Advances in data-driven analytics for protein crystallization across different scales

---

Process analytical technology for protein crystallization

Zur Erlangung des akademischen Grades einer
Doktorin der Ingenieurwissenschaften (Dr.-Ing.)

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

Dissertation

von
Christina Henriette Mei Ling Wegner, M. Sc.
aus Wiesbaden, Deutschland

Tag der mündlichen Prüfung: 17.05.2024

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

„ Ob mir durch Geistes Kraft und Mund
Nicht manch Geheimnis würde kund;
Dass ich erkenne, was die Welt
Im Innersten zusammenhält. “

Faust — Johann Wolfgang Goethe

Danksagung

Mehrere Personen haben mich während meiner Promotion begleitet, inspiriert und maßgeblich zu dieser Arbeit beigetragen. Diesen möchte ich hier meinen Dank aussprechen:

- Besonders möchte ich Prof. Dr. Jürgen Hubbuch danken, der mir ermöglicht hat, meine Doktorarbeit an seinem Institut anzufertigen. Ich habe deinen wissenschaftlichen Rat, die Freiheit bei den Themen und das Vertrauen in meine Arbeit sehr geschätzt.

- Ich möchte mich bei dem Prüfungskomitee aus Prof. Dr. Gisela Guthausen als Zweitgutachterin und ihr Interesse an meiner Arbeit, Prof. Dr. Sabine Enders für die Übernahme des Prüfungsvorsitzes und Prof. Dr. Dirk Holtmann als Prüfer bedanken.

- Meine Kooperationspartner und die Co-Autoren Brigitte Walla, Daniel Bischoff, Prof. Dr. Dirk Weuster-Botz, Ines Zimmermann und Sebastian Eming haben zu dem Erfolg der einzelnen Projekte beigetragen und mich im Publikationsprozess unterstützt.

- Im Labor wurde ich tatkräftig von meinen Studenten Ines Zimmermann, Sebastian Eming, Christopher Berg, Clara Schmedt, Michael Liu und Bernadette Pichler unterstützt. Bei den jeweiligen Projekten schätzte ich das angenehme Arbeiten im Team und euren eingebrachten Intellekt bei meinem Forschungsprojekt.

- In unserem Ing-Büro konnte ich mich stets an Kristina, Birgit, Jasmin und Nicola mit Fragen zu Geräten, Bestellungen und zum Labor wenden.

- Dank Jan M. und seinen IT-HiWis konnten diverse Probleme mit Laborgeräten, drei defekten Festplatten, Laborrechner oder Server zeitnah gelöst werden. Bester IT-Support!

- Den Weg ans MAB habe ich durch Nils und Philipp gefunden. Ihr habt mir gezeigt, dass sich stressige Forschung für wissenschaftlichen Ruhm und Ehre lohnt und man dabei auch viel lachen kann.

- Meine Bürokollegen Sarah, Sandra, Lukas, Carsten, Annabelle, Jan W., Rafaela und Jakob haben mir den Einstieg in die Promotion erleichtert und dazu beigetragen, dass ich für die Kaffeepausen, Bürostreiche mit Chromeleon & Wackelaugen oder auch mal konstruktives Feedback zur Arbeit gerne ans Institut geradelt bin.

- Annabelle, Robin, Jan W., Svenja, Lukas und Nils habe ich für ihr ehrliches Feedback zu Projektideen, Manuskripten und Vorträgen geschätzt.

- Dank Robin, Jan W., Annabelle, Nils und Angi habe ich die besten Erinnerungen an die Trips zu den Konferenzen und DFG Treffen. Porto Tonico und Peanut Cliff Bars werde ich auch in Zukunft mit euch verbinden.

- Danke Carsten, du hattest immer ein offenes Ohr für meine Belange, hast mich motiviert und mit mir im Labor an den Robbis geschraubt.

- Ich möchte Nici meinen Dank aussprechen für die fesselnden Gespräche im Laborflur sowie ihre Empathie und Unterstützung bei dem Sorgenkind fmlx oder Knut. Ebenso möchte ich Annabelle und Jan für die aufheiternde Stimmung und die Aktiv-Mittagspausen im Fasanengarten danken.

- Die gesamte Arbeitsgruppe MAB mit all ihren Studenten hat in den unterhaltsamen Mittagspausen, mit der Tradition des CWFs, bei den unvergesslichen Dokfeiern, Grillabenden, Fachschaftssommerfesten und Schnabelhausparties die Stimmung stets hoch gehalten. Ich hatte eine echt gute Zeit am MAB.

Außerdem möchte ich Annette Berg und meiner X-Ment Gruppe für die Unterstützung im letzten Promotionsjahr und die guten Ratschläge zur persönlichen Weiterentwicklung danken. Danke auch an meine Freunde aus der Heimat und Karlsruhe, die mir geholfen haben, nach einem langen Arbeitstag oder am Wochenende den Kopf freizubekommen.
Besonderen Dank gebührt meinen Eltern und meinen beiden Schwestern, die mich in all den Jahren meiner Ausbildung unterstützt haben, Verständnis für stressige Projektphasen aufgebracht haben und mit mir mehrfach über Forschung an CHO Zellen und das Wort *precipitation* gelacht haben.
Zuletzt möchte ich meinem größten Fan und Unterstützer Moritz danken. Du hattest immer ein offenes Ohr für die großen und kleinen Probleme auf der Arbeit, hast mich bedingungslos unterstützt und meinen Alltag mit versteckter Schokolade im Arbeitsrucksack versüßt. Danke für alles!

Abstract

Biotechnological innovations have revolutionized the landscape of therapeutics, shifting from medication based on small, chemical molecules to larger, biological molecules, e.g. protein-based drugs. Thanks to the enhanced specificity of biological molecules to receptors, effective treatments could be developed to tackle so far unmet medical needs. Not only the drug specificity, but the production processes changed since these biopharmaceutical products are commonly expressed in living host cells and produced in bio-reactors. Along with the product harvest, these processes are defined as upstream processing (USP). Besides the desired target molecule, the process liquid contains process- and product related impurities, e.g. cell metabolites, nucleic acids, host cell proteins (HCPs), or cell culture fluid compounds. This variety and the amount of impurities need to be depleted in several, consecutive downstream processing (DSP) steps to purify the product and to ensure the patients' safety when the medication is administered. As advances in the USP have led to optimized cell growth or cell metabolism, and to higher product titers in the cell culture fluids, the production bottleneck moved from USP to DSP. Today, purification of therapeutics relies mainly on chromatography as the standard DSP step in the biopharmaceutical industry due to its high selectivity leading to high purity. Achieving this purity comes with economic challenges as the resins are expensive, limited in their capacity, and regenerative capability. Still, to respond to the increased product expression in USP cultivation, new, cost-effective process alternatives for purification processes need to be considered.

Protein crystallization is an alternative DSP step which has been researched in academia for protein structure analysis or industry for the production and formulation of insulin. The self-organization of molecules into a crystal structure is caused by non-covalent interactions and influenced by solution parameters, e.g. pH or temperature. This process comes with a high purity and product yield making it suitable for DSP. Recent advances in protein engineering to improve crystallizability promise a broader application of crystallization processes in the industry and elevate the potential of protein crystallization as an efficient purification step. As a second, alternative DSP step, protein precipitation has the potential to isolate proteins from complex feedstocks in amorphous, unstructured precipitate. Process design for the mentioned DSP alternatives often includes resource-saving, empirical high-throughput (HT) screenings, and thus, fast and reliable analytics. These methods can

meet the needs for implementing more process analytical technology (PAT) tools by the U.S. Food and Drug Administration (FDA). According to the guiding principle quality by design (QbD), quality should be built into the process design to ensure a high product quality. PAT supports this superior goal by designing, monitoring and controlling processes with measurements of critical quality attributes (CQAs), and controlling critical process parameters (CPPs). For this purpose, academia and industry often employ multi-variate, non-destructive, spectroscopic sensors since the recorded spectra contain information about the molecules in the processed fluid on different structural levels, and thus, important CQAs. To analyze the multi-variate spectra, chemometric analysis applies mathematical or statistical methods to extract important information and identify patterns in biochemical systems. Additionally, the development of novel, biological products, e.g. virus-like particles (VLPs), monoclonal antibodies (mAbs) and their variants, antibody-drug conjugates (ADCs), gene or cell therapeutics, requires new process workflows, and thus, process design tools tailored to the new targets. Potential solutions to this issue can involve process design, based on process knowledge, and the implementation of process monitoring or process control strategies. As a result, PAT workflows need to be developed bearing the product characteristics in mind and showing the potential of transferring process knowledge to new modalities. In the past, PAT development in DSP focused on chromatographic separation while accepting the disadvantages, e.g. the high costs, the difficult scale-up, or the low volumetric throughput. However, the implementation of PAT in alternative DSP strategies has received less attention.

Therefore, the objective of this thesis was to develop data-driven PAT for protein crystallization processes which are applicable to various biological products. All analytics were based on multi-variate, spectroscopic measurements and chemometric analysis. Aiming to advance PAT for protein crystallization, this thesis presents (I) a HT-compatible, analytical workflow for screenings of model protein mixtures based on regression modeling with calibration samples, (II) a calibration-free approach for screenings of various modalities in complex feedstocks, and (III) a comprehensive PAT set-up to monitor crystallization in complex lysate on a larger scale. Demonstrated in diverse studies, the developed analytics could quantify the target molecule in heterogeneous, crystalline slurries across different scales from low-volume, HT screenings to lab-scaled crystallization vessels.

Chapter 1 describes the theoretical fundamentals relevant for this thesis regarding the production of biopharmaceutical or biotechnological products, spectroscopy, and data analysis. A special focus is laid on the phase behavior of proteins, protein crystallization for DSP, the influencing parameters, and forces leading to protein crystals or precipitate. Phase diagrams as a tool to visualize phase behavior of proteins are introduced. The spectroscopic techniques used for this thesis – ultraviolet-visible light (UV/Vis) and Raman spectroscopy – are elaborated in this chapter emphasizing their application to analyze proteins. The analysis of multi-variate spectra using multi-variate data analysis (MVDA) and its potential to interpret highly correlated, biochemical data sets are highlighted. Lastly, the advantages of PAT implementation in biopharmaceutical or biotechnological processes are explained, and the idea of QbD and its relation to important key parameters, e.g. CQAs and CPPs, are pointed out. The presented fundamentals are supported with current research in each section focusing on their application to DSP and protein crystallization.

When developing protein crystallization processes, multiple factors need to be considered, e.g. pH, temperature, protein concentration, precipitant concentration, which all influence the crystallization process, process time, yield, and purity. Since resources are scarce during process development of pharmaceutical products, low-volume, empirical HT screenings are popular and can test various different conditions. As a consequence, a large number of samples need to be analyzed, demanding fast, HT-compatible analytics that can be easily transferred to other crystallization studies. In general, most crystallization screening analytics focus on the qualitative assessment of the crystal size distribution, or the characterization of the crystals themselves. When protein crystallization is used for DSP, the concentration of the target molecule is crucial to calculate the CQAs crystal purity and yield, but the presence of impurities in DSP screenings complicates individual protein quantification. Therefore, Chapter 3 presents a rapid, quantitative, and HT-compatible analytical workflow for HT crystallization screenings of model protein mixtures (lysozyme, ribonuclease A, and cytochrome C). Here, lysozyme was treated as the target molecule and the two other model proteins as contaminants. The new analytical tool was based on UV/Vis spectroscopy and chemometric model development with partial least squares (PLS) regression models to quantify the proteins individually in the crystallization supernatant. As a proof of concept, three model proteins were mixed to calculate one PLS model per protein by regressing the recorded UV/Vis spectra to the reference concentrations from cation-exchange chromatography (CEX). The model was then applied to the analysis of supernatants in a protein crystallization screening to find optimal process conditions. The salt concentration, protein concentration, and pH were screened to show the broad applicability of the method to changes in the aqueous environment of the examined proteins. Finally, a kinetic study of two selected screening conditions was performed where samples were analyzed over time to show the transferability of the generated models to different, experimental set-ups. The PLS models showed high accuracy during calibration, the crystallization screening, and the kinetic study. The saturation concentration could be determined as a function of pH and the precipitant concentration, and the crystal yield and purity could be calculated. The results were visualized in a phase diagram to support selecting optimal crystallization conditions. In summary, the data-driven workflow demonstrated that chemometrics paired with low-volume, HT-compatible UV/Vis spectroscopy can be applied to different crystallization studies to specifically quantify proteins in mixtures.

New, biological product classes broaden the therapeutic spectrum and demand fast adaptation of process development workflows to produce the target molecule in high quality for the patient. One strategy to address this issue is phase behavior based process development as these alternative DSP steps can keep up the productivity while maintaining costs at a low level. Since processes based on phase behavior, e.g. protein crystallization or precipitation, rely on screening studies, screening methods and analytics transferable and applicable to new modalities need to be developed. When multi-variate sensors are used to record screening data, large inter-correlated data sets are generated that can be structured in each measurement dimension, e.g. time, wavelength, and sample number. The higher dimensionality of the data imposes new analytical challenges and can be utilized when suitable chemometric methods are applied. The multi-way, chemometric parallel factor analysis (PARAFAC) method can

handle such multi-dimensional data sets and highly overlapping spectra, while revealing the contribution of individual species with regard to each dimension to the overall spectra. Without the need of an extensive calibration procedure, the models can provide process knowledge about the solution composition or pure component spectra when quantitative concentration analytics or purified components are missing. The application of PARAFAC models as a calibration-free analytical tool for phase behavior-based screenings is described in Chapter 4 and demonstrated with crystallization or precipitation screenings of various biological products in chemically defined or complex process fluids. Three screenings were conducted analyzing the liquid phase during crystallization or precipitation, the wash step, and the redissolution step. In fact, the first case study used the screening data of the study described in Chapter 3 to show that PARAFAC models can be used to characterize the solution composition of protein mixtures. The generated model could estimate the concentration and the spectra of two components – of the target molecule lysozyme and the contaminants as a mixture. Only species demonstrating different phase behavior and contributing to the protein spectrum were distinguishable, treating protein clusters expressing similar phase behavior as one species. Quantitative CEX analytics and pure component spectra of the target molecule could validate the estimated model outcomes. The second and third case studies served as real-case scenarios dealing with complex process fluids in a capture step. In the second case study, mAbs were precipitated in harvest cell culture fluid (HCCF), and the screening supernatants were analyzed with UV/Vis spectroscopy during different process steps. The PARAFAC model for the mAb case study could estimate the concentration changes as well as the pure component spectra of the target molecule mAb and other contaminants in the analyzed samples. The model was validated with analytical protein A chromatography and with the pure component spectra of the target. The third case study covered a precipitation screening of VLPs in *Escherichia coli* (*E.coli*) lysate. The VLP case study was conducted similarly to the mAb case study, but quantitative reference analytics were not available. The learnings from the first and second case study could be used to generate a third model that estimated the VLP concentration transferring the preprocessing and model parameters of the first two case studies. For qualitative validation, sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE) proved the presence or absence of the different protein species in the supernatants. Additionally, the model-estimated pure component spectrum of VLPs was compared with a spectrum of purified VLPs. Demonstrating the flexible application to different biological modalities, a multi-way chemometric model called PARAFAC was used to analyze multi-dimensional UV/Vis spectra in screening studies and to estimate the protein concentrations and pure component spectra. The data-driven workflow for calibration-free screening model development contributes to the overall goal of making protein crystallization or precipitation studies more feasible and easier to apply to capture processes, especially when multiple contaminants are present in the analyzed fluids.

On a larger scale than HT screenings, PAT sensors can also be implemented to monitor production processes in real-time. Regarding DSP with chromatography, PAT has been widely employed for measuring CQAs even in complex liquids. However, new challenges arise when PAT should be used to monitor crystallization processes in DSP. Solid crystals and the broad variety of impurities in liquid or potentially precipitated form may be present in the

heterogeneous, complex process liquid and can influence the employed sensors. Furthermore, the crystallization process can induce great concentration changes in the supernatant. These challenges demand an adapted sensor set-up when PAT shall be implemented in protein crystallization processes. Chapter 5 describes the development of an adapted PAT set-up to monitor protein crystallization processes in complex process fluids. The set-up consisted of a Raman probe placed *in situ* in a lab-scaled crystallization vessel, UV/Vis on-line measurements with a variable pathlength (VP) flow cell, installed in a cross-flow filtration (CFF) based bypass to monitor the particle-free supernatant, and additional samples for off-line analysis. The off-line samples were analyzed with immobilized metal ion affinity chromatography (IMAC), enzyme-linked immunosorbent assay (ELISA), SDS-PAGE, and microscopic imaging to evaluate the target molecule concentration, HCP content, protein purity, and the crystals, respectively. The image analysis provided information about the crystal count and crystal geometry. Using this analytical set-up, the crystallization of *Lactobacillus kefir* alcohol dehydrogenase (*Lk*ADH) in clarified *E.coli* lysate was monitored and characterized. In total, five experiments were conducted on a 300 mL scale varying the lysis protocol, precipitant concentration, and whether or not the particle-free bypass was implemented. The Raman probe enabled in-line measurements of the liquid phase in the vessel despite the presence of solids in the crystalline slurry. The development of a PLS model, based on the preprocessed Raman spectra and off-line *Lk*ADH concentration measurements, enabled monitoring the concentration of the target molecule in the heterogeneous lysate supernatant in real-time. The predicted concentration decline in the supernatant indicated that protein crystals were formed and this concentration decline coincided with the detection of the first crystals in the microscopic images of the off-line samples. Concentration mismatches between the model and the reference data were visible at higher target concentration levels when the model was directly transferred to new experiments. Possible reasons may involve batch-to-batch variations and the heterogeneous lysate composition. The particle-free bypass facilitated the implementation of particle-sensitive analytics, here UV/Vis spectroscopy. As a qualitative purity indication, the UV/Vis absorption data could be used to determine the nucleic acid to protein ratio. Thanks to the implemented VP flow cell technology, the sensor could adapt quickly to changes in the UV/Vis absorption values. In the presented study, the CFF-based loop did not lead to crystal breakage as shown by the image analysis. Furthermore, the crystallization and redissolution of selected samples reached a 2-$\log_{10}$ reduction in HCP content in one experiment. As a whole, the objective of this study was the development of a comprehensive PAT set-up adapted to the challenges of protein crystallization processes as a capture step. Regardless of the presence of impurities in soluble or precipitated form, the combination of Raman spectroscopy and chemometrics monitored the specific target concentration in the complex process fluids in a multi-phase systems.

Based on multi-variate spectra and chemometrics, this thesis provides promising data-driven analytics tailored to protein crystallization process development for DSP. Even though various biologics were investigated in process solutions with a varying degree of complexity, the presented PAT tools could quantify the target molecule in each study and be used across different scales. These findings let us rethink conventional process design and move towards a more flexible implementation of protein phase behavior-based processes in DSP.

Zusammenfassung

Biotechnologische Fortschritte haben die Verwendung von therapeutischen Medikamenten grundlegend verändert. Zuvor beruhten Medikamente auf kleinen, chemischen Molekülen, nun sind größere, biologische, proteinbasierte Medikamente in den Vordergrund gerückt. Biologische Moleküle weisen eine verbesserte Rezeptor-Spezifität auf, sodass wirksame Medikamente gegen zuvor schlecht behandelbare Krankheiten entwickelt werden konnten. Dies wirkt sich auch auf die Produktionsprozesse aus, da biopharmazeutische Produkte in lebenden Wirtszellen exprimiert und in Bioreaktoren im größeren Maßstab hergestellt werden. Dieser Prozess wird als *Upstream*-Prozessierung (USP) bezeichnet. Während der Produktion enthalten die Prozesslösung neben dem gewünschten Zielmolekül verschiedene Verunreinigungen, wie z.B. Zellmetabolite, Nukleinsäuren, Wirtszellproteine (*host cell proteins*, HCPs) und Bestandteile des Zellkulturmediums. Diese müssen in der *Downstream*-Prozessierung (DSP) entfernt werden, um eine hohe Reinheit des Produktes und die Patientensicherheit zu gewährleisten. Fortschritte in der USP erzielten ein optimiertes Zellwachstum oder Zellmetabolismus sowie höhere Produktausbeuten. Dadurch hat sich der Produktionsengpass von der USP zu der DSP verschoben. Häufig werden Chromatographieschritte als Standard-DSP-Schritte in der biopharmazeutischen Industrie gewählt, da diese Schritte höchst selektiv sind und eine hohe Reinheit erzielen. Diese Schritte haben jedoch wirtschaftliche Nachteile, da die Chromatographie-Harze teuer sind, die Kapazität und die Fähigkeit die Harze zu regenerieren begrenzt sind. Um die erhöhte Produktausbeute in der USP zu bewältigen, müssen daher neue, kostengünstige Alternativen für Aufreinigungsprozesse verwendet werden.

Die Proteinkristallisation ist ein alternativer DSP-Schritt, der sowohl in der akademischen Forschung zur Analyse von Proteinstrukturen als auch in der Industrie für die Produktion oder Formulierung von Insulin bereits untersucht wurde. Moleküle ordnen sich wegen der nichtkovalenten Wechselwirkungen in einer Kristallstruktur an, wenn geeignete Prozessparameter, z.B. pH-Wert oder Temperatur in der Prozesslösung eingestellt sind. Gerade die hohe Kristallreinheit und die Produktausbeute machen die Kristallisation zu einem geeigneten Schritt für die DSP. Zusätzlich bewirken Innovationen im Bereich des Protein-Engineerings eine verbesserte Kristallisierbarkeit, wodurch eine breitere Anwendung der Proteinkristallisation als effizienter Aufreinigungsschritt in der Industrie möglich wäre. Eine weitere DSP-Alternative ist

die Proteinfällung, bei der Proteine aus komplexen Prozesslösungen zu amorphem, unstrukturiertem Niederschlag aggregieren. Bei der Prozessentwicklung der genannten DSP-Alternativen werden häufig ressourcenschonende, empirische Hochdurchsatz-Screenings durchgeführt, die wiederum schnelle und zuverlässige Analytik benötigen. Diese Methoden können genutzt werden, um Forderungen der US-amerikanischen Behörde *Food and Drug Administration* (FDA) nach Prozessanalytischer Technologie (*process analytical technology*, PAT) in der pharmazeutischen Produktion zu erfüllen. Laut dem Leitlinie *Quality by Design* (QbD) muss Qualität in die Prozessentwicklung integriert werden, um die Produktqualität zu sichern. Durch die Implementierung von PAT kann ein Prozess mit einer erhöhten Produktqualität entwickelt, überwacht und kontrolliert werden, indem kritische Qualitätsattribute (*critical quality attributes*, CQAs) gemessen und kritische Prozessparameter (*critical process parameters*, CPPs) gesteuert werden. Hierfür wird in der akademischen oder industriellen Forschung häufig multivariate, nicht-invasive Spektroskopie verwendet, da die gemessenen Spektren molekulare und strukturelle Informationen auf unterschiedlichen Ebenen über die Moleküle in der Prozesslösung enthalten und damit CQAs bestimmen können. In einem weiteren Schritt werden die multivariaten Spektren mithilfe von chemometrischen Methoden ausgewertet, um wichtige Informationen zu extrahieren und Muster in den biochemischen Systemen zu erkennen. Die Entwicklung neuartiger, biologischer Produkte, wie z.B. virusartige Partikel (*virus-like particles*, VLPs), monoklonalen Antikörpern (*monoclonal antibodies*, mAbs) und seine Varianten, Antikörper-Wirkstoff-Konjugate (*antibody-drug conjugates*, ADCs), Gen- oder Zelltherapeutika, erfordert zusätzlich anpassungsfähige, moderne Techniken der Prozessentwicklung. Tiefes Prozesswissen und Strategien zur Prozessüberwachung oder Prozessregelung können sich hier als nützlich erweisen. In Konsequenz müssen PAT-Werkzeuge so entwickelt werden, dass die Eigenschaften der neuen Medikamentenklasse berücksichtigt werden und im Idealfall bereits generiertes Prozesswissen auf neue Medikamente übertragen werden kann. Bisher konzentrierte sich die Entwicklung von PAT trotz der hohen Kosten, der schwierigen Skalierung oder des geringen, volumetrischen Durchsatzes auf die Chromatographie. PAT für alternative DSP-Schritte wurde jedoch nicht ausgiebig untersucht.

Daher war es das Ziel dieser Arbeit, neue PAT Analytik für Proteinkristallisationsprozesse zu entwickeln, die auf verschiedene, biologische Produkte angewendet werden können. Jegliche Analytik basierte auf multivariater Spektroskopie und Chemometrie. Diese Arbeit beschäftigte sich mit der Entwicklung von PAT für die Proteinkristallisation in drei Studien: (I) die Entwicklung einer hochdurchsatzfähigen Analytik für Screenings von Modellproteinmischungen mithilfe von Regressionsmodellierung und Kalibrierungsproben, (II) die Untersuchung eines kalibrierungsfreien Ansatz für Analytik in Screenings verschiedener, biologischer Produkte in komplexen Prozesslösungen und (III) der experimentelle PAT-Aufbau zur Überwachung der Kristallisation in komplexem Lysat im größeren Maßstab. Die verschiedenen Studien zeigten, dass das Zielmolekül in jeder Studie im Mikro- und Labormaßstab trotz der Komplexität durch heterogenes, biologisches Material in der Kristallsuspension quantifiziert werden konnte.

In Kapitel 1 werden die für diese Thesis relevanten Grundlagen zur Herstellung von biopharmazeutischen oder biotechnologischen Produkten, zur Analytik und Sensortechnik sowie zur Datenauswertung beschrieben. Dabei behandeln die Kapitel detaillierter das Pha-

senverhalten von Proteinen, die Einflussfaktoren und die Visualisierung in Phasendiagrammen mit Fokus auf der Proteinkristallisation für DSP. Die im Rahmen dieser Arbeit verwendeten Spektroskopiearten – die ultraviolett-sichtbares Licht (UV/Vis)- und Raman-Spektroskopie – werden hinsichtlich ihrer Fähigkeit zur Proteinanalyse erläutert. Techniken der multivariaten Datenanalyse (*multi-variate data analysis*, MVDA) werden hinsichtlich multivariaten Spektren in biochemischen Systemen erklärt. Zuletzt werden die Vorteile des Einsatzes von PAT in biopharmazeutischen oder biotechnologischen Prozessen unter der Verwendung der beschriebenen Methoden beschrieben. Hierfür wird die Beziehung von PAT zum Konzept QbD, zu CQAs und CPPs herausgestellt. Aktuelle Forschungsergebnisse sind in jedem Abschnitt mit Fokus auf ihre Anwendung in DSP und Proteinkristallisation aufgelistet und beschrieben.

Bei der Entwicklung von Proteinkristallisationsprozessen müssen mehrere Faktoren berücksichtigt werden, wie z.B. der pH-Wert, Temperatur, Proteinkonzentration, und Konzentration des Fällungsmittels. Diese einstellbaren Parameter beeinflussen den Kristallisationsprozess hinsichtlich Prozess- und Produktcharakteristika, z.B. die Kristallisationsdauer, Kristallausbeute und -reinheit. Aufgrund der geringen Verfügbarkeit von Produktmaterial während der Prozessentwicklung werden in der Regel verschiedene Prozessbedingungen in empirischen Hochdurchsatz-Screenings mit minimalem Produktverbrauch getestet. Infolgedessen fallen viele, zu analysierende Proben an, die von neuer, schneller und hochdurchsatzfähiger Analytik profitieren können. Idealerweise lassen sich diese Methoden flexibel auf unterschiedliche Kristallisationsstudien anwenden. Häufig werden in Kristallisationsstudien die Kristallgrößenverteilung oder Kristallstrukturcharakteristika qualitativ bestimmt. Wenn die Proteinkristallisation jedoch als Aufreinigungsschritt verwendet werden soll, ist die Quantifizierung des Zielmoleküls entscheidend, da damit die CQAs Kristallreinheit und -ausbeute berechnet werden können. Die Quantifizierung gestaltet sich als schwierig, da die Prozesslösungen im DSP viele Kontaminanten enthalten. Daher wurde in Kapitel 3 eine schnelle, quantitative Analytik für Hochdurchsatz-Kristallisationsscreenings von Mischungen von Modellproteinen (Lysozym, Ribonuklease A und Cytochrom C) entwickelt. Hierbei wurde Lysozym als Zielmolekül behandelt und die beiden anderen Modellproteine als Verunreinigungen. Auf Basis von UV/Vis-Spektroskopie und einem chemometrischen Regressionsmodell der partiellen kleinsten Quadrate (*partial least squares*, PLS) sollte die Konzentration jedes Protein in den untersuchten Kristallisationsüberstand gemessen werden. Pro Protein wurde ein PLS-Modell erstellt, indem die aufgezeichneten UV/Vis-Spektren und die Referenkonzentrationen aus der Kationenaustauschchromatographie (*cation-exchange chromatography*, CEX) von Mischungen der drei Modellproteine für die Modellkalibrierung verwendet wurden. Im Anschluss wurden die Modelle in Screenings mit variierter Salzkonzentration, Proteinkonzentration und pH-Wert angewendet, um die breite Anwendbarkeit der Methode trotz Unterschieden in der wässrigen Umgebung der Proteine zu demonstrieren. Schließlich wurden zwei ausgewählte Screening-Bedingungen in einer kinetischen Studie mit Probenahmen über die Zeit untersucht, um die Übertragbarkeit der generierten Modell auf weitere Kristallisationsstudien zu zeigen. Die PLS-Modelle erzielten während der Kalibrierung, des Kristallisationsscreenings und der kinetischen Studie eine hohe Genauigkeit. In Abhängigkeit des pH-Wertes und der Salzkonzentration konnten die Sättigungskonzentration des Zielproteins und damit die Kristallausbeute und -reinheit bestimmt werden. Die Visualisierung der Ergebnisse in Phasendiagrammen

erleichterte die Auswahl optimaler Kristallisationsbedingungen. In diesem Kapitel konnte eine neu entwickelte, schnelle Hochdurchsatz-Analytik auf Basis von chemometrischen Modellen und ressourcenschonender UV/Vis-Spektroskopie auf verschiedene Kristallisationsstudien angewendet werden, um Proteine in Mischungen spezifisch zu quantifizieren.

Die Entwicklung neuer, biologischer Produkte erfordert die schnelle Anpassung von Abläufen in der Prozessentwicklung, um den Patienten das Zielmolekül in hoher Qualität bereitzustellen. Prozessentwicklung auf Basis von Phasenverhalten kann eine Lösung darstellen, da diese alternativen DSP-Schritte eine hohe Produktivität erzielen können, während die Kosten niedrig bleiben. Für die Entwicklung solcher Prozessalternativen, wie z.B. die Proteinkristallisation oder -fällung, werden viele, unterschiedliche Prozessbedingungen getestet. Hochdurchsatz-Screenings und -analytik müssen daher schnell auf neue Molekül- oder Produktklassen übertragen und angewendet werden. Beim Einsatz von multivariaten Sensoren werden große, interkorrelierte Datensätze generiert, die in jeder Messdimension, z.B. Zeit, Wellenlänge und Probennummer, strukturiert werden können. Die erhöhte Dimensionalität der Daten erfordert neue Techniken der MVDA. Die parallele Faktoranalyse (*parallel factor analysis*, PARAFAC) ist eine solche multi-dimensionale, chemometrische Methode, die aus stark überlappenden Spektren den Beitrag der einzelnen Spezies zum Gesamtspektrum hinsichtlich jeder Dimension berechnen kann. Ohne eine umfangreiche Kalibrierung können die berechneten Modelle Prozesswissen wie die Lösungszusammensetzung oder reine Komponentenspektren bereitstellen. Das ist besonders hilfreich, wenn eine quantitative Konzentrationsanalytik oder Reinlösungen der einzelnen Komponenten nicht verfügbar sind. Die Anwendung von PARAFAC-Modellen als kalibrierungsfreie Analytik für Screenings von Phasenverhalten wird in Kapitel 4 beschrieben. In drei Hochdurchsatz-Screenings wurden verschiedene, biologische Produkte in chemisch definierten oder komplexen Prozessflüssigkeiten kristallisiert oder gefällt. Hierfür wurden Überstandsproben während der Kristallisation oder Fällung, den Waschschritten und der Rücklösung spektroskopisch untersucht. Die erste Studie nutzte die Daten des Kristallisationsscreenings aus Kapitel 3, um zu zeigen, dass PARAFAC-Modelle die Lösungszusammensetzung von Proteingemischen bestimmen können. In der ersten Studie konnte das Modell die Überstandskonzentration und zwei Komponentenspektren - des Zielmoleküls Lysozym und der Verunreinigungen als Mischung - abschätzen. Es konnten nur Spezies identifiziert werden, die zum Gesamtspektrum beitrugen und unterschiedliches Phasenverhalten aufwiesen. Spezies mit ähnlichem Phasenverhalten wurden als ein Proteincluster, bzw. als eine Spezies im Modell, behandelt. Quantitative CEX-Analytik und Reinspektren der Ausgangsproteine konnten zur Modellvalidierung verwendet werden. Die zweite und dritte Studie stellten realistische Szenarien in der Prozessentwicklung dar, da Zielmoleküle aus einer komplexen Prozesslösung aufgereinigt werden sollten. In der zweiten Studie wurden mAbs in Zellkulturüberstand (*harvest cell culture fluid*, HCCF) gefällt und Überstände von verschiedenen Prozessschritten während des Screenings wurden UV/Vis-spektroskopisch analysiert. Das erstellte PARAFAC-Modell konnte das Reinspektrum des mAb sowie die Konzentrationsänderungen des Zielmoleküls und anderer Kontaminanten in den Überstandsproben abschätzen. Im Anschluss wurde das Modell mit analytischer Protein-A-Chromatographie validiert. Die dritte Studie untersuchte die Fällung von VLPs in Escherichia coli (E.coli) Lysat. Die VLP-Studie wurde experimentell

analog zur mAb-Studie durchgeführt, aber eine quantitative Referenzanalytik für die Konzentrationsbestimmung des Zielmoleküls stand nicht zur Verfügung. Die Erkenntnisse aus der ersten und zweiten Fallstudie in Bezug auf die Datenvorbereitung und Modellparameter konnten verwendet werden, um ein drittes VLP-Modell zu berechnen, das die VLP-Konzentration in den analysierten Überstandslösungen abschätzen konnte. Der qualitative Nachweis erfolgte mit Natriumdodecylsulfat-Polyacrylamidgel-Elektrophorese (SDS-PAGE), um anzuzeigen, ob Proteinspezies in den Überständen vorlagen. Zusätzlich konnte das VLP-Reinspektrum des Modells mit dem Spektrum einer aufgereinigten VLP-Probe verglichen werden. Multidimensionale PARAFAC-Modelle konnten flexibel auf multi-dimensionale UV/Vis-Spektren angewendet werden, um die jeweilige Proteinkonzentration und die Reinspektren abzuschätzen und um damit das Phasenverhalten von verschiedenen, biologischen Produkten zu untersuchen. Die datengetriebene Entwicklung von kalibrierungsfreien Modellen trägt dazu bei, dass alternative DSP-Schritte in Hochdurchsatz-Screenings von Proteinkristallisation und -fällung besser entwickelt werden können, insbesondere wenn viele Kontaminanten in komplexen Lösungen vorhanden sind.

In größeren Maßstäben können PAT-Sensoren in Produktionsprozesse eingebaut werden, um die Produktion und Produktqualität in Echtzeit zu überwachen. Für Chromatographie in der DSP ist der Einsatz von PAT weit verbreitet, um CQAs in komplexen Flüssigkeiten zu messen. Wenn Kristallisationsprozesse überwacht werden sollen, steht PAT aufgrund der Vielzahl an Störgrößen vor neuen Herausforderungen. Feste Kristalle, Kontaminanten in flüssiger oder potentiell ausgefällter Form und die heterogene, komplexe Prozesssuspension können die PAT-Sensoren beeinflussen. Zusätzlich können während der Kristallisation große Konzentrationsunterschiede in der Flüssigphase auftreten. Daher muss ein PAT-Sensoraufbau spezifisch auf die Anforderungen in Proteinkristallisationsprozessen in der DSP angepasst werden. Kapitel 5 beschreibt die Entwicklung eines solchen, angepassten PAT-Sensoraufbaus zur Überwachung von Proteinkristallisationsprozessen in komplexen Prozessflüssigkeiten im Labormaßstab. Der Aufbau bestand aus einer *in situ* Raman-Sonde im Kristallisationsgefäß, UV/Vis-*On-line*-Messungen mit einer Durchflusszelle mit variabler Pfadlängen-Technologie (*variable pathlength*, VP) in einem partikelfreien Bypass – ermöglicht durch einen Querstromfiltration-Aufbau (*cross-flow filtration*, CFF) – und zusätzlicher Analyse von *Off-line*-Proben. Letzteres wurde mit analytischer, immobilisierter Metallionenaffinitätschromatographie (*immobilized metal ion affinity chromatography*, IMAC), *enzyme-linked immunosorbent assay* (ELISA), SDS-PAGE und Fotomikroskopie untersucht, um jeweils die Konzentration des Zielmoleküls, den HCP-Gehalt, die Proteinreinheit und die Kristalle zu bewerten. Die Bildanalyse lieferte Informationen über die Kristallanzahl und Kristallgeometrie. Mit dem beschriebenen Aufbau an Analytik wurde die Proteinkristallisation von *Lactobacillus kefir* Alkoholdehydrogenase (*Lk*ADH) in geklärtem E.coli-Lysat überwacht und charakterisiert. Im 300 mL Maßstab wurden fünf Experimente durchgeführt, die sich in der Lyse-Prozedur, der Präzipitant-Konzentration und der Implementierung des partikelfreien Bypasses unterschieden. Die Raman-Sonde konnte trotz des Feststoffanteils in einer kristallinen Suspension *In-line*-Messungen der Flüssigphase durchführen. Auf Basis der vorverarbeiteten Raman-Spektren und *Off-line*-Konzentrationsmessungen des Proteins *Lk*ADH konnte ein PLS-Modell entwickelt werden und die Konzentration des Zielmoleküls im heterogenen

Lysatüberstand in Echtzeit überwacht werden. Die Abnahme der vorhergesagten Überstandskonzentration deutete darauf hin, dass sich Proteinkristalle gebildet hatten. Die Auswertung der mikroskopischen Bilder der *Off-line* Proben bestätigte die Bildung von Proteinkristallen. Konzentrationsunterschiede zwischen dem Modell und den Referenzdaten waren bei erhöhter Zielkonzentration sichtbar, wenn das Modell direkt auf neue Experimente übertragen wurde. Mögliche Ursachen könnten *Batch-to-Batch*-Variationen und die durch die Lyse bedingt heterogene Zusammensetzung der Prozessflüssigkeit sein. Dank des partikelfreien Bypasses konnte auch partikelsensitve Analytik, wie z.B. die UV/Vis-Spektroskopie, angewendet werden. Die gemessene UV/Vis-Absorption gab qualitativ Aufschluss zum Nukleinsäure-Protein-Verhältnis und damit zur Reinheit. Die Flusszellentechnologie der VP reagierte schnell auf Veränderungen der Absorptionswerte, die in einem Kristallisationsprozess auftreten können. Zusätzlich konnte die Bildanalyse zur Kristallgeometrie zeigen, dass der CFF basierte Bypass keinen Kristallbruch verursachte. Darüber hinaus wurde in der Kristallisations- und Rücklösungsanalyse gezeigt, dass ein Experiment eine 2-$\log_{10}$-Reduktion des HCP-Gehaltes erreichte. Ziel dieser Studie war die Entwicklung eines umfassenden PAT-Aufbaus mit unterschiedlichen Sensoren, angepasst an einen Proteinkristallisationsschritt in der DSP. Trotz gelöster oder ausgefällter Verunreinigungen konnte die Kombination aus Raman-Spektroskopie und Chemometrie die Konzentration des Zielmoleküls in einer komplexen Prozesssuspension mit mehreren Phasen bestimmen.

Auf Basis von multivariaten Spektren und Chemometrie bietet die vorgelegte Thesis daher vielversprechende, datengetriebene Analytik, angepasst an die Prozessentwicklung für Proteinkristallisation in der DSP. Auch wenn unterschiedliche, biologischen Produkte in Prozesslösungen mit variierendem Komplexitätsgrad untersucht wurden, konnten die entwickelten PAT-Methoden das Zielmolekül in jeder Studie quantifizieren und über mehrere Maßstäbe hinweg eingesetzt werden. Die Fortschritte regen dazu an, die etablierte Prozessentwicklung für biopharmazeutische Produkte neuzugestalten und ermöglichen eine flexiblere Entwicklung von DSP, basierend auf dem Phasenverhalten von Proteinen.

Contents

1

Introduction

Biotechnological and biopharmaceutical products have caused a substantial hype in the modern industry and healthcare system as they offer innovative solutions to existing challenges. Biopharmaceutical therapeutics cover, e.g. cell therapeutics, gen therapeutics, mAbs, recombinant proteins, and VLPs. Applied in medical therapies, they can pave the way to personalized medical approaches with higher efficacy and lower toxicity compared to traditional pharmaceuticals resulting in a better drug compatibility for patients. Beyond healthcare, biotechnological products, in detail enzymes, find multiple applications in diverse industries, e.g. food and beverage, agriculture or bioenergy. As an alternative to traditional chemical production methods, biotechnological processes can save resources, and thus, help to minimize the industrial, ecological footprint.

Advancements in biotechnological USP, e.g. cell line or media optimization, greater process understanding and process control, and the usage of single-use reactors, have enhanced the protein expression in cellular systems. However, this progress intensified the demands for increased productivity in DSP regarding costs and process output. In this context, complementary options to conventional chromatography-based purification methods need to be explored resulting in research focused on alternative techniques, e.g. protein crystallization or protein precipitation processes, in recent decades (see Figure 1.1).

Simultaneously, regulatory authorities have strongly encouraged to introduce PAT in production processes to ensure a high product quality and patients' safety. To meet these requirements, process analyzers are often coupled with multi-variate data analysis to build model-supported analytics for real-time monitoring of CQAs. The product concentration in a heterogeneous, complex solution is a CQA and has often been described for chromatography-based processes combining spectroscopy and chemometrics whereas PAT for protein crystallization focused mainly on systems of less complexity using pure protein solutions or solely

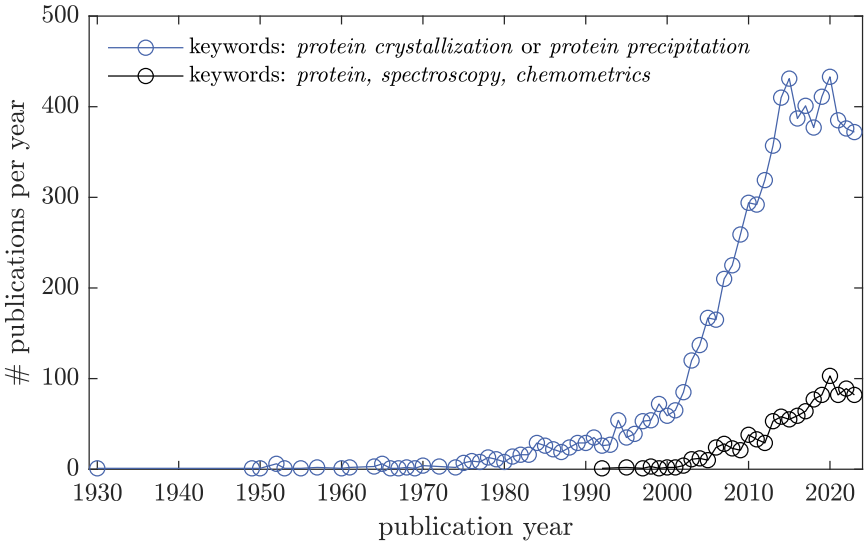


Figure 1.1 Number of publications in the PubMed database [1] in the keyword search with *protein crystallization* and *protein precipitation*, or *protein chemometrics*. The PubMed database selectively lists publications from the fields of biotechnology, biomedicine, bioinformatics and related disciplines. The blue and black circles represent the results of the keyword search with the keywords *protein crystallization* and *protein precipitation*, or *protein, spectroscopy, chemometrics*, respectively.

model proteins. Furthermore, low-volume, HT screenings and HT analytics as a modern methodologies for empirical process development have provided an efficient platform to investigate numerous conditions and identify optimal biotechnological process sweet spots for protein expression and purification. Therefore, merging HT methodologies, spectroscopy, and real-time process analyzers with chemometrics has the potential to advance PAT for phase behavior based process development of proteins at different scales.

The following sections provide a basic, theoretical understanding of protein phase behavior in Section 1.1, of spectroscopic methods in Section 1.2, of MVDA in Section 1.3, and PAT in Section 1.4.

1.1 Protein phase behavior and its influencing factors

Knowledge of protein phase behavior is fundamental in biotechnological process development as the physical states and the phase transitions, that the proteins potentially undergo, directly affect USP and DSP, formulation, protein solubility and stability, and thus, the product quality. As dynamic molecules, proteins can exist or co-exist in different phases in solution, as aggregates or crystalline states [2]. The physical state of the molecule and the interplay of two phase behavior influencing factors can be visualized in a phase diagram, explained in more detail in Section 1.1.3 and exemplified in Figure 1.3. Divided by the solubility line [2], the protein containing solution can either be stable or supersaturated leading to unordered

aggregation, i.e. amorphous precipitation, or structured protein crystallization. The protein phase behavior is influenced by multiple parameters, namely the protein properties (size, shape, hydrophobicity etc.) [3], pH [4, 5], temperature [6–8], ionic strength [9], protein concentration [5, 9–11], precipitant and additives concentration [5, 11, 12], agitation [13, 14] and the presence of contaminants [15–17].

The parameters within the scope of this thesis are discussed in more detail in the next Section 1.1.1 since they are required for the theoretical understanding of the presented thesis. The differences between the mechanism and structure between protein crystals and protein precipitates are outlined in Section 1.1.2. In Section 1.1.3 the visualization of phase behavior in protein phase diagrams is summarized.

1.1.1 Influencing factors on protein phase behavior

To understand the complex interplay of various influences on protein crystallization, the individual parameters are discussed stressing the attractive, molecular forces leading to crystal nucleation and growth.

pH: The pH greatly impacts protein phase behavior as this value determines the charge of each amino acid residue in a protein and the resulting charge distribution on the protein surface. A protein specific characteristic is the pH, at which the net charge of the protein is neutral, and it is referred to as the isoelectric point (pI). At this value the inter-protein repulsion is the lowest caused by equally charged molecules leading to the lowest solubility [18, 19]. Further away from the pI, long-range electrostatic forces prevent attractive forces increasing the protein solubility and stabilizing the molecule [20, cit. on p. 7, 140]. At low salt concentration, the influence of the pH is stronger whereas at elevated salt concentrations the molecule is completely shielded by ions and specific salt effects become more apparent [19].

Inorganic salts: The dissociation of inorganic salts into their respective ions affects the electrostatic interactions between the proteins due to ionic shielding of the protein. The impact of specific ions, namely the Hofmeister series, was first described by Franz Hofmeister [21]. At low salt concentrations, a stabilizing "salting-in" effect is evident. At higher salt concentration, the protein demonstrates quite different phase behavior as the protein undergoes exclusion from the solvent and a destabilizing "salting-out" effect becomes apparent. The order of the ions was determined empirically [22, 23] and can be related to their ability to introduce inter-protein attractive forces. Depending on the position of the ion in the Hofmeister series, specific ions can be classified as kosmotropic ions strengthening hydrophobic interactions and stabilizing the tertiary protein structure, or as chaotropic ions preventing protein aggregation by weakening the hydrophobic protein core and potentially leading to protein unfolding. Even though, the effect of the Hofmeister series has been researched extensively, the mechanism is not completely understood.

Polymers: As precipitants or additives, polymers, mostly polyethylene glycol (PEG), are employed to induce certain phase transitions caused by the volume exclusion effect [24, 25]. As the addition of the polymer lowers the available solvent space for the proteins, proteins are locally isolated from the surrounding media and the increased, attractive protein-protein

interactions may result in protein aggregation or phase transition. This leads to mild, native phase transition which means that the protein structure is not harmed.

Concentration of the protein and precipitant: With increasing protein concentration, the distance between the protein molecules decreases and short-range protein interactions, namely van der Waals and hydrophobic interactions become more significant [20]. When the concentration surpasses the saturation concentration, phase transitions may take place and precipitate or crystals are formed [2, 26].

Presence of contaminants: Contaminants have shown to influence the protein phase behavior as they affect the solubility [15, 16, 27], crystal formation or the likelihood for aggregation. Especially, crystal formation is prone to contaminants as these molecules can compete with the target protein for available nucleation sites on the crystal impeding or slowing down nucleation [16] or growth [15]. Furthermore, molecules with a similar molecular structure may induce protein-protein interactions leading to undesired aggregation or may be built into a crystal detrimental to the crystal purity.

1.1.2 Protein crystallization and precipitation

Protein crystallization and protein precipitation differ in their morphology, structure and mechanism. Protein crystals are well-ordered structures which are commonly used for detailed structure analysis on a molecular level, for formulation, or for DSP purposes more recently.

For the formation of protein crystals, a critical number of molecules need to accumulate and form a crystal nucleus of a specific, critical size. This process is termed crystal nucleation and is a prerequisite for the crystal growth phase when more molecules connect to the crystal surface in a structured order. The crystal nucleation can either happen with only one specific molecule as homogeneous nucleation or in interplay with external molecules as heterogeneous nucleation [26, 28, 29]. The thermodynamic basis behind protein crystallization can be described through Gibbs free energy ΔG for the formation of a spherical nucleus of the radius r when the surface term $\Delta G_{\text{surface}}$ exceeds the volume term ΔG_{volume} in Equation 1.1 [29].

$$\Delta G(r) = \Delta G_{\text{surface}} - \Delta G_{\text{volume}} = 4\pi \cdot r^2 \cdot \gamma - \frac{4}{3}\pi \cdot r^3 \frac{k_{\text{B}} \cdot T_{\circ\text{C}}}{\overline{v}} \cdot \ln(S) \tag{1.1}$$

The interfacial free energy between the crystal and the bulk solution is described with γ, the Boltzmann constant with k_{B}, the temperature with $T_{\circ\text{C}}$, the volume occupied by one molecule with $\overline{v}$, and the supersaturation with S. If the critical nucleus size reaches the radius at the maximum $\Delta G_{\text{nucleation}}$, nucleation is favored and crystals can grow. This process is schematically illustrated in Figure 1.2. The factors of Equation 1.1 influence the critical nucleus size, can be manipulated to influence the crystallization process, and are usually screened to find optimal process parameters. Depending on the level of supersaturation, either crystals at lower supersaturation or precipitate at higher supersaturation occur [26, cit. on p. 183]. For the latter, r_{crit} falls under the size of the smallest structural unit allowing amorphous precipitation instead of ordered crystallization [26, cit. on p. 182]. The mechanism of disordered, amorphous precipitation is illustrated in Figure 1.2 as well [26].

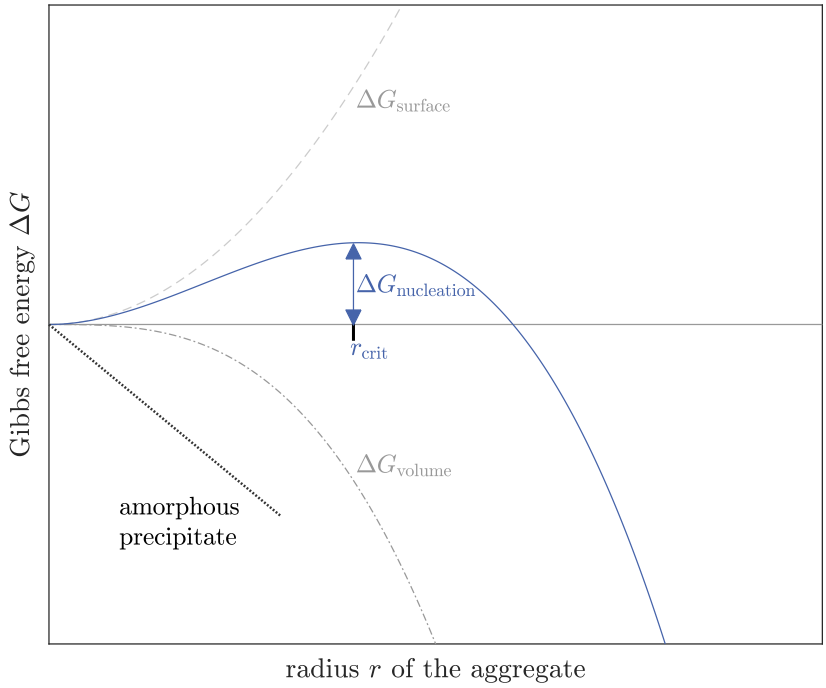


Figure 1.2 Schematic representation of Gibbs free energy for nucleation (adapted from [26, 28, 29]). The dashed, gray lines represent the contribution of the surface and volume term to Gibbs free energy $\Delta G(r)$ which is visualized with the solid, blue line (see Equation 1.1). At the critical aggregate size r_{crit}, the $\Delta G_{\text{nucleation}}$ is overcome. Only then, a stable crystal nucleus is formed and crystal growth can occur. Amorphous precipitate does not require a critical aggregate size, as illustrated with a dotted, black line.

