Computational methods for long-term protein phase behavior analysis

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

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

DISSERTATION

von
Marieke Elisabeth Klijn, M.Sc.
aus Leiderdorp, Die Niederlande

Tag der mündlichen Prüfung: 06.09.2019

Erstgutachter: Prof. Dr. Jürgen Hubbuch
Zweitgutachter: Associate Prof. Dr. Marcel Ottens, PDEng

Acknowledgements

During the course of my PhD research project there were several people who deserve a grateful acknowledgement, as without their effort, presence, and knowledge this thesis would not have been completed.

First and foremost, I would like to express my gratitude to prof. dr. Jürgen Hubbuch. The enthusiasm with which he receives and gives ideas, for my project and others, is contagious. I admire the group he has created and I am grateful to have been a part of it.

Dr. Marcel Ottens evaluated this thesis as co-promotor. I would like to thank him for the time and effort it took to assess this thesis and the subsequent discussions, as well as leading me towards the opportunity to conduct my PhD in Karlsruhe.

The BE-Basic foundation has funded the work performed for this thesis under project number FS2.003. The project meetings and scientific discussions throughout my PhD have been helpful and enlightening.

To complete my work, I collaborated with Nicolai Bluthardt, Monika Desombre, Birgit Roser, Kristina Schleining, Susanna Suhm, Philipp Vormittag, and Jana Zabel. I have enjoyed working together and I am thankful that you have shared your knowledge and experience with me. Part of my teaching responsibilities I fulfilled together with Nils Hillebrandt, whom I would like to thank for the great lessons. Steffen Grosshans, Adrian Sanden, and Susanna Suhm, I am grateful to them because of the time and effort they have invested into the German aspect of this thesis. In addition, I would like to express my gratitude to the MAB group as a whole, for it has been a pleasure to go to work every day. The group dynamic that is created at the MAB is, by all means, a reflection of each of you.

While living in Germany, parts of my life went on in The Netherlands. I am especially grateful that my family always made time and room for me when I came over, not to mention the trips they took to come over to visit Luuk and me. The same accounts for my friends, whom I am grateful to that four years abroad did not water down our connection.

In particular I would like to mention my parents, Maria and Nico, for their support regarding my education and personal life. Thank you for your advice and help throughout the years.

The person who deserves an acknowledgement the most is Luuk van Oosten. The person who, without second thoughts, moved to Germany with me. I am utterly happy with the memories that we have collected during our time in Heidelberg. Thank you for your unconditional listening, editing, love, and support. I am looking forward to our upcoming adventures, wherever they may be.

10.09.2019, Karlsruhe

Marieke Klijn

We are the cosmos made conscious
and life is the means by which the universe understands itself.

Prof. Brian Cox
in *Wonders of the Universe*

Abstract

The work presented in this thesis contributes to long-term protein phase behavior research. Long-term protein phase behavior plays a pivotal role in biotechnological product and process development. One of the largest product groups in biotechnological industries are recombinant proteins. For instance, lactases for dairy processing in the food industry and insulin to manage diabetes in the biopharmaceutical industry. The strong market position of recombinant proteins is earned due to their highly specialized functionalities, such as the ability to target specific cells and processes in the human body. However, proteins have a narrow stability window. The influence of environmental conditions, such as temperature and pH, on conformational and colloidal stability causes proteins to lose their functionality outside the required environment. Deviations from their physiological environment do not only jeopardize product efficacy, but also product safety. The extreme sensitivity to sub-optimal environmental conditions is particularly problematic at the end of production, after which protein-based products are required to remain stable during a defined shelf life of typically 18-24 months. Formulation studies, where protein phase behavior is monitored over time as a function of different environmental conditions, are needed to ensure the defined shelf life. These experiments must be conducted in real-time, meaning the experimental time equals the required shelf life of 18-24 months. Experimental time and effort can be reduced by means of knowledge-based experimental design, where prior knowledge on protein stability defines the experimental setup. Knowledge-based experimental design requires an understanding of the responsible forces and protein properties that regulate long-term protein stability as a function of environmental conditions. The obtained understanding can subsequently be employed to identify short-term properties to report on long-term protein phase behavior and to work towards the development of predictive approaches. In order to obtain such knowledge, protein phase behavior studies are coupled to biophysical screenings, where responsible forces and protein properties are characterized. Both these experimental setups are commonly performed in high-throughput to screen wide ranges of environmental conditions. Such high-throughput experimental setups generate large data sets, which inevitably leads to a bottleneck on the side of data processing, analysis, and utilization. This bottleneck is currently present in long-term protein phase behavior analysis, and has created a need for computational methods to support and advance data processing and analysis.

Studying protein phase behavior is commonly performed by means of protein phase diagrams, where protein formulations are stored for a prolonged period of time to obtain knowledge about the effects of the applied environmental conditions. After the storage period, data evaluation of such experiments typically involves manual inspection of obtained formulation images, to report on the presence of instable proteins in the form of insoluble aggregates and their morphology. Due to the workload involved with manual image evaluation, most studies focus solely on the end point image, which is the image taken on the last day of storage. However, state-of-the-art automated imaging can record data during the whole storage period. For a typical protein stability experiment, this means that end point image evaluation only accounts for roughly 1% of the generated data. The end point evaluation thereby neglects the kinetic aspect of protein phase behavior represented by the other 99% of the images. Yet this kinetic aspect is also influenced by the applied solution conditions and can contribute to a more complete understanding of protein phase behavior. The challenge with incorporating kinetic data into a protein phase diagram lies in the resulting multidimensionality of the data. This dimensionality issue was resolved in this thesis by employing the empirical phase diagram (EPD) method, an unsupervised machine learning approach developed to combine multidimensional data for visualization purposes. The EPD method was applied to compile kinetic and end point image-based data from protein phase diagram studies into one figure. The resulting so-called multidimensional protein phase diagram (MPPD) allowed for a comprehensive and more in-depth evaluation of protein phase behavior data under different environmental conditions. Another issue encountered during protein phase diagram evaluation, is the time-intensive task of manual extraction of kinetic and end point data in order to construct MPPDs. This workload can be significantly reduced with computational approaches, such as image recognition by supervised machine learning algorithms. Utilization of image recognition algorithms in the field of protein phase behavior studies is not uncommon, nevertheless their performance can still be improved. More advanced machine learning approaches, such as artificial neural networks, can be employed to improve the performance. However, increasing computational complexity decreases user-transparency and requires more expertise. Another image recognition performance enhancing approach is the generation of more diverse information to capture more distinctive properties between different protein phase behavior morphologies, without complicating the computational method itself. In this thesis, the advantages of incorporating multiple light sources (visible, UV, and cross polarized light) and kinetic data was explored to obtain higher image recognition accuracy. Compared to standard visible light end point image classification (balanced accuracy of 69.3%), an increase in balanced accuracy of 17.3 percent point (to 86.6%) was obtained upon implementation of multi-light source and kinetic data. In addition, the image recognition algorithm was coupled to the construction of MPPDs. This combination led to a computational workflow which uses raw protein phase behavior images captured over time to automatically generate an MPPD, thereby visualizing both kinetic and end point protein phase behavior.

The subsequent study presented in this thesis investigated the correlation between long-term protein phase behavior and short-term measurable forces and protein properties, where short-term empirical data was obtained on the same day as the formulation preparation. This was done as the ultimate goal in the field of formulation development is to control and predict protein phase behavior within a time frame of weeks, instead of real-time storage experiments that take months to years. As the underlying forces and properties are not known beforehand, the required analytical techniques cannot be decided on prior to experimentation. Therefore, multiple analytical techniques were employed to cover a range of potentially responsible pathways. In this work, static light scattering, dynamic light scattering, Fourier transform infrared spectroscopy, intrinsic fluorescence spectroscopy, mixed mode measurement of phase analysis light scattering, and the stalagmometric method were employed. These analytical techniques monitored protein aggregation onset temperature, apparent hydrodynamic radius, secondary structure content, melting temperature, zeta potential, and apparent surface hydrophobicity, respectively, of hen egg-white lysozyme (HEWL) in 120 different formulations. This led to a collection of short-term empirical data from the sources mentioned above, which was processed, evaluated, and subsequently correlated to long-term phase behavior. The EPD multidimensional visualization method was applied to represent short-term empirical data in a so-called empirical protein property diagram (EPPD). This allowed for a systematic and data-dependent identification of short-term parameters that reported on forces and properties responsible for observed long-term protein phase behavior.

The aforementioned computational methods were developed with the use of HEWL formulations. The applicability of the developed workflow for other protein formulations was investigated by the means of an industry case study. The objective of this case study was to screen glycerol-poor and glycerol-free food protein formulations, as EU authorities have raised concerns about the safety of high glycerol content in food products. Redesigned formulations were required to maintain a similar long-term protein stability compared to the original product, but with reduced or no glycerol content. The combination of short-term empirical protein properties and long-term protein phase behavior revealed that several redesigned formulations presumably resulted in long-term stability via a similar pathway as the original product, as they displayed a comparable apparent protein surface hydrophobicity. In addition, the workflow also identified an increase of electrostatic repulsive forces as an alternative approach to achieve long-term stability. This case study illustrated how short-term empirical studies can be used to create an in-depth product understanding and support the design of long-term stable formulations, leading the way towards knowledge-based experimental design.

Knowledge-based experimental design is also aided by prior knowledge obtained from computational methods such as protein property extraction from three-dimensional (3-D) structures. These approaches utilize molecular dynamics (MD) to simulate 3-D protein structures under varying environmental conditions, which allows for the extraction of protein properties as a function of the simulated environment. The required 3-D protein structures need to fulfill certain quality parameters prior to MD simulation, as poor structure quality causes unreliable results. The refinement of 3-D protein structures is currently a bottleneck when multiple different structures are needed, since refinement is done via a manual multi-step procedure. An illustrative example is screening virus-like particle (VLP) drug candidates. VLPs are a promising biopharmaceutical product that is able to target multiple high-profile diseases, such as cancer and Alzheimer's disease. Relatively small structural changes to the viral capsid proteins that constitute the complete VLP affect its efficacy, safety, and manufacturability, which typically leads to screening hundreds of modified structures. The required refinement of hundreds of modified structures to support preliminary drug candidate screening is not feasible with the current manual 3-D structure refinement protocol. In this thesis, an automated, data-dependent, and high-throughput compatible computational pipeline for 3-D structure preparation is presented and applied to dimeric VLP capsid protein structures. Efficiency of the computational pipeline was demonstrated by refining 31.2%-69.2% of the structural errors in only 3.6-12.5% of the total refinement time. The complete refinement was performed within 6.6-37.5 hours per structure and sufficient 3-D protein structure quality for MD simulations was obtained. In addition, a robust protein property extraction approach was developed, which takes into account the contribution of inherent structural fluctuations during an MD simulation. This was done to work towards an improved correlation between in silico and empirical protein properties.

This thesis contributes to the field of long-term protein phase behavior by combining unsupervised machine learning for multidimensional data visualization and supervised machine learning for image recognition purposes to automatically extract end point and kinetic data from long-term protein phase behavior studies. Subsequently, the unsupervised multidimensional data visualization technique was applied to investigate the correlation between short-term empirical properties measured directly after formulation preparation and long-term protein phase behavior. This workflow was applied on a case study, which led to the identification of product-specific short-term properties related to long-term stability. This thesis also contributed to the advancement of in silico approaches and their role in long-term stable formulation development. This was achieved by developing a high-throughput computational pipeline to efficiently produce high quality 3-D protein structures for MD simulations. The presented computational methods contribute to the

design of an infrastructure required to advance long-term protein phase behavior analysis and its prospective prediction.

Zusammenfassung

Die in dieser Thesis vorgestellten Arbeiten tragen zum Feld des Langzeitphasenverhaltens von Proteinen bei. Das Langzeitphasenverhalten von Proteinen spielt in der Produkt- sowie Prozessentwicklung biotechnologischer Produkte eine ausschlaggebende Rolle. Eine der größten Produktgruppen in der biotechnologischen Industrie sind rekombinante Proteine. Beispiele hierfür sind Laktase zur Verarbeitung von Milchprodukten oder Insulin als Pharmazeutikum zur Behandlung von Diabetes. Rekombinante Proteine haben eine starke Marktposition, die sich aus ihrer hochspezifischen Funktionalität ergibt, die es ihnen ermöglicht, Zellen und Prozesse im menschlichen Körper gezielt zu beeinflussen. Zu beachten ist allerdings das schmale Stabilitätsfenster der meisten Proteine. Umgebungsbedingungen wie Temperatur oder pH - Wert haben einen signifikanten Einfluss auf ihre Konformations- und Kolloidstabilität und können zum Verlust der Proteinfunktion außerhalb passender Bedingungen führen. Abweichungen von den physiologischen Bedingungen des jeweiligen Proteins können sowohl die Wirksamkeit als auch die Sicherheit des Produktes gefährden. Die extreme Sensitivität gegenüber suboptimalen Umgebungsbedingungen ist besonders gegen Ende des Produktionsprozesses problematisch, da finale Formulierungen typischerweise für 18 bis 24 Monate stabil sein müssen. Um die Haltbarkeit der finalen Formulierung über die Zeit zu gewährleisten, sind Formulierungsstudien nötig, bei welchen das Phasenverhalten von Proteinen in Abhängigkeit von verschiedenen Umgebungsbedienungen überwacht wird. Diese Experimente müssen in Echtzeit durchgeführt werden, was bedeutet, dass die Zeit für das Experiment der Haltbarkeitsdauer des Produkts von 18 bis 24 Monate entspricht. Der Aufwand und die benötigte Zeit für diese Experimente kann durch wissensbasiertes Versuchsdesign reduziert werden, bei dem Vorkenntnisse über die Proteinstabilität den Versuchsaufbau bestimmen. Derartige wissensbasierte Experimente erfordern ein Verständnis der Kräfte und Proteineigenschaften, welche die Langzeitstabilität beeinflussen. Das daraus gewonnene Wissen kann anschließend genutzt werden, um kurzfristig messbare Eigenschaften mit der Langzeitstabilität der Proteine zu verknüpfen und Methoden zur Vorhersage von Proteinstabilität zu treffen. Um dieses Verständnis zu erlangen, wird das Phasenverhalten der Proteine im Zusammenhang mit biophysikalisch messbaren Größen untersucht, wobei die relevanten Kräfte und Proteineigenschaften charakterisiert werden. Um eine weite Bandbreite an Umgebungsbedingungen testen zu können, werden sowohl das Phasenverhalten als auch die biophysikalischen Eigenschaften üblicherweise in Hochdurchsatzexperimenten untersucht. Hochdurchsatzexperimente führen unweigerlich zu sehr großen Datensätzen die verarbeitet und analysiert werden müssen. Dieser Engpass in der Datenverarbeitung hat die Entwicklung computergestützter Methoden zum Vorantreiben der Datenverarbeitung und -analyse notwendig gemacht.

Das Phasenverhalten von Proteinen wird in der Regel mit Hilfe von Phasendiagrammen untersucht. Hierzu werden verschiedene Formulierungen angesetzt und über einen längeren Zeitraum gelagert um Informationen über den Einfluss der Umgebungsbedingungen auf die jeweilige Formulierung zu erhalten. Nach der Lagerzeit beinhaltet die Datenauswertung typischerweise eine manuelle Auswertung der Fotos der Formulierungen, um das Vorhandensein von instabilen Proteinen in Form unlöslicher Aggregate und deren Morphologie zu erkennen. Aufgrund des großen Arbeitsaufwands bei der manuellen Auswertung solcher Bilder fokussieren sich die meisten Studien zu diesem Thema auf das Auswerten des Endpunktes, also des Fotos vom letzten Tag der Lagerzeit. Moderne, automatisierte Bilderfassungssysteme ermöglichen es allerdings während der gesamten Lagerzeit Fotos aufzunehmen. Für ein typisches Proteinstabilitätsexperiment beutetet dies, dass die Auswertung des Endpunktes nur ungefähr 1 % der erzeugten Daten verwendet. Des Weiteren vernachlässigt die reine Endpunktanalyse alle kinetischen Aspekte des Proteinphasenverhaltens, die von den restlichen 99 % des Datensatzes abgebildet werden. Da diese kinetischen Eigenschaften auch von den gewählten Lösungsbedingungen abhängen, könnten sie genutzt werden, um ein umfassenderes Verständnis des Proteinphasenverhaltens zu erhalten. Die Herausforderung der Integration kinetischer Daten in ein Proteinphasendiagramm liegt in der resultierenden Mehrdimensionalität der Daten. In dieser Arbeit wird das empirische Phasendiagramm (empirical phase diagram, EPD) vorgestellt, das als Ansatz für nicht-überwachtes, maschinelles Lernen entwickelt wurde, um multidimensionale Daten zu Visualisierungszwecken zu kombinieren. Das EPD wurde angewandt, um die Kinetikdaten und Endpunktanalyse aus den Proteinphasendiagrammen in einer gemeinsamen Grafik zu visualisieren. Das daraus resultierende multidimensionale Proteinphasendiagramm (multidimensional protein phase diagram, MPPD) ermöglichte eine tiefergehende Analyse des Phasenverhaltens von Proteinen bei verschiedenen Umgebungsbedingungen. Eine weitere Herausforderung beim Auswerten der Proteinphasendiagramme ist die sehr zeitintensive manuelle Extraktion der Kinetik- und Endpunktdaten zur Konstruktion der MPPDs. Dieser Arbeitsaufwand kann durch die Verwendung von computergestützten Methoden, wie der Verwendung von maschinellen Lernalgorithmen zur Bilderkennung, deutlich reduziert werden. Bilderkennungsalgorithmen zu verwenden, ist im Bereich der Proteinphasendiagramme nicht ungewöhnlich, allerdings kann ihre Leistungsfähigkeit noch gesteigert werden. Fortgeschrittene Methoden für maschinelles Lernen, wie künstliche neuronale Netze (artificial neural network, ANN), können verwendet werden, um die Ergebnisse zu verbessern. Die zunehmende Komplexität der Methode senkt jedoch die Transparenz und erfordert mehr Fachwissen vom Nutzer. Eine weitere Möglichkeit, die Leistung der Bilderkennung zu erhöhen, ist es von vornherein mehr Information aus den Phasendiagrammen zu extrahieren, um die Unterschiede prominenter herauszuarbeiten ohne die Auswertemethode an sich zu verkomplizieren. In dieser Thesis wurden zu diesem Zweck die Vorteile der Integration von mehreren Lichtquellen (sichtbares, UV- und kreuzpolarisiertes Licht) mit Kinetikdaten untersucht, um eine Erhöhung der Genauigkeit der Bilderkennung zu erreichen. Im Vergleich zu herkömmlicher Analyse von Endpunktdaten basierend auf sichtbarem Licht (durchschnittliche Genauigkeit von 69,3 %) konnte durch die Verwendung von mehreren Lichtquellen und Kinetikdaten eine Verbesserung um 17,3 Prozentpunkte auf 86,6 % erreicht werden. Zusätzlich wurde der Bilderkennungsalgorithmus an die Erstellung von MPPDs gekoppelt. Durch diese Kombination wurde ein automatisierter computerbasierter Arbeitsablauf entwickelt, der zeitaufgelöste Rohbilder von Proteinphasendiagrammen verwendet und daraus MPPDs erstellt, wobei sowohl die Kinetik- als auch Endpunktdaten des Phasenverhaltens visualisiert werden.

Die darauffolgende Studie, die in dieser Arbeit vorgestellt wird, untersucht den Zusammenhang kurzfristig messbarer empirischer Kräfte und Proteineigenschaften der Proteinformulierung mit dem Langzeitphasenverhalten, wobei die kurzfristigen empirischen Daten am selben Tag aufgenommen wurden, an dem die Formulierung erstellt wurde. Die Korrelation von kurzfristig messbaren Eigenschaften mit der Langzeitstabilität ist relevant, da das große Ziel in der Formulierungsentwicklung die Kontrolle und Vorhersage von langfristigem Proteinphasenverhalten aus kurzen Messkampagnen anstelle von aufwändigen Echtzeitversuchen ist. Da die zugrundeliegenden Kräfte und Eigenschaften vorher nicht bekannt sind, können die benötigten analytischen Methoden nicht vor den Experimenten festgelegt werden. In dieser Arbeit wurden deshalb verschiedene Analytiken verwendet, um eine große Bandbreite an möglichen Ursachen für das Phasenverhalten abzudecken. Dazu wurden statische Lichtstreuung, dynamische Lichtstreuung, Fouriertransformations-Infrarotspektroskopie, intrinsische Fluoreszensspektroskopie, Mixed-Mode-Messung zur Phasenanalyse der Lichtstreuung (M3-PALS) und die Stalagmometermethode verwendet. Mit diesen Analytiken konnten die Aggregationsstarttemperatur, effektiver hydrodynamischer Radius, Sekundärstrukturelemente, Schmelzpunkt, Zetapotenzial und die effektive Oberflächenhydrophobizität von Lysozym aus Hühnereiweiß (hen egg-white lysozyme, HEWL) in 120 verschiedenen unterschiedlichen Formulierungen gemessen werden. Diese führte zu einer Sammlung von kurzfristig messbaren, empirischen Datenpunkten aus verschiedenen Quellen, welche zusammengefasst, ausgewertet und mit dem Langzeitphasenverhalten korreliert wurden. Die multidimensionale Visualisierungsmethode EPD wurde dann verwendet, um ein sogenanntes empirisches Proteineigenschaftsdiagramm (empirical protein property diagram, EPPD) zu erstellen. Dies ermöglichte eine systematische und datenabhängige Identifizierung von kurzfristig

messbaren Parametern, die Hinweise auf Kräfte und Eigenschaften geben, die das Proteinphasenverhalten beeinflussen.

