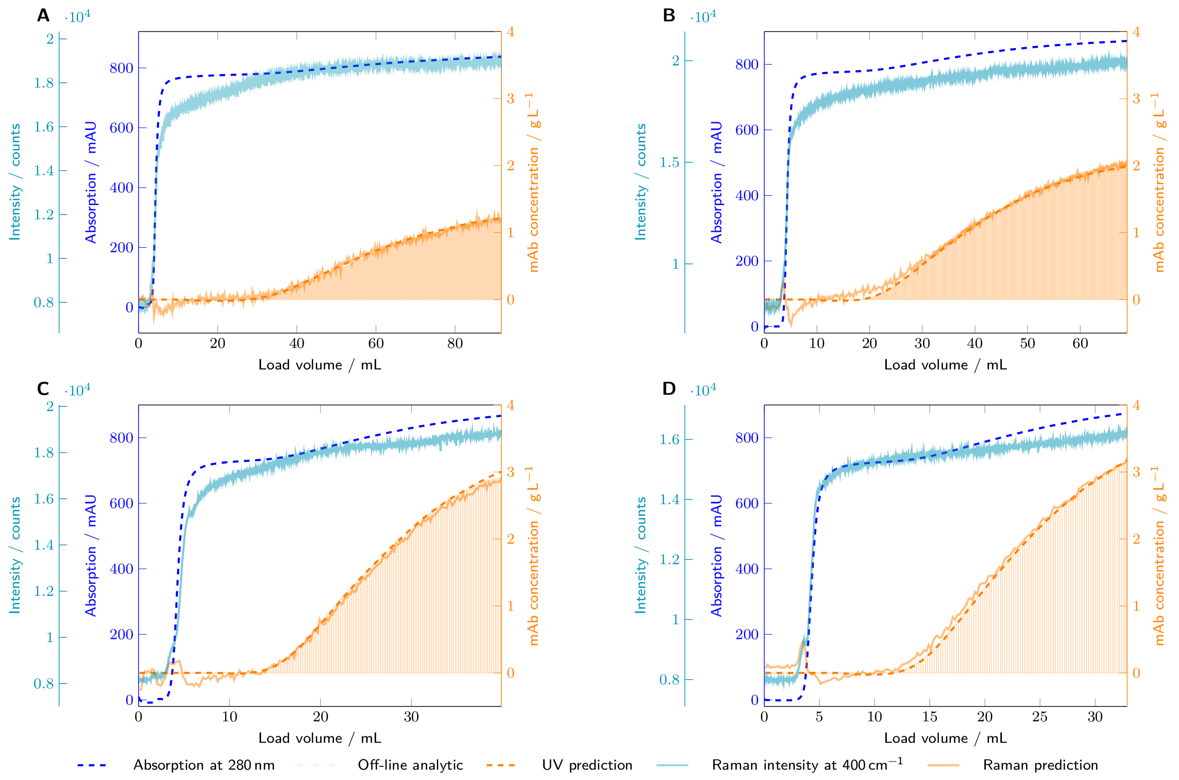

Figure 6.5: Results of the PLS model calibration for Raman and UV-based PLS models. The UV absorption at 280 nm A_{280} (displayed as dashed blue line) and Raman intensity at 400 cm^{-1} (displayed as solid cerulean line) are compared with the results of the off-line analytics for mAb quantification (orange bars). The UV-based PLS model prediction is illustrated as dashed orange line. The Raman-based PLS model prediction is illustrated as orange line. The four runs exhibited variable mAb titers in the feed A: 1 g/L, B: 1.5 g/L, C: 2.5 g/L, and D: 3 g/L.

Figure 6.5 compares the raw signals of UV absorption at 280 nm with the Raman intensity at 400 cm^{-1} over the run time. At a wavenumber of 400 cm^{-1}, no relevant Raman scattering of proteins exists [131, 151], i.e. any change may be considered a background effect. A distinct increase over the process run time is visible for the Raman intensity similar to the trend of the UV absorption. This background effect is sometimes attributed to fluorescence of cell culture components [280, 281]. However, the same background effect is seen in aqueous protein solutions with increasing protein concentration [282]. As the intrinsic protein fluorescence does not reach above 500 nm, the observed background effect is probably not caused by fluorescence [129]. It seems more likely that Rayleigh scattered light is incompletely blocked by the notch filter and optical grating [282]. The increase in scattered light could also be attributed to the change in refractive index, which is correlated to protein concentration. During the load phase, impurities with large molecular weight (e.g. DNA, HCPs) flow through the column and lead to an increased amount of Rayleigh scattering, before the mAb breaks through.

6.3.2 Comparison of UV- and Raman-based PLS Models

For the UV-based PLS model, it was previously established that a background subtraction significantly improves the precision of the UV-based PLS model [169, 266]. Based on the high quality of the prediction, the conductivity-based background subtraction was chosen as preprocessing. No further preprocessing was performed for the UV spectra.

For the calibration of the Raman-based PLS model, different preprocessing methods were evaluated. The model with the best calibration results by cross-validation was chosen as base model. The tested preprocessing methods were conductivity-based background subtraction, derivatives, and baseline removal by extended multiplicative scatter correction and asymmetric Whittaker smoothing. However, the raw data provided the best results during cross-validation. This could be caused by the noise increase in the data due to a subtraction of a noisy background spectrum or due to the amplification of noise by derivation, respectively. It is also interesting, that a baseline removal did not yield a better model compared to the raw data. Apparently, the PLS model uses the background scattering effect to improve the prediction quality.

In Figure 6.5, the calibration results of the UV-based and the Raman-based PLS models are plotted and compared to the reference analytics.

Additionally, as discussed in section 6.3.1, the UV absorption at 280 nm and the Raman intensity at 400 cm^{-1} are compared. The results of the UV-based and Raman-based PLS models are listed in Table 6.2.

The UV-based PLS model has a better prediction accuracy with a higher coefficient of determination $R2$, a higher coefficient of determination during cross-validation $Q2$, and a lower RMSECV. Regarding the RMSEP, the difference between the models is even more pronounced. The RMSEP of the UV-based PLS model is 0.013 g/L while it is 0.232 g/L for the Raman-based PLS model. In Figure 6.6, the model predictions are depicted. The UV-based model prediction and the reference mAb concentration show only minimal differences. The Raman-based prediction shows an offset to the reference mAb concentration. Additionally, the difference between prediction and measured concentration increases starting at a mAb concentration higher than 1.9 g/L. This seems to be a nonlinear behavior. When looking at the loadings of the Raman-based PLS model, the first loading has a high similarity to the background effect and the following loadings show protein bands. It seems, that the PLS model uses both the background effect and the protein bands to estimate the mAb concentration. Even though the background effect increases with increasing mAb concentration, the background effect alone cannot be used as sole predictor for the mAb concentration in this data set, because the initial intensity of the background spectrum depends on the feedstock composition. The use of the background effect, which has an offset between the different runs, could impede the linearity between spectra and protein concentration. The deviation from the linearity between concentration and certain Raman peaks could also be cause by the measurement with the ball probe, the influence of the refractive index when protein concentration is increasing or inhomogeneities in the sample flow in the flow cell.

The UV-based PLS model has a better prediction accuracy with a higher coefficient of determination $R2$, a higher coefficient of determination during cross-validation $Q2$, and a lower RMSECV. Regarding the RMSEP, the difference between the models is even more pronounced. The RMSEP of the UV-based PLS model is 0.013 g/L while it is 0.232 g/L for the Raman-based PLS model. In Figure 6.6, the model predictions are depicted. The UV-based model prediction and the reference mAb concentration show only minimal differences. The Raman-based prediction shows an offset to the reference mAb concentration. Additionally, the difference between prediction and measured concentration increases starting at a mAb concentration higher than 1.9 g/L. This seems to be a nonlinear behavior. The deviation from the linearity between concentration and certain Raman peaks could be cause by the measurement with the ball probe, the influence of the refractive index

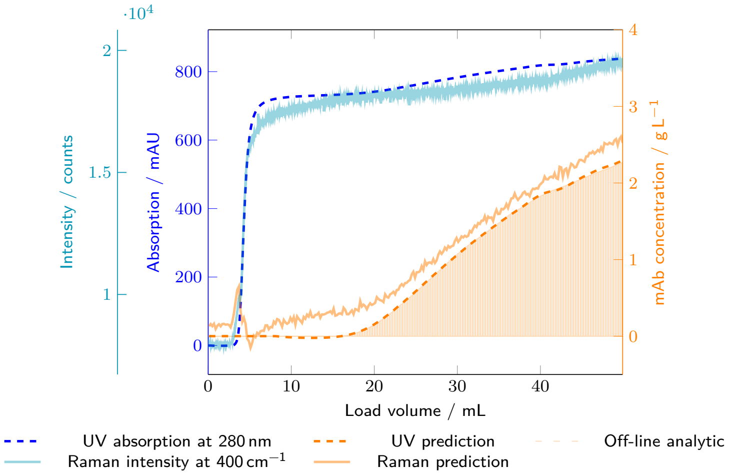

Figure 6.6: Results of the PLS model validation of run 4 for Raman and UV-based PLS models. The UV absorption at 280 nm A_{280} (displayed as dashed blue line) and Raman intensity at 400 cm^{-1} (displayed as solid cerulean line) are compared with the results of the off-line analytics for mAb quantification (orange bars). The UV-based PLS model prediction is illustrated as dashed orange line. The Raman-based PLS model prediction is illustrated as orange line.

when protein concentration is increasing or inhomogeneities in the sample flow in the flow cell.

In the performed experiments, the RMSEPs of both PLS models are expected to be comparable with the RMSECV or lower, because the validation run lays in the middle of the calibration design space. For the Raman-based model, the RMSEP is, however, higher compared to RMSECV, which can indicate an overfitting as the validation run should be in the center of the design space. The increased RMSEP of the Raman-based model could be caused by the relatively high number of seven LVs in comparison to two LVs used by the UV-based PLS model.

It is also worth noting that the prediction of the Raman-based model appears to be more corrupted by noise (less precise) than the prediction of the UV-based model. This indicates that the Raman-based prediction is more strongly affected by measurement noise than the UV-based predictions. Improvements in measurement quality of the Raman spectra could thus potentially improve the prediction quality.

Additionally, the correlation of prediction of the Raman-based model and mAb reference concentration starts to deviate from the linear relation, especially for run 3 and mAb concentration above 1.9 g/L (see also supplemental data 6.8 for an observed versus predicted plot). The UV-based model shows only very little deviation from the linear relation, probably caused by errors in the reference analytic. The stronger deviation from the linear correlation of the Raman-based model could explain why a higher number of LVs is necessary for the Raman-based model in comparison to the UV-based model. PLS models can approximate non-linearities by including additional LVs [172].

In summary, for the investigated experimental conditions, UV spectroscopy is better suited for monitoring the mAb breakthrough during protein A chromatography than used Raman spectroscopy setup. The UV-based PLS model reaches a more than 10-fold lower RMSEP compared to the Raman-based PLS model. While there might still be chromatographic capture steps, where a Raman-based PLS model performs better (e.g. high mAb concentration and high variation in UV absorbing background species), the distinctively lower RMSEP of the UV-based model indicates a competitive advantage for most applications involving mAbs. The competitive advantage is further supported by the simpler equipment requirements for UV spectroscopy which may simplify implementation in production environments. Additionally, the used Raman setup might not work for all feedstocks due to autofluorescence [283]. The only solution in the case of large autofluorescence is to switch to a longer laser wavelength by using a different equipment. As longer laser wavelengths will cause a weaker Raman signal, the exposure times need to be longer to achieve the same signal-to-noise ratio, which might not be feasible for the typical measurement times in chromatography.

6.3.3 Data Fusion for UV- and Raman-based PLS models

The results of the different data fusion levels and data pretreatments are compared in Table 6.2. For low-level data fusion, both spectra were scaled individually and block scaling was eventually applied. With only mean centering, an RMSEP of 0.290 g/L is achieved in comparison to an RMSEP of 0.092 g/L with Pareto scaling and an RMSEP of 0.044 g/L with unit variance scaling. When comparing the results of the low level data fusion models without block-scaling, it is noticeable, that the less influence the Raman data have on the model prediction, the better the fused model gets. This is expected as the solely UV-based model has better performance than the corresponding Raman model. Without scaling, the Raman spectra reach intensities of more than 30 000 c in comparison to the around 200 mAU reached by the UV spectra. The absolute change in variables of the Raman spectra are larger as well due to the scale of the spectra. When only applying mean-centering, this larger variance in the Raman spectra biases the PLS model to mostly include Raman-based signals into the first LVs (i.e. the high variance variables).

Table 6.2: Input data, data fusion level, scaling, block scaling, R^2, Q^2,Root Mean Square Error of Calibration (RMSEC), RMSECV, RMSEP and number of LVs for the PLS models.

Input data	data fusion level	hierarchical level	scaling	block scaling	R^2	Q^2	RMSEC in g/L	RMSECV in g/L	RMSEP in g/L	number of LVs
UV	-	base	center	-	0.999	0.999	0.025	0.025	0.013	2
Raman	-	base	center	-	0.992	0.992	0.073	0.076	0.232	7
both	low	-	center	-	0.986	0.986	0.100	0.101	0.290	6
both	low	-	Pareto	-	0.976	0.976	0.129	0.129	0.092	4
both	low	-	Unit var.	-	0.999	0.999	0.025	0.025	0.044	5
both	low	-	center	1/sqrt	0.987	0.987	0.096	0.096	0.155	4
scores	mid	top	center	-	0.976	0.975	0.013	0.131	0.433	4
scores	mid	top	Pareto	-	0.986	0.986	0.100	0.100	0.313	3
scores	mid	top	Pareto	1/sqrt	0.990	0.990	0.082	0.082	0.231	3
scores	mid	top	Unit var.	-	0.998	0.998	0.040	0.040	0.118	1
scores	mid	top	Unit var.	1/sqrt	0.998	0.998	0.040	0.040	0.129	2
output	high	top	center	-	0.998	0.998	0.040	0.040	0.129	1
output	high	top	Unit var.	-	0.998	0.998	0.040	0.040	0.129	1

In contrast to mean-centering, unit variance scaling additionally divides each variable by their standard deviation. Therefore, the scale of the variables gets removed. The advantage of unit variance scaling is, that not a few variables dominate the total variance of all variables. Thus, also variables with smaller variance and a good correlation to the response may become relevant for model building. The disadvantage of the unit variance scaling is the noise inflation, which usually reduces the performance of PLS models [198]. Pareto scaling is an intermediate between mean centering and unit variance scaling as variables are scaled by the square root of the standard deviation. When little is known about the importance of the different blocks for the response prediction, unit variance scaling seems a good option even though a less accurate model is achieved than by only using the UV block for prediction.

As the Raman spectra have 3101 variables in comparison to the UV spectra with 171 variables, the contributed variance of the Raman spectra to the complete X block is larger even after unit variance scaling. To avoid this bias after preprocessing, the different blocks can be multiplied by different weights. These weights typically consist of a term to make the scale of the different blocks more even. Here, the mean centered blocks were scaled by the reciprocal square root of the number of variables in each block [284]. By block scaling, the RMSEP of 0.290 g/L of the mean centered model was lowered to 0.155 g/L as the large number of variables from the Raman spectrum had less influence on the prediction.

As an approach for mid-level data fusion, hierarchical PLS modeling was chosen. In hierarchical modeling, the individual spectra are multiplied by the loadings of each LV to calculate the scores of each spectrum. The different loadings of the UV- and Raman-based PLS model are displayed in the supplemental data 6.7. When using hierarchical modeling, the same consideration for the scaling are necessary as in low-level data fusion. Again, as with low-level data fusion, the closer the scores are scaled to unit variance, the lower the RMSEP becomes. With only mean centering and mid-level data fusion, an RMSEP of 0.433 g/L is achieved in comparison to an RMSEP of 0.313 g/L with Pareto scaling and an RMSEP of 0.118 g/L with unit variance scaling. Interestingly, the RMSEPs of the unit variance scaled and Pareto scaled mid-level data fusion models are higher than the original RMSEP of the Raman-based PLS model. An explanation for this could the low linearity of the Raman spectrum with regard to the mAb concentration. The Raman base model uses the background effect to a certain degree to allow for a better prediction. With mid-level data fusion, the number of LVs are generally lower and an approximation of the non-linearities is more difficult, because fewer co-linear parameters are available for the fit.

High-level data fusion was realized as output fusion in this study, where the predictions of the base models were fused by a PLS model. In the case of output fusion, the scaling of the variables is not important as they are already on the same scale. Therefore, different scaling methods, have the same result in our case. An RMSEP of 0.118 g/L is achieved. This RMSEP is almost the average of the two base models with leveraging the UV-based model more due to a regression coefficient of 0.503 in comparison to 0.497. As an alternative to PLS, other techniques like Bayesian belief networks could be used as well.

We conclude, that the best way of optimizing a prediction is to choose the right sensor from the start [175, 285]. For the purpose of monitoring the mAb concentration in the effluent of an Protein A column, UV spectroscopy is better suited than Raman spectroscopy due to a higher sensitivity and better linearity. Often the limited selectivity of UV spectroscopy is mentioned as a drawback, but for this application case the sensitivity seems to be no issue possibly due to the applied background subtraction. Even though data fusion has been reported as a useful tool, when combining a good sensor with a sensor with limited observation ability of the effect in focus, data fusion can do very little beyond the capacity of the best sensor. We therefore would like to issue a word of caution on the application of data fusion for data sets with poor sensors or without understanding the possible benefit of data fusion. Even though we have seen an increasing body of literature where data fusion is applied [116, 117, 286], data fusion methods should be considered skeptically. If a sensor cannot quantify a concentration on its own, a fusion with a different sensor will likely not lead to meaningful results in regression. The risk of coincidental correlations and overfitting is increased. In our case, the prediction were always worse when combining UV and Raman spectra than the UV-based prediction alone. A solution could be the application of non-linear models, like ANNs to improve the prediction ability of the Raman models and thereby the accuracy of the fusion models.

6.3.4 CNNs for UV and Raman Data

Table 6.3 shows the hyperparameters after the Bayesian optimization.

Even though the UV-based CNN and Raman-based CNN were given similar boundaries for the optimization, the optimum of the UV-based CNN has less convolutional layers, less filters and smaller window widths, which implies that less data ‘preprocessing’ is required for the UV-based CNN. The first convolutional layer in the Raman-based CNN was initialized by wavelets which imitate a first and second derivation. Otherwise the optimization did

Table 6.3: Hyperparameter found by Bayesian Optimization for the Raman and the UV-based CNNs

Hyperparameter	Raman	UV
Number of convolutional layers	3	2
Window width convolutional layer zero	90	4
Pooling width convolutional layer zero	11	1
Number of filters in convolutional layer zero	8	2
Window width convolutional layer one	16	6
Pooling width convolutional layer one	1	1
Number of filters in convolutional layer one	8	8
Window width convolutional layer two	28	-
Pooling width convolutional layer two	1	-
Number of filters in convolutional layer two	6	-
Number of neurons in fully-connected layer	46	31
Learning rate	0.001	0.001

not converge on an optimum of comparable quality as a PLS model. The output of the convolutional layers for the UV- and Raman based model are displayed in Section 6.9. Figure 6.7 shows the predictions of the UV-based, Raman-based and combined CNN model for the external validation run.

Table 6.4: RMSEC, RMSEP of the Raman, UV-based and combined CNNs

Input data	RMSEC in g/L	RMSEP in g/L
UV	0.019	0.013
Raman	0.078	0.220
both	0.047	0.050

Table 6.4 lists the RMSEC and RMSEP of the CNN models. The UV-based CNN predicts the mAb concentration accurately with an RMSEP of 0.013 g/L. The Raman-based CNN has a prediction, which is more corrupted by noise in comparison to the UV-based CNN. The higher RMSEP of 0.220 g/L is not only caused by the increased noise, but also by an offset. Both CNNs deliver comparable results to the base PLS models. The CNN with the combined data had 21 neurons in the additional fully connected layer after optimization. With this, an RMSEP of 0.050 g/L was reached. The CNN with the combined data lays between the results of the individual models with regard to noise in the prediction and RMSEP.

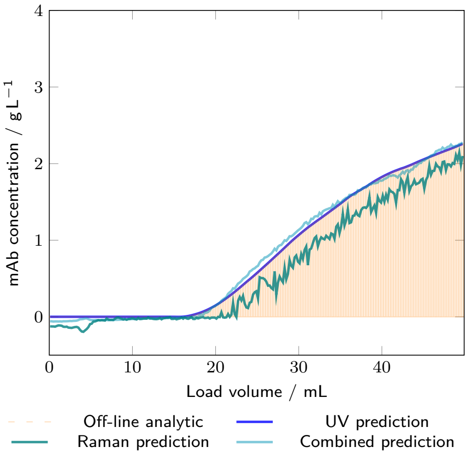

Figure 6.7: Results of the CNN model validation of run 4 for the Raman, UV-based and combined CNN models. The UV-based model prediction (displayed as solid blue line), the Raman-based model prediction (displayed as solid teal line) and the combined model prediction (displayed as solid cerulean line) are compared with the results of the off-line analytics for mAb quantification (orange bars).

For the presented study, the use of CNNs in comparison to PLS models only offers a limited benefit. The training of CNNs needs more resources and wrong setting of the initial start conditions can lead to a divergence of the training. In our case, the training set with 1169 training spectra was bigger compared to usual spectroscopic training sets. A lower amount of training spectra will probably cause problems for CNNs due to the high number of parameters.

6.4 Conclusion and Outlook

In this study, Raman and UV spectroscopy have been compared in their ability to predict the mAb concentration in the column effluent during the load phase of the Protein A capture step. Additionally, data fusion strategies based on PLS models and CNNs were presented and compared to the single sensor models.

We conclude, that UV spectroscopy achieves a better prediction accuracy in comparison to Raman spectroscopy. UV- and Raman-based PLS models required two, respectively seven LVs. The high number of LVs of the Raman-based PLS model may be related to nonlinearities, which are more difficult to fit by the linear PLS model. Of all fusion approaches, no model was better than the simple UV PLS model or the corresponding CNN model, which both achieved an RMSEP of 0.013 g/L. Data fusion for regression purposes seems not to be beneficial, if one sensor already provides a very good accuracy and additional sensor could only contribute noise. For Raman spectroscopy, the application of CNNs in comparison to traditional PLS models improved the prediction of the mAb concentration from 0.232 g/L (PLS model) to 0.220 g/L. The training and optimization of CNNs for both UV and Raman data was time-consuming. The success was dependent on establishing proper boundaries and starting conditions for model optimization. In our opinion, it seems generally not worth the effort to apply non-linear models to the monitoring of the mAb breakthrough, because a similar prediction accuracy can be reached with traditional PLS models.

For future technology evaluations for the implementation of real-time monitoring of the Protein A capture step, we consider UV spectroscopy to have a competitive advantage compared to Raman spectroscopy due to the better prediction quality and the simpler equipment. Raman spectroscopy may be of interest, if alternative chemicals should be monitored in the column effluent which do not have a UV absorption

6.5 Appendix: Raman Measurements

785 nm was chosen as laser wavelength for the Raman spectrometer as this is the most common laser wavelength for PAT applications. Shorter wavelengths offer the benefit of higher scattering efficiencies, but often also have a lower laser power and a higher risk of fluorescence. Therefore, there is no benefit in going to a lower laser wavelength. Different commercial spectrometer were evaluated based on literature information, supplier case studies and personal communication with other Raman users.

To find good measurement settings, the pure HCCF was measured with the Raman spectrometer before the experiments to find appropriate setting for the measurement. To achieve the best possible signal in the short amount of time, the maximum laser power (495 mW) was chosen. The use of 495 mW is a standard equipment setting during fermentation monitoring, so we concluded that it would not induce protein degradation. Then the auto exposure function of the Raman spectrometer was used to determine the exposure time in order to maximize the recorded counts without oversaturating the detector. For this study, the number of exposures per spectra was not important, because the spectra were averaged by fraction size later on.

6.6 Appendix: Background Effect in Raman Spectra

Figure 6.8 shows the Raman spectrum of the last fraction before the mAb concentration increases. The intensity of the background signal differs between runs. Therefore, the background intensity alone cannot be a sole indicator of the mAb concentration.

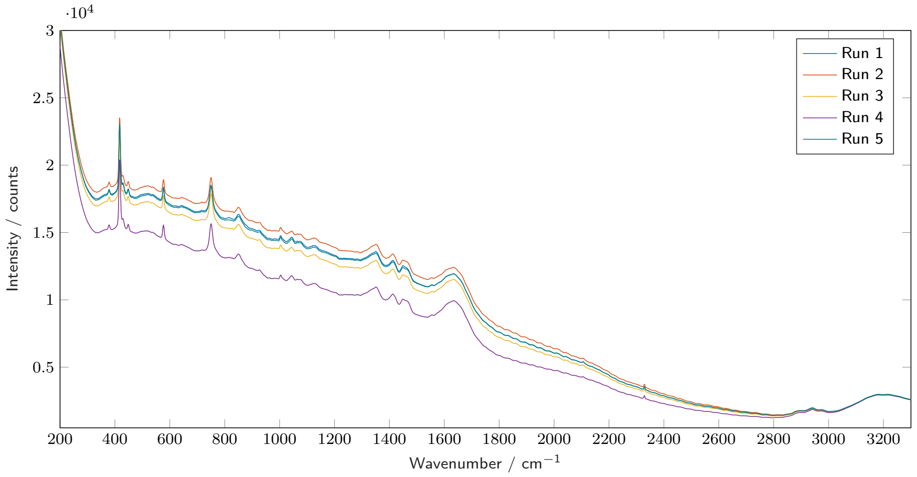

Figure 6.8: Last Raman spectrum before the mAb concentration starts to increase for Run 1 to Run 5.

6.7 Appendix: Loadings of the PLS Models

The loadings of a PLS model are the basis for calculation of the regression matrix. Thereby, the loadings can indicate which parts of the spectra are important for calculation of the y variable from x, in this case the calculation of the mAb concentration from the UV or Raman spectra.

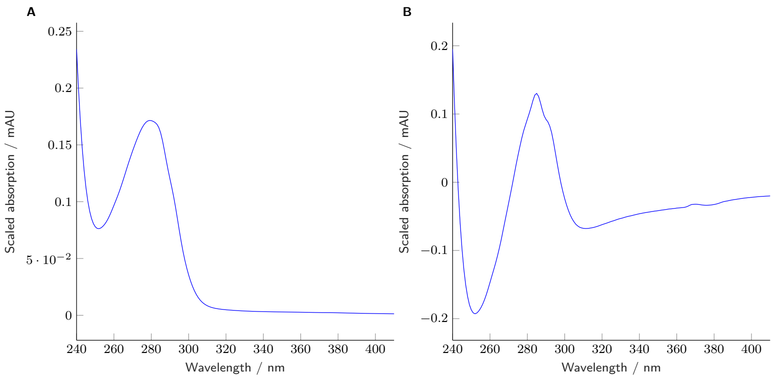

Figure 6.9: Loadings over wavelengths for the first (A) and second (B) LV of the UV-based PLS model.

In Figure 6.9, the loadings of the LVs of the UV-based PLS model are displayed. The first loading seem to describe the overall UV spectrum, while the second loading weights the spectral difference between 250 nm to 300 nm. The second loading is typical for compensation of overlapping spectra with two components [181]. Even though a background subtraction was performed, the observed loadings indicate a spectral contribution from contaminants.

In Figure 6.10, the loadings of the LVs of the Raman-based PLS model are displayed. The loading of the first LV shows the background effect in the Raman spectra. The loading of the second LV still accounts for a background effect, especially below 400 cm^{-1}. From the loading of the third LV on, the loading seem to be based on protein structure elements. The loadings are alternating between positive and negative absolute value depending on the LV. Again, this could be caused by the correlation between those bands [175].

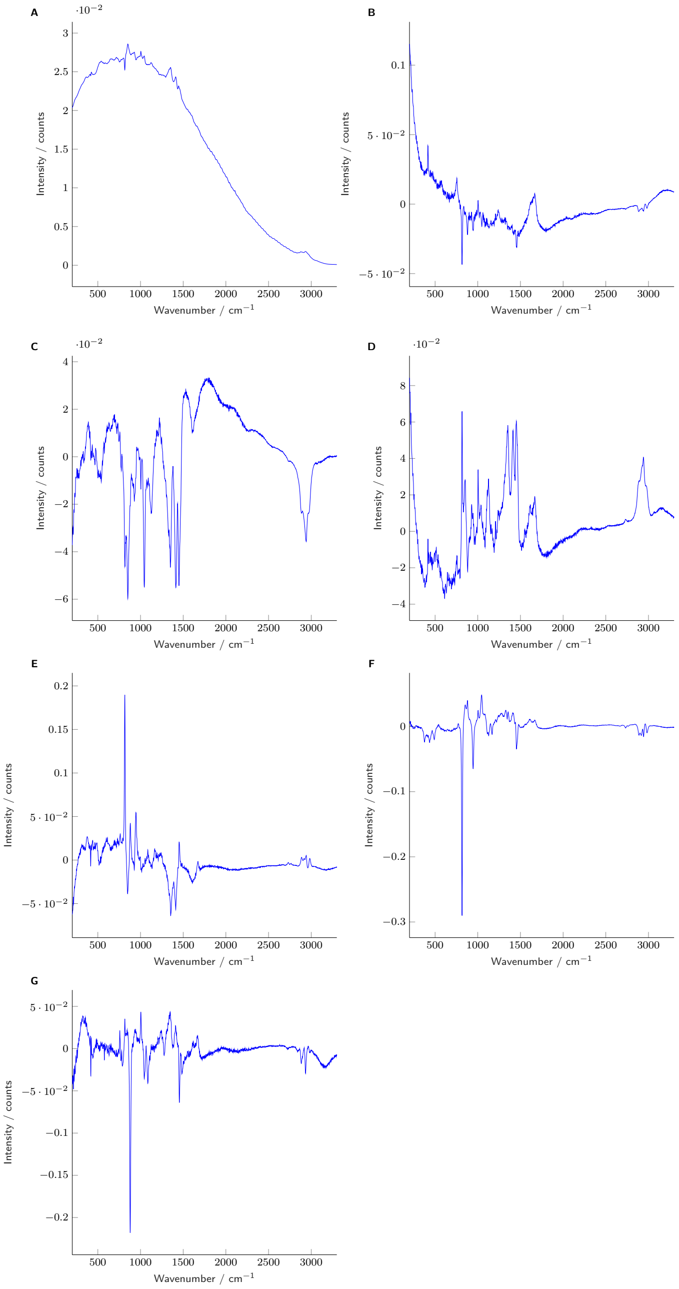

Figure 6.10: Loadings over wavenumbers for the first (A) to seventh (G) LV for the Raman-based PLS model.

6.8 Appendix: Linearity of the Base PLS Models

Figure 6.11 shows the predicted mAb concentration over the observed/measured mAb concentration for all runs of the UV- and Raman-based PLS models. Predicted and observed mAb concentrations show for both models a mostly linear relationship. For the Raman-based PLS model, the validation run shows an offset and a deviation from the linear behavior above 1.9 g/L. Other deviations from the linear relationship could be possibly caused by the off-line analytic due to carry-over between samples.

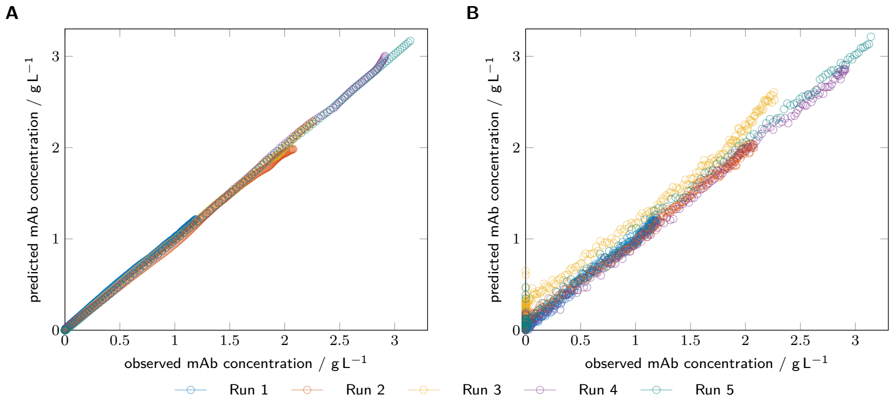

Figure 6.11: Predicted mAb concentration by the PLS model over the measured (observed) mAb concentration for (A) the UV-based PLS model and (B) the Raman-based PLS model.

6.9 Appendix: Output of the Convolutional Layers

Figure 6.12 show the output of the convolution layer of the UV-based CNN colored after mAb concentration. The output of the convolutional layers have a resemblance to the loadings of the UV-based model, which could be caused by similar function.

The output of the convolution layers of the Raman-based CNN, colored after mAb concentration, are depicted in Figure 6.13. Overall, the output of the convolutional layers seem noisy . Due to the wavelet initialization, no strong background effect is visible in the output.