Since the mechanism behind protein precipitation is not unraveled completely [24], empirical screenings with HT methods are popular.

1.1.3 Phase diagrams

To develop efficient phase behavior based processes, phase diagrams can visualize protein phase behavior, and determine the solubility line. They provide insights into the protein's propensity to stay in solution, crystallize, or precipitate, and thus, facilitate finding optimal experimental conditions for a specific molecule. A schematic representation of a phase diagram with varied protein and precipitant concentration is depicted in Figure 1.3. Phase transition can only occur at protein concentrations above the solubility line within the supersaturation zone which can further be divided into the precipitation zone, labile zone and metastable zone. In the precipitation zone, the high supersaturation level results in disordered, amorphous precipitate due to high attractive forces. The labile zone promotes crystal nucleation and

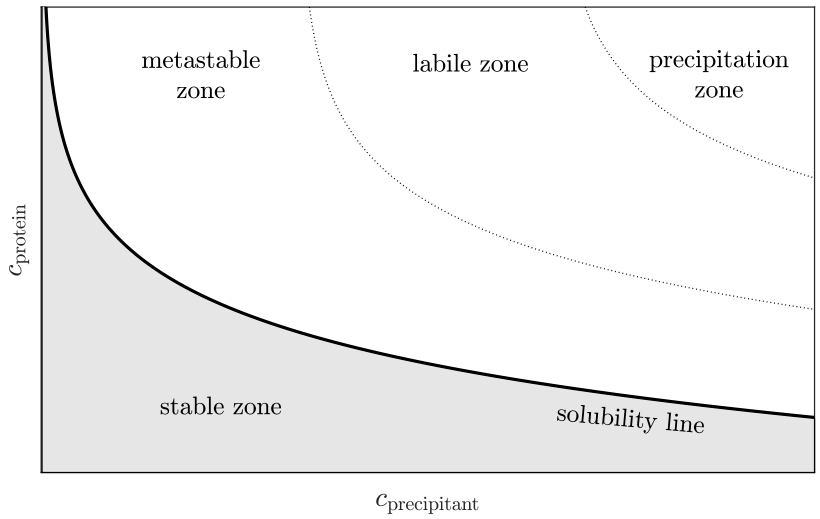


Figure 1.3 Schematic representation of protein phase behavior in a phase diagram (adapted from [2, 30]). Protein phase behavior can be visualized in a phase diagram with the protein concentration c_{protein} and precipitant concentration $c_{\text{precipitant}}$ as influencing factors on both axes. The solubility line is represented by the solid line and separates the phase diagram into a stable, undersaturated zone and a supersaturated zone. The latter can be further divided with the dotted lines into the precipitation zone, the labile zone and the metastable zone. The precipitation zone produces amorphous precipitate, whereas in the labile zone crystal nuclei can form and grow. The metastable zone has a lower degree of supersaturation and allows crystal growth, but not crystal nucleation.

crystal growth whereas only crystal growth occurs in the metastable zone. The undersaturated zone represents stable solution where the molecules are well shielded from each other.

In short, phase diagrams of biologics serve as essential tools for efficient phase behavior based process development by visualizing protein phase behavior, and determining solubility lines. As a basis for scale-up they can provide insights into process sweet spots and have already been used to scale-up purification processes of biologics [14, 31–34]. For the generation of phase diagrams, HT methodologies [9, 35, 36] and HT-compatible analytics [37, 38] are popular as they can save time and resources while increasing the reproducibility [26, cit. on p. 196].

1.2 Spectroscopic methods

Spectroscopy is a powerful, analytical technique that can capture various chemical, physical or biological phenomena. Its sensitivity to these characteristics and non-destructive measurement technique make it especially valuable for real-time monitoring of complex systems, e.g. DSP of biotechnological products. Spectroscopic measurements are based on absorption, emission, or light scattering behavior of molecular systems in response to electromagnetic radiation.

The structural composition, molecular vibration, or the transition of electrons of the sample can cause the changes in the measured data offering a versatile analytical technique on a molecular level.

This chapter delves into the theoretical principles and the applications of spectroscopy in biotechnological processes with a focus on UV/Vis in Section 1.2.1 and Raman spectroscopy in Section 1.2.2 as they are the key sensors used in this thesis.

1.2.1 UV/Vis spectroscopy

UV/Vis spectroscopy is a common technique in analytical biochemistry and is based on the interactions of electromagnetic radiation with the electronic structures of a molecule [39]. The measurement principle relies on absorption of light in the ranges of ultraviolet light and visible light over 200 to 380 nm and 380 to 720 nm, respectively [39, 40]. Initially, the energy level of molecules is at the ground state, but can be elevated to a higher energy level if the incoming electromagnetic radiation corresponds to the difference between the ground and elevated excitation state, also known as electronic transition [41, 42]. Depending on the molecule and its electronic structure, different excitation states can be reached. When exposed to light, energy-specific spectra are created, visualizing how light of different wavelengths is absorbed, revealing the electronic transitions caused by the molecule. Chromophores are the reason for these electronic transitions and are UV/Vis active functional groups on a molecular level [42]. In the case of proteins, the chromophores are generally the amino acids Tryptophan (Try), Tyrosine (Tyr), and Phenyalanine (Phe) [43], demonstrating strong absorption behavior in the wavelength range of 255 to 285 nm [44]. However, peptide bonds in the spectral range over 200 to 230 nm [45, 46], the secondary or tertiary protein structure [46–48], and disulfide bridges [46] influence the UV/Vis spectrum, but can be neglected if any aromatic residues are present [46].

The quantitative relation between the measured absorbance A_{analyte} and concentration c_{analyte} of an analyte depends on the path length d_{path}, wavelength λ specific extinction coefficient ϵ, and the initial I_0 to measured intensity I and can be described using the law of Lambert-Beer (see Equation 1.2) [40, 41]. Note that the Lambert-Beer law is valid solely for solutions with low concentrations of UV/Vis active analytes [41].

$$A_{\text{analyte}}(\lambda) = \log(\frac{I_0}{I}) = c_{\text{analyte}} \cdot d_{\text{path}} \cdot \epsilon(\lambda) \tag{1.2}$$

Throughout various DSP, UV/Vis spectroscopy can offer real-time information about the concentration and purity of the processed solution containing e.g. different proteins or nucleic acids. Typically for proteins and nucleic acids, their absorbance maximum lies around 280 nm [44] and 260 nm [49], respectively. Thus, the calculation of the ratio $A_{260\ \text{nm}}/A_{280\ \text{nm}}$ is a measure for the protein-to-nucleic acid ratio and offers the possibility to evaluate samples regarding their purity. This property has been useful in multiple, biotechnological separation processes when purity is crucial [34, 50, 51]. When multiple UV/Vis active analytes are present, e.g. in an intermediate process solution, it is assumed that the absorption of each analyte can be summed up over the number of analytes and the chromophores do not interact

strongly. Selective quantification is still possible due to the spectral differences between the molecules and the application of MVDA (see Section 1.3) [52, 53]. In food, environmental or pharmaceutical sciences, multiple studies have been conducted combining UV/Vis and MVDA [54]. For chromatography processes in biopharmaceutical production processes, this combination has been used successfully to monitor elution profiles of proteins and quantify the specific elution peaks using Lambert-Beer's law [55–57], or to determine break-through curves [58, 59] making it suitable for real-time monitoring and process control. However, the application of UV/Vis spectroscopy is not only limited to chromatography processes, but has also been applied in ultrafiltration/diafiltration (UF/DF) processes of e.g. mAbs [60] or VLPs [61]. Technological advances of UV/Vis spectrophotometers have contributed to a broader applicability in the field of biotechnology due to VP technology [56, 60, 62] or attenuated total reflection (ATR) probe technology [63, 64] and promote further PAT development with UV/Vis sensors. In summary, the listed applications underline that UV/Vis spectroscopy is a popular, quantitative, and non-destructive sensor flexibly employed in biotechnological processes of different biological products.

1.2.2 Raman spectroscopy

Raman spectroscopy is a vibrational spectroscopic technique that provides detailed information about the vibration, rotation, and other low-frequency movements of molecules. Unlike other spectroscopic methods involving light absorption, e.g. UV/Vis spectroscopy, its measurement principle is based on light scattering, namely inelastic scattering of monochromatic light of a laser. This phenomenon can be used to find unique fingerprints of the vibrational and rotational modes of the molecules in the sample through the analysis of Raman shifts [66]. In the case of biological samples, C=0 and C–N bonds typical for proteins as well as C–C, C–O or aromatic side chains, e.g. of tryptophan, tyrosine, and phenylalanine, cause molecular vibration and contribute to the unique, molecular fingerprint [67, 68].

Most scattered light, i.e. photons, undergo elastic scattering, known as Rayleigh scattering, maintaining the energy of the incoming photon. Only a small fraction of the photons experience inelastic scattering (see Figure 1.4) which means that a change in the vibrational mode, and thus, energy level, occurred. Inelastic scattering can be further divided into Stokes Raman scattering and Anti-Stokes Raman scattering where a change of the emitted energy of the photon occurs either to a higher or lower wavenumber, respectively. Due to the Boltzmann distribution law, the first effect is easier to detect and measure. The described energy shift can be recorded as the wavenumber shift from the wavenumber of the initial laser. As only one photon in 10^{10} photons undergoes inelastic Raman scattering [65], the Raman effect is difficult to obtain. Advances in the measurement technology [66, 69] have increased the sensitivity to measure Raman-active molecules, functional groups and secondary structures of proteins [70] and enabled researchers to use Raman spectroscopy in more complex processes with multiple overlaying species vibrating in the Raman spectrum.

As Raman spectra are constructed from many, inter-correlated variables and show nonlinear behavior [71], empirical model calibration using MVDA [72] is advised to quantify the analyte or investigate a sample qualitatively. Real-time in-line monitoring with *in-situ* Raman

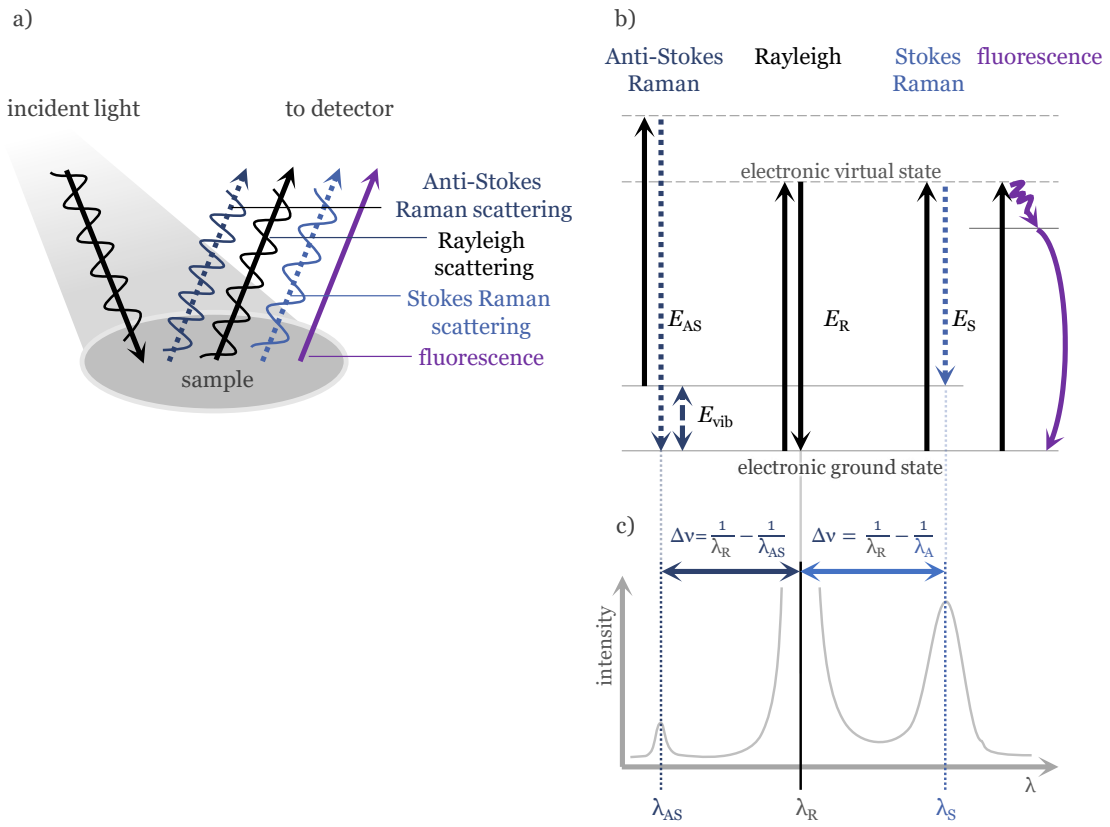


Figure 1.4 Schematic representation of elastic and inelastic scattering (adapted from [65]). Different scattering effects of monochromatic light on a sample in a) and their respective ground and excited energy states in b) are visualized. Variations in the wavelength $\lambda_{AS,\ R,\ S}$ are schematically illustrated in c) and account for the wavenumber shifts $\Delta\nu$. Some light undergoes elastic scattering called Rayleigh scattering (visualized in black color) and scattered light has the same wavelength as the incident light. The effect of fluorescence (visualized in purple color) occurs when the energy of the incoming light matches the molecular, electronic energy level and a photon of a longer wavelength is emitted during de-excitation. When the incident photons cause molecular vibrations or oscillations, molecule-specific, inelastic Raman scattering (Visualized in light blue color) can take place under emission of a photon of a lower wavelength. Photons as inelastic Anti-Stokes Raman scattering (visualized in dark blue color) can be emitted with additional energy if the molecule was initially in an excited, vibrational state with the energy transition E_{vib} prior excitation with the incident light. In reaction to the monochromatic light, molecules can undergo the different energy transfers caused by Rayleigh scattering E_R, by Stokes Raman scattering E_S, or by Anti-Stokes Raman scattering E_{AS}.

probes applied in USP has been in the focus in the biotechnological industry [72–76], whereas fewer, specific case studies have been reported in the last decade using Raman spectroscopy in later states of the purification process, e.g. during harvest of a continuous mAb cultivation process [73], chromatography [77, 78], redissolution of an active pharmaceutical ingredient (API) [64], during freeze-drying [79] or freezing [80] of API or to investigate membrane fouling [81]. Regarding the crystallization of chemical pharmaceuticals, multiple studies have been conducted investigating real-time monitoring with *in-situ* probes [82–86]. In conclusion, Raman spectroscopy is versatile, valuable and non-destructive sensor in biotechnological processes and offers molecular fingerprints due to the unique, vibrational movements of the investigated materials.

1.3 Multi-variate data analysis

MVDA is a statistical approach to design and analyze complex data sets with multiple variables. It aims to reduce the dimensionality of variables, unravel underlying patterns, relations, and trends within the data, and visualize the experimental data. By considering and analyzing multiple variables, MVDA can support the operator to understand the investigated system thoroughly, design or optimize processes with QbD, and increase reliable process control of pre-existing production plants. In the field of USP and DSP, MVDA is commonly applied with in-line or on-line spectroscopic sensors to monitor CQAs in real-time, control the product quality, or optimize processes [43, 87–89].

MVDA methods can be classified as unsupervised or supervised. Regarding unsupervised MVDA, no outcome or variable is predefined and the focus is on data exploration, pattern recognition, and clustering without prior knowledge of the response variable. Contrary, supervised methods are guided by a response variable during method development, and aim to predict the outcome or response variable for unknown data.

When biological or chemical data are at hand, MVDA is referred to as chemometrics and often applied on spectra, chemical data, or experimental designs. In the case of spectroscopic sensors, data sets can be recorded across different wavelengths, frequencies, or wavenumber, and capture information on the absorption, emission, or scattering of electromagnetic radiation by molecules depending on the analyzed sample composition, or structure. As spectra are strongly correlated, data reduction strategies are advised to extract meaningful insights into the molecule. One of these strategies is principal component analysis (PCA), which can condense spectral variance into variables termed principal component (PC). Each of these PC is built from original variables containing comparable information. When a data set can be structured in more than two dimensions, new methods are necessary to explore the data set, e.g. with the multi-way MVDA method named PARAFAC. Specifically in DSP, a three-dimensional (3D) data set may be generated when spectra are recorded over time and the data can be structured along the wavelength, absorption, and time. When specific target variables in DSP need to be predicted on unknown data, PLS regression models are suitable. These models can monitor processes in real-time when a model was calibrated before.

The next sections discuss the MVDA techniques with a focus on the application on spectral data sets. PCA and PARAFAC are both unsupervised and further explained in the Sections 1.3.1 and 1.3.3. Their core difference lays in the structure of the input data as the methods PCA and PARAFAC analyze two-dimensional (2D) and 3D data sets, respectively. The Section 1.3.2 describes the theory behind supervised PLS regression model calculation regressing 2D data to target variables.

1.3.1 Principal component analysis

PCA is an unsupervised MVDA technique used for exploratory reasons. The dimensionality of the data is reduced, trends or patterns within the data set can be extracted. Furthermore, it can be used to represent data in a simpler way, select important variables, or detect outliers [90, 91]. Originally, it was first used in economics or social sciences, but bears the opportunity to analyze chemical or biological data. This section focuses on the application of PCA on spectral data as spectroscopic sensors are often implemented in biotechnological processes.

$$X = TP^{\top} + E_{\mathrm{X}} \tag{1.3}$$

By decomposition of the mean-centered spectra X into linear combinations of the original data according to Equation 1.3 and minimizing the error matrix E_{X}, the scores matrix T and the loadings matrix P capture the maximum variance in the spectra [91]. For biological or chemical applications, the centered spectra X are structured with m wavelength variables and n measurements. The column number of the scores matrix T denotes the number of PCs which represent the transformed data into a new coordinate system defined by the PCs. These PCs are orthogonal to each other meaning that they are uncorrelated to each other and, in fact, perpendicular in the transformed space. Furthermore, the PCs are ordered by their corresponding contribution to the variance of the spectral data set. It is advised to focus on the first few PCs as they are often enough to represent the original spectra and specific PCs can correlate to biochemical or physical phenomena, e.g. the presence of a new reaction product, a specific compound, or aggregation [90]. A schematic representation of the decomposition

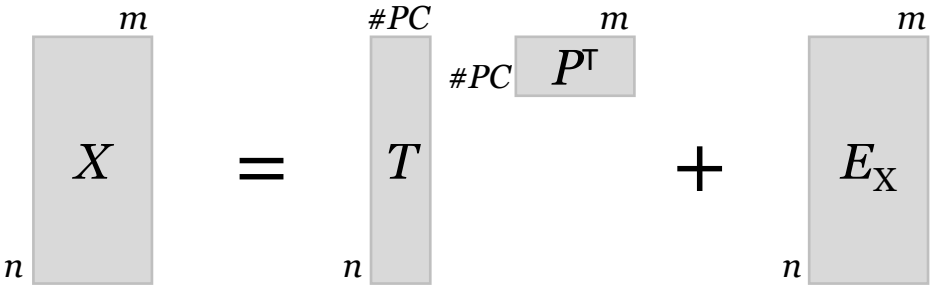


Figure 1.5 Schematic representation of PCA (adapted from [90]). The centered data set X is built from n observations and m variables and can be decomposed into the scores matrix T and loadings matrix P with respect to the error matrix E_{X}. The number of PC is $\#PC$.

by PCA is depicted in Figure 1.5. The PCA scores correspond to spectral measurements projected into the new coordinate system and represent the variance of the measurements within the data set along the PCs. This means that in a scores plot, measurements located close to each other are similar within the data set whereas measurements far away from each other demonstrate opposite effects. The loadings show how each variable, in our case each wavelength, contributes to the spectral data set variance by representing the correlation coefficients between the wavelengths and the PCs.

The application of PCA can be limited in biotechnological processes when the data behaves in a non-linear manner, a large number of outliers distort the results, or important, biological information are potentially lost due to the dimensionality reduction with too little PCs. However, PCA found multiple application using mainly optical sensors in USP [87, 92, 93] and DSP [92, 93] for outlier detection, pattern recognition, or classification of observations. Selected applications of PCA are the evaluation of a microbial fermentation process [94], mammalian cell cultures [95, 96], spectral similarities in multi-component solutions [97] or PAT sensor set-ups of crystallization processes [98].

1.3.2 Partial least squares regression

PLS regression (also known as projection to latent structures) is a supervised MVDA technique designed for the reduction of correlated variables to a few latent variables, for an improved data interpretation, or for the prediction of one or multiple target variables [99]. In biotechnological processes, multi-variate spectra are commonly used as input data, and thus, this sections focuses on the application of PLS regression on spectroscopic data. For PLS model calculation, a data set with multiple variables X, in our case a spectral data set, is regressed to its corresponding response variables Y. Analogous to Section 1.3.1, both data sets are structured with m variables and n observations. PLS models can either be used to classify into groups or predicting the response variables by reducing the multi-variate spectral data set to a low number of latent variables which explain the most variation in the spectra. Explaining as much covariance as possible between the mean-centered spectral data set X and response variable Y, X is decomposed into T and P, Y is decomposed simultaneously into the scores matrix U and loadings matrix Q according to Equation 1.3 and 1.4, respectively. Schematically, the model calculation is illustrated in Figure 1.6.

$$Y = UQ^{\top} + E_{\mathrm{Q}} \tag{1.4}$$

The error matrix E_{Q} describes the residuals from the PLS decomposition of response variables Y. Different from PCA, scores and loadings are determined to maximize the covariance between scores matrices T and U. The derived principal components are named latent variables. Finally, for the response data set Y a regression coefficient matrix B_{reg} with m rows is determined to fit the Equation 1.5 with respect to the error matrix E_{Y}.

$$Y = XB_{\mathrm{reg}} + E_{\mathrm{Y}} \tag{1.5}$$

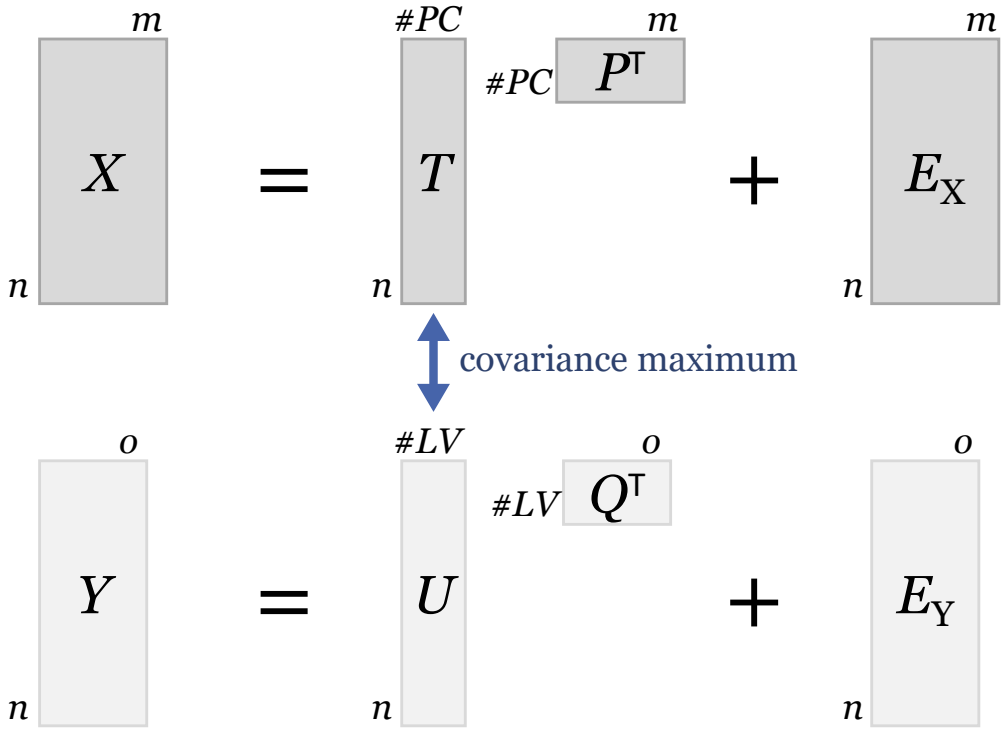


Figure 1.6 Schematic representation of PLS regression. The centered data set X is built from n observations and m variables and can be decomposed into the scores matrix T and loadings matrix P with respect to the error matrix E_X. The number of PC is $\#PC$. The response variable data set Y is built from n observations and o variables and can be decomposed into the scores matrix U and loadings matrix Q with respect to the error matrix E_Y. The number of latent variables is $\#LV$. The decomposition of X and Y is performed simultaneously and aims to maximize the covariance between the scores matrices T and U.

Analogous to PCA (see Section 1.3.1), the scores plots visualizes the distribution of recorded spectra in the space of the new coordinates, and aids clustering or outlier detection whereas loading plots illustrate the relationship between the variables, i.e. the wavelengths, and the latent variables. Ideally, the error matrices E_X, E_Y, and E_Q contain solely noise caused by the detector, the experiment itself, or irrelevant, biological or chemical phenomena.

For the implementation of PAT in biotechnological processes, PLS regression models are commonly used in USP and DSP [92, 93]. In USP, the content of nutrients, APIs, e.g. mAbs [93, 100], or metabolites could be monitored in fed-batch mammalian cell cultures [75, 76, 101], or filamentous cultivation systems [102]. In DSP, PLS models found multiple application in monitoring and potentially controlling mAb conjugation [103] or VLP (dis)-assembly reactions [61, 104], chromatography [56–59], and UF/DF processes. Furthermore, crystallization of small, organic compounds or chemical pharmaceuticals in relatively pure solutions was monitored using PLS and different spectroscopic methods investigating API content [105, 106] or drug crystallinity [107]. As PLS regression models could be applied in

multiple steps of a biotechnological production process, they have proven as valuable tools to predict target variables, and support the goal of monitoring and controlling processes.

1.3.3 Parallel factor analysis

PARAFAC is an unsupervised technique in multi-way MVDA to analyze a multi-dimensional data set and commonly finds its application in chemometrics or sensory analysis. Thanks to the decomposition into a sum of rank-one tensors, hidden structures within a multi-dimensional data set can be unraveled without the necessity of further response variables. Especially with complex interactions of inter-correlated variables and with noisy data, PARAFAC may be able to identify underlying patterns when other methods fail. Specifically for the application in biotechnological processes, spectroscopic sensors are popular as they offer biochemical information on different, structural levels [43], and thus, are the focus for the following theoretical description of PARAFAC models within the scope of this thesis.

The multi-way technique PARAFAC can make use of the *second-order advantage* [108]. This means that a three-way tensor is decomposed into three numerically equally treated vectors and that, in contrast to PCA, PARAFAC models do not have to be orthogonal. The PARAFAC model of a three-way tensor X with the three dimensions $j,\ k,\ l$ can be described with Equation 1.6 for each element $x_{j,k,l}$ in X as followed:

$$x_{j,k,l} = \sum_{f=1}^{F} (a_{j,f} \cdot b_{k,f} \cdot c_{l,f} + e_{j,k,l}) \tag{1.6}$$

For this notation, F is the number of species, $e_{j,k,l}$ is the element of the error tensor $E_{3\mathrm{D}}$. The variables $a_{j,f},\ b_{k,f},\ c_{l,f}$ are elements of loading matrices $A,\ B$, and C. The PARAFAC model is found when the sum of squares of the errors $e_{j,k,l}$ is minimized. For an exemplary, biotechnological application, a spectral tensor could be structured along the dimensions of the wavelengths, time, and sample number. A three-way PARAFAC model with two species is visualized in Figure 1.7.

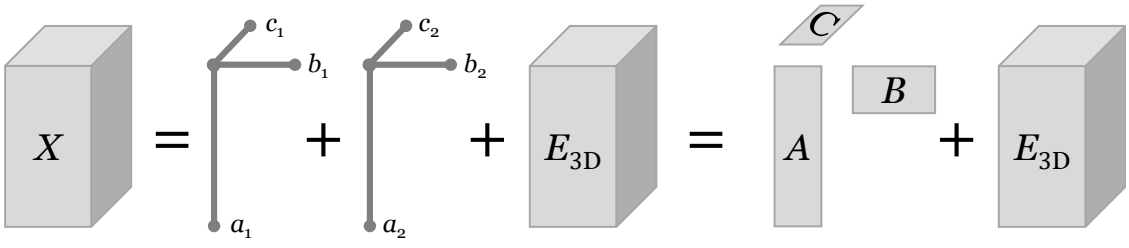


Figure 1.7 Schematic representation of a three-way PARAFAC model. The decomposition of a three-way tensor X along the dimension $j,\ k$ and l into the loading matrices $A,\ B$, and C with the respective vectors $a_f,\ b_f$, and c_f for the species f, in this case two species. The error tensor $E_{3\mathrm{D}}$ contains ideally only noise. Each vector product per species and the matrix product are of the same structure as X.

The number of components for a PARAFAC model needs to be selected with care as it has a great affect on the model outcome. To evaluate if the right number of component was chosen and the model is suitable for the selected data set, the metric core consistency diagnostic (CORCONDIA) can be used [109]. Compared to residual-based metrics, CORCONDIA assesses the appropriateness of the multi-way model using the *second-order advantage*. This value is less or equal to 100 % with high values indicating a high model appropriateness with the selected number of components.

Assuming linearity between the variables, PARAFAC models are limited to represent complex, non-linear biological phenomena. Furthermore, these models are challenging to interpret requiring additional validation analytics and biological knowledge on the investigated process. Despite these challenges, PARAFAC models have been employed successfully in the water industry to characterize e.g. drinking or waste water quality [110, 111], antibodies for structural protein analysis [112] with fluorescence spectroscopy, or to quantify metal complex in catalyst research [113], and API in a pharmaceutical formulation [114] with UV/Vis spectroscopy. In conclusion, PARAFAC as a multi-way MVDA technique offers a versatile approach to uncover underlying structures in a multi-dimensional data set and has great potential to be applied in more diverse industries than the mentioned ones.

1.4 Process analytical technology

Since the FDA and European Medicines Agency (EMA) both request the implementation of PAT in the pharmaceutical industry [43, 115, 116] to monitor, analyze and control CPP and CQA, a paradigm shift in manufacturing and process control from traditional batch-based quality control methods towards the usage of real-time PAT has taken place to ensure product quality and process efficiency. Throughout the entire production from raw material analysis to the final product quality evaluation [117], more PAT methods are applied. In general, PAT tools can consist of one or a combination of the following techniques: process analyzers, data acquisition and MVDA, process control tools and continuous knowledge management tools [115]. The process analyzers can further be classified as off-line measurements of drawn samples analyzed further away from the process, at-line measurements of drawn samples near the process, on-line measurements in a bypass process stream, or in-line measurement when the process stream is directly measured. Especially the latter is desired as real-time monitoring allows for immediate process adjustments reducing the likelihood of product deviations.

With respect to DSP, PAT tools have been developed and discussed, investigating e.g. harvest steps [73], chromatography of biopharmaceuticals [56, 62, 77, 118, 119], freeze concentration [80], crystallization processes of small biologics [120, 121] or chemical products [122], UF/DF processes [60, 104], or formulation [123]. In summary, high product quality and process efficiency is demanded by the authorities and has increased the research interest in PAT tools for monitoring or potentially controlling processes throughout the whole production of biopharmaceuticals. As sensor technologies or MVDA methods continue to develop, PAT

will be more versatile in the future and crucial to securing productivity and quality in biopharmaceutical manufacturing.

2

Thesis outline

2.1 Research proposal

Biopharmaceuticals have revolutionized modern medicine due to their increased specificity to receptors. As the biopharmaceutical and biotechnological product portfolio is growing quickly and the high pressure of costs has reached the biopharmaceutical industry, production processes need to be more flexible and cost-effective while maintaining a high product quality. In the past, protein crystallization was associated rather with the structural analysis of protein, but knowledge on the protein phase behavior has proven useful throughout multiple process steps during recovery, purification, or formulation. Among various techniques for downstream processing (DSP), phase behavior based processes, namely protein crystallization or precipitation, have become a cost-effective alternative to traditional, chromatographic purification resulting in products of high purity and efficacy.

As product quality is crucial for the patient safety, governmental agencies, namely the U.S. Food and Drug Administration (FDA) or the European Medicines Agency (EMA), request quality built into the process design by analyzing, monitoring, and controlling the pharmaceutical production. To comply with these demands, process analytical technology (PAT) is applied during process design and realized with multi-variate, spectroscopic sensors. Often the recorded spectra are coupled with data-driven, multi-variate data analysis (MVDA) methods to monitor the process aiming to adjust the critical process parameters (CPPs), and thus, control critical quality attributes (CQAs). Information about the molecule or the process can be derived from supervised and unsupervised MVDA techniques, e.g. the individual quantification of one species in mixtures of multiple ones. In traditional DSP development, PAT has been used extensively to design, monitor, and control chromatography. However, these established tools cannot be directly transferred to protein crystallization

processes as the presence of crystals and the solid-liquid interfaces impose new challenges calling for adaptions of sensor set-ups, sampling, and MVDA techniques.

The objective of this thesis is to establish data-driven PAT applicable to various biological targets to design and monitor protein crystallization processes effectively. MVDA techniques are applied to interpret the multi-variate spectra and measure CQAs of the target despite the presence of impurities or solid crystals. This thesis aims to quantify the target specifically in crystallization processes by developing (I) a quantitative PAT workflow for crystallization screenings, (II) a calibration-free PAT approach for screenings of various products in complex feedstocks, and (III) a broad PAT set-up, consisting of multiple in-line, on-line, and off-line analytics, to monitor protein crystallization in complex lysate on a larger scale. The conducted studies will involve mixtures of model proteins in chemically defined solutions, recombinant enzymes, monoclonal antibodies (mAbs), or virus-like particles (VLPs) in complex feedstocks. Investigations of different phase behavior based processes, in detail protein crystallization or precipitation, at micro- and lab-scale will further highlight the flexibility of the developed analytics across different scales.

When processes need to be characterized and optimized, high-throughput (HT) screenings come in handy as they involve rapid, automatable experiments in small-scale with minimal material consumption. They can be conducted systematically to test a large number of process conditions or compounds to increase process understanding, thus, complying with quality by design (QbD). Often used in biotechnological process development, HT methods help finding optimal process sweet spots and can be easily transferred to different biological molecules. Regarding protein crystallization, HT screenings are commonly used to investigate the phase behavior in pure protein solutions. The established HT analytics in literature primarily cover qualitative characteristics with automated image analysis or protein structure analysis focusing on the protein crystal or crystal size distribution. However, these methods cannot provide the quantitative information about the target molecule when protein crystallization should serve as a purification process step. Other quantitative methods may determine the purity or crystal yield as relevant CQAs, but they require a considerable amount of resources, sample preparation and analysis time, especially for a large number of samples. Thus, the first study (Chapter 3) aims to develop a rapid, quantitative, and HT-compatible analytical tool for HT crystallization screenings of a target molecule in a mixture of model proteins. As a first step prior the crystallization screenings, three model protein solutions will be mixed according to a selected design of experiments (DoE) approach. The ternary protein solutions will be analyzed with ultraviolet-visible light (UV/Vis) spectroscopy and a suitable reference method. The generated data will be used to calibrate and validate a chemometric model, i.e. a partial least squares (PLS) regression model. The second and third steps will involve a selective crystallization screening and kinetic study in micro-liter scale. The crystallization supernatants shall be analyzed using UV/Vis spectroscopy, and the specific protein concentration should determined using the recorded data and calculated model to demonstrate the transferability of the chemometric model to the kinetic study. This study will serve as a proof-of-concept that a PAT tool based on UV/Vis spectroscopy and PLS regression can selectively quantify species in a mixture in HT crystallization screenings with a low analysis time and sample consumption.

When HT screenings for crystallization or precipitation DSP are conducted using spectroscopic sensors, large data sets are generated that can be structured in multiple dimensions. To explore the data set at hand and reveal underlying patterns, specific chemometric methods can make use of this underlying structure. The unsupervised, multi-way chemometric model parallel factor analysis (PARAFAC) has found many applications in the analysis of chromatographic or kinetic data sets. It aims to extract quantitative information about the composition of the sample containing multiple species. In DSP capture steps, process solutions hold various impurities or potential modifications of the target, and are subject to variations during the upstream processing (USP) steps or raw material quality. These real-case scenarios are challenging for HT analytics, especially in the early stages of process development with crystallization or precipitation screenings, where CQAs need to be measured quickly regardless of the biological target, the impurity composition, or batch-to-batch variations. Therefore, the second study (Chapter 4) shall investigate a calibration-free, HT-compatible, analytical workflow that can be applied to crystallization or precipitation HT screenings of various biopharmaceuticals. Screening supernatants are analyzed with UV/Vis spectroscopy during the process steps of crystallization or precipitation, wash, and redissolution steps. Using all UV/Vis spectra of one screening study, a multi-dimensional data set will be generated spanning across the dimensions time, wavelength, and sample. In a second step, PARAFAC models will be calculated making use of the underlying higher structure in the multi-dimensional data set and revealing hidden patterns in the recorded spectra. These PARAFAC models can provide valuable process information about the impurities present in the analyzed solutions, the sample composition and the pure component spectra of the target molecule. Three screening studies will be conducted to demonstrate the applicability of this analytical workflow to various biopharmaceuticals and to different DSP steps based on protein phase behavior. In detail, protein crystallization or precipitation screenings of enzymes, mAbs, or VLPs shall be examined. Additional reference analytics will be used to validate the sample composition or the pure component spectra of the target molecule. In summary, this study aims to develop an analytical model-based approach universally applicable to crystallization or precipitation HT screenings of different biological products without the need for reference analytics. The model outcome shall quantify the target molecule in the sample to determine appropriate process conditions in the crystallization or precipitation screenings with respect to the product purity.

Real-time monitoring is an important key feature of PAT with the aim of QbD and a thorough process understanding. For these purposes, spectroscopic sensors are applied in biotechnological processes and coupled with MVDA to provide valuable process information over time. In DSP, extensive research has been conducted to monitor chromatography processes using multiple or combinations of the aforementioned techniques. However, quantitative monitoring PAT tools for protein crystallization in complex process solutions are rare and impose special challenges due to the presence of solid crystals when measured in-line. To overcome these limitations, the third study (Chapter 5) covers the development of a monitoring PAT set-up tailored to protein crystallization out of complex process solutions. This PAT set-up will consist of spectroscopic in-line sensors, on-line sensors and additional off-line analytics for validation. The difficulties of solid and liquid phases during crystalliza-

tion need to be addressed, e.g. through sophisticated sensor selection or sampling strategies. As an example process, a crystallization capture step of a recombinant protein in complex process liquids shall be investigated. Different crystallization experiments with varying conditions will be conducted and spectroscopic data sets will be recorded. In a next step, MVDA methods will be applied to analyze the spectroscopic data and will be validated with reference analytics to facilitate real-time monitoring of specific CQAs. To comply with QbD for biotechnological crystallization processes, this study will aim to monitor protein crystallization in heterogeneous, complex, multi-phase solutions using different in-line and on-line spectroscopic sensors.

2.2 Manuscript overview

This section presents an overview of the manuscripts prepared in line with this thesis. The manuscripts in the Chapters 3 and 4 deal with PAT analytics applied to HT screening studies and were published as outlined below. The manuscript of Chapter 5 describes PAT for monitoring protein crystallization in lab-scale and is in the submission process. Chapter 3 describes a HT screening study and a kinetic study of enzyme crystallization in mixtures of three model proteins. Using UV/Vis spectroscopic measurements as a basis, chemometric PLS regression models are used to selectively quantify each model protein in the crystallization supernatant. In Chapter 4, three screening studies are presented investigating enzyme crystallization in mixtures of three model proteins, mAb precipitation in *Escherichia coli* (*E.coli*) lysate, and VLP precipitation in harvest cell culture fluid (HCCF). The chosen studies demonstrate the versatility of calibration-free MVDA to different modalities when UV/Vis spectroscopy and the multi-way chemometric model PARAFAC are coupled. Chapter 5 covers a PAT set-up to monitor enzyme crystallization out of clarified lysate in lab-scale using in-line Raman spectroscopy, on-line UV/Vis spectroscopy, various off-line analytics, and PLS regression modeling.

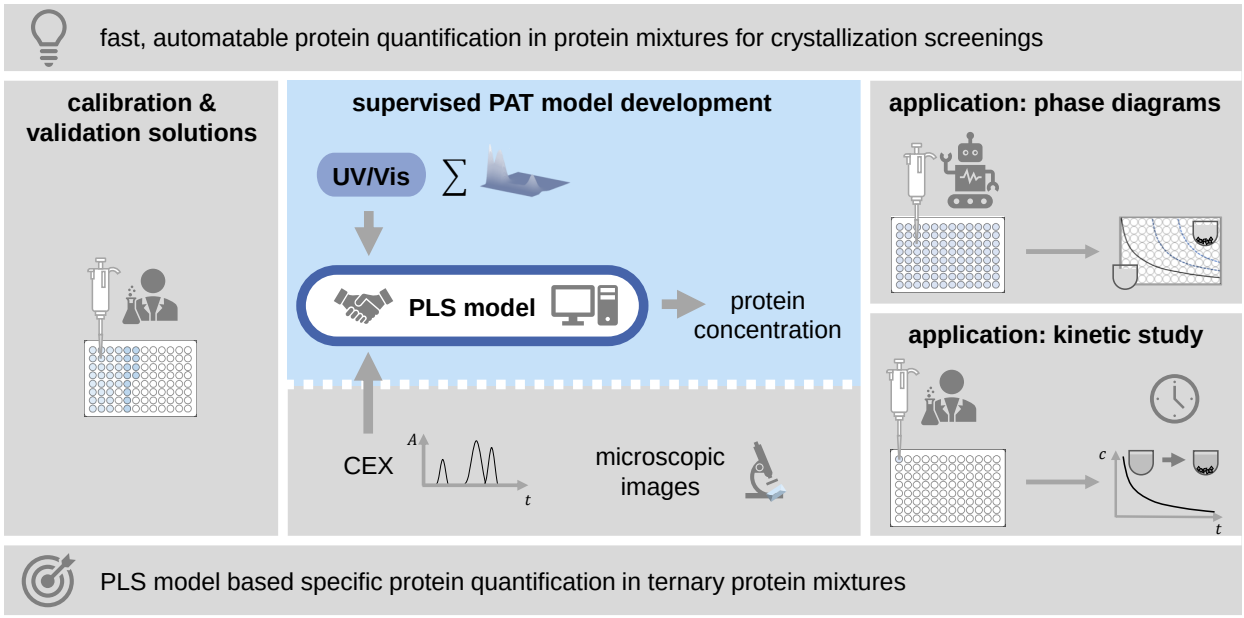
In the following, the prepared manuscripts are listed with their publication status, a graphical overview, a short summary, and the author contribution statement for each publication. The detailed lists of the author contributions were signed by the respective co-authors and are enclosed with the examination copy of this thesis.

Chapter 3

Rapid analysis for multi-component high-throughput crystallization screening: Combination of UV/Vis spectroscopy and chemometrics 29

Christina Henriette Wegner, Ines Zimmermann, and Jürgen Hubbuch

published in *Crystal Growth and Design*, Volume 22, 2022, p. 1054–1065
https://doi.org/10.1021/acs.cgd.1c00907



Protein crystallization is commonly used for protein structure analysis, but has gained more interest for an application in DSP due to its lower production costs, high product purity, and better scalability. To develop crystallization processes with the aim of protein purification, empirical HT screenings are commonly conducted. This calls for fast, quantitative, HT-compatible, and automatable analytics. This study introduces a novel, analytical workflow based on summed up UV/Vis spectra and supervised PLS regression models. These models, applied to HT crystallization screening supernatants, can predict the specific protein concentration of model protein mixtures containing lysozyme, ribonuclease A, and cytochrome C. Further, the provided information can be used to visualize the phase behavior in phase diagrams. Compared to established, quantitative analytics, the proposed method could quantify the model proteins with high precision and a 3 min analysis time per sample. Using cation-exchange chromatography (CEX) and microscopic images, the model-predicted protein concentrations and hence generated phase diagrams could be validated, respectively. To demonstrate the flexibility of the calculated models, a kinetic study was investigated on a 10 times larger scale. The described approach is a proof-of-concept proving that UV/Vis spectroscopy and chemometrics are a powerful combination when applied to phase behavior based screenings of heterogeneous protein mixtures.

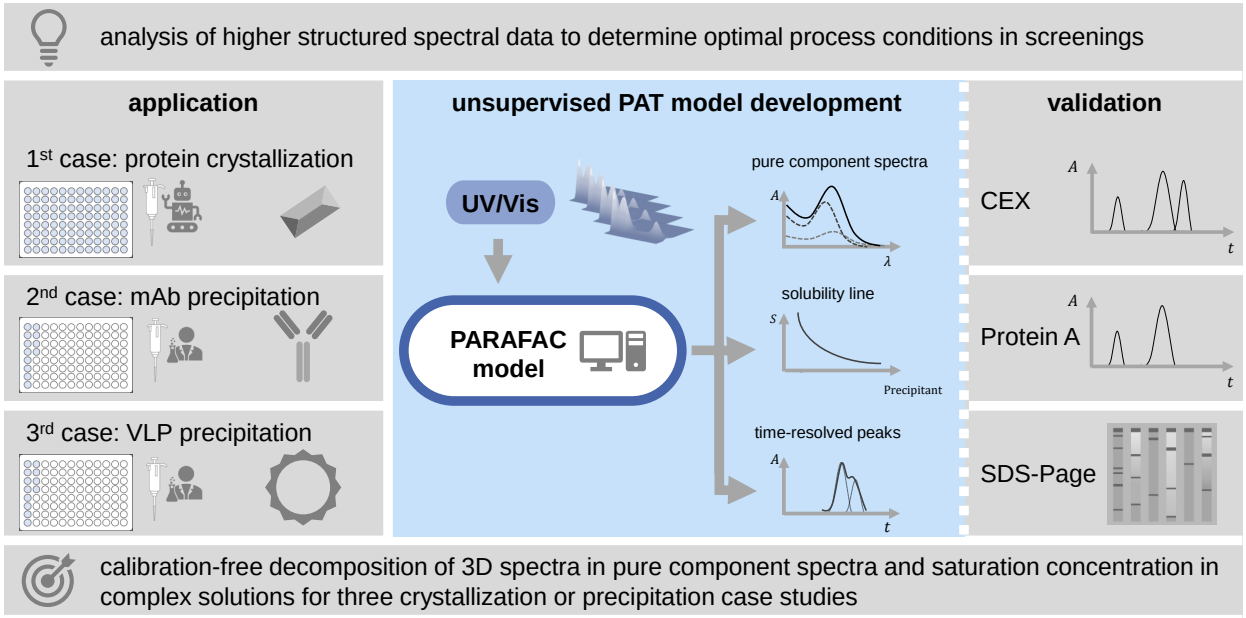
Author contributions:
Christina Henriette Wegner: conceptualization (initial idea, study design), methodology (experimental design and methods), investigation (literature review, experiments, and analysis), supervision (experiments), data curation (data preparation), formal analysis and validation (analysis and interpretation of data), visualization (figures and tables), writing (original draft, review, and editing), **Ines Zimmermann:** methodology (experimental methods), investigation (experiments and analysis), data curation (experimental preparation), writing (review), **Jürgen Hubbuch:** conceptualization (consultation), methodology (consultation), supervision, funding acquisition, writing (review)

Chapter 4

Calibration-free PAT: Locating selective crystallization or precipitation sweet spot in screenings with multi-way PARAFAC models 49

Christina Henriette Wegner, and Jürgen Hubbuch

published in *Frontiers in Bioengineering and Biotechnology*, Volume 10, 2022, p. 1-18
https://doi.org/10.3389/fbioe.2022.1051129



When HT screenings are conducted and evaluated using multi-variate analytics, e.g. UV/Vis spectroscopy over multiple wavelengths and time, the generated data can be structured in a multi-dimensional data set. Multi-way MVDA techniques are adapted to the higher data structure, can potentially make use of the multi-dimensional structure and support revealing underlying patterns. This study explores the application of the unsupervised, multi-way chemometric approach called PARAFAC on UV/Vis data generated in crystallization, or precipitation HT screenings. Three different biopharmaceutical modalities are either selectively crystallized or precipitated in chemically defined or complex solutions. In detail, one protein is crystallized in mixtures of three model proteins, and mAbs or VLPs are precipitated from HCCF or *E.coli* lysate, respectively. Without the need of prior calibration, one PARAFAC model per case study was constructed based on the UV/Vis spectra of supernatant samples during crystallization, precipitation, wash steps, or redissolution. The PARAFAC models could estimate the specific pure component spectra and specific concentration, which could identify the solubility line for optimal process conditions regarding yield and product purity. Finally, the models were validated either with spectra of purified species, with quantitative analytics, i.e. CEX or Protein A chromatography, or with qualitative analytics, i.e. sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE). The approach

proves effective regardless the modality, the contaminants, sample size, or the number of different species, and can be valuable in early-stage process development of phase behavior based processes, especially when robust analytics are missing. In summary, the conducted study provides a useful, analytical workflow for calibration-free PAT which supports the process design of selective crystallization, or precipitation processes of biopharmaceuticals.