Die oben genannten computergestützten Methoden wurden mit Hilfe von HEWL Formulierungen entwickelt. Die Anwendbarkeit der entwickelten Arbeitsabläufe auf andere Proteinformulierungen wurde anhand einer Fallstudie in Zusammenarbeit mit einem Industriepartner gezeigt. Das Ziel dieser Fallstudie war es, glyzerinarme und glyzerinfreie Lebensmittelproteinformulierungen zu untersuchen, da EU-Regulierungsbehörden Bedenken bezüglich hoher Glyzeringehalte in Lebensmitteln haben. Die Anforderung an die für die Fallstudie zu entwickelnde neue Formulierung war es, trotz niedrigem oder keinem Glyzeringehalt ähnliche Langzeitstabilität zu erreichen. Die Kombination der Ergebnisse aus kurzzeitig messbaren empirischen Proteineigenschaften und langfristigem Proteinphasenverhalten ergab, dass der Langzeitstabilität einiger der neu entwickelten Formulierungen wahrscheinlich ein ähnlicher Mechanismus zu Grunde liegt wie der des Originalprodukts, da die Oberflächenhydrophobizität ähnliche Werte zeigten. Zusätzlich wurde durch die Steigerung der elektrostatisch abstoßenden Kräfte eine mögliche Alternative zur Steigerung der Langzeitstabilität identifiziert. Diese Fallstudie zeigt somit einen Weg auf, wie kurzfristig durchführbare empirische Studien verwendet werden können, um ein tiefer gehendes Produktverständnis zu gewinnen, die Entwicklung von langzeitstabilen Formulierungen zu unterstützen und den Weg zur wissensbasierten Experimentalplanung zu ebnen.

Wissensbasiertes Experimentaldesign kann auch durch Vorwissen aus anderen computergestützten Methoden wie der Extraktion von Merkmalen aus dreidimensionalen (3D) Proteinstrukturen unterstützt werden. Bei derartigen Ansätzen werden Molekulardynamiksimulationen (MD) verwendet, um dreidimensionale Proteinstrukturen unter verschiedenen Umgebungsbedingungen zu simulieren und Eigenschaften der Proteine als Funktion der simulierten Umgebungsbedingungen zu extrahieren. Hierbei ist es essenziell, dass die Proteinstrukturen, die als Grundlage für die Simulationen verwendet werden, gewissen Qualitätsanforderungen entsprechen um verlässliche Ergebnisse zu erhalten. Die Verfeinerung der 3D Strukturen wird in einem schrittweisen manuellen Verfahren durchgeführt und kann dadurch leicht zum zeitlimitierenden Schritt werden. Ein anschauliches Beispiel ist die Kandidatenvorauswahl von virusähnlichen Partikeln (virus-like particle, VLP) die als Pharmazeutika eingesetzt werden sollen. VLPs sind ein neues vielversprechendes biopharmazeutisches Produkt, das für diverse hochrelevante Krankheiten wie Krebs oder Alzheimer eingesetzt werden könnte. Eine Herausforderung bei der Produktion von VLPs ist, dass kleine strukturelle Änderungen am Kapsid, aus

welchen die VLPs bestehen, ihre Wirksamkeit, Sicherheit und Produzierbarkeit beeinflussen. Dies führt wiederum dazu, dass für eine Vorauswahl möglicher Kandidaten eine große Anzahl unterschiedlicher Strukturen vorbereitet werden muss, was mit bisherigen manuellen Methoden zur Strukturvorbereitung nicht praktikabel ist. In dieser Arbeit wird eine automatisierte, datenbasierte und hochdurchsatzkompatible Vorbereitungsmethode für die Vorbereitung der dreidimensionalen Strukturdaten vorgestellt und für dimere VLP Hüllenproteine angewendet. Die Effizienz der automatisierten Methode zeigte sich dadurch, dass 31,2 %-69,2 % der Strukturfehler in 3,6 %-12,5 % der Vorbereitungszeit beseitigt werden konnten. Die gesamte Vorbereitungszeit betrug damit 6,6-37,5 Stunden pro Struktur um Strukturdaten von geeigneter Qualität zu erhalten. Zusätzlich wurde eine robuste Methode zur Extraktion von Proteineigenschaften aus Strukturdaten entwickelt, die die Fluktuationen, die während der MD-Simulation auftreten, berücksichtigt. Dies wurde durchgeführt, um auf eine verbesserte Korrelation zwischen in silico und empirischen Proteineigenschaften hinzuarbeiten.

Diese Thesis trägt zum Feld der Untersuchung des Langzeitphasenverhaltens von Proteinen bei, indem nicht überwachte Methoden für maschinelles Lernen zur Datenvisualisierung und überwachte Lernmethoden zur Bilderkennung kombiniert wurden, um automatisch Endpunkt- und Kinetikdaten aus Langzeitstudien zum Phasenverhalten von Proteinen zu extrahieren. Anschließend wurde die unüberwachte multidimensionale Datenvisualisierungsmethode zur Untersuchung der Korrelation zwischen kurzfristig messbaren Eigenschaften, die direkt nach der Vorbereitung der Formulierung gemessen wurden und dem Langzeitphasenverhalten der Proteine in der jeweiligen Formulierung verwendet. Der Arbeitsablauf wurde auch auf eine Fallstudie angewandt bei der produktspezifische, messbare Eigenschaften identifiziert werden konnten die einen Einfluss auf die Langzeitstabilität haben. Weiterhin leistet diese Arbeit einen Beitrag zu computergestützten Methoden und ihrer Rolle für die Entwicklung bei der Verbesserung von langzeitstabilen Formulierungen. Erreicht wurde dies durch die Entwicklung einer hochdurchsatzkompatiblen Methode für die Vorbereitung qualitativ hochwertiger Proteinstrukturdaten für MD-Simulationen. Die vorgestellten computergestützten Methoden tragen dazu bei, die benötigte Infrastruktur zu etablieren, die für die Analyse des Langzeitphasenverhaltens von Proteinen und schlussendlich für dessen Vorhersage unerlässlich ist.

Content

1

Introduction

With the insertion of functional recombinant DNA into *Escherichia coli* in 1973[1], the era of modern biotechnology had started. Since this first genetically modified organism, biotechnology has expanded into a wide variety of fields. The molecular genetics field targeted the $1000 genome to rapidly unravel more genetic codes[2]. Obtained genetic codes map out the possible cellular protein content and contribute to the identification proteins of interest, for instance disease-related proteins[3]. Fields such as proteomics[4] and systems biology[5] put the possible cellular protein content into perspective, which allows for an understanding of protein production pathways, functionality, and physiological effects. In its turn, the field of genetic engineering uses this information to modify organisms for the application of interest, such as the production of biopharmaceuticals[6] or crop modification[7]. The relatively young field of synthetic biology is moving towards engineering entire cellular systems, in order to expand the range of applications[8,9]. The field of biochemical engineering utilizes modified organisms for manufacturing of biotechnological products and is responsible for the design of production processes to match the consumer demand[10].

Considering specifically the biopharmaceutical industry, technological advances seen in all aforementioned fields have led to a shift of focus from drug discovery to drug development[11,12]. One of the most widely used biopharmaceutical products are recombinant proteins[13]. Beside their use as biopharmaceuticals, recombinant proteins are also used in other industries, such as agriculture[14] and food[15]. Recombinant proteins are proteins expressed using recombinant DNA techniques for a specific environment and application, which is of interest for most biopharmaceuticals due to the high biological complexity of the human body. For example, bacterial cells are used as expression system to produce recombinant human insulin in order to circumvent the immune response of the human body to animal-derived insulin[16]. For successful recombinant protein drug development, the following three criteria should be met[11]: (1) safety and efficacy for the patient population, (2) scalable and economic manufacturing to meet production demands, and (3) a demonstrated shelf life of at least 18-24 months[17]. However, the unique combination of amino acid chains, the higher dimensional structure, and possible chemical modifications, which allow for proteins' specific functionalities, also causes their marginal stability[16]. This means that proteins are prone to chemical and physical degradation when environmental conditions deviate from their native, physiological environment in which they should perform their function[18–20]. Chemical and physical degradation usually leads to protein aggregation, which in turn may negatively affect product safety and efficacy[21]. Degradation susceptibility is of importance throughout product process development, as environmental conditions, such as temperature, pH, and additives, change throughout the entire manufacturing process. This includes the final product formulation[22]. In order to demonstrate the product's relatively long shelf life, drug development efforts are challenged to overcome physical and chemical instability over time[11].

To tackle issues concerning protein product shelf life, an understanding of the environmental effects in relation to protein characteristics on long-term protein phase behavior is needed. Such formulation studies require an experimental approach to probe a wide range of variables, which includes the protein itself, formulation additives, and solutions conditions[20]. High-throughput biophysical characterization to monitor the effects of a wide variety of experimental conditions is therefore becoming increasingly important[20]. Knowledge obtained from high-throughput biophysical characterization experiments is also useful for long-term protein phase behavior modification, and may allow for its prediction in the future[23]. However, the typically large data sets produced by such high-throughput biophysical characterization experiments cannot be interpreted without well-designed data evaluation workflows, as was also recognized when generating massive data sets during the race towards $1000 genome[24]. With the ability to generate more data to study and understand long-term phase behavior of protein products, a bottleneck occurred on the side of data analysis[20]. In the following sections, a theoretical background of protein stability is presented, as well as an overview of analytical techniques employed in this thesis to monitor protein stability. The final section addresses data analysis techniques required to correlate theoretical information to empirical data in more detail.

1.1 Protein stability

Protein stability is a complex phenomenon, as each protein is chemically and physically unique, which consequently results in unique stability behavior[18,21]. Despite proteins' distinct behavior, common stability characteristics have been identified. In order to evaluate protein stability and the effects on long-term protein phase behavior, an overview of general aggregation pathways and environmental effects is presented in the following sections.

1.1.1 Aggregation pathways

The ability of proteins to form aggregates is determined by either conformational stability or colloidal stability, or both, where the predominant factor depends on environmental conditions[18]. In Figure 1.1 an overview of the main aggregation pathways is shown. A schematic describing colloidal and conformation stability of a protein is shown in Figure 1.1 as well.

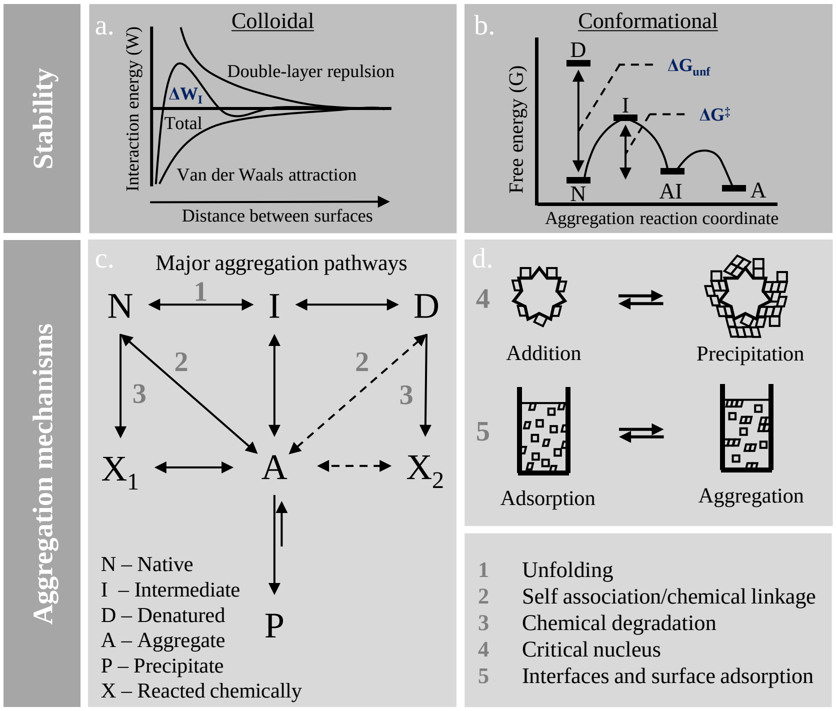


Figure 1.1: (a) Colloidal stability: interaction energy (W) as a function of surface distance between two spherical particles. The total W is the sum of the electrostatic repulsion and the van der Waals attraction, and ΔW_1 represents the maximum interaction energy barrier. (b) Conformational stability: protein aggregation reaction coordinate diagram with curved lines as transition energy barriers, and an indication of the activation energy for unfolding (ΔG_{unf}) and aggregation ($\Delta G^{‡}$). (c) Flowchart of aggregation pathway 1, 2 and 3. Different protein state abbreviations are listed. (d) Schematic of aggregation pathway 4 and 5. This figure is adapted from literature[18,25,26].

Figure 1.1a depicts colloidal stability, where the protein-protein interaction as a function of the distance between particles is described by the Derjaguin-Landau-Verwey-Overbeek (DLVO) theory[27,28]. The DLVO theory takes into account the contribution of electrostatic double-layer repulsive forces and van der Waals attractive forces on colloidal stability of proteins, which are considered to be hard spheres. The schematic in Figure 1.1a shows that the total interaction energy becomes smaller for shorter distances, as van der Waals attractive forces become more dominant. At larger distances, a positive total interaction energy indicates prevention of particle interaction by electrostatic double-layer repulsive forces. In short, colloidal stability decreases with decreasing particle distance, where particle interaction occurs when the maximum interaction energy barrier (ΔW_1) is overcome. This achieved by screening of electrostatic repulsive forces, domination of van der Waals forces, hydrophobic interactions, crowding effects, or excluded volume repulsion[12].

Conformational stability of protein is illustrated in Figure 1.1b by means of activation energies for conformational transitions of a protein. The native structure (N) can transform to a denatured state (D) or intermediate state (I) when overcoming the unfolding activation energy (ΔG_{unf}) or activation energy ($\Delta G^{\ddagger}$), respectively. For the intermediate state (I), it is assumed to be thermodynamically favorable to move towards an aggregated product (AI and A)[18]. The value for ΔG_{unf} and $\Delta G^{\ddagger}$ depends on multiple forces contributing to protein folding, such as disulfide bonds, electrostatic interactions, hydrophobic interactions, hydrogen bonds, and van der Waals interactions[19]. These forces can be influenced by environmental factors, such as pH, salt type and concentrations, and temperature.

Simply stated, a protein molecule needs to overcome ΔW_1 or $\Delta G^{\ddagger}$ in order to form aggregates. The formation of aggregates can follow one of the five main aggregation pathways[26,29]**,** which are depicted in Figure 1.1c and Figure 1.1d. The first pathway results in aggregation through partial unfolding, which means $\Delta G^{\ddagger}$ should be overcome. The protein surface can become more hydrophobic upon unfolding, which consequently decreases colloidal stability[20]. Previously, it was thought that completely unfolded protein structures (indicated by state D in Figure 1.1d) should display similar behavior as the intermediate state, but it has been noted that mainly intermediate states are prone to aggregation[18,20,26]. Nevertheless, the D state are able readily aggregate and undergo chemical degradation[26], as indicated by Figure 1.1d. The second pathway results in aggregation through self-association. This means that ΔW_1 is overcome and colloidal instability dominates. Self-association of proteins is usually an effect of environmental conditions, but self-association can also occur by chemical linkage, such as disulfide bonds. The third pathway involves chemical degradation, for instance oxidation, deamidation, or hydrolysis[23]. Chemical degradation changes the protein molecule and may cause a decrease in colloidal and conformation stability[20]. The fourth pathway typically occurs during the formation of visible or insoluble precipitates. Here, addition of monomers to a critical nucleus, an aggregate of a particular size or an impurity, is thermodynamically favored over the formation of smaller aggregates out of monomers. The fifth pathway occurs due to presence of surfaces or interfaces, such as an air-water interface. Interaction with a surface or interface decreases conformation stability in order to increase the protein-surface contact area. Conformational changes can result in aggregate formation while the protein is still in contact with the surface/interface or when the altered structure is released back into solution. Within these five common pathways, a distinction between reversible and irreversible aggregation can be made. Reversible aggregation, where aggregates can dissociate into the native form, typically occurs when no or minor structural changes lay at the foundation of aggregation and when aggregates are still small[26].

1.1.2 Factors influencing protein aggregation

The previously presented protein aggregation pathways can be induced or prevented by various environmental factors, such as temperature, pH, salt, protein concentration, and additives[18]. Combining information on the aggregation pathway with the point of engagement of these environmental factors contributes to understanding long-term protein phase behavior, as it allows for the identification of the underlying causes. Therefore, the effects of common environmental variables are presented in the following sections.

1.1.2.1 Temperature

Increasing the temperature influences conformational and colloidal stability of proteins by secondary structure disruption, activation energy (ΔG_{unf} and $\Delta G^{\ddagger}$) reduction, increased diffusion which leads to more energetic collisions between protein molecules (thereby overcoming ΔW_1), and increased chemical degradation rates[18,21,26,30]. Not only high temperatures, but also low temperatures affect colloidal and conformational stability. For example, at temperatures below the freezing point of water, proteins can adsorb to the solid-liquid interface of ice crystals[31]. In addition, ice formation changes the liquid environment in such a way that unfavorable high protein, salt, or additive concentrations may occur[32]. Conformational stability can decrease as a result of cold denaturation. This is caused by preferential hydration of the protein's nonpolar groups at low temperatures, which makes the unfolded state thermodynamically favorable[33].

1.1.2.2 Solution pH

The solution pH influences the type (positive or negative), the distribution, and net protein charge by determining the protonation state of amino acid residues[34]. The resulting charge characteristics affect both conformational and colloidal stability[18]. For increasing net charge within the protein, conformation stability decreases. This is due to a greater charge density of the native structure in comparison to the charge density of the denatured structure. This means a state of lower electrostatic free energy is found for the unfolded structure, and therefore thermodynamically favorable[35]. Contrarily, conformational stability can be enhanced by specific interactions of charged ion pairs present on the protein surface[35]. Colloidal stability is equally dependent on the net protein charge type and its distribution, and therefore influenced by solution pH as well[18,20]. For a relatively smaller net charge, the electrostatic double-layer repulsion is minimized and ΔW_1 is reduced. From this point of view, the isoelectric point (pI) should theoretically result in the lowest colloidal stability as the net protein charge equals zero at this solution pH. However, different protonation states of amino acids along the protein surface can still result in a charge distribution over the protein surface at a pH equal to the protein pI. Such a protein surface charge distribution has been shown to maintain colloidal stability[36,37].

1.1.2.3 Salt

Salt ion concentration can affect both conformational and colloidal stability. Colloidal stability can be decreased by lowering ΔW_1 as a result of screening protein surface charges, and thereby diminishing electrostatic repulsive forces[28]. At low salt concentrations shielding is the predominant effect, but at higher salt concentrations preferential binding can decrease the thermodynamic stability of the native conformation[18]. In addition to concentration, the salt type plays a role as well. Depending on the salt ion type, conformational instability may result as an effect of preferential binding to the nonnative protein state[18]. On the other hand, conformation stability may be enhanced when salt ions are preferentially excluded by the protein[38].

Not only the various influences of salt ions on protein stability, but also its dependence on protein charge, makes the influence of salt a highly complex phenomenon. For example, solution pH influences the effect of anions on colloidal stability, which is often ranked according to the Hofmeister series[39]. Initially, the Hofmeister order was considered an effect of disruption or ordering the hydrogen bond structure of water, thereby preventing or promoting protein-protein interaction, respectively. However, it has been demonstrated that the hydrogen bond structure of water is not influenced outside of the first solvation shell of the salt ion[40]. Other research demonstrated the specific influence of pH for different anions, where the direct Hofmeister order is followed when the solution pH is above the protein pI and anions serve as counterions. A reverse order of the Hofmeister series was found for solution pH values below the protein pI, when anions serve as co-ions[41]. In addition, salt concentration effects on the Hofmeister order have been reported[42]. Currently, there is no universal explanation of the direct or reverse Hofmeister series under varying conditions[43].

1.1.2.4 Protein concentration

When higher protein concentrations are used, the distance between protein molecules decreases and their collision frequency increases[26,44]. Both these effects decrease colloidal stability. In addition, the self-association aggregation pathway is promoted due to excluded volume effects, where the systems free energy is reduced by minimizing the total excluded volume[44].

1.1.2.5 Additives

Additives can be used to prolong shelf life by influencing protein stability, where some additives stabilize the native structure, some destabilize the native structure, and others are used to suppress protein-protein interactions[45]. In this work, additives are defined as all molecules that are added to formulations in addition to protein, salt, and buffer components. A comprehensive overview of additives used for protein-based products can be found elsewhere, where key mechanisms such as electrostatic interactions, preferential hydration, dispersive forces, and hydrogen bonding are discussed as well[46].