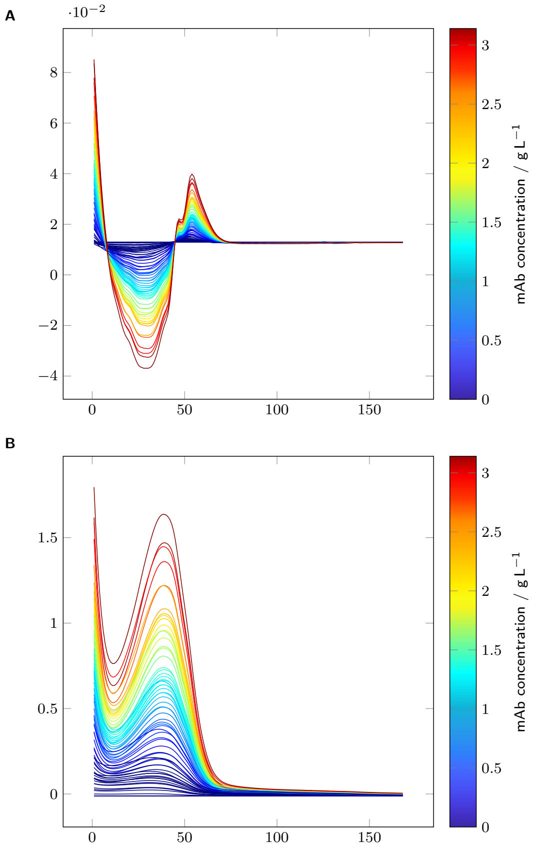

Figure 6.12: The output of the two convolutional layers (A-B) are colored according to the mAb concentration.

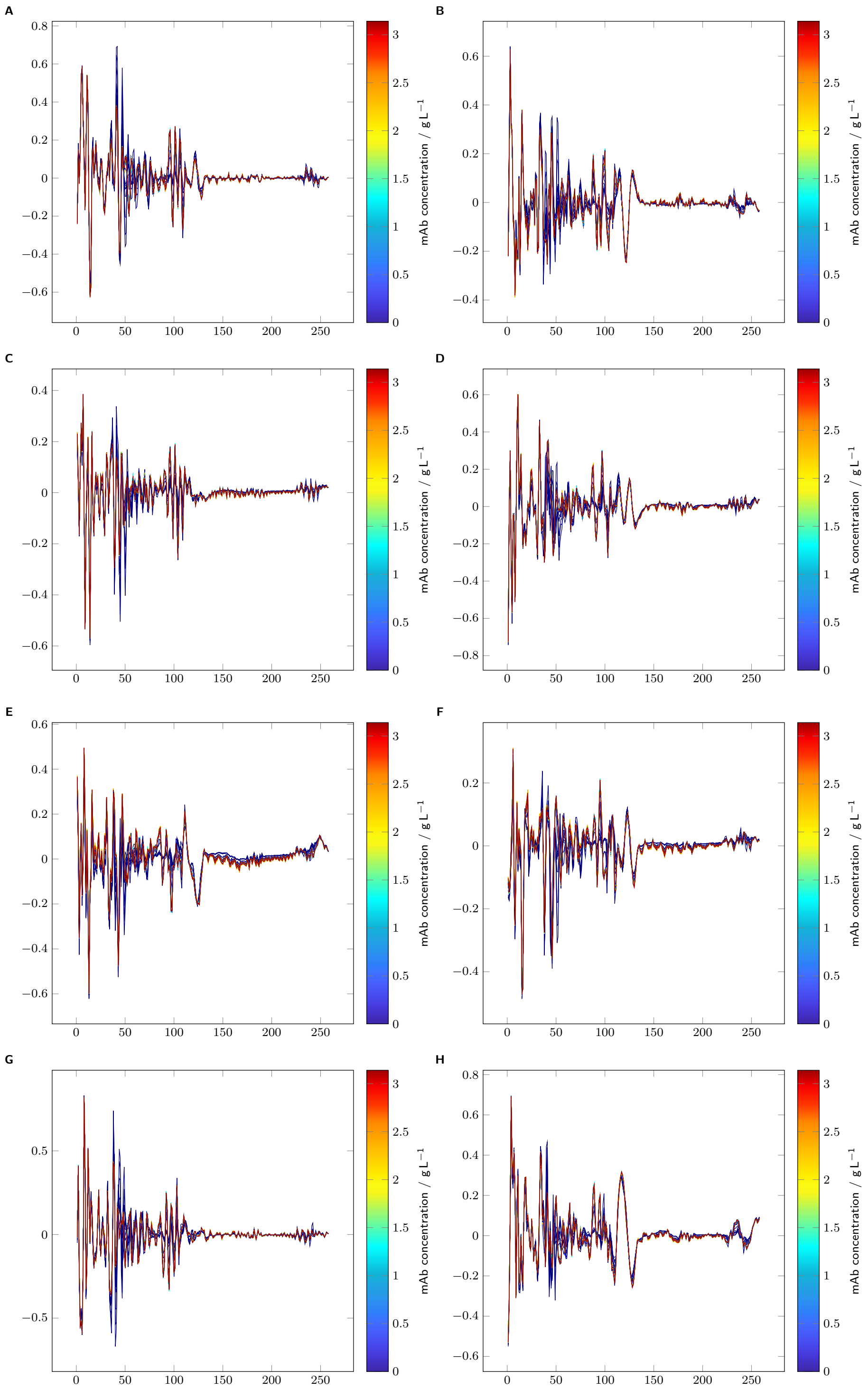

Figure 6.13: The output of the eight convolutional layers (A-H) are colored according to the mAb concentration.

7

Multi-attribute PAT for UF/DF of Proteins—Monitoring Concentration, Particle Sizes, and Buffer Exchange

Laura Rolinger[*,1],Matthias Rüdt[*,1], Juliane Diehm[1],Jessica Chow-Hubbertz[2], Martin Heitmann[2], Stefan Schleper[2], Jürgen Hubbuch[1]

[*] Contributed equally

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

[2] Sanofi-Aventis Deutschland GmbH, Frankfurt am Main, Germany

abstract

UF/DF plays an important role in the manufacturing of biopharmaceuticals. Monitoring critical process parameters and quality attributes by PAT during those steps can facilitate process development and assure consistent quality in production processes. In this study, a lab-scale CFF device was equipped with a VP UV/Vis spectrometer, a light scattering photometer, and a microLDS. Based on the measured signals, the protein concentration, buffer exchange, apparent molecular weight, and hydrodynamic radius were monitored. The setup was tested in three case studies. First, lysozyme was used in a UF-Diafiltration (DF)-UF run to show the comparability of on-line and off-line measurements. The corresponding correlation coefficients exceeded 0.97. Next, urea-induced changes in protein size of Glucose Oxidase (GOx) were monitored during two DF steps. Here, correlation coefficients were $\geq$0.92 for SLS and DLS. The correlation coefficient for the protein concentration was 0.82, possibly due to time-dependent protein precipitation. Finally, a case study was conducted with a mAb to show the full potential of this setup. Again, off-line and on-line measurements were in good agreement with all correlation coefficients exceeding 0.92. The protein concentration could be monitored in-line in a large range from 3 g/L to 120 g/L. A buffer-dependent increase in apparent molecular weight of the mAb was observed during DF, providing interesting supplemental information for process development and stability assessment. In summary, the developed setup provides a powerful testing system for evaluating different UF/DF processes and may be a good starting point to develop process control strategies.

7.1 Introduction

PAT has been an area of active research in chemical industry for several decades [62]. However, in biopharmaceutical DSP, the development and implementation of PAT tools has only recently received more attention [17, 64]. Up to the present, most research in DSP focused on monitoring and control of chromatographic steps [18–20, 67, 72, 73, 123, 124, 169]. Fewer studies evaluated PAT for CFF, even though it is generally applied at least once during the production process [57]. CFF is implemented in DSP for multiple reasons. Before chromatography, CFF may be used to reduce the process volume and to improve the process performance by shifting towards a more favorable region of the adsorption isotherm [8]. At the end of most DSP processes stands a UF/DF step , usually implemented as CFF, to adjust the final formulation and protein concentration. CFF thus builds an essential part of most biopharmaceutical production processes.

During development, purified material, especially for the final UF/DF step, is expensive and only available in limited quantities. In-line measurements of quality attributes and process parameters help to diminish material consumption by reducing sampling and off-line analytics [62]. Furthermore, as human intervention is reduced, a common source of errors is depleted. PAT can also help to improve process understanding since higher measurement frequencies are generally achieved compared to standard off-line analytics [62]. In manufacturing, UF/DF steps are often used to adjust the product to a higher concentration than targeted and based on off-line analytical results the final concentration is adjusted in a subsequent step. Additionally, contaminants that might be introduced by this step, like aggregates or sub-visible particles, cannot be depleted afterwards. A PAT-based process control allows to counteract deviations or can be used to streamline the formulation process.

Previously, monitoring of a DF step was reported by the implementation of a simple pH probe [287]. In the studied application, the pH correlated with the depletion of a contaminant and thus provided a straight-forward means for effective process monitoring. Despite being simple, the approach suffers a number of drawbacks. As the authors themselves pointed out, pH probes are prone to inaccuracies. Furthermore, the correlation of pH to the concentration of a contaminant is specific to the given process and not generally applicable to e.g. the Gibbs-Donnan-effect [288]. Recently, the application of VP UV/Vis spectroscopy for single-pass UF/DF has been reported [289]. The published data however focused on a mechanistic understanding of the process and did not aim for a PAT solution. Off-line sample analytics were used for data confirmation. Monitoring of a full UF/DF process including in-line and real-time measurements to monitor the product concentration, particle formation, and buffer exchange has not been achieved so far.

In this study, a flexible setup for monitoring a wide range of different UF/ DF processes was developed. The setup consists of a lab-scale CFF device equipped with an in-line VP spectrometer and an on-line measurement loop with a light scattering photometer as well as a microLDS. Sensor signals were processed as shown in Figure 7.1. VP UV/Vis spectroscopy provided information on the protein concentration in a large dynamic range and on changes in the tertiary structure of the protein. Viscosity/density measurements allowed to follow DF progress and buffer exchange. Furthermore, based on the scattered-light intensity and the protein concentration, an apparent molecular weight was calculated. DLS measurements in conjunction with the process fluid viscosity were used to monitor the hydrodynamic radius of the product. This setup was applied in three case studies to evaluate its performance in different processes.

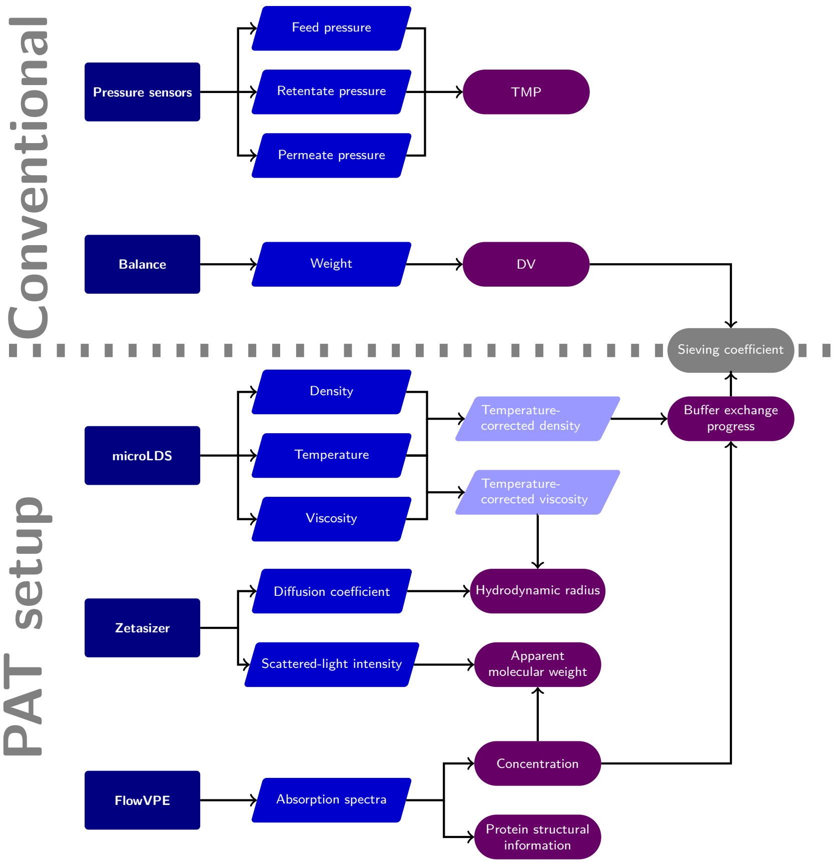

Figure 7.1: The information flow is pictured to show how the different signals are processed. The different sensors are shown as rectangular boxes in dark blue. Measurement signals are denoted as trapezoids. Derived signals are denoted by rounded rectangles. Sieving coefficients were not evaluated in this article.

7.2 Materials and Methods

7.2.1 UF/DF experiments

Experimental Setup

The custom made setup from [144] was adjusted for UF/DF experiments with higher protein concentrations. Figure 7.2 shows the setup as a Piping and Instrumentation Diagram (P&ID). A KrosFlo KRIIi CFF unit (Spectrum Labs, Rancho Dominguez, US) was equipped with a FlowVPE VP UV/Vis spectrometer (C Technologies, Bridgewater, US) and a T-piece with injection plug (Fresenius Kabi, Bad Homburg, DE) placed after the retentate reservoir of the CFF unit for representative concentration determination and for drawing samples for off-line analytics. A Topolino magnetic stirrer (IKA Werke GmbH & Co. KG, Staufen im Breisgau, DE) and a stir bar ensured homogeneous mixing in the retentate reservoir. The retentate reservoir was modified with two additional Polyether Ether Ketone (PEEK) capillaries (GE Healthcare, Chalfont St Giles, GB) to supply the on-line measurement loop with liquid from the process. In the direction of flow, the on-line measurement loop consisted of a peristaltic pump (Minipuls 3, Gilson, France) controlled via a NI USB-6008 data acquisition device (National Instruments, Austin, US), a 0.5 µm particle retention Minisart glass fiber syringe filter (Sartorius Stedim Biotech, Göttingen, DE), a non-bypass version of a flow through microLDS (Integrated Sensing Systems, Inc, Ypsilanti, US), a Zetasizer Nano ZSP photometer (Malvern Panalytical, Herrenberg, DE) with 10 mm path length ZEN0023 quartz flow cuvette (Hellma Analytics, Müllheim, DE), and a FR-902 flow restrictor (GE Healthcare). The particle retention was implemented to prevent the clogging of the microLDS and improve the signal-to-noise ratio of the Zetasizer by trapping air bubbles.

Lysozyme

Lysozyme was purchased from Hampton Research (Aliso Viejo, US). The UF/DF experiment was conducted with a modified Polyethersulfone (mPES) hollow fiber membrane module (3 kDa cutoff, 20 cm^2 membrane area) from Spectrum Labs. 50 mM acetate buffer (Merck, Darmstadt, DE) at pH 5 was used to solubilize the protein and as DF buffer. All buffers were filtered through 0.2 µm Cellulose Acetate (CA) filters (Sartorius Stedim Biotech). The 150 mL lysozyme stock solution at 10 g/L was filtered through a 0.2 µm Polyethersulfone (PES) syringe filter (VWR International, Darmstadt, DE). Before the experiment, the system was run for 5.5 min with an open back-pressure valve to equilibrate all compartments at a feed flow of 45 mL/min. Then, the TMP was set to 1.5 bar. The UF/DF experiment consisted of an UF step to approximately 20 g/L, a DF step for three Diafiltration Volumes (DVs) before concentrating again to 40 g/L. This process scheme was chosen as it is often applied in industry to reduce the DF buffer consumption.

GOx

GOx was purchased from Sigma Aldrich (St. Louis, US) and the mPES hollow fiber membrane module (3 kDa cutoff, 20 cm^2 membrane area) from Spectrum Labs. GOx was dissolved in 10 mM sodium phosphate buffer (VWR) at pH 6.5, which was used as UF buffer, to a concentration of 10 g/L.

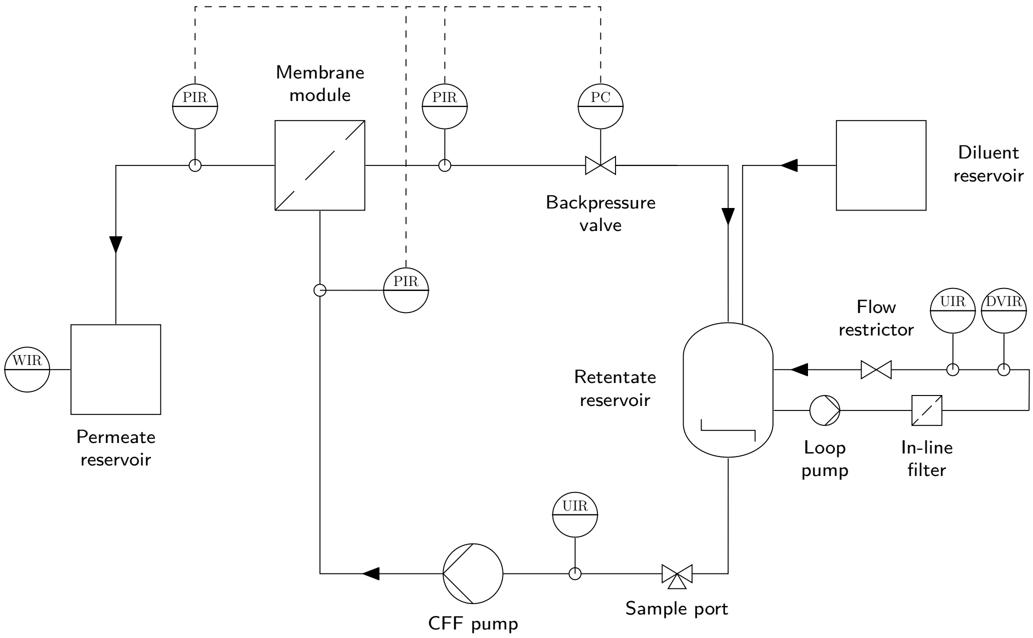

Figure 7.2: Piping and instrumentation diagram of the experimental setup. At the bottom right, the on-line measurement loop is shown. The remaining piping is required for the CFF. All sensors are connected to a computer for capturing the data centrally. Electronic communication lines are indicated by dashed lines. The letters indicate: C control, D density, I indicate, P pressure, R record, U multivariable, V viscosity, W weight.

As DF buffer, 6 M urea was dissolved in the ultrafiltration buffer. The TMP was set to 1.5 bar and the feed flow rate was set to 45 mL/min. The 30 mL of GOx stock solution were added to the retentate tank. The system was run for 11 min to equilibrate the different compartments before the DF was started. A DF into urea buffer was performed for four DVs, before the DF was continued with equilibration buffer four DVs. This process was chosen to investigate the measurability of buffer induced structural changes with the setup.

mAb

mAb stock solution was provided by Sanofi (Frankfurt, DE). The mAb stock solution (concentration 2.79 g/L) was filtered before use through a Stericup mode of PES with a pore size 0.2 µm (Merck). A Pellicon 3 Cassette with an Ultracel membrane (type C screen30 kDa cutoff, 88 cm^2 membrane area) in a Pellicon Mini Cassette Holder was used (both Merck). The system was run with an open backpressure valve and a feed flow rate of 45 mL/min for 5.5 min to equilibrate all compartments. To start the process, the TMP was increased to 1.5 bar. In a first UF step, the concentration was raised to 17 g/L. Next, during an eight DVs DF step, the buffer was exchanged to the formulation buffer. A second UF step concentrated the product to 120 g/L. The process is based on the production process of the mAb.

7.2.2 Data Acquisition and Analysis

During experiments, all integrated sensors and devices communicated with and were controlled by a custom-made application developed in MATLAB (version R2017a, The Mathworks, Natick, US) and adapted from [144]. Besides connecting the devices and starting and stopping measurements, the application gathered the signals from the integrated sensors (cf. Figure 7.1) and calculated quality attributes and process parameters, as explained in the following sections. Communication and control were performed through software libraries provided by the different instrument manufacturers. The pump of the on-line loop was timed to the measurements of the light scattering photometer, such that the fluid in the loop had been replaced before new batch measurements were started. Signals were displayed on the Graphical User Interface (GUI) and stored on the hard drive with a time stamp. Data acquisition and analysis of density and viscosity measurements, light scattering data and UV/Vis measurements were performed as described below.

UV/Vis Absorbance Measurements and Processing

UV/Vis slope spectra were recorded from 270 nm to 320 nm for lysozyme and mAb with a resolution of 1 nm and from 270 nm to 460 nm with a resolution of 2 nm for GOx. For concentration calculations, the absorbance at 280 nm was scatter-corrected by subtracting the absorbance at 320 nm. To obtain information on the local environment around the aromatic amino acids, the spectra were smoothed with a moving average over 20 measurements and second derivatives were calculated with a Savitzky-Golay filter [290] of order 5 with a 9-point window [34]. The resulting second derivative spectra were interpolated with a cubic spline to a resolution of 0.01 nm. From the interpolated spectra, the location of the minimum between 290 nm to 295 nm was used as a measure for the mean solvatization of tryptophans [34, 291]. The solvatization of tyrosines was assessed by the a/b-ratio (r_{ab}) which is calculated by dividing the trough-to-peak distance near 285 nm by the trough-to-peak distance near 294 nm [292].

Temperature Correction of Viscosity and Density Measurements

In general, the viscosity η and density ρ of solutions are affected by the buffer components, protein concentration, and temperature. For the obtained data, this was important as the used microLDS dissipates a noticeable amount of heat into the measured liquid. To obtain comparable results, the measured viscosity and density were corrected to a standard process temperature yielding η_{T_0} and ρ_{T_0}, respectively. As the temperature differences were relatively small ($\Delta T \leq 5\mathrm{K}$), it was assumed that the deviations from the ideal solution behavior were neglectable [293–295]. The temperature corrections were thus performed by cross-multiplication for viscosity and density measurements.

$$\eta_{T_0} = \frac{\eta_{\mathrm{water},T_0}}{\eta_{\mathrm{water},T}}\eta \tag{7.1}$$

$$\rho_{T_0} = \frac{\rho_{\mathrm{water},T_0}}{\rho_{\mathrm{water},T}}\rho \tag{7.2}$$

This approach is similar to the temperature correction of the sedimentation coefficient performed in analytical ultracentrifugation [34, 296]. Reference values for the density/viscosity of water were obtained from the National Institute of Standards and Technology (NIST) chemistry webbook [297].

Measurement of the Hydrodynamic Radius

DLS and SLS were measured with the Zetasizer Nano ZSP in batch mode with the Protein Size Standard Operating Procedure (SOP) of the Zetasizer Software (Version 7.12, Malvern Panalytical). The on-line measurement loop was filled with a flow rate of 3 mL/min for 1 min. The measurements were performed in stopped flow at the fixed angle of 173°, a laser wavelength of 633 nm and at a temperature of 25 °C. Each measurement duration was 10 s. Measurements were repeated three times to obtain one data point. Before each UF/DF experiment, the feed sample was measured to determine the best attenuator setting. The attenuator was adjusted such that the count rate was in the range of 200 kc/s to 500 kc/s for the experiments.

DLS measurements yielded the autocorrelation of the light intensity over time $\langle g(\Delta t)\rangle$. The mean diffusion coefficient D_0 was extracted by the method of cumulants [89]. Based on the Einstein relation [293], the hydrodynamic diameter d_h of the particles in solution can be estimated:

$$D_0 = kT/f_0, \tag{7.3}$$

$$f_0 = 3\pi\eta d_h \tag{7.4}$$

where k is the Boltzmann constant, T is the absolute temperature, and f_0 is the friction coefficient for spherical particles with a hydrodynamic diameter. If the hydrodynamic diameter is calculated from the diffusion coefficient obtained by the method of cumulants, an intensity-weighted harmonic mean hydrodynamic diameter is obtained which is called the z-average [298]. It is worth noting that the friction coefficient for classical dilute measurements is a function of the diluent viscosity η. Equations 7.3 and 7.4 are restricted to compact, diluted and significant larger molecules than the surrounding solvent. When using not ideally diluted samples and backscattering, the bulk solution viscosity η_{T_0} (as obtained from Equation 7.1) should be used instead of the diluent viscosity to account for restricted diffusion [299, 300]. Even-though particle-particle interactions are excluded from this theory, the equation in general holds true for globular proteins much smaller than the incident light wavelength [301].

Measurement of the Apparent Molecular Weight

For the calculation of the apparent molecular weight, the scattered-light intensity obtained from the Zetasizer was used. A refractive index increment $\frac{\delta n}{\delta c}$ of 0.185 mL/g was used for lysozyme and mAb. Due to the carbohydrate content of GOx, a refractive index increment of 0.177 mL/g was used [302]. A calibration was necessary for the calculation of the apparent molecular weight from the scattered-light intensity to obtain the second virial coefficient A_2 for each protein in the appropriate buffer. These measurements were done according to the procedure for off-line analytics by light scattering.

7.2.3 Off-line Analytics by Light Scattering

The off-line DLS and SLS measurements were carried out similar to the on-line measurements, except that each sample was measured with three runs with 15 sub measurements. For the calculation of the second viral coefficient A_2, concentration series from 1 g/L to 100 g/L with 8 calibration points for lysozyme, 1 g/L to 25 g/L with 7 calibration points for GOx in equilibration and DF buffer, and 1 g/L to 100 g/L with 7 calibration points for mAb buffer were prepared and measured with the off-line analytics method. Additionally, concentration series with 3 different mixtures of equilibration and DF buffer were carried out for GOx and mAb to evaluate the influence of the buffer on A_2. A Debye plot was drawn for each calibration curve and the gradient of the regression was divided by two and used as A_2 (see Appendix 7.5).

7.2.4 Off-line Analytics by SEC

Samples were analyzed on a Vanquish Flex Binary HPLC system (Thermo Fisher Scientific, Wilminton, US) by Size-Exclusion Chromatography (SEC). The system consisted of a Binary Pump F, Split Sampler FT, Column Compartment H and a Diode Array Detector HL. Chromeleon Version 7.2 SR4 (Thermo Fisher Scientific) was used to control the HPLC.

Lysozyme and GOx

The run duration was 7.5 min with a flowrate of 0.3 mL/min. 20 mM sodium phosphate and 500 mM sodium chloride (Merck) at pH 7.0 was used as a mobile phase. 2 µL were injected into a 4.6 mm × 150 mm TSKgel SuperSW mAb HTP column (Tosoh Bioscience GmbH, Griesheim, Germany). Samples were analyzed in duplicates.

mAb

The run duration was 15 min with a flowrate of 0.3 mL/min. 50 mM sodium phosphate and 300 mM sodium chloride at pH 7.0 was used as a mobile phase. 2 µL were injected into a 4.6 mm × 300 mm ACQUITY UPLC BEH200 SEC column with a pore size of 1.7 µm (Waters Corporation, Milford, Massachusetts, US). Samples with concentrations higher than 40 g/L were diluted 10-fold. All Samples were analyzed in duplicates.

7.3 Results and Discussion

In this study, a CFF set-up was developed to monitor biopharmaceutical UF/DF processes. The setup allows to monitor product concentration, buffer exchange, changes in apparent molecular weight, and changes in hydrodynamic radius. UF/DF processes with three proteins were performed to test the versatility of the setup. Lysozyme was selected due to its low aggregation tendency. A process with a DF and two concentration steps was performed. GOx was studied due to the possibility to disassemble the protein into its two subunits and to reassemble them subsequently. Finally, the CFF of a mAb in a standard UF/DF process was studied.

7.3.1 In-line Concentration Measurements by VP UV/Vis Spectroscopy

For monitoring the protein concentration in-line, the scatter-corrected absorbance at 280 nm from the VP spectrometer was used. In Figure 7.3, the measured protein concentrations are compared to the results obtained from off-line SEC analysis. In all three processes, only small absolute deviations occurred between the two measurement methods (cf. Table 7.1).

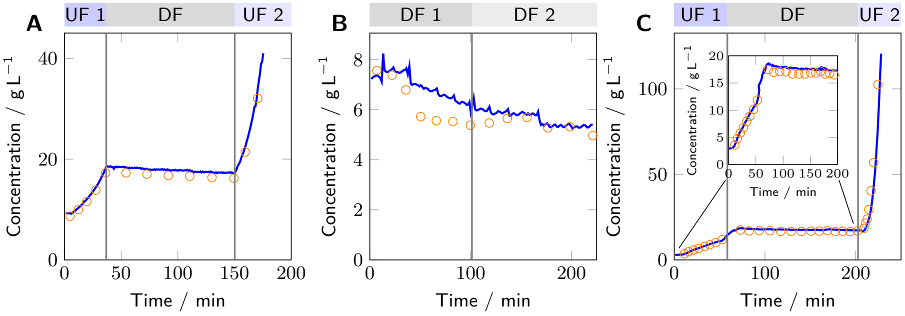

Figure 7.3: The total protein concentration is shown as measured by the in-line FlowVPE VP spectrometer (blue lines) and off-line SEC (orange circles). The different subplots show the results for lysozyme (A), GOx (B), and mAb (C). The insert in subplot C shows a magnification of the first UF and DF steps of the process.

Table 7.1: RMSE and correlation coefficients of in- and on-line measurements compared to off-line analytics.

	Concentration		z-average		M_w	
	RMSE / g/L	R	RMSE / nm	R	RMSE / kDa	R
Lysozyme	1.01	0.9996	4.08	0.9724	0.4	0.9984
GOx	0.71	0.8247	2.06	0.9224	1.2	0.9541
mAb	0.82	0.9982	11.46	0.9240	7.0	0.9862

This good agreement of the methods is remarkable as protein concentrations covered a range from 3 g/L up to 120 g/L. The most pronounced deviations were observed for GOx in the presence of urea (cf. Figure 7.3B). This was presumably due to the decreased long-term stability of the protein solution, i.e. the protein partially precipitated before the off-line analysis could be performed. Similar observations were previously reported in literature [303]. Under such circumstances, in-line measurements provide more reliable information on the process than off-line analytics as the measurement is performed directly on the process liquid without additional time delays.

Given the results obtained in the three case studies, VP UV/Vis spectroscopy provides a powerful tool for quantifying proteins in-line in (near-)real time during UF/DF processes.

7.3.2 On-line Density and Viscosity Measurements

The integrated microLDS provided information on the current density and viscosity of the process liquid. Both density and viscosity are of interest for computing process parameters and quality attributes such as the buffer exchange progress or the hydrodynamic radius (cf. Figure 7.1). As the microLDS does not maintain a constant temperature, the measured viscosity and density were mathematically corrected to the standard temperature (see also Section 7.2.2).

Figure 7.4 shows the observed density profiles during the different UF/DF processes. In all cases, differences between the raw densities and the temperature-corrected densities remained small. The density during the UF/DF process of lysozyme (Figure 7.4A) reflected the trends already observed by UV/Vis spectroscopy. This is due to the fact that in this process the buffer composition did not change and thus the density was only affected by the protein concentration.

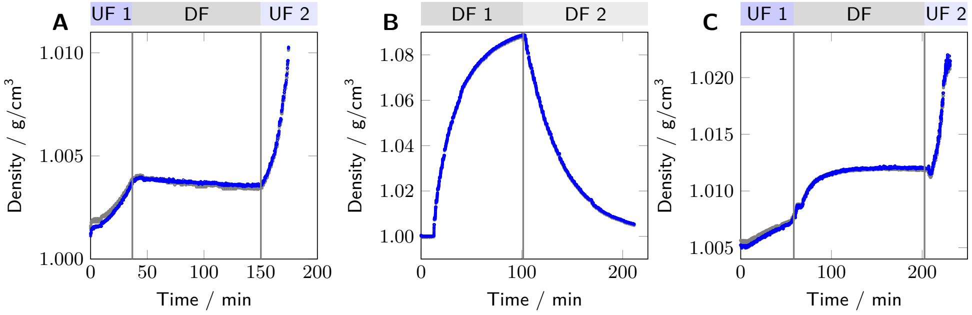

Figure 7.4: On-line density measurements were obtained from the microLDS. Since the temperature of the microLDS drifted during the measurements, a temperature correction was performed (c.f. Section 7.2.2). The raw densities are shown in gray. The temperature-corrected densities are shown in blue. The different subplots show the results for lysozyme (A), GOx (B), and mAb (C).

In contrast, the density during the GOx DF was mainly affected by the changing buffer composition, i.e. the changing urea concentration (Figure

7.4B). As the protein concentration remained almost constant, the effect of the protein concentration influenced the trends only negligibly. Interestingly, the typical exponential decay of the density towards a new baseline could be observed, as it is expected for a DF process. Furthermore, it seems clear that the buffer exchange was not complete after the performed DF steps with four DVs since the density over time still displayed a noticeable slope. For a complete buffer exchange, the DF step would have to be further extended.

The mAb process (Figure 7.4C) combined UF and DF steps. It was thus expected that the density trends would be influenced by both protein concentration and buffer concentration. Indeed, the two UF phases were dominated by density changes due to the changing protein concentration while the DF step showed an exponential growth towards a new baseline typical for a buffer exchange.

Based on the above observations, density seems to be a suitable detector for observing the progress of DF steps. Relying on a golden batch approach, it may be interesting to define a density corridor for a given process within which it should be run. It is worth to keep in mind that density is only a univariate response and the microLDS provides measurements only with limited sensitivity. Thus, the exchange of a specific buffer component cannot be observed selectively and to the concentration levels which are considered a complete depletion. However, as protein-interaction with the buffer components would change the expected decrease in density, time-series analysis could be used to monitor deviations from the permeability of the buffer components through the membrane. Additionally, other orthogonal sensors, e.g. conductivity, pH, or Raman, could further narrow down the process corridor.