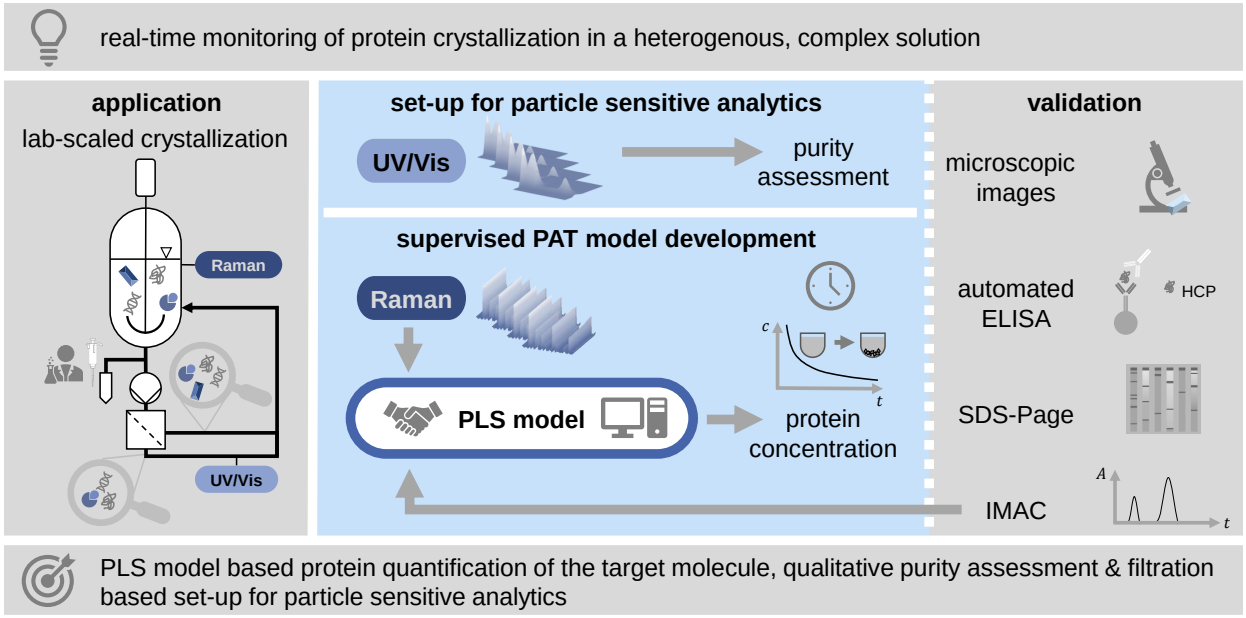
Author contributions:
Christina Henriette Wegner: conceptualization (initial idea, study design), methodology (experimental design and methods), investigation (literature review, experiments, and analysis), data curation (data preparation), formal analysis and validation (analysis and interpretation of data), visualization (figures and tables), writing (original draft, review, and editing), **Jürgen Hubbuch:** conceptualization (consultation), supervision, funding acquisition, writing (review)

Chapter 5

Spectroscopic insights into multi-phase protein crystallization in complex lysate using Raman spectroscopy and a particle-free bypass 75

Christina Henriette Wegner, Sebastian Mathis Eming, Brigitte Walla, Daniel Bischoff, Dirk Weuster-Botz, and Jürgen Hubbuch

submitted to *Frontiers in Bioengineering and Biotechnology*



Regarding DSP capture steps, new challenges arise when PAT needs to be implemented in protein crystallization processes as solid particles can interfere with measurements. The heterogeneity of multiple phases and the impurities in the initial process solution demand careful considerations for the selection of sensors and sampling techniques. To overcome these limitations, this research project applies two different spectroscopic methods aiming to monitor and improve the understanding of the crystallization process of *Lactobacillus kefir* alcohol dehydrogenase (*Lk*ADH) from clarified *E.coli* lysate on a 300 mL scale. The study employed a combination of in-line Raman spectroscopy with a probe placed *in-situ* in the crystallization vessel, on-line UV/Vis spectroscopy in a bypass, and off-line analytics (microscopic images, automated enzyme-linked immunosorbent assay (ELISA), SDS-PAGE, immobilized metal ion affinity chromatography (IMAC)) and should provide a comprehensive overview of the conducted crystallization experiments. The experimental set-up using a cross-flow filtration based bypass allowed the liquid phase analysis with particle sensitive analytics, e.g. UV/Vis spectroscopy which could evaluate the purity of the crystallization supernatant. Chemometric analysis of the Raman spectra with principal component analysis (PCA) and PLS regression enabled the quantification of the target molecule concentration in real-time, even in the presence of solid crystals, and impurities in soluble or precipitated form. Due to

potential batch-to-batch variations, the favored PAT sensor consisting of a Raman probe and a calibrated PLS model could only be transferred to new experiments with some reservations. To sum it up, a PAT set-up, tailored to protein crystallization, was developed with the aim to quantify the target during crystallization in a complex solution.

Author contributions:
Christina Henriette Wegner: conceptualization (initial idea, study design), methodology (experimental design and methods), investigation (literature review, experiments, and analysis), supervision (experiments), data curation (data preparation), formal analysis and validation (analysis and interpretation of data), visualization (figures and tables), writing (original draft, review, and editing), **Sebastian Mathis Eming:** methodology (experimental design and methods), investigation (experiments and analysis), writing (review), **Brigitte Walla:** methodology (consultation), investigation (provision of material), writing (review), **Daniel Bischoff:** methodology (consultation), data curation (image data preparation), formal analysis and validation (image analysis and its interpretation), writing (review), **Dirk Weuster-Botz:** supervision (BW & DB), funding acquisition, writing (review), **Jürgen Hubbuch:** conceptualization (consultation), supervision, funding acquisition, writing (review)

3

Rapid analysis for multi-component high-throughput crystallization screening: Combination of UV/Vis spectroscopy and chemometrics

Christina Henriette Wegner[1], Ines Zimmermann[2], and Jürgen Hubbuch[1]

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

[2] Chair of Bioseparation Engineering Group, Department of Mechanical Engineering, Technical University of Munich, Germany

Abstract

Selective protein crystallization is a trending alternative to preparative chromatography in biotechnological downstream processing. To save time and resources in early-stage process development, fast and reliable analytics are required. This work aimed to develop and assess a low-volume, quantitative, analytical tool for faster development of crystallization processes. The analytical tool was based on ultraviolet-visible spectroscopy and partial least squares modeling and aimed to selectively quantify protein concentrations in heterogeneous supernatants during crystallization process development. For this purpose, a ternary model protein system consisting of hen-egg-white Lysozyme, bovine Ribonuclease A, and equine Cytochrome C was used for model calibration and subsequent crystallization studies for application. In a high-throughput screening, Lysozyme was selectively crystallized varying pH, precipitant concentration, and Lysozyme concentration at 8 °C for 13 d. During a kinetic study, the composition of two selected conditions was monitored over a time range of 7 d. In both studies, the developed tool quantified the different species in the supernatant with high precision. Crystal yield, purity, and selectivity were evaluated with a sensitivity of 96.23 % and a short analysis time of 3 min per sample. The studies were carried out in 96-well plates. This said, the methodology could be easily adapted to higher throughput scales, i.e., 384-well or 1536-well plates.

3.1 Introduction

In the last decades, selective crystallization of biopharmaceutical products has gained increasing attention as a cost-effective, alternative downstream process step to chromatography-based approaches [14, 124, 125]. High purity [15, 126], preservation of activity [127, 128], ease of scale-up [14], and high target protein concentration are key advantages. Furthermore, it complies with the trend towards higher production titers, process intensification and process integration [125], and crystallized products demonstrate preferable formulation properties [129]. Its relatively low viscosity despite the high concentration [130], high stability [131, 132], and controlled drug release [133] make biopharmaceutical process crystallization a promising research field.

Despite previous work on protein crystallization of industrially relevant proteins, e.g., antibodies [126, 127, 134], antibody fragments [14], or enzymes [30, 131], most fundamental research on protein crystallization was carried out with the model protein lysozyme (Lys) extracted from poultry egg white due to its availability and good economics [2, 4, 9, 15, 135–137]. Many crystallization studies of Lys focused on determining morphologies of crystals from pure protein solutions at low supersaturation [15, 27, 135–137] accepting long process time to reproducibly grow large crystals. To overcome these limitations and make protein crystallization feasible for industrially relevant downstream processes, the saturation concentration of the target protein was varied, e.g., by varying the temperature [6, 8, 26, 138], pressure [139], or precipitant concentration [26, 27, 31].

Occurrences during protein crystallization are commonly investigated in empirical phase diagrams inhigh-throughput (HT) screenings. Applied analytics are image-based analytical tools [37, 140], single crystal X-ray diffraction [6, 140, 141], or dynamic light scattering (DLS) measurements [142–144]. These analytics are optimized for pure solutions or crystals, do not allow fast differentiation between different species, and can only distinguish between salt and protein crystals to some extent. When protein crystals are grown in protein mixtures or complex harvest broth, a fast, accurate, and reliable analytical method is essential for screening purposes. The combination of ultraviolet-visible light (UV/Vis) spectroscopy and chemometrics serves as a reliable tool to selectively quantify light absorbing species. In previous work, acuvvis spectroscopy paired with partial least squares (PLS) regression modeling demonstrated high prediction performance for selective quantification in preparative chromatography-based processes [53, 55–57].

Various research was performed to monitor and control preparative crystallization applying focused beam reflectance measurement (FBRM) [86, 138, 145], and in-line spectroscopy, i.e., attenuated total reflection (ATR) UV/Vis spectroscopy [86, 121, 138, 145, 146], or Raman [86, 146, 147]. In some of these studies the in-line analytics were paired with chemometric approaches, e.g., principal component regression (PCR) [147], PLS [147], or principal component artificial neural networks (PC-ANN) [145]. Chemometric approaches intend to extract valuable information out of a large data set and can be further studied here [39, 43].

However, these crystallization studies mostly focused on monitoring and control of preparative crystallization processes of small, chemical molecules and cannot be directly transferred to crystallization processes of biopharmaceuticals as biological molecules are heterogeneous, more complex, and prone to process deviations and lot-to-lot variation of raw materials [117]. Thus, empirical screenings and robust analytics are still substantial for biopharmaceutical crystallization process development. To the best of our knowledge, there have been no attempts to quantify protein species selectively using UV/Vis spectroscopy and PLS models for HT crystallization screenings.

This study is designed to develop a new and accurate analytical tool to speed up process development for HT selective protein crystallization screenings. The method aims to selectively quantify individual species in the supernatant allowing calculation of process performance indicators in early stage process development. Due to limited resources and the large number of screening conditions, the method is designed to require minimal product intake and analysis time. In the chosen ternary protein system, one protein - purely for the purpose of this study - was specified as the target protein in the crystallization process and the other two as impurities. To selectively quantify individual species during a HT screening, one PLS model per protein was calibrated and used for various crystallization conditions. The required analyses were performed on an ultra high performance liquid chromatography (UHPLC) system equipped with a diode array detector (DAD) to record the UV/Vis spectra. Exemplarily, the influences of environmental conditions, i.e., target protein and precipitant concentration, as well as pH, were screened and evaluated regarding yield, selectivity, and purity. Finally, two screening conditions were further analyzed over time to demonstrate the suitability of the developed technology to gather more information on crystallization kinetics. The presented results and

data visualization aid knowledge-based crystallization process design and stress the broad applicability of the developed method in early stage process development.

3.2 Materials and methods

3.2.1 Proteins and buffer preparation

All chemicals were purchased from Merck KGaA (Darmstadt, DE), unless stated otherwise. All buffer solutions were prepared at room temperature with ultrapure water (PURELAB Ultra, ELGA LabWater, Lane End, High Wycombe, UK). The pH was adjusted with 4 M NaOH or 32 %(w/w) HCl using a pH electrode (SenTix® 62, Xylem Analytics Germany Sales GmbH & Co. KG, Weilheim, DE) at a pH bench meter (HI 3220, Hanna Instruments, Woonsocket, RI-US). Finally, the buffers were filtered using a 0.2 µm CA membrane filter (Sartorius Stedim Biotech GmbH, Göttingen, DE).

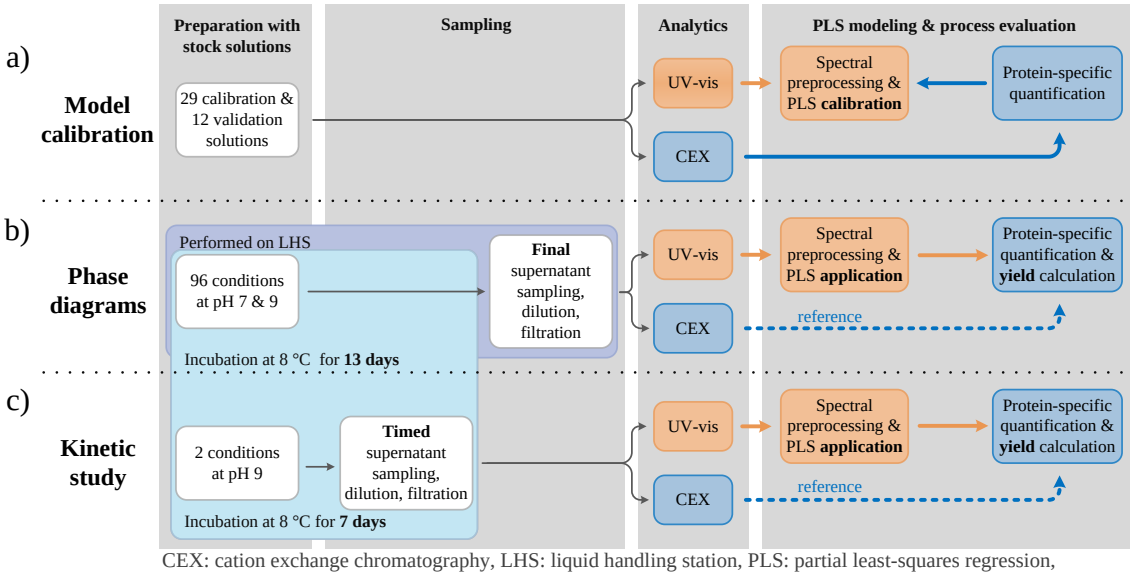
The lyophilized proteins Lys from chicken egg white (Hampton Research, Aliso Viejo, CA-US), ribonuclease A (RibA) from bovine pancreas and cytochrome C (CytC) from equine heart were each dissolved in 2 mL multi-component buffer (MCB: 21 mM N-1,1-dimethyl-2-hydroxyethyl-3-amino 2-hydroxypropanesulfonic acid (AMPSO), 17 mM 3-N-morpholino propansulfonic acid (MOPS, Carl Roth GmbH + Co. KG, Karlsruhe, DE), 15 mM succinate acid (AppliChem GmbH, Darmstadt, DE)) at pH 7 or 9. After dialysis (17 kDa Slide-A-Lyzer™, Thermo Fisher Scientific Inc., Waltham, MA-US) to a MCB at the target pH according to the manufacturer's specification, the protein solution concentrations were adjusted to the required stock solution concentrations to an accuracy of 5 %. Hereby, experimentally determined extinction coefficients at a wavelength of 280 nm and a NanoDrop™ 2000 spectrophotometer (Thermo Fisher Scientific Inc) were used. Before preparation of the PLS calibration solutions and phase diagrams, the protein and buffer stock solutions were filtered with 0.2 µm (Pall Corporation, Port Washington, NY-US) and 0.02 µm syringe filters (Cytiva, Marlborough, MA-US), respectively.

3.2.2 PLS modelling and data processing

This subchapter deals with the selection and preparation of the PLS calibration solutions, spectral preprocessing, PLS model regression, and the calculation of crystallization process performance indicators. All analytics and data collection of the samples are described in subchapter 3.2.4. 29 calibration solutions were selected according to a full factorial design with three factors for the three studied proteins on three concentration levels and twelve validation solutions according to the protein concentrations of the star points of a central-composite-circumscribed and a central-composite-inscribed design (distance factor between center of design space and star point $\alpha = +\sqrt{3}$) [148]. The calibration range (Lys: 0 to 1.5 mg/mL mg/mL, RibA and CytC: 0 to 0.2 mg/mL) was adjusted to the assumed concentration ranges of the diluted phase diagram supernatants. The validation solution concentrations were partly set outside the calibrated range (Lys: 0.317 to 2.049 mg/mL, RibA and CytC:

0.042 to 0.273 mg/mL) to improve the model prediction near the calibration limits [88]. For this purpose, protein stock solutions ($c_{\mathrm{Lys}} = 4.5\,\mathrm{mg/mL}$, $c_{\mathrm{RibA}} = c_{\mathrm{CytC}} = 0.6\,\mathrm{mg/mL}$) were prepared at pH 7 and the calibration and validation solutions were manually mixed.

The recorded UV/Vis spectral data of the samples were background subtracted, summed up along the time axis to impede diffusion effects in the spectra, cut to the required wavelength range according to Table 3.1 and treated with a Savitzky Golay (SG) [149] in the case of RibA (2^{nd} derivative, window of 7 data points). The preprocessed data were correlated to protein concentrations calculated from the reference analytics cation-exchange chromatography (CEX) with PLS models and then validated with an external validation data set (see Figure 3.1 a)). Data analysis, model calibration, validation and application were performed in MATLAB,



CEX: cation exchange chromatography, LHS: liquid handling station, PLS: partial least-squares regression, UV-vis: ultraviolet-visible spectroscopy

Figure 3.1 PLS models were calibrated with the calibration solutions by regressing the protein-specific concentration of the reference analytics CEX on the preprocessed UV/Vis spectra (a). In the second and third step, the models were used to determine the protein concentration of each species in the diluted supernatant of selective crystallization studies (b-c). The model predicted concentrations were additionally validated with the reference CEX method.

R2019b (The MathWorks, Inc., Natick, MA-US). For the application of the PLS models (see subchapter 3.2.3), the spectral data and reference data were preprocessed analogously (see Figure 3.1 b) and c)).

The relative protein-specific concentrations $\hat{c}_{\mathrm{PLS},i,j}$ in the phase diagrams were calculated solely from the PLS model predicted concentrations $c_{\mathrm{PLS},i,j}$ in the supernatant for each protein species i and each well j in 3.1. The stable conditions in the phase diagram, showing no phase

transition, in each row were used to describe the well-specific decline of protein concentration in the supernatant and calculate the mean protein concentration of stable conditions per row $\overline{c}_{\mathrm{PLS},i,\mathrm{stable}}$ (see A3.1). The protein-specific target concentration of each condition (see Figure 3.2) was not used for normalization as the actual concentration was largely affected by measurement and pipetting errors during protein stock solution preparation, pipetting of the crystallization batches, and preparation and analysis of the samples.

$$\hat{c}_{\mathrm{PLS},i,j} = \frac{c_{\mathrm{PLS},i,j}}{\overline{c}_{\mathrm{PLS},i,\mathrm{stable}}} \tag{3.1}$$

The protein-specific yield $Y_{i,j}$ for each well could be calculated assuming that the missing protein amount formed crystals. This aided the identification of successful phase transition – both crystallization and precipitation.

$$Y_{i,j} = 1 - \hat{c}_{\mathrm{PLS},i,j} \tag{3.2}$$

Additionally, the model derived purity of produced Lys crystals $P_{\mathrm{Lys},j}$ was calculated from the ratio of the missing Lys concentration to the total missing concentration in the supernatant for each well.

$$P_{\mathrm{Lys},j} = \frac{\overline{c}_{\mathrm{PLS,Lys,stable}} - c_{\mathrm{PLS,Lys},j}}{\sum_{i=1}^{3}(\overline{c}_{\mathrm{PLS},i,\mathrm{stable}} - c_{\mathrm{PLS},i,j})} \tag{3.3}$$

3.2.3 Crystallization experiments

For faster, quantitative assessment of crystallization screening conditions and optimal process time, the PLS models were applied to diluted supernatants of phase diagrams and of a kinetic study. The diluted supernatants were analyzed and preprocessed according to subchapter 3.2.4 and 3.2.2, respectively, enabling elaborate crystallization yield and purity estimation.

For the phase diagrams, the protein stock concentrations were 180 mg/mL for Lys and 48 mg/mL for RibA and CytC at pH 7 or 9. In addition to the MCB, a crystallizing solution at target pH was required, which was composed of the MCB compounds and 3.5 M AMS (ammonium sulfate, AppliChem GmbH, Darmstadt, DE). In duplicates, 96 conditions were prepared in 24 µL batches varying the ammonium sulfate (AMS) concentration in column (1-12) and target protein concentration Lys in row (A-H) in MCR Under Oil Crystallization Plates (Hampton Research, Aliso Viejo, CA-US). With a liquid handling station (Tecan Freedom Evo 100, Tecan, Männedorf, CH), the ternary protein mix and the precipitant dilutions were prepared with three protein stock and two buffer stock solutions, respectively (see Figure 3.2). Finally, the screening conditions were prepared by mixing 20 µL of the ternary protein mix dilutions and 4 µL of the precipitant dilutions and the plates were sealed with a transparent foil (HDclear, ShurTech Brands, Avon, US). During 13 d at 8 °C in a cooled incubation system (RI 54, FORMULATRIX, Bedford, MA-US; T 1000 mytron Bio- und Solartechnik GmbH, Heiligenstadt, DE), the phase diagrams were automatically imaged. After incubation, only one of the phase diagrams was further analyzed due to limited time resources. The supernatants of the screened conditions were 50 times diluted with MCB (pH

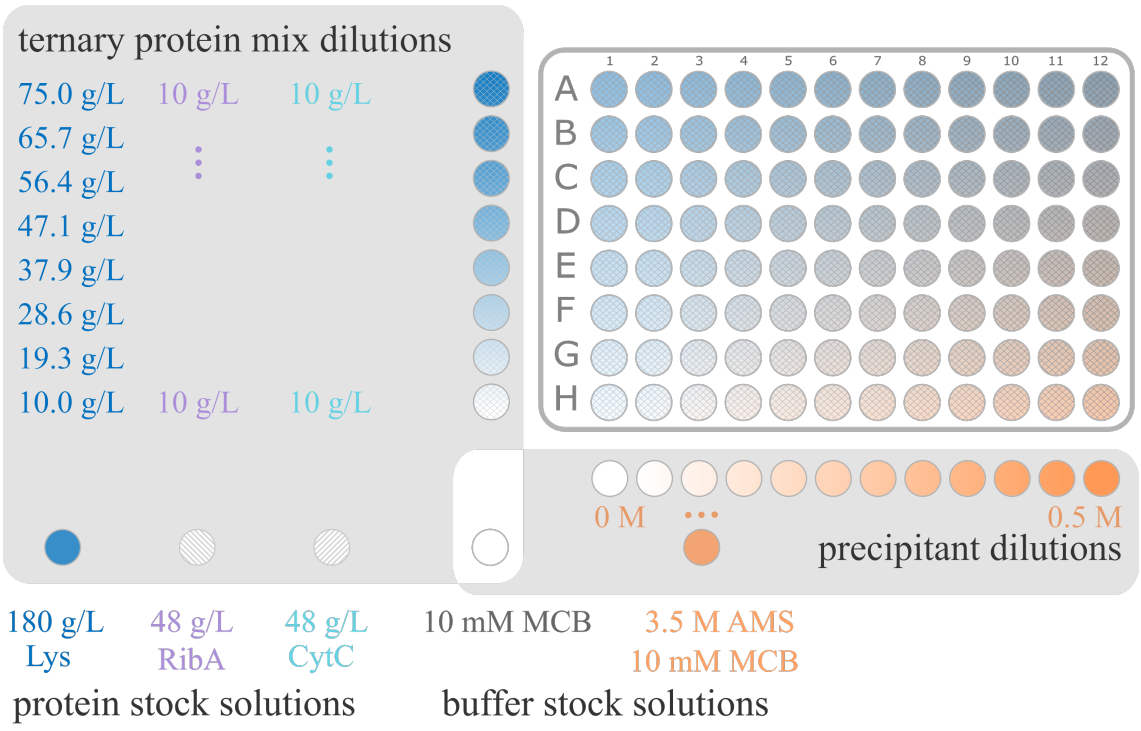


Figure 3.2 Crystallization plates were prepared by mixing eight ternary protein dilutions with three protein stock solutions and multi-component buffer (MCB) at target pH. Twelve precipitant dilutions varied in AMS concentration by mixing MCB and AMS stock solution. The protein and AMS dilutions were combined to 96 conditions in 24 µL micro-batches and stored at 8 °C for 13 d.

of the plate) and filtered in 0.2 µm filter plates (Pall Corporation). The dilution was required to impede further crystallization.

For the kinetic study, two conditions of the phase diagram at pH 9 were selected (A5: $c_{\mathrm{Lys}} = 75\,\mathrm{mg/mL}$, $c_{\mathrm{RibA}} = c_{\mathrm{CytC}} = 10\,\mathrm{mg/mL}$, $c_{\mathrm{AMS}} = 0.1364\,\mathrm{M}$; D6: $c_{\mathrm{Lys}} = 47.14\,\mathrm{mg/mL}$, $c_{\mathrm{RibA}} = c_{\mathrm{CytC}} = 10\,\mathrm{mg/mL}$, $c_{\mathrm{AMS}} = 0.2273\,\mathrm{M}$;) and additionally prepared in 300 µL crystallization batches in duplicates. During incubation for 7 d at 8 °C, the crystallization batches were covered with a semi-transparent film (Parafilm, Bemis company, Inc., Neenah, WI-US). The supernatant samples were taken manually and processed analogously to the phase diagram samples.

3.2.4 Analytics

For the PLS model calibration and application, UV/Vis spectra were recorded and CEX was performed as a reference. In detail, 20 µL samples of defined ternary protein solutions or diluted supernatant were analyzed for the model calibration or the crystallization studies, respectively.

First, the samples were analyzed using a Dionex Ultimate 3000 RS UHPLC system (Thermo Fisher Scientific Inc.), equipped with a diode array detector. The spectral analysis was performed with a pre-column filter cartridge (0.5 µm OPTI-SOLV EXP, Supelco, Bellefonte, PA-US) in the mobile phase (20 mM Tris, 100 mM NaCl, pH 8.0) but no chromatography column installed. The UV/Vis absorbance spectra between 240 to 450 nm were recorded with 1 nm resolution and 100 Hz frequency (see A3.2 b).

Secondly, as a reference, the same samples were CEX analyzed using a ProSwift SCX-1S 4.6x50mm column (Thermo Fisher Scientific Inc.) and the same UHPLC system with a low salt buffer (20 mM Tris, pH 8.0) and a high salt buffer (20 mM, 1000 mM NaCl, pH 8). The column was loaded with 20 µL sample and eluted with a gradient method (see A3.2 and A3.2 a). The flow rate was 1.5 mL/min and the recorded absorbance at the wavelength 280 nmand at 100 Hz frequency was used for further calculation.

Images taken during incubation served as a validation method for crystallization and phase behavior detection.

3.3 Results and discussion

HT screenings are crucial for selective crystallization process development and require fast and reliable analytics. This research project focuses on the application of PLS models in combination with UV/Vis spectroscopy to quickly quantify individual species in a multi-component matrix during HT screenings.

The model accuracy must not be affected by temperature or varying aqueous conditions such as precipitant concentration or pH. This was tested in an extended HT screening and a kinetic study to show the versatility of the analytical tool.

3.3.1 Data analysis and model accuracy

To selectively quantify the different species in a multi-component mixture during batch crystallization, spectral data preprocessing and model accuracy were evaluated.

When comparing the absorbance values integrated over time at a wavelength of 280 nm of the spectral and the reference analysis (CEX analysis), increased absorbance areas were observed for the latter, especially for the outer boundary wells of the 96 well plate. The observed difference could be traced back to sample evaporation caused by an extended sample storage time prior to analysis. Thus, the CEX analysis required concentration correction with the well-specific correction factor f_j for each well j.

$$f_j = \frac{\int_{t_0}^{t_{\text{end}}} A_{280,\text{DAD}} dt}{\sum_{i=1}^{3} \int_{t_{0,i}}^{t_{\text{end},i}} A_{280,\text{CEX}} dt} \tag{3.4}$$

The absorbance at 280 nm of the spectral ($A_{280,\text{DAD}}$) and the reference measurement ($A_{280,\text{CEX}}$) was integrated from the analysis start t_0 to the end time t_{end} for each species i. Assuming the same relative concentration change for each species due to evaporation, the time-wise

integrated absorbance areas were used to calculate the corrected, specific concentrations $c_{\mathrm{CEX,corr},i}$ from the original specific concentration $c_{\mathrm{CEX},i}$ derived from the CEX analytics.

$$c_{\mathrm{CEX,corr},i} = f_j * c_{\mathrm{CEX},i} \tag{3.5}$$

This data preprocessing improved model prediction and allowed neglecting differing sample storage times between spectral DAD and CEX analytics when an increased number of samples were analyzed for the HT screening. Further information on preprocessing of the spectral data and the calibrated PLS models are listed in Table 3.1. The model calibration was evaluated with the parameters root mean squared error of cross-validation ($RMSECV$) and the coefficient of determination (R^2); the model validation with the root mean squared error of prediction ($RMSEP$) and predictive relevance (Q^2).

The chosen wavelength range differed for each PLS model as the examined proteins showed individual absorption behavior in the spectra. Larger wavelength ranges increase the risk of overfitting [150] but demonstrated the highest model accuracy, regarding the $RMSECV$ and Q^2, in the case of RibA (data not shown).

Table 3.1 Parameters for preprocessing of the UV/Vis spectral data and PLS model calibration.

	Lys	RibA	CytC
wavelength range / nm	240 - 300	250 - 430	385 - 425
latent variables	4	4	1
R^2	1.0000	0.9949	0.9972
$RMSECV$ / mg/mL	0.0019	0.0045	0.0038
Q^2	0.9999	0.9955	0.9956
$RMSEP$ / mg/mL	0.0048	0.0037	0.0040

Spectral derivation has been widely reported to enhance spectral differences [151–153] and the SG is a simple and common preprocessing technique for smoothing and derivation [88] and facilitates PLS model calibration for challenging protein systems, e.g. when the different species display similar spectra [56]. Using this filter, the calculation of the 2$^{\mathrm{nd}}$ spectral derivative enabled a robust RibA model as the absorption spectra of all investigated proteins overlay in the wavelength range chosen for the RibA model. Due to this spectral overlay, four latent variables for the model calibration of Lys and RibA were necessary when compared to CytC with only one latent variable.

Models with more latent variables include more spectral information which can aid identifying spectral differences related to protein-specific absorption but bears the risk of overfitting. Compared to previous studies conducted with the same model protein system, the number of latent variables could be reduced due to different concentration ranges and ratios, the implementation of individual preprocessing strategies, and the usage of reduced

wavelength ranges [53, 55]. However, a direct comparison of these studies is complicated and not in the focus of the present study.

Similar and high values for the coefficient of determination R^2 and predictive relevance Q^2 indicated low residuals in the calibration as well as in the validation data set and therefore good predictive behavior for all three models. In the case of RibA and CytC the $RMSECV$ and $RMSEP$ showed minor differences. The Lys model showed a twice as high $RMSEP$ value compared to its $RMSECV$. This is acceptable as the used validation solutions were partly out of the calibrated range. An explanation for the higher $RMSECV$ and $RMSEP$ for the Lys model lies in the chosen protein concentration range for the calibration system. The upper concentration limit of Lys was chosen more than 7 times higher than for the contaminant species.

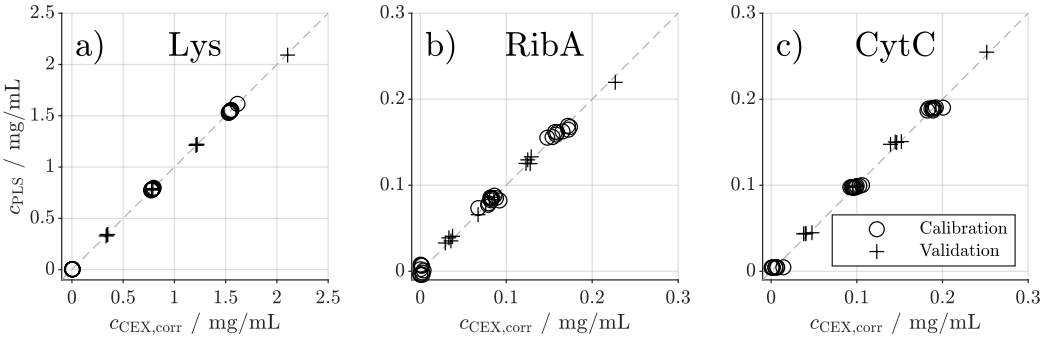


Figure 3.3 The PLS model predicted concentrations c_{PLS} for each species is shown over $c_{CEX,corr}$, measured off-line by CEX and corrected according to Eq. 3.5. The circles and crosses represent the calibration and validation data set, respectively (see Figure 3.1 a)). The dashed lines represent the ideal relationship between the predicted and reference measurements.

Figure 3.3 shows the concentration prediction performance of the calibrated PLS models in relation to the data obtained from the reference analytics CEX, corrected using Eq. 3.5. The PLS predicted values of the validation data set show a very good agreement with the reference data. The model of RibA showed the largest discrepancies, probably due to the overlapping absorption areas of Lys, CytC and RibA.

In Figure 3.4, the prediction performance of the models is displayed when applying it to the diluted supernatants in the HT screening and the kinetic study. The PLS models performed well as shown by the good agreement between the predicted values and the reference. The CytC model application on the HT screening showed $RMSEP$s comparable to the validation set. The Lys and RibA models showed 1.8 to 4.9 times higher $RMSEP$s and lower accuracy in the case of RibA at pH 7 (see Figure 3.4 b)). These discrepancies could be caused by minor differences in the protein spectrum due to changes of the protein's tertiary structure during the HT screening. Structural changes could be caused by incubation over 13 d in the presence of the precipitant AMS [36] at lower temperatures [154] as this can induce changes in the aqueous micro-environment of the protein. As opposed to that, the

calibration solutions were analyzed directly after preparation. In particular, the RibA model is expected to be prone to subtle changes in the spectrum due to the 2nd derivative data preprocessing [155].

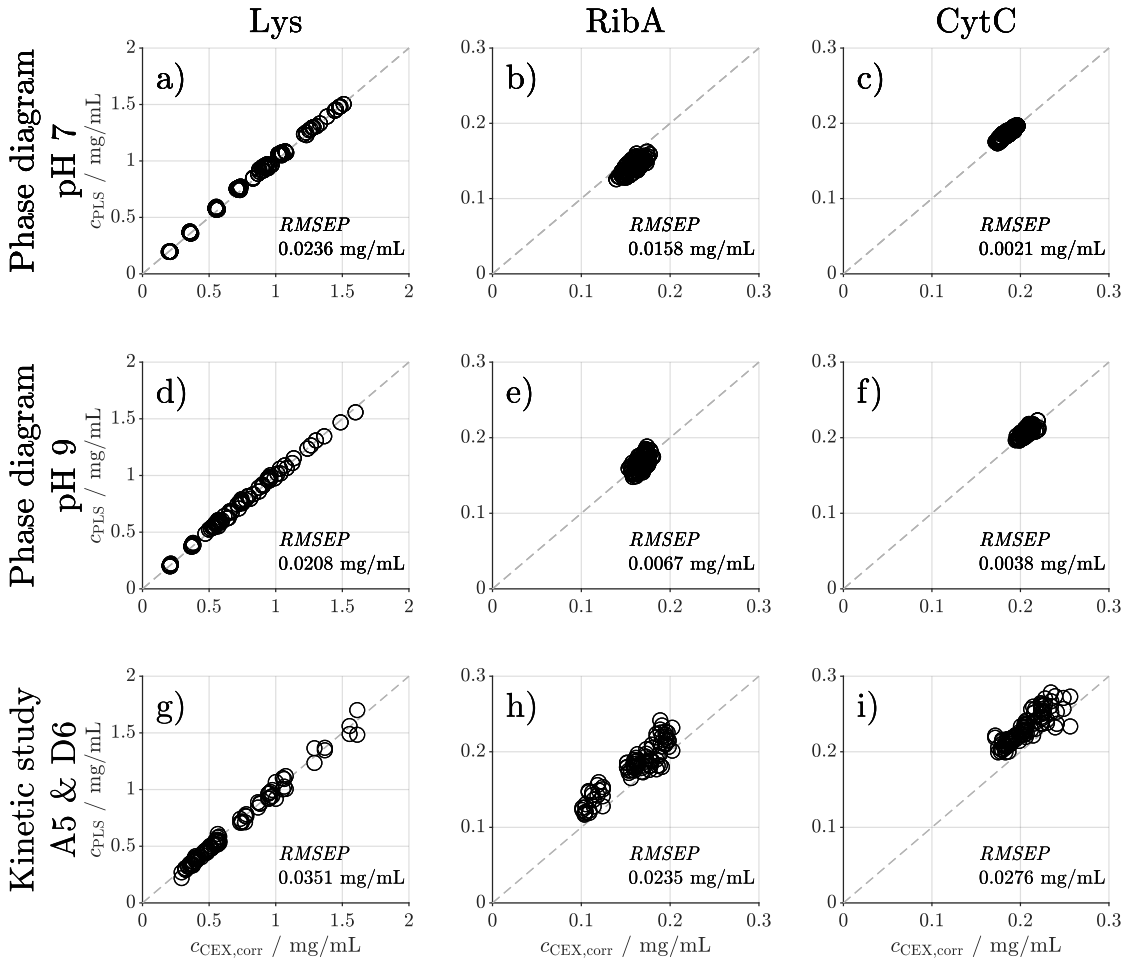


Figure 3.4 The calibrated PLS models were applied on two phase diagram studies at two pH levels ((a-c) pH 7; (d-f) pH 9) and a kinetic study of two conditions at pH 9 (g-i). The model predicted concentrations c_{PLS} for each species are shown over its corrected reference concentration $c_{\mathrm{CEX,corr}}$, measured off-line by CEX and corrected according to Eq. 3.5. The dashed line represents the ideal relationship between predicted and reference measurements.

The data of the kinetic study displayed 6.4 to 7.3 times larger $RMSEP$s than during model validation for all species. Especially the contaminating species indicated low model accuracy. However, this could be traced back to a declining performance of the CEX column used as the reference analytics. Despite the dynamic process of crystal nucleation and growth, the Lys model enabled accurate quantification.

Regarding the model application, the Lys model stood out with the highest $RMSEP$s and concentration data points over the whole concentration range. This could be explained with the higher average concentration of this species in the phase diagrams. Comparing the models for the contaminants RibA and CytC, both PLS models displayed similar performance during the HT screening and kinetic study.

To sum it up, the PLS models could be applied to a HT screening investigating the pH and AMS concentration in a phase diagram. Additionally, a kinetic study with timed sampling allowed for observation of the crystallization kinetics. It is assumed that the dilution of the supernatant shifts the solution composition below the saturation concentration of each component and thus stops the crystallization process immediately. This ensures timed snapshots of the supernatant composition.

During the application, the incubation conditions, i.e., varying precipitant concentration, pH, and temperature, lowered the PLS model performance slightly due to possibly induced conformational changes of the examined proteins during incubation.

3.3.2 Crystallization process parameters

In the following, the analytical tool is applied to a simple crystallization process screening using Lys as the product molecule and RibA and CytC as the contaminating species. The crystallization process and crystal purity can be visualized in phase diagrams offering a deeper understanding of the phase behavior of multi-component mixtures at various conditions.

Figure 3.5 displays the protein-specific phase diagrams at two pH levels, revealing areas of phase transition. The circle areas provide information on the crystallization yield calculated according to Eq. 3.2 (see A3.1 for details on yield calculation). By this, protein-specific successful phase transition and stable undersaturated conditions could be quickly distinguished. Throughout the HT screening, images were taken and served as a validation for successful phase transition out of the supersaturated liquid. This said, due to condensation on the sealing tape, 4 conditions could not be visually analyzed; furthermore, in 9 wells particle structures were formed where a discrimination between precipitate and micro-crystals was not possible due to the resolution of the camera. These conditions were located at the boundary of the crystallization window. However, as mostly crystallization was observed, it is expected that nucleation and crystal growth were the driving forces of the phase transition in all wells.

Regarding Lys behavior in Figure 3.5 a) and d), a window of selective crystallization was visible at lower salt concentration. The crystallization tendency, however, decreased with decreasing Lys start concentration and increasing salt concentration. Comparing the investigated pH levels, the higher pH led to a larger crystallization window with Lys crystal yields up to 64.6 % at pH 9 and 36.6 % at pH 7.

Note, that the conditions in the Lys crystallization window showed column-wise similar final concentration, indicated by the colors in Figure 3.5 a) and d). This observation was a consequence of the precipitant concentration dependent saturation concentration. In supersaturated liquids, the protein-specific supernatant concentrations decrease to their saturation concentration undergoing phase transitions, i.e., crystallization, and thus demonstrating the solubility curve depending on the precipitant concentration [2, 3, 9].

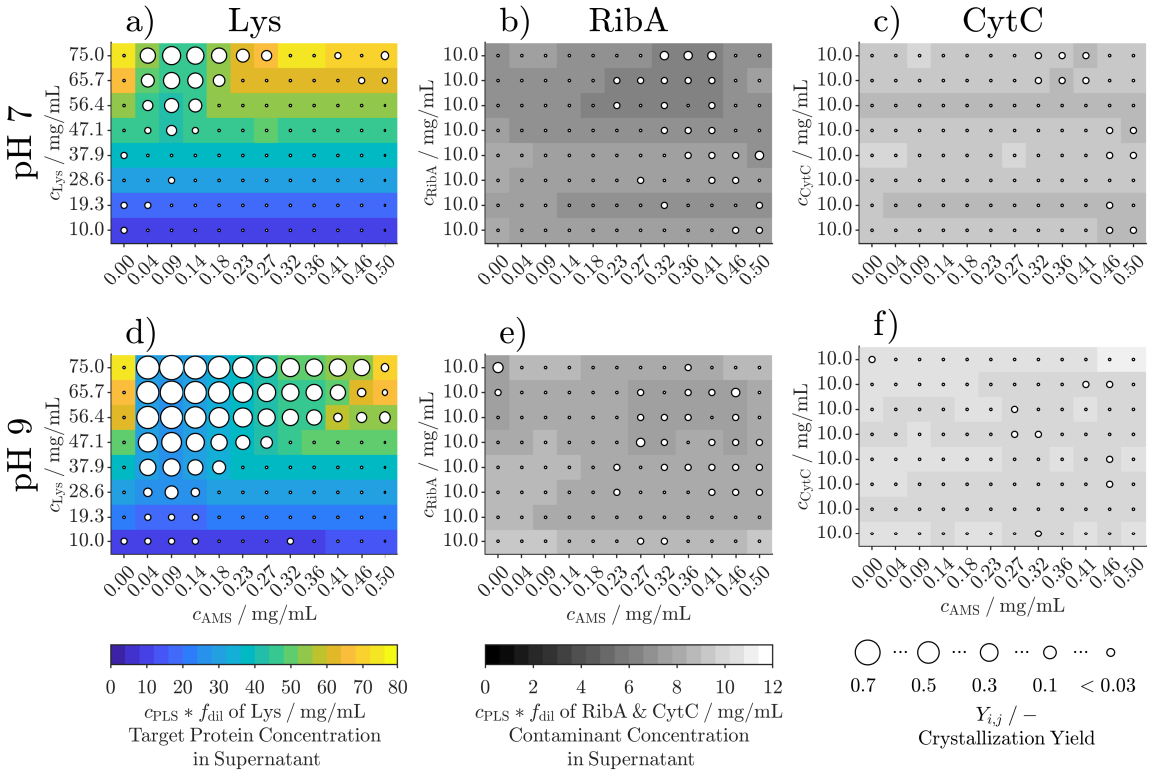


Figure 3.5 The subfigures show the protein-specific phase behavior, e.g., crystallization windows, obtained from the examined plates at pH 7 (a-c) and at pH 9 (d-f). The protein and precipitant concentrations in each well were arranged as illustrated in Figure 3.2. Each circle represents an examined condition and its position in the phase diagram. The area of the circle illustrates protein- and well-specific yield $Y_{i,j}$ after 13 d of incubation, calculated according to Eq. 3.2. The circle background is colored according to the protein-specific final supernatant concentration, calculated from the PLS predicted protein concentration and the dilution factor f_{dil}. The color ranges represent individual concentration ranges.

In case of the contaminant species RibA (Figure 3.5 b) and e)) and CytC (Figure 3.5 c) and f)), the phase diagrams showed no structured area of phase transition. Isolated, smaller circles indicated local protein-specific concentration declines. As the larger circles were not positioned in the crystallization window of Lys, it is assumed that no systematic integration of contaminants took place during the crystallization process of the target protein Lys. Supposably, the scattered circles indicating a relative decline in protein concentration were artefacts of dilution inaccuracy, analytics or model uncertainty. Regarding CytC, the highest, relative concentration drop, compared to the initial concentration, was 4.10 %.

The more prominent, scattered distribution of larger circles (up to 11.12 % concentration drop) in the RibA phase diagram lay outside the Lys crystallization window and was located at higher Lys and higher precipitant concentration. They were presumably artefacts of the lower model accuracy (see Figure 3.4 b) and e)), especially, when large quantities of Lys were present in the supernatant. The spectral measurement of a biological replicate and the reference CEX analytics did not show the same decrease in protein concentration in the supernatant (data not shown). Thus, the exceptional values were either caused by experimental errors during the measurement or model inaccuracy of single samples. However, it has to be noted that, compared to Lys, in the case of RibA and CytC measurement inaccuracies and pipetting errors had a larger impact on the yield calculation due to the smaller concentration range and may create a false impression of phase transition.

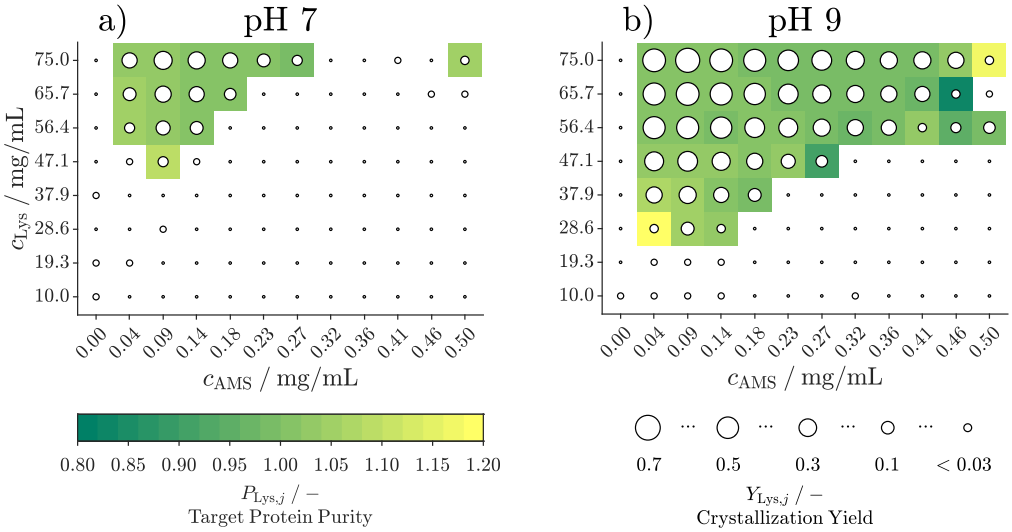


Figure 3.6 The model derived purity of the produced target protein crystals is calculated with Eq. 3.3 and illustrated with the background color of each condition. The conditions varied in pH, and protein and precipitant concentration (see Figure 3.2). The crystal yield $Y_{i,j}$ is calculated with Eq. 3.2 and illustrated as in Figure 3.5. Only the screening conditions demonstrating crystallization yields above 5.0 % are colored.

Figure 3.6 visualizes both, the calculated Lys yield in the screened conditions at both pH levels, similarly to Figure 3.5, and the model derived purity of produced Lys crystals. At

pH 7 the calculated purity varied between 99.68 % and 109.16 %; at pH 9 between 81.62 % and 142.91 %. Most screening conditions, demonstrating crystallization yields above 5.0 %, indicated high Lys crystal purity near 100 %. Outliers were only visible at the edge of the crystallization window at pH 9.

Purity values above 100 % were identified as artefacts of the model derived purity calculation according to Eq. 3.3. The predicted protein-specific concentrations of all three species present in the supernatants were included in this calculation making the purity determination prone to experimental inaccuracies. However, the model derived purity rather aims to provide a fast, qualitative estimation of crystal purity, than an exact determination.

pH: In Figure 3.5, the effect of the pH on the crystallization of Lys was clearly visible as the window of phase transition was larger at elevated pH. As the pI of Lys is above 9.95 [156], the increased tendency to form crystals can be explained by the minimized repulsive forces due to decreased inter-protein electrostatic forces near the pI. An additional experiment at pH 5 showed no decline in protein concentration in the supernatant (data not shown) supporting this statement. This trend is in good agreement with previous reports on pure Lys crystallization studies [9] keeping in mind the longer incubation time and higher storage temperature.

The effect of pH was not visible in the case of CytC or RibA presumably due to the lower protein concentrations.

Lys and AMS concentration: Especially at low AMS and high Lys start concentrations, Lys showed the tendency to undergo crystallization. The supersaturation-driven process led to the highest crystallization yields. This was expected as higher initial Lys concentration would lead to a higher supersaturation [9] and thus more crystals [27].

Using the sitting drop technique, Forsythe at al. investigated Lys crystal morphologies with different sulfate ions at high protein concentrations above 100 mg/mL between pH 4.0 and pH 7.8 [136, 137]. Similarly to the present work, they stated that low AMS and high protein concentrations initiate crystallization.

At lower AMS concentration, Lys could crystallize at lower initial Lys concentration due to the kosmotropic nature of AMS. Hydrophobic interactions are strengthened favoring the folded state and self-association. This can support nucleus formation and crystal growth. With increasing AMS concentration salting-in becomes more visible at both investigated pH levels. As salting-in effects were already described in previous work [136, 137, 157], this is not further discussed in detail.

Presence of contaminants: The presence of impurities can have a great effect on the protein crystallization window as different species prevent the formation of a critical nucleus or crystal growth [15, 16]. To assess the effects caused by impurities, the crystallization windows in Figure 3.5 a) and a reference plate with pure Lys solutions at pH 9 were compared visually using the recorded images. The impeding effect of both contaminants could not be observed in the case of Lys at high supersaturation but showed minor differences near the edges of the crystallization window. The pure Lys solution phase diagram generated a larger crystallization window in the case of 9 conditions (data not shown). This confirms previous findings of Judge et al. [15] that structurally unrelated impurities only affect Lys crystallization at low supersaturation levels.

Selectivity and purity: Comparing the protein-specific windows of phase transition in Figure 3.5 at each pH, co-crystallization or systematic integration of contaminants into the formed Lys crystals were not observed. The scattered declines in supernatant concentration of the contaminants did not overlay with the Lys crystallization window and therefore high selectivity of the examined crystallization conditions is assumed. This is further supported by the high model derived purity of the crystallizing conditions, depicted in Figure 3.6. These findings match previous work of Judge et al. [15] in which structurally unrelated proteins were not detected in Lys crystals when grown in a protein-contaminated environment and a high crystal purity was achieved.

Sensitivity and specificity: The reliability to detect Lys crystallization was evaluated qualitatively by comparing Figure 3.5 a) and d) with the final images of incubation. The conditions displaying precipitate or condensation at the sealing tape were left out for the evaluation of the sensitivity and specificity (see A3.4 and A3.5 for the equations and data). All conditions with yields above 5.0 % were scored as successful crystallization conditions in order to compensate for the impact of measurement errors and model inaccuracies on the yield calculation. Note, that one drawback of visible light image-based analysis is the inability to distinguish salt and protein crystals, and this often requires further analysis of the solid crystal or supernatant [158].

The sensitivity to detect crystallization correctly was 96.23 % for all investigated conditions. Only two conditions at the boundary of the crystallization window at pH 7 produced crystals but a yield below the chosen threshold (4.3 % and 2.3 %).

The specificity to detect soluble conditions correctly was 95.24 %. 6 conditions were falsely identified as crystallization conditions. The false-positive detection occurred especially at high initial Lys concentration (above 56.4 mg/mL) or at the boundary of the crystallization window (see A3.5). As the residuals between the PLS predicted and the measured concentrations were low for the false-positive conditions (data not shown), it is assumed that either measurement deviations of the analytical system and dilution errors may be the cause or that subvisible particle formation was detected.

Considering the above, protein crystallization conditions were detected reliably and fast with the new method. Crystallization windows could be visualized and the high sensitivity and specificity enabled elaborate yield calculation.

3.3.3 Crystallization kinetics

Information on crystallization kinetics are valuable for fast process development in order to determine the optimal process time. Limited time and product resources are available during process development, thus fast and accurate analytics with minimal product consumption are desired. Therefore, the applicability of the developed method was examined for a kinetic study conducted in microliter scale by screening two conditions displaying different kinetic behavior.

The protein-specific concentration development over time for both examined conditions (A5 and D6 at pH 9, 8 °C) is shown in Figure 3.7. For both conditions, the exponential decline of Lys concentration over time could be followed within the first day of incubation. The

concentration dropped from its initial concentration due to saturation-driven crystallization and converged towards the saturation concentration of the equilibrium state. Over the first day, the Lys concentration decline in the condition A5 was 3 times steeper than in D6 as a result of the higher supersaturation due to higher initial Lys concentration and lower saturation concentration caused by the AMS concentration. The saturation concentration was derived from the mean concentration (day 2 to day 7) and was 16.9 mg/mL for A5 and 26.7 mg/mL for D6, which were both lower than the Lys concentrations determined during the HT screening over 13 d.