Widely used solution additives include sugars, polyols, amino acids, surfactants, and preservatives. Additives such as sugars and polyols are preferentially excluded from the protein surface[46–48]. This promotes conformation stability, as the total effect is to increase the energy gap between the native and denatured state, which translates to an increase in ΔG_{unf} and $\Delta G^{\ddagger}$ [23]. However, this may decrease colloidal stability as the free energy is also reduced by self-association[45]. Amino acids, such as histidine and methionine, can increase protein stability via preferential binding, their buffering capacity, or chemical degradation prevention, such as oxidation[46]. Addition of non-ionic surfactants, such as polysorbate 80[49], prevents aggregation as a result of surface adsorption. Protein unfolding is prevented by outcompeting protein molecules for hydrophobic surfaces, such as air-water interfaces or hydrophobic surfaces during processing. Non-ionic surfactants can also directly interact with hydrophobic regions of the protein, thereby preventing hydrophobic protein-protein interactions[46].

Besides chemical and physical protein stability, microbial stability is also of importance for successful product development. The focus of this thesis is on protein stability, but additives utilized to prolong product shelf life by ensuring microbial stability, such as benzyl alcohol, have been reported to affect protein aggregation[18]. This indicates the importance of evaluating all additives that are part of the final product formulation, and not only additives that are used to ensure long-term chemical and physical protein stability.

1.2 Analytical characterization

Information on the nature and magnitude of inter- and intramolecular forces that determine long-term protein phase behavior contributes to its overall understanding. Analytical techniques applied in this work to obtain such information, are presented in the following four sections. In the first section, long-term protein phase behavior experiments are discussed. The second and third sections present the applied analytical techniques to monitor colloidal and conformational stability, respectively. The fourth section presents the analytical techniques applied to determine protein surface properties.

1.2.1 Protein phase behavior

Protein phase behavior as a function of time is studied in formulations that contain water, protein, and additives in combination with buffer components to obtain a protein phase diagram[50,51]. Four main techniques can used to generate such diagrams, namely vapor diffusion, free interface diffusion (FID), batch, and dialysis[52]. A schematic protein phase diagram is shown in Figure 1.2, as well as the pathways of these four main techniques through the protein phase diagram.

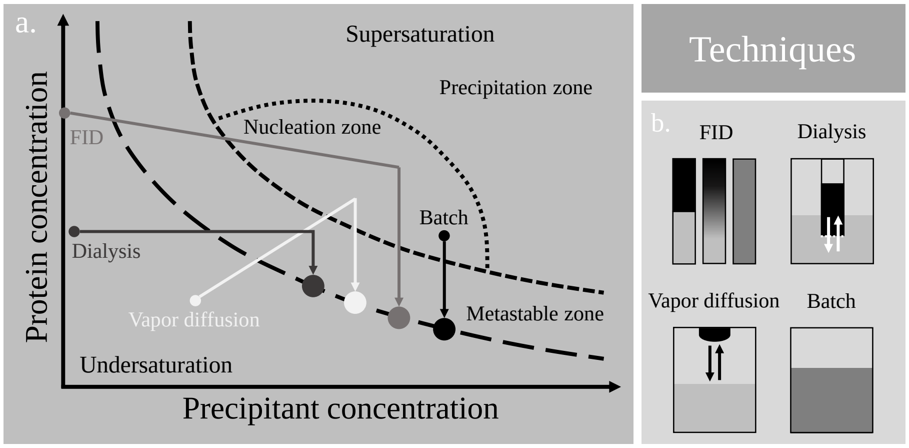


Figure 1.2: (a) Schematic protein phase diagram indicating the pathway of four main crystallization techniques: free interface diffusion (FID), dialysis, vapor diffusion, and batch. Undersaturation and supersaturation are separated by the solubility line (broad dashed line). In the supersaturation zone, the metastable (above the broad dashed line), nucleation (above the dashed line), and precipitation zone (above the dotted line) are indicated. (b) Pictograms of the four crystallization techniques. This figure is adapted from literature[52].

As depicted in Figure 1.2a, a protein phase diagram contains two main zones, namely the undersaturated zone and the supersaturated zone. The undersaturated zone contains conditions where proteins remain in solution, while the supersaturated zone contains solutions where protein aggregation occurs. The supersaturation zone is thermodynamically unstable and will move towards an equilibrium[53]. This equilibrium is found at the protein solubility concentration where aggregation and dissociation rates are equal[51]. This equilibrium is represented by the solubility line (broad dashed line in Figure 1.2a). Subzones such as the metastable, nucleation, and precipitation zone have been incorporated in Figure 1.2a to represent different aggregation kinetic stages[52]. In the metastable zone, supersaturation is too low for nucleation to take place within a reasonable amount of time. In the nucleation zone, crystal nuclei are spontaneously formed. In the precipitation zone, the level of supersaturation is too high for structured aggregation and precipitation occurs. It should be noted that the qualitative boundaries of these zones are kinetic phenomena and not well-defined[50,52].

Different techniques can be employed to obtain information about protein phase behavior, of which four are shown in Figure 1.2b. The difference between the techniques is the method to reach supersaturation to induce protein aggregation. FID uses diffusion to mix a protein solution with a precipitant solution, thereby allowing the solution to move into supersaturation. Dialysis uses diffusion to increase the precipitant concentration in the protein solution to reach supersaturation as well, but by means of a semi-permeable membrane. Vapor diffusion is based on the diffusion of volatile species (usually water) in a closed system. Dehydration of a sitting or hanging formulation drop occurs as vapor equilibrium occurs between the undersaturated drop and the reservoir containing a higher precipitant concentration. Dehydration of the formulation drop results in a protein and precipitant concentration increase towards supersaturation. For batch experiments, solution conditions are not altered during the experiment and supersaturation is reached directly after formulation preparation. Batch experiments are employed in this work. Despite the different approaches to reach supersaturation, a common disadvantage of the presented techniques is the lack of fundamental information that can be extracted from the observed protein phase behavior. Protein phase diagrams may provide some insight in the responsible aggregation pathways, as some assumptions about protein conformation or colloidal instability have been correlated to observed insoluble aggregate morphologies[54]. However, fundamental knowledge, such as the forces responsible for conformational or colloidal instability, cannot be directly extracted from protein phase diagrams. Another common disadvantage of the presented techniques, the element of time, may be eliminated with a more complete understanding of the fundamental forces. Currently, accelerated studies are performed but remain solely applicable as support for long-term stability studies[17]. This is due to the inherent assumption during accelerated studies that the external stress applied to induce protein aggregation in a shorter amount of time (such as temperature increase or pH stress) results in a comparable mechanism of physical or chemical degradation as the protein would experience over prolonged periods of time. However, this is usually not the case[26,30]. Understanding fundamental forces which lie at the basis of observed protein phase behavior may allow for the identification of short-term parameters, obtained under comparable conditions, which correlate well to long-term protein phase behavior.

1.2.2 Colloidal stability

This work employs static light scattering (SLS) and dynamic light scattering (DLS) as analytical techniques to monitor colloidal stability. A variety of other analytical techniques to monitor colloidal stability are elaborately discussed elsewhere[29,55,56].

1.2.2.1 Static light scattering

SLS uses the relationship between light scattering intensity and particle mass and concentration[57], where the scattering intensity is proportional to the sixth power of the particle diameter[58]. This technique is often used to determine the colloidal stability of proteins in defined formulations under the influence of thermal stress. A schematic representation of SLS is shown in Figure 1.3.

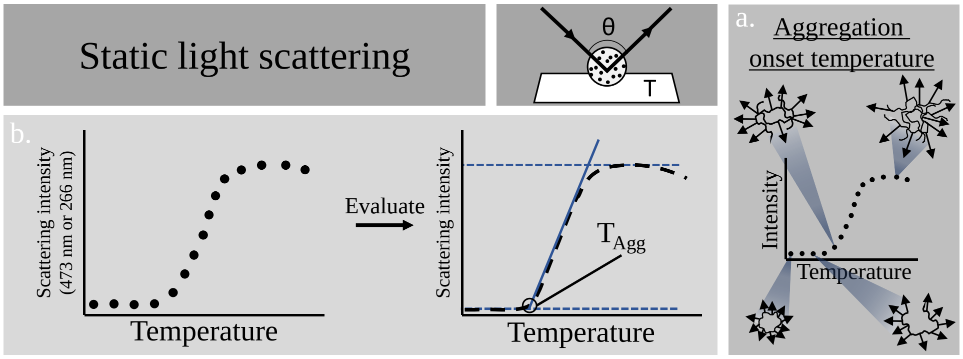


Figure 1.3: Overview of static light scattering (SLS) measurement results to determine the aggregation temperature (T_{Agg}). (a) Depiction increasing scattering intensity as a result of aggregation due to increasing temperature. (b) Schematic of a typical SLS measurement result. Scattering intensity at 473 nm or 266 nm is obtained for different temperatures, where the start of the intensity gradient is defined as T_{Agg}. Pictogram is adapted from [59].

As mentioned previously, increasing temperature causes protein structure unfolding and increases aggregation propensity. As scattering intensity is related to the size of particles, the scattering intensity increases for increasing temperature due to the increasing presence of the aggregated proteins, as depicted in Figure 1.3a. Temperature ramps are used to extract the onset aggregation temperature (T_{Agg}), which is defined as the starting point of scattering intensity increase. This is depicted in Figure 1.3b. T_{Agg} is used as a measure of colloidal stability, where a higher T_{Agg} value reflect a higher colloidal stability.

1.2.2.2 Dynamic light scattering

DLS monitors scattered light fluctuations as a result of the Brownian motion of protein molecules[57]. Brownian motion causes scattered light fluctuations over time due to changing distances between particles. Monitoring such fluctuations allows DLS to obtain information about the time scale of the movements, which in turn correlates to particle size. A schematic overview of DLS is shown in Figure 1.4.

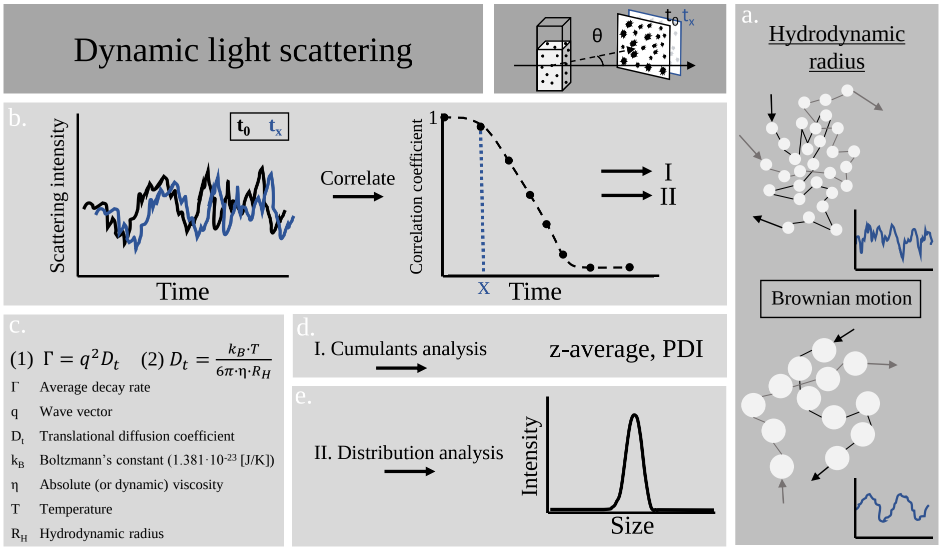

Dynamic light scattering
θ
t_0 t_x
a.
Hydrodynamic radius
Brownian motion
b.
Scattering intensity
Time
t_0 t_x
Correlate
Correlation coefficient
1
X
Time
I
II
c.
(1) $\Gamma = q^2 D_t$ (2) $D_t = \frac{k_B \cdot T}{6\pi \cdot \eta \cdot R_H}$
Γ Average decay rate
q Wave vector
D_t Translational diffusion coefficient
k_B Boltzmann's constant ($1.381 \cdot 10^{-23}$ [J/K])
η Absolute (or dynamic) viscosity
T Temperature
R_H Hydrodynamic radius
d.
I. Cumulants analysis
z-average, PDI
e.
II. Distribution analysis
Intensity
Size

Figure 1.4: Overview of a dynamic light scattering measurement to determine the (apparent) hydrodynamic radius (R_H). (a) Effect of Brownian motion on the scattering intensity over time for two particle sizes. (b) Scattering intensity over time for multiple measurements are correlated into a correlogram. The correlogram can be used in combination with (c) Equation 1 and 2 for (d) cumulant analysis to obtain an average R_H (z-average) and polydiversity index (PDI), or for (e) distribution analysis to obtain a R_H distribution. Graphs were adapted from literature[60]. Equations were obtained from literature[57].

As depicted in Figure 1.4a, smaller particles move relatively fast, which leads to a faster change in scattered light. Figure 1.4b shows how the scattering changes over time, where the decrease of the correlation coefficient relates to particle speed. With the use of the decay rate (Equation 1, Figure 1.4c) and Einstein-Stokes equation (Equation 2, Figure 1.4c), the diffusion coefficient and hydrodynamic radius (R_H) can be calculated, respectively. Two types of analysis are applicable to the measured data, the cumulants analysis and distribution analysis. As depicted in Figure 1.4d, the former is used to determine the sample's average particle size (z-average) and the polydiveristy index (PDI), which represents the sample's diversity of particle sizes. The latter analysis results in an intensity-based size distribution, as depicted in Figure 1.4e. This allows for the identification of multiple particle sizes present in a sample instead of solely an average particle size.

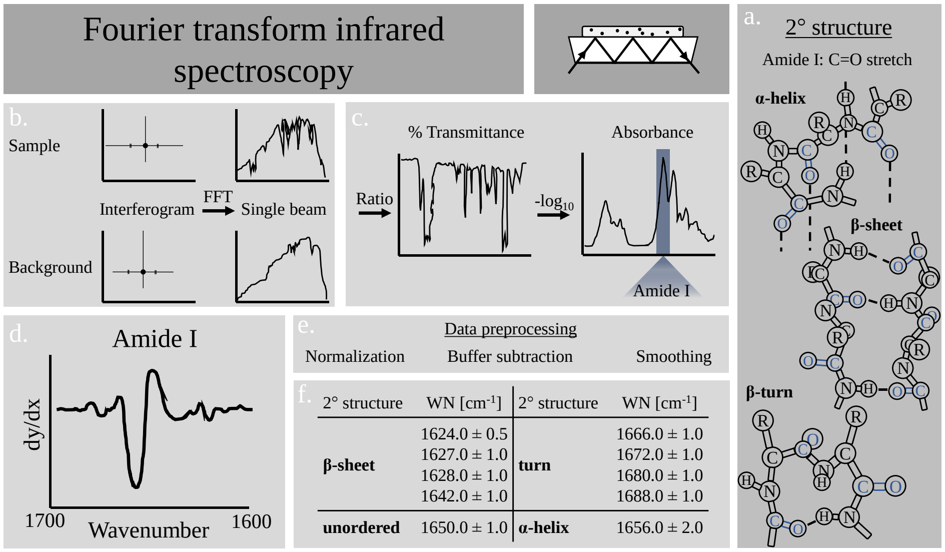
Equation 2 in Figure 1.4c reveals that DLS measurements are influenced by solution viscosity, temperature, protein size, and protein shape. What is not covered by this equation, is the fact that R_H is also dependent on protein-protein interactions[61]. Protein-protein interactions result in larger apparent particle sizes, as movement is reduced by attractive forces. Vice versa, repulsion between particles may cause an underestimation of particle size. Therefore, it must be considered that DLS often provides solely an apparent R_H when measuring at non-dilute conditions, and not a true R_H, as found for dilute systems[55]. For this work, DLS is applied to determine the presence of larger protein species, to represent aggregates, and detect minor changes in hydrodynamic radius as an effect of protein-protein interactions.

1.2.3 Conformational stability

An overview of available techniques to monitor conformational stability can be found elsewhere[19,23,56]. Fourier transform infrared (FTIR) spectroscopy and intrinsic fluorescence spectroscopy were used in this thesis and are described below.

1.2.3.1 Fourier transform infrared spectroscopy

FTIR spectroscopy uses the unique stretch vibrations of molecular bonds to identify the presence of secondary structural elements in proteins, such as an α-helix or β-sheet[19]. The measured stretch vibrations are a result of molecular bonds undergoing a change in dipole moment upon infrared radiation absorbance. The most distinct stretch vibration for proteins is found for the carbonyl group, referred to as the C-O stretch[62,63]. Secondary motifs present in protein structures have been empirically correlated to particular wavenumbers that fall within the FTIR spectral range of the C-O stretch, referred to as the Amide I region[64]. A schematic overview of an FTIR spectroscopy measurement is shown in Figure 1.5.



2° structure	WN [cm^{-1}]	2° structure	WN [cm^{-1}]
β-sheet	1624.0 ± 0.5	**turn**	1666.0 ± 1.0
	1627.0 ± 1.0		1672.0 ± 1.0
	1628.0 ± 1.0		1680.0 ± 1.0
	1642.0 ± 1.0		1688.0 ± 1.0
unordered	1650.0 ± 1.0	**α-helix**	1656.0 ± 2.0

Figure 1.5: Overview of a Fourier transform infrared (FTIR) spectroscopy measurement. (a) Depiction of the carbonyl groups (highlighted in blue) in secondary structure motifs that influence the Amide I band. (b) Sample and background interferograms are transformed with a fast Fourier transform (FFT) into single beam spectra. (c) The background and sample single beam spectra are ratioed to obtain a transmittance spectrum, which can be converted into an absorbance spectrum. (d) 2nd derivative of the absorbance spectrum in the Amide I range is used to identify present secondary structure motifs listed in (f). (e) Common data preprocessing steps used for the construction of (d). Graphs and pictogram were adapted from literature[65]. Data in (f) was obtained from literature[64].

Figure 1.5a shows a schematic overview of the α-helix, β-sheet, and β-turn, and highlights the different position of the carbonyl group. The different positions of the amino acids and corresponding hydrogen bonds influence the stretch vibrations. Figure 1.5b shows an interferogram, the raw signal obtained from an FTIR measurement. This signal is transformed using a fast Fourier transform (FFT) into a single beam spectrum. To obtain an infrared absorbance spectrum, the single beam spectrum of the background and the sample are ratioed to obtain the transmittance spectrum. In turn, the transmittance spectrum is converted into an absorbance spectrum. This procedure is depicted in Figure 1.5c. The C-O stretch vibrations are mainly captured in the Amide I wavenumber range, 1700 cm^{-1} to 1600 cm^{-1}, which is highlighted in the exemplary absorbance spectrum in Figure 1.5c. With use of the listed data preprocessing steps in Figure 1.5e and the second derivative of the absorbance for the Amide I band, the spectrum in Figure 1.5d is obtained. The second derivative absorbance spectrum is the typical data format used to identify peak area and location, which corresponds to the relative amount and type of secondary structure motifs, respectively[64]. Figure 1.5f shows a comprehensive overview of wavenumber (WN) positions of several secondary structural motifs.

1.2.3.2 Intrinsic fluorescence spectroscopy

Intrinsic fluorescence spectroscopy utilizes the influence of environmental polarity on tryptophan fluorescence, namely the red shift (shift to a higher wavelength) of its intensity maximum for increasing polarity[66,67]. This shift is used to determine tertiary structure stability as a function of thermal stress. Thermal stress is gradually applied to induce structural unfolding, which can be monitored by fluorescence as tryptophan's environment becomes more polar due to increasing solvent exposure. A depiction of an intrinsic fluorescence measurement is shown in Figure 1.6.

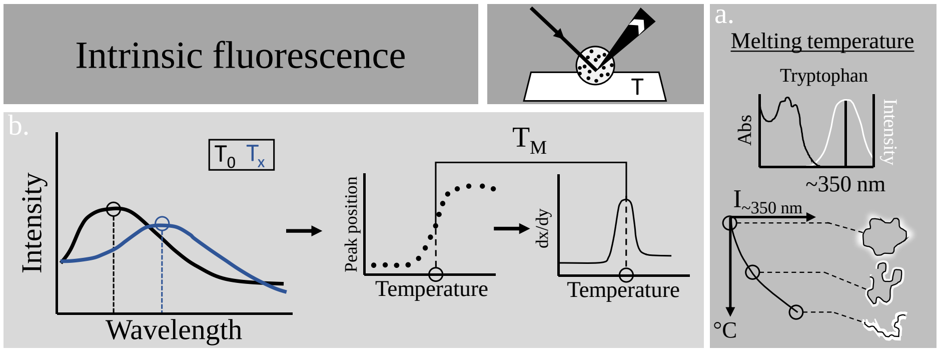


Figure 1.6: Overview of intrinsic fluorescence spectroscopy measurement results to determine the protein melting temperature (T_M). (a) Depiction of the change in fluorescence intensity wavelength upon protein unfolding. (b) Exemplary plot of fluorescence intensity obtained for two different temperatures. The wavelength peak position for each temperature is plotted and the derivative is calculated to determine the maximum gradient of the peak position plot. The maximum gradient is defined as T_M. The pictogram is adapted from literature[59].

Figure 1.6a depicts how the maximum tryptophan fluorescence intensity shifts towards a wavelength of 350 nm due to protein unfolding. Figure 1.6b shows a simplified spectrum at a starting temperature and a higher temperature, where a shift in absolute intensity and intensity maximum is depicted. The maximum intensity wavelength, referred to as the peak position, is extracted and plotted against the applied temperature gradient. This data is used to extract the temperature at which ΔG_{unf} is zero. At this point, where native and unfolded proteins are present in equal amount, is referred to as the melting temperature (T_M). T_M is defined as the maximum gradient in the peak position-temperature plot and can be determined by means of its derivative, as shown in Figure 1.6b. The T_M of a protein is a measure of its conformational stability, where a higher T_M value indicates higher conformational stability. It has been observed that aggregation propensity often correlates inversely to the relative T_M[20]. However, it should be noted that T_M cannot be used as an aggregation propensity predictor when the dominant aggregation pathway is not dependent on (partial) protein unfolding[68].