Regarding viscosity, similar trends as for density could be observed for all three processes (Figure 7.5). Here, the effects of the temperature correction were more pronounced. This went furthest for the lysozyme process (Figure 7.5A) where the viscosity trends did only show the expected behavior after temperature correction.

In general, viscosity of any biopharmaceutical solution is an important attribute for its manufacturability and ‘syringeability’ [304]. Especially at very high concentrations, the viscosity of mAb solutions may be problematic [305]. The final concentration of biopharmaceuticals are normally reached by a CFF processes. On-line viscosity measurements thus not only provide important information for calculating the hydrodynamic radius but are also of interest for assessing the manufacturability of a protein during the development of a CFF step.

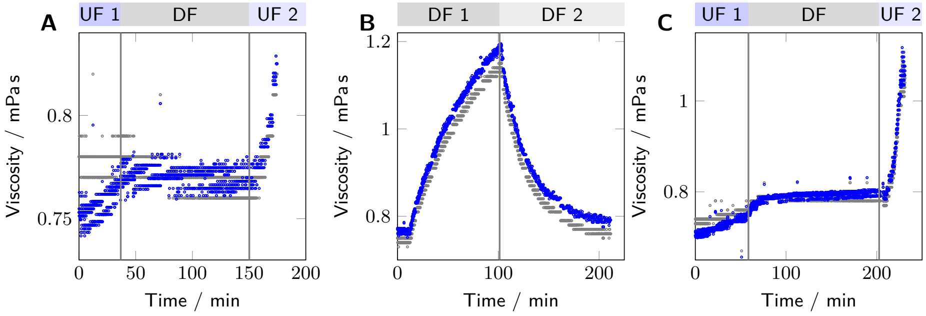

Figure 7.5: On-line viscosity measurements were obtained from the microLDS sensor. Since the temperature of the microLDS drifted during the measurements, a temperature correction was performed (c.f. Section 7.2.2). The raw viscosity is shown in gray. The temperature-corrected viscosity is shown in blue. The different subplots show the results for lysozyme (A), GOx (B), and mAb (C). The stepped behavior of the signal is due to the digital signal accuracy of the microLDS.

7.3.3 On-line Apparent Molecular Weight and Z-Average Measurements by Light Scattering

Figure 7.6 shows the observed z-average during the different UF/DF processes. As described in Section 7.2.2, all measured z-averages were calculated with the measured viscosity to prevent biased z-average estimations due to changing solution compositions.

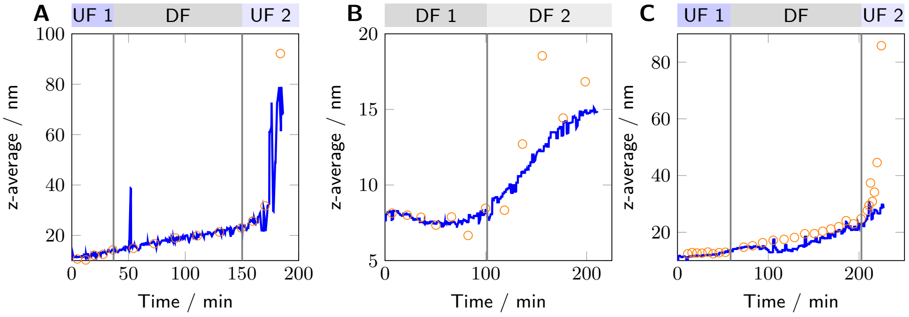

Figure 7.6: The on-line z-average measurements (blue lines) and off-line measurements (orange circles) were obtained from the Zetasizer. The different subplots show the results for lysozyme (A), GOx (B), and mAb (C).

For lysozyme (Figure 7.6A), an elevated initial z-average of 8.95 nm was observed compared to a z-average of 3.9 nm reported in literature [306]. A natural explanation for the elevated z-average could be the presence of larger sub-micrometer particles in the process stream [307]. The intensity-weighted size distribution of the DLS measurements indeed showed a main peak at 3.55 nm and a smaller peak around 127 nm (data not shown). Those sub-micrometer particles can either consist of non-biological particles (e.g. membrane abrasion) or protein particles [307]. In a volume-weighted size distribution, the peak around 127 nm is negligible. This emphasizes the advantages and disadvantages of DLS. Particles scatter light depending on the diameter to the power of six [307]. Therefore, already a very small amount of large particles can be detected, but these particles can also blind the measurement to any comparably small particles [84]. Since submicrometer particles are precursor for further aggregation, it is however important to monitor the amount of those particles to prevent later stability issues of a batch [84].

During the GOx DF process, the protein was disassembled by a chaotropic agent. In its native form, GOx is a homodimer with a molecular weight of 160 kDa [308]. The native holoprotein carries two noncovalently linked Flavin-adenine Dinucleotides (FAD) molecules per homodimer. When adding denaturing agents (here urea) to GOx solutions, FAD dissociates from GOx [309]. From 2 M to 6 M urea, unfolding of GOx is observed accompanied by the loss of the cofactor FAD. Completed unfolding occurs at concentrations greater than 6 M urea. The denaturation does not seem to be reversible by just lowering the urea concentration [310]. The initial z-average of 7.87 nm (Figure 7.6B) is in good agreement with literature [311]. After starting the DF, the z-average decreased until a minimum (7.23 nm) was reached after 71.24 min. This corresponds to a concentration of urea of 5.5 M calculated by the density sensor data. At high urea concentrations GOx is not only dissociating, but also denaturating. According to [312], the hydrodynamic radius of denaturated GOx monomers is larger than the one of the native form. Because dissociation and denaturation are occurring simultaneously, this could explain the little change in z-average. Reducing the concentration of urea again leads to an increase in z-average. Literature indicates that after the dissociation of FAD from the enzyme, the apoenzyme tends to form aggregates [303].

Figure 7.6C shows the z-average of a mAb during the process. The z-average was continuously increasing during the process. Interestingly, in the second UF step from 214.6 min on, the results from off-line and on-line measurements started to drift apart as the z-average determined by off-line measurement was increasing faster than the z-average determined by on-line measurement. It was reported in literature that concentration-dependent aggregation is also time-dependent [313]. This could explain the discrepancy between on-line and off-line measurements as the off-line measurements were performed the next day and additional aggregation may have taken place.

Figure 7.7 shows the observed scattering intensity and the calculated apparent molecular weight during the different UF/DF processes. For lysozyme (Figure 7.6A), an apparent molecular weight of 16.8 kDa at the beginning of the process was observed. This is in agreement with the elevated z-average . During the whole process a steady increase in the apparent molecular weight was observed with a similar trend as the z-average. The scattering intensity over the process followed the concentration profile with an underlying steady increase due to the changing particle sizes. As the scattering intensity is proportional to the concentration and molecular weight of the scatterer, the variations due to concentration effects or the mean molecular weight are not distinguishable without additional information. If only the protein aggregation is of interest, the apparent molecular weight has a better interpretability, because the concentration-dependency of the process is removed in comparison to the raw light scattering signal.

The measured apparent molecular weight of GOx in the beginning was 132.9 kDa (Figure 7.7B) while the molecular weight specified by the vendor is 160 kDa. One reason for the differences could be the used refractive index increment $\frac{dn}{dc}$ of 0.177 which was obtained from literature. During the process, the concentration of GOx decreased from 7.23 g/L to 5.40 g/L which made the changes in molecular weight the main driving force of changes to the scattered-light intensity. Therefore, the apparent molecular weight correlates with the scattered-light intensity. Differences between off-line and on-line measurements were observed. In contrast to the z-average measurements, a distinct decrease in apparent molecular weight was observed with increasing urea concentration. The apparent molecular weight decreased to 50 kDa at the highest urea concentration. This is lower than the expected molecular weight of the monomer, even when considering that the refractive index increment might be off due to the carbohydrate content of GOx. The results illustrate that the calculation of the apparent molecular weight gives a trend of the process performance. Determining the absolute molecular weight is however challenging because of the many measurement parameters (e.g. refractive index increment, refractive index of the solution) that need to be determined.

The measured apparent molecular weight of the mAb was 105.6 kDa in the beginning of the first UF step as depicted in Figure 7.7D, which is a zoom of Figure 7.7C . Again, this apparent molecular weight is lower than expected for a mAb. However, the starting material contained fragments. Furthermore, the refractive index increment $\frac{dn}{dc}$ was not optimized for the mAb as the degree of glycosylation of the mAb was unknown. During the first UF, the apparent molecular weight increased to 132.9 kDa (Figure 7.7D). During the DF the apparent molecular weight increased from

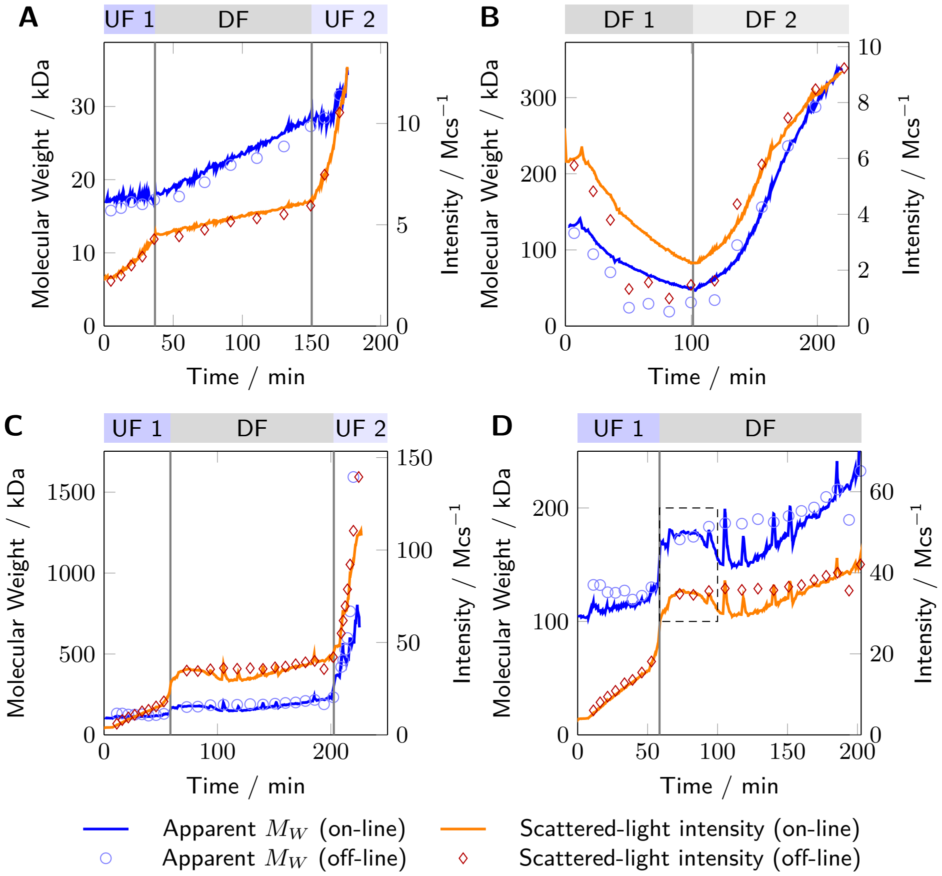

Figure 7.7: The on-line scattering intensity (orange lines) and off-line measurements (orange circles) were obtained from the Zetasizer. From these measurements, an apparent on-line molecular weight (blue lines) and offline molecular weight (blue circles) were calculated. The different subplots show the results for lysozyme (A), GOx (B), the whole mAb process (C), and a zoom on the UF 1 and DF steps of the mAb process (D). In subplot D, a dashed rectangle highlights a transient increase in the on-line measured signals in the beginning of the DF step.

169.4 kDa to 227.8 kDa. The increase was not monotonous. Instead, an initial overshoot followed by a slight decrease was observed (highlighted by a dashed rectangle) before a second phase of steady increase. The overshot could be caused by the change in buffer conditions which could cause an undesired aggregation tendency in the product at certain buffer compositions. As aggregation should be minimized during formulation, such information should be considered for choosing appropriate buffer conditions. During the DF, due to the changing buffer, the second virial coefficient could change as well. This information could be used to understand the protein-interaction and enhance formulation buffer development. Additionally, the sample drawing in the on-line loop seemed to cause a peak and disturbance of the measurement, possibly due to air bubble entry or changes in flow. During the last UF step, the apparent molecular weight increased to 672.8 kDa. Due to the calculation of the apparent molecular weight with the protein concentration, the signals were noisier than the raw scattered-light intensity. Furthermore, the process fluid was changing very quickly in the second UF step. Thus, already slight inaccuracies in the delay time between the in-line protein concentration measurement and the on-line light scattering measurements could have a noticeable influence on the apparent molecular weight calculation. Both effects may explain the discrepancies to off-line measurements.

7.3.4 Comparison of SEC to Light Scattering Measurements

For lysozyme, no aggregation was observed by SEC (data not shown). This is interesting, because Figure 7.6A and Figure 7.7A show an increase in z-average and apparent molecular weight probably due to sub-visible particle formation. Similar observations have previously already been reported in literature [144]. Thus, on-line light-scattering measurements provide an orthogonal analysis on particle formation in the process stream. For GOx, dimers and monomers could not be selectively quantified by SEC due to the formation of a large number of fragmented product-related impurities. A direct comparison of SEC and light scattering measurements was therefore not possible.

Figure 7.8 shows the trend of the apparent molecular weight and z-average in comparison to the fraction of aggregates and fragments of the mAb. The fraction of aggregates increases during the first UF from 2.42 % to 3.48 %, while the fraction of fragments increases slightly from 3.60 % to 3.89 %. During DF, the fraction of aggregates slightly decreases to 3.44 % while the fraction of fragments increases to 4.25 %. In general, the off-line fragment concentration seems to be noisier than the off-line aggregate concentration. This is likely related to the accuracy of the reference analytics as the fragment peak is a shoulder on the product peak in the SEC chromatogram. The apparent molecular weight and z-average are slightly increasing during the first UF. During the second UF, the fractions of fragments and aggregates increase to 6.32 % and 5.92 %, respectively. As the aggregates are overall increasing more strongly than the fragments, an increase in apparent molecular weight and z-average is expected as both the molecular weight and z-average are sum signals of all contributing species.

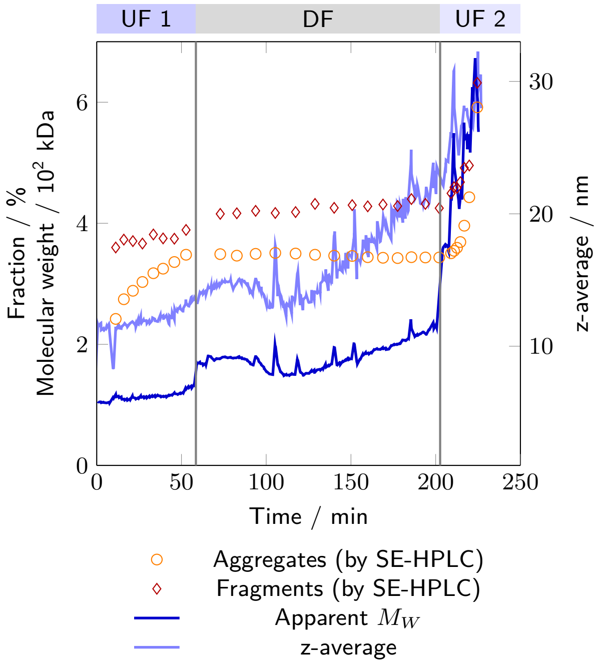

Figure 7.8: Comparison of the apparent molecular weight and z-average to off-line fraction analytics by SEC for the mAb. This Figure compares the on-line data from Figure 7.7C and 7.6C.

During DF, the aggregate and fragment fractions are fairly constant but the apparent molecular weight and the z-average raise from 169.4 kDa to 227.8 kDa and from 13.3 nm to 23.1 nm, respectively. An explanation for this increase could be the formation of sub-visible particles, which are not detected by SEC, but influence both SLS and DLS.

In summary, even though an overall increase in aggregated species can be detected by SEC, SLS, and DLS, the extent of the increases varies for all methods. There is already an abundance of papers discussing the differences between several methods for aggregate quantification [314–316]. This study shows that SLS and DLS measurements can provide on-line information on the formation of sub-visible particles and protein aggregates. SEC analysis on the contrary provides selective and quantitative information on aggregates and fragments while excluding sub-visible particles. As such, the three methods are complementary to each other and all measurements should be taken into account to fully understand the underlying process.

7.3.5 Protein Tertiary Structure Evaluation

A further exciting option of the experimental setup is the acquisition of spectral information up to very high protein concentrations by the VP UV/Vis spectrometer. Based on the protein spectra, information on the solvatization of the aromatic amino acids can be gathered [34]. This approach was implemented for the GOx DF process (Figure 7.9). Based on the a/b-ratio, changes in the solvatization of tyrosines were evaluated. The first DF step consisted of a lag phase where GOx retained the hydrophobic environment around tyrosines. Only in a second phase, the a/b-ratio started to raise indicating an increased solvent exposure of tyrosines. In the second DF step, the a/b-ratio decreased. This shows that the hydrophobic environment around tyrosines was partial restored towards the end of the process.

Contrary to this behavior, the tryptophan minimum displayed a red shift during the whole process after an initial lag phase which translates to a continuous increase in the local hydrophobicity around tryptophans [291]. These results seem to indicate that a hydrophobic core is retained around tryptophans despite the increasing urea concentrations. Furthermore, the red-shift may indicate that hydrophobic interactions of those aromatic amino acids are involved in aggregate formation. To prevent aggregation in a similar process development scenario, it could thus be beneficial to add a detergent which reduces hydrophobic interactions and may prevent aggregate formation [317].

7.4 Conclusion

In this study, a versatile setup for monitoring UF/DF processes was developed and applied to three different proteins and processes. The setup allows to measure—among other values—the protein concentration, density, viscosity, z-average, and the apparent molecular weight. The on-line protein concentration, z-average, and scattered-light intensity were compared to off-line measurements and achieved high correlation coefficients. Results

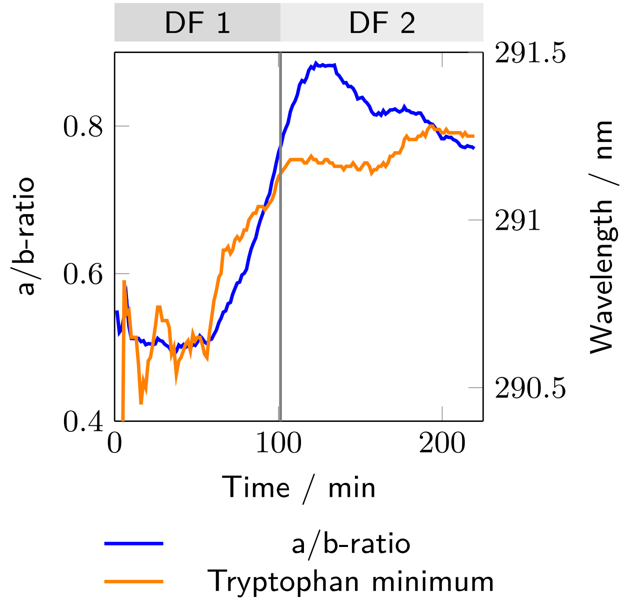

Figure 7.9: Time evolution of the a/b-ratio and the tryptophan minimum over the DF process of GOx. The a/b-ratio shows changes in the mean local polarity around tyrosines, while the mean local polarity around tryptophans is assessed by the position of the local minimum of the second-derivative UV/Vis spectrum around 291 nm.

show that the developed setup provides valuable information on protein concentration, aggregation, particle formation, and the progress of the buffer exchange. The ability to measure those parameters makes this setup interesting for research and development to test different buffer systems and may help to evaluate the Gibbs-Donnan-effect. However, additional orthogonal sensors, like pH, conductivity or Raman, may be needed for DF end-point prediction as density measurements alone are not selective to different buffer components. Additionally, the setup can be used to monitor production processes and detect anomalies instantly and to avoid deviations. Due to the many influence parameters for the calculation of the molecular weight, which are changing during the process, the calculation of an exact molecular weight is challenging. However, the apparent molecular weight could be used to compare between different runs, e.g. to compare sub-visible particle formation. In summary, the developed setup provides a powerful testing system for evaluating different UF/DF processes and may be a good starting point to develop process control strategies.

7.5 Appedix: Debye Plots

Figure 7.10 shows the regressions curves for the calculation of the second virial coefficient and Table 7.2 shows the obtained values. Every measurement of the calibration points consist of the average of three runs. The standard deviation of those three runs is not shown in the graph because they were consistently less than 1%. As mentioned in Section 7.2.3, five concentration series were measured for GOx and mAb to obtain different A_2 for each buffer composition. Due to the instability of GOx in urea buffer and the relatively low concentration, buffer related changes in the A_2 were neglected. Instead the A_2 of GOx in the feed buffer was used for the whole process. For the mAb concentrations of up to 120 g/L were reached during the second UF. Therefore, an extra A_2 for the DF buffer was used because with higher concentrations the A_2 value becomes more important for the calculation of the apparent molecular weight.

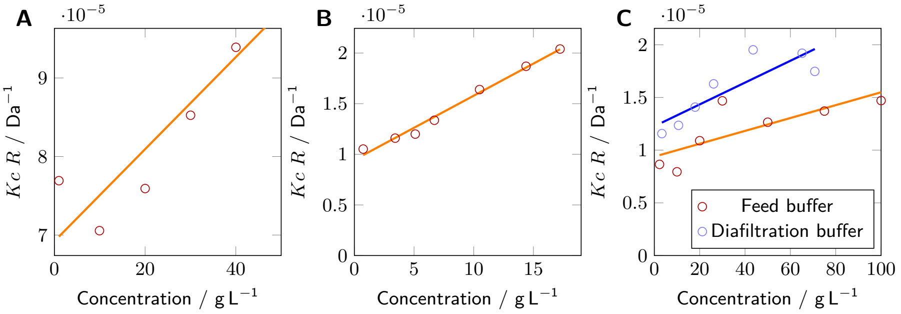

Figure 7.10: For the calcualtion of the second virial coefficient A_2, concentration series for lysozyme (A), GOx (B) and mAb (C) were measured by SLS (circles). The gradient of the regression line was divided by two and used as A_2. For lysozyme and GOx, the concentration series was only measured with feed buffer (blue). For the mAb, a concentration in both the feed (orange) and DF buffer (blue) was measured.

Table 7.2: Second virial coefficients A_2 from Debye plots.

	A_2 / L mol/g^2	
	feed buffer	DF buffer
Lysozyme	2.9e-07	-
GOx	3.2e-07	-
mAb	3.1e-08	5.2e-08

8

Monitoring of Ultra- and Diafiltration Processes by Kalman-filtered Raman Measurements

Laura Rolinger[1,2], Jürgen Hubbuch[1], Matthias Rüdt[1,3],

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

[2] Hoffmann-La Roche AG, Basel, Switzerland

[3] Haute Ecole d'Ingénierie (HEI), HES-SO Valais-Wallis, Switzerland

Abstract

Monitoring the protein concentration and buffer composition during the UF/DF step enables the further automation of biopharmaceutical production and supports Real-time Release Testing (RTRT). Previously, in-line UV and IR measurements have been used to successfully monitor the protein concentration in a large range. The progress of the diafiltration step has been monitored with density measurements and IR. Raman spectroscopy is capable of measuring both the protein and excipient concentration while being more robust and suitable for production measurements in comparison to IR. Regardless of the used spectroscopic sensor, the low concentration of

excipients poses a challenge for the sensors. By combining sensor measurements with a semi-mechanistic model through an EKF, the sensitivity to determine the progress of the diafiltration can be improved. In this study, Raman measurements are combined with an EKF for three case studies. The advantages of Kalman-filtered Raman measurements for excipient monitoring are shown in comparison to density measurements. Furthermore, Raman measurements showed a higher measurement speed in comparison to VP UV measurement at the trade-off of a slightly worse prediction accuracy for the protein concentration. However, the Raman-based protein concentration measurements relayed mostly on an increase in the background signal during the process and not on proteinaceous features, which could pose a challenge due to the potential influence of batch variability on the background signal. Overall, the combination of Raman spectroscopy and EKF is a promising tool for monitoring the UF/DF step and enables process automation by using adaptive process control.

8.1 Introduction

Biopharmaceuticals are an important asset to the modern pharmaceutical industry due to their potential to address diseases that were previously difficult to treat and, from an economical point of view, due to their high retail prices [12, 318]. Biopharmaceuticals are most often produced by genetically modified cells in bioreactors [319]. After the cultivation, the biopharmaceuticals are purified during the Downstream Processing (DSP) to a target purity to allow an administration to patients. The DSP most importantly incorporates centrifugation, chromatography, and filtration steps [320].

Among the listed DSP unit operations, Cross-Flow Filtration (CFF) is used at least once at the end of the production process to set the final protein concentration and transfer biopharmaceuticals into their formulation buffers [57]. The unit operation uses the large hydrodynamic diameter of proteins to retain them in a recycling system, while buffer components, water, and contaminants are forced through a membrane [59]. Typically, the process is performed in multiple steps. First, the biopharmaceutical is concentrated to an intermediate concentration to reduce the initial volume during a first Ultrafiltration (UF) step. Second, a buffer exchange into the formulation buffer is performed during a Diafiltration (DF) step. Normally, the protein concentration remains stable during this step. A preset volume of formulation buffer (e.g. five times the pool volume) is forced over the membrane to ensure a sufficient depletion of the original buffer. In the case of highly concentrated drug substance solutions, a second UF step is subsequently used to concentrate the biopharmaceutical to its target concentration. The second UF step helps to avoid concentration-related gel formation on the membrane during the previous DF step which would decrease the process performance [57]. Additionally, the DF buffer should be designed to reduce the viscosity of the protein solution to reduce the process time of the second UF step [321].

It is common practice during process development and production to rely on mass balances to monitor the progress of the Ultrafiltration/Diafiltration (UF/DF) steps. For example, the DF step is completed if a certain number of DF volumes have been exchanged. Typically, either scales or mass flow meters (e.g. Coriolis sensors) are used as input for the mass balances. While this allows monitoring the overall progress, it is only an indirect measurement of important metrics such as the exchange of buffering species or the current protein concentration. During development, effects such as the Donnan effect [322] and protein adsorption to the CFF membrane [323] need to be investigated. The Donnan effect may prevent the full depletion of product counter ions due to the build-up of an electrostatic potential over the membrane [288]. Some buffer components thus might be inadvertently retained despite a diafiltration step. Protein adsorption on membranes is caused by concentration polarization [323]. Proteins are advectively transported to the membrane reaching very high concentrations. Consequently, the proteins may adsorb or interact with other proteins. The conditions may lead to protein aggregation, decreased permeate flow and protein loss.

Off-line analytics are often required to measure the concentration of the target protein and buffer components. In-line and real-time measurements promise to more easily detect said effects and may potentially speed up process development [264]. During production, a control strategy needs to ensure that the product concentration is within the normal operating range during DF and that the final protein concentration complies with the specifications. Especially for subcutaneously administered monoclonal antibodys (mAbs), the high concentrations and low volumes make an in-line control attractive. Not achieving the required protein concentrations during DF and at the end of the process may result in reprocessing or even batch loss. In-line and real-time measurements can reduce this risk and are useful to reduce manual interventions. Additionally, real-time measurements can be used to automate the process resulting in better-controlled processes and improved process times.

Previously, several studies have already investigated Process Analytical Technology (PAT) methods for the UF/DF step. Most studies focused on monitoring at least one of the typical critical quality attributes (protein concentration, excipient concentration, and aggregate content) during UF/ DF. Rolinger et al. used a combination of multiple process analyzers, which were mathematically connected, to calculate protein concentration, buffer exchange progress, and the apparent molecular weight [156]. While in this approach a density signal allowed to monitor the buffer exchange, the effect of the changing protein concentration was neglected, thus potentially resulting in a biased observation. Furthermore, the apparent molecular weight is based on light-scattering measurements which does not allow the independent quantification of aggregates and monomeric species. West et al. [324] used on-line Ultra High Performance Liquid Chromatography (UHPLC) to monitor the protein concentration, aggregate content, and the UV-active excipients. The benefit of an on-line UHPLC is the measurement accuracy, the downsides are long measurement times (5 min to 15 min), the preset dilution factors of the on-line samples and the limited measurability of excipients when using Ultraviolet (UV) absorption for detection. Thakur et al. demonstrated the use of Near-Infrared (NIR) for monitoring and controlling protein and excipient concentrations during CFF in a conventional [325] and a single-pass setup [326]. Both applications are interesting as NIR is well suited for in-line applications in the manufacturing area [62]. However, the water absorbance is strong in the NIR [327] and Infrared (IR) spectral region and shows a significant temperature dependence [264]. The chemometric model thus needs to be validated against temperature variations during a given process but also against long-term variations (e.g. seasonal fluctuations) [328]. Wasalathanthri et al. [329] used Fourier-Transform Infrared (FTIR) to monitor the protein and excipient concentration. While FTIR is more selective compared to NIR, the measurement time was 45 s compared to the 15 s presented by Thakur et al. with NIR. Both measurement speeds can be too slow for UF/DF runs if rapid concentration changes occur in processes due to large membrane areas.

In this study, Raman measurements were used to monitor the protein concentration and buffer exchange. Raman features advantages such as little interference from water and sharp spectral features for the different molecules. The results of the Raman measurements were compared to UV absorption and density measurements as a benchmark. As the changes in buffer and excipient concentrations during the DF are decreasing with increasing process time, an Extended Kalman Filter (EKF) was implemented to estimate the process state based on a semi-mechanistic process model with the predictions on Raman and density measurements. This setup was applied in three case studies to evaluate its performance in different processes and to show the benefits and the limitations of the setup.

8.2 Materials and methods

8.2.1 UF/DF experiments

Experimental setup

The custom made setup from Rüdt et al. [144] and Rolinger et al. [156] was adjusted for automation of the UF/DF process. Figure 8.1 shows the setup as a Piping and Instrumentation Diagram (P&ID). A KrosFlo KRIIi CFF unit (Spectrum Labs, Rancho Dominguez, US) was equipped with a FlowVPE Variable Pathlength (VP) Ultraviolet/Visible (UV/Vis) spectrometer (C Technologies, Bridgewater, US), a non-bypass version of a flow-through micro Liquid Density Sensor (microLDS) (TrueDyne Sensors AG, Reinach, CH), a MarqMetrix BioReactor Ballprobe (MarqMetrix, Seattle, US) inserted into an in-house made flow cell for Raman measurements and a T-piece with injection plug (Fresenius Kabi, Bad Homburg, DE) placed after the retentate reservoir of the CFF unit for drawing samples for off-line analytics. The ball probe was connected to a HyperFlux PRO Plus 785 Raman analyzer with Spectralsoft 2.8.0 (Tornado Spectral Systems, Toronto, CA). Additionally, a fractionation valve of an Äkta prime (Cytiva, Chicago, US) was connected to a relay module, which was controlled via a NI USB-6008 data acquisition device (National Instruments, Austin, US) to switch between air and DF buffer. A Topolino magnetic stirrer (IKA Werke GmbH & Co. KG, Staufen im Breisgau, DE) and a stir bar ensured homogeneous mixing in the retentate reservoir.

Lysozyme

The protocol for the UF/DF process for Lysozyme from our previous publication [266] was slightly adjusted by changing the DF buffer to 50 mM phosphate buffer (VWR Chemicals, Leuven, B) at pH 7.1. In short, the process consisted of an UF phase concentrating the protein from 10 g/L to 20 g/L, a DF phase, where a buffer exchange from citrate buffer at pH 6.0 to a phosphate buffer at pH 7.1 occurred, and a second DF phase to achieve a final concentration of 40 g/L.

mAb

The mAb UF/DF process was adjusted from our previous publication [266]. In the first UF phase, the filtered mAb stock solution at a concentration of 2.79 g/L was concentrated to 25 g/L. A Pellicon 3 Cassette with an Ultracel membrane (type C screen with 30 kDa cutoff, 88 cm^2 membrane area) in a

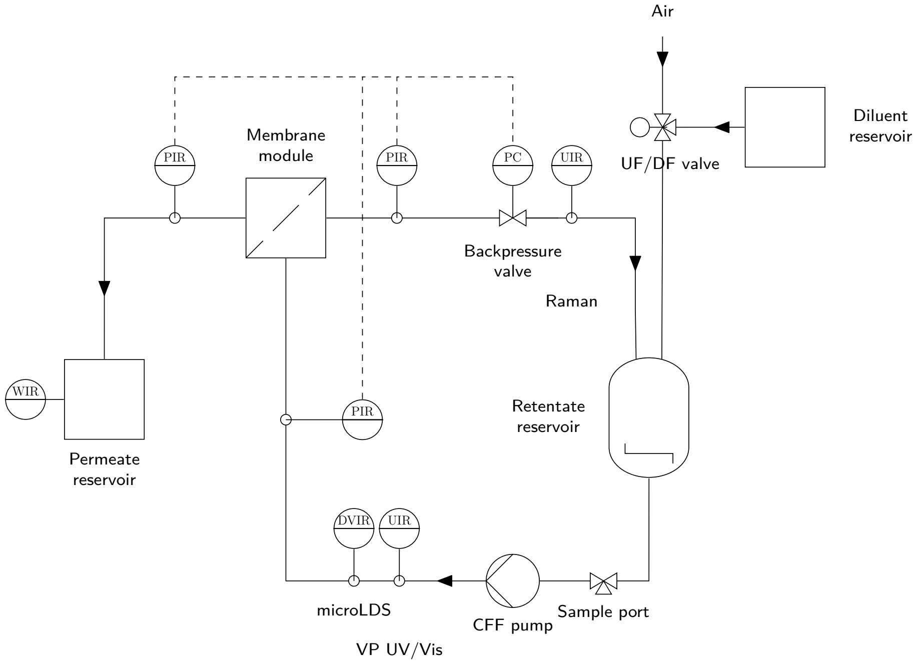

Figure 8.1: Piping and instrumentation diagram of the experimental setup. A VP UV/Vis spectrometer,a microLDS and a Raman probe are incorporated into the flow of the Tangential Flow Filtration (TFF). Additionally, a three-way valve is incorporated to change between UF and DF phase. All sensors are connected to a computer for capturing the data centrally. Electronic communication lines are indicated by dashed lines. The letters indicate: C control, D density, I indicate, P pressure, R record, U multivariable, V viscosity, W weight.