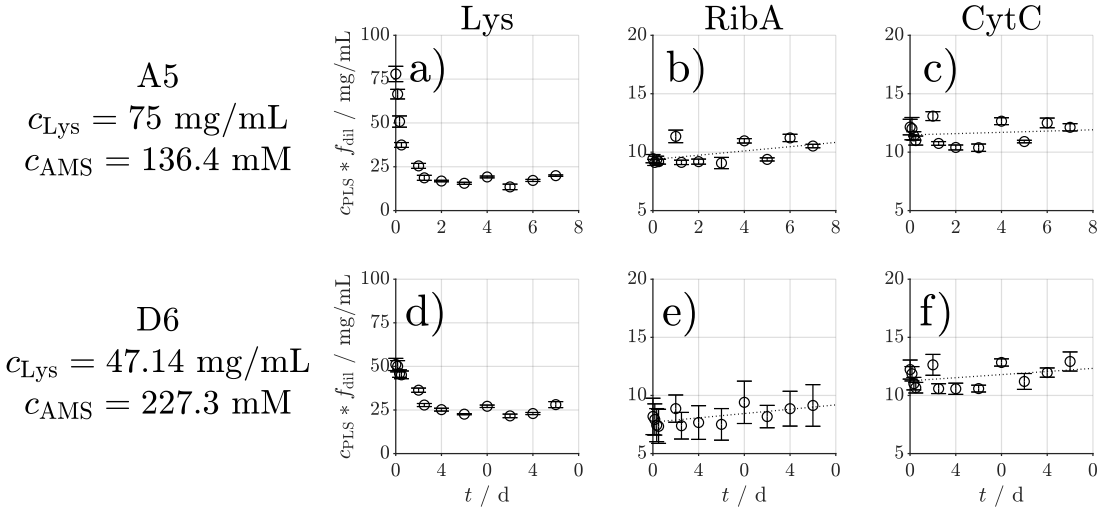


Figure 3.7 The subfigures display the protein-specific concentration development of diluted supernatant samples over time of two screened conditions A5 (a-c) and D6 (d-f). Their starting conditions are described on the left. The mean and standard deviation are illustrated by the circles and whiskers, respectively. The kinetic study was conducted with both biological and technical duplicates over 7 d. Dashed, linear trend lines are included to guide the eye.

Hebel et al. [14] successfully performed a scale-up of an antibody fragment crystallization process from a 10 µL static vapor diffusion screening to stirred 5 mL and 100 mL vessels of the same geometry. The same saturation concentration was achieved for stirred vessels of the same geometry, but a lower saturation concentration was observed during the vapor diffusion micro-batch screening. The difference in process handling, scale and geometry can lead to these discrepancies, similar to the present work, as the working volume in this kinetic study was about 15 times higher compared to the phase diagram experiments. Furthermore, Asherie [2] observed different saturation concentrations of a crystallizing Lys suspension, when shortly shaken, and traced this phenomenon back to improperly oriented proteins at the crystal surface. Agitation then reinitiates crystal growth which could have occurred during timed sampling for the kinetic study.

In contrast, the concentrations of RibA and CytC scattered in Figure 3.7, but increased slightly over the observed time span as indicated by the linear trend. Evaporation of the liquid

over 6 d may be the cause for this increase. The standard deviations of the contaminating proteins were larger compared to Lys. A reasonable explanation is the stronger impact of measurement errors on the lower absolute protein concentrations compared to the higher Lys concentrations.

It is assumed that the contaminating proteins RibA and CytC were not included into the Lys crystals as their concentration did not decrease during the crystal growth phase. The protein solutions reached their equilibrium within the first day and the saturation concentration could be determined.

3.3.4 Potential of PLS-UV/Vis spectroscopy for crystallization

The developed analytical method demonstrates robust quantification of individual species out of a ternary protein mix requiring only 3 µL out of a 24 µL micro-batch. The workflow is HT compatible, can be transferred to other protein systems and allows quick assessment of crystallization conditions regarding performance, purity, and selectivity. The transfer to other protein systems only requires a calibration data set. When material is scarce, this could be realized with a subset of the actual screening if the subset conditions show a protein solid-liquid separation.

Furthermore, crystallization kinetic data can be obtained with this method which would facilitate the knowledge-based development of crystallization processes, the optimization of existing processes or the control of crystal properties without the need of time and material consuming off-line analytics.

The visualization of the more dimensional data in a phase diagram per species can reveal crystallization windows of one species present in a complex mixture, e.g., harvest broth, occurring impurity integration or co-crystallization. The high dilution of the supernatant samples is required to fit the linear absorbance range of the DAD detector, but minimizes the required sample volume and impedes further nucleus formation and crystal growth. Only by this, the method adaption to case specific concentration ranges and timed measurements are possible. The UV/Vis analysis can be conducted with a plate-based spectrophotometer but bears the difficulty to accurately determine the optical path length of the diluted samples.

Compared to standard image-based analytics, the combination of UV/Vis spectroscopy and PLS modeling can easily distinguish between salt and protein crystals, and allows yield calculation. As opposed to X-ray diffraction, the crystal size is not a limiting factor and crystal harvesting is not required. The developed, quantitative method is fast, widely applicable and easy to implement in existing workflows as the analytical devices are present in most laboratories.

3.4 Conclusion

In this study, we have shown that ultraviolet-visible light (UV/Vis) spectroscopy paired with chemometrics is a fast and versatile analytical tool. It can selectively quantify the individual species and be applied to high-throughput (HT) crystallization screenings of

protein mixtures. Three partial least squares (PLS) models were calibrated with ternary protein mixtures and applied on a HT screening of highly concentrated protein mixtures. During the screening, the pH, and the concentrations of the precipitant and target protein were altered. The application of the calibrated PLS models allowed for elaborated yield and purity calculation, and selectivity evaluation. Crystallization kinetics at two different conditions could be monitored with minimal sample intake over time.

Compared to well-established, mostly qualitative crystallization analytics, e.g., X-ray crystallography or image analysis, the newly developed tool stands out in speed, accuracy and simplicity in handling. In 3 min per sample, the spectral analysis offers selective quantification and yield calculation, and specifically detects protein phase transition. The integration of contaminants can be examined and this can be used for purity assessment. Timed sampling and analysis provide knowledge on crystallization kinetics and, by this, crystal properties control.

In future, the combination of UV/Vis spectroscopy and PLS modelling could be used in small scale for phase transition HT screenings, i.e., precipitation or crystallization studies of product in harvest broth. Further research in crystallization kinetics are essential to accelerate process development and scale-up. At larger scales, the provided method could serve as a new tool for on-line monitoring of selective crystallization processes. All of this may encourage alternative process development to well-established chromatography-based processes.

Acknowledgment

This work received funding from the German Research Foundation (Deutsche Forschungsgemeinschaft DFG) in the framework of SPP 1934, project number 315315694. The authors are grateful to Nils Hillebrandt and Robin Schiemer for reviewing the manuscript draft.

4

Calibration-free PAT: Locating selective crystallization or precipitation sweet spot in screenings with multi-way PARAFAC models

Christina Henriette Wegner[1], and Jürgen Hubbuch[1]

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

Abstract

When developping selective crystallization or precipitation processes, biopharmaceutical modalities require empirical screenings and analytics tailored to the specific needs of the target molecule. The multi-way chemometric approach called parallel factor analysis (PARAFAC) coupled with ultraviolet-visible light (UV/Vis) spectroscopy is able to predict specific concentrations and spectra from highly structured data sets without the need for calibration samples and reference analytics. These calculated models can provide exploratory information on pure species spectra and concentrations in all analyzed samples by representing one model component with one species.

In this work, protein mixtures, monoclonal antibodies and virus-like particles in chemically defined and complex solutions were investigated in three high-throughput crystallization or precipitation screenings with the aim to construct one PARAFAC model per case. Spectroscopic data sets of samples after the selective crystallization or precipitation, washing, and redissolution were recorded and arranged into a four-dimensional data set per case study. Different reference analytics and pure species spectra served as validation. Appropriate spectral preprocessing parameters were found for all case studies allowing even the application of this approach to the third case study in which quantitative concentration analytics are missing. Regardless of the modality or the number of species present in complex solutions, all models were able to estimate the specific concentration and find the optimal process condition regarding yield and product purity. It was shown that in complex solutions, species demonstrating similar phase behavior can be clustered as one component and described in the model. PARAFAC as a calibration-free approach coupled with UV/Vis spectroscopy provides a fast overview of species present in complex solution and of their concentration during selective crystallization or precipitation, washing, and redissolution.

4.1 Introduction

The variety and number of biopharmaceutical products are constantly increasing. There are e.g. monoclonal antibodies (mAbs) [159], vaccines [160, 161], and new therapeutics [162]. Each new therapeutic drug is accompanied by new physico-chemical properties, which need to be assessed with target molecule-specific analytics to ensure drug purity and safety for the patient. Broadly applicable analytical technologies are preferred as they can characterize various products and process steps. This may lead to deeper product and process knowledge, together with cost- and risk-based decisions during process development.

Downstream processes of biopharmaceutical products commonly rely on preparative chromatographic processes, which are costly or difficult to scale-up. In general, selective protein crystallization or precipitation can be an alternative to costly chromatography capture steps [127, 163, 164] and bear their advantages, e.g. high purity, concentration, and stability during product storage [124, 125] . Given that the process conditions are selected appropriately, these processes can provide highly concentrated products and can be scaled at lower costs compared to chromatographic process steps. To speed up the process of

finding optimal process conditions, empirical high-throughput (HT) studies are common for early-stage process development and require HT-compatible analytics. In this context, fast, non-destructive, versatile methods, e.g. spectroscopic methods, are preferred and they can be used to determine critical process parameters, e.g. target protein concentration, yield, and purity. When combining HT studies and spectroscopy, though, a situation often arises where large data sets are recorded which are difficult to interpret and are strongly correlated; the information sought-after is hidden in a data jungle. To overcome these limitations, scientists commonly apply chemometric methods to large spectral data sets, e.g. partial least squares (PLS) regression [43, 145, 147], convolutional neural networks (CNNs) [165], or Gaussian process regression [166], and generate process analytical technology (PAT) models to improve the design, analysis, and control during product manufacturing [117]. The mentioned regression models, however, generally require robust reference analytics for calibration. Specific PAT research on crystallization processes mainly focused on mechanistic models for crystal nucleation or growth implementing physical or empirical equations and is discussed elsewhere [167–169].

In the case of spectroscopy measurements recorded over time, three-dimensional (3D) data sets are generated, which are ordered along three dimensions, e.g. wavelength, time, and absorbance. When the spectra of several samples are recorded, four-dimensional (4D) data sets are formed. This multi-dimensionality further complicates the data analysis and calls for multi-way chemometrics. To process data sets of higher order, multi-way chemometric approaches, e.g. generalized rank annihilation method (GRAM), unfolded partial least-squares (U-PLS), and multi-way partial least-squares (N-PLS) regression models, require external calibration [170, 171]. They cannot be applied when accurate reference analytics are missing, e.g. in product capture process steps due to the variety of product- and process-related impurities.

On the contrary, PARAFAC models can analyze data sets of higher order without the need for calibration samples. Given the number of components in the data set, the PARAFAC model can decompose a linear, spectral data set of second or higher order into the signal contribution of each component and regress the model towards a minimal model error compared to the original data set. In this application, one PARAFAC component represents one species in the data set. As a result, the initial data set can be described as the sum of loading vectors of each species in each dimension and the model error [172–174]. PARAFAC was successfully applied to qualitative and quantitative data analysis on excitation emission spectra of fluorescence spectroscopy [112, 175, 176] using data sets structured along excitation wavelength x emission wavelength x samples. Other possible applications are the flow injection analysis (FIA) [177, 178] and high-performance liquid chromatography (HPLC) runs equipped with multi-variate detector, e.g. diode array detector (DAD) [114, 179] or mass spectrometry (MS) [176, 180].

The mentioned work on PARAFAC models focused on the deconvolution of overlapping peaks in chromatography runs or the quantification of chemical analytes in fluorescence spectroscopy. With regard to the rising number of new biopharmaceuticals and early stage process development, HT screenings for crystallization and precipitation processes are time-consuming and need to be evaluated quickly with versatile analytics. This calls for the

investigation of the PARAFAC model application to identify sweet spots in the phase behavior of biopharmaceuticals for crystallization or precipitation processes. This research project thus investigates how PARAFAC models can predict specific spectra and concentration profiles in a screening of unknown species from UV/Vis data.

To show the broad applicability of PARAFAC to HT screenings, three case studies on phase behavior were conducted. The case studies covered one selective protein crystallization process of a defined ternary protein system and two selective precipitation processes of mAbs and virus-like particles (VLPs) in complex solutions. Depending on the case study, UV/Vis spectra were recorded from supernatant samples taken from different process steps, e.g. crystallization, precipitation, washing, and redissolution. Time-resolved spectroscopic data were obtained by injecting samples into a HPLC system equipped with a DAD. No chromatographic column was installed to save analysis time and generate the data with a universal method unaffected by the investigated molecule. This analytical setup led to a second-order data set of three dimensions (wavelength x time x samples). The PARAFAC model calculated the loadings in the mentioned dimensions for each component describing the spectral, time, and concentration profile of the different species.

The presented results demonstrate how multi-way chemometrics can explore spectroscopic screening data sets of higher order. Different case studies with varying product characteristics may be examined with little experimental effort and in a calibration-free way. The PARAFAC models can help to assess selective crystallization and precipitation conditions with regard to purity and yield while increasing process knowledge in early stage process development of new biopharmaceutical products. Reference analytics for calibration are not required for the model calculation making it suitable for use in early stage process development. Additionally, qualitative information on spectra and phase behavior increase process knowledge and may be used for process development according to quality by design (QbD).

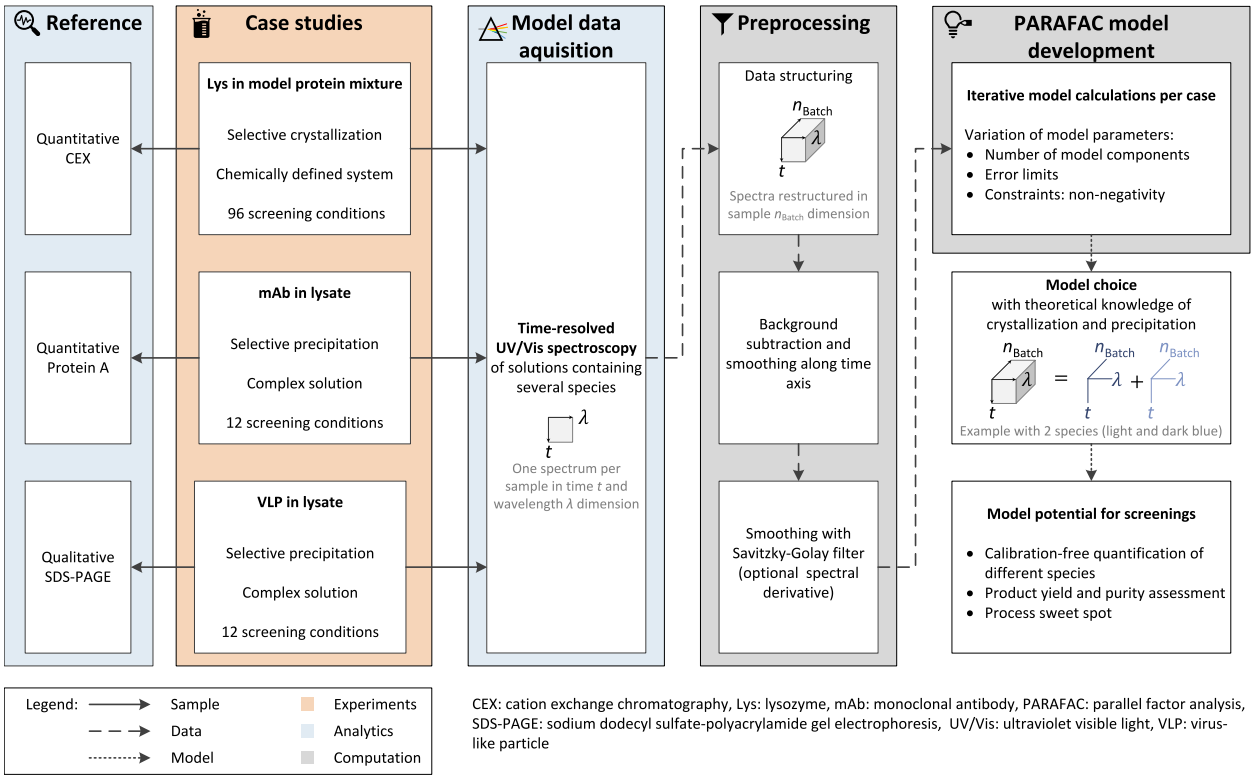
4.2 Materials and methods

The preparation and execution of the first case study were described in detail by Wegner et al. [181] and are described in brief in this work. An overview of the experimental setup, analytics, and computation is visualized in Figure 4.1.

4.2.1 Experiment buffer and protein preparation

All chemicals were purchased from Merck KGaA (Darmstadt, DE), unless otherwise stated. The buffer solutions were prepared at room temperature with ultrapure water (PURELAB Ultra, ELGA LabWater, Lane End, High Wycombe, U.K.), pH-adjusted with 32 % hydrochloric acid (HCl) or 4 M sodium hydroxide (NaOH).

In the first case study, lyophilized model proteins lysozyme (Lys) from chicken-egg-white (Hampton Research, Aliso Viejo, CA), ribonuclease A (RibA) from bovine pancreas, and cytochrome C (CytC) from equine heart were dissolved in multi-component buffer (MCB, 21 mM N-1,1- dimethyl-2-hydroxyethyl-3-amino2-hydroxypropanesulfonicacid (AMPSO), 17 mM 3-



CEX: cation exchange chromatography, Lys: lysozyme, mAb: monoclonal antibody, PARAFAC: parallel factor analysis, SDS-PAGE: sodium dodecyl sulfate-polyacrylamide gel electrophoresis, UV/Vis: ultraviolet visible light, VLP: virus-like particle

Figure 4.1 The workflow for the PARAFAC model calculation can be divided into the experimental work of three different case studies, the analytics, and the computational work. Screening samples are UV/Vis-analyzed and the recorded spectral data set is restructured in the dimensions time t, wavelength λ, and supernatant sample n_{Batch}. Subsequent preprocessing allowed the calculation of one PARAFAC model per case study. The reference analytics validate the generated models and vary depending on the target molecule, purification process, i.e. selective crystallization or precipitation, and the composition of the initial material.

N-morpholino propansulfonic acid (MOPS, Carl Roth GmbH + Co. KG, Karlsruhe, DE), 15 mM succinate acid AppliChem GmbH, Darmstadt, DE) at pH 9. After dialysis to the target multi-component buffer (MCB), the protein concentrations were adjusted as required and the protein solutions were filtered (0.2 µm, Pall Corporation, Port Washington, NY).

For the second case study, Byondis B.V. (Nijmegen, NL) kindly provided frozen cell culture supernatant (CCS) of a mAb harvest of chinese hamster ovary (CHO) cells. The material was thawed, filtered (0.2 µm, Pall Corporation), aliquoted, and stored at −20° until later usage.

The required amount of CCS was thawed and a buffer exchange was performed to a phosphate-buffered saline (PBS) buffer (58.4 mM sodium chloride (NaCl), 74.6 mM potassium

chloride (KCl), 136.1 mM potassium dihydrogenphosphate (KH2PO4), 142.0 mM disodium hydrogen phosphate (Na2HPO4), pH 7.4) using a PD MiniTrap™G-25 column (GE Healthcare, Chicago, IL)). The CCS stock solution was filtered (0.2 µm, Pall Corporation) prior to screening.

The third case study involved truncated Hepatitis B core antigen (HBcAg) VLPs [182]. The VLPs were produced in-house in *E. coli* as previously described by Hillebrandt et al. [34]. After filtering the lysed material with a glass fiber, a 0.45 µm, and a 0.2 µm cellulose acetate (CA) syringe filter (Sartorius Stedim Biotech GmbH, Göttingen, DE), the material was 3 times diluted, aliquoted, and stored at −30° until further usage. For the screenings, the material was thawed and filtered (0.2 µm, CA, Pall Corporation).

The used crystallization solution was the MCB at pH 9 and contained additional 3.5 M ammonium sulfate (AMS). The precipitation solution of the second and third case studies contained only 3.6 M AMS. The redissolution buffers were PBS buffer, pH 7.4 in the second (mAb) and 50 mM Tris buffer, pH 7.2 in the third case study (VLP).

4.2.2 Crystallization and precipitation experiments

The following subchapter describes the experimental conditions of the three HT screening case studies. The second and third paragraphs deal with selective crystallization in a ternary protein mixture and with the selective precipitation of mAbs and VLPs in complex solutions, respectively.

The prepared protein solutions for the ternary phase diagram were mixed and crystallized in 24 µL micro-batches as described by Wegner et al. [181]. 3 µL samples for the analysis were drawn after 13 d of incubation at 8 °C and 50 times diluted with MCB, pH 9.

The selective precipitation screenings were conducted by mixing 278 µL of 12 differently diluted precipitation solutions with 222 µL of the initial mAb or VLP protein stock solutions leading to twelve 500 µL batches. The desired screening range of AMS was between 0 and 2 M. The precipitation solutions were shaken using a thermo shaker at 300 rpm for 30 to 60 min and then centrifuged (17000 g, 2 min). The shaking and centrifugation conditions were used for all steps. The supernatant (S1) was removed, and a wash step was performed by adding 500 µL of a buffer containing the same components as the respective screening condition. Then, the supernatant solutions were centrifuged and the wash step supernatant (S2) was removed. Adding 500 µL of the respective redissolution buffer (see Subchapter 4.2.1) and shaking for 2 h redissolved the precipitate. Eventually, the redissolution batches were centrifuged (S3).

Supernatant samples (S1 - S3) were drawn after each centrifugation step, diluted (mAb: 2 times; VLP: 10 times) with redissolution buffer, and cooled at 8 °C until the analysis at the end of the experiment.

4.2.3 Analytics

4.2.3.1 Multi-way UV/Vis spectra

First, the samples were UV/Vis-analyzed using a Dionex Ultimate 3000 RS HPLC system (Thermo Fisher Scientific, Inc., Waltham, MA) equipped with a RS diode array detector. The UV/Vis spectra were recorded by injecting 20 µL sample volume into the device with no column installed. The injection volume stayed constant for all HPLC measurements. The detector data acquisition was performed with 100 Hz frequency and in the wavelength range of 240 to 450 nm for the first and 220 to 550 nm for the remaining case studies. A filter cartridge (pore size 0.5 µm, OPTI/SOLV EXP, Merck KGaA (Darmstadt, DE)) was integrated to impede aggregates in the detector. The mobile phase was a (50 mM Tris, 100 mM NaCl, pH 8.0) buffer for the first case study or the respective redissolution buffer of the case study and the flow rate was 200 µL/min in the first or 50 µL/min for the remaining case studies.

4.2.3.2 Reference analytics

Different analytics were applied depending on the case study and target protein. The reference data of the first study were derived from cation-exchange chromatography (CEX) performed with a ProSwift SCX-1S 4.6 x 50 mm column using the aforementioned HPLC system (see Subchapter 4.2.3.1 with a low salt buffer (50 mM Tris, pH 8.0) and high salt buffer (50 mM Tris, 1 M NaCl, pH 8.0) with a flow rate of 1.5 mL/min [181].

A 2.1 x 30 mm POROS™ protein A column (Applied Biosystems, Waltham, MA) was used to separate the mAbs from the contaminants, and it allowed species quantification. After sample injection, the column was equilibrated with equilibration buffer (PBS buffer, pH 7.4) for 16 column volumes (CVs) and eluted with elution buffer (PBS buffer, pH 2.6) for 28 CVs. The flow rate was set to 2 mL/min.

For the third case study, the sample purity was assessed only qualitatively with sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE). The analysis was performed with lithium dodecyl sulfate (LDS) sample buffer, 2-(N-morpholino)ethanesulfonic acid (MES) running buffer, and NuPage 4-12 % BisTris Protein Gels (all Thermo Fisher Scientific, Inc.). The addition of reducing 50 mM dithiothreitol (DTT) was the only adaption to the manufacturer's protocol.

The pure species spectra of Lys, RibA, and CytC were recorded by measuring single protein solutions using the setup described in Subchapter 4.2.3.1. In line with this, the pure VLP spectrum was derived from a re-dissolved and sterile-filtered VLP solution purified by diafiltration and multimodal size-exclusion chromatography according to Hillebrandt et al. [183]. The contaminant and the pure mAb species spectra were calculated from the protein A analysis flow-through and elution peak.

4.2.4 Data analyses

All data analyses, preprocessing, and model calibration were performed in MATLAB, R2019b (The MathWorks, Inc., Natick, MA), including the MATLAB N-way toolbox [184] to construct the chemometric models.

4.2.4.1 Data structure and preprocessing

Each UV/Vis-analyzed sample measurement led to a 3D spectral data set spanned over the system retention time, wavelength measuring the absorbance, similar to a 3D chromatographic data set with strongly overlaying species peaks. When multiple supernatant samples per case study were analyzed, the generated data were arranged along the sample number leading to a 4D data set. For each case study, one 4D data set was constructed, preprocessed, and used for the model calculation.

Preprocessing (see Figure 4.1) consisted of the background subtraction and smoothing the absorbance data set along the time axis. The preprocessed data were cut to a wavelength range of 255 to 410 nm for the first and 255 to 310 nm for the remaining case studies to leave out the non-absorbing wavelength ranges and thus improve the model development. For each case study, the preprocessing parameters were varied and tested for the spectral and time-wise smoothing (see Table 4.1) with a Savityky-Golay smoothing filter [149]. The third

Table 4.1 Preprocessing and model development parameters: These parameters were varied for each case study to find optimal calculation parameters. The final calculation parameters are listed as well.

		Data preprocessing			**Model parameters**	
		Derivative	Time smoothing range	Wavelength smoothing range	Number of model components	Error limit
Case 1	max	2	10	13	3	0.010000
	min	0	1	3	2	0.000001
Case 2	max	0	51	7	4	0.008000
	min	0	10	5	3	0.000010
Case 3	max	2	35	7	4	0.008000
	min	0	10	5	2	0.000100
Case 1		0	3	7	2	0.000001
Case 2	final	0	10	7	3	0.000100
Case 3		2	10	7	3	0.000100

data set required the calculation of the second derivative with the Savitzky-Golay filter to enhance spectral differences as the species present in the examined solutions showed strongly overlapping spectra.

4.2.4.2 PARAFAC model construction

The calculation of the PARAFAC models (see Figure 4.1) was performed varying the model parameters, i.e. error limits, and number of PARAFAC components. Especially, the latter needs to be selected with care as this parameter is essential for a valid model. These model calculation parameter ranges are listed in Table 4.1. Additionally, the non-negativity constraint was imposed in time, wavelength, and concentration dimension in all case studies with one exception. For the third model, this constraint was left out in the wavelength dimension due to the second-derivative preprocessing data treatment (see Subchapter 4.2.4.1). Due to instability reasons of the PARAFAC model algorithm, ten different models for each selected preprocessing and model parameter set were calculated. The model with the highest core consistency diagnostic (CORCONDIA) value [109] was chosen if the loadings in the concentration mode were sensible and agreed with the theoretical knowledge of protein crystallization and precipitation. In detail, this means that the calculated concentration loadings of all protein species were assumed to decrease to their protein-specific solubility lines with increasing precipitant concentration. The inverse behavior was expected for the analyzed redissolution solutions.

The used PARAFAC algorithm kept the data variance only in the first mode - the time loadings - leading to normalized spectral and concentration loadings.

4.3 Results

4.3.1 Case 1 - Selective crystallization of lysozyme in a ternary protein solution

As a proof of concept, the PARAFAC model construction was first applied to UV/Vis spectral data of a phase transition process of a chemically defined system. In a system of three model proteins, the target molecule (Lys) was selectively crystallized in a HT screening with 96 different conditions. The other two species (CytC and RibA) are arbitrarily treated as contaminants and were preferred to stay in the supernatant to achieve a high Lys purity in the crystals.

The supernatants of the screened conditions holding different protein-specific concentrations were UV/Vis-analyzed. The generated data was used for the model construction. The selected model required two PARAFAC components - one for the target molecule Lys, and the second one for clustering the contaminants. Figure 4.2 A shows the PARAFAC-predicted single species time profiles compared to the measured absorbance of the initial material at the wavelength $\lambda = 280\,\mathrm{nm}$ over time. The dashed and solid lines visualize the model-predicted data (right axis) and the measured data (left axis), respectively. This remains consistent

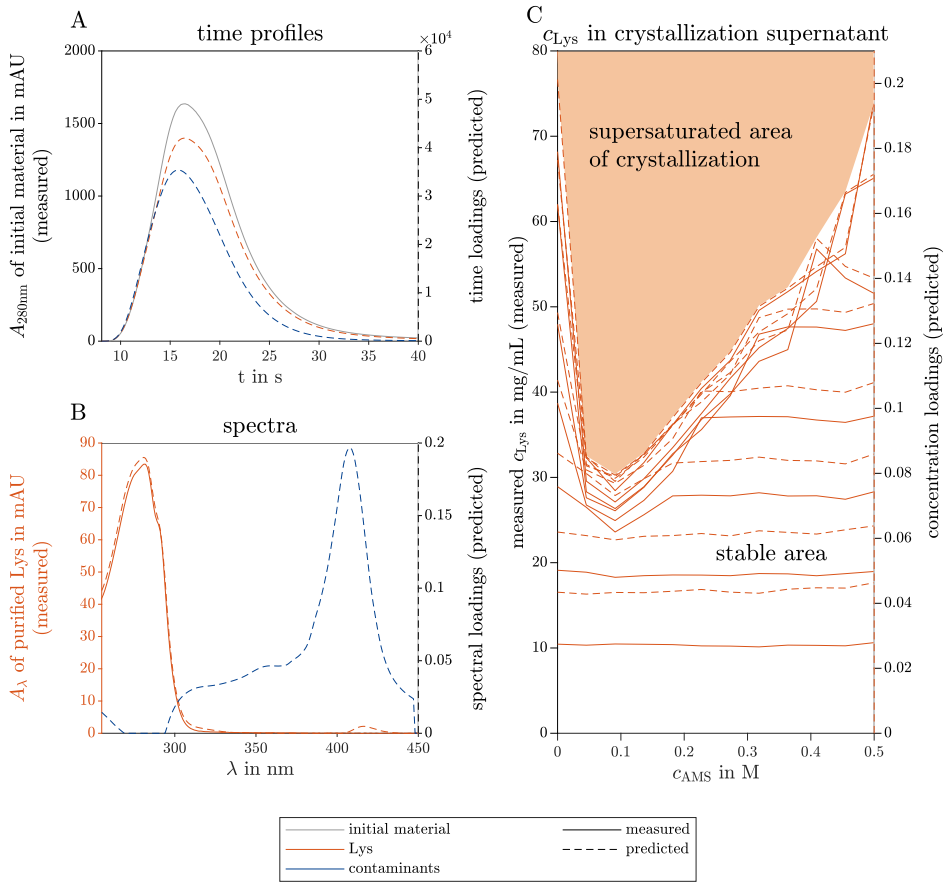


Figure 4.2 PARAFAC model results of the selective crystallization screening of Lys in a ternary model protein system. The measured reference data (left axis) and the predicted loadings (right axis) are illustrated with solid and dashed lines, respectively. The colors gray, orange, and blue indicate the initial raw material, the target species Lys, and the contaminating species, respectively. The time course loadings in (A) show the PARAFAC model predictions of the species absorption loadings over time t in the flow cell of the UV/Vis detector. Additionally, the spectral absorption of the initial solution A_{280nm} is shown at wavelength 280 nm over time. The spectral loadings in (B) demonstrate the similarity between the predicted and the measured Lys absorption spectra A_λ over the wavelength λ. From the concentration loadings in (C), the predicted saturation curve can describe the phase behavior of Lys in the investigated ternary model system and can distinguish the screened conditions into the supersaturation and stable area. The variables c_{Lys} and c_{AMS} represent the concentrations of Lys and AMS, respectively.

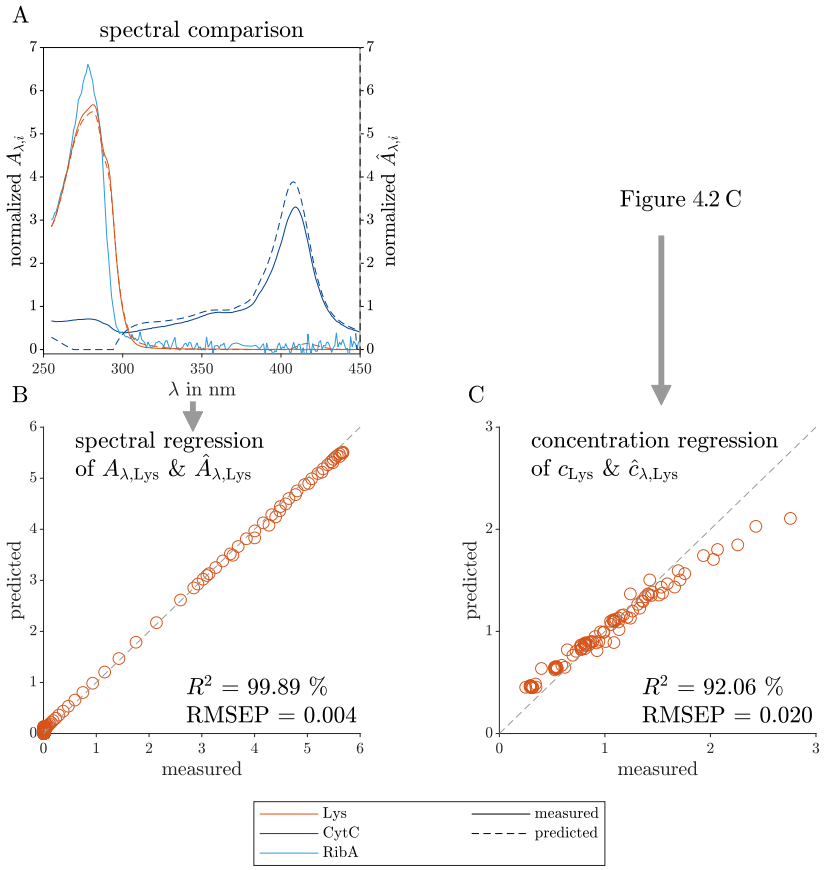


Figure 4.3 Comparison between predicted and measured data of the spectral and concentration loadings of the selective crystallization screening of Lys. The measured reference data (left axis) and the predicted loadings (right axis) are illustrated with solid and dashed lines, respectively. The colors orange, dark blue, and light blue indicate the contaminating species, target species Lys, the model proteins CytC and RibA, respectively. The predicted spectral loadings, and the measured reference data are used to calculate the mean-normalized predicted and measured absorption ($\hat{A}_{\lambda,i}$ & $A_{\lambda,i}$) which are plotted over the wavelength λ for each species i in (A). The predicted absorption $\hat{A}_{\lambda,\mathrm{Lys}}$ of Lys is shown over the measured absorption $A_{\lambda,\mathrm{Lys}}$ of Lys in (B). Figure 4.2 C is used to calculate the mean-normalized predicted concentration loadings $\hat{c}_{\mathrm{Lys}}$ and the measured concentration data c_{Lys} of Lys in (C). The gray dashed lines visualize the ideal fit of the predicted to the measured data (B & C). The calculated, high coefficient of determination R^2 values support the PARAFAC model validity.

throughout this research work. The predicted spectra of the two components are illustrated in Figure 4.2 B in different colors for each species. As a reference, the pure Lys spectrum is included with solid lines for identification of the target molecule component. The predicted and measured Lys concentration of the supernatants of the screened conditions are depicted in Figure 4.2 C. This plot illustrates the phase behavior of Lys in a phase diagram depending on the AMS and initial Lys concentration of the screened condition, and distinguishes between the supersaturation and stable area. The loading vectors in all three modes are unitless, and one component represents one species in each mode. The concentration of the contaminant species did not change (data not shown). The phase behavior of this HT screening is described and explained in detail by Wegner et al. [181]. The time courses of the predicted two species match the position of the overall absorbance at $\lambda = 280\,\text{nm}$ of the analyzed initial material. Both predicted species demonstrate a similar flow behavior through the HPLC system during the no-column runs and resemble the Gaussian shape due to axial diffusion in the analysis system. The spectral prediction of the Lys component fits the measured spectrum of pure Lys, only the shoulder at $\lambda = 290\,\text{nm}$ is slightly less pronounced than in the measured spectrum. The predicted concentration loadings and measured concentrations overlay and indicate the saturation curve of the phase diagram clearly. This curve distinguishes the screened condition into the stable area showing no Lys concentration decline in the supernatant and the supersaturation area, in which the Lys concentration drops to the saturation curve, possibly due to crystallization.

To compare the predicted PARAFAC loadings and the measured reference data, Figure 4.3 depicts the model and measurement data sets in two ways. First, the data sets in Figure 4.3 A show the predicted spectral loadings and measured species, similarly to Figure 4.2 B, but with the spectra of all three model proteins (Lys, CytC, and RibA) present in the screening solutions. Second, the spectral data of the Lys species were mean-normalized to overcome the difference in axis scale. Finally, the data sets were plotted against each other and used for the coefficient of determination (R^2) and root mean squared error of prediction ($RMSEP$) calculation (see Figure 4.3 B for the Lys spectrum and Figure 4.3 C for the concentration comparison). Figure 4.3 C is derived from the mean-normalized concentration data of Figure 4.2 C. The $RMSEP$ in this work is given without a unit as the variable is calculated from normalized values.

The RibA UV/Vis spectrum shows a noisy spectrum above 300 nm, which is a normalization artefact as the overall absorption of the pure RibA spectrum was low due to its low extinction coefficient and the measured concentration of 0.2 mg/mL. It is visible that the predicted contaminant spectrum is similar to the pure CytC spectrum between 300 - 450 nm. According to the model, below 300 nm, the two contaminant species (CytC and RibA) do not contribute to the measured UV/Vis absorbance which differs from the measured pure species spectra. PARAFAC models with three components did not lead to reasonable models, so that the species RibA was not modeled as an own species due to its low contribution to the overall UV/Vis absorbance. However, RibA and CytC together can be clustered as impurities and can be described by one contaminant component as they demonstrate similar phase behavior.

The mean-normalized model prediction and the measured mean-normalized spectrum of pure Lys overlay as indicated by the high R^2 value. The Lys concentration loadings of the PARAFAC model are slightly underestimated at higher protein concentrations, which is quantified with a lower R^2.

4.3.2 Case 2 - Selective precipitation of monoclonal antibodies in a complex solution

As the second case study, a mAb was selectively precipitated out of a clarified, complex solution (CCS) consisting of several different species. In total, 12 different precipitant concentrations were investigated, and the supernatants of the precipitation (S1), wash (S2), and redissolution (S3) process steps were UV/Vis-analyzed to finally construct a valid PARAFAC model.

The results of the constructed model with three different components are shown in Figure 4.4. The three components could be identified as the mAb, contaminants, and AMS. The predicted time profiles of each component and the measured absorbance at $\lambda = 280\,\text{nm}$ are shown in Figure 4.4 A. The predicted spectral profiles and the measured spectrum of purified mAb are depicted in Figure 4.4 B. The predicted, specific concentration in the supernatant of precipitation (Figure 4.4 C), wash (D), and redissolution process step (E) are colored according to the species. As a reference, the measured peak area of the mAbs and the contaminant are included in Figures 4.4 C-E and represent the concentration profile throughout the investigated screening conditions.

The predicted time profiles in Figure 4.4 A show a Gaussian curve for the contaminant species, two Gaussian curves for the AMS species, and an irregular profile for the mAb component resembling multiple overlaying species. The predicted AMS time profile overlaps with the measured time profiles of pure AMS solution measurements (see Figure A4.1).

The predicted spectrum of the target molecule mAb fits the measured spectrum of protein A purified mAb (see Figure 4.4 B). The predicted concentration profile of the AMS during the precipitation and wash step agrees with the experimental AMS concentration as the precipitant concentration was linearly increased over the investigated conditions from 0 to 2 M during the precipitation and wash process step (see Figure A4.2). The predicted and the measured mAb concentrations in the precipitation supernatants decrease strongly above 1.2 M AMS in Figure 4.4 C and match the increase in mAb concentration in the redissolution solutions above the same AMS concentration in Figure 4.4 E. The predicted and the measured contaminant concentrations behave likewise with a different threshold at 1.6 M AMS. A slight increase in the mAb concentration at 1.6 M AMS during the washing step is visible in the predicted and the measured data sets. A slight increase in the contaminant concentration with rising AMS concentration was only seen in the reference analytics and indicates contaminant removal during the wash step. The predicted mAb concentration in Figure 4.4 C is overestimated at AMS concentration between 0 and 0.4 M AMS whereas the contaminant concentration is underestimated. Similarly, the behavior of overestimated mAb and underestimated contaminant concentrations is visible in the redissolution samples at higher AMS screening conditions in Figure 4.4 E.

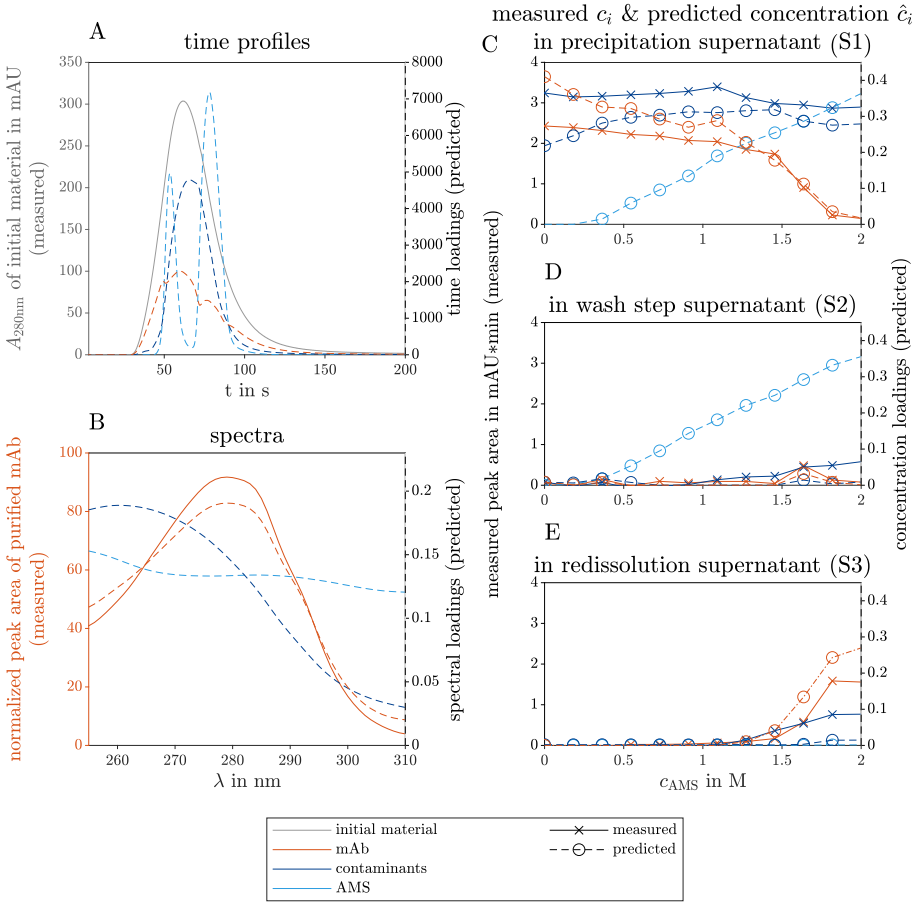


Figure 4.4 PARAFAC model results of the selective mAb precipitation screening from clarified CHO CCS. The measured reference data (left axis) and the predicted loadings (right axis) are illustrated with solid and dashed lines, respectively. The colors gray, orange, dark blue, and light blue indicate the initial raw material, the target mAb, the contaminating species, and the precipitant AMS. The time course loadings in (A) show the PARAFAC model predictions of the species absorption loadings over time t in the flow cell of the UV/Vis detector. Additionally, the spectral absorption of the initial solution $A_{280\text{nm}}$ is shown at wavelength 280 nm over time. The spectral loadings in (B) illustrate the predicted contaminant spectrum over the wavelength λ and the similarity between the predicted and the measured mAb spectrum. The predicted concentration loadings $\hat{c}_i$ are shown over varying precipitant concentration c_{AMS} during the precipitation in (C), wash step (D), and redissolution process step (E). The measured concentration c_i is derived from the peak area of the reference analytics. The peak areas of a reference analytic represent the concentrations of the mAb and the contaminant. They are shown in (C-E) with solid lines.

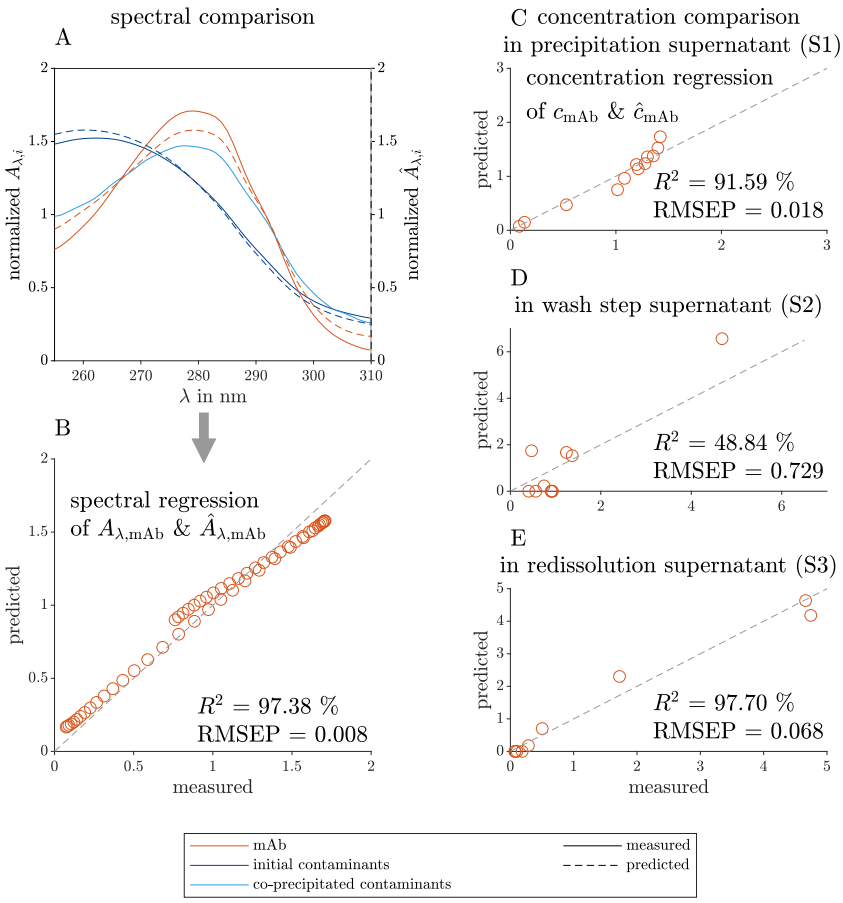


Figure 4.5 Comparison between predicted and measured data of the spectral and concentration loadings of the selective mAb precipitation screening. The measured reference data (left axis) and the predicted loadings (right axis) are illustrated with solid and dashed lines, respectively. The colors orange, dark blue, and light blue represent the target species mAb, the contaminating species before precipitation, and the remaining contaminant species after redissolution, respectively. The spectral predictions $\hat{A}_{\lambda,i}$ and measurements $A_{\lambda,i}$ are mean-normalized and depicted over the wavelength λ in (A). The predicted spectral mAb loadings $\hat{A}_{\lambda,\mathrm{mAb}}$ and the measured reference spectrum of purified mAb $A_{\lambda,\mathrm{mAb}}$ are used to plot the predicted over measured data in (B). The predicted mAb concentration loadings $\hat{c}_{\mathrm{mAb}}$ and measured concentration reference c_{mAb} from mAb peak areas (see Figures 4.5 C-E) are mean-normalized and plotted against each other for the process steps of precipitation (C), washing (D), and redissolution (E). These data were used to calculate the coefficient of determination R^2 values to quantify the validity of the constructed model. The gray dashed lines visualize the ideal fit of the predicted to the measured data (B-E).

To further validate the constructed PARAFAC model, comparisons of the predicted loadings, and measured data of the mAb spectrum and concentration are illustrated in Figure 4.5. The predicted spectral loadings of the mAb and the predicted contaminant are shown in Figure 4.5 A, as well as the spectrum of the initial contaminants, present in the precipitation supernatant, and of the co-precipitated contaminants, which are still present after redissolution. The initial contaminants, which are present in large excess and remain in solution despite the presence of the precipitant AMS, are well described by the blue contaminant component of the PARAFAC model. The co-precipitated contaminants could not be described by the model as these contaminants underwent phase transition at similar precipitant concentration as the target molecule. The mean-normalized, predicted spectral loadings and the measured spectrum of the mAb species are depicted in Figure 4.5 B and agreed as indicated by the R^2 value of 97.38 % and a low $RMSEP$ of 0.009.

To further visualize the model agreement, the predicted, mean-normalized concentration loadings and measured peak area of the mAb are shown during the different process steps in Figures 4.5 C, E, and F with their process-specific R^2 and $RMSEP$ values. The concentration loadings show moderate agreement with the measured data for the precipitation and wash step samples. In the precipitation supernatant analysis, the presence of the different contaminants at high mAb concentration (especially at lower AMS concentration) might be the cause. The wash step analysis samples showed very low mAb concentration except for one outlier. The lowest R^2 and the highest $RMSEP$ values among the investigated process steps might be caused by a mathematical artefact and the outlier. The high R^2 and low $RMSEP$ values for the precipitation and redissolution supernatant indicate that the model could produce valid mAb concentrations.

4.3.3 Case 3 - Selective precipitation of virus-like particles in a complex solution

The third case study dealt with the selective precipitation of VLPs in *E.coli* lysate. In line with the second case study, a screening was performed over different precipitant concentrations, and the UV/Vis-analyzed precipitation (S1), wash (S2), and redissolution step (S3) supernatants were used to construct a PARAFAC model.

The results of the constructed model with three different components are shown in Figure 4.6. The three components are identified as the VLPs and two contaminant clusters.