1.2.4 Protein surface properties

Protein surface properties, such as charge and hydrophobicity, play a significant role in protein aggregation. Two surface properties are monitored in this work, namely the zeta potential and apparent surface hydrophobicity. These properties are determined by mixed mode measurement of phase analysis light scattering (M3-PALS) and the stalagmometric method, respectively. Both methods are described below. A presentation of other analytical techniques to determine protein charge and hydrophobicity can be found elsewhere[69,70].

1.2.4.1 Mixed mode measurement of phase analysis light scattering

M3-PALS utilizes light scattering to track the movement of charged proteins under the influence of an electric field. Protein motion caused by means of an electric field is called electrophoretic mobility[71]. Electrophoretic mobility is induced during M3-PALS by switching poles, referred to as field reversals, which causes the protein to change its direction. The magnitude and nature of the protein surface charge can be extracted by tracking protein movement as a function of field reversal, as it determines the speed and direction of the motion, respectively. An overview of an M3-PALS is depicted in Figure 1.7.

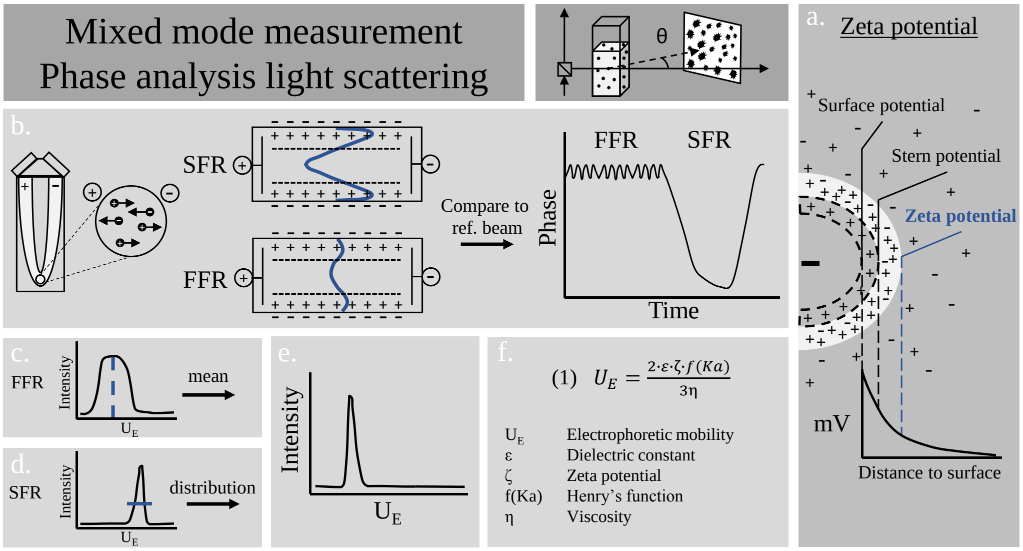


Figure 1.7: Overview of a mixed mode measurement of phase analysis light scattering (M3-PALS) to determine the zeta potential (ζ). (a) Depiction of different potentials as a function of the protein surface distance. (b) M3-PALS includes a slow field reversal (SFR) and fast field reversal (FFR), where the blue line indicates electroosmosis. Phase shift between the sample and reference beam is recorded over time. This provides information on (c) the mean electrophoretic mobility (U_E) and (d) its distribution, respectively. The combination of FFR and SFR results in (e) a U_E-intensity distribution. The ζ can be calculated with (f) the Henry equation, Equation 1. Pictograms and graphs were adapted from literature[60,72].

Figure 1.7a shows a schematic representation of the electrical double layer that exist around a charged protein. The first layer, the Stern layer, is a layer of relatively strong bound ions, which carry the opposite charge of the protein charge. The second layer is called the diffuse layer, where weakly associated ions are found. Upon movement of a protein through a solution, there is a boundary between the ions in the diffuse layer that move with the protein and ions that do not. This boundary is the so-called slipping plane. The potential at the slipping plane is the zeta potential. The nature and magnitude of the zeta potential play a role in electrostatic repulsive forces, which in turn influence colloidal interactions[55].

Figure 1.7b depicts how M3-PALS is used to determine electrophoretic mobility. Two field reversal methods are applied during a measurement, namely the fast field reversal (FFR) and the slow field reversal (SFR). The two field reversal approaches are used in order to determine an accurate and precise electrophoretic mobility, respectively. Light scattering fluctuations due to the subsequent protein movement is determined by comparing the phase of the light from a beam passing through the sample to the phase of a reference beam that did not pass the sample. This results in a phase plot as depicted in Figure 1.7b. FFR allows for the determination of an accurate mean electrophoretic mobility (Figure 1.7c), as it prevents the influence of electroosmosis[73]. Electroosmosis is the movement of liquid under the influence of an electric field. For each field reversal speed, electroosmosis is schematically depicted by the blue line in Figure 1.7b. This liquid movement influences protein movement and causes an overestimation of its electrophoretic mobility. However, the order of magnitude at which a liquid reacts to a field reversal is roughly in the order of ten milliseconds[72]. As the response of protein particles to the field reversal is faster, an accurate particle electrophoretic mobility can be determined with FFR. Nonetheless, a precise electrophoretic mobility cannot be obtained with FFR as the velocity distribution is unavailable. A precise electrophoretic mobility distribution is measured with SFR (Figure 1.7d). A both accurate and precise electrophoretic mobility distribution can only be obtained by combining both results, as depicted in Figure 1.7e. The zeta potential can be calculated from the electrophoretic mobility, using the Henry equation (Equation 1 in Figure 1.7f), wherein the measured electrophoretic mobility depends on the zeta potential, the applied electric field strength, dielectric constant of the solution, and the solution viscosity. In addition, Henry's function (*f(Ka)*) shows the influence of the protein radius (*a*) and the Debye parameter (*K*), which represents the thickness of the electrical double layer.

1.2.4.2 Stalgmometric method

The stalagmometric method depends on the relationship between the required gravitational force to detach a formulation droplet and the adhesive force of the droplet to the dispensing tip to determine the apparent protein surface hydrophobicity[74]. A schematic overview of the stalagmometric method is shown in Figure 1.8.

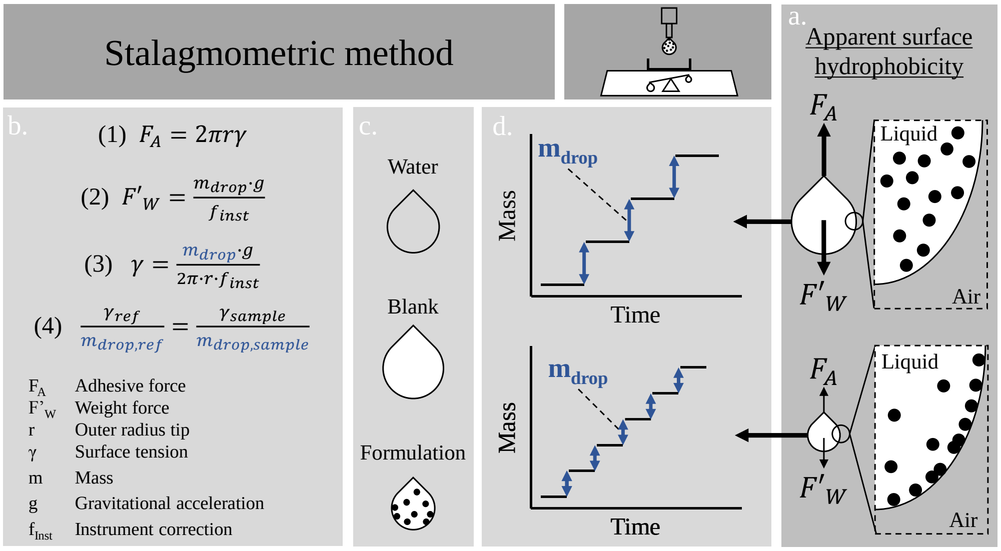

Stalagmometric method
a.
Apparent surface hydrophobicity
F_A
F'_W
Liquid
Air
b.
(1) $F_A = 2\pi r\gamma$
(2) $F'_W = \frac{m_{drop} \cdot g}{f_{inst}}$
(3) $\gamma = \frac{m_{drop} \cdot g}{2\pi \cdot r \cdot f_{inst}}$
(4) $\frac{\gamma_{ref}}{m_{drop,ref}} = \frac{\gamma_{sample}}{m_{drop,sample}}$
F_A Adhesive force
F'_W Weight force
r Outer radius tip
γ Surface tension
m Mass
g Gravitational acceleration
f_{Inst} Instrument correction
c.
Water
Blank
Formulation
d.
m_{drop}
Mass
Time

Figure 1.8: Overview of the stalagmometric method to determine the apparent surface hydrophobicity. (a) Depiction of the air-liquid interface for different droplet sizes. (b) Overview of equations used to determine the surface tension. (c) Solutions required during a measurement with the stalagometric method. (d) Schematic of typical measurement results, where the mass of dispensed droplets is plotted over time. The average mass of a single droplet (m_{drop}) is defined as the difference between the mass plateaus.

Figure 1.8a depicts two formulation droplets, where the upper droplet shows a formulation without protein interface adsorption, and the lower droplet shows a formulation with protein interface adsorption. The adhesive forces, F_A, are dependent on the radius of the dispensing tip (r) and the formulation surface tension (γ), as shown by Equation 1 in Figure 1.8b. F_A equals the weight forces (F'_W) upon droplet detachment, where F'_W is defined by the droplet mass (m_{drop}), the gravitational acceleration (g), and an instrument correction (f_{inst}), as indicated by Equation 2 in Figure 1.8b. Thus, the mass of a detached droplet is proportional to the surface tension, which described by Tate's law[75] (Equation 3, Figure 1.8b). The resulting formulation droplet mass decreases for decreasing surface tension. Figure 1.8d shows an exemplary result of a stalagmometric measurement, where droplet mass is measured over time by means of an automated liquid handling station[76]. Each plateau indicates a droplet falling on a precision scale, which measures the weight continuously. The obtained average droplet mass, defined as the average difference between the plateaus, is used to determine the surface tension. This is done by comparison to a reference solution, typically water, and Equation 4 in Figure 1.8d. Figure 1.8c shows three different droplets to be measured to gain information on the apparent protein surface hydrophobicity. Besides the reference solution (water) and the formulation itself, a blank sample is measured. The difference between a blank droplet and a formulation droplet provides information on the contribution of protein molecules to the surface tension.

1.3 Data handling

Experiments employing multiple analytical techniques to monitor various biophysical protein properties, in combination with kinetic long-term protein phase behavior, result in large and multidimensional data sets. Such multidimensional data sets to study protein stability can be found throughout literature[77–82], but interpretation of these data sets is not straightforward. The following three sections describe procedures of computer-assisted mining and interpretation of such data sets. It has been stated that this type of computational support is of great importance for the advancement of the biotechnological field[83]. The first section presents data processing steps, the second section discusses approaches for data visualization, and the third section presents approaches to utilize the generated data for mining and prediction purposes.

1.3.1 Data processing

Raw experimental data is usually not immediately suitable for evaluation. Data preprocessing is the first step to obtain data that can be utilized for information mining purposes. Data preprocessing usually involves, among others, data transformation, data normalization, standardization, smoothing, and outlier detection[84]. After preprocessing, one can decide whether further processing is needed or not. One optional additional processing step is called feature extraction, which can be performed in a supervised or unsupervised manner. The terminology supervised and unsupervised will be discussed in more detail in section 1.3.3. Feature extraction is often applied when the data format is inconvenient, such as the case for image data, or when data includes redundant information[85,86]. Besides extraction of relevant information, feature extraction has also shown to reduce model training times, enhance prediction performance, and lower computation expenses associated with generation, storing, and processing of data sets[87].

1.3.2 Data visualization

Creating a comprehensive visual representation of experimental data is of importance for obtaining a deeper understanding, as well as to present data to external parties. The applicability of particular data visualization techniques is dependent on data dimensionality. For example, histograms are applicable for univariate data sets, while 2-D scatterplots are suitable for bivariate data sets. An overview of applicable visualization techniques when handling multidimensional data sets can be found elsewhere[88]. In this work, the empirical phase diagram (EPD) visualization technique was used to represent multidimensional data sets[78,89]. This unsupervised machine learning method employs a data dimension reduction approach. This allows for the representation of multidimensional data in three dimensions, which is converted into an RBG color code for simplified interpretation of trends based on colors. Other biopharmaceutical studies that have successfully employed the EPD method are listed elsewhere[90]. An expansion of the EPD method to enhance data interpretation, involving the representation of reduced multidimensional data by means of radar charts[91], has also been applied in this work.

1.3.3 Data utilization

The combination of machine learning and experimental data can be applied for data visualization, but also for data utilization. The term data utilization means employing mined data sets for pattern recognition or training of predictive models. In general, this is achieved using two main machine learning approaches, namely supervised and unsupervised learning. A schematic workflow for each machine learning approach is shown in Figure 1.9.

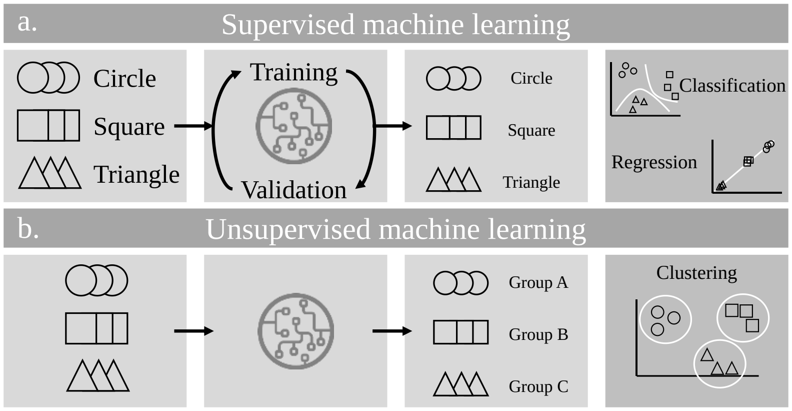


Figure 1.9: Schematic workflow of (a) supervised and (b) unsupervised machine learning algorithms, where supervised machine learning involves model training and validation. The last panel indicates what types of algorithms are used for each machine learning approach, where (a) employs classification or regression algorithms and (b) employs clustering algorithms.

Figure 1.9a illustrates supervised machines learning and Figure 1.9b depicts unsupervised machine learning. The main difference between supervised and unsupervised is the labelling of the input data for supervised learning. The goal of supervised machine learning is to predict a class or value, which is achieved by training and validating algorithms with said labelled input data. The resulting predictive models can be categorized as classification or regression models, which are applicable for discrete or continuous output data, respectively. Typical algorithms that are used for classification are support vector machines, discriminant analysis, decision trees, or neural networks. Regression models can include linear regression, ensemble methods, or neural networks. An overview of classification and regression algorithms can be found elsewhere[92–95]. To prevent overfitting, model bias, or overoptimistic model performance, it is essential to carefully select the size and class or value distributions of the training and validation sets[96,97]. In addition, corresponding model evaluation parameters need to be selected in order to describe the outcome of both internal and external validation of the model, and to evaluate the overall model performance. It is also possible to evaluate the prediction confidence, which allows for the identification of possible prediction errors[92].

For unsupervised machine learning approaches, the goal is to identify patterns in the data. This approach is often applied for data exploration to extract labels that may define different groups[92]. Such labels, often comprised of different data features, should therefore contain information that can separate objects from one another and identify similarities to recognize similar objects. There are several algorithms that can be used for this application, such as *k*-means clustering, principal component analysis, singular value decomposition, and neural networks. An overview of applicable algorithms can be found elsewhere[92,93].

Protein phase behavior is a multidimensional phenomenon and therefore the corresponding data required for investigating protein phase behavior is multidimensional as well. Even though this has been recognized in the field, there are still opportunities to implement computational methods, such as machine learning approaches, to advance protein phase behavior analysis. Opportunities lie in the comprehensive visualization of protein phase diagrams, where not only protein phase behavior after a prolonged storage time should be visualized, but where also the kinetic phenomena monitored during storage are depicted. Data mining from protein phase diagram experiments (image-based data) remains a manual, time-consuming, and subjective procedure[98]. Therefore, data mining in protein phase behavior studies would greatly benefit from supervised image recognition algorithms, especially in combination with automated feature extraction approaches. Opportunities in data utilization can also be found for characterizing protein phase behavior by means of the responsible forces and protein surface properties. As multiple analytical techniques are needed to monitor the possible aggregation pathways, this thesis presents a machine learning workflow that can correlated short-term empirical properties to long-term protein phase behavior in order to generate a better understanding of environmental effects on the observed protein phase behavior.

Another interesting subject where computational workflows could be incorporated in the field of protein phase behavior, is in silico extraction of protein properties obtained from molecular dynamics (MD) simulations. By employment of supervised machine learning approaches, protein surface properties based on three-dimensional (3-D) protein structures can be used to model experimental observations or serve as an additional, new source of knowledge to obtain an efficient experimental design. However, the computational extraction of such properties is highly dependent on the quality of the 3-D protein structures that are used during MD simulations[99]. In addition, in order to obtain reliable models, a large amount of structures is needed to determine the statistical significance of the correlation between theoretical and experimental data[100,101]. Preparation of 3-D structures for MD simulations is typically a manual procedure. This hampers the implementation of MD simulations as supportive computational method during formulation development, as it currently cannot compete with the screening numbers achievable with high-throughput experiments.

2 Thesis outline

2.1 Research proposal

Protein-based products are developed in biotechnology sectors such as agriculture, food, and biopharmaceutics. Stability for each protein-based product needs to be demonstrated, where the desired protein phase state, such as soluble or crystallized protein, should be maintained over a prolonged period of time (e.g., months or years) in order to assure its safety and functionality. High-throughput and accelerated experiments were implemented during product development to minimize material use as well as experimental time required to perform long-term protein phase behavior screenings. Consequently, optimizing these screenings led to an increase of available experimental data. Upon evaluation, interpretation, and understanding the generated data, a knowledge-based approach can be adopted to further accelerate protein-based product development. This leads to the identification of short-term empirical parameters which correlate to long-term protein phase behavior, targeted phase behavior modification, and partially paves the way for protein phase behavior prediction. However, the generated information is not used in its complete capacity due to insufficient data extraction and data analytical techniques. The full potential of knowledge-based experimental design will only be realized when the gap between data acquisition and data utilization is bridged.

The objective of this research is to develop computational methods that advance long-term protein phase behavior analysis. Protein phase behavior is a multidimensional phenomenon, where an interplay between a multitude of environmental factors, such as additives and pH, and protein-specific properties, such as charge and size, determines the protein phase outcome. To understand, modify, and ultimately predict a multidimensional phenomenon such as protein phase behavior, computational approaches are required that incorporate complexity while maintaining a straight-forward data interpretation. In addition, a data-dependent design of said approaches is required to ensure its applicability for a wide range of protein-based products. However, extensive data extraction and processing increases the experimenters' workload. This is prevented by implementation of automated approaches, which simultaneously controls objectivity and standardization, which is hard to enforce with manual approaches. Besides protein phase behavior data evaluation, the ability to characterize the observed protein phase behavior is required to obtain a more in-depth understanding. A more in-depth understanding of protein phase behavior will allow for identification of short-term protein phase behavior optimization targets and a decrease in experimental work. Furthermore, a broader understanding serves as a stepping stone towards the design of in silico descriptors suitable for protein phase behavior prediction. To generate such in silico descriptors, a computational pipeline that allows one to process the in silico equivalent of high-throughput experimental data is required in order to become a helpful tool during product development.

To bridge the gap between data acquisition and data utilization for long-term protein phase behavior experiments, the first part of this work focusses on data-dependent visualization of image-based protein phase behavior data. The aim was to prevent information loss during data evaluation by combining static and kinetic long-term protein phase behavior data. The visualization method was subsequently coupled to an automatic data extraction algorithm, where protein phase behavior was classified using data obtained with a hardware combination of multiple light sources. Identification of short-term protein properties responsible for long-term protein phase behavior was established by combining the multidimensional image-based protein phase behavior data with empirical protein properties obtained directly after formulation preparation. A case study was performed to illustrate the application of this approach for formulation development. In this case study, novel formulations were identified based on short-term properties which led to similar phase behavior over time as the original formulation properties. Short-term protein properties that determine long-term protein phase behavior are potential parameters that can be used for protein phase behavior prediction based on in silico descriptors extracted from 3-dimensional (3-D) protein structures. To move towards such short-term descriptor predictions, a high-throughput 3-D protein structure preparation pipeline was developed. This computational pipeline was designed to support high-throughput structure processing, data-dependent structure curation, and robust in silico descriptor extraction.

2.2 Manuscript overview

This subsection presents a compendious list of manuscripts written within the scope of the thesis. The corresponding chapter and page number are indicated per manuscript, followed by a brief summary.

Chapter 3. Application of empirical phase diagrams for multidimensional data visualization of high-throughput microbatch crystallization experiments..25

M.E. Klijn, J. Hubbuch

Journal of Pharmaceutical Sciences (107), 2018, p. 2063-2069

This paper presents a data visualization methodology, which allows for the representation of kinetic and static protein phase behavior data in one figure without the loss of information or comprehensiveness. Subsequently, this multidimensional protein phase diagram was used to discuss the observed static and kinetic protein phase behavior of model protein hen egg-white lysozyme in the context of environmental changes. This resulted in a detailed interpretation and understanding of the obtained protein phase behavior data.