Pellicon Mini Cassette Holder was used (both Merck) in the UF/DF setup. The process was run at a Transmembrane Pressure (TMP) of 1.5 bar and a feed flow of 45 mL/min. In the DF phase, the solution was diafiltrated with eight Diafiltration Volumes (DVs) of DF buffer (250 mM glycine, 25 mM histidine at pH 5.8). In the second UF phase, the solution was concentrated to approximately 100 g/L.

bsAb

For the bispecific antibody (bsAb), the membrane, TMP and feed flowrate settings from the mAb process were used. The bsAb stock solution (concentration 11.49 g/L) was adjusted with a 2 M Tris(hydroxymethyl)aminomethane (TRIS) buffer to pH 7.1 and filtered before use. In a first UF step, the concentration was raised to 25 g/L. Next, the solution was diafiltrated with eight DVs of DF buffer (2.2 mM sodium phosphate, 1.3 mM TRIS). A second UF step concentrated the product to approximately 80 g/L.

8.2.2 Data acquisition and analysis

During experiments, all integrated sensors and devices communicated with and were controlled (except for the Raman analyzer) by a custom-made application developed in MATLAB (version R2020a, The Mathworks, Natick, US) and adapted from Rüdt et al. [144] and Rolinger et al. [156]. Besides connecting the devices and starting and stopping measurements, the application gathered the signals from the integrated sensors and calculated quality attributes and process parameters. Communication and control were performed through software libraries provided by the different instrument manufacturers. In contrast to the previous publications, no Graphical User Interface (GUI) was used to display the signals to save computational power. Data acquisition and analysis of the density and viscosity measurements, Raman measurements, and UV measurements were performed as described below.

UV absorbance measurements and processing

UV slope spectra were recorded from 280 nm to 300 nm for lysozyme, mAb, and bsAb with a resolution of 5 nm. For concentration calculations, the absorbance at 280 nm was used without scatter correction. The settings resulted in a measurement speed of 0.9 min per spectrum. To improve the measurement speed, measuring at a wavelength of 280 nm would be sufficient. Measuring more wavelengths can give information about the formation of aggregates in the solution, as large aggregate scatter increases the background scatter signal in the UV range.

Temperature and protein concentration correction of density measurements

In general, the density ρ of solutions is affected by the buffer components, protein concentration, and temperature. For the obtained data, this was important as the used microLDS dissipates a noticeable amount of heat into the measured liquid. To obtain comparable results, the measured viscosity and density were corrected to a standard process temperature yielding η_{T_0} and ρ_{T_0}, respectively. As the temperature differences were relatively small ($\Delta T \leq 5\,\mathrm{K}$), it was assumed that the deviations from the ideal solution behavior were neglectable [293–295]. The temperature correction was thus performed by cross-multiplication for viscosity and density measurements.

$$\rho_{T_0} = \frac{\rho_{\mathrm{water},T_0}}{\rho_{\mathrm{water},T}}\rho \tag{8.1}$$

This approach is similar to the temperature correction of the sedimentation coefficient performed in analytical ultracentrifugation [34, 296]. Reference values for the density/viscosity of water were obtained from the National Institute of Standards and Technology (NIST) chemistry webbook [297].

To calculate the buffer density ρ_{buffer,T_0}, the influence of the protein concentration on the density was subtracted from the temperature corrected density ρ_{T_0}.

$$\rho_{buffer,T_0} = \rho_{T_0} - a_{prot} \cdot c_{prot} \tag{8.2}$$

where c_{prot} is the protein concentration and a_{prot} is a buffer-dependent factor, also referred to as partial specific volume of the protein. To obtain a_{prot} serial dilutions of the protein in buffer solutions were performed and a_{prot} was estimated as the slope of an ordinary linear regression of $\rho_{T_0} = \rho_{buffer,T_0} + a_{prot} \cdot c_{prot}$ since a linear relationship is expected [330]. As the applied buffer conditions in this paper are fairly narrow in terms of pH range and ionic buffer strength, only small changes in a_{prot} are expected during the DF phase [331]. We therefore used a_{prot} for the DF buffer as an approximation for the whole process phase.

Raman measurements

The laser power during acquisition was set to 495 mW with an exposure time of 800 ms and 10 acquisitions per spectrum for lysozyme and the bsAb. Due to the lower concentration of the mAb, an initial exposure time of 1200 ms was chosen. As the mAb showed a significant level of background scattering, which increased with increasing mAb concentration, the exposure time was step-wise lowered, every time the maximum intensity reached the saturation limit of the detector. X-axis, Y-axis, and laser calibration were done before the experiment according to the manual.

For Partial-Least Squares (PLS) modeling, Solo 8.9 (Eigenvector Research, Inc., Wenatchee, US) was used. First, different spectral preprocessing steps were evaluated to improve the model prediction and linearity based on the recorded dilution series. However, the raw spectra provided the best model accuracy during cross-validation and initial optimization. Consequently, no spectral preprocessing was done and no wavelength selection was done. Only mean centering was applied as it is a standard treatment for spectral data. More information on the PLS models is provided in the Appendix.. For visualization purposes, the automatic asymmetric Whittaker Filter was used along with the Savitzky-Golay filter (15 points, second-order, no derivative) to remove the background/baseline signal and to smooth the data.

Extended Kalman filter implementation

An EKF was used to smooth the data during DF. The EKF concept was selected, because it is the classical concept for extending the Kalman filter concept to non-linear state transitions and observer models, where the direct derivation of the Hessian and Jacobian matrix is possible [332, 333]. However, other alternatives like Particle filters, the Unscented Kalman Filter, or an EKF based on a second order Taylor expansion [334] would have been also a valid choice for smoothing the data during the DF phase. The basic idea behind the EKF is to combine measurements with a non-linear process model to estimate the current true state of the process. This approach also makes predictions into the future possible by leveraging the predictive abilities of the non-linear process model. Predictions may be used to timely terminate reactions, anticipate unwanted behavior or control the process in other ways.

For DF processes, the process may be approximated by the buffer exchange in a Continuously Stirred Tank Reactor (CSTR) under the assumption that the retentate flow is much bigger than the permeate flow and the process volume remains constant. We thus describe the buffer exchange in our CFF setup by following differential equation:

$$\frac{dc}{dt} = c_{in}\frac{F}{V} - c\frac{F\kappa}{V}, \tag{8.3}$$

where c and c_{in} are the concentration of the considered species in the retentate tank resp. the DF buffer, F is the constant permeate flowrate, V is the constant volume of the retentate tank and κ is an empirical sieving coefficient. For free membrane passing ions, κ is close to 1 [61] If a Donnan effect occurs, the sieving coefficient κ can increase or decrease depending on the kind of interaction between the ions of the excipient, protein and membrane [61].. For the differential equation integration and for the EKF transfer function, we assume that κ is constant over time. Since κ is recursively estimated by the EKF, the estimate may change over the course of the run. By integration from t_{k-1} to t_k, we obtain:

$$\frac{c_{in}}{\kappa} - c_k = \left(\frac{c_{in}}{\kappa} - c_{k-1}\right)\exp\left(-\frac{\kappa F}{V}\Delta t\right) \tag{8.4}$$

with c_{k-1} and c_k being the concentration at t_{k-1} and t_k, respectively, and Δt being the step in time. Consequently, a buffer signal during DF follows an exponential decay towards a new steady-state concentration. It is worth noting that Equation 8.4 can directly be used for the EKF as long as a measurement calibration is available. This allows to directly estimate the empirical sieving coefficient κ by the EKF. For the current application, the goal was to implement an EKF which does not require prior calibration. To this end, we now replace the concentrations c with the more general concept of a signal linearly correlated to the concentration. The signal may either be a Raman band intensity, a density measurement, or indeed also a buffer component concentration. The signal may either be increasing or decreasing depending on the nature of the measurement. Transforming Equation 8.4 and lumping the signal terms $\frac{x_{in}}{\kappa} - x(t) = \Delta x(t)$ results in:

$$\Delta x(t_k) = \Delta x(t_{k-1})\exp\left(-\frac{\kappa F}{V}\Delta t\right). \tag{8.5}$$

Starting from Equation 8.5, the EKF is now implemented as described in [332]. Equation 8.6 is used to predict the state vector $\hat{\boldsymbol{x}}_{k|k-1}$ at the time point k based on the measurements up to the time point $k-1$. The first entry in the state vector $\hat{\boldsymbol{x}}_{k|k-1}$ is the estimated delta buffer signal $\hat{x}_1 = E(\Delta x)$.

$\hat{x}_2$ is the estimated buffer exchange rate $E\left(-\frac{\kappa F}{V}\Delta t\right)$. As discussed above, the model assumes that the buffer exchange rate $\hat{x}_2$ is constant over time. $\hat{x}_3$ is the estimated offset, i.e. the terminal signal height $E\left(\frac{x_{in}}{\kappa}\right)$. The offset of the measurement signal $\hat{x}_{3,k|k-1}$ and the buffer signal $\hat{x}_{1,k|k-1}$ is then used to predict the observation $\hat{z}_{k|k-1}$ with Equation 8.7.

$$\text{Predict state of buffer signal} \qquad \hat{\boldsymbol{x}}_{k|k-1} = \begin{bmatrix} \hat{x}_{1,k-1|k-1} \cdot e^{\hat{x}_{2,k-1|k-1}} \\ \hat{x}_{2,k-1|k-1} \\ \hat{x}_{3,k-1|k-1} \end{bmatrix} \tag{8.6}$$

$$\text{Predict state of observation} \qquad \hat{z}_{k|k-1} = \hat{x}_{1,k|k-1} + \hat{x}_{3,k|k-1} \tag{8.7}$$

Equation 8.8 is used to predict the covariance matrix $\boldsymbol{P}_{k|k-1}$ from the previous covariance matrix $\boldsymbol{P}_{k-1|k-1}$ and the Jacobian matrix $\boldsymbol{F}_k$ to linearize the state function on the local point by a first-order Taylor series expansion. The process covariance matrix $\boldsymbol{Q}_k$ is added to account for the model uncertainty. σ_v is the covariance coefficient of the process error.

$$\text{Predict covariance matrix} \qquad \boldsymbol{P}_{k|k-1} = \boldsymbol{F}_k \boldsymbol{P}_{k-1|k-1} \boldsymbol{F}_k^{\intercal} + \boldsymbol{Q}_k \tag{8.8}$$

$$\text{with } \boldsymbol{F}_k = \begin{bmatrix} e^{-\hat{x}_{2,k-1|k-1}} & -e^{-\hat{x}_{2,k-1|k-1}} \cdot \hat{x}_{1,k-1|k-1} & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \text{ and } \boldsymbol{Q}_k = \begin{bmatrix} \sigma_v^2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

The innovation covariance matrix $\boldsymbol{S}_k$ is calculated via Equation 8.9 based on the Jacobian of the sensor transfer functions $\boldsymbol{H}_k$, the covariance matrix $\boldsymbol{P}_{k|k-1}$ and the sensor covariance matrix $\boldsymbol{R}_k$. σ_w is the covariance coefficient of the sensor error.

$$\text{Predict innovation covariance} \qquad \boldsymbol{S}_k = \boldsymbol{H}_k \boldsymbol{P}_{k|k-1} \boldsymbol{H}_k^{\intercal} + \boldsymbol{R}_k \tag{8.9}$$

$$\text{with } \boldsymbol{H}_k = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \text{ and } \boldsymbol{R}_{\boldsymbol{k}} = \begin{bmatrix} \sigma_w^2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Now, the Kalman gain $\boldsymbol{K}_{\boldsymbol{k}}$ can be calculated via Equation 8.10 from the covariance matrix $P_{k|k-1}$ and the sensor transfer functions $\boldsymbol{H}_{\boldsymbol{k}}$, scaled by the innovation covariance matrix $\boldsymbol{S}_{\boldsymbol{k}}$.

$$\text{Predict Kalman gain} \qquad \boldsymbol{K}_k = \boldsymbol{P}_k \boldsymbol{H}_k^{\intercal} \boldsymbol{S}_k^{-1} \tag{8.10}$$

With the calculated Kalman gain $\boldsymbol{K}_k$, the prediction of the state estimate $\hat{\boldsymbol{x}}_{k|k}$ and the covariance matrix $\boldsymbol{P}_{k|k}$ can be updated via Equation 8.11 and Equation 8.12, respectively.

$$\text{Updated state estimate} \qquad \hat{\boldsymbol{x}}_{k|k} = \hat{\boldsymbol{x}}_{k|k-1} + \boldsymbol{K}_k(z_k - \hat{z}_{k|k-1}) \tag{8.11}$$

$$\text{Updated covariance estimate} \qquad \boldsymbol{P}_{k|k} = (I - \boldsymbol{K}_k \boldsymbol{H}_k) \boldsymbol{P}_{k|k-1} \tag{8.12}$$

In principle, the peak height of the buffer component in question in the Raman spectrum may be used as an input signal for the EKF. To improve the prediction and reduce noise levels, Raman spectra were factorized by a Principal Component Analysis (PCA) and the principal component score of the buffer component was used as input for the EKF.

8.2.3 Off-line analytics by SEC

The off-line Size-Exclusion Chromatography (SEC) analytic was done according to our previous publication [266], with the difference that already mAb and bsAb samples with concentrations higher than 30 g/L were diluted 10-fold. bsAb samples were analyzed according to the protocol for the mAb.

8.3 Results and discussion

In this study, three different case studies are investigated to compare Raman spectroscopy, UV spectroscopy, and density measurement for their ability to measure the protein concentration and buffer exchange progress. First, the Raman spectra are discussed in detail. Then, Raman spectroscopy and UV spectroscopy are compared towards their prediction accuracy for the protein concentration. Finally, density measurements and Raman spectroscopy will be compared towards their ability to monitor the buffer exchange progress.

8.3.1 Raman spectra

In Figure 8.2, every 50th spectrum of the lysozyme process and every 40th spectrum of the mAb and bsAb process are shown. For lysozyme, the protein features are well visible in the Raman spectra with bands in the range from 500 cm^{-1} to 1700 cm^{-1} and around 2900 cm^{-1}. The sapphire bands at 384 cm^{-1}, 418 cm^{-1}, 452 cm^{-1} and 753 cm^{-1} are visible in the Raman spectra of all case studies and are, as expected, constant. The protein bands, especially at 1006 cm^{-1} originating from phenylalanine, 1360 cm^{-1}, 1448 cm^{-1}, and 1549 cm^{-1} originating from tryptophane and C-H deformation [151, 279], and at 2942 cm^{-1} originating from C-H stretching [34], are distinct from other components by the fact that they also increase during the second UF step. Additionally, an increase in background signal correlates with the increase in protein concentration. This phenomenon was already discussed in a previous publication and is likely related to increased Rayleigh scattering [335]. For the bsAb, the spectra look comparable to the lysozyme spectra, even though the height of the protein features is lower. The mAb shows an increased background signal in comparison to the other two proteins and, therefore, lower intensities in the protein bands compared to the background signal. Already in the last publication, an increased molecule to molecule interaction of the mAb was detected, which lead to buffer-induced light scattering increase and gel formation [335]. Here, the change in the background signal is more pronounced than the change in protein features. Again, the intensity of the background signal correlates with the protein concentration. Measurement-wise, the large amount of background signal made a reduction of exposure time necessary to prevent the over-saturation of the detector. The spectra were subsequently normalized by the exposure time.

8.3.2 In-line protein concentration measurements

For monitoring the protein concentration, the absorbance at 280 nm from the VP UV spectrometer and the full spectra of the Raman analyzer in combination with a PLS model were used. In Figure 8.3, the predicted protein concentrations are compared to the results obtained from off-line SEC analysis. Qualitatively, both the predicted protein concentration from the Raman analyzer and the VP UV spectrometer are in good agreement with the off-line analytics for all three processes. For lysozyme, towards the end of the second UF, the FlowVPE signal starts to deviate from the Raman signal. This was attributed to an increasing amount of air bubbles in the solution, which impaired the FlowVPE measurements.

During the whole process, the Raman predictions showed a few outliers, probably caused by air bubbles in the measurement chamber. In manufacturing, this could be mitigated by rejecting the predictions based on the Hotelling's T^2 value or the distance to the model hyperplane. The PLS model for the Raman-based protein concentration predictions is mostly influenced by the background signal. This is in agreement with [335] and

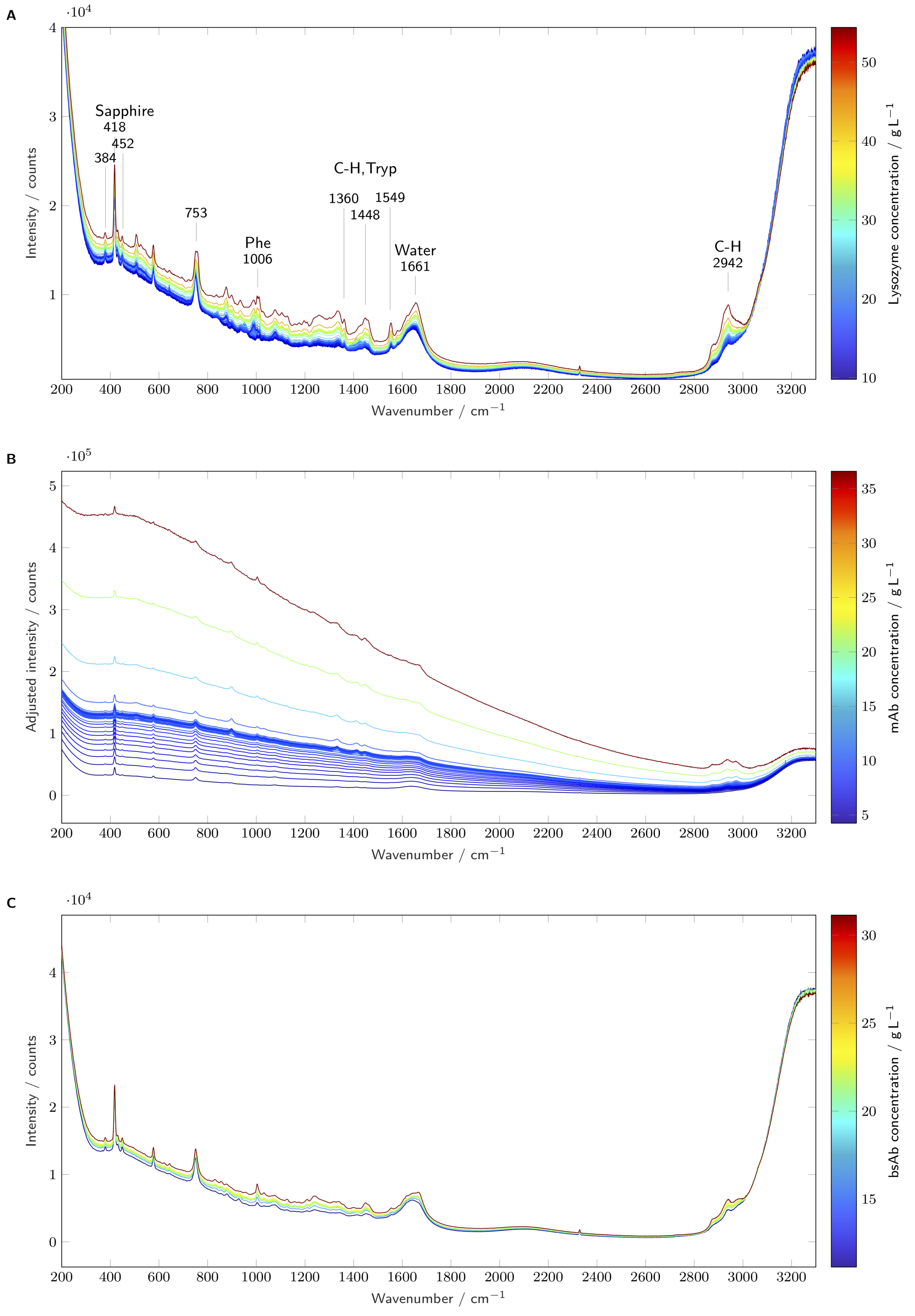

Figure 8.2: The raw Raman spectra recorded by the in-line Raman analyzer are plotted and colored according to the protein concentration. The different subplots show the results for lysozyme (A), mAb (B), and bsAb (C).

can already be seen by comparing the concentration prediction to Figure 8.8, which shows the intensity trend at 700 cm^{-1}, where no protein vibrational band is located. Theoretically, a PLS model might not be necessary to predict the protein concentration as a single intensity already correlates well to the protein concentration. However, a single wavenumber/wavelength measurement has usually lower accuracy compared to the PLS model based on several wavenumber [335]. The dependence of the PLS model on the background reduces the specificity of the model for the protein of interest. For example, an increased aggregate content likely increases the background signal disproportionately and thereby affects the protein concentration prediction. In routine production, the reduced selectivity is however not a problem since any manufacturing process must be reproducible regarding the feed composition and will work with highly pure protein solution, especially towards the end of the process. UV/Vis absorption relies on the more specific absorption of the aromatic amino acids [264]. Quantification is normally robust and not significantly impacted by batch-to-batch variability. In the current experimental results, the UV-based protein concentration measurements show fewer outliers in comparison to the Raman measurements. However, UV measurements were already filtered based on a coefficient of determination higher than 0.97 during the VP regression.

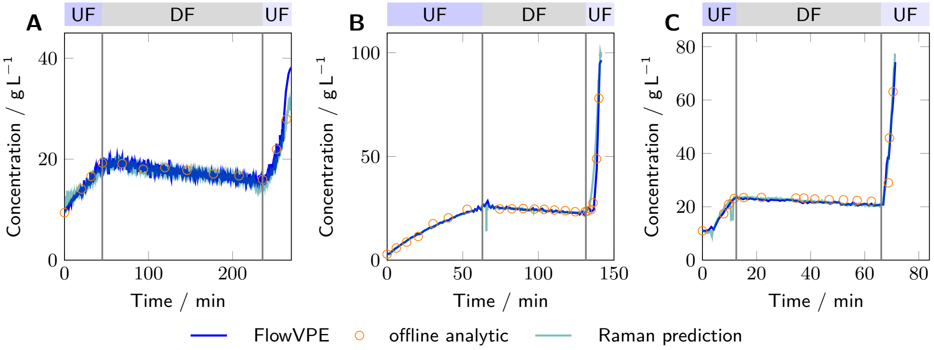

Figure 8.3: The total protein concentration is shown as measured by the in-line FlowVPE VP spectrometer (blue lines), Raman analyzer (teal lines) and off-line SEC (orange circles). The different subplots show the results for lysozyme (A), mAb (B), and bsAb (C).

Quantitatively, the Root Mean Square Error (RMSE) for lysozyme and the bsAb of the UV- and Raman-based measurements are very similar (cf. Table 8.1). However, for the mAb, the RMSE of the Raman predictions is with 4.59 g/L more than twice as high as the RMSE of the UV-based measurements at 1.73 g/L. This difference is mostly driven due to the residuals in the second UF phase. Due to the fast change in protein concentration, the uncertainty in the sampling time affects the measurement accuracy more strongly than during the rest of the process. Furthermore, as shown in the P&ID (Figure 8.1), the Raman spectra were measured in the retentate (due to pressure constraints of the flow cell), while sampling and UV-based measurements were conducted in the feed. Normally, if the feed and retentate flow are similar, this does not pose a problem. However, at the beginning of the second UF phase, the process progressed very quickly introducing a systematic offset and increasing the overall RMSE for the Raman measurements.

Table 8.1: RMSE , normalized RMSE, and coefficients of determination for UV and Raman measurements compared to off-line analytics.

	Concentration prediction based on FlowVPE			Concentration prediction based on Raman		
		normalized		normalized		
	RMSE	R^2	RMSE	RMSE	R^2	RMSE
	/ g/L		%	/ g/L		%
Lysozyme	0.87	0.9874	4.87	0.99	0.9799	5.54
mAb	1.73	0.9943	3.31	4.59	0.9788	8.80
bsAb	2.88	0.9709	3.83	2.67	0.9771	3.55

Given the results obtained in the three case studies, both VP UV/Vis spectroscopy and Raman spectroscopy are useful tools for quantifying proteins in-line in real-time during UF/DF processes. Raman spectroscopy was quicker compared to VP UV/Vis spectroscopy, which takes about 8 s when measuring at one wavelength at four pathlengths. UV/Vis spectroscopy may be more robust towards process variability (e.g. changing aggregate content), because the background effect in the Raman measurements seems to mostly origin from the molecular weight and interaction between molecules. Additionally, UV/Vis spectroscopy works by simple determination of the absorption coefficient without the need to calibrate a chemometric model. Although the data analysis of the Raman spectra is more complex in comparison to UV/Vis measurements, Raman spectroscopy allows for simultaneous insights into the protein and excipient concentrations. The ability of Raman spectroscopy to selectively measure different excipients will be used in Section 8.3.3 to monitor the buffer exchange process. Both methods measure protein concentration in the investigated range without any major deviations from linearity. Noise levels remain comparably small. In this study, the traditional limit of quantification could not be applied for comparing the concentration predictions. This is due to the fact that the concentration predictions from Raman spectra relied on a multivariate PLS model which does not permit traditional limit of quantification calculations. It is worth considering that the limit of quantification for Raman-based concentration prediction changes due to changing measurement settings (e.g. exposure time). We therefore consider the comparison of Root Mean Square Error of Cross Validation (RMSECV) as most insightful.

Predicting the concentration of the different aggregate and fragment species was attempted with Raman spectroscopy in this study, but it was ultimately unsuccessful. The concentrations of the individual species might be too low and the structural changes between differently sized species not prominent enough to be picked up by Raman spectroscopy in the short measurement times. However, Wei et al. [336] showed promising results to quantify aggregates and fragments with a multi-product PLS model based on offline Raman measurements with a measurement time of 22.5 minutes.

8.3.3 Buffer exchange progress monitoring

In Figure 8.4, the preprocessed Raman spectra of the DF phase are plotted. For the lysozyme case study, the change from citrate buffer to phosphate is most prominently visible at 840 $\mathrm{cm^{-1}}$, 952 $\mathrm{cm^{-1}}$, 990 $\mathrm{cm^{-1}}$ and 1412 $\mathrm{cm^{-1}}$. The citrate buffer has a significant number of bands (see teal line). The most prominent band at 952 $\mathrm{cm^{-1}}$ can be attributed to $\mathrm{O-HO}$ out-of-plane deformation vibration of the carboxylic acid group [337]. Also prominent is the carboxylate symmetric stretching band at 1412 $\mathrm{cm^{-1}}$ and the carbon-carbon stretching mode at 840 $\mathrm{cm^{-1}}$[338]. Phosphate shows a major band at 990 $\mathrm{cm^{-1}}$ due to the P-O stretching of phosphate [337, 339–341] along with to bands at 1078 $\mathrm{cm^{-1}}$ and 877 $\mathrm{cm^{-1}}$, which can be attributed to the symmetrical $\mathrm{P(OH)_2}$ stretching vibration and the in-plane $\mathrm{PO_2}$ [340].

For the mAb case study, the DF buffer consists of histidine and glycine which have the most dominant peaks at 899 $\mathrm{cm^{-1}}$, 1332 $\mathrm{cm^{-1}}$, 1413 $\mathrm{cm^{-1}}$, 1446 $\mathrm{cm^{-1}}$ and 2972 $\mathrm{cm^{-1}}$ (Figure 8.4B, black line). Glycine has a strong $\mathrm{C-C}$ stretching band at 899 $\mathrm{cm^{-1}}$ [342]. The other two intense Raman bands 1332 $\mathrm{cm^{-1}}$ and 1413 $\mathrm{cm^{-1}}$ can be attributed to the twisting of the $\mathrm{NH_3}$ and $\mathrm{CH_2}$ groups and a $\mathrm{NH_3}$ wagging mode coupled with COO stretching [342]. A smaller band is located at 1448 $\mathrm{cm^{-1}}$ and is caused by $\mathrm{CH_2}$ scissoring [342]. The peak at 2972 $\mathrm{cm^{-1}}$ is caused by the symmetric stretching of $\mathrm{CH_2}$ [343]. These peaks are expected to build up during the DF. No distinct bands for histidine are visible. This might be caused by the 10-fold lower concentration.

The phosphate and TRIS DF buffer for the bsAb case study is very low concentrated, therefore no changing peaks attributed to the DF buffer are visible in the process spectra. TRIS has to a CH_2 deformation band at $1470\,cm^{-1}$ and CO stretching at $1066\,cm^{-1}$ [337, 344] However, a chemical with two peaks at $881\,cm^{-1}$ and $930\,cm^{-1}$ is depleted during the DF. This chemical presumably originates from the previous production step, a chromatographic separation.

Figure 8.5 shows the normalized buffer signal derived from the Raman spectra over time. The normalized signal consists of the normalized scores of the principal component collecting the spectral variability due to the buffer exchange. This approach was used to improve the signal-to-noise ratio. In principle, also a unique peak of a buffer component could be chosen. However, due to the various proteineous Raman peaks, the peaks are often overlapping with the protein peaks. A PCA allows separating the protein signal from the buffer signal.

For the first case study with lysozyme, the phosphate peaks of the DF buffer are too weak and overlapping with the citrate peak, so that an individual monitoring of the two species is not possible. Instead, the principal component representing the citrate buffer was used. The signal itself follows a decay curve as expected during DF. Interestingly, the signal is still changing around the end of the DF at four DV. Raman spectroscopy provides this information in real-time allowing for an immediate evaluation of the DF process. Based on the observed behavior, a decision may be taken to extend the DF phase.

For both the mAb and bsAb, the buffer signal seems stable towards the end of the DF. For the mAb, the principal component analysis did not differentiate between the two components of the DF buffer, glycine and histidine. Depending on the net-charge of the mAb at the given pH, a Donnan-effect was previously observed with an accumulation of histidine during the DF for negatively charged mAbs [345, 346]. The observed accumulation was within 3 mM after eight DV [345]. Either the higher overall ion-concentration [347], a positively charged mAb or the quantification limit of the Raman could have led to the non-observability of the effect.