The time profiles in Figure 4.6 A show a flat, broad peak for the VLP species. The calculation of the second derivative of the spectra along the wavelength dimension improved the model validity (data not shown). The second spectral derivative of a reference spectrum of purified VLPs validated the spectral PARAFAC loadings (see Figure 4.6 B). The reference data illustrate how well the peak position is found by the PARAFAC model estimation of the spectra. The concentration loadings of the different species during the precipitation, wash, and redissolution process step are depicted in Figures 4.6 C, D, and E, respectively. The VLP species concentration decreases with rising AMS concentration above 1 M concentration and approaches a limit (see Figure 4.6 C). The VLP concentration loadings of the redissolution step show the inverse behavior above the same threshold (see Figure 4.6 E). The first contaminant

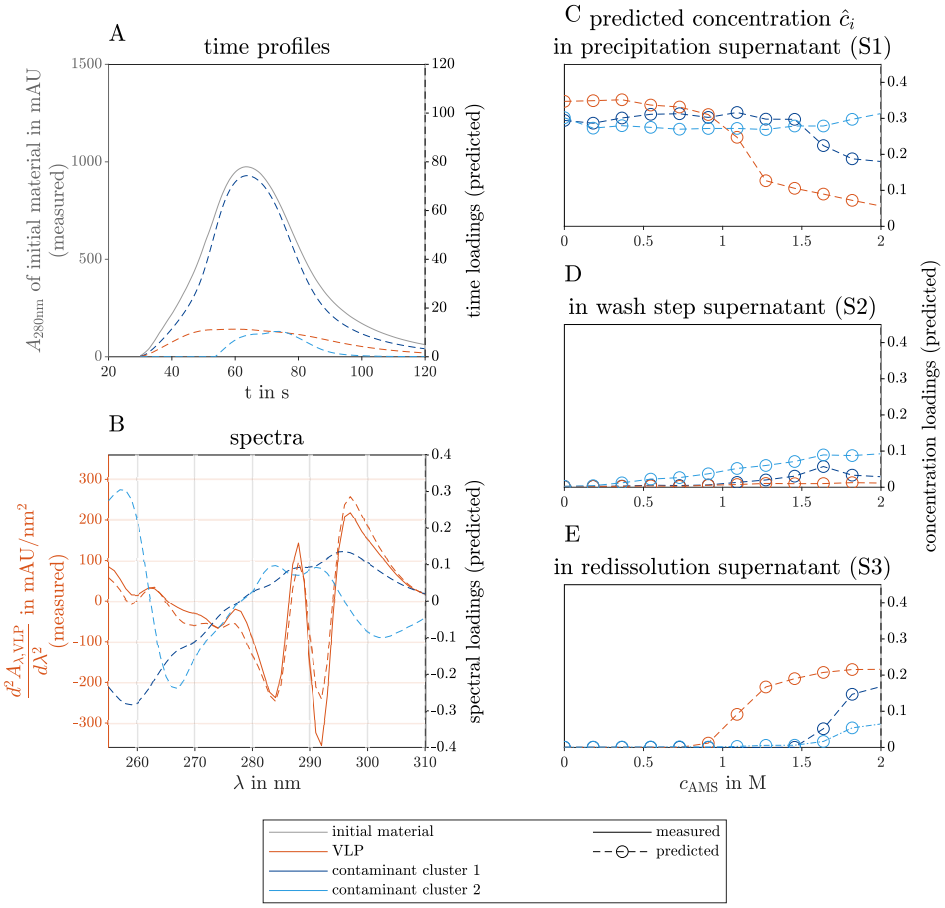


Figure 4.6 PARAFAC model results of the selective VLP precipitation screening from *E.coli* lysate. The measured reference data (left axis) and the predicted loadings (right axis) are illustrated with solid and dashed lines, respectively. The colors gray, orange, dark blue, and light blue indicate the initial raw material, the VLPs, and two contaminant clusters. The time course loadings in (A) show the PARAFAC model predictions of the species absorption loadings over time t in the flow cell of the UV/Vis detector. Additionally, the spectral absorption of the initial solution A_{280nm} is shown at wavelength 280 nm over time. The spectral loadings in (B) illustrate the predicted contaminant spectra over the wavelength λ and the similarity between the predicted loadings and the measured second derivative of the VLP spectrum $\frac{d^2 A_{\lambda,VLP}}{d\lambda^2}$. The predicted concentration loadings $\hat{c}_i$ are shown over varying precipitant concentration c_{AMS} during the precipitation in (C), wash step (D), and redissolution process step (E).

cluster shows a similar behavior above 1.5 M AMS with a higher limit in the precipitation solutions and a lower limit during the redissolution step. Presumably, this contaminant cluster

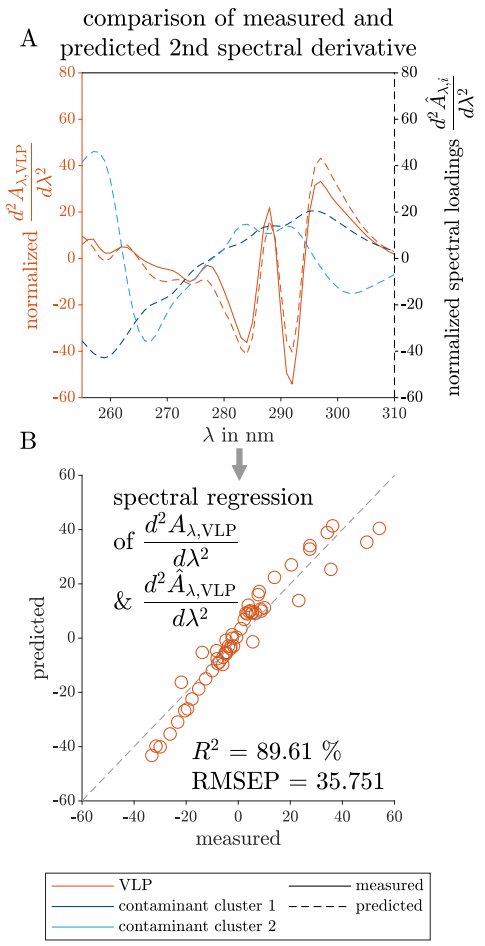


Figure 4.7 Comparison between predicted and measured data of the spectral loadings of the selective VLP precipitation screening. The measured reference data (left axis) and the predicted loadings (right axis) are illustrated with solid and dashed lines, respectively. The colors orange, dark blue, and light blue indicate the VLPs, and two contaminant clusters. The predictions of the second derivative spectra $\frac{d^2 \hat{A}_{\lambda,i}}{d\lambda^2}$ of the species i and the spectral second derivative measurements $\frac{d^2 A_{\lambda,\mathrm{VLP}}}{d\lambda^2}$ of purified VLP solutions are mean-normalized and depicted over the wavelength λ in (A). The predicted spectral second derivative loadings of VLPs $\frac{d^2 \hat{A}_{\lambda,\mathrm{VLP}}}{d\lambda^2}$ and the reference $\frac{d^2 A_{\lambda,\mathrm{VLP}}}{d\lambda^2}$ are used to plot the predicted over the measured data in (B). The gray dashed line visualizes the ideal fit of the predicted to the measured data. The measured and the predicted spectra are used to calculate the coefficient of determination R^2 values to quantify the validity of the constructed model.

precipitates to the solubility line above the threshold. During redissolution, the precipitate of screened conditions with high AMS concentration is redissolved. The AMS concentration does not strongly affect the concentration loadings of the second contaminant cluster in the precipitation solutions, but the concentration loadings of this component increase slightly to a limit in the redissolution solutions. The second contaminant cluster represents species that are stable at higher AMS concentration. Similar results were achieved by Hillebrandt et al. [34] for a chimeric VLP construct. The concentration loadings during the wash step show no significant increase in the VLPs and the first contaminant cluster. The second contaminant cluster shows a slight concentration loadings increase and is probably washed out of the precipitate with the rising AMS concentration.

Scanned SDS-PAGE gels of the precipitation and redissolution step are included in the Supplementary Material (see Figure A4.3) analyzing the conditions between 0 to 1.27 M and 2 M AMS concentration. The findings on the concentration profile of the predicted species match the scanned gel of the reference SDS-PAGE analysis (see Figure A4.3).

The similarity between the predicted and measured second derivative of the VLP spectrum is visible in Figure 4.7 A. The estimated wavelength position of the peak maxima and minima fits the measured data in the wavelength range below 265 nm and above 275 nm, but the absolute values at the peak maxima and minima do not overlay. Between the mentioned wavelengths, the curve characteristics of the predicted spectral loadings show a flattened curve and differ from the measured data. The absolute values at the peak maxima and minima do not overlay. This may be the result of the applied preprocessing techniques as smoothing can eliminate or broaden peaks, whereas the spectral derivative calculation is sensitive to subtle differences in spectra.

To visualize the fit of the predicted to the measured data, the mean-normalized predicted VLP loadings and the second derivative data of a measured VLP spectrum are plotted against each other in Figure 4.7 B and used for the calculation of R^2 and $RMSEP$ values. Closer to the center, the predicted data overlay strongly with the measured data. At the boundaries of the spectral loadings, the predicted and the measured data differ more. Still, the spectral loadings showed a high R^2, but the highest $RMSEP$ for the spectral regression among the three investigated case studies.

4.4 Discussion

To prove the overall applicability of PARAFAC models to HT screenings, the three conducted case studies are discussed regarding the choice of the valid PARAFAC model, the process parameters yield and purity, and the differences between the investigated case studies.

4.4.1 PARAFAC model choice

A PARAFAC model can decompose a data set into the signal contribution of each species if the experimental data set has a truly trilinear structure [171, 172]. In the case of spectral data sets, this means that an experimental data set can estimate e.g. the spectrum and

concentration profile of each species present. Considering the physical logic that the spectra and concentration profiles are positive, the non-negativity constraints can be included in the calculation of chemometric models. This is a common practice to find stable, correct multi-way chemometric models during model calculation [112, 113, 172, 185, 186].

Still, valid PARAFAC models can only be constructed if the appropriate number of components [109, 175, 176], preprocessing techniques, and suitable model calculation parameters are used. In the case of biological, complex solutions containing several different species, the requirement of an appropriate number of PARAFAC components imposes a problem for the model calculation. As not every single UV/Vis-absorbing species can be described by one model component, the different species need to be categorized in clusters. These clusters are formed on the basis of their similar phase behaviors among the species and shall be described by one PARAFAC component accepting inaccuracies in the spectral prediction. This simplification of the variety of species to several clusters introduces an error into the model. However, if the target molecule undergoes a phase transition and contributes strongly to the measured spectral data set, the focus of the PARAFAC models is to find the target molecule in any phase behavior screening study. Further strategies [187] to determine the correct number of PARAFAC components are e.g. half-splitting and comparing the experiments [172], evaluating residuals [172, 188], and the CORCONDIA value [109]. More information on finding suitable preprocessing [189, 190] and model calculation parameters [172, 185] can be found elsewhere.

In crystallization or precipitation screenings, it can be expected that the protein concentration decreases to the solubility line with increasing precipitant or protein concentration due to the decreased protein solubility, which results in protein crystallization [2, 36] or precipitation [7, 191, 192]. In the case of selective crystallization or precipitation processes, the phase behavior is protein-specific and can be used for protein purification. This theoretical process knowledge can be included in the choice of the PARAFAC model.

The spectral data set for the first case study was recorded for a HT-selective crystallization screening of Lys in a ternary protein system. In total, 96 conditions were screened varying the initial Lys concentration and precipitant concentration. The initial concentrations of the two other proteins (RibA, CytC) were maintained constant in all screened conditions. As the calculation of PARAFAC models with three components did not lead to a robust model, a model with two components was calculated (see Table 4.1). Evaluating Figure 4.3 A, one component can be identified as the target molecule Lys; the other one as a contaminant cluster resembling mainly CytC. It is assumed that the absorbance contribution of the third species RibA is built into a contaminant cluster [110], and that this third species is not described as a single model component. It contributes to a smaller extent to the UV/Vis spectra due to the lower extinction coefficient in the investigated wavelength range (3.8 and 2.8 times lower at 280 nm than for Lys and CytC) and lower concentration (up to 7.5 times lower than the Lys concentration). Furthermore, the protein concentrations of CytC and RibA do not change during the screening, contrary to the target protein Lys (see Wegner et al. [181] for further explanation). As a consequence, the model cannot distinguish species demonstrating similar phase behavior. This shows that low-absorbing species are difficult to

describe with an own model component, and that species with similar phase behavior can be clustered justifying species clustering in screenings with complex solutions.

The selective precipitation study of mAbs leads to a spectral data set, which can be described by a PARAFAC model with three model components (see Table 4.1). One component represents the target molecule mAb, the other two the AMS concentration and a contaminant cluster. The time profile of the mAb component in Figure 4.4 A may be caused by the changing light refraction when a solution with a high AMS passes the detector (see Subchapter 4.4.2). Another possible source could be different product-related impurities, e.g. aggregates, fragments, as they would show a mAb resembling spectrum, but different retention times in the analysis system due to diffusion. Below the AMS concentration of 0.5 M the mAb species is overestimated and the contaminant cluster is underestimated by the PARAFAC model in Figure 4.4 C. In Figure 4.4 E, the two model components show the same effects above 1.4 M AMS. A possible explanation of these contrasting model discrepancies of the measured to the predicted data is that the predicted mAb UV/Vis spectrum is overestimated below 270 nm leading to inverse effects on the concentration loadings of the mAb and contaminant component. As a result, the spectral loadings of the contaminants may be incorporated in the predicted mAb spectrum and distort the concentration loadings of both species - the target molecule and the contaminant cluster. This effect is more pronounced at higher absorbance values and thus higher protein concentrations. The protein A chromatography gave further information on the composition of the contaminants during the precipitation, wash, and redissolution step. Figure 4.5 A provides information on the main contaminant cluster during the precipitation and during the redissolution step. This means that the co-precipitated contaminant cluster during redissolution cannot be distinguished from the target molecule.

The PARAFAC model of the selective VLP precipitation HT screening could be calculated with three model components (see Table 4.1). One component describes the VLP species while the other two describe two contaminant clusters. Assessing the concentration loadings of all three PARAFAC components in Figure 4.6 C, the predicted species show different phase behaviors with increasing precipitant concentration. This enables the use of a selective VLP precipitation step for purification. Regarding the screened redissolution samples in Figure 4.6 E, the predicted concentration loadings of the VLPs and first contaminant cluster increase above the same precipitation threshold in Figure 4.6 C. The second contaminant cluster shows a slight concentration increase at higher precipitant concentration meaning that this cluster was redissolved and thus precipitated at a higher precipitant concentration. This does not comply with the phase behavior during the precipitation step, and it is expected that this discrepancy is caused by model inaccuracies. This assumption is supported by the highest residuals of this model to the measured summed up spectra for the investigated redissolution samples above the stated threshold (data not shown). Overall, the predicted VLP spectral loadings match the measured VLP spectrum (see Figure 4.7 A). Discrepancies are visible in the regression plot (see Figure 4.7 B) only at the higher or lower values of the spectral loadings. Compared to the first and second case studies, the R^2 value of the third case study for the spectral loadings is lower indicating a greater deviation of the predicted spectra to the measured spectrum. The highest $RMSEP$ is partially caused by the different scale and the model mismatch which can be seen in Figure 4.7 B. Additionally, the required

preprocessing of the VLP screening data included the second derivative to enhance subtle spectral differences between the screened solutions. The spectral preprocessing may lead to higher discrepancies in Figure 4.7 A and lower accuracy compared to the first and second case studies, but led to a robust model.

In summary, the choice of the correct model component and preprocessing techniques is crucial for the model outcome. These need to be selected with care when the investigated screening solutions involve complex solutions. Theoretical knowledge of selective precipitation and crystallization processes helps finding valid PARAFAC models. Nonetheless, the species in complex solutions demonstrating similar phase behavior can be clustered and described by one model component. In the case of co-precipitation of contaminants with the target molecule, the model may merge the spectra of these species in the predicted spectral loadings.

4.4.2 Screening for optimal yield and purity

The developed models provided information on the solubility line, protein phase behavior, and selectivity of the screened conditions. In the first case study, the solubility line of Lys is visible in the phase diagram in Figure 4.3 B and can be used for further yield calculations. As the concentration of the contaminating species stayed constant in the supernatant, it can be assumed that the produced Lys crystals demonstrate a high purity. The research on mAb crystallization screenings spiked with model protein contaminants showed that a high mAb crystal purity is accompanied by contaminants present in the crystallization supernatant [126]. In general, this selective crystallization process depends strongly on the impurity and its concentration [15, 17, 27]. Regarding yield, optimal process conditions were achieved in a precipitant range between 0.05 and 0.15 M AMS.

Assessing the selective mAb precipitation study in Figure 4.4, a high AMS concentration above 1.8 M leads to the highest precipitate yield. Under the same precipitant conditions, the concentration loadings of the contaminant species decrease indicating co-precipitation above 1.5 M AMS, but with a lower yield due to the higher specific solubility concentration. According to the model, the mAb purity of redissolved precipitate is greatly improved when the predicted concentration loadings of the redissolution and the precipitation solutions are compared. Comparing the predicted to measured concentrations, the redissolution solutions show an over- and underestimation of the mAb and contaminant species, respectively. Purity calculations based solely on the predicted concentration loadings would be overestimated. This may be caused by the co-precipitated contaminants (see Figure 4.5 A) as they were not separated during the screening process.

Regarding the selective VLP precipitation process (see Figure 4.6), the model predicts optimal process parameters when the precipitant concentration lies between 1 to 1.5 M to assure a high purity. The predicted concentration loadings of both contaminant clusters did not indicate co-precipitation and, as a result, are not present in the redissolution samples. To increase the product yield, the concentration above 1.2 M is desired, as the VLP concentrations in the precipitation and redissolution samples are near the limit. As quantitative reference analytics are missing for the third case study, these results are based

purely on model predictions and the qualitative validation with the VLP spectrum and the solution composition with the SDS-PAGE analysis (see Figure A4.3).

4.4.3 Experimental and preprocessing differences between the case studies

The experimental setup and the spectral data preprocessing of each case study required adjustments to the specific protein system. This subchapter focuses on the preprocessing differences between the investigated case study, the experimental screening variations between selective crystallization and selective precipitation studies, and their possible effect on the calculated PARAFAC models.

The time smoothing range for the final models of the crystallization case study was lower than for the precipitation case studies (see Table 4.1). The four times higher flow rate of the UV/Vis spectral analysis in the first case study is the reason, as the sample passed by the detector in a shorter time (compare Figures 4.2 A, 4.4 A, and 4.6 A) as the time-resolved, spectral information of the sample is comparable between the case studies after preprocessing. Longer time-wise smoothing may lead to the removal of important information for the model calculation. The selected wavelength range for the first case study was broader than for the other two (see Subchapter 4.2.4.1) since CytC was present in the first case study and has a second absorption maximum at 410 nm. The third case study required the calculation of the second derivative (see Table 4.1). Possible reasons could be that the target molecule VLP did not present distinct spectral differences to the contaminants [193] or contributed less to the measured spectra compared to target molecules of the first and second case studies. The target protein absorption shares of the initial material was high with 89.24 % and 42.82 % for the first and second case study, respectively. The VLP absorption share could not be determined as quantitative UV/Vis absorption data as a reference were missing. The large amount of UV/Vis-absorbing contaminants in the VLP lysate may interfere with the identification of the component representing VLPs. The differences in the time profile peak maxima of the target molecules compared to the contaminants support this assumption (see Figures 4.2 A, 4.4 A, and 4.6 A).

For each case study, the buffer system was adapted to the requirements of the target molecule. The buffer substances were not UV/Vis-active in the used concentration and did not affect the model calculation. On the contrary, the precipitant AMS showed UV/Vis-absorbing behavior in the second case study and had an impact on the constructed models. A possible reason could be that the light refraction occurs when solutions of different density (mobile phase and sample solvent) pass the detector [194]. This strongly depends on the screening AMS concentration and the sample dilution prior to the UV/Vis analysis. In the first case study dealing with the selective crystallization of Lys, the maximal screening AMS concentration was four times lower than in the second and third case studies. The dilution factors for the first, second, and third case studies varied (see Subchapter 4.2.2) and were adjusted according to the total absorbance of the initial material at wavelength 280 nm. Taking all these factors into account, the analyzed samples of the second case study (mAb) contained the highest AMS concentration and thus the AMS concentration contributed to a

greater extent to the recorded UV/Vis spectra. The constructed model compensated this by describing the precipitant concentration with its own model component (see Figures 4.4 C and D). UV/Vis data recorded of buffer solutions containing different amounts of AMS is shown in Figure A4.1 and support this explanation.

The screening volume, screening size, and the analyzed process step solutions differed. The first case study (Lys crystallization) investigated 96 different conditions in 24 µL batches with eight different Lys starting concentrations and twelve precipitant concentrations. Only the supernatant samples of the crystallization step were analyzed. The spectral data set size was varied in this case study. Screening conditions that did not show concentration changes of the target molecule were excluded for model calculation. It was found that a large screening size with little variety in species composition and concentration ratios does not improve the model robustness but decreases the CORCONDIA value and increases the model error (data not shown). Preferably, the model error is low and the CORCONDIA high indicating an appropriate component number [109] and, hence, a valid model. The second and third case studies screened twelve different precipitant concentrations in 500 µL batches for the selective precipitation of mAbs and VLPs. Samples were analyzed during the precipitation, the wash, and the redissolution step leading to a variety of 36 analyzed samples per screening differing in species compositions and concentration ratios. This sample variety improved the model calculation as the CORCONDIA of the final models was higher and the model error lower for the second and the third case studies. The screening volume did not affect the spectral data set or the model calculation as long as there is enough supernatant for sampling. When selective crystallization or precipitation processes are characterized with the PARAFAC approach, the models cannot detect if the proteins crystallized or precipitated, as the generated models rely solely on the UV/Vis spectroscopic data set and specific protein concentration reductions. Regarding the experimental differences between the two processes, an additional centrifugation step is required to separate precipitate from the supernatant. Furthermore, the crystallization process requires more time than precipitation processes due to the time-intensive crystal nucleation and crystal growth of macromolecules [3, 26].

In summary, these three case studies illustrate how the chemometric multi-way approach of PARAFAC can be applied to different phase behavior screenings with varying process conditions. The differences in spectral data preprocessing could be explained leading to a general preprocessing approach for future crystallization and precipitation screenings. Experimental differences in scale, sample dilution, screening size, and changes of the used chemicals did not interfere with the model calculation as long as the spectra of the target molecule and contaminant species contribute to the UV/Vis spectral measurement and differ in their spectral profiles. A broad variation of the different species concentrations and ratios in the data set was found to be preferred and can be achieved by analyzing different process solutions during selective precipitation or crystallization, washing, and redissolution process steps.

4.5 Conclusion

In this research project, multi-way chemometrics were successfully applied to three high-throughput (HT) screenings for the characterization of selective crystallization and precipitation processes. Supernatant samples were taken after crystallization in the first case study, and after precipitation, washing, and redissolution for the second and third case studies. Besides model proteins, different modalities, e.g. virus-like particles (VLPs), monoclonal antibodies (mAbs), were investigated. The recorded ultraviolet-visible light (UV/Vis) spectra of the samples of each case study were structured as a four-dimensional (4D) data set and preprocessed to eventually calculate one parallel factor analysis (PARAFAC) model per case study. The models of the first and second case studies were compared with quantitative reference data on specific concentrations and spectra of the purified species to test the model validity and to find general preprocessing and model parameters. This knowledge of the calculation parameters was used for the third study when only the spectrum of the purified target molecule could serve as a quantitative reference. The concentration profile was only validated with the qualitative sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE) analysis.

Without prior calibration, these models coupled with UV/Vis spectroscopy could quickly provide species spectra and concentration estimations for selective crystallization in chemically defined solutions or precipitation screenings in complex solutions. The calculated PARAFAC components were supposed to represent the various species present in the solution. Still, low-absorbing species or species with similar phase behaviors could not be described with a single model component per species as shown in the first case study. This bears the advantage of clustering species depending on their phase behavior and to better describe multiple impurity species in complex solutions with one model component per cluster. This said, only species which crystallize or precipitate at various precipitant concentrations can be distinguished.

With quantitative insights calculated from the concentration estimations, the generated models could visualize the influence of the precipitant on the different species. Thus, they could be used to evaluate the screened conditions in terms of purity and yield and could potentially find optimal process conditions in all three case studies.

When a suitable model component number was used, reasonable and valid models could be calculated regardless of the modality, screening scale, and other experimental parameters. This supports the assumption that the approach of coupling PARAFAC and UV/Vis spectroscopy can be transferred to other modalities and purification processes based on phase behavior. At an exploratory stage of process development, this approach can support process analytical technology (PAT) and it may be especially valuable as deeper process knowledge can be generated without refined analytics and with reduced input of resources. Different impurity clusters and the target molecule can be characterized regarding their differences in spectra and phase behavior. The PAT models estimated yield and purity and can be a basis for detailed process engineering. This process knowledge helps designing selective crystallization and precipitation processes and finding optimal process conditions while complying with the quality by design (QbD) guidelines and the high standard of biopharmaceutical processes.

Acknowledgment

The authors would like to thank Michel Eppink and Byondis B.V. for the material supply. Furthermore, the authors thank Prof. Adam Zlotnick for the provision of the VLP production plasmids. The authors are grateful to Annabelle Dietrich and Jan Tobias Weggen for proofreading the manuscript, Robin Schiemer for the fruitful discussions on the topic, and Ines Zimmermann for her laboratory assistance in the first case study and its experimental setup. We acknowledge support by the KIT Publication Fund of the Karlsruhe Institute of Technology.

5

Spectroscopic insights into multi-phase protein crystallization in complex lysate using Raman spectroscopy and a particle-free bypass

Christina Henriette Wegner[1], Sebastian Mathis Eming[1], Brigitte Walla[2], Daniel Bischoff[2], Dirk Weuster-Botz[2], and Jürgen Hubbuch[1]

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

[2] Chair of Biochemical Engineering, TUM School of Engineering and Design, Technical University of Munich, Germany

Abstract

Protein crystallization as opposed to well-established chromatography processes has the benefits to reduce production costs while reaching a comparable high purity. However, monitoring crystallization processes remains a challenge as the produced crystals may interfere with analytical measurements. Especially as a method for capturing proteins from complex feedstock containing various impurities, establishing reliable process analytical technology (PAT) to monitor protein crystallization processes can be complicated. In heterogeneous mixtures, important product characteristics can be found using multivariate analysis and chemometrics which contribute to the development of a thorough process understanding.

In this project, an analytical set-up is established combining off-line analytics, on-line ultraviolet-visible light (UV/Vis) spectroscopy, and in-line Raman spectroscopy to monitor a stirred batch crystallization process when there are multiple phases and species present. As an example process, the enzyme *Lactobacillus kefir* alcohol dehydrogenase (*Lk*ADH) was crystallized from clarified *Escherichia coli* (*E.coli*) lysate on a 300 mL scale in five distinct experiments, with the experimental conditions changing with regard to the initial lysate solution preparation method or the precipitant concentration. Since UV/Vis spectroscopy is sensitive to particles, a cross-flow filtration (CFF)-based bypass enabled the on-line analysis of the liquid phase providing information on the lysate composition regarding the nucleic acid to protein ratio. A principal component analysis (PCA) of *in situ* Raman spectra supported identifying spectra and wavenumber ranges associated with product-specific information and revealed that the experiments followed a comparable, spectral trend when crystals were present. Based on preprocessed Raman spectra, a partial least squares (PLS) regression model was optimized to monitor the target molecule concentration in real-time. The off-line sample analysis provided information on the crystal number and crystal geometry by automated image analysis, as well as the concentration of *Lk*ADH and host cell proteins (HCPs).

In spite of a complex suspension containing lysate and scattering crystals, and various impurities, it was possible to monitor the target molecule concentration in a heterogeneous, multi-phase process using spectroscopic methods. With the presented analytical set-up of off-line, particle-sensitive on-line and in-line analyzers, a crystallization capture process can be better characterized regarding the geometry, yield and purity of the crystals.

5.1 Introduction

Proteins, e.g. biopharmaceuticals, enzymes, and other biologically active molecules offer a wide range of therapeutic applications and have reinvented the treatment of various diseases and disorders. Essential to the success of biologics are efficient production, isolation, and purification using mostly chromatography as an expensive standard technique to ensure a high purity. Alternatively, other downstream processes, e.g., protein crystallization [127] or precipitation [163, 164], can be developed which are easier to scale, can achieve high purity and yield, and decrease production costs while maintaining high productivity. Whereas protein crystallization is traditionally associated with fundamental knowledge on the protein

structure, the application for formulation and purification reasons has drawn more interests in the past years reducing the number of process steps saving both time and resources in the production of biologics. With respect to formulation, crystalline suspensions are beneficial due to their lower viscosity at high product concentration [130], higher stability [131], and potentially controlled release properties [133].

Saturation is the primary cause behind crystallization [195] , and it is influenced by a variety of environmental factors, e.g. protein concentration [136, 195], pH [18, 136], precipitant concentration [25, 196, 197], or temperature [6, 195]. Compared to crystals of a chemical substance, the larger size of a biological molecule increases the complexity of the protein crystal. Therefore, extensive empirical screenings [2, 195], precise, automated high-throughput (HT) techniques [31, 36] and HT analytics [37, 181, 198, 199] are essential to find optimal process conditions.

In the past, protein engineering introduced the possibility to produce proteins with different abilities or processing properties, e.g. increased crystallizability and solubility behavior [200–202]. Especially increased crystallizability may make protein crystallization attractive for larger production scales due to its higher productivity and higher probability to form crystals [10]. For this purpose, research in micro-liter [10, 201, 202] and milli-liter scale [128, 203] has proven that protein crystal contacts in an enzyme can be improved to increase crystal occurrence and yield in pure protein solutions or clarified harvest leading to a high product purity. In practice, harvest broth from biotechnological processes involves mixtures of proteins, impurities, and only a small quantity of the target molecule. While a lot of research was reported on scaled-up protein crystallization in solutions containing only traces of impurities or even none [14, 32, 33, 204–208], the challenges imposed by complex solutions in capture processes have received relatively little attention [127, 128, 209, 210].

As suggested by the U.S. Food and Drug Administration (FDA) [115], PAT is crucial for ensuring the products safety to the patient, and the quality of a pharmaceutical manufacturing process. To accomplish this, the critical process parameters and quality attributes need to be controlled by selecting real-time analytics, and suitable variables. Possible real-time process analyzers are spectroscopic measurements which are commonly applied in biotechnological processes. A lot of PAT research focuses on the particle-sensitive UV/Vis spectroscopy [57, 103], water-sensitive fourier-transform infrared spectroscopy (FTIR) [83, 118] or process solutions with lower concentration [53, 57]. These restrictions impose challenges when multiple phases and heterogeneous mixtures need to be monitored in a protein crystallization process in an aqueous environment. Especially when particles are present possibly caused by aggregation, precipitation, or crystallization, PAT faces difficulties with the choice of an adequate process analyzer. Light scattering, the heterogeneity of the suspension, and size distribution may affect the measurement and need to be considered for the data analysis. Raman spectroscopy may be a possible solution to this problem as it was shown to monitor crystallization processes of chemical target molecules [84, 147, 208] or the enzyme lysozyme [211] out of pure component solutions [82]. When examining heterogeneous, more complex solutions, as e.g. in upstream processes [72, 76], Raman spectroscopy has demonstrated its suitability for process monitoring despite possible interferences from sample turbidity [66], stirring, temperature [64], or pH fluctuation [76]. In this context, the integration of Raman spectroscopy holds promise for

improving our understanding of protein crystallization processes in heterogeneous mixtures. As an alternative, UV/Vis spectroscopy is a promising analytical tool as it was implemented in-line with an attenuated total reflection (ATR) probe in pure solution crystallization processes [86, 138, 146, 204, 212] and is often used for strongly absorbing solutions. UV/Vis transmission measurements with variable pathlength (VP) technology are more flexible in terms of solution absorption and was used when molecule concentration varied during the process [56, 60, 62] similar to crystallization processes. However, this technique is prone to particle scattering and solid crystalline particles would interfere with the measurements.

Due to high correlation within the data set, the spectra produced by the above stated techniques are commonly processed using chemometric techniques, e.g. PCA [147], PLS [147] regression models, or gaussian process regression (GPR) [213], just to name a few techniques. Further explanation on chemometric methods can be looked up in published literature [90, 99, 165, 166]. Additional preprocessing of Raman spectra [214] improves the chemometric analysis and helps to reduce the complexity of the data set, extract essential information from spectral data, and remove spectral noise or unwanted experimental disturbances, particularly in situations involving multiple species and interferential effects. Regarding crystallization processes, crystallization processes of mostly chemical substances were monitored spectroscopically in the past using PCA [120], principal component artificial neural networks (PC-ANN) [145], principal component regression (PCR) and PLS [147, 208, 215], or multiple linear regression (MLR) [211]. For the purpose of PAT, there have been numerous attempts to monitor crystallization of chemical compounds in pure [85, 212] or relatively pure mixtures [208, 216, 217]. With respect to biologics, PAT studies investigating the crystallization process of the benchmark crystallization protein lysozyme in model protein solutions [204, 211] or of small, biological molecules with low levels of impurities [120] have been discussed before. To the best of the authors' knowledge, however, no research has been conducted developing PAT for protein crystallization as a capture step with larger biological targets in heterogeneous, complex mixtures, i.e. clarified lysate. The implementation of real-time monitoring would extend our understanding of protein crystallization in complex solutions and facilitate process control.

To find a suitable PAT set-up for protein crystallization in a heterogeneous mixture, this research project investigates different spectroscopic methods, their limitations, and possible implementation for the application to crystalline slurries. The molecule of interest is the enzyme *Lk*ADH and is crystallized from clarified lysate in a stirred vessel on lab-scale. To increase the variety of the recorded data sets, five batch experiments are conducted with varying crystallization conditions, namely the precipitant concentration, initial absorption value of the clarified lysate, and changes in the lysis protocol. An *in situ* Raman probe is immersed directly into the crystallization vessel, and records in-line spectra which are processed with chemometric methods to predict product characteristics, e.g. target molecule concentration in the liquid phase. An analytical bypass of the crystallization vessel is realized with a CFF-based set-up to facilitate the use of particle-sensitive analytics, i.e. UV/Vis spectroscopy with a VP flow cell. Microscopic imaging, *Lk*ADH and HCP quantification of off-line samples - and optionally redissolved crystals - assist in developing a comprehensive process understanding of the crystallization process in complex lysate. In short, the results

demonstrate how a protein crystallization PAT can be realized for process and product characteristics in a heterogeneous, complex solution where multiple phases are present.

5.2 Materials and methods

5.2.1 Experiment buffer and protein preparation

All chemicals were purchased from Merck KGaA (Darmstadt, DE), unless otherwise stated. The buffer solutions were prepared at room temperature with ultrapure water (PURELAB Ultra, ELGA LabWater, Lane End, High Wycombe, U.K.), pH-adjusted with 32 % hydrochloric acid (HCl) or 4 M sodium hydroxide (NaOH) and filtered using a 0.2 µm cellulose acetate (CA) membrane filter (Sartorius Stedim Biotech GmbH, Göttingen, DE).

The *Lk*ADH protein (wild-type (WT); protein data bank (PDB) ID: 7P36) was produced with *E.coli* BL21(DE3) in a fed-batch process in 1.5 L stirred-tank bioreactors (DASGIP, Eppendorf GmbH, Hamburg, DE) as described in Schmideder et al. [218]. The process is divided in three consecutive phases at pH 7.0: batch phase (5.0 g/L glucose, 4 h at 37 °C), exponential feeding phase (growth rate 0.15/h, 18 h at 37 °C), and protein production phase (500 µM isopropyl β-d-1-thiogalactopyranoside (IPTG), 3.0 g/(L h) glucose, 48 h at 30 °C). The harvested *E.coli* cells were kindly provided by the research group of Prof. Weuster-Botz. The cell pellets were further processed as described in Walla et al. [203], with variations listed in the Supplementary Table A5.1. The cell pellets were sonified in an ice bath by the sonifier SFX550 (Branson Ultrasonic Corporation, Danbury, US-CT, tapered microtip 101-148-062, 70 % amplitude, 40 s twice to three times with 50 % pulse and with 3 min breaks between each cycle. Cell debris were removed from the supernatant by centrifugation at 4 °C with 17 418 rcf for 1 h and by filtration with a glass fiber, a 0.45 µm, and a 0.2 µm CA syringe filter (Sartorius Stedim Biotech GmbH, Göttingen, DE). After dialysis (SnakeSkin™ dialysis tube, ID 34 mm, 3.5 kDa molecular weight cut-off (MWCO), Thermo Fisher Scientific, Inc., Waltham, MA) the filtered supernatant to the protein buffer (20 mM 4-2-hydroxyethyl-1-piperazineethanesulfonic acid (HEPES), 1 mM magnesium chloride ($MgCl_2$) at pH 7.0), the initial absorption value A_{initial} at 280 nm of the clarified lysate was adjusted with protein buffer according to Table 5.1 using a NanoDrop™ 2000 spectrometer (Thermo Fisher Scientific, Inc.).

The crystallization buffer was a 100 mM tris(hydroxymethyl)aminomethane (Tris), 50 mM $MgCl_2$ buffer with a varying polyethylene glycol monomethyl ether 550 (PEG MME 550) concentration depending on the experiment (see Table 5.1). The redissolution (RD) buffer was a 20 mM HEPES, 2 M $MgCl_2$ buffer at pH 7.0. The required buffers for the immobilized metal ion affinity chromatography (IMAC) analysis contained 50 mM phosphate, 500 mM sodium chloride (NaCl) and 20 mM imidazole for the equilibration or 200 mM imidazole for the elution buffer (both pH 7.0).

5.2.2 Protein crystallization experiment

The batch crystallization process was initiated in a 300 mL jacketed glass vessel (CG-1929-X11) equipped with an overhead stirrer with a stirrer speed of 80 rpm (CG-2024-10, both provided by Chemglass Life Sciences, Vineland, US-NJ, anchor style stir paddle) by placing 150 mL clarified lysate in the vessel and adding 150 mL crystallization buffer. The initial crystallization conditions varied according to Table 5.1 for the five conducted experiments Exp1 - Exp5. After 30 to 90 min the vessel content was centrifuged (15 min, 3225 rcf) to remove initial HCP and nucleic acid precipitate. The supernatant was placed in the glass vessel, the experiment continued and the target molecule crystallized after the centrifugation step.

Table 5.1 Crystallization conditions, HCP content and crystal yield: The variations of the crystallization conditions are the initial absorption at 280 nm A_{initial}, $c_{\text{PEG, initial}}$, and the number of lysis cycles. The experiments were performed with or without the analytical bypass. The HCP content of the first sample is compared with the content of the washed and redissolved crystals. To account for differences in dilution, the HCP content was normalized to the target molecule concentration. The crystal yield is estimated by the ratio of the initial to equilibrium *Lk*ADH concentration in the supernatant and derived from the IMAC analysis (see section 5.2.3.3).

exp.	**crystallization conditions**			bypass	**HCP removal**			crystal yield in %
	A_{initial} at 280 nm in AU/cm	$c_{\text{PEG, initial}}$ in g/L	number of lysis cycles		$c_{\text{HCP, initial}}$ in µg/L	$c_{\text{HCP, RD}}$ in µg/L	HCP reduction factor	
Exp1	20.0	200	3	w/	53460	n.d.	n.d.	46.6
Exp2	20.5	250	2	w/	27412	358	77	77.9
Exp3	20.8	300	2	w/	42438	n.d.	n.d.	77.9
Exp4	28.5	200	2	w/	38644	n.d.	n.d.	27.9
Exp5	20.5	250	2	w/o	42982	437	98	76.4

abbreviations - HCP: host cell protein;PEG: polyethylene glycol; RD: redissolution;
n.d.: not determined; w/: with; w/o: without

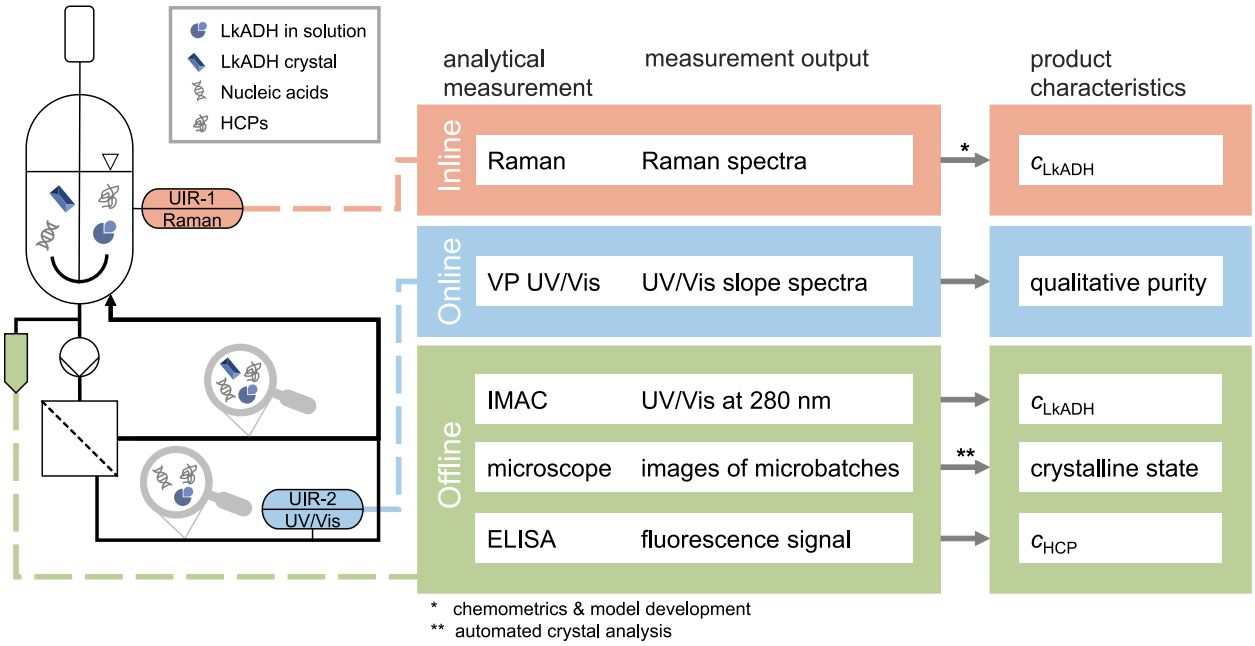


Figure 5.1 Experimental and analytical set-up of the protein crystallization experiments as a scheme. The desired product characteristics are listed on the right and paired up with the respective analytical measurement method and output. The vessel contains the clarified lysate containing HCPs, nucleic acids, the molecule of target *Lk*ADH, and later during the process *Lk*ADH crystals while the installed Raman probe records in-line spectra. The CFF-based set-up facilitates the solid-liquid separation which makes the implementation of an on-line VP UV/Vis measurement in the permeate stream possible. Both, retentate and permeate stream are directed back to the vessel. All off-line samples are analyzed with IMAC, and microscopic imaging. Selected samples are further analyzed with automated enzyme-linked immunosorbent assay (ELISA) to determine the HCP content.

5.2.3 Analytics

The following section describes the PAT set-up to monitor protein crystallization using in-line Raman spectroscopy, a filtration-based on-line UV/Vis set-up and off-line samples. The set-up of the different analytics is visualized and listed in Figure 5.1.

5.2.3.1 In-line Raman spectroscopy

To monitor the crystallization process by in-line Raman spectroscopy, a MarqMetrix Bioreactor Ballprobe (MarqMetrix®, Seattle, US-WA) was immersed in the crystallization suspension and connected to a HyperFlux™ PRO Plus 785 Raman analyzer with Spectralsoft 3.3.600.1 (Tornado Spectral Systems, Mississauga, CA). The measurement was performed with a laser power of 495 mW at the laser wavelength 785 nm and an exposure time of 8553 ms averaging 15 spectra every 12 min between the Raman shift range 200 to 3300 cm^{-1}.

5.2.3.2 Analytical bypass and on-line analytics

As an analytical bypass, CFF-based set-up was installed to separate the solid precipitate and crystals from the supernatant to facilitate the implementation of particle-sensitive devices and avoid their possible blockages by solid particles.

A KrosFlo Research KRIIi CFF system was equipped with an automatic backpressure valve, pressure transductors (all Spectrum Labs, Rancho Dominguez, US-CA) and CFF membrane (modified polyethersulfone (mPES), 0.2 µm pore size, 13 cm^2 surface area, C02-P20U-10-N, Spectrum Laboratories, Inc., Rancho Dominguez, US-CA). The feed flow rate and the desired transmembrane pressure (TMP) were set to 20 mL/min and 0.05 bar, respectively. Overnight the bypass was switched off and the bypass suspension was pumped into the crystallization vessel. Subsequently, the bypass and the membrane were cleaned with water at 40 °C. The liquid flow meter SLS-1500 (Sensirion AG, Stäfa, CH) was installed at the permeate plug in the analytical bypass and recorded the permeate flow averaged over a time range of 5 s. As UV/Vis spectroscopy is sensitive to larger particles and light scattering, the on-line FlowVPE flow cell (C Technologies, Inc., Bridgewater, US-NJ) with a Cary 60 spectrometer (Agilent Technologies, Inc., Santa Clara, US-CA) was implemented in the analytical bypass and measured the UV/Vis absorption slope spectra in the permeate flow between 220 to 400 nm.

5.2.3.3 Off-line analytics

Off-line samples were taken during the crystallization process through an injection plug (Fresenius Kabi AG, Bad Homburg, DE) in the feed flow. For visual crystal detection, suspension samples were 10 times diluted to prevent proceeding crystallization. For the supernatant analysis, the samples were centrifuged (2 min, 12 000 rcf) and the diluted supernatants (2 times) for IMAC and ELISA were stored at −20 °C until analysis. Grown crystals were redissolved by removal of the supernatant after centrifugation, washing with protein buffer, a second centrifugation step, redissolving in RD buffer and a third centrifugation step. The centrifugation procedure is described above.

For visual inspection of the crystalline suspension, 24 µL-quadruplicates of the undiluted and 10 times diluted suspension were placed a MCR Under Oil Crystallization Plate (Hampton Research, Aliso Viejo, CA), sealed with a transparent foil (Shurtape Technologies, LLC, Hickory, US-NC) and imaged using a tempered microscopic system (RI 54, FORMULATRIX LLC, Bedford, US-MA, T 1000 mytron Bio- und Solartechnik GmbH, Heiligenstadt, DE) at 20 °C. As the sampling time for the microscopic imaging was less than 20 min and is short compared to the protein crystallization time, crystal nucleation or growth in the static micro-batch samples is not expected. Next to manual, visual inspection, a machine learning (ML) model based on augmented, synthetic images of crystals [219] counted and measured the crystal height and width to detect crystals objectively and automatically. Images were treated as outliers and taken out of the analysis when they were out of focus or showed large bubbles. Using the model as a basis, the following small adaptions were applied to adjust the detection method to the setup used in the presented experiments. The border due to the circular well

geometry was removed from the microscope images by cropping the image to the central region with a size of 1400x1200 pixels. Furthermore, large-area false positive detections could be eliminated by applying a threshold for the maximum crystal size of 10^3 px^2. The confidence threshold for accepted detections was set to 0.2. Finally, inference was performed on a GTX 1080 GPU.

The IMAC analysis was performed as a reference for the *Lk*ADH concentration c_{LkADH} as the target molecule contained a His-tag. A TSKgel® Chelate-5PW column (Tosoh Corporation, Shiba, JA) with a pre-column filter (0.2 µm, OPTI-SOLV EXP, Supleco™, Bellefonte, US-PA) was installed in a Dionex Ultimate 3000 RS high-performance liquid chromatography (HPLC) system (Thermo Fisher Scientific, Inc.) equipped with a diode array detector. The supernatant samples were thawed, filtered with a AcroPrep™ Advance filter plate (3.0 µm glass fiber/0.2 µm Supor® membrane, Pall Corporation, Port Washington, NY). Either 20 µL supernatant samples or 40 µL of the redissolved crystals (filtered as above) were analyzed with a two step elution protocol at 100 mM and 200 mM imidazole eluting loosely bound impurities and the target molecule, respectively (see Supplementary Material A5.1). The absorption was used to quantify *Lk*ADH. The elution peak absorption and the extinction coefficient 0.8596 AU $*$ L/(g $*$ cm) (derived from the web-tool ProtParam [220]) at 280 nm and was used to quantify *Lk*ADH.

The HCP concentration of selected supernatant and redissolved crystals samples was determined using the Gyrolab XPlore station with its software Gyrolab Control 7.0.3.133 (Gyros Protein Technologies AB, Uppsala, SE) following the manufacturer's protocol and used to evaluate the HCP removal by *Lk*ADH crystallization and RD.

5.2.4 Data analysis

Data analysis including spectral preprocessing, model calibration, and data plotting was performed in MATLAB, R2019b (The MathWorks, Inc., Natick, MA). To contrast different sampling approaches, the Kennard-Stone (KS) data split algorithm [221, 222] and a manual data split approach were tested for model validation. Spectral preprocessing for Raman spectra were implemented to highlight significant spectral features which can then be correlated to the desired process parameter to enable PAT.

The *mdatools* toolbox [223] was applied on the Raman spectra using the baseline correction with asymmetric least squares (smoothness 10000, penalty value 0.01). The spectra were treated with the Savitzky Golay (SG) filter (Savitzky and Golay, 1964, KS data split: window size 29, 2$^{\mathrm{nd}}$ derivative, manual data split: window size 17, 1$^{\mathrm{st}}$ derivative) and cut to the Raman wavenumber regions 300 to 490, 750 to 1040, 1210 to 1320, and 1600 to 1640 cm^{-1}. The optimal parameters for the preprocessing, namely window size and SG filter derivative, and model calibration, i.e. number of latent variables, were optimized using a genetic algorithm (GA). For details on the methodology, the authors refer to Andris et al. [103].

The Raman spectra of all experiments were baseline-corrected and analyzed with the unsupervised learning method PCA to reduce the dimensionality of the data set and visualize correlation between the spectra and the crystallization process. The PLS regression model, as a supervised learning method, was employed to predict the concentration of the target

molecule in the supernatant. As a first step, the Raman spectra closest to the sampling time were selected and grouped into a calibration and a validation data subset consisting of 29 and 5 experimental samples, respectively. To assess the impact of two distinct data splitting methods on the model prediction, the samples were initially divided using the KS algorithm which selects a representative data subset from a larger data set. As an alternative, manual data split approach, the samples of Exp5 were selected as the external validation set to examine the model predictability on new experiments and batch-to-batch variations. Finally, the *Lk*ADH concentrations obtained from the IMAC analysis were regressed against the preprocessed spectra of the calibration subset via leave-one-out cross-validation. The model calibration procedure with KS or manual data splitting resulted in 8 or 10 latent variables, respectively.

5.3 Results

5.3.1 Off-line: Image analysis, *Lk*ADH and HCP quantification

In this project, the crystal yield is estimated by the decrease from the initial to the equilibrium concentration of the target molecule and can be used to evaluate and compare processes. Furthermore, the HCP concentration of the initial solution and the redissolved crystals was determined and normalized to the target molecule concentration providing information on the purity and HCP removal for this process. These values and the experimental conditions are listed in Table 5.1 and demonstrate a 77-fold, and 98-fold HCP removal in Exp2 and Exp5 while achieving a yield of 77.9 % and 76.4 %.

The mean of detected crystals per off-line sample of the five conducted experiments are depicted over time with their standard deviation in Figure 5.2 in (A-E). The light green and gray shaded areas indicate off-line samples in which crystals of larger size were visible and micro-crystals were assumed as the latter are difficult to detect due to the image resolution. For Exp1 and Exp4, the mean count of detected crystal fluctuates between 200 to 500 whereas Exp2 and Exp5 start with detected crystal counts less than 100 and increase after 20 h to values above 1000. The crystal count of Exp3 rises after 6 h to values above 400. A trend towards lower crystal detection points in time with increasing polyethylene glycol (PEG) concentration is visible from left to right. The largest number of detected crystals were achieved in Exp2 and Exp5 which also demonstrated slightly higher crystal heights (see Supplementary Material Figure A5.3). In two exemplary microscopic images of the same off-line sample of Exp2 after 19.7 h, the ML-based model detected crystals are highlighted in (F, G) in Figure 5.2. The microscopic images differed in the dilution factor to account for high crystal densities and reduce overlapping crystals. As not all visible crystals are highlighted in Figure 5.2 (F, G), the image analysis tool is used in this project complementary to the manual inspection as an objective, qualitative tool to narrow down the crystal induction time and to provide insight into the crystal geometry.