Chapter 4. Time-dependent multi-light source image classification combined with automated multidimensional protein phase diagram construction for protein phase behavior analysis…………………..………………………………**37**

M.E. Klijn, J. Hubbuch

Journal of Pharmaceutical Sciences, In Press

This study shows an automated workflow which starts with raw images obtained from long-term protein phase behavior experiments and visualizes multidimensional protein phase behavior. This was accomplished by enhancing automated image classification with static and kinetic image-based features obtained from visible light, cross polarized light, and ultraviolet light. Predicted classification data was subsequently used to automatically construct a multidimensional protein phase diagram.

Chapter 5. Correlating multidimensional short-term empirical protein properties to long-term protein physical stability data via empirical phase diagrams..**57**

M.E. Klijn, J. Hubbuch

International Journal of Pharmaceutics (560), 2019, p. 166-174

This paper investigates the usability of the empirical phase diagram method to overcome common shortcomings during protein phase behavior characterization, such as limited data set size or simplistic visualization. This resulted in a systematic and data-dependent workflow which created a comprehensive overview of short-term protein properties in an empirical protein property diagram. These short-term properties could be partially related to trends observed in long-term multidimensional protein phase diagrams, which led to a straight-forward characterization of long-term protein phase behavior.

Chapter 6. Redesigning food protein formulations with empirical phase diagrams: A case study on glycerol-poor and glycerol-free formulations…...**75**

M.E. Klijn, J. Hubbuch

Food Research International (125), 2019, p. 108609

The case study presented in this manuscript served as an example of an industrial application for the methodology presented in Chapter 5. The combination of short-term empirical protein property data and long-term protein phase behavior was utilized to redesign a protein-based food product formulation. The combination of an empirical protein property diagram and multidimensional protein phase diagram led to the identification of new long-term stable formulations based on short-term properties similar to the original formulation.

Chapter 7. High-throughput computational pipeline for 3-D structure preparation and in silico protein surface property screening: A case study on HBcAg dimer structures……………………………..……………..…..……**95**

M.E. Klijn[†], P. Vormittag[†], N. Bluthardt, J. Hubbuch ([†]contributed equally)

International Journal of Pharmaceutics (563), 2019, p. 337 - 346

This manuscript describes a computational pipeline which automatically performs homology modelling and subsequent data-dependent curation of 3-D protein structures. Such a preparative pipeline is required before predictive descriptors can be extracted from the 3-D protein structures, but this is usually a manual procedure. An automated approach is of interest when a large amount of candidate structures is screened, as seen during the development of hepatitis B core antigen (HBcAg) virus-like particles. Fast curation simulations, relatively high quality structures, and surface charge descriptors that correlated to experimental data indicated the potential of the proposed pipeline.

Application of empirical phase diagrams for multidimensional data visualization of high-throughput microbatch crystallization experiments

Marieke E. Klijn[1] and Jürgen Hubuch[1]

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

Abstract

Protein phase diagrams are a tool to investigate cause and consequence of solution conditions on protein phase behavior. The effects are scored according to aggregation morphologies such as crystals or amorphous precipitates. Solution conditions affect morphological features, such as crystal size, as well as kinetic features, such as crystal growth time. Commonly used data visualization techniques include individual line graphs or phase diagrams based on symbols. These techniques have limitations in terms of handling large data sets, comprehensiveness or completeness. To eliminate these limitations, morphologic and kinetic features obtained from crystallization images generated with high-throughput microbatch experiments have been visualized with radar charts in combination with the empirical phase diagram method. Morphologic features (crystal size, shape, and number, as well as precipitate size) and kinetic features (crystal and precipitate onset and growth time) are extracted for 768 solutions with varying chicken egg white lysozyme concentration, salt type, ionic strength, and pH. Image-based aggregation morphology and kinetic features were compiled into a single and easily interpretable figure, thereby showing that the EPD method can support high-throughput crystallization experiments in its data amount as well as its data complexity.

3.1 Introduction

Protein phase behavior is of interest for formulation, purification process development, and 3-dimensional (3D) structure analysis. Protein phase behavior is dependent on protein properties, such as net surface charge and structural stability, which are in turn influenced by physical and chemical parameters of the solution[18,102–104]. Resulting protein-protein (PP) and protein-solvent (PS) interactions determine protein phase behavior. The desired protein phase behavior can differ between applications. For example, interplay of PP and PS interactions ideally results in long-term stable pharmaceutical formulations[105,106], whereas particular purification processes are dependent on phase transitions such as crystallization[107]. Crystallized proteins can also be used to obtain 3D structure information[108,109]. Although the application defines the desired phase behavior, finding corresponding solution conditions is carried out using a similar approach. Solution conditions are varied to map their effect on protein phase behavior, where protein concentration, additive type, additive concentration, pH, pressure, or temperature are altered[110–115]. Ternary (e.g., water, additive, and protein concentration)[112,116] or binary (e.g., temperature-pressure or protein-additive concentration)[51,117] protein phase diagrams are frequently used to present optically visible protein phase behavior effects. This information is not only used to identify optimal solution conditions but is also used as a basis for understanding[107,112,115,118], manipulation[119,120], and prediction[77,82,121,122] of protein phase behavior. The visible effect on protein phase behavior is scored in the following morphology categories: clear solution, crystallization, precipitation, skin formation, gelation, or phase separation[54,123]. Different morphological subtypes have been observed within these morphology categories[77,114,124]. For example, crystal subtypes may include micro crystals, sea urchins, needles, plates, and 3D crystals[123,125,126]. These crystal types differ in growth rates, size, and morphology, which are dependent on the growth mechanism and thus dependent on underlying PP and PS interactions[110,111,118,127–130]. Aggregation mechanisms can also determine precipitation size, color, and texture. Amorphous precipitation is considered to originate from nonnative aggregation and appears in darker colors while crystalline precipitation permits native conformation and has a more sandy appearance[54,131,132]. Details of phase behavior morphology provide necessary information that leads to better understanding of PP and PS interactions.

Detailed information on protein phase behavior under multiple conditions, such as morphology subtypes and kinetic features, results in a multidimensional dataset and that is not easily visualized or interpreted. The most simple visualization approach is plotting a single morphologic or kinetic feature as a function of 1 to 3 different solution conditions[110,111,127,130,133–137]. Visualizing high-throughput data with this approach, where more than 3 solution conditions are varied, would result in multiple different figures per feature. This leads to a loss of overview on trends between observed features as well as their connection to all tested solution conditions. Alternatively, symbols are used to capture the morphology effects of solution conditions in binary phase diagrams[82,115,138]. Here, different symbols are used to represent 6 general morphology categories. Adaptations of symbol-based phase diagrams use more symbol types to represent morphology subclasses[121,139]. Symbol-based phase diagrams are easy to interpret and compatible with large data sets. However, kinetic features are not included, and subsetting of morphologic categories is prone to subjectivity[98]. To account for the former, symbol sizes have been scaled to represent kinetic information[140]. Capturing morphology subclasses as well as kinetic parameters results in numerous symbol types and sizes. This makes figures more difficult to interpret, and subjective subsetting is not eliminated. Next to symbols, crystallization images itself have been used to show morphological features and kinetic features, where kinetic features are presented by showing images taken at multiple time points[114,124,125,127,129,135,137,141,142]. Crystallization images contain all desired information, but the data format is inconvenient. Next to the required image size for proper morphology visualization, the use of multiple images over time for high-throughput data becomes highly impractical. This leads to showing examples instead of the complete dataset. The need and generation of experimental protein phase behavior data has become a multidimensional problem but means for data visualization and interpretation are currently insufficient. This makes data evaluation challenging and potentially incomplete. Therefore, a high-throughput compatible comprehensive figure that can present a complete nonsubsetted data set is required[143,144].

A method of combining multidimensional data into one comprehensive figure is the empirical phase diagram (EPD). The EPD was originally developed to combine data obtained from high-throughput experiments on protein conformational states as a function of solvent conditions and stress[78,89]. Multidimensional data is reduced to 3 dimensions, which provides the means to visualize and interpret data with the use of colors, where changes in colors represent differences in underlying features. Three adaptations of the EPD have been previously explored[91]. The first alternative includes a color indexed EPD using predefined colors that correspond to specific protein structural states. The other two proposed adaptations use arbitrary colors in combination with radar plots or Chernoff diagrams to represent underlying multidimensional data. The Chernoff diagrams do not allow for easy interpretation as facial features are used to represent underlying data changes instead of axis, as seen for radar charts. Combining the EPD and radar chart offers the possibility to visualize a large image-based protein phase behavior data set without compromising in data completeness or ease of interpretation. This has not been achieved in previous protein phase behavior studies.

This study uses data on phase behavior of chicken egg white lysozyme during a long-term storage (40 days) microbatch crystallization experiment under 768 different solution conditions at 20 °C. Solution conditions cover 4 pH values (pH 3, 5, 7 and 9), 2 salts (sodium chloride and ammonium sulfate), 12 ionic strengths (0 – 275 mM), and 8 protein concentrations (5-125 g/L). These solution conditions are selected to capture a wide pH range and incorporate 2 commonly used salts which reportedly have ion specific effects[116]. Increments of 25 mM in ionic strength have been chosen to challenge visualization of subtle ionic strength effects, and a wide range of lysozyme concentration is to cover both low and highly concentrated protein solutions. Long-term storage under the selected variety of conditions results in multidimensional data, which challenges previously discussed shortcomings of visualization techniques. Morphologic and kinetic feature extraction was used to prevent loss of protein phase behavior information and minimize subjective morphology category subsetting. Morphologic features describe absolute average crystal size, variation in crystal axial ratio, and the amount of crystals. Precipitation features describe size and color intensity. Kinetic features include onset and growth time of precipitates and crystals. This data set is used to show the benefits of multidimensional visualization techniques for comprehensive and complete presentation of detailed protein phase behavior data using the EPD method in combination with radar plots. For convenience, the term multidimensional protein phase diagram (MPPD) is used for the EPD, which represents the protein phase behavior information.

3.2 Material and Methods

3.2.1 Buffer preparation

The effect of buffer components on protein phase behavior was excluded by using a multicomponent buffer with a 10 mM buffer capacity[145]. Buffer components were CHES (6.13 mM; Applichem), TAPS (14.61 mM; Applichem), MOPS (7.00 mM; Roth), sodium acetate trihydrate (3.01 mM; Merck) and citric acid monohydrate (13.86 mM; Merck). The pH was adjusted using 4 M sodium hydroxide (Merck) as titrant, using a five-point calibrated pH-meter (HI-3220; Hanna Instruments, Woonsocket, RI) equipped with a SenTix 62 pH electrode (Xylmen Inc., White Plains, NY). The pH was adjusted to 3, 5, 7, or 9 with ±0.05 pH unit accuracy. The effect of ionic strength for each buffer was excluded by adjusting conductivity of each buffer to the conductivity measured for the pH 9 buffer. Buffer conductivity was measured with a conductivity probe (Radiometer Analytical, Lion, France). Sodium chloride (Merck) or ammonium sulfate (Applichem) was used for conductivity adjustment. Afterwards, the buffers were filtered using a 0.2 µm cellulose acetate filter (Sartorius, Göttingen, Germany). This buffer served as buffer with a relative ionic strength of 0 mM. A stock buffer with a relative ionic strength 1050 mM was made with sodium chloride and ammonium sulfate for each pH.

3.2.2 Protein stock preparation

A stock solution of 150 g/L lysozyme from chicken egg white (Hampton Research, Aliso Viejo, CA) was made. Lysozyme was weighed and dissolved in the appropriate 0 mM ionic strength buffer. The protein solution was filtered using a 0.2 μm cellulose acetate filter (VWR, Radnor, PA). The filtered protein solution was desalted with a PD-10 column (GE Healthcare Life Sciences, Uppsala, Sweden), employing the manufacturer's centrifugation protocol. The concentration of the stock solution was determined with a Nanodrop 2000c UV-Vis spectrophotometer (Thermo Fischer Scientific, Waltham, MA). An E1% (280 nm) extinction coefficient of 22.00 was used. The protein stock solution was prepared on the same day of crystallization plate preparation.

3.2.3 Long-term storage

All protein solutions were stored in duplicate for 40 days at 20°C in a Rock Imager 54 (Formulatrix, Bedford, MA). The long-term storage experiment was carried out according to the method described by Baumgartner et al. with the following adaptations. Salt and protein dilutions were mixed in a 1:5 ratio by pipetting up and down twice, with a final volume of 24 μL per well. This resulted in lysozyme concentration of 125, 112, 100, 75, 50, 25, 12, and 5 g/L and a relative ionic strength of 0 to 275 mM sodium chloride or ammonium sulfate with a 25 mM step size. Each well was photographed daily during storage with visible light, and more frequently during the first 8 days. All visible light photographs consisted of five focus levels to obtain an averaged sharp picture. After 40 days of storage, UV light imaging was used additional to visible light. UV light exposure time was set to 400 ms, signal amplification (gain) was set to 14.92 dB, and midtone contrast adjustment (gamma) was set to 1.4. A 2.5X zoom was used, and 12 focus levels per well were taken.

3.2.4 Image Features

Precipitation onset time is defined here as the time point when first light precipitation was observed. Precipitation growth cessation time is defined as the time point when the precipitate stopped to change in size and intensity. Crystal onset time is the time point when nuclei were first optically visible (minimum detectable size of ~5 μm). The corresponding growth cessation time is defined as the time point when crystal dimensions did not increase anymore. The difference between the onset and cessation time of precipitates and crystals was calculated as their growth time. Crystal dimensions and precipitate diameter were measured in μm with the ruler tool in the Rock Maker software (version 2.3.0.83). Four crystals were selected to extract their dimensions to form a representative sample group. Four crystal lengths were averaged to represent the absolute average crystal size. In addition, 4 axial ratios were calculated as quantification of crystal shapes. Subsequently, the interquartile range of these values was calculated to represent the diversity of crystal shapes.

3.2.5 Multidimensional data visualization

Each extracted image feature was averaged between the duplicates and normalized between zero and one. A representation of duplicate data is shown in Supplementary Figure A1. Before construction of the MPPDs, all features were evaluated based on internal correlation using the Pearson correlation coefficient. The EPD construction method used is described in literature[78,91]. In short, the dimensionality of the data was reduced to 3 dimensions using singular value decomposition (SVD). The (x,y,z)-values of the 3D data were normalized between zero and one to obtain (x,y,z)-values that can be used as a RGB color value. The optimal number of clusters between 1 and 9 was selected by the function *evalcluster*, available in Matlab, version 2016b. Cluster evaluation used the *k*-means cluster algorithm with a silhouette criterion based on squared Euclidean distance metric. The optimal cluster number was used as input for the *k*-means clustering function (*kmeans*, available in Matlab, version 2016b) to cluster the 3D SVD data using 100 replicates, a maximum of 1000 iterations, and randomly chosen initial cluster centroid positions. The average RGB color value for each cluster was calculated using the normalized (x,y,z)-values of each data point within the cluster. With R (version 1.0.136, using *ggplot2* and *fmsb* library) each data point was plotted against all solution conditions (pH, salt, ionic strength, and protein concentration) and colored with their corresponding average cluster color. In addition, a radar plot was constructed for each cluster to represent the median value of the image feature, as well as the median absolute deviation to represent distribution of the image feature.

3.3 Results and Discussion

3.3.1 Image scoring

Images obtained with visible and UV light were used for morphology and kinetic feature extraction after 40 days of storage. One application of UV light images was to determine whether observed change in phase behavior was a result of protein or non-protein insolubility[109,146]. Examples of visible and UV light images for comparison between clear solutions, crystallized solution, non-protein precipitation, and protein precipitation are depicted in Supplementary Figure A2. Absence of UV signal indicated that all precipitation observed under visible light images were non-protein. On the contrary, all crystals were visible under UV and thus consist of protein. Apart from 3D crystals, no other morphological crystal subtypes were seen in this dataset.

3.3.2 Data treatment and clustering

The use of various features may result in features to correlate to one another and subsequently favor a certain phase behavior property within the dataset. Overrepresentation of a phase behavior property was evaluated based on internal correlation using the Pearson correlation coefficient. The Pearson correlation coefficient matrix is shown Supplementary Table A1. The set threshold of 0.850 was violated by precipitate intensity and precipitation diameter, indicating that precipitate intensity increased as size increased. Precipitate intensity was therefore removed from the dataset. Remaining image features, their corresponding absolute value ranges, and phase behavior property descriptions are summed up in Figure 3.1a. Normalized values of the listed image features were used for SVD dimension reduction. This resulted in a 3D data set with an energy value of 95%, which means a 5% loss of information on data variance. This falls within the general rule, where an energy value of 90% is considered the minimum for reduced dimension data representation[147]. An optimal number of 5 clusters was determined with this three dimensional dataset.

The median cluster value for each image feature is represented within the radar charts using a colored surface, shown in Figure 3.1b. Dispersion of image features in each cluster, represented by the median absolute deviation, is shown as a dotted lined. Cluster 1 corresponds to clear solutions, as all extracted features are equal to zero in images without crystals and precipitates. Coexistence of protein crystals and non-protein precipitation is represented by cluster 2. Here, protein crystals fill roughly 50% of the crystallization well after storage. Crystal nuclei form after 350 ± 40 h and grow for 300 ± 70 h to a reach median size of 48 ± 6 μm. Non-protein precipitation appears after 120 ± 15 hours and grows 65 ± 10 h. The median precipitation diameter is 345 ± 30 μm. Solely crystallized solutions are represented by clusters 3, 4, and 5. Similar to cluster 2, cluster 3 crystals fill half of a crystallization well. Crystal nuclei are observed after 175 ± 20 h of storage, which is earlier compared to cluster 2. Crystals grow for 240 ± 20 h to a size of 36 ± 5 μm. In cluster 4 an increase in crystal number is seen, where wells were almost completely filled after 40 days of storage. A crystal onset time of 90 ± 17 h is lower compared to cluster 3. Growth time is increased to 720 ± 50 h, which resulted in a median crystal size of 96 ± 6 μm. Cluster 5 shows entirely filled wells, where crystals are formed within the first few hours of storage. Growth time has a median of 3 ± 3 h, and the median crystal size is 36 ± 3 μm. The axial ratio interquartile range varies between 0.124 ± 0.06 for cluster 5 to 0.279 ± 0.235 for cluster 2. Crystal shape diversity shows only small deviations for each identified cluster.

3.3.3 Multidimensional protein phase diagram

Results of long-term chicken egg white lysozyme storage at 20 °C using chicken egg white lysozyme under 768 different solution conditions are represented by MPPDs in Figure 3.1c. The top row of Figure 3.1c shows MPPDs for protein solutions containing ammonium sulfate and the bottom row presents the MPPDs where sodium chloride was added to protein solutions. For each individual MPPD, the *y*-axis indicates lysozyme concentration (ranging 5-125 mg/mL) and *x*-axis indicates ionic strength of the corresponding salt (ranging 0-275 mM, 25mM increments).

In protein phase diagrams an undersaturated and supersaturated zone can be identified[50]. The undersaturated zone, represented by cluster 1, contains solutions conditions where no change in protein phase behavior is observed. The supersaturation zone, clusters 2-5, represent solution conditions causing protein aggregation. Figure 3.1c shows cluster transformation and increasing supersaturated zone area for increasing pH values, for both ammonium sulfate and sodium chloride. Solubility dependency on pH is expected as pH affects amino acid residue protonation states[34]. This determines the type, total, and distribution of protein surface charge. In turn, these surface properties influence how the protein electrostatically interacts with the solvent and other solutes. Under these conditions, the surface charge of chicken egg white lysozyme (theoretical isoelectric point (pI) of 11.35[148]) shifts from highly positively charged towards less positively charged (pH 3-pH 9). Protein solubility decreases for solution pH values close to the protein pI as repulsive electrostatic forces diminish. Solubility is at a minimum at a pH equal to the pI[26,34]. In this work, increasing pH values decrease protein solubility and therefore cause larger supersaturated zones.