The Raman signal in all case studies shows significant noise, which makes the signal more difficult to interpret and to use to control the process. To reduce the noise level, an EKF was used to approximate the real process state from the noisy measurements. Kalman filters allow for some plant variability. They are also applicable in real-time for recursive state estimation and control. The orange lines in Figure 8.5 and 8.6 indicate the EKF-filtered results. It is worth noting that the EKF successfully suppresses a significant part of the measurement noise. Furthermore, with the used estimates for

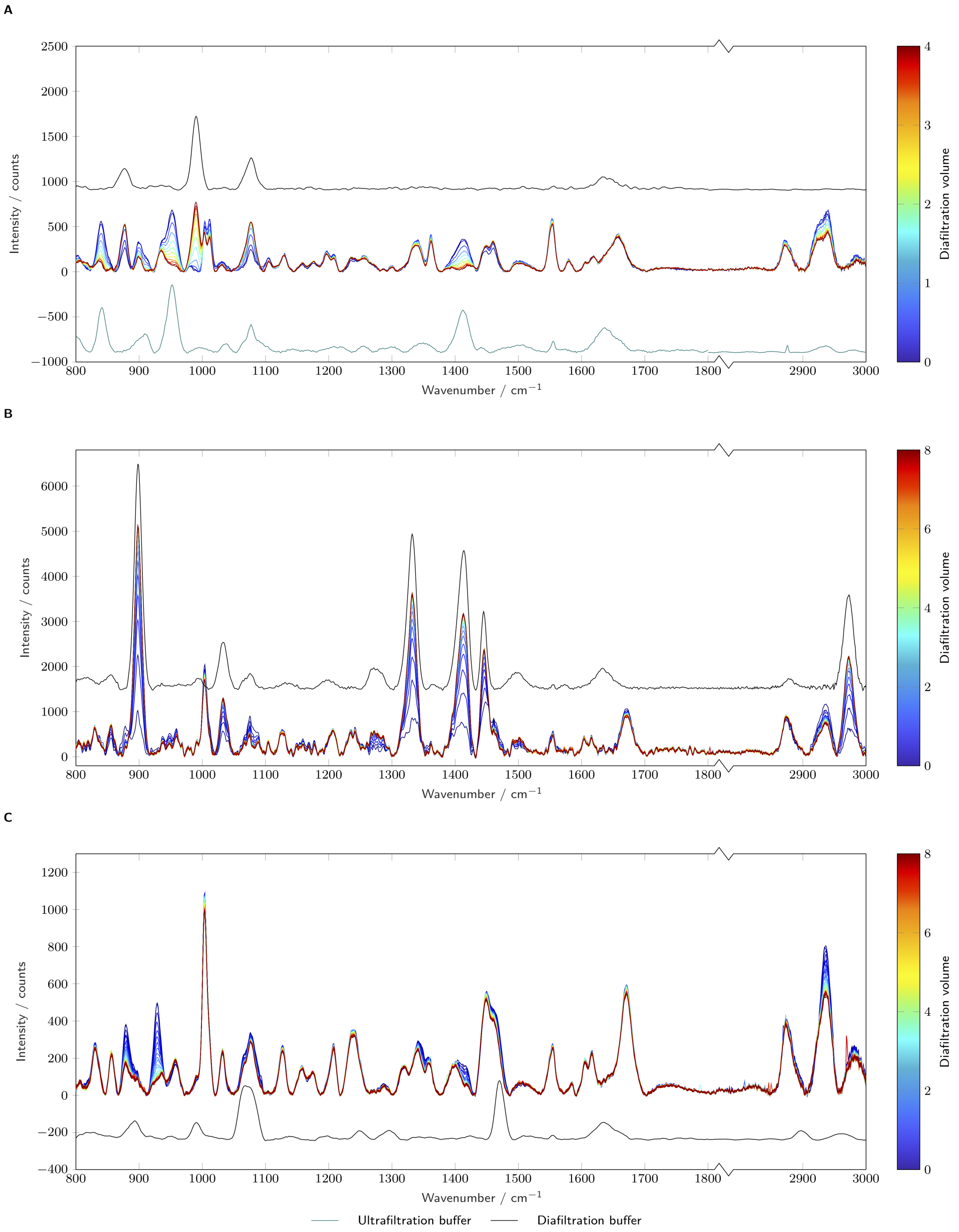

Figure 8.4: The preprocessed Raman spectra recorded during the DF phase are plotted and colored by diafiltration volumina. The diafiltration buffer (black line) are plotted with an offset. For lysozyme, additionally the ultrafiltration buffer (teal line) is depicted. The different subplots show the results for lysozyme (A), mAb (B), and bsAb (C).

the system and measurement noise, the EKF is still flexible enough to adjust the prediction dynamically to changing conditions. For example, during the diafiltration of the mAb (Figure 8.5B and 8.6B), the buffer exchange initially starts more slowly than expected. The EKF incorporates this into its prediction by adjusting the two other state variables (offset and exponential decay constant). The estimated state variables may also be used to gain insight into changes of the filtration behavior, e.g. due to a changing sieving coefficient. This approach provides a real-time mechanistic insight into the process performance and may help to improve the understanding of the ongoing process. The EKF, thus, provides an interesting tool for real-time recursive evaluation of the buffer exchange progress and a method for an improved understanding of the ongoing process variability.

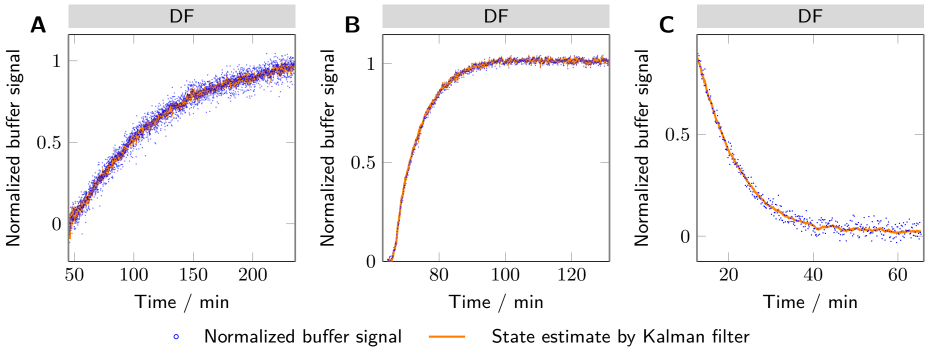

Figure 8.5: The normalized buffer signal derived from a PCA of the Raman spectra are plotted over time with the state estimate of the Kalman filter. The different subplots show the results for lysozyme (A), mAb (B), and bsAb (C).

Next to the Raman signal, the buffer exchange was monitored by the evolution of the density signal. The density signal is univariate, collecting information from all components in the solution in one variable. Therefore, no separate monitoring of individual species is possible. Next to the buffer components, the density signal is also affected by changing protein concentrations. In Figure 8.6, the temperature and protein concentration-corrected density are plotted. The concentration-correction was done with the protein concentration predictions from the Raman due to the higher measurement frequency of the Raman signal. For lysozyme, the density measurements were corrupted by air bubbles due to the increasing viscosity of the solution over time. Repeating the experiments led to the same phenomena. The air bubbles seem to decrease the liquid density and remain in the feed due to the viscosity of the solution. Only filters could help to remove bigger air bubbles from the solution, but might introduce more aggregation due to shear forces. For viscose protein solutions, the density measurements seem difficult due to the air entrapment. Therefore, the concentration-correction does not work for the lysozyme case study, because during the DF the corrected density is decreasing, which is not physically reasonable and not in agreement with the Raman data. For the mAb and bsAb case study, the protein concentration-correction of the density leads to a stable signal towards the end of the DF. The measured density for the mAb and bsAb case study seems to decline over the whole DF phase due to the protein concentration decrease, whereas the Raman-based buffer signal already indicates a stabilization and thereby completion of the DF process. The measured density signal alone is, therefore, only of limited use to monitor and control the DF phase. The protein concentration-correction density seems to agree with the Raman-based buffer signal. However, a direct comparison between the protein concentration-corrected density and the normalized buffer signal by Raman is difficult to make based on Figure 8.6. Therefore, the comparison is directly plotted in Figure 8.7.

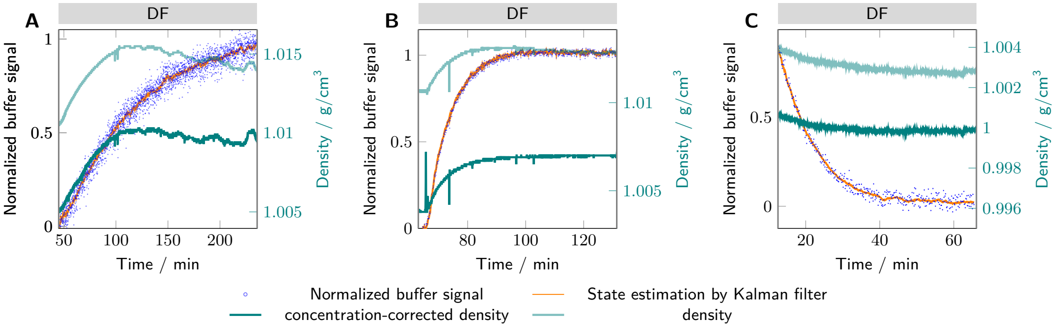

Figure 8.6: The density (light teal) and the concentration-corrected density (teal) are plotted over the DF run time along with the normalized buffer signal from the Raman measurements and the EKF prediction (orange). The different subplots show the results for lysozyme (A), mAb (B), and bsAb (C).

Figure 8.7 shows the comparison between the Raman measurements and the concentration-corrected density measurements. The lysozyme data is not plotted due to the unreliability of the density measurements as discussed above. Both are in good agreement, even though a significant noise level is apparent for both measurements. For the density data, the Kalman filter can improve the DF progress prediction as well. The density signal has the benefit of even observing Raman-inactive components in the solution, like NaCl, under the prerequisite of a density difference between buffers.

However, the needed protein concentration correction makes a second sensor necessary, which adds complexity and room for failure.

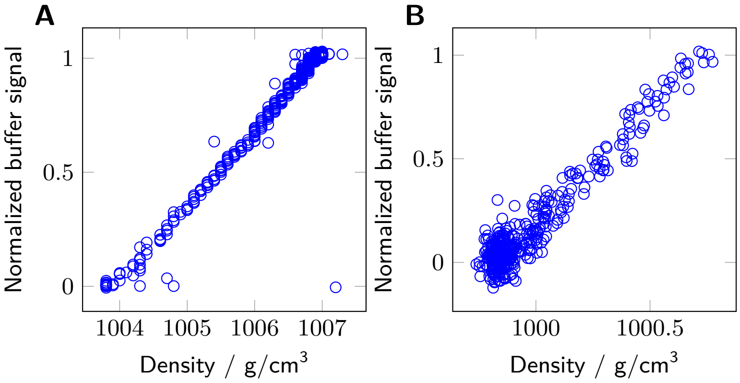

Figure 8.7: The concentration-corrected density is compared to the normalized Raman signal. The different subplots show the results for mAb (A), and bsAb (B).

8.4 Conclusion

In this study, the advantages and disadvantages of Raman spectroscopy for monitoring UF/DF processes were shown in three case studies and compared to UV absorption and density measurements as a benchmark. To improve the sensitivity of the measurements, an EKF was implemented to estimate the process state during the DF based on a semi-mechanistic process model combined with the predictions of Raman and density measurements. Raman spectroscopy and VP UV/Vis spectroscopy were compared for their prediction accuracy of the protein concentration in comparison to off-line measurements. VP UV spectroscopy showed slightly better or comparable coefficients of determination in comparison to the Raman measurements. UV concentration measurements were derived based on the absorption coefficient at 280 nm, while Raman measurements required a PLS model to predict the protein concentration. Raman measurements took less than a second in comparison to eight seconds for the VP UV measurements. The higher measurement speed of the Raman spectrometer may be an advantage for fast processes. However, the Raman measurements were more prone to outliers in comparison to the UV measurements.A drawback of the Raman spectroscopy is that the prediction of the protein concentration seems to rely on the unspecific background effect, that correlates with the protein concentration. In addition to the protein concentration prediction, the Raman spectra provided the concentration of Raman-active buffer components. These concentration predictions were used to monitor the buffer exchange progress. To reduce the measurement noise, an EKF was used for state estimation. The prediction of the buffer exchange progress by Raman was less noisy compared to the density measurement. Another advantage of Raman spectroscopy is the ability to monitor individual buffer components.

Among other applications, Raman measurements thus pave a further step on the way towards the real-time control of the protein concentration during and at the end of the UF/DF process ensuring the final product concentration and buffer composition within the processes. Raman measurements thus pave a further step on the way towards Real-time Release Testing (RTRT) by replacing off-line in-process controls of critical quality attributes by their in-line equivalents.

8.5 Appendix: Exposure time correction

In Figure 8.8, the Raman intensity at $700\,\mathrm{cm}^{-1}$ and the exposure time-adjusted Raman intensity at $700\,\mathrm{cm}^{-1}$ are plotted over time for the mAb run. The exposure time adjustment yields a fairly smooth curve, which correlates to the protein concentration in the run, even though there is no protein band at $700\,\mathrm{cm}^{-1}$. It seems, that the increase in background signal is mainly driven by the protein concentration. It is interesting to note that the background effect seems not be influenced by the buffer change during the DF phase.

8.6 Appendix: Additional information on Raman-based PLS models

8.6.1 Selection of number of Latent Variables

In Figure 8.9, the RMSECV and Root Mean Square Error of Calibration (RMSEC) over the number of Latent Variables (LVs) is shown. It can be seen, that increasing the number of LVs above one does not improve the prediction ability of the model. The main concentration information seems to be already captured in the first latent variable.

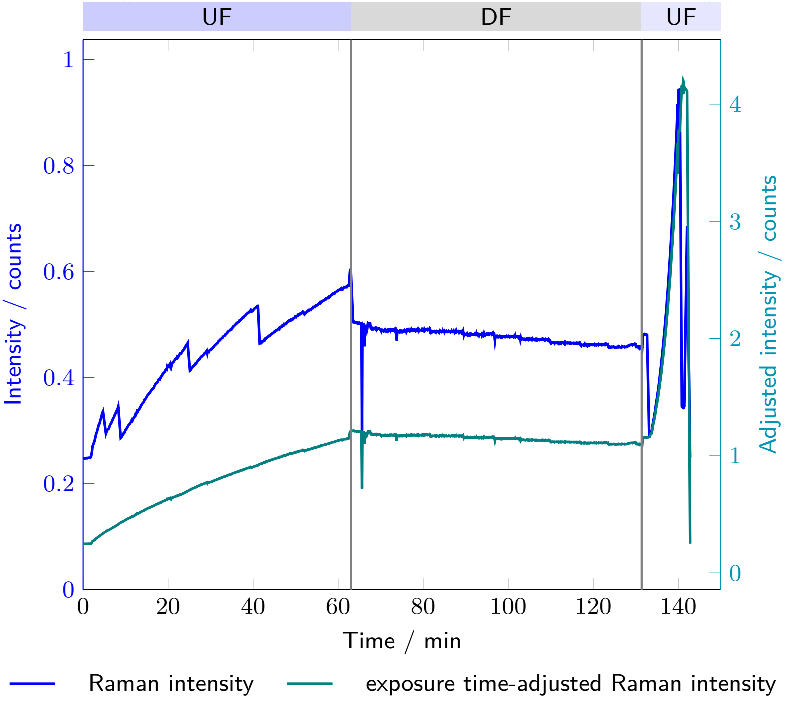

Figure 8.8: The Raman intensity at $700\,\mathrm{cm}^{-1}$ and the exposure time-adjusted Raman intensity at $700\,\mathrm{cm}^{-1}$ are plotted during the mAb run.

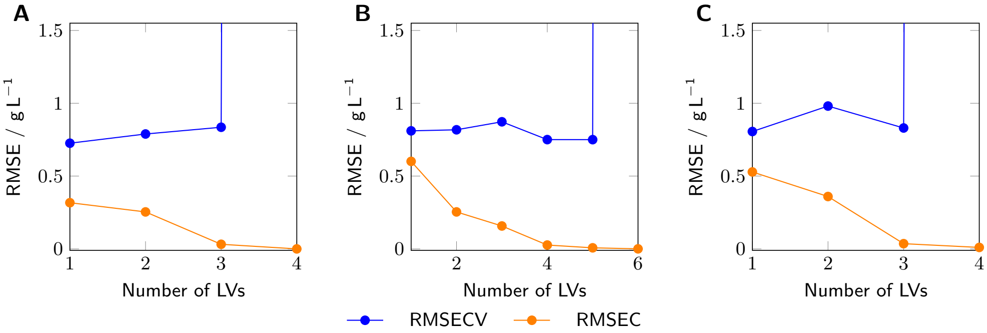

Figure 8.9: The RMSECV and RMSEC are plotted over the number of LVs for the Raman-based PLS models. The different subplots show the results for lysozyme (A), mAb (B), and bsAb (C).

Table 8.2: Inlfuence of different preprocessing options on RMSEC and RMSECV of the Raman based PLS model for the mAb

Preprocessing	Number of LVs	Wavenumbers / cm^{-1}	RMSEC / g/L	RMSECV / g/L
Mean center	1	200-3300	0.60	0.81
1st deriv., mean center	1	200-3300	2.53	3.47
2st deriv., mean center	1	200-3300	5.27	11.52
EMSC, mean center	1	200-3300	9.33	12.59
Mean center	1	300-1800	0.60	0.80
1st deriv., mean center	1	300-1800	3.16	4.68
2st deriv., mean center	1	300-1800	5.18	9.18
EMSC, mean center	1	300-1800	11.43	15.45
1st deriv., mean center	2	300-1800	0.91	1.54
2st deriv., mean center	2	300-1800	1.44	7.78

8.6.2 Preprocessing evaluation

As the Raman spectra for the mAb mainly contain the background information, which is correlated to the protein concentration, the model accuracy is reducing, when removing more of the background information. Either by removing the background through preprocessing or by reducing the wavelength range. When removing the background effect, an increase in LVs improves model prediction. More preprocessing options could be evaluated to further improve the model performance, for instance by using a Genetic Algorithm (GA). However, the PLS model calibration was based on a simple dilution series resulting in a small calibration data set. Therefore, a simple model was built to fulfill the prediction requirements.

8.6.3 Scores plot

In Figure 8.10, the scatter plots show the scores on the first Principal Component (PC) against the scores on the second and third PC. The first PC always contains mainly the concentration information. The second PC seems to be mostly influenced by the buffer exchange.

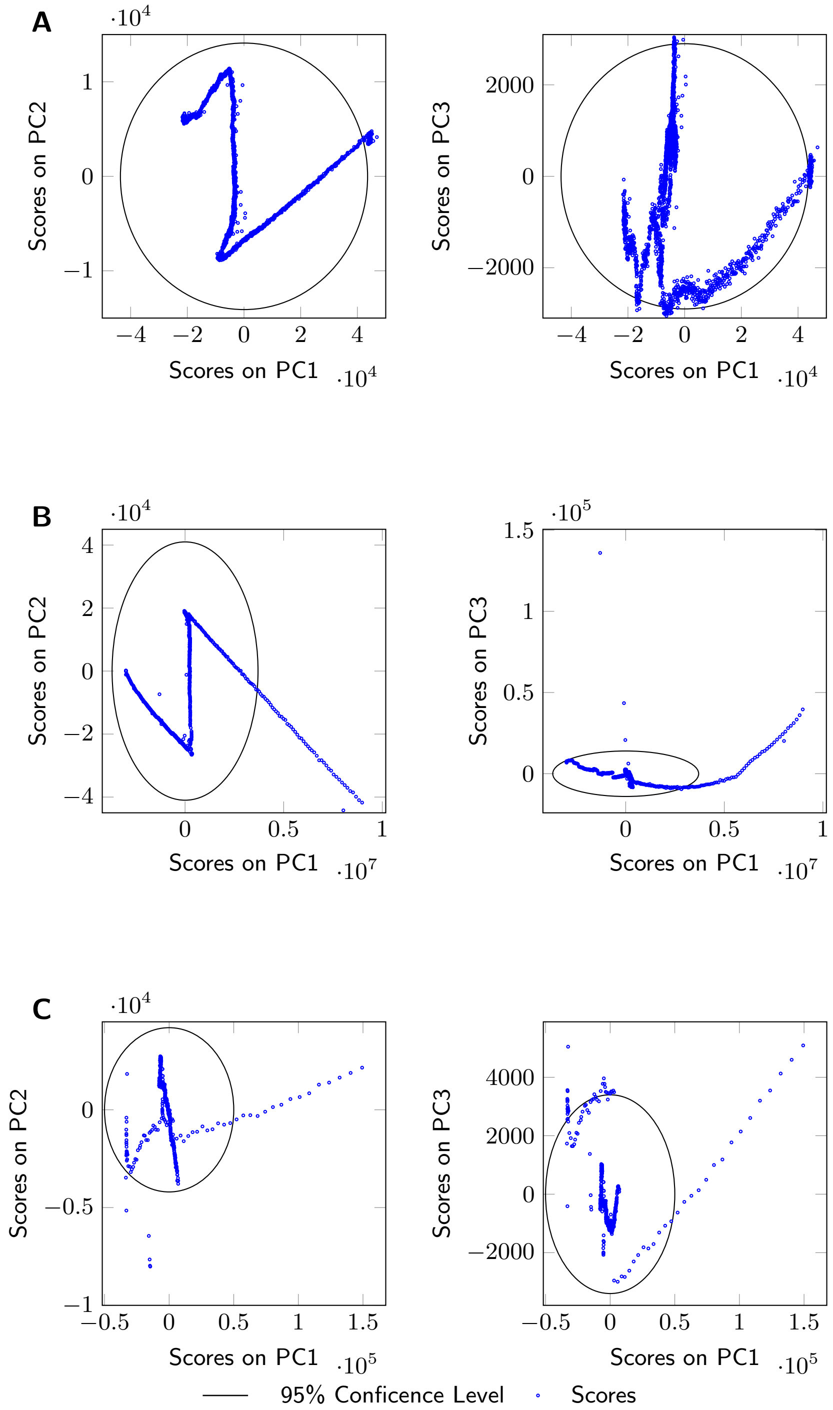

Figure 8.10: Scatter plots of scores from PCA. The different subplots show the results for lysozyme (A), mAb (B), and bsAb (C).

Table 8.3: Inlfuence of different preprocessing options on RMSEC and RMSECV of the Raman based PLS model for the bsAb.

Preprocessing	Number of LVs	Wavenumbers / cm^{-1}	RMSEC / g/L	RMSECV / g/L
Mean center	1	200-3300	0.53	0.81
1st deriv., mean center	1	200-3300	0.41	0.80
2st deriv., mean center	1	200-3300	0.13	2.01
EMSC, mean center	1	200-3300	1.29	1.99
Mean center	1	300-1800	0.52	0.85
1st deriv., mean center	1	300-1800	0.38	0.82
2st deriv., mean center	1	300-1800	0.53	0.80
EMSC, mean center	1	300-1800	1.70	3.25

Table 8.4: Inlfuence of different preprocessing options on RMSEC and RMSECV of the Raman based PLS model for Lysozyme.

Preprocessing	Number of LVs	Wavenumbers / cm^{-1}	RMSEC / g/L	RMSECV / g/L
Mean center	1	200-3300	0.53	0.81
1st deriv., mean center	1	200-3300	1.27	2.53
2st deriv., mean center	1	200-3300	1.51	4.65
EMSC, mean center	1	200-3300	0.40	1.24
Mean center	1	300-1800	0.26	0.90
1st deriv., mean center	1	300-1800	1.53	2.71
2st deriv., mean center	1	300-1800	0.91	2.60
EMSC, mean center	1	300-1800	1.88	3.05

9 General discussion and conclusion

The goal of this thesis was to develop smart sensor concepts as PAT tools for the downstream process of biologics. First, an evaluation of optical spectroscopic sensors was done in Chapter 3 to evaluate the capabilities of the methods for the application to the downstream process. UV spectroscopy in combination with PLS modeling was identified as a promising technique to measure product concentration in the column effluent of a Protein A column (Chapters 4, 5, and 6). The initial approach described in Chapter 4, was further refined in Chapter 5, where the approach was combined with a conductivity-based background spectrum subtraction to improve the precision of the approach and make it more applicable to real processes. Thereby the limited selectivity of UV spectroscopy was overcome. An alternative to UV measurements is Raman spectroscopy, which is more selective toward different structural elements of proteins, but lacks sensitivity at low protein concentrations and requires longer measurement times. In Chapter 6, UV spectroscopy was compared to Raman spectroscopy. Additionally, the value of multi-block data fusion methods and Convolutional Neural Networks (CNNs) as non-linear methods is evaluated as an attempt to combine the advantages of UV and Raman spectroscopy. In a last study in Chapter 7, VP UV spectroscopy, light scattering and density/viscosity measurements are combined to monitor three different UF/DF processes.

Chapter 3 evaluates the sensitivity and selectivity of optical spectroscopic techniques toward the quantification of proteins in the downstream process. Due to the close chemical relation of the desired product to the contaminants, also the specificity for measuring different protein structure attributes is important for the applicability of spectroscopic techniques.

Therefore, a focus lies on the measurability of different structure levels of proteins and thereby differentiation between proteins by different spectroscopic methods. Additionally, a guidance for PLS model calibration and validation is given to avoid common pitfalls in data analysis of spectroscopic data. From this review, generally UV spectroscopy seems to be the most sensitive spectroscopic technique for measurement of proteins in aqueous solutions due to the high absorption coefficients of proteins and low absorption of water in the UV region. However, UV spectroscopy lacks a high selectivity to differentiate between different proteins due to the broad and overlapping bands of different structure elements. Here, Raman spectroscopy seems to be promising due to the availability of information on the primary, secondary, and tertiary protein structure in the spectra. The only spectroscopic technique, which can measure the most common form of aggregated proteins with main changes in quaternary structure level is light scattering. Unfortunately, there is no sensor, which can measure all Critical Quality Attributes (CQAs) in the downstream process. This makes sensor combination necessary. Therefore, Chapter 3 includes a summary of data fusion techniques to cope with the multi-block data from different sensors.

As a first example, the implementation of UV spectroscopy combined with PLS modeling is shown as a proof-of-concept for the real-time monitoring and control of the Protein A load phase in Chapter 4. It was demonstrated that PLS modeling on UV absorption spectra can be applied to quantify the mAb concentration in the column effluent during the load phase despite the influence of many protein and non-protein-based impurities on the UV spectra. Based on the quantification, the load phase was automatically terminated, when a previously specified mAb concentration was reached. Consequently, the proposed method has the potential for monitoring and control of capture steps, like Protein A chromatography, at large scale production for both batch and continuous processes. In batch chromatography, the loading volume can be determined dynamically with the proposed method, which allows for increased resin capacity utilization while keeping the product loss small. Additionally, the time-consuming off-line determination of the mAb titer in the Harvested Cell Culture Fluid (HCCF) could be eliminated. For continuous chromatography, the proposed method may also be interesting for controlling the column switch for capture steps. For continuous chromatography, the proposed method may also be interesting for controlling the column switch. A drawback of the study is that only a variation in mAb titer in the upstream was included into the study. Other variations, like the contaminant content in the HCCF or media component changes, were not investigated.

In a next step, the method with UV spectroscopy combined with PLS modeling was applied in a second study (see Chapter 5) to a revised design space, which included large process variation due to the use of different feedstocks with different mAb compositions, to test the robustness of the method and applicability to different products. The study showed, that the error of the method is increased due to the large design space. To overcome this challenge, a dynamic UV background subtraction based on the leveling out of the conductivity signal during the load was implemented to increase the prediction ability of the PLS model. It was demonstrated that by subtracting the background spectrum during the breakthrough, the prediction of the mAb concentration is facilitated and improved compared to models using the raw spectra.

The conductivity-based background subtraction in combination with PLS modeling on UV spectra offers a robust quantification of the product breakthrough regardless of large variability in the cell culture fluid. Additionally, it was shown, that by using the conductivity-based background subtraction, the use of a single absorption wavelength instead of a multivariate spectrum becomes feasible for the mAb quantification. This smart sensor concept shows great potential for the application to production processes as the required sensors are already implemented in most processes.

Even though the results of the study presented in Chapter 5 are promising it also shows, that UV spectroscopy lacks the sensitivity to distinguish between the mAb and contaminants, which makes the background subtraction necessary. Other spectroscopy methods, especially Raman spectroscopy, have proven to be more selective. Therefore, Raman spectroscopy is frequently used in upstream processing to differentiate between various cell culture components and the product. A drawback of Raman spectroscopy is the long measurement time as the Raman effect is very weak compared to absorption phenomena. In recent years, the application of Raman spectroscopy to the downstream process became feasible due to an increased measurement speed by instrumentation improvements.

In Chapter 6, the application of both Raman and UV spectroscopy for monitoring the Protein A capture step was presented to compare both methods and evaluate the benefit of a combination of both methods by data fusion. As data fusion techniques, hierarchical PLS modeling, and CNNs were tested. If no preprocessing was applied to the spectral, it was shown that UV spectroscopy has a slightly better prediction accuracy in comparison to Raman spectroscopy. However, when the dynamic background spectrum subtraction (developed in Chapter 5) is applied, the prediction accuracy of the UV-based models improves 20-fold. For Raman spectroscopy, the background subtraction does not improve the prediction ability, probably due to the increased noise. It seems, that Raman spectroscopy is more selective than UV spectroscopy, which might make a background subtraction not helpful. A main drawback of Raman spectroscopy were the observed non-linearities in the spectra, which led to an increased number of LVs of the PLS models and a lower model prediction in comparison to the UV-based model. PLS models as linear regression techniques are only able to fit non-linearities to a certain extend. The larger the design space, the worse the linear approximation. CNNs were applied for non-linear regression to overcome this problem. CNNs can improve the prediction ability slightly in comparison to PLS-based methods, but the training of CNNs is challenging, requires a larger amount of data and might converge to different solutions. Besides the evaluation of PLS and CNN models, data fusion algorithms were tested to potentially improve the prediction accuracy by combining the sensitivity of UV spectroscopy with the selectivity of Raman spectroscopy. However, no improvement was observed in comparison to the solely UV-based models. Even though, the combination of the high signal-to-noise ratio of the UV measurements with the selectivity of the Raman measurements seems promising, for the purpose of quantifying only the mAb concentration, UV-based methods, especially in combination with a background spectrum subtraction, seem to be the best option. Nevertheless, UV spectroscopy cannot monitor other attributes of interest, like the buffer composition, aggregate content or disulfide bridges formation which makes other sensor concepts necessary.

An example for a process step, where multiple attributes need to monitored to facilitate process development and assure consistent quality in production processes, is the the combined process step UF/DF. For UF/DF processes not only the monitoring of the product concentration, but also of the buffer composition and aggregate content is important. In this study presented in Chapter 7, a lab-scale CFF device was equipped with a VP UV/Vis spectrometer, a light scattering photometer, and a microLDS. The protein concentration was measured by VP UV/Vis spectrometer. Due to the large concentration range of the UF/DF step, the use of the VP technology was necessary to avoid a detector saturation. The buffer exchange was monitored by density measurements of the microLDS. To calculate the apparent molecular weight, both the protein concentration determined by the VP UV/Vis spectrometer and the Static Light Scattering (SLS) signal measured by the light scattering photometer were necessary. The average hydrodynamic radius was calculated by the Dynamic Light Scattering (DLS) signal of the light scattering photometer, which was corrected by the viscosity determined by the microLDS. The setup was tested in three case studies. First, lysozyme was used in a UF-DF-UF proof-of-concept run to show the comparability of on-line and off-line measurements. Next, urea-induced changes in the protein size of Glucose Oxidase (GOx) were monitored during two DF steps. Finally, a case study was conducted with a mAb to show the full potential of this setup. Again, off-line and on-line measurements were in good agreement. The protein concentration could be monitored in-line in a large concentration range. The buffer-dependent increase in apparent molecular weight of the mAb could be shown during diafiltration, giving valuable information for process development and stability assessment. The developed sensor concept has shown to be a powerful tool for monitoring protein concentration, buffer exchange, apparent molecular weight and hydrodynamic radius. The in Chapter 7 presented study shows, that often it is not possible to measure all quality attributes of interest with only one sensor. Therefore smart sensor concepts are necessary to measure as many critical quality attributes as possible with the lowest amount of sensors.

Protein and buffer components concentration are critical quality attributes, that drive the design of the UF/DF process. Monitoring of the protein and buffer components concentration, therefore enables process automation of the UF/DF process by switching to the next process phase, when either the desired protein concentration or buffer component concentration are reached. Chapter 8 build on the process monitoring foundations presented in Chapter 7 with the addition of a relay valve to switch automatically between process phases based on the calculated quality attributes. Additionally, a Raman analyzer was implemented in the setup, because Raman spectroscopy is capable of measuring the protein concentration and a variety of Raman-active buffer components simultaneously. As the protein concentrations observed in UF/DF are significantly higher than during the Protein A load phase (presented in Chapter 6), Raman showed comparable results to UV spectroscopy for quantification of the protein concentration, even though the quantification was again based on the increase in the background spectrum/baseline. However, the noise level in both the buffer signal of the Raman spectra and density were too large to allow for a process automation without data preprocessing. Therefore, an EKF was implemented to combine mechanistic process knowledge with the data to estimate the state of the process more accurately and thereby allow for process automation.

In summary, the potential of different spectroscopic methods to monitor the downstream process was evaluated in this thesis. Commonly implemented univariate sensors were evaluated to close the gaps of spectroscopic techniques or facilitate the implementation of the PAT methods. Smart sensor concepts for the Protein A capture step and the UF/DF step were introduced. Additionally, data fusion techniques and new concepts in machine learning, especially CNNs and EKF were evaluated for their ability to improve the prediction ability of spectroscopic methods. While CNNs can automate the preprocessing optimization in the convolutional layers and apply non-linear regression techniques in the fully connected layers, the performance in the tested use case did not justify the computational effort in comparison to PLS models. This thesis facilitates the implementation of PAT in the downstream process of biologics, because solutions to specific monitoring needs of the capture and UF/DF step are presented. Therefore, a contribution to make critical biopharmaceutical drugs more affordable is made by the presented smart sensor concepts to improve production processes.

10 Outlook

As the high cost of pharmaceutical drugs and especially biopharmaceutical drugs are putting a significant economic burden on the healthcare systems worldwide, PAT methods seem to be a promising way to improve process robustness, product quality, and operational excellence by reducing manufacturing costs [108]. The implementation of PAT in the synthetic, pharmaceutical molecule field is rapidly advancing as PAT is a key enabler of the automated process control and digitalization. Automated process control is a key element for continuous processes, which promises additional great savings compared to traditional batch processes without PAT methods. As the need for PAT instruments, like spectrometers, is increasing, also the instrument manufacturers are increasing the efforts to supply the demand. This ranges from efforts to facilitate the communication of the instruments with control systems by the standardized implementation of Open Platform Communications (OPC) or to develop special instruments or methods, which work under Good Manufacturing Practice (GMP).