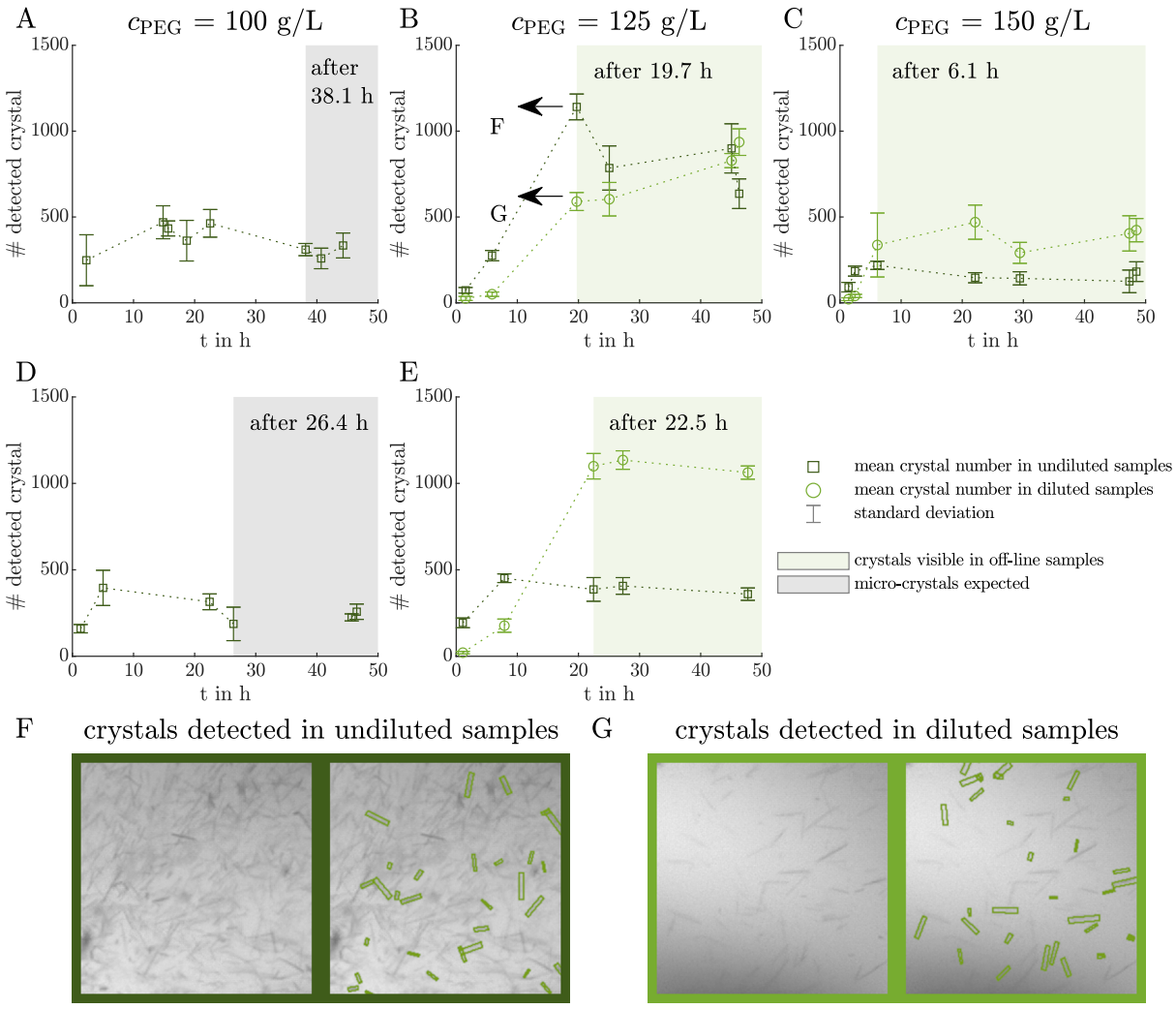


Figure 5.2 Counted crystals in microscopic images from off-line samples. The microscopic images of off-line samples are analyzed with a ML-based image analysis tool [219] counting the crystals and providing information on the crystal geometry (see Supplementary Material A5.3). The mean crystal count per imaged well, and its standard deviation of undiluted and diluted off-line samples are visualized over the experimental time with dark green squares and light green circles with dotted lines to guide the eye and with their respective error bars. The off-line samples with micro-crystals present are shaded in gray as they are difficult to detect due to the image resolution (A, D). The off-line samples showing larger crystals are shaded in a light green box (B, C, E). Exemplary, the results of the automated image detection are shown for an undiluted (F), and a diluted (G) off-line sample of Exp2 after 19.7 h.

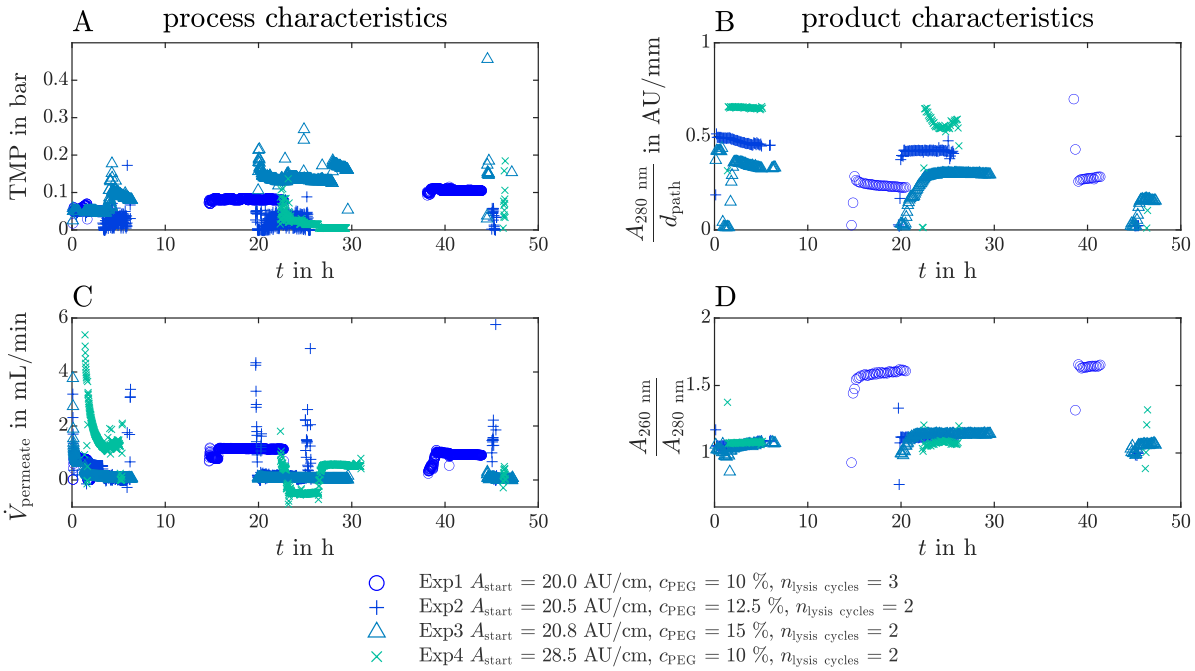


Figure 5.3 Performance and application of the analytical bypass. The analytical bypass can be characterized by the measurement of the TMP (A) or permeate flow rate $\dot{V}_{\mathrm{permeate}}$ (C) over time for the five conducted experiments. For clearer visualization, TMP and the $\dot{V}_{\mathrm{permeate}}$ are averaged with a moving mean over each minute and only one value per minute is shown as the data were recorded with a high frequency. The missing data are caused by the fact that the bypass was switched off overnight. Spikes in the recorded flow rate are artefacts from starting or turning off the analytical bypass. The recorded VP UV/Vis slope $\frac{A_{280\mathrm{nm}}}{d_{\mathrm{path}}}$ (B) and the $\frac{A_{260\mathrm{nm}}}{A_{280\mathrm{nm}}}$ (D) ratio are shown over time. The different experiments are visualized with different markers and four different shades of blue.

5.3.2 On-line: Analytical bypass and UV/Vis spectroscopy

The analytical bypass was installed to make the implementation of particle-sensitive analytics feasible in crystallization processes. The bypass characteristics and the results of the VP UV/Vis spectroscopy are depicted in Figure 5.3 over time for Exp1 to Exp4 as Exp5 did not have a bypass installed. The different colored markers each represent one experiment differing in the crystallization conditions (see Table 5.1). The TMP over the CFF membrane in Figure 5.3 (A), and the flow rate of the permeate stream in Figure 5.3 (C) can help to evaluate the reliability of the on-line sensor implemented in the analytical bypass as the sensor can only measure reliably if the solution in the bypass represents the current particle-free vessel content. The absorption at 280 nm $A_{280\mathrm{nm}}$ is derived with respect to the path length d_{path} and shown over time t as $\frac{A_{280\mathrm{nm}}}{d_{\mathrm{path}}}$ in Figure 5.3 (B). The ratio between the absorption values at 260 nm and 280 nm is depicted over time as $\frac{A_{260\mathrm{nm}}}{A_{280\mathrm{nm}}}$ in Figure 5.3 (D)).

The TMP for all experiments was mainly below 0.15 bar and remained constant for several hours on each day. Outliers are visible in all experiments which occurred when the bypass was blogged requiring manual blockage removal. Due to potential overnight tube blockage and damage on the devices, the bypass was switched off overnight which explains the missing data during nighttime. Exp2, Exp3, and Exp4 do not demonstrate stable TMP values on the third day and, consequently, the bypass was switched off. The start and stop time varied as the experiments started at different time points during the day.

The flow rates in the permeate stream in all experiments drop from a value between 3 to 6 mL/min to a value of 0.1 to 1.1 mL/min within the first four hours of each experiment. On the second and third day, the flow rate of Exp1 and Exp3 remain on a constant level between 0.01 to 1.1 mL/min. The flow rate values for Exp2 and for Exp4 fluctuate between −0.5 to 0.5 mL/min.

The absorption slope $\frac{A_{280\text{nm}}}{d_{\text{path}}}$ can indicate changes in the concentration of UV/Vis absorbing material in the supernatant. After switching on the analytical bypass, the absorption slope data required between 0.5 to 4 h to stabilize to a constant value. On the first day within the first 7 h the absorption slope decreases during the Exp2 and Exp3 whereas Exp4 does not show decreasing absorption values. The absorption data of Exp1 were not recorded. The absorption slopes of Exp1 to Exp3 experiment stabilized after 2 to 4 h on the second day. The third day shows stable values in Exp1, Exp2 and Exp3 whereas the absorption values of Exp4 did not stabilize due to tube blockage.

The absorption ratio $\frac{A_{260\text{nm}}}{A_{280\text{nm}}}$, as an indicator for nucleic acid and protein content, stabilized to values around 1 in the case of Exp2 to Exp4. The highest $\frac{A_{260\text{nm}}}{A_{280\text{nm}}}$ ratios of 1.15 were achieved on the second day. Analogous to the absorption slope at 280 nm in Figure 5.3 (B), the ratios stabilized on the second and third day after switching on the analytical bypass but required less time. The ratio in Exp1 was higher around 1.6 and was the only experiment which included three lysis cycles.

5.3.3 In-line: Raman spectroscopy and exploratory analysis

To monitor the stirred batch crystallization process, a Raman probe was installed in the vessel and recorded in-line spectra over time. Spectral preprocessing is advised to enhance spectral differences and remove baseline drifts, background signals, or detector noise. Generally, several different techniques are tested to find a matching set of preprocessing steps in the most cases, e.g. baseline correction, background subtraction, normalization, centering. Figure 5.4 shows the effects of the preprocessing steps on the spectra later used for the regression model. All recorded spectra are preprocessed and are visualized in gray to black color with one arbitrary spectrum in orange to better visualize the preprocessing effects on one exemplary spectrum. The raw spectra (see Figure 5.4 (A)) are baseline-corrected (see Figure 5.4 (B)), and treated with a SG filter, and 2$^{\text{nd}}$ derivative for the KS or 1$^{\text{st}}$ derivative for the manual data split (see Figure 5.4 (C, D)). The selected wavenumber ranges for the PLS regression model development are illustrated with gray shaded boxes in (see Figure 5.4 (C, D)). The selection of preprocessing steps reduce the baseline drift, which is visible in Figure 5.4 (A, B), align the spectra and help to increase spectral differences. The calculation of the derivatives

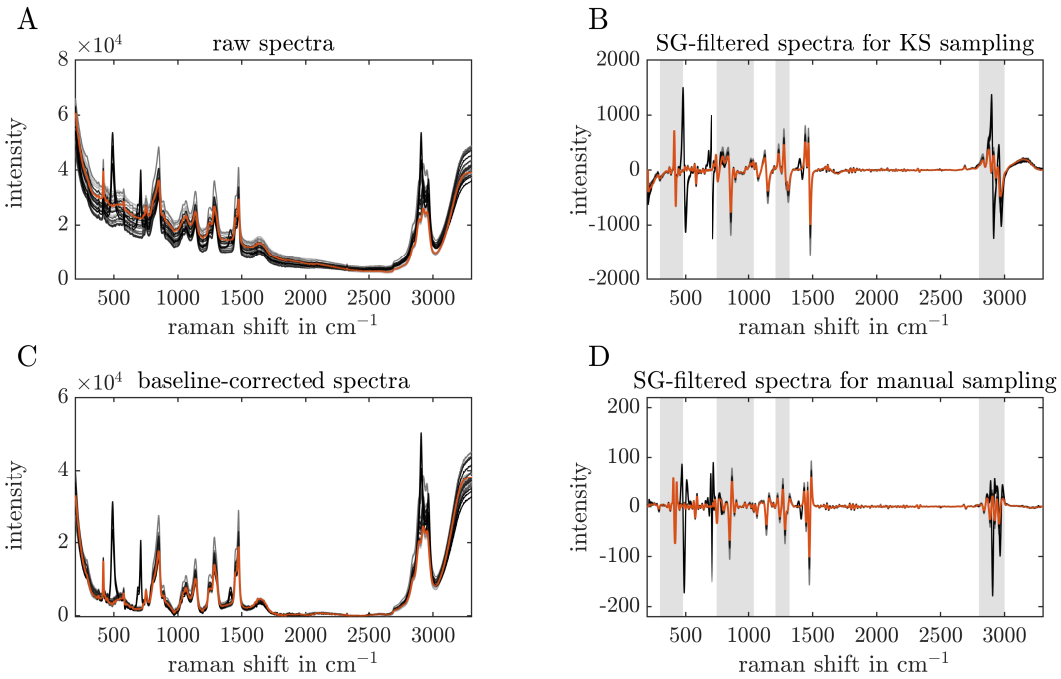


Figure 5.4 Preprocessing of Raman spectra. The raw Raman spectra, used for the regression modeling, are shown over the recorded wavenumber range from gray to black to visualize the time (A). The orange line represents one specific spectrum to better visualize the effect of the preprocessing effects. Preprocessing techniques, namely the baseline correction (B), and the application of the SG filter for the KS (C) or manual data split (D), are applied to enhance spectral differences. The gray boxes in (C, D) depict the Raman shift ranges that are used for the PLS model development.

emphasizes peak shifts in the examined spectra near 790 cm^{-1}, 1260 cm^{-1}, or 2970 cm^{-1}. Beforehand, different normalization, derivative and baseline correction methods were tested, but did not improve the interpretability of the data. To demonstrate the preprocessing effects on the experimental data, a zoom into the selected wavenumber regions of Exp3 is included in the Supplementary Material A5.5 as an example.

A PCA analysis of a large data set can aid to visualize trends and cluster observations in groups, and was performed on the preprocessed Raman data of all 5 experiments in this study using the whole spectral range (200 to 3300 cm^{-1}). Figure 5.5 depicts PC2 over PC1 of the PCA and each subfigure depicts the spectra of one experiment. The colors blue, orange, and yellow represent observations before the centrifugation step (see section 5.2.2), before and after the first crystals were detected in the microscopic images in the off-line examined samples. The PCA loadings can be found in the Supplementary Material A5.6. With passing time of the crystallization experiment, PC1 decreases whereas PC2 increases as indicated by the arrows. The arrows demonstrate a comparable slope when all observations are visualized in one diagram (figure not shown). The experiments Exp2 and Exp5 show

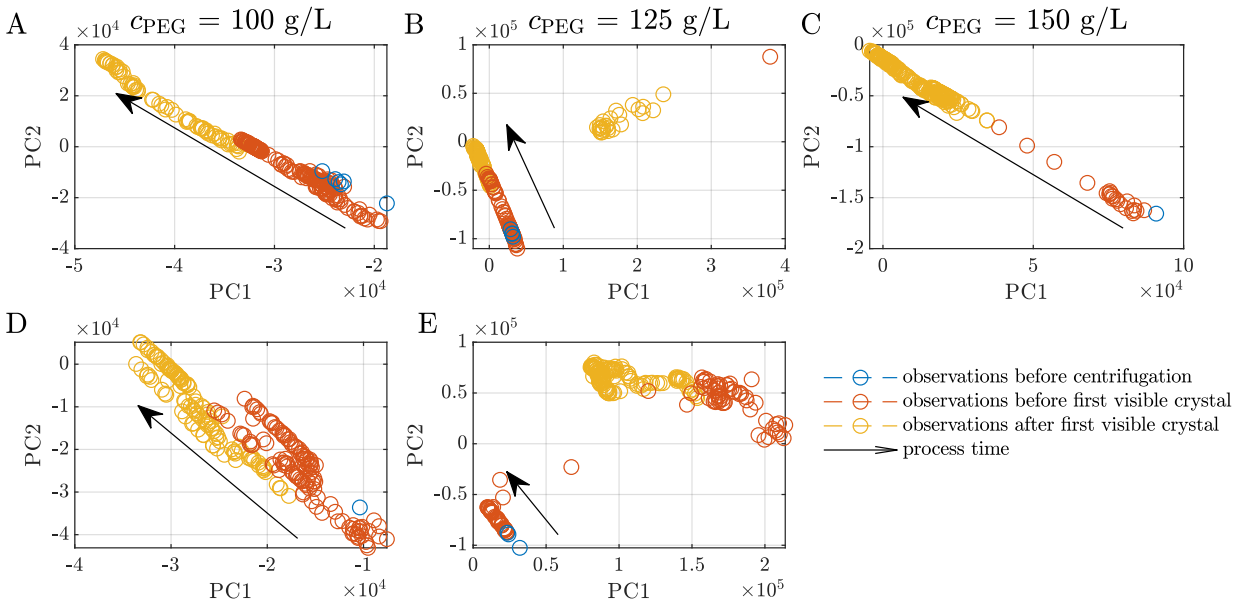


Figure 5.5 PCA scores of Raman spectra. The scores of the 1st & 2nd principal component (PC) are identified as PC1 and PC2 and are shown for the five conducted experiments (A-E). The observation of each experiment are classified by investigation of the off-line microscopic images. Observations before the initial centrifugation step, after the centrifugation step until the first, visual occurrence of crystals and after the first detected crystals are shown in blue, orange, and yellow, respectively.

two clusters. The left clusters follow the direction of the arrows. The right clusters do not follow the same direction and could be traced back to the irregular peak appearances which can be seen in Figure 5.4 near 493 cm^{-1}, 708 cm^{-1}, 1410 cm^{-1}, 2909 cm^{-1}, and 2970 cm^{-1}. Among the five experiments, the observations of Exp4 stand out as they follow the direction of the marked arrow, but are more widely scattered. Exp3 observations move quickly from observations in the lower right to the upper left corner of Figure 5.5 (C). The observations before crystallization was detected in the off-line samples are located at the end of the illustrated arrows whereas the observations after crystallization are located near the tip of the arrows. Further inspection of the preprocessed spectra over time showed that the changes before and after crystallization are visible in the spectra by gradually reduced peak heights (data not shown).

5.3.4 PLS model development and application on protein concentration monitoring

For the development of a PLS model, the preprocessed spectra are regressed on the off-line measured concentration c_{LkADH} from the IMAC analysis. Then, the developed model

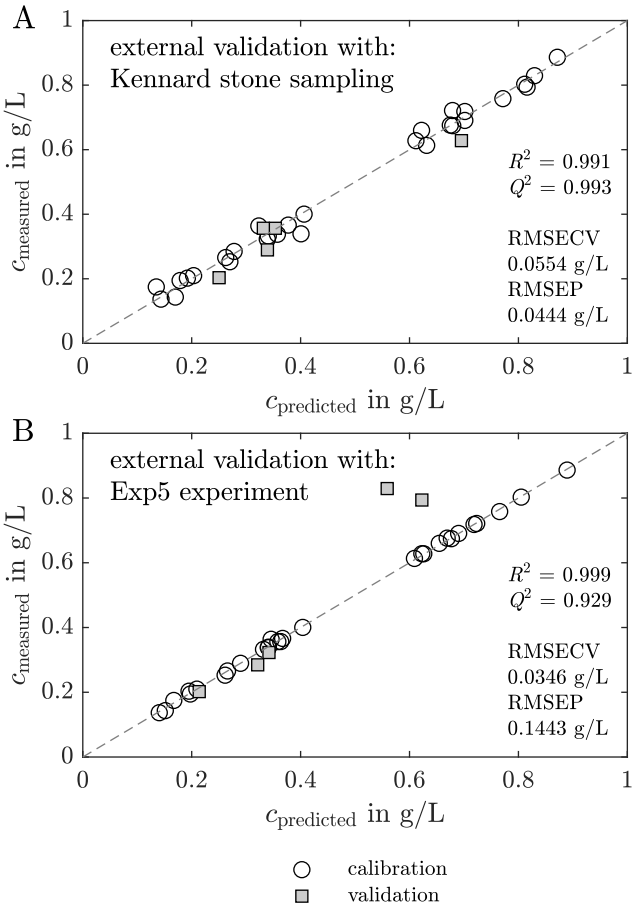


Figure 5.6 Chemometric regression model on Raman spectra and effects of validation sampling techniques. The preprocessed Raman spectra are regressed on the IMAC derived *Lk*ADH concentration with PLS models. Two models differing in the choice of the validation data set are compared. The white circles, gray squares, and dashed line represent the calibration, validation data, and theoretical values, respectively. First, the measured over model-predicted concentrations are visualized in (A) for a model with KS data split. Analogous to that, the measured over model-predicted concentrations are shown in (B) for a model where Exp5 was chosen manually as the validation data set. High coefficient of determination (R^2) andpredictive relevance (Q^2), and low root mean squared error of cross-validation ($RMSECV$) and root mean squared error of prediction ($RMSEP$) values indicate an applicable model.

is applied on all spectra which were recorded during the batch experiments to assess its potential to monitor real-time concentrations of the target molecule.

Figure 5.6 shows the results of two separately calculated PLS model which differed only in the choice of the data split for the external validation. The KS algorithm chooses the external validation samples according to the uniform distribution within the data set. For the model with KS data split, the measured over predicted concentrations of the target molecule

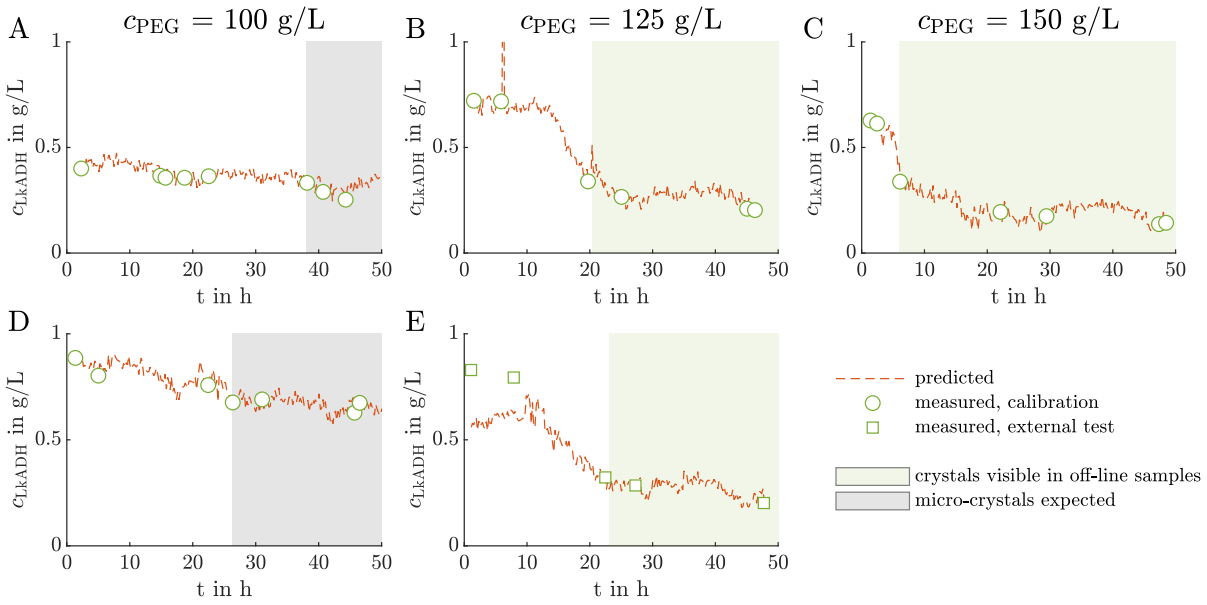


Figure 5.7 PLS model application on crystallization processes out of clarified lysate. The PLS model calculated with the manual data split predicts the *Lk*ADH concentration on the basis of the in-line recorded Raman spectra in orange for the five conducted experiments (A-E). Off-line *Lk*ADH calibration and validation concentrations are calculated from the IMAC analysis and are depicted with green circles and squares, respectively. Analogous to Figure 5.2, the light green boxes (B, C, E) indicate the time range when crystals are expected in the crystallization vessel as crystals are detected in the microscopic images in the off-line samples. The off-line samples in the time range illustrated with the gray boxes (A, D) showed only micro-crystals which were difficult to distinguish from precipitate visually.

are shown in Figure 5.6 (A). The white circles and gray squares represent the calibration and external validation set, respectively. For the second model, Exp5 was manually selected as the external validation set to evaluate the PLS model transferability to new experiments. The measured over predicted concentrations of the second model with the manual data split are illustrated in Figure 5.6 (B).

In the case of both models, the calibration data fit the dashed line, representing the a suitable model, well indicating that the preprocessed spectra and c_{LkADH} correlate in the calibration data set. The external validation data set chosen with a KS data split fits the ideal line overall very well with maximum discrepancies of 0.07 g/L (see Figure 5.6 (A)). The external data of the PLS model with the manual data split fits the ideal, dashed line below concentrations of 0.4 g/L well, but two outliers are visible at higher concentrations around 0.8 g/L with maximum deviations of 0.27 g/L. PLS model metrics, i.e. R^2, Q^2, $RMSECV$, and $RMSEP$, are added to the Figure 5.6 (A, B).

The PLS model with manual data split is discussed in more detail in this section to emphasize the importance of data splitting whereas the results of the PLS model calibrated with the KS data division can be found in the Supplementary Material A5.7. The model predicted and measured *Lk*ADH supernatant concentration of the five experiments are visualized in Figure 5.7 from (A-E). Analogous, the application of the PLS model with KS data split on the crystallization process can be found in the Supplementary Material A5.7.

The white circles, the gray squares, and the orange lines represent the off-line measured calibration concentrations, validation concentrations, and the in-line model-predicted concentrations derived from the Raman spectra, respectively. The concentrations c_{LkADH} are shown over the time t with the start time being the moment when the crystallizing solution was added to the crystallization vessel. The gray shaded box indicates the time slot when crystals were assumed in the crystallization vessel as the micro-crystals or crystals of a larger size were visible in the microscopic images of the off-line samples. The predicted data demonstrates fluctuations and outliers of up to 1.2 g/L deviation in Figure 5.7 (B) after 6 h. The predicted *Lk*ADH supernatant concentration decreases, notably, in Figure 5.7 (B, C, E) after 13 h, 5 h, and 10 h to 0.23 g/L, 0.18 g/L, and 0.24 g/L, respectively. After that, the concentration stays on the mentioned level fluctuating by 0.05 g/L. Exp1 and Exp4 do not show a steep decrease in supernatant concentration, but a gradual decrease by 33 % and 28 % till the end of the experiment, respectively.

Comparing the model-predicted concentration values of the Figures 5.7 (B, C, E), a *Lk*ADH concentration drop from the initial to the equilibrium concentration is clearly visible. The crystallization induction time of the experiments Exp2 and Exp5 with similar absorption value at 280 nm and PEG concentration required 12 h and 10.5 h until the *Lk*ADH concentration decreased (see Figures 5.7 (B, E)). The concentration decrease of Exp3, conducted with the highest PEG concentration of 150 g/L, is visible after 4.5 h. The light green shaded area starts after the aforementioned concentration drop and indicates the area when crystals were present in the microscopic images of the off-line samples indicating that the concentration drop is caused by protein crystallization. As the concentration drop in Exp1 and Exp4 are not pronounced, the time till the micro-crystals are first detected in the off-line samples can be used to compare the experiments. Micro-crystals were detected 11.7 h earlier in Exp4 than in Exp1 as the latter started with a lower 280 nm absorption value of the lysate and, thus, a lower, initial *Lk*ADH concentration influencing the supersaturation. Comparing the time till the first (micro)-crystals were visible, a trend towards lower points in time with increasing PEG concentration becomes apparent.

5.4 Discussion

In this work, the implementation of an analytical bypass for particle-sensitive analytics, as well as the implementation of an in-line Raman probe are discussed for batch protein crystallization process monitoring in real-time with the focus on their applicability and limitations. Employing the developed PAT and additional off-line analytics, the protein crystallization process itself can be assessed.

5.4.1 Analytical bypass and UV/Vis spectroscopy

To implement particle-sensitive analytics in a crystallization process, the analytical bypass consisted of a CFF-based set-up and enables monitoring the crystallization supernatant free of crystal particles. Inspecting Figure 5.3 (A, C) the recorded TMP and permeate flow rate showed irregularities and spikes which could be solved by stopping the CFF pump and cleaning the tubes manually. It is assumed that crystals or precipitated impurity blocked the membrane leading to a varying TMP levels in the experiments from day to day. Permeate flow rates at a constant level indicate that the bypass and the implemented sensor were filled with material which is representative of the liquid phase in the crystallization vessel. The comparison of Exp2 and Exp5 provides insight into the effects of the bypass on the crystallization process as these experiments only varied in the implementation of the bypass. Crystal breakage due to the CFF-based set-up can be excluded as the number, width, and height of the crystals were not reduced (see the Figure 5.2, and Supplementary Material A5.2 and A5.3). Other process characteristics, e.g. yield and purity, were not influenced as both experiments demonstrated comparable yields and a high HCP reduction factor (see Table 5.1).

Different levels of the initial UV/Vis absorption slope at 280 nm are visible in Figure 5.3 (B) for Exp2, Exp3, and Exp4. A decreasing trend of the absorption slope is noticeable from day to day, but it does not directly correlate with the decreasing *Lk*ADH concentration in the supernatant derived from the IMAC analysis in the off-line samples (data not shown). Note that impurities, e.g. nucleic acids or HCPs, were present (see Figure 5.3 (D) and $c_{\text{HCP, initial}}$ in Table 5.1) and absorb at 280 nm [44, 49] which complicates a direct measurement from the absorption at the selected wavelength. In Figure 5.3 (D) Exp1 demonstrates an increased $\frac{A_{260\text{nm}}}{A_{280\text{nm}}}$ ratio of 1.6, which indicates a higher content of nucleic acids [51], compared to the other experiments with a $\frac{A_{260\text{nm}}}{A_{280\text{nm}}}$ ratio around 1.0. The increased lysis cycle number may be the reason for this observation as more nucleic acids were released during a longer lysis duration and higher energy input. The slight increase of the $\frac{A_{260\text{nm}}}{A_{280\text{nm}}}$ ratios from the first to the second day may be caused by crystallized protein which leads to a higher impurity proportion in the liquid phase. The decreasing $\frac{A_{260\text{nm}}}{A_{280\text{nm}}}$ ratio on the third day can be an effect of the insufficient permeate flow in the bypass.

In the past, Smejkal [224, Chapter 4.2] used a similar CFF-based set-up to take samples during a crystallization process automatically, but required sample dilution when the UV/Vis absorption value exceeded the detector saturation. The VP technology circumvents the additional dilution step and allows automated UV/Vis analysis in real-time when the analytical bypass is switched on. [204] described a different approach to implement UV/Vis spectroscopy as monitoring PAT for a pure lysozyme crystallization process using an ATR probe directly placed in the crystal slurry. This approach bears the difficulty that the ATR technology is limited to applications with strongly absorbing or highly concentrated solutions [63] and cannot adjust to concentration changes as the VP technology. Furthermore, the real-time concentration using the 280 nm absorption from the ATR UV/Vis spectroscopy could not be determined during the crystallization process because particle scattering obstructed the measurement [225] as soon as small crystals were formed [204]. These challenges can be

tackled with our approach separating the liquid phase from the crystals to implement UV/Vis spectroscopy in a protein crystallization process.

Taking into account the gained information about the supernatant composition by the $\frac{A_{260\text{nm}}}{A_{280\text{nm}}}$ ratio from the implemented VP UV/Vis spectroscopy in the analytical bypass, the outliers of the bypass-related analytics, and the blockage of the tubing, the new insights by the implemented UV/Vis spectroscopy did not justify the increased complexity of the bypass in the set-up for this project. The variation in the UV/Vis spectrum could not be correlated with process parameters, e.g. target molecule concentration in the supernatant, as the data usability was lowered by interfering impurities in the UV/Vis spectrum, missing data overnight, and difficulties during the start of the bypass. However, the implementation of other particle-sensitive analytics should be possible, e.g. fluorescence or nuclear magnetic resonance (NMR) spectroscopy, if applicable in the specific crystallization process.

5.4.2 Raman spectroscopy and chemometrics

To characterize and potentially monitor crystallization, a Raman spectroscopy probe was immersed directly into the stirred crystallization vessel. The probe was in direct contact with the crystal suspension and may show variation in the spectrum over the process time as the liquid phase composition changes.

Spectral differences are visible between 450 to 1500 cm^{-1} and between 2800 to 3000 cm^{-1}. The latter is contributed by C-H stretching [214]. The former is described as the fingerprint region of proteins [214]. Comparing the spectra to Raman spectra of air, protein buffer and crystallization buffer, preprocessed Raman peaks could be traced back to different compounds. The crystallization buffer spectrum shows distinct peaks near 850, 1065, 1140, 1250, 1286 and 1475 cm^{-1} (see Supplementary Material A5.4 (B)). As PEG contributes strongly to the Raman spectrum compared to the protein, the spectral analysis is hampered with respect to the desired process characteristics, i.e. crystal yield and target molecule concentration in the supernatant. Differences between the spectra of the crystallization buffer and during the experiment are visible near 970 to 1030 cm^{-1}, between 1170 to 1230 cm^{-1}. This may be caused by the amino acid contribution of Phenyalanine (Phe) (1000, 1030 and 1205 cm^{-1} Tuma, 2005; Huang *et al.*, 2006) and Tyrosine (Tyr) (1174, 1205 cm^{-1} Tuma, 2005). The wavenumber 757, 853 cm^{-1} and 1225 to 1525 cm^{-1} are associated with Tryptophan (Try), Tyr, and the amide III bands, respectively [70].

The chemometric analysis of the preprocessed Raman spectra with PCA showed that the experiments followed a trend (see Figure 5.5) as indicated by the arrows. Note that Exp3 showed a faster transition from the lower right to the upper left corner (see Figure 5.5 (C)) and may be linked to the arising crystallization accompanied with a decreasing *Lk*ADH concentration. The PCA of the UV/Vis spectra could only cluster the experiments according to different experiment conditions, namely lysis cycle number, and varying initial absorption of the clarified lysate at 280 nm, but did not show a trend within each experiment (data not shown). Five peaks at 493 cm^{-1}, 708 cm^{-1}, 1410 cm^{-1}, 2909 cm^{-1}, and 2970 cm^{-1} occurred during Exp2 and Exp5 (see Figure 5.4 (B)) and could not be traced back to protein crystals. Looking at the PC1 and PC2 of the PCA over the whole spectral range, the

supposedly defective observations are captured in clusters which are clearly separated from the crystallization-associated trends indicated by the arrows (see Figure 5.5 (B, E)). As these peaks do not correlate with the protein buffer or crystallizing buffer, the *Lk*ADH concentration in the supernatant, or the crystal yield, and appeared or disappeared spontaneously, it is assumed that precipitated impurities may have aggregated and accumulated near or detached from the spectroscopy probe, arbitrarily, due to the agitation in the stirred vessel.

Based on these findings, the Raman spectra and the presented preprocessing procedure were used for the model development for PAT. The aforementioned wavenumber ranges (see Figure 5.4 (C)) are selected for the PLS model development as it is assumed to correlate with product characteristics, for instance *Lk*ADH concentration.

The KS algorithm-based and manual data split with Exp5 were investigated and compared to evaluate the extrapolation capability of the calculated PLS models. The metrics for chemometric models demonstrate a high model validity in both model cases (see Figure 5.6). The values R^2 and Q^2 are near 1 implying a good transferability on the external spectral data set. The $RMSECV$ and $RMSEP$ are desired to be low. In this case, the PLS model with KS data split demonstrates higher R^2, Q^2 and lower $RMSEP$ values - both suggesting that the PLS model with KS data split is superior. The KS data split method is depending strongly on the specified number of validation samples and selects the validation samples based on a uniform distribution of the data split using a distance metric. The assumption that the data was split in data subsets with high similarity may not hold in all situations, limiting its applicability to certain types of data. The spectra in the second cluster of the PCA were not included in the validation set by the KS algorithm which improved the model evaluation metrics. Data points of only Exp1, and each one data point of Exp2 and Exp4 are selected to represent the validation data set (see Supplementary Material A5.7). The calibration of the model on data of each experiment may potentially incorporate variations of the experiments and batch-to-batch variations into the model. Batch-to-batch variations are caused during the *Lk*ADH production in *E.coli*, by variations during lysis and clarification, and, in our case, the different initial crystallization conditions (see Table 5.1). To challenge the model applicability to extrapolate on a new experiment, the authors decided to proceed with the model calibrated with the manual data split using Exp5 for validation. By this, the PLS model prediction performance could be evaluated on data possibly prone to experimental variations.

The model calibrated and validated with the manual data split underestimates the *Lk*ADH concentration in Exp5 at higher concentrations above 0.7 g/L (see Figure 5.7 (E)). Within the first 10 h crystallization was not visible in the microscopic images of the off-line samples, but the Raman spectrum may be influenced by other processes occurring in the crystallization vessel which leads to an underestimated concentration prediction in the first discussed time slot. The spectra of samples with high *Lk*ADH concentration in Exp4 were not representative for spectra in Exp5. As the *Lk*ADH concentration was at a comparable level, other species present in the supernatant may interfere. The crystallization solution contains PEG which is known to induce aggregation or precipitation. Aggregation processes of impurities, namely nucleic acids or HCPs, in the examined time slot may affect the Raman spectrum and lead to the visible discrepancies.

Regarding lower concentrations, the model performed well, even though the spectra of Exp5 varied strongly (see the second cluster in Figure 5.5 (E) compared to (A-D)). The selected preprocessing parameters and wavenumber ranges were able to cope with the disturbance in lower concentration ranges. However, off-line samples and their analysis cannot be left out entirely, but the sample number may be reduced combining Raman spectroscopy and PLS modeling for protein crystallization PAT. More experiments with varied crystallization conditions, more samples analyzed, and different cultivation batches can increase the variety of spectra and reduce the effect of outliers which is beneficial for model calibration. This may lead to a better chemometric model which can predict reliably over the whole concentration range.

5.4.3 Assessment of the crystallization process using multiple PAT tools

The presented analytical set-up as a whole provides the possibility to examine the conducted experiments regarding the type and amount of impurities, target molecule, and different initial absorption value at 280 nm. The shortened induction time of protein decrease in the supernatant marking the start of protein crystallization (see Figure 5.7 from left to right) is expected as the supersaturation and, thus, phase behavior change with varying PEG concentration [24]. Baumgartner et al. [36] examined the effect of two PEG additives on different proteins and observed an increasing depletion attraction effect with increasing polymer concentration [197]. The enzyme *Lactobacillus brevis* alcohol dehydrogenase (*Lb*ADH) which is a homologous protein to *Lk*ADH of this project showed an increased tendency to form crystals with increasing PEG concentration and was studied in detail in Nowotny et al. [10]. Furthermore, the PEG concentration influences the crystal geometry as well leading to larger crystal sizes (see Supplementary Material A5.3) for lower supersaturation level [9, 11] in a crystallization buffer with 12.5 % PEG in Exp2 and Exp5. These experiments led to the highest crystals count per well in Figure 5.2. This contradicts the fact that higher supersaturation levels result in a larger amount of smaller crystals, but can be explained as smaller crystals are difficult to detect automatically by the ML-based tool due to the low ratio of the crystal size to the camera resolution.

The presence of impurities becomes apparent in Figure 5.3 (D), and in Table 5.1 regarding the nucleic acid content, and HCP, respectively. The HCP reduction was achieved in a similar magnitude for crystal redissolution of a homologous enzyme [10]. Even though the $\frac{A_{260\text{nm}}}{A_{280\text{nm}}}$ ratio and HCP quantification are based on on-line and off-line analytics, they provide valuable process knowledge and help to understand crystallization processes in complex, heterogeneous solutions.

With a higher absorption $A_{280\text{nm}}$ of the clarified lysate in Exp4, a high *Lk*ADH concentration was achieved in the beginning of the crystallization process (see Figure 5.7 (D)). Compared to Exp1 with the same PEG concentration, protein crystals could be detected earlier, but the supernatant concentration of *Lk*ADH did not drop to the same value. It is assumed that the equilibrium and the maximum crystal yield were not reached within the conducted experimental time. These findings contradict the results of Walla et al. [203]

who observed that *Lk*ADH WT reached the equilibrium within 48 h for the screened PEG concentration. Note that the analytical frame differed in the mentioned project as the total protein concentration was determined. The crystal yields of the experiments performed with 125 or 150 g/L PEG were lower than the yields achieved in Walla et al. [203]. In this work, the crystal yield and crystallization process time were derived from the individual *Lk*ADH concentration with IMAC, which makes a direct comparison difficult. Furthermore, variations during the cultivation, lysis, lysate clarification procedure, or crystallization vessel (see Supplementary Table A5.1) may change the product or impurity profiles leading to a lower, initial *Lk*ADH concentration when the $A_{280\text{nm}}$ is adjusted to the same value.

5.5 Conclusion

This research project aimed to examine and monitor stirred *Lactobacillus kefir* alcohol dehydrogenase (*Lk*ADH) enzyme crystallization out of clarified *Escherichia coli* (*E.coli*) lysate on a 300 mL scale to increase process understanding of a multiphase process. The implemented analytics consisted of an in-line Raman spectroscopy probe, on-line cross-flow filtration (CFF) bypass for the liquid phase analysis in a variable pathlength (VP) flow cell for ultraviolet-visible light (UV/Vis) spectroscopy, and high-performance liquid chromatography (HPLC) immobilized metal ion affinity chromatography (IMAC), enzyme-linked immunosorbent assay (ELISA) and microscopic analysis for off-line samples.

Chemometric analysis of the preprocessed Raman spectra could identify similar process trends in the spectra of the experiments with principal component analysis (PCA), and could monitor the *Lk*ADH concentration in clarified lysate with a partial least squares (PLS) regression model built on selected wavenumber regions containing product-relevant information. The presented, analytical set-up led to a comprehensive overview of the conducted batch experiments which is in agreement with theoretical considerations of protein crystallization.

Despite the complexity of the clarified lysate, a suspension containing scattering crystals, impurities in the supernatant, and precipitate, spectroscopy could be used to monitor the target molecule concentration in the liquid phase during a multi-phase process. The analytical bypass facilitated the implementation of particle-sensitive analytics, i.e. VP UV/Vis spectroscopy which indicated changes in the contaminant profile with the absorption ratio at two specific wavelengths typical for proteins and nucleic acids. The off-line analysis of microscopic images allowed objective evaluation of crystal nucleation, or crystal breakage. In our case, the crystal number, and geometry did not vary when a CFF bypass was installed meaning that crystal breakage was not observed with the chosen CFF process parameters. Regarding model limitations, batch-to-batch variations and the heterogeneous components in the clarified lysate complicated the direct model transfer to new experiments without additional validation samples.

The suggested process analytical technology (PAT) set-up with in-line Raman spectroscopy can be applied to other processes based on phase behavior, e.g. precipitation or flocculation, if the molecule of interest contributes to the recorded spectrum sufficiently. Good calibration

procedure and carefully considered data splitting for the model development help to unravel the underlying spectral nuances associated with the desired product characteristics. The increased process understanding and possibility to monitor phase behavior based processes can help the operator to optimize the process, e.g. regarding crystal yield. When protein solutions of high purity need to be crystallized, the installation of an on-line bypass with VP UV/Vis measurements can be especially useful to determine the supernatant concentration directly. The ability to monitor protein crystallization processes is essential for process control and process adaptations as biotechnological processes are often subject to batch-to-batch variability.

Acknowledgment

The authors would like to thank Bernadette Pichler and Kristina Schleining for conducting substantial pre-experiments, as well as Annabelle Dietrich and Robin Schiemer for proofreading the manuscript. Furthermore, the authors express their gratitude to Egbert Müller for the material supply of the analytical chromatography column. The support of Brigitte Walla and Daniel Bischoff by the TUM Graduate School is acknowledged as well. We acknowledge support by the KIT Publication Fund of the Karlsruhe Institute of Technology.

6

General discussion and conclusion

Recent advances in upstream processing (USP) have shifted the productivity bottleneck towards downstream processing (DSP) steps in biotechnological or biopharmaceutical production processes. Therefore, alternative process steps have gained more interest due to their higher productivity and lower production costs compared to DSP with standard chromatography. As protein crystallization can produce pure crystals with a high yield and comparably low production costs, it can be an excellent alternative DSP step. However, protein crystallization processes are rarely developed as the optimal process conditions need to be screened in extensive high-throughput (HT) screenings resulting in a large number of samples and long analysis time. Implementing process analytical technology (PAT) into other DSP processes has shown to improve process development as these methods support designing, monitoring, and controlling processes ideally in real-time. Thus, the employment of PAT sensors for protein crystallization in DSP is advised. This initiative complies with the paradigm of quality by design (QbD) which states that quality should be built into the process design to ensure a high product quality. Contrary to other DSP steps, solid particles are present and large concentrations shifts occur during the process which need to be considered in the PAT sensor choice and set-up. Therefore, the objective of this thesis is the development of data-driven, analytical solutions to challenges in protein crystallization processes by implementing PAT for effective process design.

HT screenings are commonly used when phase behavior based DSP steps need to be developed. Regarding protein crystallization screenings, the samples are commonly analyzed using qualitative analytics. However, quantitative analytics are crucial when protein crystallization screenings should find the optimal process condition regarding yield or purity. To quantify the target reliably and deal with the large number of screening samples, fast and HT-compatible analytics are required. Ideally, these analytics should consume only

a minimal sample volume. To meet these criteria, the first research project investigated a novel, analytical workflow based on ultraviolet-visible light (UV/Vis) spectroscopy and chemometric partial least squares (PLS) regression modeling. As a proof of concept, the calculated regression model was then applied to mixtures of three model proteins in a crystallization screening. Furthermore, two conditions were selected for a kinetic study to transfer the generated models to a different experimental set-up and to monitor the crystallization kinetics over time. For each protein, one PLS model was calculated and further applied to specifically quantify each protein in the crystallization supernatant using solely the UV/Vis spectra and the PLS model. Regardless of the experimental conditions and variations (pH, precipitant concentration, target protein concentration) during the screening or kinetic study, the models predicted the concentrations accurately. By visualizing the model-predicted, specific concentrations in a phase diagram, the process sweet spots for crystallization could be identified with respect to the crystal yield. The crystal purity was overall high and inclusion of contaminants could be excluded in the center of the crystallization window whereas at its borders variations in the purity could be traced back to the calculation method. In the kinetic study, the visible target concentration drop in the supernatant could be tracked back to the formation of crystals. The concentrations of all proteins increased throughout the kinetic study which was caused by solvent evaporation over time. Overall, this study describes a workflow of a rapid, versatile, low-volume analysis based on spectroscopy and chemometrics for the design of protein crystallization processes. As a proof-of-concept, it investigated model proteins in chemically defined solutions. With a four times shorter analysis time compared to the reference method, crystal yield and purity could be quantified and used for finding optimal crystallization process conditions.

In empirical HT screenings, large, complex data sets are generated and can be organized along the screened conditions or recorded variables. When these data sets of higher structure need to be analyzed, chemometric analysis require adaptions. One potential solution is the analysis with multi-way methods as a subclass of chemometrics which take advantage of this multi-dimensionality. Multi-way parallel factor analysis (PARAFAC) models explore data sets and decompose multi-dimensional data into the contribution of each species in each dimension. Furthermore, this model operates in an unsupervised manner and can potentially quantify concentrations without calibration. Preparing calibration samples and the analysis with a reference method for model calibration becomes unnecessary when PARAFAC models are created. This can prove effective in early stages of process development of capture steps when accurate, quantitative analytics are not yet available or influenced by a greater number of impurities. However, multi-way methods have not been tested on phase behavior based screenings of complex, biological solutions. To demonstrate the transferability to screenings of new molecules or new capture steps, three screening studies on protein crystallization and precipitation were conducted in the second research project (Chapter 4) investigating three different molecule classes in chemically defined and complex solutions. In particular, the first case study was based on the screening data generated in the first research project (Chapter 3). Two precipitation studies on monoclonal antibodies (mAbs) in harvest cell culture fluid (HCCF) and virus-like particles (VLPs) in clarified *Escherichia coli* (*E.coli*) lysate were conducted and supernatant samples from the precipitation, wash step, and

redissolution step were analyzed. The data set consisted of UV/Vis spectra structured along the dimensions time, wavelength, and sample number. The first and second PARAFAC models could be validated with quantitative reference concentrations and provided suitable spectral preprocessing and model parameters. Consequently, this knowledge could be used for the third screening where a quantitative concentration reference method was missing. Using only the higher structured UV/Vis spectral data set, the calculated PARAFAC models could estimate the pure component spectra and concentrations in the screenings. Furthermore, the first crystallization screening of three model proteins showed that only species, expressing different phase behavior and specific spectral differences, can be distinguished by the model. As two of the model proteins stayed mainly in solutions and did not show any concentration changes in the investigated screening solutions, they were clustered as one species. Consequently, PARAFAC models can be applied to biological solutions where a large number of species and species variants are present. The results of the second and third precipitation screenings proved that PARAFAC models could be used to quantify concentrations and determine pure species spectra regardless of the modality, impurities, or examined process steps. The estimated concentration and spectral profiles corresponded closely to off-line analytics. As a result, process sweet spots with the screened conditions could be located with respect to the yield and purity of the target molecule. Despite the absence of reference analytics, the presented, analytical workflow was universally applicable to crystallization or precipitation HT screenings and provided valuable process knowledge regardless of the biological target.

Next to efficient process design, additional objectives of PAT are process monitoring and process control. PAT tools for monitoring are commonly implemented in DSP steps, especially in chromatography steps. Regarding PAT for protein crystallization, first advances were achieved investigating the crystals itself or polymorphism in rather pure process liquids, solely containing the target molecule. However, applications of protein crystallization PAT in the presence of numerous species and contaminants is rare. Established PAT tools have to be tailored to the increased complexity of multi-phase crystallization processes as particles and precipitated contaminants are present. In addition, crystallization processes can cause high concentration shifts, potentially trespassing the sensor detection limits. Therefore, the third part of this thesis (Chapter 5) aimed to develop a broad monitoring PAT set-up for the crystallization of *Lactobacillus kefir* alcohol dehydrogenase (*Lk*ADH) from clarified *E.coli* lysate in 300 mL scale using in-line probes, on-line sensors, and off-line analytics. Similar to the second and third case study of Chapter 4, complex process solution were investigated, though this time for a crystallization process. For real-time monitoring, a Raman probe was placed *in situ* in the crystalline process liquid. To analyze liquid phase and deploy particle sensitive analytics, a cross-flow filtration (CFF) based bypass was developed and a variable pathlength (VP) flow cell connected to a UV/Vis spectrophotometer was installed. The off-line analysis complemented the PAT set-up providing information on the protein purity, the host cell protein (HCP) content, the target molecule concentration, and the crystal formation utilizing sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE), enzyme-linked immunosorbent assay (ELISA), immobilized metal ion affinity chromatography (IMAC), and automated image analysis of the drawn samples, respectively. Based on the off-line concentration measurements and in-line measured

Raman spectra, a PLS model predicted the target protein concentration in the supernatant. The combination of Raman spectroscopy and chemometrics proved effective despite the crystalline particles in a multi-phase process, the biological complexity of clarified lysate, and the broad variety of soluble or precipitated impurities. The calculated model achieved adequate concentration predictions when the data-splitting procedure selected calibration samples of all model runs for the model calculation. However, when one experiment was used as an independent test set, the predictive model performance was lower, especially for the high target concentrations. This leads to the assumption, that with in-line Raman spectroscopy, the number of validation samples for new experiments can be reduced but their analysis is still crucial to monitor the whole crystallization process accurately as crystallization processes from lysate are prone to batch-to-batch variations. The implemented UV/Vis sensor in the particulate-free bypass supported the purity assessment of the initial process liquids and demonstrated that particle sensitive sensors can be implemented in crystallization process using the described set-up. The HCP reduction from the initial process liquid to the redissolved crystals showed that the experimental set-up generated pure crystals and might be a good alternative to chromatography based processes. Consistent crystal geometry was shown by the image analysis for both crystallization processes, with and without the CFF-based bypass, underlining that no crystal breakage occurred. Considering the aforementioned limitations of the increased complexity of multi-phase crystallization processes in clarified lysate, this PAT framework was able to monitor the crystallization process by coupling Raman spectra and chemometric regression modeling. Additionally, the analytical set-up presented in Chapter 5 provides a comprehensive overview of the crystallization process by characterizing critical quality attributes (CQAs) of the generated protein crystals.