Protein solubility can also be influenced by specific ions and their ionic strength[26,118]. Proteins can display increasing and decreasing solubility for increasing ionic strength, called salting-in and salting-out, respectively[149]. Solutions containing sodium chloride, depicted in bottom MPPDs in Figure 3.1c, show only salting-out for all pH values. Salting-out becomes more effective for increasing pH indicated by increasing supersaturation zone area. Both observations are in agreement with previously published work[116,150]. On the contrary, ammonium sulfate has no effect on protein solubility at pH 3 and pH 5, whereas salting-in for pH 7 and pH 9 is observed. Salting-in becomes less effective for increasing pH values. Salting-in and salting-out effectiveness of anions and cations is related to its position in the Hofmeister series[41]. Currently, there are many conditions for which a direct or an inverse effectiveness order has been identified[151]. Lysozyme with a net positive surface charge and in presence of relatively low ionic strength range (<300 mM) follows a reversed Hofmeister series[42,152].

a

Symbol	Min	Max	Description
L_C	0	192	Absolute crystal size. Defined as average length of four crystals [μm]
$Đ_{L:W}$	0	3.1	Diversity in crystal shape. Defined as inter quartile range of four crystal axial ratios [-]
Δt_P	0	520	Growth time precipitate [h] ($t_{C.cess}$– $t_{C.onset}$)
t_P	0	324	Onset time precipitate [h]
D_P	0	920	Absolute precipitation size. Defined as diameter of precipitate [μm]
Δt_C	0	959	Growth duration [h] ($t_{C.cess}$– $t_{C.onset}$)
t_C	0	696	Onset time crystal [h]
n_C	0	100	Number of crystals. Scored between 0 and 100, where 100 is a well filled with crystals
●----●	*n.a.*	*n.a.*	$\text{Median}_x \pm$ Median absolute deviation$_x$

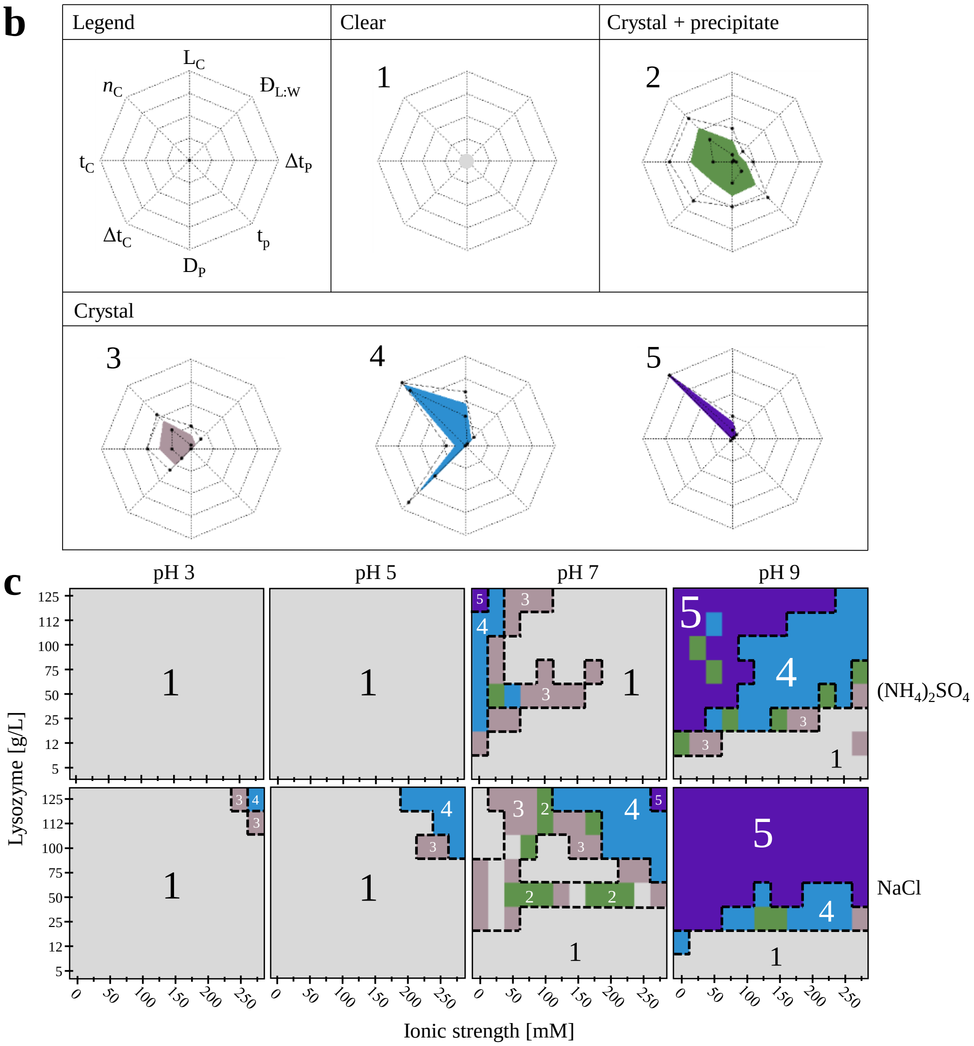


Figure 3.1: (a) Overview of symbols and descriptions of image features including the absolute value range. (b) Radar charts for color based clusters with a legend to indicate the position of image features. The colored surface indicates the normalized median image feature for each cluster. The dotted line indicates the image feature median absolute deviation. (c) Empirical protein phase diagram for lysozyme under varying protein concentrations (*y*-axis), salt ionic strengths (*x*-axis), pH values (grid columns), and salts (grid rows). Five identified color clusters are indicated by mean color as well as a cluster number. Dashed lines are added to highlight regions and guide the eye.

The existence of reversed Hofmeister series has indicated that the classic theory, where only salt hydration properties influences protein solubility based on ion-water interactions, should be expanded by including cosolutes-protein surface interactions[38,41,151,153,154]. In this work, a dominant role of protein surface charge is highlighted by increasing salting-out and decreasing salting-in effectiveness for increasing pH values. However, a unified molecular interaction mechanism explaining salting-in and salting-out is still unknown[154]. Possible molecular salting-in and salting-out mechanisms of lysozyme under tested conditions are interesting phenomena, but it is considered to be outside the scope of this study and therefore not further discussed. The focus lies on combining morphologic and kinetic image-based features into a complete and comprehensive multidimensional protein phase diagram to support straight-forward accessibility of protein phase behavior information via appropriate data visualization.

Figure 3.1c shows cluster transformations from cluster 3 to 5 in direction of higher lysozyme concentrations and higher sodium chloride ionic strength. For ammonium sulfate, a similar cluster transformation is seen in direction of higher lysozyme concentration but, in contrast to sodium chloride, for lower ammonium sulfate ionic strength. The similarity in cluster transformation shows that there is no ion specific effect on the resulting morphology and kinetics. The transformations are similar, but respective cluster areas are different for each salt. For example, cluster 5 is dominating the supersaturation zone at pH 9 in combination with sodium chloride, whereas cluster 4 and 5 are equal in size for pH 9 in combination with ammonium sulfate. This shows that the degree of supersaturation is affected by ion type when all other solution conditions are kept constant. Higher degrees of supersaturation, increasing lysozyme concentration at similar solubility or lowering solubility (e.g. by pH or salt) at equal lysozyme concentration, increases the probability for spontaneous homogeneous crystal nucleation[50,155]. Higher probability results in faster nuclei formation and higher crystal numbers, which corresponds to the differences in both crystal onset time and crystal abundance between cluster 3, 4, and 5 as well as the respective cluster identification at increasing protein concentration, ionic strength, or pH. Previously reported positive correlations between the effects of increased crystal nucleation and degree of supersaturation are in agreement with this observation[111,134–136,156,157]. The change in crystal onset time and crystal abundance between cluster 3, 4, and 5 is also accompanied by a difference in average absolute crystal size. Crystal size is dependent on growth duration as well as speed, which in turn relies on underlying growth mechanisms, response on surface poising (i.e., impurity incorporation or incorrect molecule positioning on the surface), or available amount of protein[127,155,158]. Differently from nucleation, crystal size is not directly correlated to the degree of supersaturation[158]. For pH 3, 5, and 7, crystal size and crystal growth time increase in the direction of higher supersaturation, but at pH 9, the crystal size and crystal growth decrease in the same direction. This is represented by the transformation of cluster 3 to cluster 4 and cluster 4 to cluster 5, respectively. Such an optimum in lysozyme crystal size has been reported before under different conditions[136]. Clusters 3 and 5 both represent similarly small crystals (~36 µm) in different parts of the supersaturated zone. The combination of kinetic data and crystal size suggests that solubility limitations and underlying growth mechanisms are responsible. Cluster 3, identified at the frontier of the supersaturation zone, shows a growth time of ~240 h before it reaches its terminal crystal size. An increase in protein concentration leads to crystals identified by cluster 4, where crystal growth time and crystal size increase simultaneously. Crystal growth cessation can occur when protein molecules still in solution reach the solubility limit as a result of crystal growth[155]. It is suggested that crystal size in cluster 3 was limited by the amount of available material as a lysozyme concentration increase for similar conditions results in further crystal growth. On the other side of the crystal size optimum, at higher supersaturation, cluster 5 is identified. Size limitation may be because of high supersaturation, which increases the number of crystal nuclei. Growth of many nuclei may not be supported by the available amount of material which causes growth cessation[135,158]. However, protein concentrations ranging from 25 to 125 g/L belong to cluster 5, but a lysozyme concentration increase under similar conditions does not influence crystal size or growth time. This indicates that the underlying growth mechanism is limiting. Formation of well-ordered crystals requires proper molecule positioning[155]. Proper molecule positioning on the crystal surface is achieved by moderate interaction between available crystal contacts and the molecule. This should allow for rearrangement of less favorable orientations before incorporation. Strong attractive interactions, which occur at higher supersaturation, may trap molecules in a less ordered state. Such changes to the crystal surface can cause growth cessation[155]. The maximum crystal size of ~96 µm, identified in cluster 4, lies within a region where growth can be supported in terms of available material without extreme nucleation and strong interactions as seen for higher supersaturation. This combination is suggested to be the cause of the crystal size optimum.

3.4 Conclusion

This study shows that multidimensional data visualization using the EPD method allows for comprehensive and complete representation of lysozyme phase behavior under 768 unique solution conditions using 4 morphological and 4 kinetic image-based features in a single figure. Data-dependent clustering resulted in 4 of 5 clusters for a single crystal subtype. This indicates that subtle changes in protein phase transitions, such as crystal onset time, growth time, and crystal size, can be identified. Feature differences as an effect of protein concentration, salt type, ionic strength, and solution pH are easily identified with colored clusters. The combination of morphologic and kinetic features gave insight into the route of crystal formation belonging to a similar morphology subcategory. It was shown that MPPDs are capable of handling large amounts of phase behavior data without challenging data interpretation, a characteristic currently missing in high throughput protein phase behavior experiments. In terms of data handling, improvements are necessary. Image feature extraction was preformed manually for this study, which is time and labor intensive and prone to subjectivity. Image recognition algorithms to substitute manual morphology and kinetic feature extraction would greatly improve time and labor consumption. Image recognition algorithms would also offer the opportunity to use image recognition based features to describe protein phase behavior more systematic and accurate.

3.5 Acknowledgements

This work was financially supported by the BE-Basic Foundation (www.be-basic.org), under project FS2.003. The authors want to thank academic and industrial partners for scientific discussions during the development of this work.

4

Time-dependent multi-light source image classification combined with automated multidimensional protein phase diagram construction for protein phase behavior analysis

Marieke E. Klijn[1] and Jürgen Hubuch[1]

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

Abstract

Image-based protein phase diagram analysis is key for understanding and exploiting protein phase behavior in the biopharmaceutical field. However, required data analysis has become a notorious time-consuming task since high-throughput screening approaches were implemented. A variety of computational tools have been developed to support analysis, but these tools primarily use end point visible light images. This study investigates the combined effect of end point and time-dependent image features obtained from cross polarized and UV light features, supplementary to visible light images, on the classification of protein phase diagram images. In addition, external validation was performed to evaluate the classification algorithm's applicability to support protein phase diagram scoring. The predicted protein phase behavior classes were subsequently used to automatically construct multidimensional protein phase diagrams (MPPDs) to prevent image information loss without complicating the employed image classification algorithm. Combining end point and time-dependent features from three light sources resulted in a balanced accuracy of 86.4 ± 4.3%, which is comparable to or better than more complex classifiers reported in literature. External validation resulted in a correct formulation classification rate of 91.7%. Subsequent automated construction of the MPPDs, using predicted classes, allowed visualization of details such as crystallization rate and protein phase behavior type co-existence.

4.1 Introduction

Protein phase behavior plays an important role in various sectors of the biopharmaceutical field. Knowledge on protein phase behavior indirectly aids in unraveling protein's three dimensional structure which requires a crystalline phase[108,109], but is also essential for downstream processing[107] and formulation development[105,106]. Protein phase behavior is often characterized with protein phase diagrams, which are used as a source of information on protein solubility[159], insoluble aggregate morphology[114,115,119], and aggregation kinetics[160] .

Currently, most protein phase diagram data is obtained via automated imaging systems[161]. Subsequent scoring and analysis of the obtained image datasets is a time-consuming task as typical screenings with an automated imaging system consist of a multitude of 96-well plates. In addition, scoring subjectivity has been raised as a concern. Subjectivity can influence the number of scoring classes that are used, but also the consistency of scoring by experimenters[162,163]. Workload and error reduction by means of computational classification algorithms has therefore been explored in the field. This resulted in a variety of protein phase behavior image classification approaches. An elaborate overview of published work regarding image classification approaches can be found elsewhere[164]. The amount of reports shows how desirable an accurate classification algorithm is for protein phase behavior research. However, some classification algorithm properties and classification performance measures are specifically designed for certain protein phase behavior applications. For example, most classification algorithms focus on the identification of optimal crystallization conditions[165], preferably applicable in real-time during experiments[166,167]. The work presented in this paper focuses on the application of an image classifier to aid protein phase behavior analysis after conclusion of experimental work. This work aims to use the retrieved information to understand effects of different environmental factors to manipulate and potentially predict protein phase behavior. Computational speed is less important here than for real-time applications, as the retrieved empirical data is examined after experiments instead of during experiments. In addition, not solely the identification accuracy of crystals but the identification accuracy of all types of protein phase behavior morphologies is considered important.

One of the major issues in protein phase behavior image classification is the number of protein phase behavior types, sub-types, and the co-existence of these (sub)types[168]. For example, crystallization can be in the form of needle crystals or three dimensional crystals. In addition, these subtypes can co-exist in a single formulation or co-exist with another protein phase behavior type, such as precipitates. Incorporation of more classes to cover the wide variety of possible morphology types has shown a decrease in classification accuracy compared to simpler class systems[169–172]. However, high accuracy for more advanced approaches is required to fully capture the complexity of protein phase behavior. A strategy to improve the accuracy of classification models is the employment of different data sources, which have been explored to find more distinctions between protein phase behavior (sub)types. Evaluated sources include, but are not limited to, protein trace-labeling in combination with green fluorescence[168,173,174], second-order nonlinear imaging of chiral crystals (SONICC)[175], ultraviolet (UV) light[176,177], and two-photon excited UV fluorescence[178]. Despite these alternatives, the main data source remains images obtained with visible light. Another property of the majority of image classification studies is the use of one image per formulation, which is usually the image taken at the end of a protein phase behavior experiment (end point image). However, incorporation of information obtained during protein phase behavior experiments (time-dependent information) has shown to aid crystallization screenings[98,179].

To explore improvements of protein phase behavior classification, this study aims to combine time-dependent and end point features obtained from multi-light source images and assess the impact by means of a random forest classification algorithm. A random forest classification algorithm was selected because the optimization of protein phase behavior classification via more complex algorithms was considered outside the scope of this work. Light sources used in this study are visible light, cross polarized light, and UV light. To the best of the authors' knowledge, no previous studies report on the effects of features extracted from this combination of light sources for protein phase behavior classification. The use of time-dependent data for protein phase behavior classification purposes has also not been reported on before. To evaluate the impact of multiple light sources and time-dependent information on protein phase behavior classification, internal validation of the classification performance has been compared to performances using only end point features obtained from visible light images and performances reported in literature. External validation, by means of scoring 96-well format microbatch experiments, was used to evaluate the applicability of this classification approach for scoring protein phase diagrams.

The classification algorithm in this work used four classes (clear, crystal, precipitate, and other). Classification of protein phase behavior images with a 4-class system generates information on aggregate morphology. However, protein phase diagram images captured over time contain information on aggregate properties and kinetics as well, which is lost when scoring solely protein phase behavior types. Construction of multidimensional protein phase diagrams (MPPDs) allows for the objective representation of aggregate properties and kinetics[180]. This can include aggregation extent, aggregate size, and aggregate growth time. In previous work, MPPDs were constructed with manually extracted image-based features[180]. In this work, the extraction of the image features has been automated. The combination of image classification and subsequent MPPD construction was investigated to determine its potential to aid and innovate computational protein phase behavior classification. By focusing on main protein phase behavior types during classification, the accuracy of the classifier is thought to remain relatively high, while an objective interpretation of the classified protein phase behavior properties can be obtained from the corresponding MPPD.

In total three topics are covered in this study. First, the impact of multi-light source and time-dependent image features on protein phase behavior classification was evaluated via interval validation. Second, an external validation is performed to assess the applicability of the image classification model for scoring protein phase diagrams. Third, MPPDs were automatically constructed using the predicted classes from external validation, where class-based extraction of aggregate growth time and dimensions was used to visualize details on protein phase behavior. The combination of these three topics exemplifies the diversity of image classification approaches, the advantages of additional image sources, and the potential of expanding classification algorithms with automated visualization of multidimensional image-based data.

4.2 Material and Methods

4.2.1 Image dataset

Images were obtained with microbatch crystallization experiments, where 96-well MRC Under-Oil crystallization plates (Swissci, Neuheum, CH) were placed in the automatic image system Rock Imager 54 (Formulatrix, Bedford, MA, USA). Storage time was either 14 days or 30 days. Visible light images were taken at least daily, and cross polarized and UV light images were taken at least six times during the storage period. A total of 57 and 67 images per well were taken during 14 and 30 days of storage, respectively. A detailed overview of the employed image schedules can be found in Supplementary Table B1. The settings for each light source have been previously described[180]. After storage, the end point image (i.e., the image taken at the last time point) of each well was scored using a 4-class system: “clear”, “precipitate”, “crystal”, and “other”. Scoring was performed based on visual inspection of the corresponding visible, cross polarized, and UV light images. In addition, the difference between initial formulation protein concentration and supernatant concentration after storage (data not shown) was taken into consideration. The class “other” was assigned to formulations which did not remain clear over time, but showed no illumination in the UV images and no change in supernatant protein concentration compared to the starting formulation protein concentration. Formulations where co-existence of precipitates and crystals occurred were scored as “crystal”. This resulted in a 4-class dataset of 4416 formulations (~63% “clear” (2797), ~3% “precipitate” (149), ~24% “crystal” (1039), and ~10% “other” (431)). An example image of each class for each light source is shown in Supplementary Material Figure B1.

4.2.2 Feature extraction

Image features were extracted from thumbnail images (200x150 pixels) using MATLAB (version R2015b, MathWorks, Natick, MA, USA). Boundaries of the well and plate were replaced with black pixels (i.e., masked) for each visible light and cross polarized image in order to remove irrelevant information from the image. An example of the masked-out region can be found in Supplementary Figure B2. This was not needed for the UV light images, as a 7x zoom was used during UV light imaging instead of 2.5x zoom during visible light and cross polarized light imaging. The applied 7x zoom eliminated the well walls from the images. Feature extraction for an entire 96-well plate took ~200 seconds (14 days of storage) and ~400 seconds (30 days of storage). Two types of images features were extracted per well: (1) end point features and (2) time-dependent features. All features were extracted for each of the three light sources. The end point features resulted in extraction of 150 features from the final image. These features can be subdivided in three categories: (1) histogram features, (2) blob features, and (3) gray-level co-occurrence matrix (GLMC) features. A complete list of extracted image features can be found in Supplementary Material Table B2. Histogram features were extracted for each color level and the gray image employing the MATLAB function *imhist*. The gray image was obtained using the MATLAB function *rgb2gray*. Blobs were identified by using the Sobel edge detection method with a diamond as structural element set with a distance of 2 pixels from the origin to the points (MATLAB function *strel*). After edge detection, the edges were closed (MATLAB function *imclose*) and filled (MATLAB function *imfill*). After closing and filling, a second mask (with a smaller radius: original mask + 2 pixels) was added to remove the edges of the initial mask from the blob identification. Similar to the initial masking, the second mask was not used for UV light images. Blob properties were retrieved with MATLAB function *regionprops*, and the total pixel area and blob count were calculated. The GLMC of each image was obtained with the MATLAB function *graycomatrix*. Corresponding features were extracted with function *GLCM_Feature1*, which is available via MathWorks file exchange[181]. For the time-dependent feature, every image was subtracted from the first image of the corresponding formulation. This was done to determine the mean pixel intensity change over time. The mean pixel intensity of the end point difference image (start point image – end point image) was used in the classifier as time-dependent feature. The course of mean pixel intensity of difference images over time was used for the construction of multidimensional protein phase diagrams. The mean pixel intensity of each individual image was also determined at each time point, where the start point and end point intensity were used for the construction of the multidimensional protein phase diagrams.

4.2.3 Feature selection

All computational steps after feature extraction were performed with MATLAB version R2018b. Feature selection was performed via two subsequent steps. First, the internal correlation between features was used as a filter to minimize overrepresentation of particular system characteristics. The Pearson correlation coefficient was determined for the complete feature dataset. A threshold of 0.950 for positive and negative linear dependency between features was set. All features with a Pearson correlation coefficient outside this threshold were eliminated. The remaining features were processed with the second step. To minimize the incorporation of noise, features were evaluated and selected based in their feature importance for the classification problem under investigation. This was done with an embedded feature selection method by employing the MATLAB function *TreeBagger*. A bagged (bootstrap aggregated) random forest consisting of 100 trees was built and the relative importance of each feature during classification was extracted. The number of trees was determined by inspection of the obtained out-of-bag error as a function of number of incorporated trees. The results used to select the number of trees can be found in Supplementary Material Figure B3. A cut-off feature importance value was set for the selection of features. The cut-off value was defined as the 50th percentile of the feature importance values of the entire feature set. All features with a feature importance value above the threshold were selected for training the classification model.