Due to the development of compact and high-power lasers, charge-coupled devices, fiber-optics probes, and further optical component enhancements in the past decades, Raman measurement times have decreased significantly [69, 161]. However as the measurement of proteins by Raman spectroscopy is still challenging due to the low scattering coefficients of proteins, further improvement will be necessary for a real-time implementation for chromatography. A possible improvement could be the development of liquid-core waveguides to improve the number of interacting proteins with the laser light. Additionally, a liquid-core waveguide could be realized as a single-use flow cell, which could facilitate the implementation of spectroscopic sensors in the future in a GMP environment. Using the resonance effect in the UV range is another possibility to selectively increase the Raman scatter coefficients of protein and thereby decrease measurement times. As UV Resonance Raman (UVRR) is a destructive technique due to the high energy input, this method could be interesting as on-line method.

Improvements in UV spectroscopy instrumentation are possible as well. An example is the commercialization of VP UV/Vis spectrometer, marketed as FlowVPE, which can measure in a dynamic concentration range. Replacing the monochromator with a polychromator and a diode array detector in the FlowVPE could improve measurement time in the future and reduce the number of moving parts in the VP spectroscopy system.

The main focus of PAT research has been the implementation of new spectroscopic techniques for the monitoring of processes. However, this thesis has shown, that even already implemented univariate sensors, like conductivity sensors, density sensor and single wavelength absorption sensors can give valuable information, if the information is extracted in the correct way and possibly fused with other measurements.

Besides the combination of univariate sensors with spectroscopic sensors, also the combination with process models is promising to compensate for selectivity issues or measurement frequency issues. Especially for UF/DF processes the combination PAT data with Model Predictive Control (MPC) seems favorable to control critical process parameters to compensate for aggregation during the process or delayed buffer exchange effects.

An unresolved challenge is the real-time quantification of low concentration contaminants, where the higher concentrated product hinders the measurement, like Host Cell Protein (HCP) or Deoxyribonucleic Acid (DNA) quantification. Different techniques, like Liquid Chromatography coupled to Mass Spectrometry (LC/MS) [348] or a microfluidic based on-line enzyme-linked immunosorbent assay (ELISA) [349], have shown first promising results for a timely quantification of HCPs. High-throughput quantitative polymerase chain reaction (qPCR) techniques have made improvements for quantification of DNA.

Regardless of the remaining challenges, the implementation of PAT methods in biopharmaceutical processes is increasing. The main application nowadays is the monitoring of the upstream process, but the gained knowledge will increase the confidence in PAT methods and demonstrate the benefits. A wide application to the downstream process is then just a matter of time.

References

[1] World Health Organization, „The selection and use of essential medicines: report of the WHO Expert Committee on Selection and Use of Essential Medicines, 2019 (including the 21st WHO Model List of Essential Medicines and the 7th WHO Model List of Essential Medicines for Children)“, 2019 (cit. on p. 1).

[2] D. Trapani and G. Curigliano, *How can biosimilars change the trajectory of breast cancer therapy?*, 2020 (cit. on p. 1).

[3] World Health Organisation. (2019). Essential medicines and health products, [Online]. Available: `https://www.who.int/medicines/services/essmedicines_def/en/` (visited on 03/26/2020) (cit. on p. 1).

[4] C. Robert, J. Schachter, G. V. Long, A. Arance, J. J. Grob, L. Mortier, A. Daud, M. S. Carlino, C. McNeil, M. Lotem, *et al.*, „Pembrolizumab versus ipilimumab in advanced melanoma“, *New England Journal of Medicine*, vol. 372, no. 26, pp. 2521–2532, 2015 (cit. on p. 1).

[5] C. Robert, G. V. Long, B. Brady, C. Dutriaux, M. Maio, L. Mortier, J. C. Hassel, P. Rutkowski, C. McNeil, E. Kalinka-Warzocha, *et al.*, „Nivolumab in previously untreated melanoma without BRAF mutation“, *New England journal of medicine*, vol. 372, no. 4, pp. 320–330, 2015 (cit. on p. 1).

[6] S. Shak, „Overview of the trastuzumab (Herceptin) anti-HER2 monoclonal antibody clinical program in HER2-overexpressing metastatic breast cancer. Herceptin Multinational Investigator Study Group.“, in *Seminars in oncology*, vol. 26, 1999, pp. 71–77 (cit. on p. 1).

[7] D. J. Crommelin, R. D. Sindelar, and B. Meibohm, *Pharmaceutical Biotechnology: Fundamentals and Applications*, Fifth Edition. Springer Nature Switzerland, 2019 (cit. on p. 1).

[8] G. Jagschies, E. Lindskog, K. Lacki, and P. M. Galliher, *Biopharmaceutical Processing: Development, Design, and Implementation of Manufacturing Processes*. Elsevier, 2018 (cit. on pp. 1, 134).

[9] R. Shapiro, K. Singh, and M. Mukim, *Generic biological treatments and the associated cost savings*, 2008 (cit. on p. 1).

[10] E. A. Blackstone and J. P. Fuhr Jr, „Biopharmaceuticals: the economic equation“, *Biotechnology healthcare*, vol. 4, no. 6, p. 41, 2007 (cit. on p. 1).

[11] J. A. DiMasi and H. G. Grabowski, „The cost of biopharmaceutical R&D: is biotech different?“, *Managerial and decision Economics*, vol. 28, no. 4-5, pp. 469–479, 2007 (cit. on p. 1).

[12] M. R. Trusheim, M. L. Aitken, and E. R. Berndt, „Characterizing markets for biopharmaceutical innovations: do biologics differ from small molecules?“, in *Forum for Health Economics & Policy*, vol. 13, 2010 (cit. on pp. 1, 156).

[13] G. Subramanian, *Biopharmaceutical production technology, 2 volume set*. John Wiley & Sons, 2012, vol. 2 (cit. on p. 1).

[14] M. Schiestl, T. Stangler, C. Torella, T. Čepeljnik, H. Toll, and R. Grau, „Acceptable changes in quality attributes of glycosylated biopharmaceuticals“, *Nature biotechnology*, vol. 29, no. 4, pp. 310–312, 2011 (cit. on p. 2).

[15] A. Rathore, R. Bhambure, and V. Ghare, „Process analytical technology (PAT) for biopharmaceutical products“, *Analytical and bioanalytical chemistry*, vol. 398, no. 1, pp. 137–154, 2010 (cit. on p. 2).

[16] J. Glassey, K. V. Gernaey, C. Clemens, T. W. Schulz, R. Oliveira, G. Striedner, and C.-F. Mandenius, „Process analytical technology (PAT) for biopharmaceuticals“, *Biotechnology Journal*, vol. 6, no. 4, pp. 369–377, 2011 (cit. on p. 2).

[17] U.S. Food and Drug Administration, „Guidance for Industry PAT - A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance“, Department of Health and Human Services, Tech. Rep., 2004 (cit. on pp. 2, 11, 35, 134).

[18] R. L. Fahrner and G. S. Blank, „Real-time control of antibody loading during protein A affinity chromatography using an on-line assay“, *Journal of Chromatography A*, vol. 849, no. 1, pp. 191–196, 1999 (cit. on pp. 2, 66, 67, 134).

[19] R. L. Fahrner, P. M. Lester, G. S. Blank, and D. H. Reifsnyder, „Real-time control of purified product collection during chromatography of recombinant human insulin-like growth factor—I using an on-line assay“, *Journal of Chromatography A*, vol. 827, no. 1, pp. 37–43, 1998 (cit. on pp. 2, 66, 134).

[20] R. L. Fahrner and G. S. Blank, „Real-time monitoring of recombinant antibody breakthrough during Protein A affinity chromatography“, *Biotechnology and applied biochemistry*, vol. 29, no. 2, pp. 109–112, 1999 (cit. on pp. 2, 134).

[21] A. S. Rathore, M. Yu, S. Yeboah, and A. Sharma, „Case study and application of process analytical technology (PAT) towards bioprocessing: Use of on-line high-performance liquid chromatography (HPLC) for making real-time pooling decisions for process chromatography“, *Biotechnology and bioengineering*, vol. 100, no. 2, pp. 306–316, 2008 (cit. on p. 2).

[22] A. S. Rathore, R. Wood, A. Sharma, and S. Dermawan, „Case study and application of process analytical technology (PAT) towards bioprocessing: II. Use of ultra-performance liquid chromatography (UPLC) for making real-time pooling decisions for process chromatography“, *Biotechnology and bioengineering*, vol. 101, no. 6, pp. 1366–1374, 2008 (cit. on p. 2).

[23] J. M. Woof, „Tipping the scales toward more effective antibodies“, *Science*, vol. 310, no. 5753, pp. 1442–1443, 2005 (cit. on p. 3).

[24] J. W. Goding, *Monoclonal antibodies: principles and practice.* Elsevier, 1996 (cit. on p. 3).

[25] S. Flatman, I. Alam, J. Gerard, and N. Mussa, „Process analytics for purification of monoclonal antibodies“, *Journal of Chromatography A*, vol. 848, no. 1, pp. 79–87, 2007 (cit. on pp. 3, 34, 40).

[26] A. A. Shukla and J. Thömmes, „Recent advances in large-scale production of monoclonal antibodies and related proteins“, *Trends Biotechnol*, vol. 28, no. 5, pp. 253–261, 2010 (cit. on pp. 3, 7).

[27] G. Köhler and C. Milstein, „Continuous cultures of fused cells secreting antibody of predefined specificity“, *nature*, vol. 256, no. 5517, pp. 495–497, 1975 (cit. on p. 3).

[28] K. C. T. Nguyen, M.-S. Yoo, S.-H. Han, S.-H. Kwon, Y.-H. Park, and B.-S. Yoon, „Generation of Specific Monoclonal Antibody against Recombinant Major Royal Jelly Protein 1 (MRJP1) from Honeybee (Apis mellifera)“, *Journal of Apiculture*, vol. 25, no. 2, pp. 129–135, 2010 (cit. on p. 3).

[29] S. Hansen, N. Brestrich, A. Staby, and J. Hubbuch, *Mid-UV Protein Absorption Spectra and Partial Least Squares Regression as QbD and PAT Tool.* Wiley Online Library, 2017, pp. 501–536 (cit. on pp. 3, 14, 15).

[30] M. M. Cox and D. L. Nelson, *Lehninger principles of biochemistry.* Wh Freeman, 2008 (cit. on pp. 3, 4).

[31] L. Pauling, R. B. Corey, and H. R. Branson, „The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain“, *Proceedings of the National Academy of Sciences*, vol. 37, no. 4, pp. 205–211, 1951 (cit. on pp. 3, 4).

[32] K. Jungermann and H. Möhler, *Biochemie.* Berlin [u.a.]: Springer, 1980, vol. [1]: Ein Lehrbuch für Studierende der Medizin, Biologie und Pharmazie (cit. on p. 4).

[33] J. T. Pelton and L. R. McLean, „Spectroscopic methods for analysis of protein secondary structure“, *Analytical biochemistry*, vol. 277, no. 2, pp. 167–176, 2000 (cit. on p. 4).

[34] W. Jiskoot and D. Crommelin, Eds., *Methods for structural analysis of protein pharmaceuticals.* American Association of Pharmaceutical Scientists, 2005 (cit. on pp. 4, 14, 16, 18–20, 39–41, 111, 139, 140, 152, 162, 167).

[35] C. A. Janeway, P. Travers, M. Walport, and D. J. Capra, *Immunobiology.* Taylor & Francis Group UK: Garland Science, 2001 (cit. on pp. 5, 6).

[36] J. M. Woof and D. R. Burton, „Human antibody-Fc receptor interactions illuminated by crystal structures“, *Nat Rev Immunol*, vol. 4, no. 2, pp. 89–99, 2004 (cit. on p. 5).

[37] B. Bröker, C. Schütt, B. Fleischer, and VISUV., *Grundwissen Immunologie.* Springer, 2019 (cit. on pp. 5, 6).

[38] J. Bratt, A. Linderholm, B. Monroe, and S. Chamow, „Therapeutic IgG-like bispecific antibodies: modular versatility and manufacturing challenges, part 1“, *Bioprocess International*, vol. 15, pp. 36–42, 2017 (cit. on p. 5).

[39] U. Brinkmann and R. E. Kontermann, „The making of bispecific antibodies“, in *MAbs*, Taylor & Francis, vol. 9, 2017, pp. 182–212 (cit. on p. 6).

[40] S. S. Hosseini, S. Khalili, B. Baradaran, N. Bidar, M.-A. Shahbazi, J. Mosafer, M. Hashemzaei, A. Mokhtarzadeh, and M. R. Hamblin, „Bispecific monoclonal antibodies for targeted immunotherapy of solid tumors: Recent advances and clinical trials“, *International Journal of Biological Macromolecules*, 2020 (cit. on p. 6).

[41] D. Müller and R. E. Kontermann, „Recombinant bispecific antibodies for cellular cancer immunotherapy.“, *Current opinion in molecular therapeutics*, vol. 9, no. 4, pp. 319–326, 2007 (cit. on p. 6).

[42] P. Chames and D. Baty, „Bispecific antibodies for cancer therapy: the light at the end of the tunnel?“, in *MAbs*, Taylor & Francis, vol. 1, 2009, pp. 539–547 (cit. on p. 6).

[43] A. D. Tustian, L. Laurin, H. Ihre, T. Tran, R. Stairs, and H. Bak, „Development of a novel affinity chromatography resin for platform purification of bispecific antibodies with modified protein a binding avidity“, *Biotechnology progress*, vol. 34, no. 3, pp. 650–658, 2018 (cit. on pp. 6, 7).

[44] S. K. Gupta, S. K. Srivastava, A. Sharma, V. H. Nalage, D. Salvi, H. Kushwaha, N. B. Chitnis, and P. Shukla, „Metabolic engineering of CHO cells for the development of a robust protein production platform“, *PloS one*, vol. 12, no. 8, e0181455, 2017 (cit. on p. 6).

[45] H. Chmiel, R. Takors, and D. Weuster-Botz, *Bioprozesstechnik*. Springer, 2018 (cit. on p. 6).

[46] A. A. Shukla, B. Hubbard, T. Tressel, S. Guhan, and D. Low, „Downstream processing of monoclonal antibodies—application of platform approaches“, *Journal of Chromatography B*, vol. 848, no. 1, pp. 28–39, 2007 (cit. on pp. 7, 8, 40, 66).

[47] S. Kozlowski and P. Swann, „Current and future issues in the manufacturing and development of monoclonal antibodies“, *Advanced drug delivery reviews*, vol. 58, no. 5-6, pp. 707–722, 2006 (cit. on p. 7).

[48] H. F. Liu, J. Ma, C. Winter, and R. Bayer, „Recovery and purification process development for monoclonal antibody production“, in *MAbs*, Taylor & Francis, vol. 2, 2010, pp. 480–499 (cit. on pp. 7, 8).

[49] R. D. R. Tarrant, M. L. Velez-Suberbie, A. S. Tait, C. M. Smales, and D. G. Bracewell, „Host cell protein adsorption characteristics during protein A chromatography“, *Biotechnology progress*, vol. 28, no. 4, pp. 1037–1044, 2012 (cit. on pp. 7, 66).

[50] A. A. Shukla and P. Hinckley, „Host cell protein clearance during protein A chromatography: Development of an improved column wash step“, *Biotechnology progress*, vol. 24, no. 5, pp. 1115–1121, 2008 (cit. on pp. 7, 9, 66, 73, 87).

[51] G. V. Research, „Downstream Processing Market Size, Share and Trends Analysis Report By Product (Chromatography Systems, Filters), By Technique (Purification, Formulation), By Application, By Region, And Segment Forecasts, 2021 - 2028“, 2021 (cit. on p. 8).

[52] M. Franzreb, E. Müller, and J. Vajda, „Cost estimation for protein A chromatography: An in silico approach to Mab purification strategy“, *BioProcess International*, vol. 12, pp. 44–52, 2014 (cit. on p. 8).

[53] Tosoh Bioscience, „Protein A Chromatography–The Process Economics Driver in mAb Manufacturing“, (cit. on p. 8).

[54] M. Graille, E. A. Stura, A. L. Corper, B. J. Sutton, M. J. Taussig, J.-B. Charbonnier, and G. J. Silverman, „Crystal structure of a Staphylococcus aureus protein A domain complexed with the Fab fragment of a human IgM antibody: structural basis for recognition of B-cell receptors and superantigen activity“, *Proceedings of the National Academy of Sciences*, vol. 97, no. 10, pp. 5399–5404, 2000 (cit. on p. 8).

[55] B. Sheth, „Characterisation of chromatography adsorbents for antibody bioprocessing“, PhD thesis, UCL (University College London), 2009 (cit. on p. 9).

[56] D. Karst, F. Steinebach, and M. Morbidelli, „Integrated continuous processing for the manufacture of monoclonal antibodies“, in *Integrated Continuous Biomanufacturing II, Chetan Goudar, Amgen Inc. Suzanne Farid, University College London Christopher Hwang, Genzyme-Sanofi Karol Lacki, Novo Nordisk Eds, ECI Symposium Series*, 2015 (cit. on pp. 10, 67).

[57] U. Gottschalk, *Process scale purification of antibodies.* John Wiley & Sons, 2017 (cit. on pp. 10, 134, 156, 157).

[58] C. Charcosset, *Membrane processes in biotechnology and pharmaceutics.* Elsevier, 2012 (cit. on p. 10).

[59] C. Charcosset, „Membrane processes in biotechnology: an overview“, *Biotechnology advances*, vol. 24, no. 5, pp. 482–492, 2006 (cit. on pp. 10, 156).

[60] K. Ahrer, A. Buchacher, G. Iberer, and A. Jungbauer, „Effects of ultra-/diafiltration conditions on present aggregates in human immunoglobulin G preparations“, *Journal of Membrane Science*, vol. 274, no. 1-2, pp. 108–115, 2006 (cit. on p. 10).

[61] A. Steele and J. Arias, „Accounting for the Donnan effect in diafiltration optimization for high-concentration UFDF applications“, *BioProcess International*, vol. 12, no. 1, pp. 50–54, 2014 (cit. on pp. 10, 164).

[62] K. A. Bakeev, *Process analytical technology: spectroscopic tools and implementation strategies for the chemical and pharmaceutical industries.* John Wiley & Sons, 2010 (cit. on pp. 11, 23, 35, 36, 41, 49, 53, 61, 134, 135, 158).

[63] J. Workman, M. Koch, and D. J. Veltkamp, „Process analytical chemistry“, *Analytical Chemistry*, vol. 75, no. 12, pp. 2859–2876, 2003 (cit. on p. 11).

[64] D. C. Hinz, „Process analytical technologies in the pharmaceutical industry: the FDA's PAT initiative“, *Analytical and Bioanalytical Chemistry*, vol. 384, no. 5, pp. 1036–1042, 2006 (cit. on pp. 11, 12, 134).

[65] ICH Quality Implementation Working Group, „Points to Consider (R2)“, International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use, Tech. Rep., 2011 (cit. on p. 12).

[66] S. Chhatre, S. S. Farid, J. Coffman, P. Bird, A. R. Newcombe, and N. J. Titchener-Hooker, „How implementation of Quality by Design and advances in Biochemical Engineering are enabling efficient bioprocess development and manufacture“, *Journal of Chemical Technology and Biotechnology*, vol. 86, no. 9, pp. 1125–1129, 2011 (cit. on p. 12).

[67] M. Rüdt, T. Briskot, and J. Hubbuch, „Advances in downstream processing of biologics-Spectroscopy: An emerging process analytical technology“, *J. Chromatogr. A*, vol. 1490, pp. 2–9, 2017 (cit. on pp. 12, 34, 35, 41, 42, 44, 134).

[68] P. W. Atkins and J. De Paula, *Atkins Physical chemistry*, 8. ed. Oxford [u.a.]: Oxford University Press, 2006 (cit. on p. 13).

[69] I. R. Lewis and H. Edwards, *Handbook of Raman spectroscopy: from the research laboratory to the process line.* CRC press, 2001 (cit. on pp. 13, 17, 18, 43, 189).

[70] M. Rüdt, „Spectroscopy as process analytical technology for preparative protein purification“, dissertation, Karlsruhe Institute of Technology, 2019 (cit. on p. 14).

[71] A. S. Rathore, R. Bhambure, and V. Ghare, „Process analytical technology (PAT) for biopharmaceutical products“, *Anal Bioanal Chem*, vol. 398, no. 1, pp. 137–54, 2010 (cit. on p. 15).

[72] N. Brestrich, T. Briskot, A. Osberghaus, and J. Hubbuch, „A tool for selective inline quantification of co-eluting proteins in chromatography using spectral analysis and partial least squares regression“, *Biotechnology and Bioengineering*, vol. 111, no. 7, pp. 1365–1373, 2014 (cit. on pp. 15, 41, 68, 69, 134).

[73] 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“, *Biotechnology and Bioengineering*, vol. 112, no. 7, pp. 1406–1416, 2015 (cit. on pp. 15, 41, 58, 68, 101, 134).

[74] D. Wetlaufer, „Ultraviolet spectra of proteins and amino acids“, in *Advances in protein chemistry*, vol. 17, Elsevier, 1963, pp. 303–390 (cit. on p. 15).

[75] P. R. Griffiths and J. A. De Haseth, *Fourier transform infrared spectrometry.* John Wiley & Sons, 2007, vol. 171 (cit. on p. 16).

[76] M. Jackson and H. H. Mantsch, „The Use and Misuse of FTIR Spectroscopy in the Determination of Protein Structure“, *Critical Reviews in Biochemistry and Molecular Biology*, vol. 30, no. 2, pp. 95–120, 1995 (cit. on p. 16).

[77] J. Kong and S. Yu, „Fourier Transform Infrared Spectroscopic Analysis of Protein Secondary Structures“, *Acta Biochimica et Biophysica Sinica*, vol. 39, no. 8, pp. 549–559, 2007 (cit. on p. 16).

[78] Adochitei and Drochiuio, „Rapid characterization of peptide secondary structure by FT-IR spectroscopy“, *Revue Roumaine de Chimie*, vol. 56, no. 8, pp. 783–791, 2011 (cit. on p. 16).

[79] F. Capito, R. Skudas, H. Kolmar, and C. Hunzinger, „At-line mid infrared spectroscopy for monitoring downstream processing unit operations“, *Process Biochemistry*, vol. 50, no. 6, pp. 997–1005, 2015 (cit. on p. 16).

[80] J. B. Wiester, „Investigating the Similarities and Differences among UV/Vis, Infrared, Fluorescence, and Raman Spectroscopies through Discussion of Light–Matter Interactions“, in *Raman Spectroscopy in the Undergraduate Curriculum*, ACS Publications, 2018, pp. 13–33 (cit. on p. 18).

[81] E. A. Carter and H. G. Edwards, *Biological applications of Raman spectroscopy.* Marcel Dekker, Inc.: New York, NY, USA, 2001, vol. 24 (cit. on p. 18).

[82] W. Burchard, „Static and dynamic light scattering from branched polymers and biopolymers“, in *Light scattering from polymers*, Springer, 1983, pp. 1–124 (cit. on p. 18).

[83] C. Tanford, „Light scattering“, *Physical chemistry of macromolecules*, pp. 275–316, 1961 (cit. on p. 19).

[84] T. Arakawa, J. S. Philo, D. Ejima, K. Tsumoto, and F. Arisaka, „Aggregation analysis of therapeutic proteins, part 2“, *BioProcess International*, vol. 5, no. 4, pp. 36–47, 2007 (cit. on pp. 19, 147).

[85] W. Schärtl, *Light scattering from polymer solutions and nanoparticle dispersions.* Springer Science & Business Media, 2007 (cit. on p. 20).

[86] U. Nobbmann, M. Connah, B. Fish, P. Varley, C. Gee, S. Mulot, J. Chen, L. Zhou, Y. Lu, F. Sheng, *et al.*, „Dynamic light scattering as a relative tool for assessing the molecular integrity and stability of monoclonal antibodies“, *Biotechnology and genetic engineering reviews*, vol. 24, no. 1, pp. 117–128, 2007 (cit. on pp. 20, 63).

[87] A. B. Leung, K. I. Suh, and R. R. Ansari, „Particle-size and velocity measurements in flowing conditions using dynamic light scattering“, *Applied optics*, vol. 45, no. 10, pp. 2186–2190, 2006 (cit. on p. 20).

[88] A. P. Minton, „Recent applications of light scattering measurement in the biological and biopharmaceutical sciences“, *Analytical biochemistry*, vol. 501, p. 4, 2016 (cit. on p. 20).

[89] D. E. Koppel, „Analysis of Macromolecular Polydispersity in Intensity Correlation Spectroscopy: The Method of Cumulants“, *The Journal of Chemical Physics*, vol. 57, no. 11, pp. 4814–4820, 12/1972 (cit. on pp. 21, 140).

[90] H. Holthoff, S. U. Egelhaaf, M. Borkovec, P. Schurtenberger, and H. Sticher, „Coagulation rate measurements of colloidal particles by simultaneous static and dynamic light scattering“, *Langmuir*, vol. 12, no. 23, pp. 5541–5549, 1996 (cit. on p. 21).

[91] R. Wehrens, *Chemometrics with R : Multivariate Data Analysis in the Natural Sciences and Life Sciences*, Berlin, Heidelberg, 2011 (cit. on pp. 21–23).

[92] C. M. Bishop, *Pattern recognition and machine learning.* Springer, 2006 (cit. on pp. 21, 25–27, 46).

[93] H. Hotelling, „Analysis of a complex of statistical variables into principal components.“, *Journal of educational psychology*, vol. 24, no. 6, p. 417, 1933 (cit. on p. 22).

[94] L. L. Thurstone, „Multiple factor analysis.“, *Psychological review*, vol. 38, no. 5, p. 406, 1931 (cit. on p. 22).

[95] W. Kessler, *Multivariate datenanalyse: für die pharma, bio-und Prozessanalytik.* John Wiley & Sons, 2007 (cit. on pp. 22, 24, 48, 49, 51, 53, 84).

[96] L. Eriksson, E. Johansson, N. Kettaneh-Wold, J. Trygg, C. Wikström, and S. Wold, *Multi-and megavariate data analysis.* Umetrics Sweden, 2006, vol. 1 (cit. on pp. 22, 23, 47, 49, 51–53, 55, 70, 84, 93, 108, 109).

[97] S. Wold, J. Trygg, A. Berglund, and H. Antti, „Some recent developments in PLS modeling“, *Chemometrics and Intelligent Laboratory Systems*, vol. 58, no. 2, pp. 131–150, 2001 (cit. on p. 23).

[98] S. Wold, A. Ruhe, H. Wold, and W. Dunn III, „The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses“, *SIAM Journal on Scientific and Statistical Computing*, vol. 5, no. 3, pp. 735–743, 1984 (cit. on p. 24).

[99] S. Wold, M. Sjöström, and L. Eriksson, „PLS-regression: a basic tool of chemometrics“, *Chemometrics and Intelligent Laboratory Systems*, vol. 58, no. 2, pp. 109–130, 2001 (cit. on pp. 24, 51, 52).

[100] W. S. McCulloch and W. Pitts, „A logical calculus of the ideas immanent in nervous activity“, *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, 1943 (cit. on p. 25).

[101] M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken, „Multilayer feedforward networks with a nonpolynomial activation function can approximate any function“, *Neural networks*, vol. 6, no. 6, pp. 861–867, 1993 (cit. on p. 27).

[102] M. Kessel, „The problems with today's pharmaceutical business—an outsider's view", *Nature biotechnology*, vol. 29, no. 1, pp. 27–33, 2011 (cit. on pp. 34, 78).

[103] J. W. Scannell, A. Blanckley, H. Boldon, and B. Warrington, „Diagnosing the decline in pharmaceutical R&D efficiency", *Nat. Rev. Drug Discovery*, vol. 11, no. 3, p. 191, 2012 (cit. on p. 34).

[104] A. L. Grilo and A. Mantalaris, „The increasingly human and profitable monoclonal antibody market", *Trends in biotechnology*, vol. 37, no. 1, pp. 9–16, 2019 (cit. on pp. 34, 78).

[105] P. Gagnon, „Technology trends in antibody purification", *Journal of chromatography A*, vol. 1221, pp. 57–70, 2012 (cit. on p. 34).

[106] M. Grebe, M. Rüßmann, M. Leyh, and M. R. Franke, „Digital Maturity Is Paying Off", *Boston Consulting Group*, 2018 (cit. on p. 34).

[107] V. Steinwandter, D. Borchert, and C. Herwig, „Data science tools and applications on the way to Pharma 4.0", *Drug Discov. Today*, 2019 (cit. on p. 34).

[108] S. Schlack, „Addressing the Challenges of Developing Biopharmaceutical Drugs", *BioProcess International*, vol. 14, no. 10, pp. 72–74, 2016 (cit. on pp. 34, 189).

[109] S. Laske, A. Paudel, O. Scheibelhofer, S. Sacher, T. Hoermann, J. Khinast, A. Kelly, J. Rantannen, O. Korhonen, and F. Stauffer, „A review of PAT strategies in secondary solid oral dosage manufacturing of small molecules", *Journal of Pharmaceutical Sciences*, vol. 106, no. 3, pp. 667–712, 2017 (cit. on p. 34).

[110] L. L. Simon, H. Pataki, G. Marosi, F. Meemken, K. Hungerbühler, A. Baiker, S. Tummala, B. Glennon, M. Kuentz, G. Steele, *et al.*, „Assessment of recent process analytical technology (PAT) trends: a multiauthor review", *Organic Process Research & Development*, vol. 19, no. 1, pp. 3–62, 2015 (cit. on p. 34).

[111] M. S. Hong, K. A. Severson, M. Jiang, A. E. Lu, J. C. Love, and R. D. Braatz, „Challenges and opportunities in biopharmaceutical manufacturing control", *Computers & Chemical Engineering*, vol. 110, pp. 106–114, 2018 (cit. on p. 34).

[112] H. Liu, G. Gaza-Bulseco, D. Faldu, C. Chumsae, and J. Sun, „Heterogeneity of monoclonal antibodies", *Journal of Pharmaceutical Sciences*, vol. 97, no. 7, pp. 2426–2447, 2008 (cit. on p. 34).

[113] A. Tiwari, N. Kateja, S. Chanana, and A. S. Rathore, „Use of HPLC as an Enabler of Process Analytical Technology in Process Chromatography“, *Anal. Chem.*, vol. 90, no. 13, pp. 7824–7829, 2018 (cit. on p. 34).

[114] 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“, *Chemical Engineering Science*, vol. 176, pp. 157–164, 2018 (cit. on pp. 35, 41–43, 58, 101).

[115] S. Großhans, M. Rüdt, A. Sanden, N. Brestrich, J. Morgenstern, S. Heissler, and J. Hubbuch, „In-line Fourier-transform infrared spectroscopy as a versatile process analytical technology for preparative protein chromatography“, *J. Chromatogr. A*, vol. 1547, pp. 37–44, 2018 (cit. on pp. 35, 41).

[116] N. Walch, T. Scharl, E. Felföldi, D. G. Sauer, M. Melcher, F. Leisch, A. Dürauer, and A. Jungbauer, „Prediction of the Quantity and Purity of an Antibody Capture Process in Real Time“, *Biotechnology Journal*, p. 1 800 521, 2019 (cit. on pp. 35, 44, 45, 54, 122).

[117] D. G. Sauer, M. Melcher, M. Mosor, N. Walch, M. Berkemeyer, T. Scharl-Hirsch, F. Leisch, A. Jungbauer, and A. Dürauer, „Real-time monitoring and model-based prediction of purity and quantity during a chromatographic capture of fibroblast growth factor 2“, *Biotechnology and Bioengineering*, 2019 (cit. on pp. 35, 45, 122).

[118] M. Liggins II, D. Hall, and J. Llinas, *Handbook of multisensor data fusion: theory and practice.* CRC press, 2017 (cit. on pp. 35, 57, 102, 109).

[119] M. Sokolov, F. Feidl, M. Morbidelli, and A. Butte, „Big data in biopharmaceutical process development vice or virtue?“, *Chimica Oggi – Chemistry Today*, vol. 36, no. 5, pp. 26–29, 2018 (cit. on p. 35).

[120] S. Wold, „Chemometrics; what do we mean with it, and what do we want from it?“, *Chemometrics and Intelligent Laboratory Systems*, vol. 30, no. 1, pp. 109–115, 1995 (cit. on p. 35).

[121] E. Borràs, J. Ferré, R. Boqué, M. Mestres, L. Aceña, and O. Busto, „Data fusion methodologies for food and beverage authentication and quality assessment–A review“, *Analytica Chimica Acta*, vol. 891, pp. 1–14, 2015 (cit. on pp. 35, 55, 56, 102, 108).

[122] R. Luttmann, D. G. Bracewell, G. Cornelissen, K. V. Gernaey, J. Glassey, V. C. Hass, C. Kaiser, C. Preusse, G. Striedner, and C.-F. Mandenius, „Soft sensors in bioprocessing: a status report and recommendations“, *Biotechnology journal*, vol. 7, no. 8, pp. 1040–1048, 2012 (cit. on p. 35).

[123] P. Roch and C.-F. Mandenius, „On-line monitoring of downstream bioprocesses“, *Current Opinion in Chemical Engineering*, vol. 14, pp. 112–120, 2016 (cit. on pp. 35, 134).