In conclusion, this thesis presents new data-driven strategies to design or monitor protein crystallization processes combining spectroscopic sensors and chemometric models. As a reaction to advances in novel, personalized biotherapeutics, the submitted studies dealt with a broad variety of biological proteins in chemically defined or complex process liquids. HT-compatible, analytical sensors were developed for screenings applying chemometric models, e.g. PLS regression, or modern multi-way techniques, e.g. PARAFAC models, for the rapid quantification of different species in mixtures. Additionally, a broad PAT set-up could selectively quantify the target molecule in a multi-phase process. The developed workflows increase knowledge-based process development which is in accordance with QbD for the biopharmaceutical or biotechnological industry. The analytical screening approaches, described in this thesis, have proven their transferability to other biologics or other process steps that depend on protein solubility and phase behavior. Next to precipitation processes, potential, new applications could be flocculation, the redissolution of inclusion bodies, crystal redissolution, or protein aggregation. Analogously, the workflow for in-process monitoring may find similar employments in the aforementioned processes or crystallization of proteins that were protein engineered for crystallization. The presented monitoring tools can be a first step towards process optimization or process control of CQAs for QbD. Since multi-way methods have shown to explore multi-variate data sets about chemical reactions, it may be interesting to analyze data sets dealing with biological reactions. Spectral data sets of enzymatic reactions, dis- and reassembly processes, un- and refolding processes of proteins, or chemically linking

reactions, e.g. reduction and conjugation reactions of antibody-drug conjugates (ADCs), can be potential applications for multi-way chemometrics when various samples are examined with multi-variate spectroscopy. In analytical chromatography of the mentioned, biological reactions, multi-way methods may be a feasible solution to quantify species in overlapping peaks individually if the samples cannot be resolved sufficiently. However, the successful application depends on a good signal-to-noise ratio and further is restricted to reactions or analytical measurements demonstrating significant spectral differences between the analytes.

Considering these conclusions, this thesis presents a broader toolbox of data-driven analytics tailored to crystallization processes using spectroscopy and chemometrics with or without calibration samples and reference analytics. Demonstrated in studies on different scales, the described approaches cover numerous process design stages from phase behavior screenings, development, and monitoring for manufacturing. The developed analytics encourage PAT for alternative DSP steps and lay the first steps towards knowledge-based process design and control.

References

[1] National Center for Biotechnology Information (NCBI) and U.S. National Library of Medicine, *PubMed database*, 01/2024.

[2] N. Asherie, „Protein crystallization and phase diagrams“, *Methods*, vol. 34, no. 3, pp. 266–272, 2004.

[3] A. McPherson, „Introduction to protein crystallization“, *Methods*, vol. 34, no. 3, pp. 254–265, 2004.

[4] L. Galm, J. Morgenstern, and J. Hubbuch, „Manipulation of lysozyme phase behavior by additives as function of conformational stability“, *International Journal of Pharmaceutics*, vol. 494, no. 1, pp. 370–380, 2015.

[5] M. T. Schermeyer, A. K. Wöll, B. Kokke, M. Eppink, and J. Hubbuch, „Characterization of highly concentrated antibody solution - A toolbox for the description of protein long-term solution stability“, *mAbs*, vol. 9, no. 7, pp. 1169–1185, 2017.

[6] S. Martínez-Caballero, M. Cuéllar-Cruz, N. Demitri, M. Polentarutti, A. Rodríguez-Romero, and A. Moreno, „Glucose Isomerase Polymorphs Obtained Using an Ad Hoc Protein Crystallization Temperature Device and a Growth Cell Applying an Electric Field“, *Crystal Growth and Design*, vol. 16, no. 3, pp. 1679–1686, 2016.

[7] E. O. Watanabe, E. Popova, E. A. Miranda, G. Maurer, and P. d. A. P. Filho, „Phase equilibria for salt-induced lysozyme precipitation: Effect of salt type and temperature“, *Fluid Phase Equilibria*, vol. 281, no. 1, pp. 32–39, 2009.

[8] Y. B. Lin *et al.*, „An extensive study of protein phase diagram modification: Increasing macromolecular crystallizability by temperature screening“, *Crystal Growth and Design*, vol. 8, no. 12, pp. 4277–4283, 2008.

[9] M. E. Klijn and J. Hubbuch, „Application of Empirical Phase Diagrams for Multidimensional Data Visualization of High-Throughput Microbatch Crystallization Experiments“, *Journal of Pharmaceutical Sciences*, vol. 107, no. 8, pp. 2063–2069, 2018.

[10] P. Nowotny *et al.*, „Rational Crystal Contact Engineering of Lactobacillus brevis Alcohol Dehydrogenase to Promote Technical Protein Crystallization“, *Crystal Growth and Design*, vol. 19, no. 4, pp. 2380–2387, 2019.

[11] A. McPherson and B. Cudney, „Optimization of crystallization conditions for biological macromolecules“, *Acta Crystallographica Section F:Structural Biology Communications*, vol. 70, pp. 1445–1467, 2014.

[12] K. C. Bauer, S. Suhm, A. K. Wöll, and J. Hubbuch, „Impact of additives on the formation of protein aggregates and viscosity in concentrated protein solutions“, *International Journal of Pharmaceutics*, vol. 516, no. 1-2, pp. 82–90, 2017.

[13] D. Hekmat, D. Hebel, H. Schmid, and D. Weuster-Botz, „Crystallization of lysozyme: From vapor diffusion experiments to batch crystallization in agitated ml-scale vessels“, *Process Biochemistry*, vol. 42, no. 12, pp. 1649–1654, 12/2007.

[14] D. Hebel, S. Huber, B. Stanislawski, and D. Hekmat, „Stirred batch crystallization of a therapeutic antibody fragment“, *Journal of Biotechnology*, vol. 166, no. 4, pp. 206–211, 2013.

[15] R. A. Judge, E. L. Forsythe, and M. L. Pusey, „The effect of protein impurities on lysozyme crystal growth“, *Biotechnology and Bioengineering*, vol. 59, no. 6, pp. 776–785, 09/1998.

[16] C. Abergel, M. P. Nesa, and J. C. Fontecilla-Camps, „The effect of protein contaminants on the crystallization of turkey egg white lysozyme“, *Journal of Crystal Growth*, vol. 110, no. 1-2, pp. 11–19, 1991.

[17] J. Liu *et al.*, „The dual function of impurity in protein crystallization“, *CrystEngComm*, vol. 24, no. 3, pp. 647–656, 2022.

[18] K. A. Kantardjieff and B. Rupp, „Protein isoelectric point as a predictor for increased crystallization screening efficiency“, *Bioinformatics*, vol. 20, no. 14, pp. 2162–2168, 2004.

[19] M. Boström, F. W. Tavares, S. Finet, F. Skouri-Panet, A. Tardieu, and B. W. Ninham, „Why forces between proteins follow different Hofmeister series for pH above and below pI“, *Biophysical chemistry*, vol. 117, no. 3, pp. 217–224, 10/2005.

[20] B. L. Neal, D. Asthagiri, O. D. Velev, A. M. Lenhoff, and E. W. Kaler, „Why is the osmotic second virial coefficient related to protein crystallization?“, *Journal of Crystal Growth*, vol. 196, no. 2-4, pp. 377–387, 01/1999.

[21] F. Hofmeister, „24. Zur Lehre von der Wirkung der Salze - Fünfte Mittheilung Untersuchungen über den Quellungsvorgang“, *Archiv für Experimentelle Pathologie und Pharmakologie*, vol. 27, no. 6, pp. 395–413, 1890.

[22] R. Majumdar *et al.*, „Effects of salts from the Hofmeister series on the conformational stability, aggregation propensity, and local flexibility of an IgG1 monoclonal antibody.“, *Biochemistry*, vol. 52, no. 19, pp. 3376–89, 05/2013.

[23] R. L. Baldwin, „How Hofmeister ion interactions affect protein stability.“, *Biophysical Journal*, vol. 71, no. 4, p. 2056, 1996.

[24] O. Galkin and P. G. Vekilov, „Nucleation of protein crystals: Critical nuclei, phase behavior, and control pathways“, *Journal of Crystal Growth*, vol. 232, no. 1-4, pp. 63–76, 2001.

[25] N. Rakel, L. Galm, K. C. Bauer, and J. Hubbuch, „Influence of macromolecular precipitants on phase behavior of monoclonal antibodies“, *Biotechnology Progress*, vol. 31, no. 1, pp. 145–153, 2015.

[26] S. D. Durbin and G. Feher, „Protein crystallization.“, *Annual review of physical chemistry*, vol. 47, no. 1, pp. 171–204, 10/1996.

[27] M. W. Burke, R. Leardi, R. A. Judge, and M. L. Pusey, „Quantifying Main Trends in Lysozyme Nucleation: The Effect of Precipitant Concentration, Supersaturation, and Impurities“, *Crystal Growth and Design*, vol. 1, no. 4, pp. 333–337, 2001.

[28] P. G. Vekilov, „Nucleation“, *Crystal Growth and Design*, vol. 10, no. 12, pp. 5007–5019, 12/2010.

[29] J. M. García-Ruiz, „Nucleation of protein crystals“, *Journal of Structural Biology*, vol. 142, no. 1, pp. 22–31, 04/2003.

[30] C. Jacobsen, J. Garside, and M. Hoare, „Nucleation and growth of microbial lipase crystals from clarified concentrated fermentation broths“, *Biotechnology and Bioengineering*, vol. 57, no. 6, pp. 666–675, 1998.

[31] H. Huettmann, S. Zich, M. Berkemeyer, W. Buchinger, and A. Jungbauer, „Design of industrial crystallization of interferon gamma: Phase diagrams and solubility curves“, *Chemical Engineering Science*, vol. 126, pp. 341–348, 2015.

[32] T. Lee, J. Vaghjiani, G. Lye, and M. Turner, „A systematic approach to the large-scale production of protein crystals.“, *Enzyme and microbial technology*, vol. 26, no. 8, pp. 582–592, 05/2000.

[33] J. Peters, T. Minuth, and W. Schröder, „Implementation of a crystallization step into the purification process of a recombinant protein“, *Protein Expression and Purification*, vol. 39, no. 1, pp. 43–53, 01/2005.

[34] N. Hillebrandt, P. Vormittag, N. Bluthardt, A. Dietrich, and J. Hubbuch, „Integrated Process for Capture and Purification of Virus-Like Particles: Enhancing Process Performance by Cross-Flow Filtration“, *Frontiers in Bioengineering and Biotechnology*, vol. 8, pp. 66–76, 05/2020.

[35] N. Harn, C. Allan, C. Oliver, and C. R. Middaugh, „Highly concentrated monoclonal antibody solutions: direct analysis of physical structure and thermal stability.“, *Journal of pharmaceutical sciences*, vol. 96, no. 3, pp. 532–46, 03/2007.

[36] K. Baumgartner *et al.*, „Determination of protein phase diagrams by microbatch experiments: Exploring the influence of precipitants and pH“, *International Journal of Pharmaceutics*, vol. 479, no. 1, pp. 28–40, 2015.

[37] M. E. Klijn and J. Hubbuch, „Time-Dependent Multi-Light-Source Image Classification Combined With Automated Multidimensional Protein Phase Diagram Construction for Protein Phase Behavior Analysis.“, *Journal of pharmaceutical sciences*, vol. 109, no. 1, pp. 331–339, 01/2020.

[38] M. E. Klijn, A. K. Wöll, and J. Hubbuch, „Apparent protein cloud point temperature determination using a low volume high - throughput cryogenic device in combination with automated imaging“, *Bioprocess and Biosystems Engineering*, no. 0123456789, pp. 1–45, 2019.

[39] Bakeev, *Process Analytical Technology*, K. A. Bakeev, Ed. Oxford, UK: Blackwell Publishing Ltd, 08/2005, vol. 41, pp. 545–553.

[40] W. Mäntele and E. Deniz, „UV–VIS absorption spectroscopy: Lambert-Beer reloaded“, *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 173, pp. 965–968, 02/2017.

[41] A. Müllertz, Y. Perrie, and T. Rades, „Analytical Techniques in the Pharmaceutical Sciences“, *Thermal Analysis of Pharmaceuticals*, pp. 253–293, 2016.

[42] F. Hinderer, *UV/Vis-Absorptions- und Fluoreszenz-Spektroskopie* (essentials). Wiesbaden: Springer Fachmedien Wiesbaden, 2020.

[43] M. Rüdt, T. Briskot, and J. Hubbuch, „Advances in downstream processing of biologics – Spectroscopy: An emerging process analytical technology“, *Journal of Chromatography A*, vol. 1490, pp. 2–9, 2017.

[44] A. R. Goldfarb, L. Saidel, and E. Mosovich, „The ultraviolet absorption spectra of proteins“, *Journal of Biological Chemistry*, vol. 193, no. 1, pp. 397–404, 11/1951.

[45] H. Mach, C. R. Middaugh, and N. Denslow, „Determining the Identity and Purity of Recombinant Proteins by UV Absorption Spectroscopy“, *Current Protocols in Protein Science*, vol. 1, no. 1, pp. 1–7, 1995.

[46] D. B. Wetlaufer, „Ultraviolet spectra Of Proteins and Amino Acids“, *Advances in Protein Chemistry*, vol. 17, no. C, pp. 303–390, 01/1963.

[47] 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*, 1994.

[48] J. M. Antosiewicz and D. Shugar, „UV–Vis spectroscopy of tyrosine side-groups in studies of protein structure. Part 2: selected applications“, *Biophysical Reviews*, vol. 8, no. 2, pp. 163–177, 2016.

[49] S. R. Gallagher, „Quantitation of DNA and RNA with absorption and fluorescence spectroscopy“, *Current Protocols in Molecular Biology*, no. SUPPL.93, pp. 1–14, 2011.

[50] A. Valentic, J. Müller, and J. Hubbuch, „Effects of Different Lengths of a Nucleic Acid Binding Region and Bound Nucleic Acids on the Phase Behavior and Purification Process of HBcAg Virus-Like Particles“, *Frontiers in Bioengineering and Biotechnology*, vol. 10, p. 929 243, 07/2022.

[51] W. W. Wilfinger, K. Mackey, and P. Chomczynski, „Effect of pH and ionic strength on the spectrophotometric assessment of nucleic acid purity.“, *BioTechniques*, vol. 22, no. 3, pp. 474–6, 03/1997.

[52] M. H. Kamga, H. Woo Lee, J. Liu, and S. Yoon, „Quantification of protein mixture in chromatographic separation using multi-wavelength UV spectra“, *Biotechnology Progress*, vol. 29, no. 3, pp. 664–671, 05/2013.

[53] S. K. Hansen, E. Skibsted, A. Staby, and J. Hubbuch, „A label-free methodology for selective protein quantification by means of absorption measurements“, *Biotechnology and Bioengineering*, vol. 108, no. 11, pp. 2661–2669, 11/2011.

[54] R. Ríos-Reina and S. M. Azcarate, „How Chemometrics Revives the UV-Vis Spectroscopy Applications as an Analytical Sensor for Spectralprint (Nontargeted) Analysis“, *Chemosensors 2023, Vol. 11, Page 8*, vol. 11, no. 1, p. 8, 12/2022.

[55] 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, 07/2014.

[56] N. Brestich, 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, 02/2018.

[57] 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. 1406–16, 07/2015.

[58] 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, 02/2017.

[59] 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*, no. October, 2020.

[60] L. Rolinger *et al.*, „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, 04/2020.

[61] 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, 06/2019.

[62] R. P. Bhangale, R. Ye, T. B. Lindsey, and L. S. Wolfe, „Application of inline variable pathlength technology for rapid determination of dynamic binding capacity in downstream process development of biopharmaceuticals“, *Biotechnology Progress*, vol. 38, no. 2, 03/2022.

[63] H. Schlemmer and J. Katzer, „ATR technique for UV/VIS analytical measurements“, *Fresenius' Zeitschrift für Analytische Chemie*, vol. 329, no. 4, pp. 435–439, 1987.

[64] B. Wan, C. A. Zordan, X. Lu, and G. McGeorge, „In-line ATR-UV and Raman Spectroscopy for Monitoring API Dissolution Process During Liquid-Filled Soft-Gelatin Capsule Manufacturing“, *AAPS PharmSciTech*, vol. 17, no. 5, pp. 1173–1181, 2016.

[65] S. Mosca, C. Conti, N. Stone, and P. Matousek, *Spatially offset Raman spectroscopy*, 12/2021.

[66] K. A. Esmonde-White, M. Cuellar, C. Uerpmann, B. Lenain, and I. R. Lewis, „Raman spectroscopy as a process analytical technology for pharmaceutical manufacturing and bioprocessing“, *Analytical and Bioanalytical Chemistry*, vol. 409, no. 3, pp. 637–649, 2017.

[67] J. M. Benevides, S. A. Overman, and G. J. Thomas, *Raman Spectroscopy of Proteins*. 2003, vol. 33.

[68] R. Tuma, „Raman spectroscopy of proteins: From peptides to large assemblies“, *Journal of Raman Spectroscopy*, vol. 36, no. 4, pp. 307–319, 2005.

[69] E. A. Gooding, E. R. Deutsch, J. Huehnerhoff, and A. R. Hajian, „The high throughput virtual slit enables compact, inexpensive Raman spectral imagers“, no. February 2018, p. 27, 2018.

[70] C. Y. Huang, G. Balakrishnan, and T. G. Spiro, „Protein secondary structure from deep-UV resonance Raman spectroscopy“, *Journal of Raman Spectroscopy*, vol. 37, no. 1-3, pp. 277–282, 2006.

[71] B. Nagy *et al.*, „Quantification and handling of nonlinearity in Raman micro-spectrometry of pharmaceuticals“, *Journal of Pharmaceutical and Biomedical Analysis*, vol. 128, pp. 236–246, 2016.

[72] K. A. Esmonde-White, M. Cuellar, and I. R. Lewis, „The role of Raman spectroscopy in biopharmaceuticals from development to manufacturing“, *Analytical and Bioanalytical Chemistry*, vol. 414, no. 2, pp. 969–991, 2022.

[73] D. Yilmaz, H. Mehdizadeh, D. Navarro, A. Shehzad, M. O'Connor, and P. McCormick, „Application of Raman spectroscopy in monoclonal antibody producing continuous systems for downstream process intensification“, *Biotechnology Progress*, vol. 36, no. 3, 05/2020.

[74] N. R. Abu-Absi *et al.*, „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, 05/2011.

[75] B. Berry, J. Moretto, T. Matthews, J. Smelko, and K. Wiltberger, „Cross-scale predictive modeling of CHO cell culture growth and metabolites using Raman spectroscopy and multivariate analysis“, *Biotechnology Progress*, vol. 31, no. 2, pp. 566–577, 03/2015.

[76] B. N. Berry, T. M. Dobrowsky, R. C. Timson, R. Kshirsagar, T. Ryll, and K. Wiltberger, „Quick generation of Raman spectroscopy based in-process glucose control to influence biopharmaceutical protein product quality during mammalian cell culture“, *Biotechnology Progress*, vol. 32, no. 1, pp. 224–234, 2016.

[77] F. Feidl *et al.*, „A new flow cell and chemometric protocol for implementing in-line Raman spectroscopy in chromatography“, *Biotechnology Progress*, vol. 35, no. 5, 09/2019.

[78] 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*, vol. 118, no. 11, pp. 4255–4268, 2021.

[79] T. Starciuc, Y. Guinet, L. Paccou, and A. Hedoux, „Influence of a Small Amount of Glycerol on the Trehalose Bioprotective Action Analyzed In Situ During Freeze-Drying of Lyzozyme Formulations by Micro-Raman Spectroscopy“, *Journal of Pharmaceutical Sciences*, vol. 106, no. 10, pp. 2988–2997, 10/2017.

[80] D. Weber and J. Hubbuch, „Raman spectroscopy as a process analytical technology to investigate biopharmaceutical freeze concentration processes“, *Biotechnology and Bioengineering*, no. May, pp. 1–12, 2021.

[81] J. Tang, H. Jia, S. Mu, F. Gao, Q. Qin, and J. Wang, „Characterizing synergistic effect of coagulant aid and membrane fouling during coagulation-ultrafiltration via in-situ Raman spectroscopy and electrochemical impedance spectroscopy“, 2020.

[82] G. Févotte, „In situ Raman spectroscopy for in-line control of pharmaceutical crystallization and solids elaboration processes: A review“, *Chemical Engineering Research and Design*, vol. 85, no. 7 A, pp. 906–920, 2007.

[83] J. Cornel, C. Lindenberg, and M. Mazzotti, „Quantitative Application of in Situ ATR-FTIR and Raman Spectroscopy in Crystallization Processes“, *Industrial & Engineering Chemistry Research*, vol. 47, no. 14, pp. 4870–4882, 07/2008.

[84] Y. Hu, J. K. Liang, A. S. Myerson, and L. S. Taylor, „Crystallization monitoring by raman spectroscopy: Simultaneous measurement of desupersaturation profile and polymorphic form in flufenamic acid systems“, *Industrial and Engineering Chemistry Research*, vol. 44, no. 5, pp. 1233–1240, 03/2005.

[85] E. Simone and Z. K. Nagy, „A link between the ATR-UV/Vis and Raman spectra of zwitterionic solutions and the polymorphic outcome in cooling crystallization“, *CrystEngComm*, vol. 17, no. 34, pp. 6538–6547, 2015.

[86] E. Simone, A. N. Saleemi, N. Tonnon, and Z. K. Nagy, „Active polymorphic feedback control of crystallization processes using a combined raman and ATR-UV/Vis spectroscopy approach“, *Crystal Growth and Design*, vol. 14, no. 4, pp. 1839–1850, 2014.

[87] M. Streefland, D. E. Martens, E. C. Beuvery, and R. H. Wijffels, „Process analytical technology (PAT) tools for the cultivation step in biopharmaceutical production“, *Engineering in Life Sciences*, vol. 13, no. 3, pp. 212–223, 05/2013.

[88] 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*, vol. 412, no. 9, pp. 2047–2064, 04/2020.

[89] R. W. Kessler, „Perspectives in process analysis“, *Journal of Chemometrics*, vol. 27, no. 11, pp. 369–378, 2013.

[90] R. Bro and A. K. Smilde, „Principal component analysis“, *Analytical Methods*, vol. 6, no. 9, pp. 2812–2831, 2014.

[91] S. Wold, K. Esbensen, and P. Geladi, „Principal component analysis“, *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1-3, pp. 37–52, 08/1987.

[92] A. S. Rathore, N. Bhushan, and S. Hadpe, „Chemometrics applications in biotech processes: A review“, *Biotechnology Progress*, vol. 27, no. 2, pp. 307–315, 03/2011.

[93] T. D. Pham, C. Manapragada, Y. Sun, R. Bassett, and U. Aickelin, „A scoping review of supervised learning modelling and data-driven optimisation in monoclonal antibody process development“, *Digital Chemical Engineering*, vol. 7, p. 100 080, 06/2023.

[94] J. Chapman, R. Orrell-Trigg, K. Y. Kwoon, V. K. Truong, and D. Cozzolino, „A high-throughput and machine learning resistance monitoring system to determine the point of resistance for Escherichia coli with tetracycline: Combining UV-visible spectrophotometry with principal component analysis“, *Biotechnology and Bioengineering*, vol. 118, no. 4, pp. 1511–1519, 2021.

[95] M. Manahan, M. Nelson, J. J. Cacciatore, J. Weng, S. Xu, and J. Pollard, „Scale-down model qualification of ambr® 250 high-throughput mini-bioreactor system for two commercial-scale mAb processes“, *Biotechnology Progress*, vol. 35, no. 6, 11/2019.

[96] S. M. Mercier, B. Diepenbroek, M. C. Dalm, R. H. Wijffels, and M. Streefland, „Multivariate data analysis as a PAT tool for early bioprocess development data“, *Journal of Biotechnology*, vol. 167, no. 3, pp. 262–270, 09/2013.

[97] 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, 02/2013.

[98] E. Simone, A. N. Saleemi, and Z. K. Nagy, „In situ monitoring of polymorphic transformations using a composite sensor array of Raman, NIR, and ATR-UV/vis spectroscopy, FBRM, and PVM for an intelligent decision support system“, *Organic Process Research and Development*, vol. 19, no. 1, pp. 167–177, 2015.

[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.

[100] A. Tsopanoglou and I. Jiménez del Val, „Moving towards an era of hybrid modelling: advantages and challenges of coupling mechanistic and data-driven models for upstream pharmaceutical bioprocesses“, *Current Opinion in Chemical Engineering*, vol. 32, p. 100 691, 06/2021.

[101] 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“, 2019.

[102] L. O. Rodrigues, L. Vieira, J. P. Cardoso, and J. C. Menezes, „The use of NIR as a multi-parametric in situ monitoring technique in filamentous fermentation systems“, *Talanta*, vol. 75, no. 5, pp. 1356–1361, 06/2008.

[103] S. Andris, M. Rüdt, J. Rogalla, M. Wendeler, and J. Hubbuch, „Monitoring of antibody-drug conjugation reactions with UV/Vis spectroscopy“, *Journal of Biotechnology*, vol. 288, no. June, pp. 15–22, 2018.

[104] N. Hillebrandt, P. Vormittag, A. Dietrich, and J. Hubbuch, „Process monitoring framework for cross-flow diafiltration-based virus-like particle disassembly: Tracing product properties and filtration performance“, *Biotechnology and Bioengineering*, vol. 119, no. 6, pp. 1522–1538, 06/2022.

[105] Z. P. Chen *et al.*, „On-line monitoring of batch cooling crystallization of organic compounds using ATR-FTIR spectroscopy coupled with an advanced calibration method“, *Chemometrics and Intelligent Laboratory Systems*, vol. 96, no. 1, pp. 49–58, 03/2009.

[106] R. F. Li, X. Z. Wang, and S. B. Abebe, „Monitoring batch cooling crystallization using NIR: Development of calibration models using genetic algorithm and PLS“, *Particle and Particle Systems Characterization*, vol. 25, no. 4, pp. 314–327, 11/2008.

[107] A. C. F. Rumondor and L. S. Taylor, „Application of Partial Least-Squares (PLS) modeling in quantifying drug crystallinity in amorphous solid dispersions“, *International Journal of Pharmaceutics*, vol. 398, pp. 155–160, 2010.

[108] G. M. Escandar, P. C. Damiani, H. C. Goicoechea, and A. C. Olivieri, „A review of multivariate calibration methods applied to biomedical analysis“, *Microchemical Journal*, vol. 82, no. 1, pp. 29–42, 2006.

[109] R. Bro and H. A. Kiers, „A new efficient method for determining the number of components in PARAFAC models“, *Journal of Chemometrics*, vol. 17, no. 5, pp. 274–286, 2003.

[110] L. Yang, D. H. Han, B. M. Lee, and J. Hur, „Characterizing treated wastewaters of different industries using clustered fluorescence EEM-PARAFAC and FT-IR spectroscopy: Implications for downstream impact and source identification“, *Chemosphere*, vol. 127, pp. 222–228, 2015.

[111] J. Schuster, J. Huber, J. Stumme, A. Grieb, and M. Ernst, „Combining real-time fluorescence spectroscopy and flow cytometry to reveal new insights in DOC and cell characterization of drinking water“, *Frontiers in Environmental Chemistry*, vol. 3, p. 931 067, 08/2022.

[112] M. Steiner-Browne, S. Elcoroaristizabal, Y. Casamayou-Boucau, and A. G. Ryder, „Investigating native state fluorescence emission of Immunoglobulin G using polarized Excitation Emission Matrix (pEEM) spectroscopy and PARAFAC“, *Chemometrics and Intelligent Laboratory Systems*, vol. 185, no. December 2018, pp. 1–11, 2019.

[113] D. Ebrahimi, D. F. Kennedy, B. A. Messerle, and D. B. Hibbert, „High throughput screening arrays of rhodium and iridium complexes as catalysts for intramolecular hydroamination using parallel factor analysis“, *Analyst*, vol. 133, no. 6, pp. 817–822, 2008.

[114] J. M. Leitão and J. C. Esteves Da Silva, „PARAFAC and PARAFAC2 calibration models for antihypertensor Nifedipine quantification“, *Analytica Chimica Acta*, vol. 559, no. 2, pp. 271–280, 2006.

[115] Food and Drug Administration, „Guidance for Industry, PAT-A Framework for Innovative Pharmaceutical Development, Manufacturing and Quality Assurance“, U.S. Department of Health and Human Services, Tech. Rep. September, 2004.

[116] European Medicines Agency (EMA), *ICH guideline Q8 (R2) on pharmaceutical development EMA/CHMP/ICH 167068/2004*, 2017.

[117] A. S. 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.

[118] 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“, *Journal of Chromatography A*, vol. 1608, p. 460 410, 12/2019.

[119] J. E. Hales, S. Aoudjane, G. Aeppli, and P. A. Dalby, „Proof-of-concept analytical instrument for label-free optical deconvolution of protein species in a mixture“, *Journal of Chromatography A*, vol. 1641, p. 461 968, 2021.

[120] E. Simone, W. Zhang, and Z. K. Nagy, „Analysis of the crystallization process of a biopharmaceutical compound in the presence of impurities using process analytical technology (PAT) tools“, *Journal of Chemical Technology and Biotechnology*, vol. 91, no. 5, pp. 1461–1470, 2016.

[121] E. Simone, W. Zhang, and Z. K. Nagy, „Application of process analytical technology-based feedback control strategies to improve purity and size distribution in biopharmaceutical crystallization“, *Crystal Growth and Design*, vol. 15, no. 6, pp. 2908–2919, 2015.

[122] L. L. Simon, E. Simone, and K. Abbou Oucherif, „Crystallization process monitoring and control using process analytical technology“, in *Computer Aided Chemical Engineering*, 1st ed., vol. 41, Elsevier B.V., 2018, pp. 215–242.

[123] C. Stillhart and M. Kuentz, „Comparison of high-resolution ultrasonic resonator technology and Raman spectroscopy as novel process analytical tools for drug quantification in self-emulsifying drug delivery systems“, *Journal of Pharmaceutical and Biomedical Analysis*, vol. 59, no. 1, pp. 29–37, 2012.

[124] A. C. A. Roque *et al.*, „Anything but Conventional Chromatography Approaches in Bioseparation“, *Biotechnology Journal*, vol. 15, no. 8, pp. 1–8, 2020.

[125] R. dos Santos, A. L. Carvalho, and A. C. A. Roque, „Renaissance of protein crystallization and precipitation in biopharmaceuticals purification“, *Biotechnology Advances*, vol. 35, no. 1, pp. 41–50, 2017.

[126] Y. Zang, B. Kammerer, M. Eisenkolb, K. Lohr, and H. Kiefer, „Towards protein crystallization as a process step in downstream processing of therapeutic antibodies: screening and optimization at microbatch scale.“, *PloS one*, vol. 6, no. 9, pp. 1–8, 2011.

[127] B. Smejkal *et al.*, „Fast and scalable purification of a therapeutic full-length antibody based on process crystallization“, *Biotechnology and Bioengineering*, vol. 110, no. 9, pp. 2452–2461, 2013.

[128] P. Grob *et al.*, „Crystal Contact Engineering Enables Efficient Capture and Purification of an Oxidoreductase by Technical Crystallization“, *Biotechnology Journal*, vol. 2000010, 2020.

[129] S. K. Basu, C. P. Govardhan, C. W. Jung, and A. L. Margolin, „Protein crystals for the delivery of biopharmaceuticals“, *Expert Opinion on Biological Therapy*, vol. 4, no. 3, pp. 301–317, 2004.

[130] M. X. Yang *et al.*, „Crystalline monoclonal antibodies for subcutaneous delivery“, *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 12, pp. 6934–6939, 2003.

[131] B. Shenoy, Y. Wang, W. Shan, and A. L. Margolin, „Stability of crystalline proteins“, *Biotechnology and Bioengineering*, vol. 73, no. 5, pp. 358–369, 2001.

[132] S. Pechenov, B. Shenoy, M. X. Yang, S. K. Basu, and A. L. Margolin, „Injectable controlled release formulations incorporating protein crystals“, *Journal of Controlled Release*, vol. 96, no. 1, pp. 149–158, 2004.

[133] M. L. Brader *et al.*, „Hybrid insulin cocrystals for controlled release delivery“, *Nature Biotechnology*, vol. 20, no. 8, pp. 800–804, 2002.

[134] H. Yang *et al.*, „Optimization of vapor diffusion conditions for anti-CD20 crystallization and scale-up to meso batch“, *Crystals*, vol. 9, no. 5, pp. 1–13, 2019.

[135] A. K. Wöll, J. Schütz, J. Zabel, and J. Hubbuch, „Analysis of phase behavior and morphology during freeze-thaw applications of lysozyme“, *International Journal of Pharmaceutics*, vol. 555, no. November 2018, pp. 153–164, 2019.

[136] E. L. Forsythe, E. H. Snell, and M. L. Pusey, „Crystallization of chicken egg-white lysozyme from ammonium sulfate“, *Acta Crystallographica Section D: Biological Crystallography*, vol. 53, no. 6, pp. 795–797, 1997.

[137] E. L. Forsythe, E. H. Snell, C. C. Malone, and M. L. Pusey, „Crystallization of chicken egg white lysozyme from assorted sulfate salts“, *Journal of Crystal Growth*, vol. 196, no. 2-4, pp. 332–343, 1999.

[138] A. N. Saleemi, C. D. Rielly, and Z. K. Nagy, „Comparative investigation of supersaturation and automated direct nucleation control of crystal size distributions using ATR-UV/vis spectroscopy and FBRM“, *Crystal Growth and Design*, vol. 12, no. 4, pp. 1792–1807, 2012.

[139] M. Groß and M. Kind, „Bulk Crystallization of Proteins by Low-Pressure Water Evaporation“, *Chemical Engineering and Technology*, vol. 39, no. 8, pp. 1483–1489, 2016.

[140] X. Li, W. Chen, H. Yang, Z. Yang, and J. Y. Heng, „Protein crystal occurrence domains in selective protein crystallisation for bio-separation“, *CrystEngComm*, vol. 22, no. 27, pp. 4566–4572, 2020.

[141] J. Hermann, P. Nowotny, T. E. Schrader, P. Biggel, D. Hekmat, and D. Weuster-Botz, „Neutron and X-ray crystal structures of Lactobacillus brevis alcohol dehydrogenase reveal new insights into hydrogen-bonding pathways“, *Acta Crystallographica Section F: Structural Biology Communications*, vol. 74, no. 12, pp. 754–764, 2018.

[142] A. Proteau, R. Shi, and M. Cygler, „Application of dynamic light scattering in protein crystallization“, *Current Protocols in Protein Science*, no. SUPPL. 61, 2010.

[143] A. George and W. W. Wilson, „Predicting protein crystallization from a dilute solution property“, *Acta Crystallographica Section D Biological Crystallography*, vol. 50, no. 4, pp. 361–365, 1994.

[144] R. B. Zhou, X. L. Lu, C. Dong, F. Ahmad, C. Y. Zhang, and D. C. Yin, „Application of protein crystallization methodologies to enhance the solubility, stability and monodispersity of proteins“, *CrystEngComm*, vol. 20, no. 14, pp. 1923–1927, 2018.

[145] A. N. Saleemi, C. D. Rielly, and Z. K. Nagy, „Monitoring of the combined cooling and antisolvent crystallisation of mixtures of aminobenzoic acid isomers using ATR-UV/vis spectroscopy and FBRM“, *Chemical Engineering Science*, vol. 77, pp. 122–129, 2012.

[146] K. Tacsi *et al.*, „Polymorphic Concentration Control for Crystallization Using Raman and Attenuated Total Reflectance Ultraviolet Visible Spectroscopy“, *Crystal Growth and Design*, vol. 20, no. 1, pp. 73–86, 2020.

[147] E. Simone, A. N. Saleemi, and Z. K. Nagy, „Application of quantitative Raman spectroscopy for the monitoring of polymorphic transformation in crystallization processes using a good calibration practice procedure“, *Chemical Engineering Research and Design*, vol. 92, no. 4, pp. 594–611, 2014.

[148] Z. Zhang and Baixiaofeng, „Comparison about the three central composite designs with simulation“, *Proceedings - International Conference on Advanced Computer Control, ICACC 2009*, no. 3, pp. 163–167, 2009.

[149] A. Savitzky and M. J. E. Golay, „Smoothing and Differentiation of Data by Simplified Least Squares Procedures.“, *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639, 07/1964.

[150] C. M. Andersen and R. Bro, „Variable selection in regression-a tutorial“, *Journal of Chemometrics*, vol. 24, no. 11-12, pp. 728–737, 11/2010.

[151] M. Rüdt, S. Andris, R. Schiemer, and J. Hubbuch, „Factorization of preparative protein chromatograms with hard-constraint multivariate curve resolution and second-derivative pretreatment“, *Journal of Chromatography A*, vol. 1585, pp. 152–160, 2019.

[152] C. Bosch Ojeda and F. Sanchez Rojas, „Recent developments in derivative ultraviolet/visible absorption spectrophotometry“, *Analytica Chimica Acta*, vol. 518, no. 1-2, pp. 1–24, 2004.

[153] 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.

[154] R. Esfandiary, J. S. Hunjan, G. H. Lushington, S. B. Joshi, and C. R. Middaugh, „Temperature dependent 2nd derivative absorbance spectroscopy of aromatic amino acids as a probe of protein dynamics“, *Protein Science*, vol. 18, no. 12, pp. 2603–2614, 2009.

[155] J. Blaffert, H. H. Haeri, M. Blech, D. Hinderberger, and P. Garidel, „Spectroscopic methods for assessing the molecular origins of macroscopic solution properties of highly concentrated liquid protein solutions“, *Analytical Biochemistry*, vol. 561-562, pp. 70–88, 2018.

[156] T. Tojo, K. Hamaguchi, M. Imanashi, and T. Amano, „Structure of Lysozyme: XI. Spectrophotometric Titration of Tyrosyl Groups of Hen and Duck Egg-white Lysozyme*“, *The Journal of Biochemistry*, vol. 60, no. 5, pp. 538–542, 11/1966.

[157] T. Arakawa and S. N. Timasheff, „Mechanism of Protein Salting In and Salting Out by Divalent Cation Salts: Balance between Hydration and Salt Binding“, *Biochemistry*, vol. 23, no. 25, pp. 5912–5923, 1984.

[158] J. R. Luft, J. R. Wolfley, and E. H. Snell, „What's in a drop? Correlating observations and outcomes to guide macromolecular crystallization experiments“, *Crystal Growth and Design*, vol. 11, no. 3, pp. 651–663, 2011.

[159] J. G. Elvin, R. G. Couston, and C. F. Van Der Walle, „Therapeutic antibodies: Market considerations, disease targets and bioprocessing“, *International Journal of Pharmaceutics*, vol. 440, no. 1, pp. 83–98, 2013.

[160] M. Saxena, S. H. van der Burg, C. J. Melief, and N. Bhardwaj, „Therapeutic cancer vaccines“, *Nature Reviews Cancer*, vol. 21, no. 6, pp. 360–378, 2021.

[161] L. H. Lua, N. K. Connors, F. Sainsbury, Y. P. Chuan, N. Wibowo, and A. P. Middelberg, „Bioengineering virus-like particles as vaccines“, *Biotechnology and Bioengineering*, vol. 111, no. 3, pp. 425–440, 2014.

[162] P. H. Tobin, D. H. Richards, R. A. Callender, and C. J. Wilson, „Protein engineering: a new frontier for biological therapeutics.“, *Current drug metabolism*, vol. 15, no. 7, pp. 743–56, 2014.

[163] P. McDonald, C. Victa, J. N. Carter-Franklin, and R. Fahrner, „Selective antibody precipitation using polyelectrolytes: A novel approach to the purification of monoclonal antibodies“, *Biotechnology and Bioengineering*, vol. 102, no. 4, pp. 1141–1151, 2009.

[164] F. Perosa, R. Carbone, S. Ferrone, and F. Dammacco, „Purification of human immunoglobulins by sequential precipitation with caprylic acid and ammonium sulphate“, *Journal of Immunological Methods*, vol. 128, no. 1, pp. 9–16, 1990.

[165] J. Acquarelli, T. van Laarhoven, J. Gerretzen, T. N. Tran, L. M. Buydens, and E. Marchiori, „Convolutional neural networks for vibrational spectroscopic data analysis“, *Analytica Chimica Acta*, vol. 954, pp. 22–31, 2017.

[166] T. Chen, J. Morris, and E. Martin, „Gaussian process regression for multivariate spectroscopic calibration“, *Chemometrics and Intelligent Laboratory Systems*, vol. 87, no. 1, pp. 59–71, 2007.

[167] B. Szilagyi, A. Eren, J. L. Quon, C. D. Papageorgiou, and Z. K. Nagy, „Application of Model-Free and Model-Based Quality-by-Control (QbC) for the Efficient Design of Pharmaceutical Crystallization Processes“, *Crystal Growth and Design*, vol. 20, no. 6, pp. 3979–3996, 2020.

[168] M. Trampuž, D. Teslić, and B. Likozar, „Process analytical technology-based (PAT) model simulations of a combined cooling, seeded and antisolvent crystallization of an active pharmaceutical ingredient (API)“, *Powder Technology*, vol. 366, pp. 873–890, 2020.

[169] M. Trampuž, D. Teslić, and B. Likozar, „Crystal-size distribution-based dynamic process modelling, optimization, and scaling for seeded batch cooling crystallization of Active Pharmaceutical Ingredients (API)“, *Chemical Engineering Research and Design*, vol. 165, pp. 254–269, 2021.

[170] A. C. Olivieri, „Analytical figures of merit: From univariate to multiway calibration“, *Chemical Reviews*, vol. 114, no. 10, pp. 5358–5378, 2014.

[171] M. B. Anzardi, J. A. Arancibia, and A. C. Olivieri, „Processing multi-way chromatographic data for analytical calibration, classification and discrimination: A successful marriage between separation science and chemometrics“, *TrAC - Trends in Analytical Chemistry*, vol. 134, pp. 2–10, 2021.

[172] R. Bro, „PARAFAC. Tutorial and applications“, *Chemometrics and Intelligent Laboratory Systems*, vol. 38, no. 2, pp. 149–171, 1997.

[173] H. Yu, L. Guo, M. Kharbach, and W. Han, „Multi-way analysis coupled with near-infrared spectroscopy in food industry: Models and applications“, *Foods*, vol. 10, no. 4, 2021.

[174] M. A. B. Levi, I. S. Scarminio, R. J. Poppi, and M. G. Trevisan, „Three-way chemometric method study and UV-Vis absorbance for the study of simultaneous degradation of anthocyanins in flowers of the Hibiscus rosa-sinensys species“, *Talanta*, vol. 62, no. 2, pp. 299–305, 2004.

[175] C. M. Andersen and R. Bro, „Practical aspects of PARAFAC modeling of fluorescence excitation-emission data“, *Journal of Chemometrics*, vol. 17, no. 4, pp. 200–215, 2003.

[176] M. Ortiz, L. Sarabia, M. Sánchez, A. Herrero, S. Sanllorente, and C. Reguera, „Usefulness of PARAFAC for the Quantification, Identification, and Description of Analytical Data“, in *Fundamentals and Analytical Applications of Multiway Calibration*, Arsenio Muñoz de la Peña, H. C. Goicoechea, G. M. Escandar, and A. C. Olivieri, Eds., vol. 29, Elsevier, 2015, pp. 37–81.

[177] N. R. Marsili, A. Lista, B. S. Fernandez Band, H. C. Goicoechea, and A. C. Olivieri, „New method for the determination of benzoic and sorbic acids in commercial orange juices based on second-order spectrophotometric data generated by a pH gradient flow injection technique.“, *Journal of agricultural and food chemistry*, vol. 52, no. 9, pp. 2479–2484, 05/2004.

[178] A. Niazi, J. Ghasemi, and A. Yazdanipour, „PARAFAC decomposition of three-way kinetic-spectrophotometric spectral matrices based on phosphomolymbdenum blue complex chemistry for nitrite determination in water and meat samples“, *Analytical Letters*, vol. 38, no. 14, pp. 2377–2392, 2005.

[179] I. García, M. C. Ortiz, L. Sarabia, and J. M. Aldama, „Validation of an analytical method to determine sulfamides in kidney by HPLC-DAD and PARAFAC2 with first-order derivative chromatograms“, *Analytica Chimica Acta*, vol. 587, no. 2, pp. 222–234, 2007.

[180] M. C. Ortiz *et al.*, „Three-way PARAFAC decomposition of chromatographic data for the unequivocal identification and quantification of compounds in a regulatory framework“, *Chemometrics and Intelligent Laboratory Systems*, vol. 200, p. 104 003, 2020.

[181] C. H. Wegner, I. Zimmermann, and J. Hubbuch, „Rapid Analysis for Multicomponent High-Throughput Crystallization Screening: Combination of UV–Vis Spectroscopy and Chemometrics“, *Crystal Growth & Design*, vol. 22, no. 2, pp. 1054–1065, 02/2022.

[182] A. Zlotnick *et al.*, „Dimorphism of hepatitis B virus capsids is strongly influenced by the C-terminus of the capsid protein“, *Biochemistry*, vol. 35, no. 23, pp. 7412–7421, 1996.

[183] N. Hillebrandt, P. Vormittag, A. Dietrich, C. H. Wegner, and J. Hubbuch, „Process development for cross-flow diafiltration-based VLP disassembly: A novel high-throughput screening approach.“, *Biotechnology and bioengineering*, vol. 118, no. 10, pp. 1–15, 10/2021.

[184] C. A. Andersson and R. Bro, „The N-way Toolbox for MATLAB“, *Chemometrics and Intelligent Laboratory Systems*, vol. 52, no. 1, pp. 1–4, 2000.

[185] K. R. Murphy, C. A. Stedmon, D. Graeber, and R. Bro, „Fluorescence spectroscopy and multi-way techniques. PARAFAC“, *Analytical Methods*, vol. 5, no. 23, pp. 6557–6566, 2013.

[186] M. H. Van Benthem, T. J. Keller, G. D. Gillispie, and S. A. DeJong, „Getting to the core of PARAFAC2, a nonnegative approach“, *Chemometrics and Intelligent Laboratory Systems*, vol. 206, pp. 1–33, 2020.

[187] A. Smilde, R. Bro, and P. Geladi, *Multi-way Analysis with Applications in the Chemical Sciences Multi-way*. John Wiley & Sons, Ltd, 2004.

[188] A. K. Smilde and D. A. Doornbos, „Simple validatory tools for judging the predictive performance of parafac and three-way PLS“, *Journal of Chemometrics*, vol. 6, no. 1, pp. 11–28, 01/1992.

[189] R. Bro and A. K. Smilde, „Centering and scaling in component analysis“, *Journal of Chemometrics*, vol. 17, no. 1, pp. 16–33, 2003.

[190] R. Bro, „Multi-way analysis in the food industry“, Ph.D. dissertation, Amsterdam, 1998.

[191] P. Wingfield, „Protein precipitation using ammonium sulfate.“, *Current protocols in protein science*, vol. Appendix 3, Appendix 3F, 05/1998.

[192] R. R. Burgess, „Chapter 20 Protein Precipitation Techniques“, *Methods in Enzymology*, vol. 463, no. C, pp. 331–342, 2009.

[193] H. Mach, J. A. Thomson, and C. R. Middaugh, „Quantitative analysis of protein mixtures by second derivative absorption spectroscopy“, *Analytical Biochemistry*, vol. 181, no. 1, pp. 79–85, 1989.

[194] K. Raval and H. Patel, „Review on Common Observed HPLC Troubleshooting Problems“, *International Journal of Pharma Research and Health Sciences*, vol. 8, no. 4, pp. 3195–3202, 2020.

[195] C. Haas and J. Drenth, „Understanding protein crystallization on the basis of the phase diagram“, *Journal of Crystal Growth*, vol. 196, pp. 388–394, 1999.

[196] N. Chayen, J. Akins, S. Campbell-Smith, and D. M. Blow, „Solubility of glucose isomerase in ammonium sulphate solutions“, *Journal of Crystal Growth*, vol. 90, no. 1-3, pp. 112–116, 1988.

[197] D. Vivarès, L. Belloni, A. Tardieu, and F. Bonneté, „Catching the PEG-induced attractive interaction between proteins“, *European Physical Journal E*, vol. 9, no. 1, pp. 15–25, 2002.

[198] M. E. Klijn and J. Hubbuch, „Application of ultraviolet, visible, and infrared light imaging in protein-based biopharmaceutical formulation characterization and development studies“, *European Journal of Pharmaceutics and Biopharmaceutics*, vol. 165, pp. 319–336, 2021.

[199] C. H. Wegner and J. Hubbuch, „Calibration-free PAT: Locating selective crystallization or precipitation sweet spot in screenings with multi-way PARAFAC models“, *Frontiers in Bioengineering and Biotechnology*, vol. 10, no. December, p. 1 051 129, 12/2022.

[200] Lawson DM *et al.*, „Solving the structure of human H ferritin by genetically engineering intermolecular crystal contacts“, *Nature*, vol. 349, no. 6309, pp. 541–544, 1991.

[201] H. E. McElroy, G. W. Sisson, W. E. Schoettlin, R. M. Aust, and J. E. Villafranca, „Studies on engineering crystallizability by mutation of surface residues of human thymidylate synthase“, *Journal of Crystal Growth*, vol. 122, no. 1-4, pp. 265–272, 1992.

[202] C. Charron, D. Kern, and R. Giegé, „Crystal contacts engineering of aspartyl-tRNA synthetase from Thermus thermophilus: Effects on crystallizability“, *Acta Crystallographica Section D: Biological Crystallography*, vol. 58, no. 10 II, pp. 1729–1733, 2002.

[203] B. Walla, D. Bischoff, R. Janowski, N. Von Den Eichen, D. Niessing, and D. Weuster-Botz, „Transfer of a rational crystal contact engineering strategy between diverse alcohol dehydrogenases“, *Crystals*, vol. 11, no. 8, pp. 8–12, 2021.

[204] W. Tian, W. Li, and H. Yang, „Protein Nucleation and Crystallization Process with Process Analytical Technologies in a Batch Crystallizer“, *Crystal Growth & Design*, vol. 23, no. 7, pp. 5181–5193, 07/2023.

[205] H. Yang, W. Chen, P. Peczulis, and J. Y. Heng, „Development and Workflow of a Continuous Protein Crystallization Process: A Case of Lysozyme“, *Crystal Growth and Design*, vol. 19, no. 2, pp. 983–991, 2019.

[206] R. Chen, J. Weng, S. F. Chow, and R. Lakerveld, „Integrated Continuous Crystallization and Spray Drying of Insulin for Pulmonary Drug Delivery“, *Crystal Growth and Design*, vol. 21, no. 1, pp. 501–511, 2021.

[207] R. A. Judge, M. R. Johns, and E. T. White, „Protein purification by bulk crystallization: The recovery of ovalbumin“, *Biotechnology and Bioengineering*, vol. 48, no. 4, pp. 316–323, 1995.

[208] I. Jul-Jørgensen, R. Oliver, K. Gernaey, and C. Hundahl, „Modernizing non-classical protein crystallization through industry 4.0: Advanced monitoring and modelling utilizing process analytical technology“, *Chemical Engineering Research and Design*, vol. 204, pp. 382–389, 04/2024.

[209] T. Takakura *et al.*, „High-level expression and bulk crystallization of recombinant L-methionine γ-lyase, an anticancer agent“, *Applied Microbiology and Biotechnology*, vol. 70, no. 2, pp. 183–192, 2006.

[210] D. Hekmat, M. Huber, C. Lohse, N. Von Den Eichen, and D. Weuster-Botz, „Continuous Crystallization of Proteins in a Stirred Classified Product Removal Tank with a Tubular Reactor in Bypass“, *Crystal Growth and Design*, vol. 17, no. 8, pp. 4162–4169, 2017.

[211] J. Mercado, M. Alcalà, K. M. Karry, J. L. Ríos-Steiner, and R. J. Romañach, „Design and in-line raman spectroscopic monitoring of a protein batch crystallization process“, *Journal of Pharmaceutical Innovation*, vol. 3, no. 4, pp. 271–279, 2008.

[212] M. R. Bakar, Z. K. Nagy, and C. D. Rielly, „Seeded batch cooling crystallization with temperature cycling for the control of size uniformity and polymorphic purity of sulfathiazole crystals“, *Organic Process Research and Development*, vol. 13, no. 6, pp. 1343–1356, 2009.

[213] R. Schiemer, J. T. Weggen, K. M. Schmitt, and J. Hubbuch, „An adaptive soft-sensor for advanced real-time monitoring of an antibody-drug conjugation reaction“, *Biotechnology and Bioengineering*, pp. 1–15, 05/2023.