4.2.4 Cross-validation

A stratified 10-fold cross-validation was used to evaluate the image classification model performance. Each cross-fold preformed feature selection, as described in Section 4.2.3, and trained a random forest classification model using the selected features and the MATLAB function *TreeBagger*. Similar to feature selection, 100 trees were used in the classification model. The evaluation parameters to quantify the classifier performance were recall, precision, accuracy, and balanced accuracy, as defined by equation 4.1, 4.2, 4.3, and 4.4, respectively.

$$Recall = \frac{TP}{TP + FN} \tag{4.1}$$

$$Precision = \frac{TP}{TP + FP} \tag{4.2}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.3}$$

$$Balanced\ accuracy = \frac{Precision + Recall}{2} \tag{4.4}$$

The evaluation parameters were defined by the true positive count (TP), the true negative count (TN), the false positive count (FP), and the false negative count (FN). These variables were obtained from a confusion matrix after classification. To determine the overall performance after cross validation, an overall average of each evaluation parameter was calculated using the evaluation parameters per class.

4.2.5 External validation

Next to 10-fold cross-validation, the performance of image classification was evaluated for an experimental dataset by means of external validation. The dataset was divided into a model training set and external test set. Three 96-well crystallization plate results (plate A, B, and C) were selected as external test set and the rest of the images (4128 images) were used as model training set. These three plates were selected based on the observed protein phase behavior after storage, to ensure the inclusion of all identified classes during external validation. Plate A contained formulations classified as “clear”, “precipitate”, and “crystal”. In addition, plate A showed formulations with precipitate and crystal co-existence. Plate B contained “crystal” and “clear” formulations, and plate C contained mostly “clear” and “other” formulations. The external validation image classification model was trained with similar settings as mentioned for the 10-fold cross-validation in Section 4.2.4. External validation was evaluated based on the correct classification rate within an entire 96-well plate, and quantified by percentage of correctly scored formulations per plate. Data was visualized by representing the classes obtained from the external validation model as symbols in a 96-well plate format: a scatter plot with eight y-axis values and twelve x-axis values. Class determination by the external validation model returned the probability of the identified class as well (i.e., the probability of the observation to truly belong to the returned class). This classification probability was incorporated in the scatter plot by adjusting the size of the symbol that indicates the position of the formulation in the 96-well plate. This was carried out by multiplying the probability value, when it fell below 0.750, with the default symbol size. This means smaller symbols were obtained for lower probability values. The overall probability of the classification was quantified by the mean value of all 96 probability values.

4.2.6 Automated multidimensional protein phase diagram (MPPD) construction

An MPPD was automatically constructed to retrieve detailed information on the classified formulations. Automated construction of an MPPD was performed with end point and time-dependent features. An overview of the employed image features for the MPPD can be found in Supplementary Figure B3. Each of the listed features were extracted for all three light sources. This means that each main feature listed in Supplementary Figure B3 consisted of three sub-features, namely one for each light source. After extraction, the three sub-features were averaged to represent the corresponding main feature. The time-dependent feature growth time was determined as follows. The intensity difference over time (extraction is explained in Section 4.2.2) was fitted with a smoothing spline function by employing the MATLAB function *fit* and a smoothing parameter of $1{\cdot}10^{-4}$. The fitted function was used to calculate the first derivate. The first zero value (in time) of the derivative was extracted, which represents the point in time at which intensity change ceased. This point in time was used as the definition for the end of aggregation growth, and thus aggregation growth time. If a zero point could not be found, the last time point was set as growth time.

The image features listed in Supplementary Table B3 were extracted for formulations that were scored as “precipitate” or “crystal” by the external validation classification model. Formulations scored as “clear” or “other” were not included during feature extraction for the MPPD data set, as aggregate properties are non-existent in these formulation. All values for “clear” and “other” classified formulations were set to zero. The data of all three plates was used to construct an MPPD for each plate. Methods used to obtain the MPPD, such as dimensionality reduction and visualization, have been previously been described by Klijn et al.[180]. The optimal cluster number was set to range from 1 to 6 and a Pearson correlation coefficient cut-off value of 0.850 was used.

4.3 Results and Discussion

4.3.1 Feature set evaluation by 10-fold cross-validation

The effect of incorporating different light sources and a time-dependent image feature on the classification of protein phase diagram images was evaluated. This was done by performing 10-fold cross-validation for multiple image feature sets and determining the accuracy, balanced accuracy, precision, and recall. Balanced accuracy (the average of recall and precision) was used as an evaluation parameter because protein phase diagram image datasets often deal with a class imbalance[168,182]. This class imbalance is not only an aspect to take into account during model training via proper class representation, but also during model evaluation. Six feature sets have been evaluated in this study: (1) image features extracted from visible light end point images (Vis); (2) Vis feature set combined with a time-dependent image feature extracted for all three light sources (Vis+Time); (3) Vis feature set combined with image features extracted from cross polarized end point images (Vis+CP); (4) Vis feature set combined with image features extracted from UV light end point images (Vis+UV); (5) all features extracted from end point images for each light source (Vis+CP+UV); and (6) time-dependent features combined with the fifth feature set (Vis+CP+UV+Time). For each feature set, and each fold during cross-validation, feature selection and classification model training was performed. It was observed that feature selection did not return different features between the 10 folds for the same feature set (data not shown). An overview of the selected image features per evaluated feature set can be found in Supplementary Figure B4. Figure 4.1 shows the mean evaluation parameters for each feature set after 10-fold cross-validation. Accuracy was added to the evaluation parameters as it is often used in studies when evaluating image classification models. Therefore, accuracy may be of interest for other work as comparable evaluation parameter. However, for the reasons mentioned above, the balanced accuracy will be used as the overall performance measure in the discussion of this work.

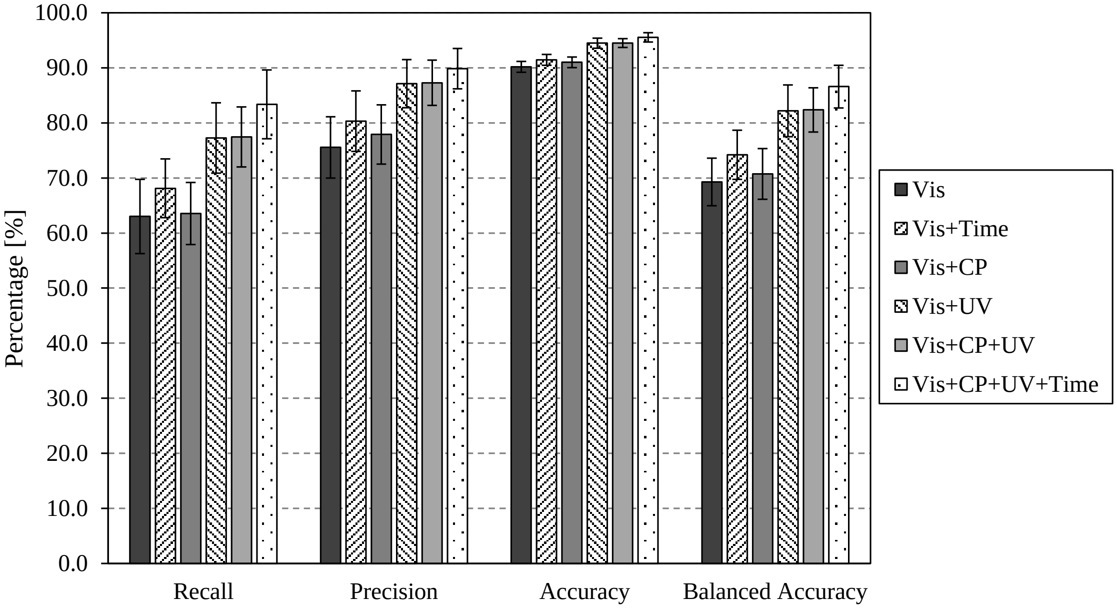


Figure 4.1: Recall, precision, accuracy, and balanced accuracy in percentages for six different feature set. Vis: visible light end point image features; Time: time-dependent feature from each light source; CP: cross polarized light end point image features; and UV: ultraviolet light end point image features. Error bars represent the standard deviation calculated for each evaluation parameter obtained after internal 10-fold cross validation.

The poorest classification model performance was obtained for the Vis feature set. The evaluation parameters show a recall, precision, and balanced accuracy of 63.0 ± 6.8%, 75.5 ± 5.6%, and 69.3 ± 4.3%, respectively. Note that evaluation parameters are mentioned as mean ± standard deviation which was calculated with the evaluation parameters of each fold and each class obtained during 10-fold cross validation. The performance of all other feature sets are compared to performance of the Vis feature set, as visible light end point images are most often used to classify protein phase behavior images. An increase in classification model performance was seen for each addition to the Vis feature set. The addition of end point image features obtained with cross polarized light (Vis+CP) showed the smallest increase. The balanced accuracy increased from 69.3 ± 4.3% (Vis) to 70.7 ± 4.6% (Vis+CP). Recall and precision increased with 0.5 percent point and 2.4 percent point, up to 63.5 ± 5.6% and 77.9 ± 5.4%, respectively. The addition of time-dependent features (Vis+Time) showed a 4.9 percent point increase in balanced accuracy (74.2 ± 4.5%) compared to the classification using solely Vis (69.3 ± 4.3%). Recall and precision increased from 63.0 ± 6.8% and 75.5 ± 5.6% to 68.1 ± 5.8 and 80.3 ± 5.5%, respectively. The largest increase compared to Vis was seen for the addition of features extracted from the end point images obtained with UV light (Vis+UV). Balanced accuracy increased up to 82.2 ± 4.7%. Recall and precision showed an increase of 14.3 percent point (77.3 ± 6.4%) and 11.6 percent point (87.1 ± 4.4%), respectively, compared to the performance when using Vis. The addition of cross polarized light feature images to the Vis+UV feature set, to obtain the Vis+CP+UV feature set, resulted in a classification performance comparable to the Vis+UV feature set. The observed increase of 0.2 percent point upon comparison of Vis+UV and Vis+UV+CP in terms of recall, precision, and balanced accuracy falls within the standard deviations of the performance evaluation parameters of the Vis+UV+CP feature set (77.4 ± 5.5%, 87.3 ± 4.1%, and 82.4 ± 4.0%, respectively). This observation corresponds to the small increase in performance when comparing the Vis and Vis+CP feature set. The lack of relevant information from cross polarized light images in this classification problem may be due to the limited dataset size and diversity. It could also be due to the extraction of similar features for each light source. Future work can determine whether specific image-based features for specific light sources increases the light source's relevance. Addition of time-dependent features (Vis+CP+UV+Time) resulted in the best performance of all tested feature sets. A balanced accuracy of 86.6 ± 3.9%, a recall of 83.4 ± 6.2%, and precision of 89.8 ± 3.7% were obtained. The evaluation parameters of Vis+CP+UV+Time showed a relatively small increase from the Vis+CP+UV feature set, namely an increase of 5.9, 2.5, and 4.2 percent point, respectively. Nevertheless, compared to the Vis feature set, addition of different light source features and time-dependent features resulted in an overall increase of 20.4, 14.3, and 17.3 percent point for recall, precision, and balanced accuracy, respectively.

Table 4.1 lists the mean evaluation parameters for all four classes obtained during 10-fold cross-validation with the Vis+CP+UV+Time feature set. The evaluation parameters per class are shown for this feature set as it showed the best classification performance in this study. Mean evaluation parameters obtained during cross-validation per class for all other feature sets can be found in Supplementary Table B5.

Table 4.1: Average recall, precision, accuracy, and balanced accuracy listed in percentage per protein phase behavior class obtained for 10-fold cross-validation using the image feature set containing visible, cross polarized, ultraviolet light image features in combination with a time-dependent feature. The values are listed as mean ± standard deviation. This was calculated based on the 10 folds obtained from internal cross-validation.

	Recall [%]	**Precision [%]**	**Accuracy [%]**	**Balanced Accuracy [%]**
Clear	96.9 ± 1.1	90.8 ± 1.5	91.8 ± 1.1	93.8 ± 0.8
Precipitate	83.2 ± 13.6	93.9 ± 5.8	99.1 ± 0.6	88.6 ± 7.8
Crystal	86.4 ± 3.0	96.8 ± 1.6	96.2 ± 0.7	91.6 ± 1.6
Other	66.9 ± 7.3	77.9 ± 5.9	95.1 ± 0.9	72.4 ± 5.3

Table 4.1 shows that the “other” class resulted in the lowest performance. This is represented by a recall and precision of 66.9 ± 7.3% and 77.9 ± 5.9%, respectively. The accuracy does not reflect this poor performance (95.1 ± 0.9%), which highlights the role of an imbalanced class distribution on performance evaluation. The balanced accuracy paints a more realistic picture of the classification model performance, with 72.4 ± 5.3% for the “other” class. The balanced accuracy also reflects that the classifier performs well for classes “clear” and “crystal”, represented by 93.8 ±0.8% and 92.6 ± 1.6%, respectively. However, recall for the “crystal” class is lower than the “clear” class (86.4 ± 3.0% versus 96.9 ± 1.1%), while the opposite is seen for precision (90.8 ± 1.5% for “clear” and 96.8 ± 1.6% for “crystal”). This reflects that crystal formulations show a higher false negative rate, which means that crystallized formulations are more often missed than clear formulations. On the other hand, “clear” formulations showed a higher false positive rate. The higher false positive rate of the “clear” class and the low performance of the “other” class are both attributed to misclassifications between “clear” and “other” images. This can be deducted from the individual confusion matrices of the Vis+CP+UV+Time feature set, shown in Supplementary Material Table B6. The “precipitate” class showed a moderate balanced accuracy of 88.6 ± 7.8%. A recall of 83.2 ± 13.6% reflects a relatively high false negative rate and a large deviation between the folds. The large deviation is presumable due to the small contribution of the “precipitate” class to the total dataset (~3%). The corresponding precision for the “precipitate” class of 93.9 ± 5.8% reflects a low false positive rate.

A considerable amount of research has been published concerning image recognition for protein phase behavior studies[164]. However, due to the wide range of classification classes, image sources, training/test sets sizes, algorithms, and classification optimization targets, it is difficult to put new protein phase behavior classification work into perspective. Three studies have been selected to put the work presented in this study into perspective. The data is shown in Table 4.2. The studies have been selected based on the available data, number of classification classes, and type of classifier. The available published performance parameters have been converted to match the definitions as described in Section 4.2.4. The number of classification classes are considered as a selection criterion, as the number of classes shows a large influence on the evaluation parameters. High class systems tend to show a poor performance, while two class systems show a relatively good performance.

Table 4.2: Overview of the average recall, precision, accuracy, and balanced accuracy in percentages for literature data and the work presented in this study. The values are given in mean ± standard deviation, which are determined based on the corresponding evaluation parameters for all considered classes.

	Bruno et al.[167]	**Cumbaa et al.**[171]	**Sigdel et al.**[168]	**This study**
Classified type	CNN	RF	RF	RF
Number of training images	442930	124816	714	4234
Number of classes	4	3	3	4
Recall [%]	88.7 ± 11.3	87.5 ± 7.9	85.6 ± 14.3	83.4 ± 6.2
Precision [%]	92.9 ± 3.2	80.0 ± 16.9	87.9 ± 8.4	89.8 ± 3.7
Accuracy [%]	97.2 ± 1.0	91.0 ± 4.0	96.5 ± 2.3	95.5 ± 0.8
Balanced accuracy [%]	90.8 ± 7.2	83.7 ± 10.9	86.8 ± 10.3	86.6 ± 3.6

The work of Bruno et al. employs similar classes (“clear”, “crystal”, “precipitate”, “other”)[167], while Sigdel et al. and Cumbaa et al. employed a 3-class system (“clear”, “crystal”, “other”)[168,171]. In addition to the classes, the type of classifier was taken into account. As listed in Table 4.2, deep convolutional neural networks (CNNs) were used by Bruno et al., while a random forest classifier was used by the Sigdel et al. and Cumbaa et al. The type of classifier is of interest because of the required computational time and expertise to design and train a classification model. Deep CNNs classification algorithms are considered advanced and computational expensive, while random forest classification algorithms are considered more transparent and computational inexpensive. For the application of a classification model in a laboratory with powerful yet ordinary computers and scientists with moderate programming skills, it would be beneficial to keep algorithms computational inexpensive and accessible. However, simpler classification algorithms tend to be less accurate.

Table 4.2 shows deep CNNs as the best performing classifier. The percent point difference between the work by Bruno et al. and the work presented here is 5.3, 3.1, and 4.2 for the average recall, precision, and balanced accuracy, respectively. Even though the deep CNNs performance is better on average, the simplistic approach and basic features that are used in this study already result in a classification performance that lies within the standard deviation range. This highlights the potential of time-dependent and multi-light source data for image classification accuracy improvement, while employing a simple and transparent classification algorithm. The use of more complex features from visible light end point image by Cumbaa et al. resulted in 4.1 percent point higher recall compared to the study presented here. Contrarily, the precision and balanced accuracy are 9.8 and 2.9 percent point lower for the work by Cumbaa et al., respectively. Performance parameters of Sigdel et al. are comparable to performance parameters of the current study. The main difference between the current study and the work presented by Sigdel et al. is the higher standard deviation reported for the latter. This is represented by an 8.1, 4.7, and 6.7 percent point higher standard deviation for recall, precision, and balance accuracy, respectively. In general, the standard deviation for the current study is smaller compared to the three literature studies, which indicates that the performance between classes is more consistent.

Presented internal validation results and the subsequent comparison with previously published work shows the advantage of information obtained from multiple light sources as well as information obtained over time for protein phase behavior image classification. The full potential of this classification approach can be assessed in future work, where the effects of more complex and light source-specific features, a rigorous feature optimization workflow, and a larger and more complex image dataset should be investigated.

4.3.2 Protein phase diagram prediction

Cross-validation was performed to obtain an estimation of the model performance, but the application of the classification model would be to classify protein phase diagram images after an experiment is completed. To evaluate the performance for this application, an external validation was carried out using the best performing feature set during internal 10-fold cross-validation, namely the combination of visible light, cross polarized light, UV light, and a time-dependent feature. To evaluate the external validation, images obtained from three 96-well microbatch crystallization plates were selected to be classified. These plates were selected so that all four classes were part of the external validation. All other 4128 formulations were used to train the classification model. Before training the external validation classification model, feature selection was performed. Feature selection resulted in the removal of 94 features based on internal correlation coefficient and 28 based on feature importance. The corresponding correlation coefficient matrix and feature importance graph can be found Supplementary Figure B4 and Figure B5, respectively. A total of 28 features remained to train the classification model, which are listed in Table 4.3. The resulting classification of the three excluded plates by the obtained classification model is depicted in Figure 4.2.

Table 4.3: Overview of selected image feature for the external validation classification model.

Number	Feature description	Light source
1 - 3	Entropy color level red, green, and blue	Visible light
4	Mean pixel intensity color level red	
5 - 6	Pixel variance color level red and blue	
7	Total count of blobs	
8	Contrast obtained from GLMC	
9	Correlation obtained from GLMC	
10	Cluster shade obtained from GLMC	
11	Energy obtained from GMLC	
12	Entropy obtained from GMLC	
13	Measure of correlation 2 obtained from GLMC	
14	Total intensity difference	
15	Entropy color level red	Cross polarized light
16	Total area of blobs	
17	Total intensity difference	
18	Entropy color level red	UV light
19	Mean pixel intensity color level red	
20 - 21	Skewness and kurtosis color level red	
22 - 23	Total area and count of blobs	
24	Contrast obtained from GLMC	
25	Correlation obtained from GLMC	
26	Cluster shade obtained from GLMC	
27	Energy obtained from GMLC	
28	Total intensity difference	

Colored symbols in Figure 4.2 represent the predicted class, where the size of the symbol is adjusted for its classification probability. A smaller symbol represents a lower classification probability, which was thought to help identify a possible misclassification. A circle around the symbol indicates that the formulation was incorrectly classified. The true class can be identified by the color of the circle, which is similar to the color of the symbol classes. The correct classification rate and mean probability of the entire plate is shown below each plate. Figure 4.2a shows a correct classification of 94.8% and a corresponding mean probability of 0.919 for plate A. For plate B in Figure 4.2b, 3 out of 96 formulations were incorrectly classified, represented by a correct identification value of 96.9%. Figure 4.2c (plate C) shows an incorrect classification of 16 out of 96 formulations, which is reflected by a correct identification percentage of 83.3%. The overall probability is also lowest for plate C, with a value of 0.770. The misclassifications in plate C correspond to confusion matrices obtained after 10-fold cross-validation (Supplementary Table B6) for the Vis+CP+UV+Time feature set. Confusion matrices show that “other” misclassifications are most often classified as “clear” and vice versa. In combination with a mean recall of 66.9 ± 7.3% it is not unexpected that 8 out of 23 (~35%) of the “other” formulations were misclassified.