[124] A. S. Rathore and G. Kapoor, „Application of process analytical technology for downstream purification of biotherapeutics“, *Journal of Chemical Technology and Biotechnology*, vol. 90, no. 2, pp. 228–236, 02/2015 (cit. on pp. 35, 134).

[125] R. W. Kessler, *Prozessanalytik: Strategien und Fallbeispiele aus der industriellen Praxis.* John Wiley & Sons, 2012 (cit. on pp. 36, 92).

[126] J. Chalmers and P. Griffiths, *Handbook of Vibrational Spectroscopy,* Wiley, 2002, vol. 5 (cit. on pp. 36, 40).

[127] D. J. Segelstein, „The complex refractive index of water“, PhD thesis, University of Missouri–Kansas City, 1981 (cit. on p. 37).

[128] T. Scientific, „Extinction Coefficients: A guide to understanding extinction coefficients, with emphasis on spectrophotometric determination of protein concentration“, *Thermo Scientific*, vol. 6, 2012 (cit. on p. 38).

[129] J. R. Lakowicz, *Principles of fluorescence spectroscopy.* Springer Science & Business Media, 2013 (cit. on pp. 38, 43, 62, 115).

[130] G. ElMasry and S. Nakauchi, „Prediction of meat spectral patterns based on optical properties and concentrations of the major constituents“, *International Journal of Food Sciences and Nutrition*, vol. 4, no. 2, pp. 269–283, 2016 (cit. on p. 38).

[131] Z.-Q. Wen, „Raman spectroscopy of protein pharmaceuticals“, *Journal of Pharmaceutical Sciences*, vol. 96, no. 11, pp. 2861–2878, 2007 (cit. on pp. 38, 40, 101, 111, 115).

[132] S. A. Oladepo, K. Xiong, Z. Hong, S. A. Asher, J. Handen, and I. K. Lednev, „UV resonance raman investigations of peptide and protein structure and dynamics“, *Chemical Reviews*, vol. 112, no. 5, pp. 2604–2628, 2012 (cit. on p. 39).

[133] A. Barth, „Infrared spectroscopy of proteins“, *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, vol. 1767, no. 9, pp. 1073–1101, 2007 (cit. on p. 39).

[134] J. Popp, V. V. Tuchin, A. Chiou, and S. H. Heinemann, *Handbook of biophotonics: Vol. 2: Photonics for health care.* John Wiley & Sons, 2011, vol. 2 (cit. on pp. 39, 40).

[135] J. E. Noble and M. J. Bailey, „Quantitation of protein“, in *Methods in enzymology*, vol. 463, Elsevier, 2009, pp. 73–95 (cit. on p. 39).

[136] D. A. Skoog, D. M. West, F. J. Holler, and S. R. Crouch, *Fundamentals of analytical chemistry.* Nelson Education, 2013 (cit. on p. 40).

[137] C. Parker and W. Rees, „Fluorescence spectrometry. A review“, *Analyst*, vol. 87, no. 1031, pp. 83–111, 1962 (cit. on p. 40).

[138] M. Swartz, „HPLC detectors: a brief review“, *Journal of Liquid Chromatography & Related Technologies*, vol. 33, no. 9-12, pp. 1130–1150, 2010 (cit. on pp. 40, 92).

[139] I. López-Peña, B. S. Leigh, D. E. Schlamadinger, and J. E. Kim, „Insights into protein structure and dynamics by ultraviolet and visible resonance Raman spectroscopy“, *Biochemistry*, vol. 54, no. 31, pp. 4770–4783, 2015 (cit. on p. 40).

[140] G. Den Boef and A. Hulanicki, „Recommendations for the usage of selective, selectivity and related terms in analytical chemistry“, *Pure and Applied Chemistry*, vol. 55, no. 3, pp. 553–556, 1983 (cit. on p. 40).

[141] J. Vessman, R. I. Stefan, J. F. Van Staden, K. Danzer, W. Lindner, D. T. Burns, A. Fajgelj, and H. Müller, „Selectivity in analytical chemistry (IUPAC Recommendations 2001)“, *Pure and Applied Chemistry*, vol. 73, no. 8, pp. 1381–1386, 2001 (cit. on pp. 40, 41).

[142] R. W. Kessler, W. Kessler, and E. Zikulnig-Rusch, „A critical summary of spectroscopic techniques and their robustness in industrial PAT applications“, *Chemie Ingenieur Technik*, vol. 88, no. 6, pp. 710–721, 2016 (cit. on p. 40).

[143] S. K. Hansen, B. Jamali, and J. Hubbuch, „Selective high throughput protein quantification based on UV absorption spectra“, *Biotechnology and Bioengineering*, vol. 110, no. 2, pp. 448–460, 2013 (cit. on pp. 40, 41).

[144] M. Rüdt, P. Vormittag, N. Hillebrandt, and J. Hubbuch, „Process monitoring of virus-like particle reassembly by diafiltration with UV/Vis spectroscopy and light scattering“, *Biotechnology and Bioengineering*, vol. 116, no. 6, pp. 1366–1379, 2019 (cit. on pp. 41, 44, 136, 139, 150, 159, 161).

[145] J. Bandekar, „Amide modes and protein conformation“, *Biochimica et Biophysica Acta*, vol. 1120, no. 2, pp. 123–143, 1992 (cit. on p. 41).

[146] M. Jackson and H. H. Mantsch, „The use and misuse of FTIR spectroscopy in the determination of protein structure“, *Critical Reviews in Biochemistry and Molecular Biology*, vol. 30, no. 2, pp. 95–120, 1995 (cit. on p. 41).

[147] A. Barth, „The infrared absorption of amino acid side chains“, *Prog. Biophys. Mol. Biol.*, vol. 74, no. 3-5, pp. 141–173, 2000 (cit. on p. 41).

[148] A. Sanden, S. Suhm, M. Rüdt, and J. Hubbuch, „Fourier-transform infrared spectroscopy as a process analytical technology for near real time in-line estimation of the degree of PEGylation in chromatography“, *J. Chromatogr. A*, p. 460 410, 2019 (cit. on p. 41).

[149] B. C. Smith, *Fundamentals of Fourier transform infrared spectroscopy.* CRC press, 2011 (cit. on p. 41).

[150] D. A. Burns and E. W. Ciurczak, *Handbook of near-infrared analysis.* CRC press, 2007 (cit. on pp. 41, 62).

[151] A. Rygula, K. Majzner, K. M. Marzec, A. Kaczor, M. Pilarczyk, and M. Baranska, „Raman spectroscopy of proteins: a review“, *Journal of Raman Spectroscopy*, vol. 44, no. 8, pp. 1061–1076, 2013 (cit. on pp. 41, 111, 115, 167).

[152] T. Hirschfeld, J. Callis, and B. Kowalski, „Chemical sensing in process analysis“, *Science*, vol. 226, no. 4672, pp. 312–318, 1984 (cit. on p. 41).

[153] J. Claßen, F. Aupert, K. F. Reardon, D. Solle, and T. Scheper, „Spectroscopic sensors for in-line bioprocess monitoring in research and pharmaceutical industrial application“, *Anal. Bioanal.Chem.*, vol. 409, no. 3, pp. 651–666, 2017 (cit. on p. 42).

[154] S. Huffman, K. Soni, and J. Ferraiolo, „UV-Vis based determination of protein concentration: Validating and implementing slope measurements using variable pathlength technology“, *BioProcess International*, vol. 12, no. 8, pp. 66–72, 2014 (cit. on p. 42).

[155] M. Jiang, K. A. Severson, J. C. Love, H. Madden, P. Swann, L. Zang, and R. D. Braatz, „Opportunities and challenges of real-time release testing in biopharmaceutical manufacturing“, *Biotechnology and Bioengineering*, vol. 114, no. 11, pp. 2445–2456, 2017 (cit. on pp. 42, 61).

[156] L. Rolinger, M. Rüdt, J. Diehm, J. Chow-Hubbertz, M. Heitmann, S. Schleper, and J. Hubbuch, „Multi-attribute PAT for UF/DF of Proteins—Monitoring Concentration, particle sizes, and Buffer Exchange“, *Analytical and bioanalytical chemistry*, vol. 412, no. 9, pp. 2123–2136, 2020 (cit. on pp. 42, 158, 159, 161).

[157] M. Pathak, K. Lintern, V. Chopda, D. G. Bracewell, and A. S. Rathore, „Fluorescence based real time monitoring of fouling in process chromatography“, *Scientific Reports*, vol. 7, p. 45 640, 2017 (cit. on p. 42).

[158] S. Zhang, K. Xu, W. Daniels, J. Salm, J. Glynn, J. Martin, C. Gallo, R. Godavarti, and G. Carta, „Structural and functional characteristics of virgin and fouled protein A MabSelect resin cycled in a monoclonal antibody purification process“, *Biotechnology and Bioengineering*, vol. 113, no. 2, pp. 367–375, 2016 (cit. on p. 42).

[159] R. L. McCreery, *Raman spectroscopy for chemical analysis*. John Wiley & Sons, 2005, vol. 225 (cit. on p. 43).

[160] C. V. Raman and K. S. Krishnan, „The production of new radiations by light scattering. Part I“, *Proc. R. Soc. Lond.*, vol. 122, no. 789, pp. 23–35, 1929 (cit. on p. 43).

[161] S. Sasic, *Pharmaceutical applications of Raman spectroscopy*. John Wiley & Sons, 2008 (cit. on pp. 43, 189).

[162] N. R. Abu-Absi, B. M. Kenty, M. E. Cuellar, M. C. Borys, S. Sakhamuri, D. J. Strachan, M. C. Hausladen, and Z. J. Li, „Real time monitoring of multiple parameters in mammalian cell culture bioreactors using an in-line Raman spectroscopy probe“, *Biotechnology and Bioengineering*, vol. 108, no. 5, pp. 1215–1221, 2011 (cit. on pp. 43, 101).

[163] R. M. Santos, P. Kaiser, J. C. Menezes, and A. Peinado, „Improving reliability of Raman spectroscopy for mAb production by upstream processes during bioprocess development stages“, *Talanta*, vol. 199, pp. 396–406, 2019 (cit. on p. 43).

[164] R. G. Harrison, P. Todd, S. R. Rudge, and D. P. Petrides, „Bioseparations science and engineering“, 2015 (cit. on p. 43).

[165] Y. Berthois, J. A. Katzenellenbogen, and B. S. Katzenellenbogen, „Phenol red in tissue culture media is a weak estrogen: implications concerning the study of estrogen-responsive cells in culture“, *Proceedings of the National Academy of Sciences*, vol. 83, no. 8, pp. 2496–2500, 1986 (cit. on p. 43).

[166] G. Walrafen and J. Stone, „Intensification of spontaneous Raman spectra by use of liquid core optical fibers“, *Applied Spectroscopy*, vol. 26, no. 6, pp. 585–589, 1972 (cit. on p. 43).

[167] J. T. Meade, B. B. Behr, and A. R. Hajian, „A new high-resolution, high-throughput spectrometer: first experience as applied to Raman spectroscopy“, in *Next-Generation Spectroscopic Technologies V*, International Society for Optics and Photonics, vol. 8374, 2012, p. 83740V (cit. on p. 44).

[168] F. Feidl, S. Garbellini, S. Vogg, M. Sokolov, J. Souquet, H. Broly, A. Butté, and M. Morbidelli, „A new flow cell and chemometric protocol for implementing in-line Raman spectroscopy in chromatography“, *Biotechnology progress*, vol. 35, no. 5, e2847, 2019 (cit. on pp. 44, 52, 78, 79, 91–94, 100, 101).

[169] M. Rüdt, N. Brestrich, L. Rolinger, and J. Hubbuch, „Real-time monitoring and control of the load phase of a protein A capture step“, *Biotechnology and bioengineering*, vol. 114, no. 2, pp. 368–373, 2017 (cit. on pp. 44, 78, 79, 100, 101, 104, 115, 134).

[170] B. A. Patel, A. Gospodarek, M. Larkin, S. A. Kenrick, M. A. Haverick, N. Tugcu, M. A. Brower, and D. D. Richardson, „Multi-angle light scattering as a process analytical technology measuring real-time molecular weight for downstream process control“, *mAbs*, vol. 10, no. 7, pp. 945–950, 2018 (cit. on pp. 44, 58).

[171] V. Centner, O. De Noord, and D. Massart, „Detection of nonlinearity in multivariate calibration“, *Anal. Chim. Acta*, vol. 376, no. 2, pp. 153–168, 1998 (cit. on p. 45).

[172] H. Martens and T. Naes, *Multivariate calibration.* John Wiley & Sons, 1992 (cit. on pp. 45, 70, 118).

[173] S. Wold, N. Kettaneh-Wold, and B. Skagerberg, „Nonlinear PLS modeling“, *Chemometrics and Intelligent Laboratory Systems*, vol. 7, no. 1-2, pp. 53–65, 1989 (cit. on p. 45).

[174] T. Næs, K. Kvaal, T. Isaksson, and C. Miller, „Artificial neural networks in multivariate calibration“, *Journal of Near Infrared Spectroscopy*, vol. 1, no. 1, pp. 1–11, 1993 (cit. on p. 45).

[175] C. M. Andersen and R. Bro, „Variable selection in regression—a tutorial“, *J. Chemom.*, vol. 24, no. 11-12, pp. 728–737, 2010 (cit. on pp. 45, 49, 50, 52, 55, 122, 127).

[176] L. Fahrmeir, T. Kneib, and S. Lang, „Penalized structured additive regression for space-time data: a Bayesian perspective“, *Statistica Sinica*, pp. 731–761, 2004 (cit. on p. 45).

[177] A. L. Tarca, V. J. Carey, X.-w. Chen, R. Romero, and S. Drăghici, „Machine learning and its applications to biology“, *PLoS Computational Biology*, vol. 3, no. 6, e116, 2007 (cit. on p. 46).

[178] C.-F. Mandenius and N. J. Titchener-Hooker, *Measurement, monitoring, modelling and control of bioprocesses.* Springer, 2013, vol. 132 (cit. on p. 46).

[179] A. P. Ferreira, J. C. Menezes, and M. Tobyn, *Multivariate analysis in the pharmaceutical industry.* Academic Press, 2018 (cit. on pp. 46, 53).

[180] P. Gramatica, „Principles of QSAR models validation: internal and external“, *QSAR & combinatorial science*, vol. 26, no. 5, pp. 694–701, 2007 (cit. on p. 46).

[181] K. Kjeldahl and R. Bro, „Some common misunderstandings in chemometrics“, *J. Chemom.*, vol. 24, no. 7-8, pp. 558–564, 2010 (cit. on pp. 46, 49, 50, 53, 127).

[182] F. Westad and F. Marini, „Validation of chemometric models—a tutorial“, *Anal. Chim. Acta*, vol. 893, pp. 14–24, 2015 (cit. on pp. 46, 47).

[183] E. J. Wolf, K. M. Harrington, S. L. Clark, and M. W. Miller, „Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety“, *Educational and psychological measurement*, vol. 73, no. 6, pp. 913–934, 2013 (cit. on p. 46).

[184] G. A. Marcoulides and C. Saunders, „Editor's comments: PLS: a silver bullet?“, *Management Information Systems Quarterly*, pp. iii–ix, 2006 (cit. on p. 46).

[185] N. Faber and R. Rajko, „How to avoid over-fitting in multivariate calibration—The conventional validation approach and an alternative“, *Anal. Chim. Acta*, vol. 595, no. 1-2, pp. 98–106, 2007 (cit. on p. 47).

[186] R. W. Kennard and L. A. Stone, „Computer aided design of experiments“, *Technometrics*, vol. 11, no. 1, pp. 137–148, 1969 (cit. on p. 47).

[187] Å. Rinnan, L. Nørgaard, F. W. J. van der Berg, J. Thygesen, R. Bro, and S. B. Engelsen, „Data pre-processing: chapter 2“, in *Infrared Spectroscopy for Food Quality Analysis and Control*, Academic Press, 2009, pp. 29–50 (cit. on pp. 47–49).

[188] R. Gautam, S. Vanga, F. Ariese, and S. Umapathy, „Review of multidimensional data processing approaches for Raman and infrared spectroscopy“, *European Physical Journal*, vol. 2, no. 1, p. 8, 2015 (cit. on p. 48).

[189] B. G. Lipták and K. Venczel, *Analysis and Analyzers*. CRC Press, 2016, vol. 2 (cit. on p. 48).

[190] J. Engel, J. Gerretzen, E. Szymańska, J. J. Jansen, G. Downey, L. Blanchet, and L. M. Buydens, „Breaking with trends in pre-processing?“, *Trends in Analytical Chemistry*, vol. 50, pp. 96–106, 2013 (cit. on p. 48).

[191] T. Bocklitz, A. Walter, K. Hartmann, P. Rösch, and J. Popp, „How to pre-process Raman spectra for reliable and stable models?“, *Anal. Chim. Acta*, vol. 704, no. 1-2, pp. 47–56, 2011 (cit. on p. 48).

[192] Å. Rinnan, F. Van Den Berg, and S. B. Engelsen, „Review of the most common pre-processing techniques for near-infrared spectra“, *TrAC Trends in Analytical Chemistry*, vol. 28, no. 10, pp. 1201–1222, 2009 (cit. on pp. 48, 49).

[193] H. Martens, S. Jensen, and P. Geladi, „Multivariate linearity transformation for near-infrared reflectance spectrometry“, in *Proceedings of the Nordic symposium on applied statistics*, Stokkand Forlag Publishers Stavanger, Norway, 1983, pp. 205–234 (cit. on p. 48).

[194] R. Bro and A. K. Smilde, „Principal component analysis“, *Analytical Methods*, vol. 6, no. 9, pp. 2812–2831, 2014 (cit. on pp. 49, 50).

[195] F. Pukelsheim, „Robustness of statistical gossip and the Antarctic ozone hole“, *Institute of Mathematical Statistics Bulletin*, 1990 (cit. on p. 50).

[196] A. S. Hadi, A. R. Imon, and M. Werner, „Detection of outliers“, *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 1, no. 1, pp. 57–70, 2009 (cit. on p. 50).

[197] T. Mehmood, K. H. Liland, L. Snipen, and S. Sæbø, „A review of variable selection methods in partial least squares regression“, *Chemometrics and Intelligent Laboratory Systems*, vol. 118, pp. 62–69, 2012 (cit. on p. 50).

[198] R. A. van den Berg, H. C. Hoefsloot, J. A. Westerhuis, A. K. Smilde, and M. J. van der Werf, „Centering, scaling, and transformations: improving the biological information content of metabolomics data“, *BMC Genomics*, vol. 7, no. 1, p. 142, 2006 (cit. on pp. 51, 109, 121).

[199] M. Clark and R. D. Cramer III, „The probability of chance correlation using partial least squares (PLS)“, *Quantitative Structure-Activity Relationships*, vol. 12, no. 2, pp. 137–145, 1993 (cit. on p. 52).

[200] S. Wold, L. Eriksson, and S. Clementi, „Statistical validation of QSAR results“, *Chemometric methods in molecular design*, pp. 309–338, 1995 (cit. on pp. 52, 54).

[201] N. Zhao, Z.-s. Wu, Q. Zhang, X.-y. Shi, Q. Ma, and Y.-j. Qiao, „Optimization of parameter selection for partial least squares model development“, *Scientific Reports*, vol. 5, p. 11 647, 2015 (cit. on p. 52).

[202] O. Devos and L. Duponchel, „Parallel genetic algorithm co-optimization of spectral pre-processing and wavelength selection for PLS regression“, *Chemometrics and Intelligent Laboratory Systems*, vol. 107, no. 1, pp. 50–58, 2011 (cit. on p. 52).

[203] R. Leardi and A. L. Gonzalez, „Genetic algorithms applied to feature selection in PLS regression: how and when to use them“, *Chemometrics and Intelligent Laboratory Systems*, vol. 41, no. 2, pp. 195–207, 1998 (cit. on p. 52).

[204] H. Narayanan, M. Sokolov, A. Butté, and M. Morbidelli, „Decision Tree–PLS (DT-PLS) algorithm for the development of process-specific local prediction models“, *Biotechnol. Progr.*, e2818, 2019 (cit. on p. 52).

[205] S. Saerens, F. Delvaux, K. Verstrepen, P. Van Dijck, J. Thevelein, and F. Delvaux, „Parameters affecting ethyl ester production by Saccharomyces cerevisiae during fermentation“, *Applied and Environmental Microbiology*, vol. 74, no. 2, pp. 454–461, 2008 (cit. on p. 54).

[206] J. Shao, „Linear model selection by cross-validation“, *J. Am. Stat. Assoc.*, vol. 88, no. 422, pp. 486–494, 1993 (cit. on p. 54).

[207] M. Farrés, S. Platikanov, S. Tsakovski, and R. Tauler, „Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation“, *J. Chemom.*, vol. 29, no. 10, pp. 528–536, 2015 (cit. on p. 54).

[208] D. Lahat, T. Adali, and C. Jutten, „Multimodal data fusion: an overview of methods, challenges, and prospects“, *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1449–1477, 2015 (cit. on p. 55).

[209] R. C. Luo, C.-C. Yih, and K. L. Su, „Multisensor fusion and integration: approaches, applications, and future research directions“, *IEEE Sensors journal*, vol. 2, no. 2, pp. 107–119, 2002 (cit. on p. 55).

[210] M. Bevilacqua, R. Bro, F. Marini, Å. Rinnan, M. A. Rasmussen, and T. Skov, „Recent chemometrics advances for foodomics“, *TrAC Trends in Analytical Chemistry*, vol. 96, pp. 42–51, 2017 (cit. on p. 55).

[211] S. Wold, N. Kettaneh, and K. Tjessem, „Hierarchical multiblock PLS and PC models for easier model interpretation and as an alternative to variable selection“, *J. Chemom.*, vol. 10, no. 5-6, pp. 463–482, 1996 (cit. on pp. 57, 109).

[212] M. Cocchi, *Data Fusion Methodology and Applications.* Elsevier, 2019, vol. 31 (cit. on pp. 57, 108).

[213] J. Liu, M. Osadchy, L. Ashton, M. Foster, C. J. Solomon, and S. J. Gibson, „Deep convolutional neural networks for Raman spectrum recognition: a unified solution“, *Analyst*, vol. 142, no. 21, pp. 4067–4074, 2017 (cit. on p. 57).

[214] J. Acquarelli, T. van Laarhoven, J. Gerretzen, T. N. Tran, L. M. Buydens, and E. Marchiori, „Convolutional neural networks for vibrational spectroscopic data analysis“, *Anal. Chim. Acta*, vol. 954, pp. 22–31, 2017 (cit. on p. 57).

[215] S. Malek, F. Melgani, and Y. Bazi, „One-dimensional convolutional neural networks for spectroscopic signal regression“, *J. Chemom.*, vol. 32, no. 5, e2977, 2018 (cit. on p. 57).

[216] Y. LeCun, Y. Bengio, *et al.*, „Convolutional networks for images, speech, and time series“, *The handbook of brain theory and neural networks*, vol. 3361, no. 10, p. 1995, 1995 (cit. on p. 57).

[217] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, and G. Penn, „Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition“, in *2012 IEEE international conference on Acoustics, speech and signal processing (ICASSP)*, IEEE, 2012, pp. 4277–4280 (cit. on p. 57).

[218] J. Yang, J. Xu, X. Zhang, C. Wu, T. Lin, and Y. Ying, „Deep learning for vibrational spectral analysis: Recent progress and a practical guide“, *Anal. Chim. Acta*, 2019 (cit. on p. 57).

[219] Y. Bengio, „Learning deep architectures for AI“, *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009 (cit. on p. 57).

[220] F. Marini, R. Bucci, A. Magrì, and A. Magrì, „Artificial neural networks in chemometrics: History, examples and perspectives“, *Microchemical Journal*, vol. 88, no. 2, pp. 178–185, 2008 (cit. on p. 57).

[221] D. G. Bracewell, R. Francis, and C. M. Smales, „The future of host cell protein (HCP) identification during process development and manufacturing linked to a risk-based management for their control“, *Biotechnology and Bioengineering*, vol. 112, no. 9, pp. 1727–1737, 2015 (cit. on p. 58).

[222] A. L. Tscheliessnig, J. Konrath, R. Bates, and A. Jungbauer, „Host cell protein analysis in therapeutic protein bioprocessing–methods and applications“, *Biotechnology Journal*, vol. 8, no. 6, pp. 655–670, 2013 (cit. on p. 58).

[223] T. Schmidberger, C. Posch, A. Sasse, C. Gülch, and R. Huber, „Progress toward forecasting product quality and quantity of mammalian cell culture processes by performance-based modeling“, *Biotechnology progress*, vol. 31, no. 4, pp. 1119–1127, 2015 (cit. on p. 59).

[224] C. D. Agarabi, B. K. Chavez, S. C. Lute, E. K. Read, S. Rogstad, D. Awotwe-Otoo, M. R. Brown, M. T. Boyne, and K. A. Brorson, „Exploring the linkage between cell culture process parameters and downstream processing utilizing a plackett-burman design for a model monoclonal antibody“, *Biotechnology progress*, vol. 33, no. 1, pp. 163–170, 2017 (cit. on p. 59).

[225] K. Severson, J. G. VanAntwerp, V. Natarajan, C. Antoniou, J. Thömmes, and R. D. Braatz, „Elastic net with Monte Carlo sampling for data-based modeling in biopharmaceutical manufacturing facilities“, *Computers & Chemical Engineering*, vol. 80, pp. 30–36, 2015 (cit. on p. 59).

[226] J. H. Lee, „Model predictive control: Review of the three decades of development“, *International Journal of Control, Automation, and Systems*, vol. 9, no. 3, p. 415, 2011 (cit. on p. 60).

[227] M. Morari and J. H. Lee, „Model predictive control: past, present and future“, *Computers & Chemical Engineering*, vol. 23, no. 4-5, pp. 667–682, 1999 (cit. on pp. 60, 61).

[228] S. J. Qin and T. A. Badgwell, „A survey of industrial model predictive control technology“, *Control Engineering Practice*, vol. 11, no. 7, pp. 733–764, 2003 (cit. on p. 60).

[229] W. Sommeregger, B. Sissolak, K. Kandra, M. von Stosch, M. Mayer, and G. Striedner, „Quality by control: Towards model predictive control of mammalian cell culture bioprocesses“, *Biotechnology Journal*, vol. 12, no. 7, p. 1 600 546, 2017 (cit. on p. 60).

[230] S. Craven, J. Whelan, and B. Glennon, „Glucose concentration control of a fed-batch mammalian cell bioprocess using a nonlinear model predictive controller“, *Journal of Process Control*, vol. 24, no. 4, pp. 344–357, 2014 (cit. on p. 60).

[231] C. Grossmann, G. Ströhlein, M. Morari, and M. Morbidelli, „Optimizing model predictive control of the chromatographic multi-column solvent gradient purification (MCSGP) process“, *Journal of Process Control*, vol. 20, no. 5, pp. 618–629, 2010 (cit. on p. 60).

[232] M. M. Papathanasiou, A. L. Quiroga-Campano, F. Steinebach, M. Elviro, A. Mantalaris, and E. N. Pistikopoulos, „Advanced model-based control strategies for the intensification of upstream and downstream processing in mAb production“, *Biotechnol. Progr.*, vol. 33, no. 4, pp. 966–988, 2017 (cit. on p. 60).

[233] M. M. Papathanasiou, F. Steinebach, M. Morbidelli, A. Mantalaris, and E. N. Pistikopoulos, „Intelligent, model-based control towards the intensification of downstream processes“, *Computers & Chemical Engineering*, vol. 105, pp. 173–184, 2017 (cit. on p. 60).

[234] K. Singh, G. Sandhu, B. Lark, and S. Sud, „Molar extinction coefficients of some carbohydrates in aqueous solutions“, *Pramana*, vol. 58, no. 3, pp. 521–528, 2002 (cit. on p. 62).

[235] Y. Jiang, C. Li, X. Nguyen, S. Muzammil, E. Towers, J. Gabrielson, and L. Narhi, „Qualification of FTIR spectroscopic method for protein secondary structural analysis“, *Journal of Pharmaceutical Sciences*, vol. 100, no. 11, pp. 4631–4641, 2011 (cit. on p. 62).

[236] S. Y. Venyaminov and N. Kalnin, „Quantitative IR spectrophotometry of peptide compounds in water (H2O) solutions. II. Amide absorption bands of polypeptides and fibrous proteins in α-, β-, and random coil conformations“, *Biopolymers*, vol. 30, no. 13-14, pp. 1259–1271, 1990 (cit. on p. 62).

[237] M. L. Groot, L. J. van Wilderen, and M. Di Donato, „Time-resolved methods in biophysics. 5. Femtosecond time-resolved and dispersed infrared spectroscopy on proteins“, *Photochemical and Photobiological Sciences*, vol. 6, no. 5, pp. 501–507, 2007 (cit. on p. 62).

[238] G. W. Faris and R. A. Copeland, „Wavelength dependence of the Raman cross section for liquid water“, *Applied Optics*, vol. 36, no. 12, pp. 2686–2688, 1997 (cit. on p. 62).

[239] J. R. Howell, M. P. Menguc, and R. Siegel, *Thermal radiation heat transfer*. CRC press, 2015 (cit. on p. 63).

[240] A. Cox, A. J. DeWeerd, and J. Linden, „An experiment to measure Mie and Rayleigh total scattering cross sections“, *Am. J. Phys*, vol. 70, no. 6, pp. 620–625, 2002 (cit. on p. 63).

[241] R. Hahn, K. Shimahara, F. Steindl, and A. Jungbauer, „Comparison of protein A affinity sorbents III. Life time study“, *J Chromatogr A*, vol. 1102, no. 1–2, pp. 224–231, 2006 (cit. on p. 66).

[242] M. Tsukamoto, H. Watanabe, A. Ooishi, and S. Honda, „Engineered protein A ligands, derived from a histidine-scanning library, facilitate the affinity purification of IgG under mild acidic conditions“, *J Biol Eng*, vol. 8, no. 1, pp. 1–9, 2014 (cit. on p. 66).

[243] M. N. Gupta, Ed., *Methods for affinity-based separations of enzymes and proteins*, 1st ed., ser. Methods and Tools in Biosciences and Medicine. Birkhäuser, 2002 (cit. on p. 66).

[244] C. Jiang, J. Liu, M. Rubacha, and A. A. Shukla, „A mechanistic study of Protein A chromatography resin lifetime“, *J Chromatogr A*, vol. 1216, no. 31, pp. 5849–5855, 2009 (cit. on p. 66).

[245] M. Angarita, T. Müller-Späth, D. Baur, R. Lievrouw, G. Lissens, and M. Morbidelli, „Twin-column capture SMB: A novel cyclic process for protein A affinity chromatography“, *J Chromatogr A*, vol. 1389, pp. 85–95, 2015 (cit. on p. 66).

[246] V. Warikoo, R. Godawat, K. Brower, S. Jain, D. Cummings, E. Simons, T. Johnson, J. Walther, M. Yu, B. Wright, J. McLarty, K. P. Karey, C. Hwang, W. Zhou, F. Riske, and K. Konstantinov, „Integrated continuous production of recombinant therapeutic proteins“, *Biotechnol Bioeng*, vol. 109, no. 12, pp. 3018–3029, 2012 (cit. on p. 66).

[247] F. Capito, A. Zimmer, and R. Skudas, „Mid-infrared spectroscopy-based analysis of mammalian cell culture Parameters“, *Biotechnology Progress*, vol. 31, no. 2, pp. 578–584, 2015 (cit. on p. 67).

[248] R. F. Steinhoff, D. J. Karst, F. Steinebach, M. R. Kopp, G. W. Schmidt, A. Stettler, J. Krismer, M. Soos, M. Pabst, A. Hierlemann, M. Morbidelli, and R. Zenobi, „Microarray-based MALDI-TOF mass spectrometry enables monitoring of monoclonal antibody production in batch and perfusion cell cultures“, *Methods*, in press, 2015 (cit. on p. 67).

[249] H. A. Chase, „Rapid chromatographic monitoring of bioprocesses“, *Biosensors*, vol. 2, no. 5, pp. 269–286, 1986 (cit. on p. 67).

[250] S. S. Ozturk, J. C. Thrift, J. D. Blackie, and D. Naveh, „Real-time monitoring of protein secretion in mammalian cell fermentation: Measurement of monoclonal antibodies using a computer-controlled HPLC system (BioCad/RPM)“, *Biotechnol Bioeng*, vol. 48, no. 3, pp. 201–206, 1995 (cit. on p. 67).

[251] S. Paliwal, T. Nadler, D. Wang, and F. Regnier, „Automated process monitoring of monoclonal antibody production“, *Anal Chem*, vol. 65, no. 23, pp. 3363–3367, 1993 (cit. on p. 67).

[252] P. Bängtsson, E. Estrada, K. Lacki, and H. Skoglar, *A method in a chromatography system*, EP Patent App. EP20,100,792,416, 2012 (cit. on p. 67).

[253] A. Höskuldsson, „PLS regression methods“, *J Chemom*, vol. 2, pp. 211–228, 1988 (cit. on p. 70).

[254] N. Aboulaich, W. K. Chung, J. H. Thompson, C. Larkin, D. Robbins, and M. Zhu, „A novel approach to monitor clearance of host cell proteins associated with monoclonal antibodies“, *Biotechnology progress*, vol. 30, no. 5, pp. 1114–1124, 2014 (cit. on pp. 73, 85, 87).