[214] T. Bocklitz, A. Walter, K. Hartmann, P. Rösch, and J. Popp, „How to pre-process Raman spectra for reliable and stable models?“, *Analytica Chimica Acta*, vol. 704, no. 1-2, pp. 47–56, 2011.

[215] M. Lin, Y. Wu, and S. Rohani, „Simultaneous Measurement of Solution Concentration and Slurry Density by Raman Spectroscopy with Artificial Neural Network“, *Crystal Growth and Design*, vol. 20, no. 3, pp. 1752–1759, 2020.

[216] T. Togkalidou, H. H. Tung, Y. Sun, A. Andrews, and R. D. Braatz, „Solution concentration prediction for pharmaceutical crystallization processes using robust chemometrics and ATR FTIR spectroscopy“, *Organic Process Research and Development*, vol. 6, no. 3, pp. 317–322, 2002.

[217] F. Wang, J. A. Wachter, F. J. Antosz, and K. A. Berglund, „An investigation of solvent-mediated polymorphic transformation of progesterone using in situ Raman spectroscopy“, *Organic Process Research and Development*, vol. 4, no. 5, pp. 391–395, 2000.

[218] A. Schmideder, J. H. Cremer, and D. Weuster-Botz, „Parallel Steady State Studies on a Milliliter Scale Accelerate Fed-Batch Bioprocess Design for Recombinant Protein Production With Escherichia coli“, *American Institute of Chemical Engineers Biotechnol. Prog*, vol. 32, pp. 1426–1435, 2016.

[219] D. Bischoff, B. Walla, and D. Weuster-Botz, „Machine learning-based protein crystal detection for monitoring of crystallization processes enabled with large-scale synthetic data sets of photorealistic images“, *Analytical and Bioanalytical Chemistry*, vol. 414, no. 21, pp. 6379–6391, 2022.

[220] E. Gasteiger *et al.*, „Protein Identification and Analysis Tools on the ExPASy Server“, in *The Proteomics Protocols Handbook*, J. M. Walker, Ed., Totowa, NJ: Humana Press, 2005, pp. 571–607.

[221] R. W. Kennard and L. A. Stone, „Computer Aided Design of Experiments“, *Technometrics*, vol. 11, no. 1, pp. 137–148, 1969.

[222] F. Westad and F. Marini, „Variable Selection and Redundancy in Multivariate Regression Models“, *Frontiers in Analytical Science*, vol. 2, 06/2022.

[223] S. Kucheryavskiy, „mdatools – R package for chemometrics“, *Chemometrics and Intelligent Laboratory Systems*, vol. 198, no. January, p. 103 937, 2020.

[224] B. Smejkal, „Aufreinigung und Formulierung eines therapeutischen Antikörpers mittels Kristallisation“, Ph.D. dissertation, 03/2013.

[225] B. Van Eerdenbrugh, D. E. Alonzo, and L. S. Taylor, „Influence of particle size on the ultraviolet spectrum of particulate-containing solutions: Implications for in-situ concentration monitoring using UV/Vis fiber-optic probes“, *Pharmaceutical Research*, vol. 28, no. 7, pp. 1643–1652, 07/2011.

List of Figures

List of Tables

Abbreviations

2D	two-dimensional	11
3D	three-dimensional	51
4D	four-dimensional	73
ADC	antibody-drug conjugate	103
AMS	ammonium sulfate	54
API	active pharmaceutical ingredient	10
ATR	attenuated total reflectance	78
ATR	attenuated total reflection	78
CA	cellulose acetate	79
CCS	cell culture supernatant	53
CEX	cation-exchange chromatography	55
CFF	cross-flow filtration	101
CHO	chinese hamster ovary	53
CNN	convolutional neural network	51
CORCONDIA	core consistency diagnostic	57
CPP	critical process parameter	17
CQA	critical quality attribute	102
CV	column volume	55
CytC	cytochrome C	52
DAD	diode array detector	51
DLS	dynamic light scattering	31
DoE	design of experiments	18
DSP	downstream processing	99

DTT	dithiothreitol	55
E.coli	*Escherichia coli*	100
ELISA	enzyme-linked immunosorbent assay	101
EMA	European Medicines Agency	17
FBRM	focused beam reflectance measurement	31
FDA	U.S. Food and Drug Administration	77
FIA	flow injection analysis	51
FTIR	fourier-transform infrared spectroscopy	77
GA	genetic algorithm	83
GPR	gaussian process regression	78
GRAM	generalized rank annihilation method	51
HBcAg	Hepatitis B core antigen	54
HCCF	harvest cell culture fluid	100
HCl	hydrochloric acid	79
HCP	host cell protein	101
HEPES	4-2-hydroxyethyl-1-piperazineethanesulfonic acid	79
HPLC	high-performance liquid chromatography	125
HT	high-throughput	99
IMAC	immobilized metal ion affinity chromatography	101
IPTG	isopropyl β-d-1-thiogalactopyranoside	79
KCl	potassium chloride	53
KH2PO4	potassium dihydrogenphosphate	54
KS	Kennard-Stone	153
***Lb*ADH**	*Lactobacillus brevis* alcohol dehydrogenase	96
***Lk*ADH**	*Lactobacillus kefir* alcohol dehydrogenase	101
LDS	lithium dodecyl sulfate	55
Lys	lysozyme	52
mAb	monoclonal antibody	100
MCB	multi-component buffer	53
MES	2-(N-morpholino)ethanesulfonic acid	55
$MgCl_2$	magnesium chloride	79
ML	machine learning	125
MLR	multiple linear regression	78

mPES	modified polyethersulfone	82
MS	mass spectrometry	51
MVDA	multi-variate data analysis	17
MWCO	molecular weight cut-off	79
Na2HPO4	disodium hydrogen phosphate	54
NaCl	sodium chloride	79
NaOH	sodium hydroxide	79
NMR	nuclear magnetic resonance	94
N-PLS	multi-way partial least-squares	51
PARAFAC	parallel factor analysis	100
PAT	process analytical technology	99
PBS	phosphate-buffered saline	53
PC	principal component	89
PCA	principal component analysis	97
PC-ANN	principal component artificial neural networks	78
PCR	principal component regression	78
PDB	protein data bank	79
PEG	polyethylene glycol	148
PEG MME 550	polyethylene glycol monomethyl ether 550	79
Phe	Phenyalanine	94
pI	isoelectric point	3
PLS	partial least squares	100
Q^2	predictive relevance	90
QbD	quality by design	99
R^2	coefficient of determination	90
RD	redissolution	79
RibA	ribonuclease A	52
$RMSECV$	root mean squared error of cross-validation	90
$RMSEP$	root mean squared error of prediction	90
SDS-PAGE	sodium dodecyl sulfate–polyacrylamide gel electrophoresis	101
SG	Savitzky Golay	83
TMP	transmembrane pressure	82
Tris	tris(hydroxymethyl)aminomethane	79

Symbols

Symbol	Meaning
α	distance factor
ϵ	extinction coefficient
γ	interfacial free energy
λ	wavelength
ν	wavenumber shift
A_λ	wavelength specific absorption
$\hat{A}_\lambda$	predicted wavelength specific absorption
A_j	loading matrix in the dimension j
$a_{j,f}$	loading vectors of component f in the dimensions j
B_k	loading matrix in dimension k
B_{reg}	regression coefficient matrix
$b_{k,f}$	loading vectors of component f in the dimensions k
C_l	loading matrix in the dimension l
$c_{l,f}$	loading vectors of component f in the dimensions l
c_f	concentration of component f
$\hat{c}_f$	predicted concentration of component f
$\overline{c}_f$	stable concentration of component f per row in phase diagram
d_{path}	path length
E	energy
E_Z	error matrix of matrix Z
$e_{j,k,l}$	error element of a 3D tensor with the dimensions $j,\ k,\ l$
f_{well}	well specific correction factor in phase diagrams
G	Gibbs free energy
k_{B}	Boltzmann constant
m	number of variables in X

Continued on next page

Symbol	Meaning
n_{Batch}	number of samples
n	number of observations
o	number of variables in Y
P	loadings matrix of X
$P_{f,\text{well}}$	well specific purity of component f
Q	loadings matrix of Y
Q^2	predictive relevance
R^2	coefficient of determination
$RMSE$	root mean squared error
r	radius
r_{crit}	critical nucleus radius
S	supersaturation
T	scores matrix of X
$T_{\circ\text{C}}$	temperature
t	time
U	scores matrix of Y
$\dot{V}$	flow rate
$\overline{v}$	molecular volume
X	input data matrix
$x_{j,k,l}$	element of a 3D tensor with the dimensions $j,\ k,\ l$
Y	response data matrix
$Y_{f,\text{well}}$	well specific yield of component f

A3

Appendix Chapter 3

Rapid analysis for multi-component high-throughput crystallization screening: Combination of UV/Vis spectroscopy and chemometrics

Christina Henriette Wegner[1], Ines Zimmermann[2] and Jürgen Hubbuch[1]

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

[2] Chair of Bioseparation Engineering, Technical University of Munich, Germany

A3.1 Explanation of $\overline{c}_{\mathrm{PLS},i,\mathrm{stable}}$ and $Y_{i,j}$ calculation in the phase diagrams

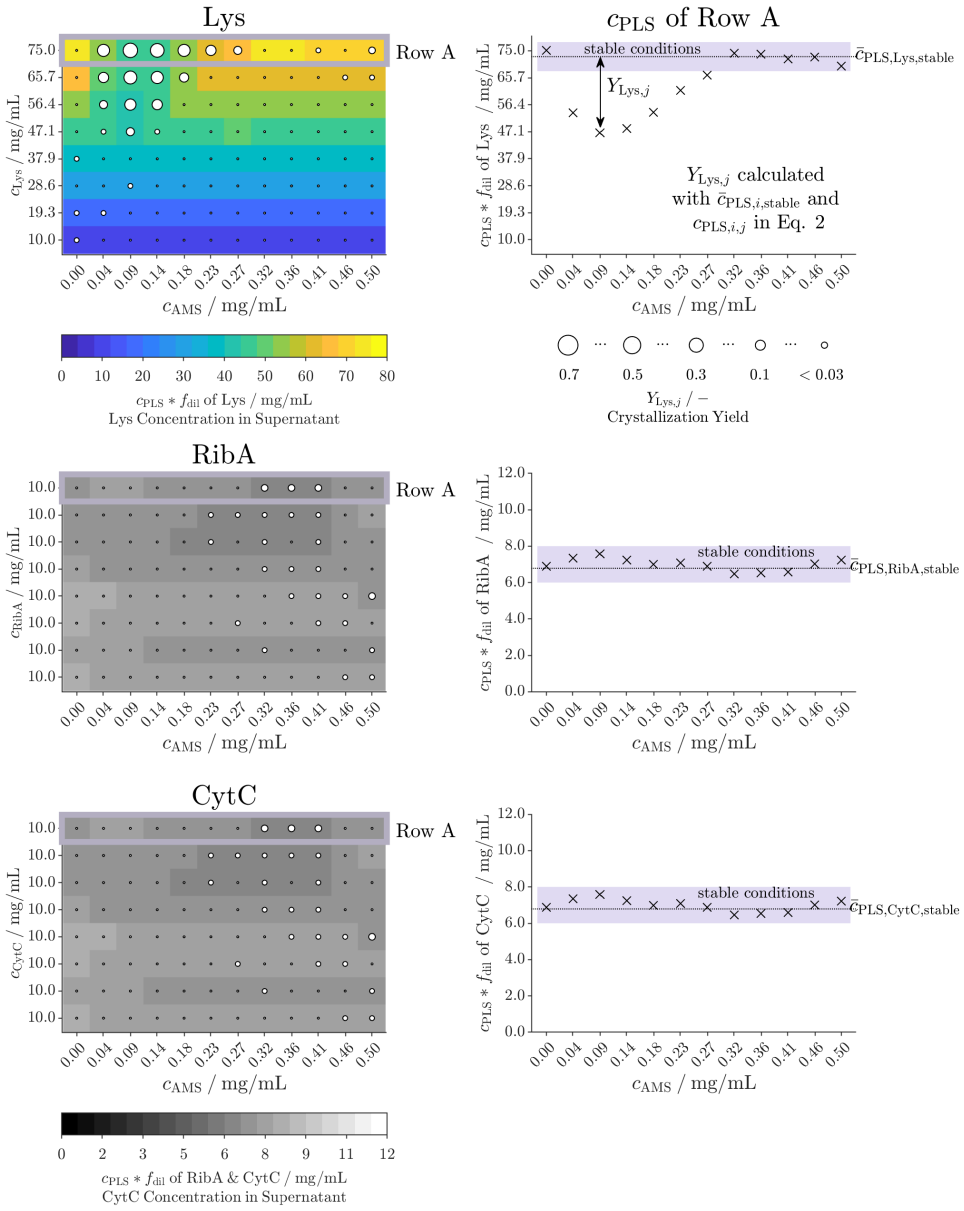


Figure A3.1 The relation between the data displayed in the phase diagram and the calculation of the yield $Y_{i,j}$ of each protein i and well j is displayed. The stable conditions are used to calculate the mean concentration of stable conditions per row $\overline{c}_{\mathrm{PLS},i,\mathrm{stable}}$. The final protein-specific concentration of each protein in row A is displayed over varying ammonium sulfate concentration c_{AMS}. Exemplarily, the calculated crystal yield of Lysozyme (Lys) is illustrated with an arrow for one condition.

A3.2 Analytical cation exchange chromatography gradient method

The cation exchange chromatography (CEX) column ProSwift SCX-1S 4.6x50mm column (Thermo Fisher Scientific Inc.) was used in a Dionex Ultimate 3000 RS ultra high performance chromatography system (UHPLC, Thermo Fisher Scientific Inc.). The column was loaded with 20 L sample in a low salt buffer (20 mM Tris, pH 8.0) for 0.5 min. The elution was performed using a high salt buffer (20 mM Tris, 1000 mM NaCl, pH 8) with a gradient to 70 mM NaCl for 2 min, a steeper gradient to 1000 mM NaCl for 3.1 min, a hold step for 0.5 min, a hold step at 500 mM NaCl for 0.5 min, a gradient to 1000 mM NaCl for 1 min, a second hold step for 0.5 min and re-equilibration for 2.5 min.

A3.3 Recorded UV/Vis spectral data

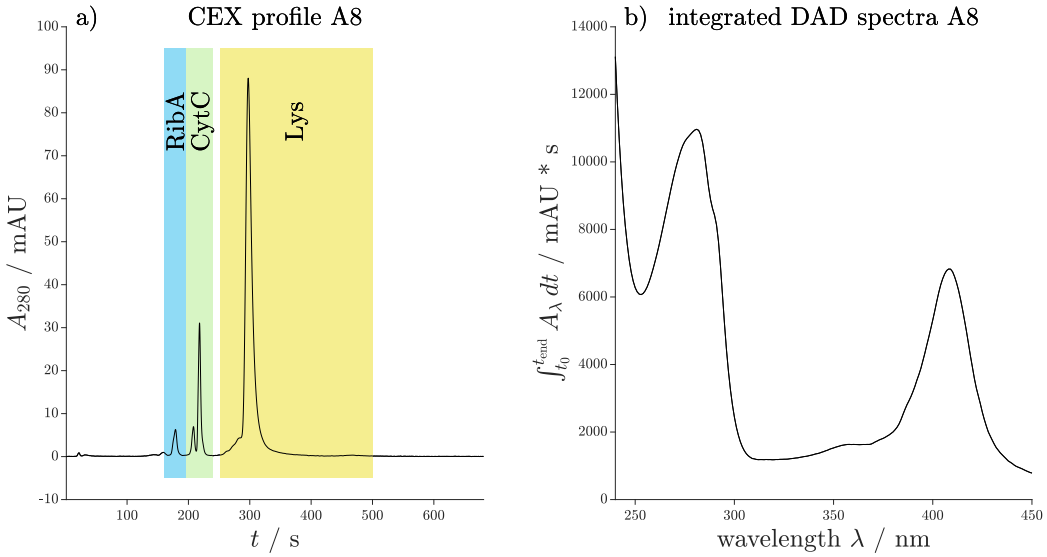


Figure A3.2 Exemplarily for the sample A8 in a phase diagram at pH 9, the chromatogram at wavelength of 280 nm of the CEX reference analytic is shown over time t in a). The colored areas illustrate the integration area for each species. In b), the time-wise summed up spectra of the corresponding DAD analysis is depicted. The starting and end of the analysis time are t_0 and t_{end}, respectively, and are used for the integration.

A3.4 Sensitivity and specificity equation

In this work, the *sensitivity* provides information on the probability that the developed method detects crystallizing conditions in the screening correctly. The variable n is the quantity of true-positives, true-negatives, false-positives, and false-negatives during the screening (see subchapter 3.2.3 for screening experiment and image scoring). Visible light images, taken at the end of the incubation of the phase diagrams, served as the validation method.

$$sensitivity = \frac{n_{\text{true-positives}}}{n_{\text{true-positives}} + n_{\text{false-negatives}}} \tag{A3.1}$$

The *specificity* provides information on the probability that the developed method detects stable conditions in the screening correctly.

$$specificity = \frac{n_{\text{true-negatives}}}{n_{\text{true-negatives}} + n_{\text{false-positives}}} \tag{A3.2}$$

A3.5 Image scoring analysis of the phase diagram

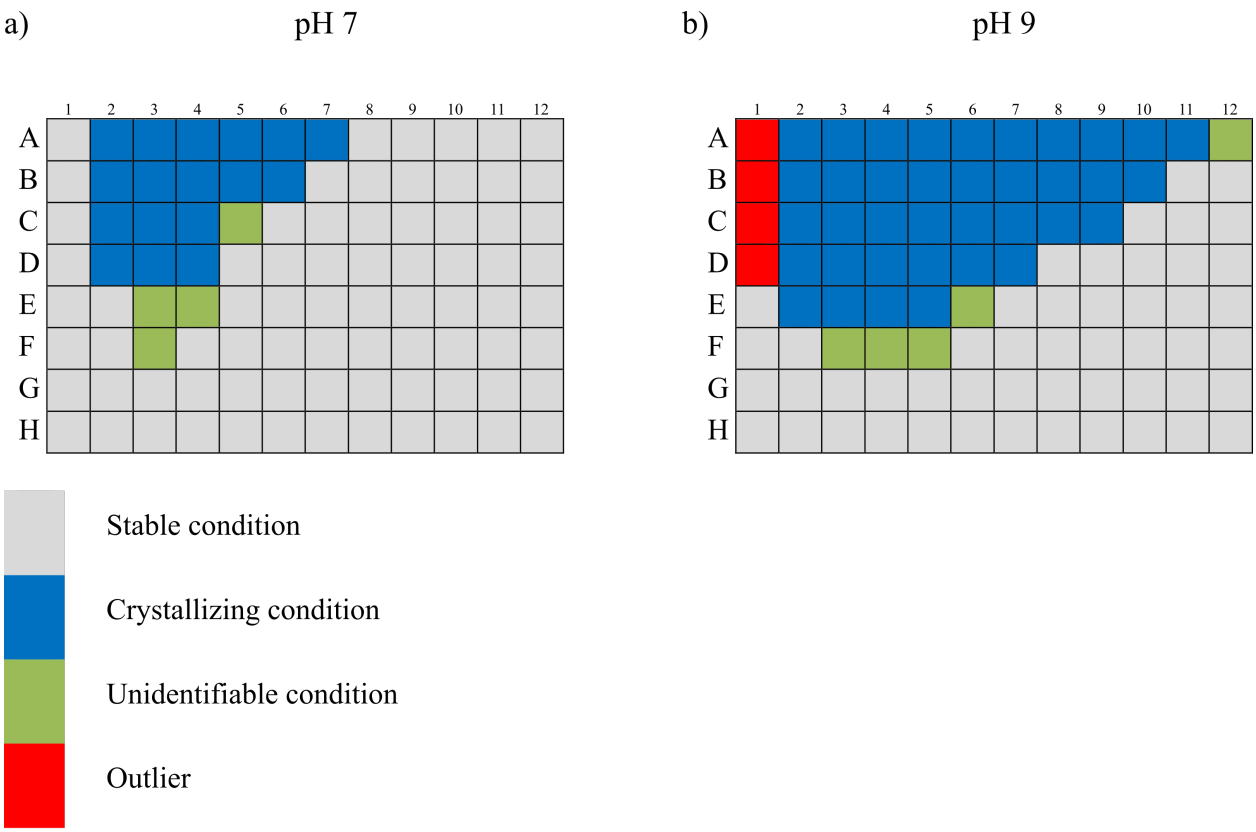
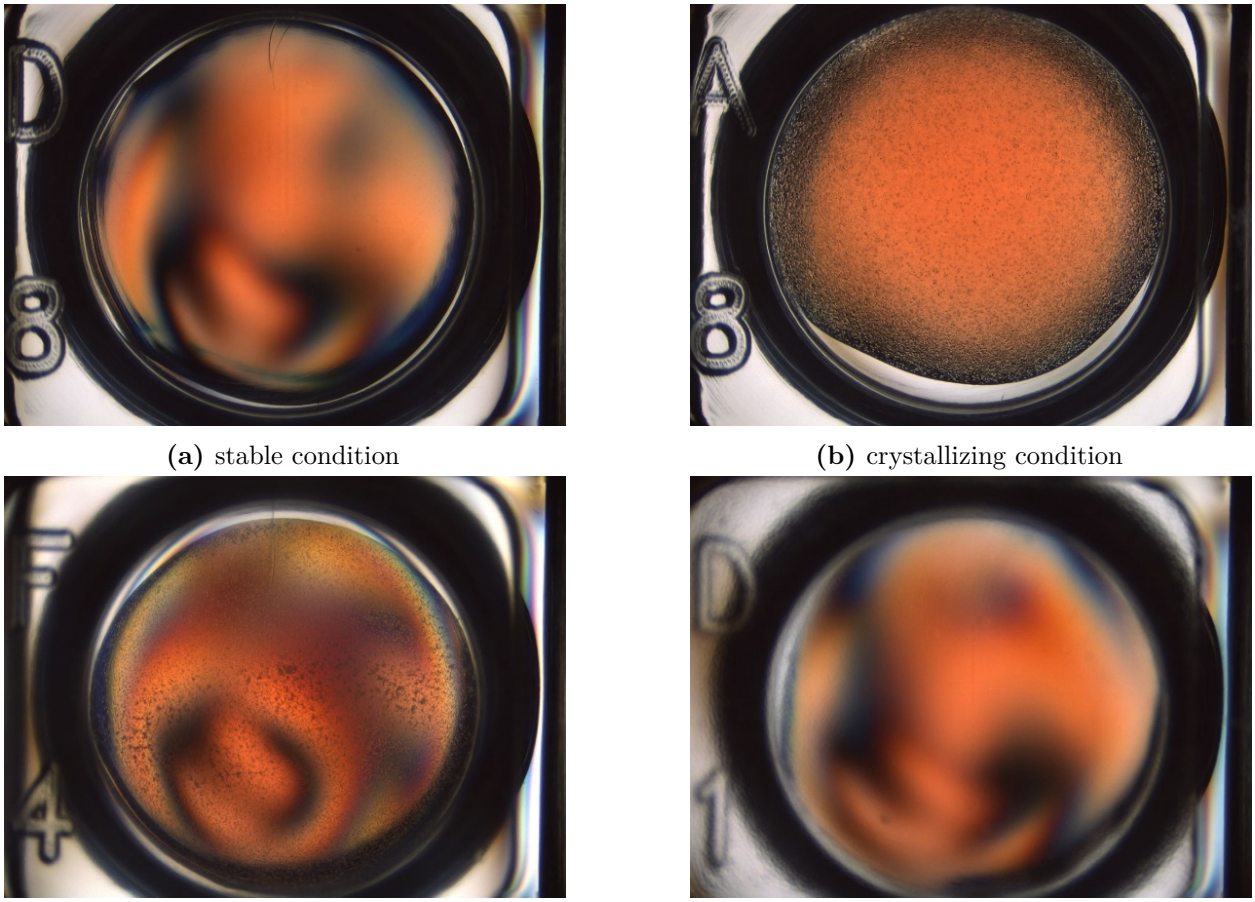


Figure A3.3 The image analysis of the phase diagrams provided scoring data which could be used to provide information on the specificity and sensitivity of the developed method to detect crystallizing conditions in a high throughput screening. The colors indicate the results of the visual analysis.



(a) stable condition

(b) crystallizing condition

(c) unidentifiable condition

(d) outlier

Figure A3.4 Representative images of the visual scoring analysis of the phase diagram at pH 9 are depicted in a) for stable, b) for crystallizing, c) for unidentifiable conditions due to blur and d) for outliers.

Table A3.1 This table displays the manually counted scoring data used for the *sensitivity* and *specificity* calculation. The developed method was used to score yields above 5.0 % as successful crystallization conditions, the image scoring serves as the validation method. Outliers were defined when the validation method only offered images out of focus or unclear images due to condensation at the covering foil. In conditions assigned as unidentifiable, it was not possible to distinguish visually between micro-crystals and precipitate. Outliers and unidentifiable conditions were not included in the calculation of the *sensitivity* and *specificity*.

	pH 7	pH 9
True-Positives / -	14	37
False-Positives / -	1	5
True-Negatives / -	75	45
False-Negatives / -	2	0
Unidentifiable / -	4	5
Outliers / -	0	4

A4

Appendix Chapter 4

Calibration-free PAT: Locating selective crystallization or precipitation sweet spot in screenings with multi-way PARAFAC models

Christina Henriette Wegner[1], and Jürgen Hubbuch[1]

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

A4.1 Case 2 - Selective precipitation of mAbs in a complex solution

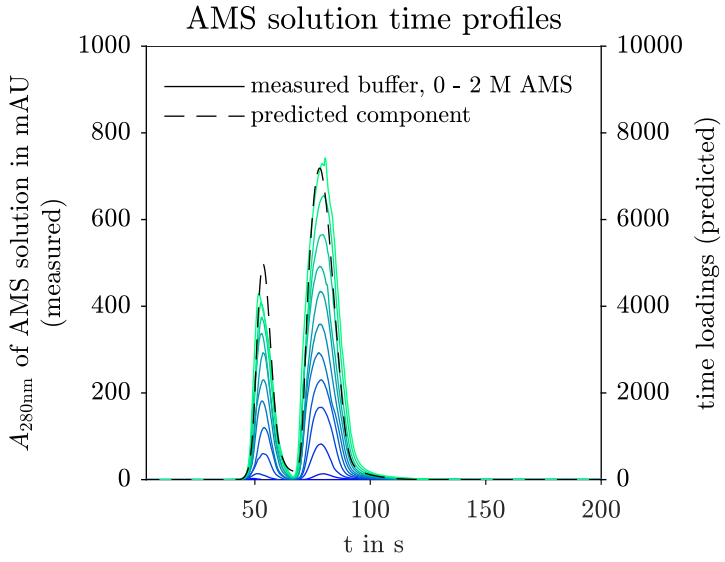


Figure A4.1 The measured UV/Vis background absorption of the solutions with 0 to 2 mol AMS is shown over time with solid lines from blue to green. The predicted time profile of the AMS component for the second case study (mAb) is illustrated with black dashed lines.

Figure A4.1 illustrates the absorption time profile of AMS solution injections which were diluted as the analyzed samples of the selective mAb precipitation screening. The position of the measured and the predicted time profiles overlay. The peak maximum rises with the AMS concentration of the analyzed sample.

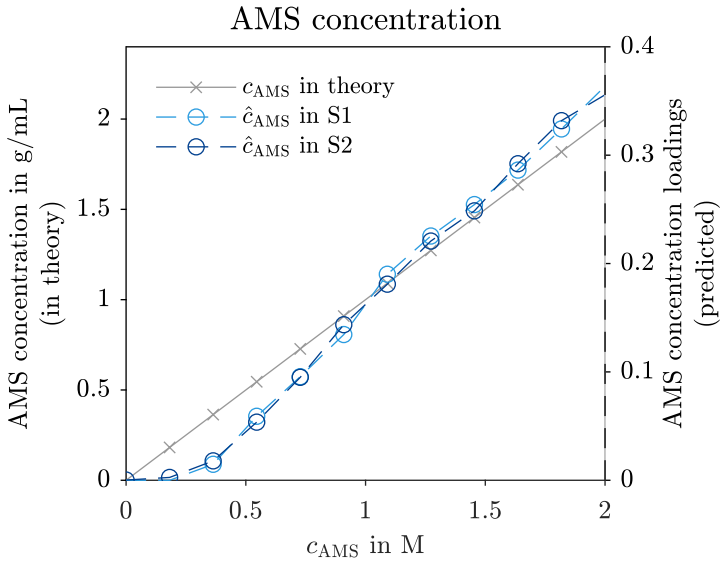


Figure A4.2 The predicted and the experimental AMS concentration in the precipitation (S1) and wash step supernatant (S2) are shown with blue dashed and gray, solid lines, respectively.

Figure A4.2 illustrates the AMS concentration during the experiment and from the PARAFAC model. The AMS concentration from the analyzed samples (S1, S2) increases in a linear manner and overlays with the experimental AMS concentration. The authors assume that the discrepancies from the ideal concentration are caused by pipetting or model errors.

A4.2 Case 3 - Selective precipitation of VLPs in a complex solution

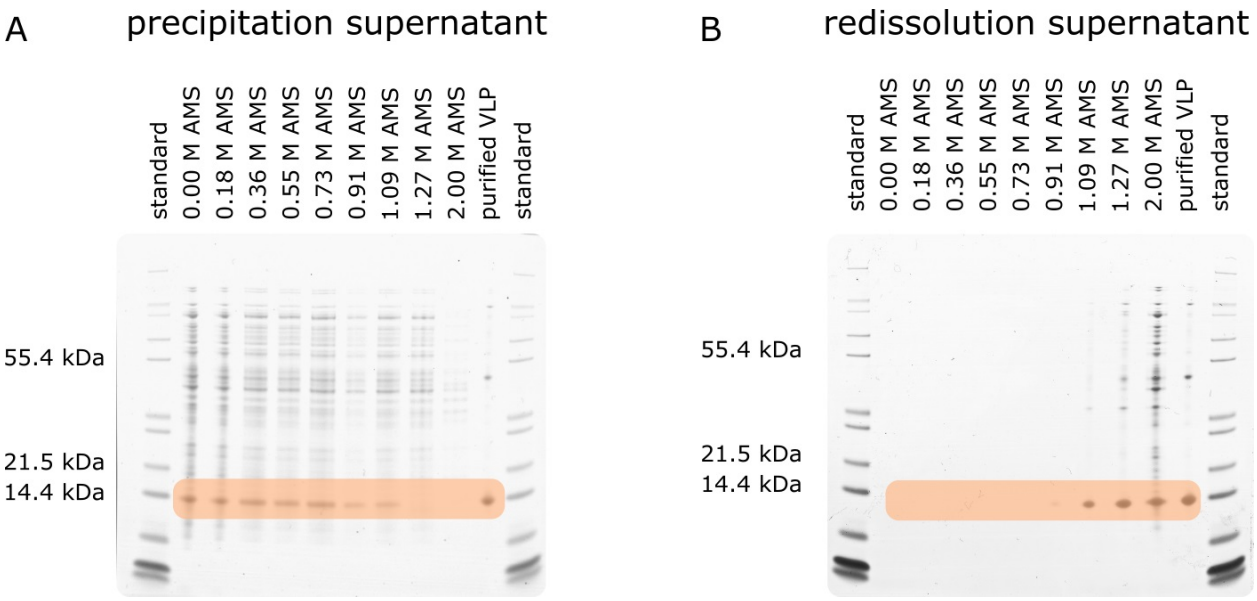


Figure A4.3 The SDS-PAGE scan of precipitation and redissolution supernatant of selected conditions are depicted in A and B, respectively. A sample of purified VLP was used as a reference to identify the VLP in the gel. The VLP is marked in orange.

Figure A4.3 shows the scanned SDS-PAGE of precipitation (A) and redissolution supernatant (B) for the selective VLP precipitation screening. The reference data of purified VLPs indicates the position of the target molecule - the VLPs. The solutions with 0 to 1.09 mol AMS still contain VLPs, but above 0.73 mol AMS, the band fades. Above 1.27 mol AMS, VLPs are not present in the precipitation supernatant. Species with larger molecular weight remain in the supernatant solution between 0 to 1.27 mol AMS which are assumed to be impurities.

The redissolution supernatant solutions indicate that VLPs are present above precipitation conditions above 1.09 mol AMS. Species with larger molecular weight are visible and were redissolved at precipitation conditions above 1.27 mol AMS. The species profile of the precipitation condition with 2 mol AMS indicates a high impurity level as many different species are present. The conditions indicating selective VLP precipitation in the precipitation supernatant agree with the conditions indicating VLP redissolution in the redissolution samples.

A5

Appendix Chapter 5

Spectroscopic insights into multi-phase protein crystallization in complex lysate using Raman spectroscopy and a particle-free bypass

Christina Henriette Wegner[1], Sebastian Mathis Eming[1], Brigitte Walla [2], Daniel Bischoff [2], Dirk Weuster-Botz [2], and Jürgen Hubbuch[1]

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

[2] Chair of Biochemical Engineering, TUM School of Engineering and Design, Technical University of Munich, Germany

A5.1 Variations of *Lk*ADH production and preparation compared to Walla et al. (2021)

Table A5.1 Differences between the protein production, preparation, and crystallization experiment: The experimental and equipment variations between the crystallization experiments of this research work and in Walla et al. [203] are listed.

process step	variations	material & methods sections 2.1 and 2.2	Walla et al. (2021)
cultivation	mode	fed-batch	batch
	vessel	1.5 L parallel fermenter	0.5 L shake flasks
cell lysis	device	Sonifier SFX550	Sonoplus HD 2070
	microtip	tapered Microtip 101-148-062 (Branson Ultrasonic Corporation)	Microtip MS 72 (BANDELIN electronic, GmbH & Co. KG)
	amplitude	70 %	90 %
	pulse	10 s	0.5 s
	pulse time	40 s	90 s
	cycle number	2 or 3	3
dialysis	membrane	SnakeSkin™ (Thermo Fisher Scientific, Inc.)	Membra-Cel(TM) Cellu. (Carl Roth GmbH + Co. KG)
	MWCO	3.5 kDa	14 kDa
	ID	22 mm	34 mm
crystallization	scale	300 mL	5 mL
	stirrer speed	80 rpm	150 rpm
	stirrer geometry	anchor style paddles	pitched-blade impellers

A5.2 IMAC analysis

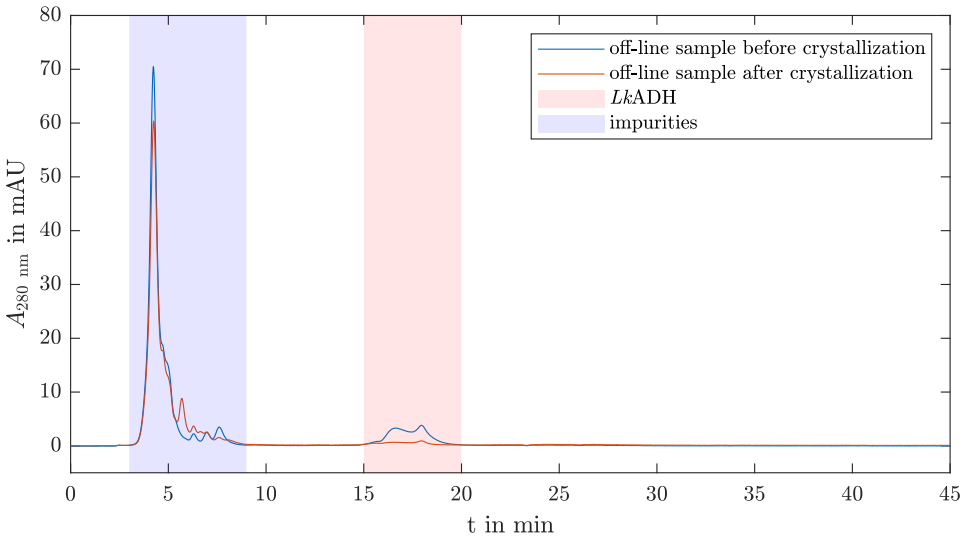


Figure A5.1 Exemplary separation with HPLC IMAC. The absorption at 280 nm of the IMAC analysis is shown over time. The blue shaded area indicates the impurities in the flow-through whereas the red shaded area is the *Lk*ADH peak which elutes with increased imidazole concentration. Exemplarily, two off-line samples at the beginning and end of the experiment are depicted in blue and orange.

A5.3 Machine-learning-based image analysis

Information about the crystal geometry need to be interpreted carefully bearing in mind the actual number of counted crystals (see Figure 2). The determined crystal widths and heights when low numbers or no crystals were visually detected can be caused by the high noise level in the images when the model falsely detects crystals. Especially the results of the automated image analysis of the experiments conducted with polyethylene glycol (PEG) concentration of 10 % may be prone to false-positive crystal detections as microcrystals can be expected, but larger crystals were not visible by human eye. The authors interpreted the provided information about the geometry in the Figures A5.2 and A5.3 as an indication that the crystal geometry does not change over time when larger crystal counts were reached.

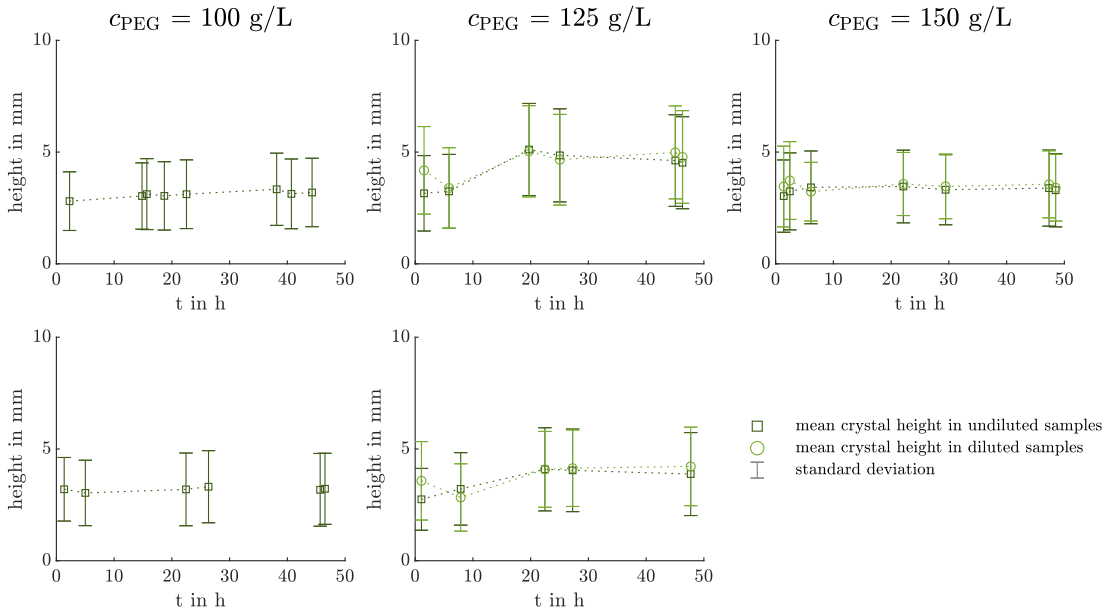


Figure A5.2 The automated ML-based image analysis [219] detected crystals, and determined the crystal height and width. These information can be used to characterize the crystal geometry throughout the experiments, i.e. crystal height and width. The mean crystal height, and the standard deviation in the undiluted and diluted off-line samples are depicted over time for five experiments in dark green squares and light green circles, respectively.

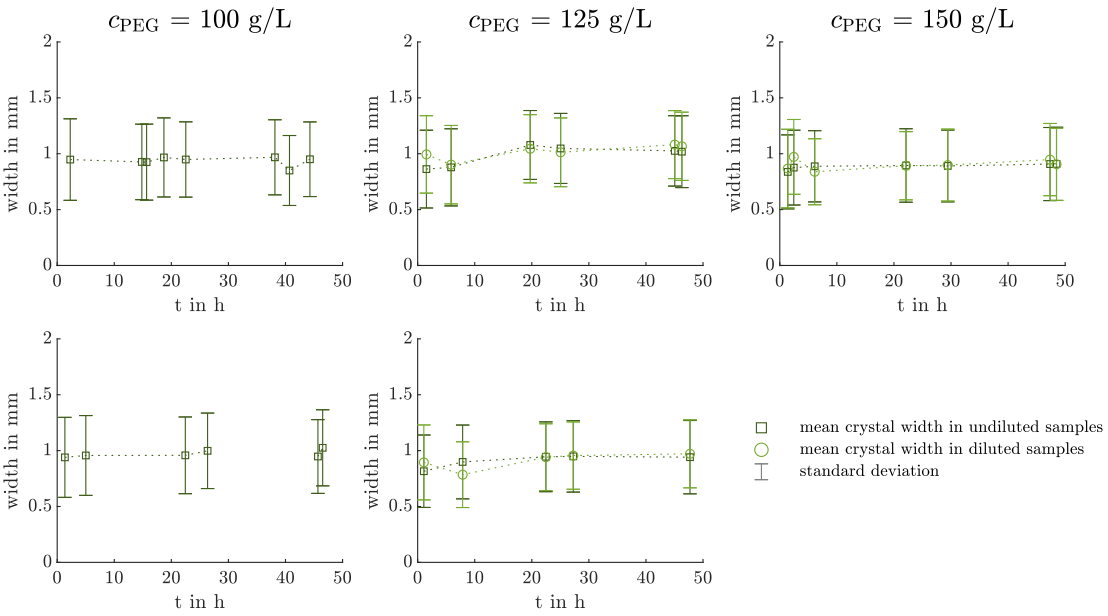


Figure A5.3 The automated ML-based image analysis [219] detected crystals, and determined the crystal height and width. These information can be used to characterize the crystal geometry throughout the experiments, i.e. crystal height and width. The mean crystal width, and the standard deviation in the undiluted and diluted off-line samples are depicted over time for five experiments in dark green squares and light green circles, respectively.

A5.4 Background Raman spectrum of protein and crystallization buffer

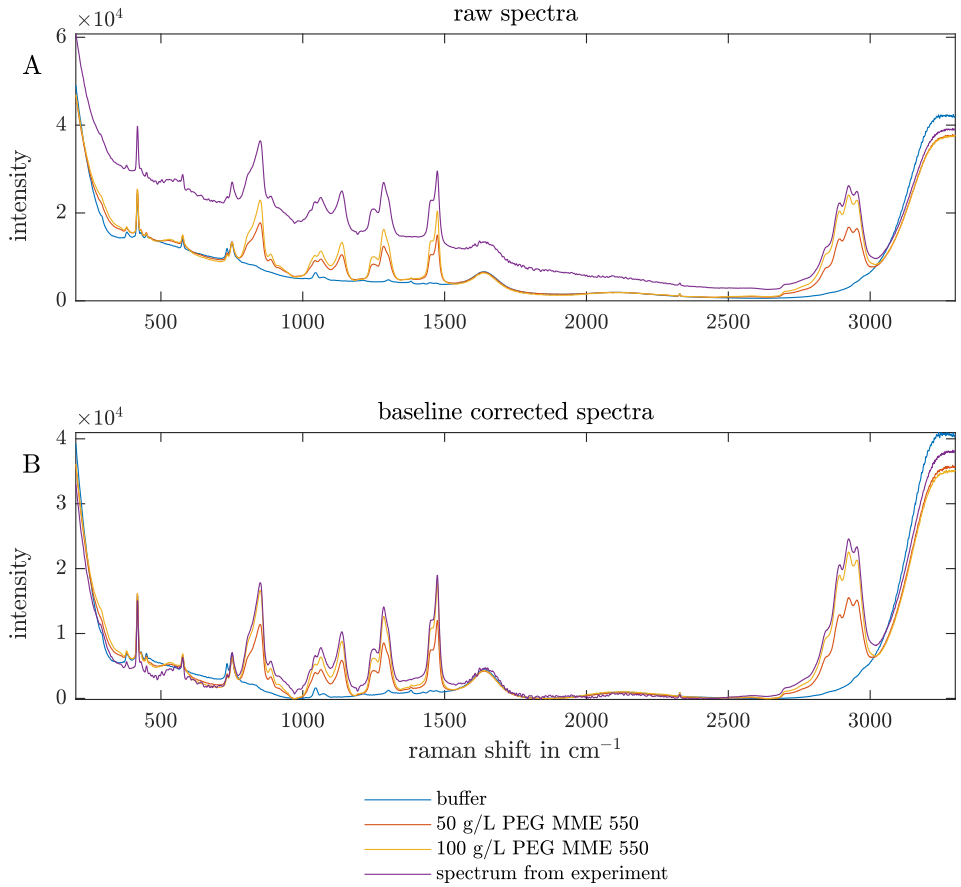


Figure A5.4 The Raman intensity of protein buffer, crystallization buffer with PEG at different concentrations, and one spectrum derived from an experiment are shown over the wavenumber shift in blue, orange, yellow and purple line color, respectively. The raw and preprocessed spectra after baseline-correction are visible in (A, B).

A5.5 Zoom into the preprocessed spectra of Exp3

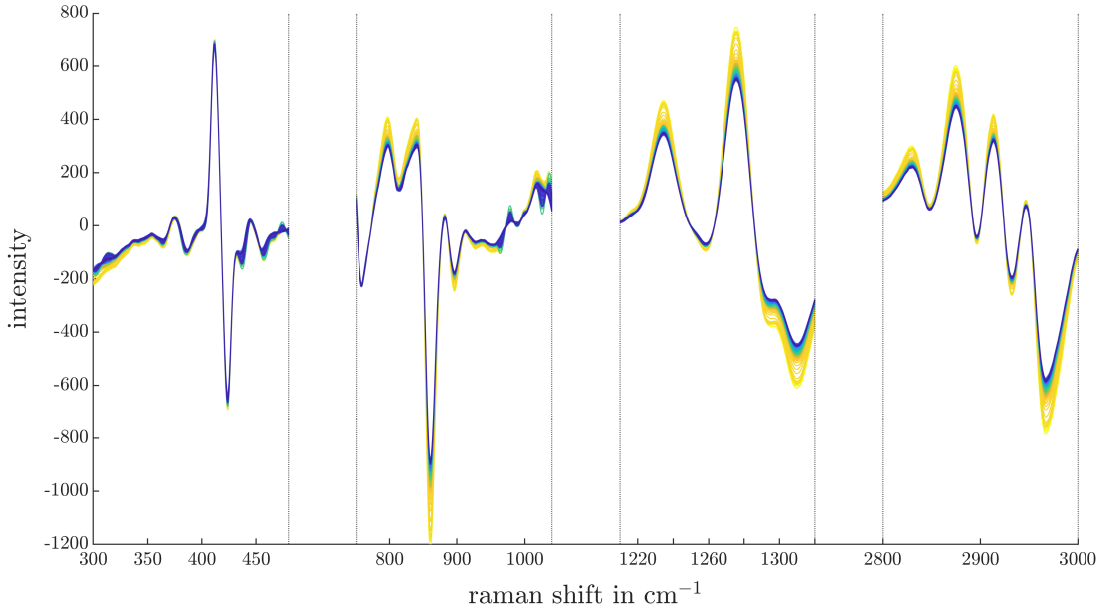


Figure A5.5 The preprocessed Raman spectra of Exp3 are illustrated over the selected wavenumber regions for the PLS model with the manual data split. The time course of the experimental spectra is visualized from yellow to blue.

A5.6 PCA loadings

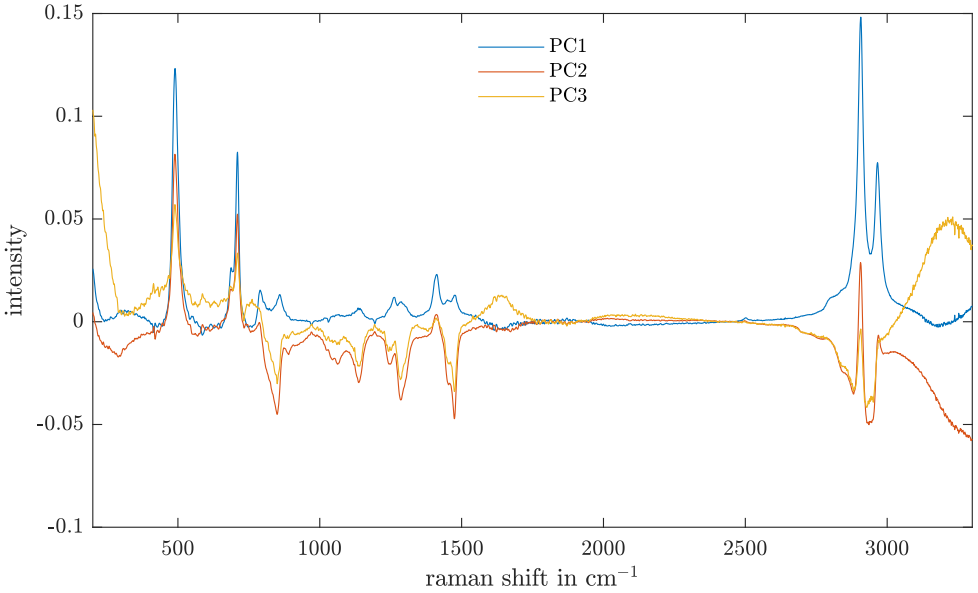


Figure A5.6 The principal components PC1, PC2, and PC3 of baseline-corrected Raman spectra of all experiments are illustrated over the recorded wavenumber range in blue, red and yellow, respectively.

A5.7 PLS model with KS algorithm applied on crystallization process spectra

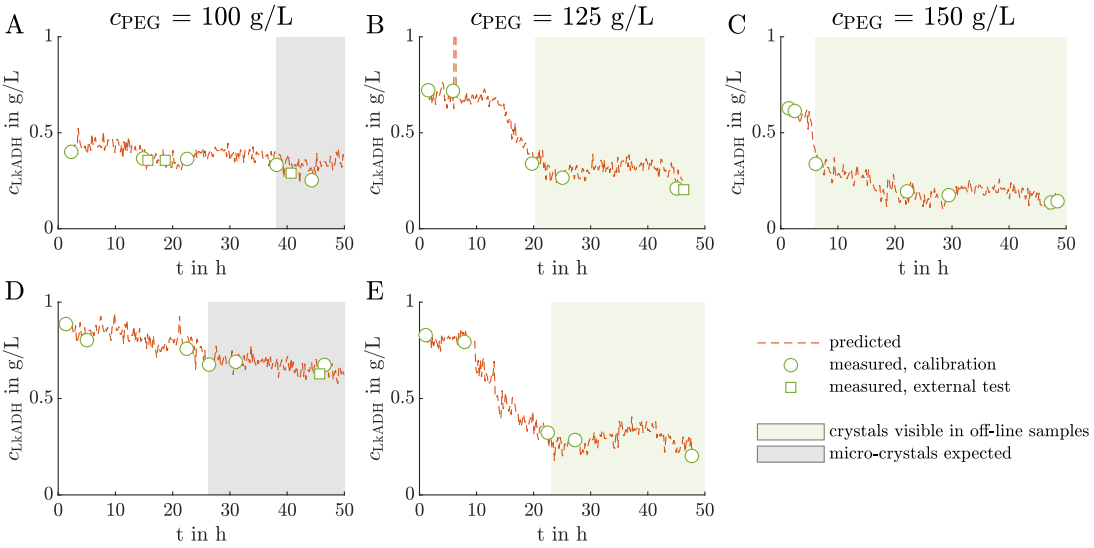


Figure A5.7 Application of PLS model on crystallization processes out of clarified lysate. The calculated PLS model with the Kennard-Stone (KS) validation data split predicts the *Lk*ADH concentration on the basis of the in-line recorded Raman spectra in orange for the five conducted experiments (A-E). Off-line *Lk*ADH calibration and validation concentrations are calculated from the IMAC analysis and are depicted with green circles and squares, respectively. The light green boxes indicate the time range when crystals were expected in the crystallization vessel as crystals were detected in the microscopic images in the off-line samples. The light gray boxes indicate time ranges in the Exp1 and Exp4 experiment when only micro-crystals were visible in the microscopic images which are difficult to distinguish from precipitate.