[255] R. T. Thakor, N. Anaya, Y. Zhang, C. Vilanilam, K. W. Siah, C. H. Wong, and A. W. Lo, „Just how good an investment is the biopharmaceutical sector?“, *Nature biotechnology*, vol. 35, no. 12, p. 1149, 2017 (cit. on p. 78).

[256] J. Rantanen and J. Khinast, „The future of pharmaceutical manufacturing sciences“, *Journal of pharmaceutical sciences*, vol. 104, no. 11, pp. 3612–3638, 2015 (cit. on p. 78).

[257] F. Feidl, S. Garbellini, M. F. Luna, S. Vogg, J. Souquet, H. Broly, M. Morbidelli, and A. Butté, „Combining Mechanistic Modeling and Raman Spectroscopy for Monitoring Antibody Chromatographic Purification“, *Processes*, vol. 7, no. 10, p. 683, 2019 (cit. on pp. 78, 79, 92–94, 100, 101).

[258] G. Thakur, V. Hebbi, and A. S. Rathore, „An NIR-based PAT approach for real-time control of loading in Protein A chromatography in continuous manufacturing of monoclonal antibodies“, *Biotechnology and Bioengineering*, 2019 (cit. on pp. 78, 91, 94, 100).

[259] C. H. Goey, „Cascading effects in bioprocessing: the impact of cell culture environment on CHO cell behaviour and host cell protein species“, 2016 (cit. on p. 79).

[260] B. Nogal, K. Chhiba, and J. C. Emery, „Select host cell proteins coelute with monoclonal antibodies in protein a chromatography“, *Biotechnology progress*, vol. 28, no. 2, pp. 454–458, 2012 (cit. on p. 85).

[261] V. N. Sisodiya, J. Lequieu, M. Rodriguez, P. McDonald, and K. P. Lazzareschi, „Studying host cell protein interactions with monoclonal antibodies using high throughput protein A chromatography“, *Biotechnology Journal*, vol. 7, no. 10, pp. 1233–1241, 2012 (cit. on p. 85).

[262] J. Van de Velde, M. J. Saller, K. Eyer, and A. Voloshin, „Chromatographic clarification overcomes chromatin-mediated hitch-hiking interactions on Protein A capture column“, *Biotechnology and Bioengineering*, 2020 (cit. on p. 85).

[263] U. Langel, B. F. Cravatt, A. Graslund, N. Von Heijne, M. Zorko, T. Land, and S. Niessen, *Introduction to peptides and proteins*. CRC press, 2009 (cit. on p. 88).

[264] 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“, *Analytical and bioanalytical chemistry*, pp. 1–18, 2020 (cit. on pp. 92, 101, 102, 111, 157, 158, 169).

[265] F. Allegrini and A. C. Olivieri, „IUPAC-consistent approach to the limit of detection in partial least-squares calibration“, *Analytical chemistry*, vol. 86, no. 15, pp. 7858–7866, 2014 (cit. on p. 96).

[266] L. Rolinger, M. Rüdt, and J. Hubbuch, „A multisensor approach for improved protein A load phase monitoring by conductivity-based background subtraction of UV spectra“, *Biotechnology and Bioengineering*, 2020 (cit. on pp. 100, 101, 108, 110, 115, 159, 166).

[267] B. Li, P. W. Ryan, B. H. Ray, K. J. Leister, N. M. Sirimuthu, and A. G. Ryder, „Rapid characterization and quality control of complex cell culture media solutions using Raman spectroscopy and chemometrics“, *Biotechnology and bioengineering*, vol. 107, no. 2, pp. 290–301, 2010 (cit. on p. 101).

[268] B. Li, B. H. Ray, K. J. Leister, and A. G. Ryder, „Performance monitoring of a mammalian cell based bioprocess using Raman spectroscopy“, *Analytica chimica acta*, vol. 796, pp. 84–91, 2013 (cit. on p. 101).

[269] K. Buckley and A. G. Ryder, „Applications of Raman spectroscopy in biopharmaceutical manufacturing: a short review“, *Applied spectroscopy*, vol. 71, no. 6, pp. 1085–1116, 2017 (cit. on p. 101).

[270] S. Zobel-Roos, M. Mouellef, C. Siemers, and J. Strube, „Process Analytical Approach towards Quality Controlled Process Automation for the Downstream of Protein Mixtures by Inline Concentration Measurements Based on Ultraviolet/Visible Light (UV/VIS) Spectral Analysis“, *Antibodies*, vol. 6, no. 4, p. 24, 2017 (cit. on p. 101).

[271] M. Saggu, J. Liu, and A. Patel, „Identification of subvisible particles in biopharmaceutical formulations using Raman spectroscopy provides insight into polysorbate 20 degradation pathway“, *Pharmaceutical research*, vol. 32, no. 9, pp. 2877–2888, 2015 (cit. on p. 101).

[272] A. Biancolillo, K. H. Liland, I. Måge, T. Næs, and R. Bro, „Variable selection in multi-block regression“, *Chemometrics and Intelligent Laboratory Systems*, vol. 156, pp. 89–101, 2016 (cit. on p. 102).

[273] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del R'ıo, M. Wiebe, P. Peterson, P. G'erard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, „Array programming with NumPy“, *Nature*, vol. 585, no. 7825, pp. 357–362, 09/2020 (cit. on p. 109).

[274] W. McKinney, „Data Structures for Statistical Computing in Python“, in *Proceedings of the 9th Python in Science Conference*, S. van der Walt and J. Millman, Eds., 2010, pp. 56–61 (cit. on p. 109).

[275] Martı́n Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Y. Jia,

Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng, *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*, Software available from tensorflow.org, 2015 (cit. on p. 109).

[276] T. O'Malley, E. Bursztein, J. Long, F. Chollet, H. Jin, L. Invernizzi, *et al.*, *Keras Tuner*, `https://github.com/keras-team/keras-tuner`, 2019 (cit. on p. 109).

[277] P. Ramachandran, B. Zoph, and Q. V. Le, „Searching for activation functions", *arXiv preprint arXiv:1710.05941*, 2017 (cit. on p. 110).

[278] D. P. Kingma and J. Ba, *Adam: A Method for Stochastic Optimization*, 2017 (cit. on p. 110).

[279] L. Silveira, C. A. Pasqualucci, B. Bodanese, M. T. T. Pacheco, and R. A. Zângaro, „Normal-subtracted preprocessing of Raman spectra aiming to discriminate skin actinic keratosis and neoplasias from benign lesions and normal skin tissues", *Lasers in Medical Science*, vol. 35, pp. 1141–1151, 2019 (cit. on pp. 111, 167).

[280] S. Goldrick, A. Umprecht, A. Tang, R. Zakrzewski, M. Cheeks, R. Turner, A. Charles, K. Les, M. Hulley, C. Spencer, *et al.*, „High-Throughput Raman Spectroscopy Combined with Innovate Data Analysis Workflow to Enhance Biopharmaceutical Process Development", *Processes*, vol. 8, no. 9, p. 1179, 2020 (cit. on p. 115).

[281] J. Whelan, S. Craven, and B. Glennon, „In situ Raman spectroscopy for simultaneous monitoring of multiple process parameters in mammalian cell culture bioreactors", *Biotechnology progress*, vol. 28, no. 5, pp. 1355–1362, 2012 (cit. on p. 115).

[282] D. R. Parachalil, B. Brankin, J. McIntyre, and H. J. Byrne, „Raman spectroscopic analysis of high molecular weight proteins in solution–considerations for sample analysis and data pre-processing", *Analyst*, vol. 143, no. 24, pp. 5987–5998, 2018 (cit. on p. 115).

[283] T. E. Matthews, J. P. Smelko, B. Berry, S. Romero-Torres, D. Hill, R. Kshirsagar, and K. Wiltberger, „Glucose monitoring and adaptive feeding of mammalian cell culture in the presence of strong aut-

ofluorescence by near infrared Raman spectroscopy“, *Biotechnology progress*, vol. 34, no. 6, pp. 1574–1580, 2018 (cit. on p. 118).

[284] L. Eriksson, E. Johansson, N. Kettaneh-Wold, J. Trygg, C. Wikström, and S. Wold, *Multi-and megavariate data analysis: part II: advanced applications and method extensions*. Umetrics Inc, 2006 (cit. on p. 121).

[285] D. L. Hall and A. Steinberg, „Dirty secrets in multisensor data fusion“, Pennsylvania State Univ University Park Applied Research Lab, Tech. Rep., 2001 (cit. on p. 122).

[286] E. Felföldi, T. Scharl, M. Melcher, A. Dürauer, K. Wright, and A. Jungbauer, „Osmolality is a predictor for model-based real time monitoring of concentration in protein chromatography“, *Journal of Chemical Technology & Biotechnology*, vol. 95, no. 4, pp. 1146–1152, 2020 (cit. on p. 122).

[287] A. S. Rathore, A. Sharma, and D. Chilin, „Applying process analytical technology to biotech unit operations“, *Biopharm International*, vol. 19, no. 8, 2006 (cit. on p. 135).

[288] F. G. Donnan, „Theorie der Membrangleichgewichte und Membranpotentiale bei Vorhandensein von nicht dialysierenden Elektrolyten. Ein Beitrag zur physikalisch-chemischen Physiologie.“, *Zeitschrift Für Elektrochemie Und Angewandte Physikalische Chemie*, vol. 17, no. 14, pp. 572–581, 1911 (cit. on pp. 135, 157).

[289] A. Arunkumar, J. Zhang, N. Singh, S. Ghose, and Z. J. Li, „Ultrafiltration behavior of partially retained proteins and completely retained proteins using equally-staged single pass tangential flow filtration membranes“, *Biotechnology Progress*, vol. 34, no. 5, pp. 1137–1148, 2018 (cit. on p. 135).

[290] A. Savitzky and M. J. Golay, „Smoothing and differentiation of data by simplified least squares procedures.“, *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964 (cit. on p. 139).

[291] H. Mach and C. R. Middaugh, „Simultaneous monitoring of the environment of tryptophan, tyrosine, and phenylalanine residues in proteins by near-ultraviolet second-derivative spectroscopy.“, *Analytical Biochemistry*, vol. 222, pp. 323–331, 1994 (cit. on pp. 139, 152).

[292] R. Ragone, G. Colonna, C. Balestrieri, L. Servillo, and G. Irace, „Determination of tyrosine exposure in proteins by second-derivative spectroscopy“, *Biochemistry*, vol. 23, no. 8, pp. 1871–1875, 1984 (cit. on p. 139).

[293] A. Einstein, „Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen“, *Annalen Der Physik*, vol. 17, no. 4, pp. 549–560, 1905 (cit. on pp. 140, 162).

[294] P. D. Godfrin, S. D. Hudson, K. Hong, L. Porcar, P. Falus, N. J. Wagner, and Y. Liu, „Short-time glassy dynamics in viscous protein solutions with competing interactions“, *Physical Review Letters*, vol. 115, no. 22, p. 228 302, 2015 (cit. on pp. 140, 162).

[295] P. D. Godfrin, I. E. Zarraga, J. Zarzar, L. Porcar, P. Falus, N. J. Wagner, and Y. Liu, „Effect of Hierarchical Cluster Formation on the Viscosity of Concentrated Monoclonal Antibody Formulations Studied by Neutron Scattering“, *Journal of Physical Chemistry B*, vol. 120, no. 2, pp. 278–291, 2016 (cit. on pp. 140, 162).

[296] J. Lebowitz, M. S. Lewis, and P. Schuck, „Modern analytical ultracentrifugation in protein science: A tutorial review“, *Protein Science*, vol. 11, no. 9, pp. 2067–2079, 2002 (cit. on pp. 140, 162).

[297] E. Lemmon, M. McLinden, and D. Friend, „NIST chemistry webbook, NIST standard reference database number 69“, in, P. Linstrom and W. Mallard, Eds. Gaithersburg MD-US: National Institute of Standards and Technology, 2005, ch. Thermophysical Properties of Fluid Systems (cit. on pp. 140, 162).

[298] J. C. Thomas, „The determination of log normal particle size distributions by dynamic light scattering“, *J. Colloid Interface Sci.*, vol. 117, no. 1, pp. 187–192, 1987 (cit. on p. 141).

[299] Z. Sun, T. Deluca, and K. Mattison, „The size and rheology characterization of concentrated emulsions“, *American Laboratory*, vol. 37, no. 12, p. 8, 2005 (cit. on p. 141).

[300] Malvern, „Application of Dynamic Light Scattering (DLS) to Protein Therapeutic Formulations: Principles, Measurements and Analysis—2. Concentration Effects and Particle Interactions“, Malvern Instruments Limited, Grovewood Road, Malvern, Worcestershire, UK, Tech. Rep., 2017 (cit. on p. 141).

[301] S. B. Dubin, J. H. Lunacek, and G. B. Benedek, „Observation of the spectrum of light scattered by solutions of biological macromolecules“, *Proceedings of the National Academy of Sciences*, vol. 57, no. 5, pp. 1164–1171, 1967 (cit. on p. 141).

[302] E. S. Forzani, M. Otero, M. A. Pérez, M. L. Teijelo, and E. J. Calvo, „The structure of layer-by-layer self-assembled glucose oxidase and Os (Bpy) 2ClPyCH2NH- poly (allylamine) multilayers: Ellipsometric and quartz crystal microbalance studies“, *Langmuir*, vol. 18, no. 10, pp. 4020–4029, 2002 (cit. on p. 141).

[303] G. Zoldák, A. Zubrik, A. Musatov, M. Stupák, and E. Sedlák, „Irreversible thermal denaturation of glucose oxidase from Aspergillus niger is the transition to the denatured state with residual structure“, *Journal of Biological Chemistry*, vol. 279, no. 46, pp. 47 601–47 609, 2004 (cit. on pp. 143, 147).

[304] S. Mitragotri, P. A. Burke, and R. Langer, „Overcoming the challenges in administering biopharmaceuticals: formulation and delivery strategies“, *Nature Reviews Drug Discovery*, vol. 13, no. 9, p. 655, 2014 (cit. on p. 145).

[305] S. Yadav, S. J. Shire, and D. S. Kalonia, „Factors affecting the viscosity in high concentration solutions of different monoclonal antibodies“, *Journal of Pharmaceutical Sciences*, vol. 99, no. 12, pp. 4812–4829, 2010 (cit. on p. 145).

[306] E. Lewis, W. Qi, L. Kidder, S. Amin, S. Kenyon, and S. Blake, „Combined Dynamic Light Scattering and Raman Spectroscopy Approach for Characterizing the Aggregation of Therapeutic Proteins“, *Molecules*, vol. 19, no. 12, pp. 20 888–20 905, 12/2014 (cit. on p. 146).

[307] H.-C. Mahler and W. Jiskoot, *Analysis of aggregates and particles in protein pharmaceuticals.* John Wiley & Sons, 2011 (cit. on pp. 146, 147).

[308] J. J. O’Malley and J. L. Weaver, „Subunit structure of glucose oxidase from Aspergillus niger“, *Biochemistry*, vol. 11, no. 19, pp. 3527–3532, 1972 (cit. on p. 147).

[309] B. E. Swoboda and V. Massey, „On the reaction of the glucose oxidase from Aspergillus niger with bisulfite“, *Journal of Biological Chemistry*, vol. 241, no. 14, pp. 3409–3416, 1966 (cit. on p. 147).

[310] M. Gouda, M. Thakur, and N. Karanth, „Reversible denaturation behavior of immobilized glucose oxidase“, *Applied Biochemistry and Biotechnology*, vol. 102, no. 1-6, pp. 471–480, 2002 (cit. on p. 147).

[311] J. Liu, J. Lu, X. Zhao, J. Lu, and Z. Cui, „Separation of glucose oxidase and catalase using ultrafiltration with 300-kDa polyethersulfone membranes“, *Journal of Membrane Science*, vol. 299, no. 1-2, pp. 222–228, 2007 (cit. on p. 147).

[312] D. K. Wilkins, S. B. Grimshaw, V. Receveur, C. M. Dobson, J. A. Jones, and L. J. Smith, „Hydrodynamic radii of native and denatured proteins measured by pulse field gradient NMR techniques“, *Biochemistry*, vol. 38, no. 50, pp. 16 424–16 431, 1999 (cit. on p. 147).

[313] J. Liu, M. D. Nguyen, J. D. Andya, and S. J. Shire, „Reversible self-association increases the viscosity of a concentrated monoclonal antibody in aqueous solution“, *Journal of Pharmaceutical Sciences*, vol. 94, no. 9, pp. 1928–1940, 2005 (cit. on p. 147).

[314] J. S. Philo, „A critical review of methods for size characterization of non-particulate protein aggregates“, *Current Pharmaceutical Biotechnology*, vol. 10, no. 4, pp. 359–372, 2009 (cit. on p. 152).

[315] K. Ahrer, A. Buchacher, G. Iberer, D. Josic, and A. Jungbauer, „Analysis of aggregates of human immunoglobulin G using size-exclusion chromatography, static and dynamic light scattering“, *Journal of Chromatography A*, vol. 1009, no. 1-2, pp. 89–96, 2003 (cit. on p. 152).

[316] J. S. Philo, „Is any measurement method optimal for all aggregate sizes and types?“, *The Aaps Journal*, vol. 8, no. 3, E564–E571, 2006 (cit. on p. 152).

[317] N. B. Bam, J. L. Cleland, J. Yang, M. C. Manning, J. F. Carpenter, R. F. Kelley, and T. W. Randolph, „Tween protects recombinant human growth hormone against agitation-induced damage via hydrophobic interactions“, *Journal of Pharmaceutical Sciences*, vol. 87, no. 12, pp. 1554–1559, 1998 (cit. on p. 152).

[318] B. Rasmussen, *Innovation and commercialisation in the biopharmaceutical industry: Creating and capturing value.* Edward Elgar Publishing, 2010 (cit. on p. 156).

[319] D. Goldstein and J. Thomas, „Biopharmaceuticals derived from genetically modified plants“, *QJM: An International Journal of Medicine*, vol. 97, no. 11, pp. 705–716, 2004 (cit. on p. 156).

[320] E. Goldberg, *Handbook of downstream processing.* Springer Science & Business Media, 2012 (cit. on p. 156).

[321] S. J. Shire, „Formulation and manufacturability of biologics“, *Current opinion in biotechnology*, vol. 20, no. 6, pp. 708–714, 2009 (cit. on p. 157).

[322] M. Holstein, J. Hung, H. Feroz, S. Ranjan, C. Du, S. Ghose, and Z. J. Li, „Strategies for high-concentration drug substance manufacturing to facilitate subcutaneous administration: A review“, *Biotechnology and Bioengineering*, vol. 117, no. 11, pp. 3591–3606, 2020 (cit. on p. 157).

[323] R. G. Harrison, P. Todd, S. R. Rudge, and D. P. Petrides, *Bioseparations science and engineering.* Oxford University Press, USA, 2015 (cit. on p. 157).

[324] J. M. West, H. Feroz, X. Xu, N. Puri, M. Holstein, S. Ghose, J. Ding, and Z. Li, „Process analytical technology for on-line monitoring of quality attributes during single-use ultrafiltration/diafiltration“, *Biotechnology and Bioengineering*, vol. 118, no. 6, pp. 2293–2300, 2021 (cit. on p. 158).

[325] 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“, *International Journal of Pharmaceutics*, vol. 600, p. 120 456, 2021 (cit. on p. 158).

[326] G. Thakur, S. Thori, and A. S. Rathore, „Implementing PAT for single-pass tangential flow ultrafiltration for continuous manufacturing of monoclonal antibodies“, *Journal of Membrane Science*, vol. 613, p. 118 492, 2020 (cit. on p. 158).

[327] P. Renati, Z. Kovacs, A. De Ninno, and R. Tsenkova, „Temperature dependence analysis of the NIR spectra of liquid water confirms the existence of two phases, one of which is in a coherent state“, *Journal of Molecular Liquids*, vol. 292, p. 111 449, 2019 (cit. on p. 158).

[328] F. Wülfert, W. T. Kok, O. E. de Noord, and A. K. Smilde, „Linear techniques to correct for temperature-induced spectral variation in multivariate calibration“, *Chemometrics and intelligent laboratory systems*, vol. 51, no. 2, pp. 189–200, 2000 (cit. on p. 158).

[329] D. P. Wasalathanthri, H. Feroz, N. Puri, J. Hung, G. Lane, M. Holstein, L. Chemmalil, D. Both, S. Ghose, J. Ding, *et al.*, „Real-time monitoring of quality attributes by in-line Fourier transform infrared spectroscopic sensors at ultrafiltration and diafiltration of bioprocess“, *Biotechnology and Bioengineering*, vol. 117, no. 12, pp. 3766–3774, 2020 (cit. on p. 158).

[330] G. S. Adair and M. Adair, „The density increments of proteins“, *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, vol. 190, no. 1022, pp. 341–356, 1947 (cit. on p. 162).

[331] P. Charlwood, „Partial Specific Volumes of Proteins in Relation to Composition and Environment1“, *Journal of the American Chemical Society*, vol. 79, no. 4, pp. 776–781, 1957 (cit. on p. 162).

[332] P. C. Young, *Recursive estimation and time-series analysis: an introduction*. Springer Science & Business Media, 2012 (cit. on pp. 163, 164).

[333] E. A. Wan and R. Van Der Merwe, „The unscented Kalman filter for nonlinear estimation“, in *Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium (Cat. No. 00EX373)*, Ieee, 2000, pp. 153–158 (cit. on p. 163).

[334] F. Gustafsson and G. Hendeby, „Some relations between extended and unscented Kalman filters“, *IEEE Transactions on Signal Processing*, vol. 60, no. 2, pp. 545–555, 2011 (cit. on p. 163).

[335] L. Rolinger, M. Rüdt, and J. Hubbuch, „Comparison of UV- and Raman-based monitoring of the protein A load phase and evaluation of data fusion by PLS models and CNNs“, *Biotechnology and Bioengineering*, accepted (cit. on pp. 167, 169).

[336] B. Wei, N. Woon, L. Dai, R. Fish, M. Tai, W. Handagama, A. Yin, J. Sun, A. Maier, D. McDaniel, *et al.*, „Multi-attribute Raman spectroscopy (MARS) for monitoring product quality attributes in formulated monoclonal antibody therapeutics“, in *Mabs*, Taylor & Francis, vol. 14, 2022, p. 2 007 564 (cit. on p. 171).

[337] G. Socrates, *Infrared and Raman characteristic group frequencies: tables and charts*. John Wiley & Sons, 2004 (cit. on pp. 171, 172).

[338] O. G. Pandoli, R. J. Neto, N. R. Oliveira, A. C. Fingolo, C. C. Corrêa, K. Ghavami, M. Strauss, and M. Santhiago, „Ultra-highly conductive hollow channels guided by a bamboo bio-template for electric and electrochemical devices“, *Journal of Materials Chemistry A*, vol. 8, no. 7, pp. 4030–4039, 2020 (cit. on p. 171).

[339] A. Zwick, F. Lakhdar-Ghazal, and J.-F. Tocanne, „Characterization of the ionization of phosphoric acid using Raman spectroscopy“, *Journal of the Chemical Society, Faraday Transactions 2: Molecular and Chemical Physics*, vol. 85, no. 7, pp. 783–788, 1989 (cit. on p. 171).

[340] K. A. Syed, S.-F. Pang, Y. Zhang, and Y.-H. Zhang, „Micro-Raman observation on the H2PO4- association structures in a supersaturated droplet of potassium dihydrogen phosphate (KH2PO4)“, *The Journal of chemical physics*, vol. 138, no. 2, p. 024 901, 2013 (cit. on p. 171).

[341] L. P. Heighton, M. Zimmerman, C. P. Rice, E. E. Codling, J. A. Tossell, and W. F. Schmidt, „Quantification of inositol hexa-kis phosphate in environmental samples“, *journal of soil science*, 2012 (cit. on p. 171).

[342] B. Sjöberg, S. Foley, B. Cardey, and M. Enescu, „An experimental and theoretical study of the amino acid side chain Raman bands in proteins“, *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 128, pp. 300–311, 2014 (cit. on p. 171).

[343] P. Freire, F. M. Barboza, J. A. Lima, F. Melo, and J. Mendes Filho, „Raman spectroscopy of amino acid crystals“, *Raman Spectroscopy and Applications*, vol. 201, 2017 (cit. on p. 171).

[344] N. N. Brandt, A. Y. Chikishev, and I. K. Sakodinskaya, „Raman spectroscopy of tris-(hydroxymethyl) aminomethane as a model system for the studies of α-chymotrypsin activation by crown ether in organic solvents“, *Journal of Molecular Structure*, vol. 648, no. 3, pp. 177–182, 2003 (cit. on p. 172).

[345] Y. Baek, N. Singh, A. Arunkumar, A. Borwankar, and A. L. Zydney, „Mass balance model with donnan equilibrium accurately describes unusual pH and excipient profiles during diafiltration of monoclonal antibodies“, *Biotechnology journal*, vol. 14, no. 7, p. 1 800 517, 2019 (cit. on p. 172).

[346] F. Miao, A. Velayudhan, E. DiBella, J. Shervin, M. Felo, M. Teeters, and P. Alred, „Theoretical analysis of excipient concentrations during the final ultrafiltration/diafiltration step of therapeutic antibody“, *Biotechnology progress*, vol. 25, no. 4, pp. 964–972, 2009 (cit. on p. 172).

[347] M. R. Stoner, N. Fischer, L. Nixon, S. Buckel, M. Benke, F. Austin, T. W. Randolph, and B. S. Kendrick, „Protein- solute interactions affect the outcome of ultrafiltration/diafiltration operations", *Journal of pharmaceutical sciences*, vol. 93, no. 9, pp. 2332–2342, 2004 (cit. on p. 172).

[348] C. Doneanu, A. Xenopoulos, K. Fadgen, J. Murphy, S. J. Skilton, H. Prentice, M. Stapels, and W. Chen, „Analysis of host-cell proteins in biotherapeutic proteins by comprehensive online two-dimensional liquid chromatography/mass spectrometry", *mAbs*, vol. 4, no. 1, pp. 24–44, 2012, PMID: 22327428 (cit. on p. 190).

[349] Q. Luan, S. Cahoon, A. Wu, S. S. Bale, M. Yarmush, and A. Bhushan, „A microfluidic in-line ELISA for measuring secreted protein under perfusion", *Biomedical microdevices*, vol. 19, no. 4, p. 101, 2017 (cit. on p. 190).

Abbreviations

r_{ab} a/b-ratio 139

Adam Adaptive Moment Estimation 110

ALS Alternating Least-Squares 48

ANN artificial neural network 21, 25–27, 30, 32, 45, 57, 122

ATR Attenuated Total Reflection 42

bsAb bispecific antibody 6, 161, 163, 166, 167, 169, 170, 172, 175

CA Cellulose Acetate 137

CEX Cation-Exchange 44

CFF Cross-Flow Filtration v, 10, 133–137, 142, 145, 156–159, 164, 186

CHO Chinese Hamster Ovary 6, 67

CNN convolutional neural network iv, vi, x, xii, 32, 57, 99, 100, 102, 105, 106, 110, 122–125, 129, 183, 185–188

CPP Critical Process Parameter 12, 29, 30

CQA Critical Quality Attribute ii, 11–14, 29, 30, 59–61, 184

CSTR Continuously Stirred Tank Reactor 163

CV Cross-Validation 45, 46, 50–54

CV Column Volume 82, 104, 105

DAD Diode Array Detector 69, 70, 81, 82, 85–87, 89, 91, 93, 104

DF Diafiltration 134, 135, 137–139, 141, 142, 144, 145, 147, 148, 150–154, 156–164, 171, 172, 175–177, 186, 187

DLS Dynamic Light Scattering v, 19, 20, 134, 135, 140, 141, 147, 151, 152, 186

DNA Deoxyribonucleic Acid 6, 45, 54, 58, 59, 61, 79, 84, 85, 88, 94–96, 115, 190

DoE Design of Experiment 46

DSP Downstream Processing 15, 134, 156

DV Diafiltration Volume 137–139, 145, 161, 172

EKF Extended Kalman Filter vi, xii, xiii, 32, 156, 158, 163, 164, 166, 172, 174, 176, 177, 187

ELISA enzyme-linked immunosorbent assay 190

EML List of Essential Medicine i, 1

EMSC Extended Multiplicative Signal Correction 48, 49, 108

Fab Fragment antigen binding 6

FAD Flavin-adenine Dinucleotides 147

Fc Fragment crystallized 6, 8, 9

FDA Food and Drug Administration 11

FTIR Fourier-Transform Infrared 4, 16, 17, 41, 48, 158

GA Genetic Algorithm 52, 179

GMP Good Manufacturing Practice 189, 190

GOx Glucose Oxidase 134, 137–139, 141–144, 147, 148, 150, 152–154, 187

GUI Graphical User Interface 139, 161

HCCF Harvested Cell Culture Fluid iii, 7, 10, 66, 68, 76, 79–81, 84, 85, 95, 102, 103, 125, 184

HCP Host Cell Protein 7, 16, 29, 45, 54, 58, 59, 61, 73, 78, 79, 84, 85, 90, 94–96, 115, 190

HIC Hydrophobic-Interaction Chromatography 44

HMW High Moleculare Weight Variant 45, 54

HPLC High Performance Liquid Chromatography 2, 34, 81, 105, 142

Ig Immunoglobulin 3

IgG Immunoglobulin G 39

IR Infrared 13, 15–18, 155, 158

IUPAC International Union of Pure and Applied Chemistry 40

LC/MS Liquid Chromatography coupled to Mass Spectrometry 190

LOD limit of detection 91, 96, 97

LOQ limit of quantitation 91, 96, 97

LV Latent Variable iv, 24, 70, 71, 73, 84, 87, 88, 90, 117–121, 124, 127, 128, 177, 179, 186

mAb monoclonal antibody ii–v, ix–xi, 6, 7, 9, 10, 16, 17, 31, 42–45, 54, 55, 58, 62, 65–76, 79–95, 97, 99–105, 108, 111–118, 121–127, 129–131, 134, 138, 139, 141–143, 145, 147, 148, 150, 154, 157, 159, 161, 163, 166, 167, 169–172, 175, 177–179, 184–187

mAbs monoclonal antibodies 2, 3, 6–8

MCSGP Multi-Column Solvent Gradient Purification 60

MCSGP Research and Development 78

microLDS micro Liquid Density Sensor v, xi, 133, 135, 137, 140, 144, 145, 159, 160, 162, 186

MIR Major Immunodominant Region 36, 38, 39, 41, 48, 62

MLR multiple linear regression 23, 24

MPC Model Predictive Control 34, 60, 61, 190

mPES modified Polyethersulfone 137

MSC Multiplicative Signal Correction 48, 49

MSE Mean Square Error 110

MVDA Multivariate Data Analysis 15, 21, 46

NIPALS Nonlinear Iterative Partial Least Squares 24, 84, 108

NIR Near-Infrared 36, 38, 39, 41, 48, 62, 79, 91–94, 100, 158

NIST National Institute of Standards and Technology 140, 162

O-PLS Orthogonal PLS 49

OPC Open Platform Communications 189

OSC Orthogonal Signal Correction 49

P&ID Piping and Instrumentation Diagram 136, 159, 170

PAT Process Analytical Technology i, ii, vi–viii, xii, xiii, 2, 8, 10–12, 29, 33–36, 42, 58–60, 67, 77, 78, 99–101, 125, 133–135, 157, 183, 187–190

PC Principal Component 21, 22, 179

PCA Principal Component Analysis 21–24, 49, 51, 166, 172

PCCC Periodic Counter Current Chromatography 10

PCR Principal Component Regression 23, 24

PEEK Polyether Ether Ketone 137

PEG Polyethylene Glycole 41

PES Polyethersulfone 137, 138

PG *Panzergewinde* 104

PLS Partial-Least Squares ii–iv, vi, viii–x, xiii, 21, 23–25, 27, 29–32, 41, 44–47, 49–53, 57, 65, 66, 68–76, 78, 82, 84, 85, 87–92, 94, 95, 97, 99, 100, 102, 104–106, 108–110, 114–119, 121–125, 127–129, 163, 167, 169, 171, 176, 179, 183–186, 188

PRESS Predicted Residual Error Sum of Squares 53

QbD Quality by Design 60

qPCR quantitative polymerase chain reaction 190

UV Ultraviolet ii–v, viii–xii, 13–15, 31, 32, 36, 38–42, 44, 48, 62, 67–69, 77, 78, 82, 84, 85, 88, 92–94, 99–102, 105, 108–110, 114–119, 121–125, 127, 129, 155, 156, 158, 161, 162, 166, 167, 169, 170, 176, 183–187, 190

UV/Vis Ultraviolet/Visible v, xi, 4, 16, 17, 65, 67, 68, 72, 75, 78, 79, 133, 135, 136, 139, 144, 152, 153, 159, 160, 169, 170, 176, 186, 190

UVRR UV Resonance Raman 40, 190

VLP Virus-Like Particle 44

VP Variable Pathlength v, xi, 32, 42, 133, 135, 136, 144, 152, 156, 159, 160, 167, 169, 170, 176, 183, 186, 190

WHO World Health Organization i, 1