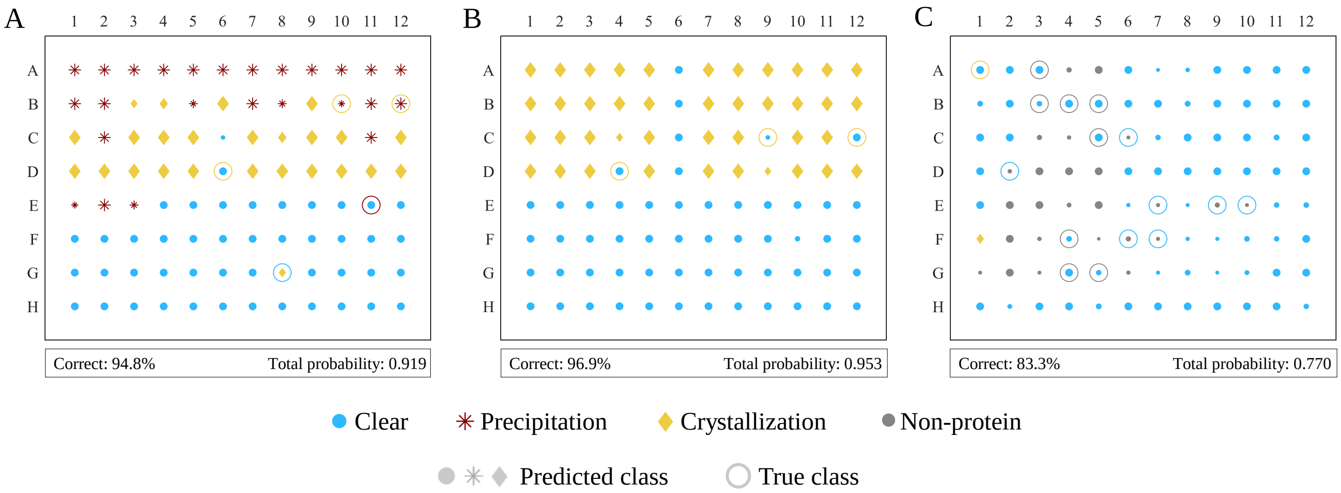


Figure 4.2: Classification results of external validation for (a) plate A, (b) plate B, and (c) plate C. The correct classification percentage and total plate probability is listed below each plate. Symbols represent the four predicted classes: (1) clear (blue dot), (2) precipitation (red asterisk), (3) crystal (yellow diamond), and (4) non-protein (gray dot). Symbol size is adjusted to the individual classification probability. Open circles represent the true class in case of misclassification. Colors represent the four true classes: (1) clear (blue), (2) precipitation (red), (3) crystal (yellow), and (4) non-protein (gray).

Misclassification of crystallized formulations in plate A (B10 and B12) was most likely due to inability to recognized ~7 small crystals (<50 µm) in co-existence with precipitation. Considering the misclassified crystallized formulation in plate B (C9, C12, D4) and plate C (A1), which also contains a few small crystals, it can be concluded that the classifier does not perform well for the classification of a few small crystals. A hardware related limitation was found when inspecting the other misclassified crystallized formulation in plate A (D6). This formulation showed small crystals near the well walls, which are missed by the UV light images. This is due to the available minimal zoom for UV light images (7x for UV light versus 2.5x for visible and cross polarized light images), which results in an image lacking the outer part of the liquid formulation. Precipitates and crystals occurring at the well wall are therefore not captured in UV light images. It is assumed this issue can be resolved by applying a smaller zoom to capture the entire liquid formulation in UV light images. Class probability was incorporated in the symbol-based protein phase diagram to target uncertain model classifications for closer manual inspection more easily. However, external validation showed that not all incorrect classification have a corresponding low probability. For plate A, the probability for the misclassifications ranges from 0.570 to 1.000 with a mean of 0.760. Plate B shows a range from 0.580 to 1.000, with a mean of 0.810. Plate C shows a wider range of 0.500 to 0.970, with a mean of 0.720.

External validation resulted in an overall correct classification value of 91.7% and overall classification probability of 0.881. These results indicate that the classification model in the current state is applicable as a fast, but solely preliminary assessment of protein phase diagrams. Elimination of human interpretation is not yet possible. Further development, as suggested in Section 4.3.1, is required to obtain a more reliable classification model. An additional approach that would be of particular interest to enhance the identification of uncertain classifications in protein phase diagrams, is the use of two or more parallel classification algorithms as presented by Buchala and Wilson[172]. The combined probability of multiple classifiers could enrich the classification outcome by a more accurate representation of classification uncertainty.

4.3.3 Automated multidimensional protein phase diagram (MPPD)

Image-based protein phase diagram studies are information-dense experiments. Besides information on the resulting protein phase behavior, imaging experiments contain information on aggregate dimensions such as crystal size, as well as information on aggregation kinetic properties, such as growth time. In previous work, construction of MPPDs showed how this rich information can be visualized and how MPPDs can aid protein phase diagram interpretation[180]. Extraction of aggregate dimensions and kinetic properties was previously done manually. In this work, results obtained from the protein phase behavior classification algorithm allowed for subsequent automated extraction of the information required to construct an MPPD. MPPD construction resulted in the clustering of features describing the total aggregated area (image feature: total blob area), the crystal count (image feature: number of blobs), the crystal length (image feature: mean blob major axis length), and the aggregation growth time (image feature: time point at which no intensity change was observed anymore). These features were reduced to three dimensions, with an energy loss of 1.4%. This energy loss falls within the accepted 10% energy loss after dimension reduction[147]. The obtained three dimensional dataset was used to cluster formulations together that show similar properties. These clusters are displayed as a uniform group in MPPDs. The results for plate A, B, and C are shown in Figure 4.3.

Figure 4.3 shows three MPPDs (left) and five radar charts to represent data-dependently identified formulation clusters (right). Radar charts show a median value for each extracted feature as a color surface and a dashed line to indicate the median absolute deviation within clusters. Cluster 1 represents the formulations that were classified as “clear” or “other” as all image feature values equal zero. Cluster 2 represent “precipitate” formulations, as crystal specific features equal zero. Cluster 3, 4, and 5 represent “crystal” formulations. The fact that three clusters were identified to represent “crystal” formulations, shows that more information can be obtained without complicating the initial protein phase behavior classification algorithm.

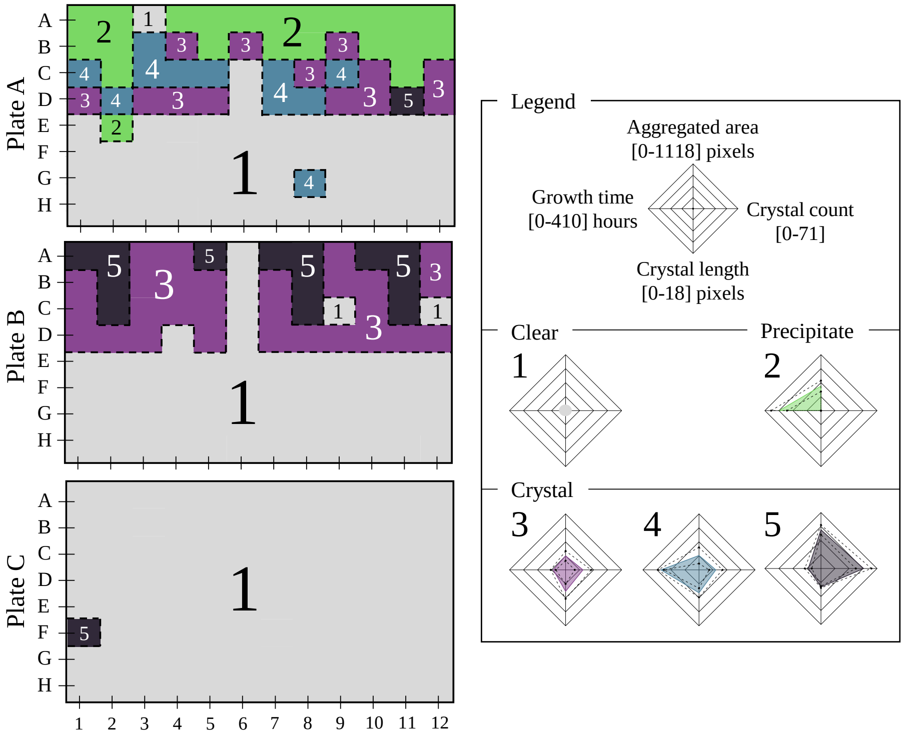


Figure 4.3: Multidimensional protein phase diagrams for plate A, B, and C. Formulation clusters are indicated by color and number. Clusters are separated by a dashed line to guide the eye. Formulation cluster numbers correspond to radar chart numbers. Each radar chart represents the median value of the image features shown in the legend radar chart with a colored surface. The black dashed line inside each radar chart represents the median absolute deviation.

Crystal length is comparable (~7 pixels) for all three crystal clusters, but differences were observed for aggregated area, growth time, and crystal count. Cluster 3 formulations showed a median aggregated area of 280 ± 101 pixels, with a median crystal count of 22 ± 11 which grew for a median of 90 ± 16 hours. Note that cluster values are mentioned as median ± median absolute deviations. Cluster 4 formulations show a larger aggregation area (median of 291 ± 157 pixels) compared to cluster 3 formulations, which was obtained over a longer period of time (median of 279 ± 21 hours). However, cluster 4 formulations contained a comparable crystal count (median of 21 ± 9) as cluster 3 formulations. A comparable crystal count and crystal length combined with a larger aggregation area is presumably a consequence of the co-existence of crystals and precipitates, which was not seen for cluster 3 formulations. Results for cluster 4 indicate that co-existence can potentially be excluded from image classifiers as a (sub)class and be identified with complementary MPPDs. This benefits protein phase behavior classification, as recognition of co-existing phases is one of the main issues[168]. Cluster 5 shows a further increase in aggregated area (median of 771 ± 101), which was obtained in 94 ± 25 hours. The increase of crystal count (54 ± 10) indicates that the increase in aggregated area is a result of increased formation of crystals, and not co-existence with precipitates as seen for cluster

4. In addition, the higher crystal count also represents an increased nucleation rate for cluster 5 formulations compared to cluster 3 and cluster 4. The quantification and visualization of nucleation rates allows one to assess the level of super saturation.

The discussed workflow from raw multi-light source images to symbol-based protein phase diagrams and complementary MPPDs showed the diversity of image-based information in the field of protein phase behavior studies. Prior to application of the proposed workflow for classification of screening experiments, the impact of the employed UV light on the protein in question should be assessed beforehand. This is noted as protein aggregation propensity may be affected by the use of UV light imaging[183]. Future work should focus on optimized MPPD information extraction for each light source and test a larger image database with more protein phase behavior (sub)types. The small scale and relative simplicity of the presented study indicates that more advanced techniques and data sets could increase its potential to aid analysis of protein phase behavior studies.

4.4 Conclusion

This study presented three topics concerning image-based data for protein phase behavior analysis. The first topic demonstrated that combining multiple light sources (visible, cross polarized, and UV light) with time-dependent features improves the classification accuracy of protein phase behavior images. A balanced accuracy of 86.4 ± 3.9% was achieved during 10-fold cross-validation. This was a 17.3 percent point increase compared to 10-fold cross-validation with only visible light image features extracted from end point images. Evaluation of multiple image feature sets showed that features obtained from UV light images were most influential, followed by the time-dependent feature. The second topic covered the external evaluation of the image classification model to determine its applicability to protein phase diagram scoring. External validation resulted in an overall correct protein phase behavior classification of 264 out of 288 formulations (91.7%). The third topic used the external validation classification results to investigate the combination of image classification and objective multidimensional data visualization to exploit the information-rich image data without complicating the classification algorithm. It was shown that automated MPPDs can complement automatically classified protein phase diagrams by distinguishing phase co-existence and changing nucleation rates within the "crystal" class. The results presented for these three topics indicate that merging different approaches allows protein phase behavior research to benefit from the strength of each aspect. Hardware variety aids the distinction between protein phase behavior types, employing different visualization techniques allows one to capture several levels of information, and implementation of automated computational approaches minimize the workload.

4.5 Acknowledgements

This work was financially supported by the BE-Basic Foundation (www.be-basic.org), under project FS2.003. The authors want to thank academic and industrial partners for scientific discussions during the development of this work.

5

Correlating multidimensional short-term empirical protein properties to long-term protein physical stability data via empirical phase diagrams

Marieke E. Klijn[1] and Jürgen Hubuch[1]

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

Abstract

Identification of long-term stable biopharmaceutical formulations is essential for biopharmaceutical product development. Reduction of the number of long-term storage experiments and a well-defined formulation search space requires knowledge-based formulation screenings and a detailed protein phase behavior understanding. To achieve this, short-term analytical techniques can serve as predictors for long-term protein phase behavior. Protein phase behavior studies that investigate this concept commonly display shortcomings such as limited and small datasets, sample adjustments, or simplistic data analysis. To overcome these shortcomings, 150 unique lysozyme solutions were analyzed using six different short-term analytical techniques. Lysozyme's structural properties, conformational stability, colloidal stability, surface charge, and surface hydrophobicity were obtained directly after formulation preparation. Employing the empirical phase diagram method, this short-term data was correlated to long-term physical stability data obtained during 40 days of storage. Short-term protein properties showed partial correlation to long-term phase behavior. Identification of different structural conformations related to changing surface properties, colloidal stability, and conformation stability as a function of formulation conditions. This study contributes to long-term protein phase behavior research by presenting a systematic, data-dependent, and multidimensional data evaluation workflow to create a comprehensive overview of short-term protein analytics in relation to long-term protein phase behavior.

5.1 Introduction

Protein phase behavior characterization is necessary to identify stable formulation conditions for long-term storage of biopharmaceuticals. Stable biopharmaceutical product formulations are found via prolonged periods of storage time (6 to 60 months) of the biopharmaceutical compound in question under varying formulation conditions[184]. This approach of searching for optimal formulation conditions is time consuming, has a trial-and-error nature, and experience is required for solution parameters selection. Therefore, it is desired to reduce development time and costs by rational design of protein characterization experiments and move towards long term protein phase behavior prediction[12,185]. This can be done by identifying short-term measurable protein properties which correlate to long-term protein phase behavior. This requires that short-term protein properties capture protein-protein, protein-solvent, and protein-cosolute interactions that induce long-term physical instability. The strength and type of these interactions are dependent on intrinsic protein properties, which are in turn influenced by physical and chemical parameters of the formulation and the protein itself [18,26,104]. Observed protein phase behavior can result from various aggregation pathways due to different underlying aggregation mechanisms[20,29]. In this study, physical instability refers to insoluble and optically visible aggregate formation, such as crystals or amorphous precipitates.

The Derjaguin-Landau-Verwey-Overbeek (DLVO) theory describes the dependency of intermolecular protein-protein interactions (i.e., colloidal stability) on long-range electrostatic repulsion and weak, short-range van der Waals attraction[27]. High protein concentration formulations have shown that non-DLVO forces, such as hydrophobic forces and hydrogen bonding, influence colloidal stability as well as excluded volume[186–189]. In addition to colloidal stability, conformational stability plays a key role in protein aggregation as well and is dependent on solvation and intramolecular properties[47]. Therefore, a quantification of colloidal and conformational stability is necessary to characterize protein aggregation. Analytical techniques to assess colloidal and conformational stability are elaborately discussed in available reviews[23,46,190]. Specific protein properties, such as protein charge and surface hydrophobicity, should be taken into account as well when investigating long-term protein aggregation. These properties can determine the occurrence and degree of colloidal or conformation stability and are therefore an important part of investigating cause and consequence of protein physical instability[21,191]. Therefore, investigation of the correlation between short-term empirical protein properties and long-term physical stability requires multiple analytical techniques. The requirement of combining analytical techniques has been recognized and has been used to investigate protein phase behavior as a function of formulation conditions in several different studies[47,77,82,139,188,192–195]. These combinations of experiments can increase protein phase behavior understanding, which in its turn aids rational formulation screening design and predictive parameter identification[23]. Some studies have already used empirical short-term properties to investigate the predictive power in relation to long-term protein phase behavior[82,193,195,196]. Based on the reported combinations of analytical techniques it can be stated there is a continuing need to generate data and knowledge on formulation-dependent long-term protein physical stability. In particular, there is a need to continue this type of research while simultaneously eliminating specific shortcomings. Reported studies show limitations that can be described with one or more of the following four issues: (1) incomplete datasets by including only conformational stability or colloidal stability, or a general lack of protein properties measurements to determine properties such as protein surface charge; (2) relative small datasets (n < 30); (3) analytical techniques that require formulation condition adjustments, such as dyes or sample dilution; (4) simplistic data analysis techniques which hinders data visualization and interpretation.

This study seeks to overcome these limitations while investigating the correlation between short-term empirical protein properties and long-term protein phase behavior. To avoid incomplete and small data sets, the first and second shortcoming, the model protein chicken egg white lysozyme was formulated in 150 unique solution conditions and analyzed using six different analytical techniques. The employed analytical techniques included dynamic light scattering (DLS), Fourier-transform infrared (FTIR) spectroscopy, static light scattering (SLS), intrinsic fluorescence (IF) spectroscopy, mixed-mode measurement phase analysis light scattering (M3-PALS), and stalagmometry to represent formulation condition effects on inter-particle interactions, secondary structure, colloidal stability, conformational stability, surface charge, and surface hydrophobicity, respectively. None of the employed techniques required formulation adjustments, thereby eliminating the third shortcoming. The extracted short-term empirical protein properties were lysozyme's apparent hydrodynamic radius, FTIR peak region areas, aggregation onset temperature, melting temperature, zeta potential, and normalized surface tension. Formulations were designed to cover a wide range of conditions. The conditions covered 4 pH values (pH 3.0, 5.0, 7.0, and 9.0), 2 salts (ammonium sulfate and sodium chloride), 4 ionic strengths (0, 50, 175, and 275 mM) and 5 lysozyme concentrations (25, 50, 75, 100, and 125 g/L). Long-term phase behavior of lysozyme under identical formulation conditions was monitored with 96-well format microbatch experiments, where formulations were stored for 40 day at 20 °C [115,180]. The fourth shortcoming, data visualization and interpretation, was resolved by using an advanced method of compiling multidimensional data into a comprehensive figure called the empirical phase diagram (EPD)[78,89]. Reducing multidimensional empirical data to three dimensions provides the means to visualize and interpret data more easily by making use of color clustering. Colors were used to distinguish differences in empirical data as a function of formulation conditions. Radar charts can complement EPDs by providing an overview of underlying empirical data for each color cluster. The effectiveness of this visualization technique was demonstrated with protein structural data[91]. Both the short-term and long-term dataset were visualized with the EPD method. Visualization of the large formulation dataset required a systematic data processing workflow, which resulted in a comprehensive presentation, exploration, and discussion of the correlation between evaluated short-term empirical protein properties and long-term phase behavior.

5.2 Material and Methods

5.2.1 Buffer preparation

A multicomponent buffer with 10 mM buffer capacity was used to exclude buffer component effects on protein phase behavior [145]. Buffer components were CHES (Applichem, 6.13 mM), TAPS (Applichem, 14.61 mM), MOPS (Roth, 7.00 mM), sodium acetate trihydrate (Merck, 3.01 mM) and citric acid monohydrate (Merck, 13.86 mM). Buffer pH was adjusted using a 5-point calibrated pH-meter (HI-3220, Hanna Instruments, Woonsocket, RI, USA) equipped with a SenTix 62 pH electrode (Xylmen Inc., White Plains, NY, USA) using 4 M sodium hydroxide (Merck) as titrant. The pH was adjusted to 3.0, 5.0, 7.0, or 9.0 with 0.05 pH unit accuracy. Equal ionic strength between buffers with different pH values was obtained by addition of sodium chloride (Merck) or ammonium sulfate (Applichem) while stirring and monitoring the conductivity with a conductivity probe (Radiometer Analytical, Lion, France). After conductivity adjustment, the buffers were filtered over a 0.2 µm cellulose acetate filter (Sartorius, Göttingen, Germany). These buffers served as buffers with a relative ionic strength of 0 mM. Buffers with a relative ionic strength of 50, 175, 275, and 1050 mM were made with either sodium chloride or ammonium sulfate. The buffers were stored for a maximum of one month and the pH was routinely checked.

5.2.2 Protein stock preparation

A 150 g/L stock solution of lysozyme from chicken egg white (Hampton Research, Aliso Viejo, USA) was made in the appropriate 0 mM ionic strength buffer. The obtained protein solution was filtered over a 0.2 µm cellulose acetate prefilter (VWR, Radnor, PA, USA). After filtering, the protein solution was desalted with a PD-10 column (GE Healthcare Life Sciences, Uppsala, Sweden). Depending on the volume of the protein solution a mini or normal PD-10 column was used, employing the centrifugation protocol as provided by the manufacturer. Lysozyme stock solution concentration was determined with a Nanodrop 2000c UV-Vis spectrophotometer (ThermoFischer Scientific, Waltham, MA, USA). An E1% (280 nm) extinction coefficient of 22.00 $g^{-1}\cdot L\cdot cm^{-1}$ was used.

5.2.3 Sample preparation

Samples for protein analytical measurements were prepared on the same day as the experiments and measured within 6 hours. A mixing ratio of 5:1 (protein:salt) was used to obtain protein samples with a protein concentration of 125, 100, 75, 50 and 25 g/L and a relative ionic strength of 0, 50, 175 and 275 mM sodium chloride or ammonium sulfate. To obtain the desired protein concentration and ionic strength for the 5:1 mixing ratio, protein and salt stock solutions were pre-diluted with the appropriate relative 0 mM ionic strength buffer.

5.2.4 Stalagmometry

Protein surface hydrophobicity was determined by measuring solution surface tension using a fully automated liquid handling station based high-throughput stalagmometer [76]. The following changes in the setup have been made to reduce the sample volume and decrease experimental time: samples were measured with four technical replicates using a sample volume of 80 µL and repeating sample drop-wise dispense twice using low volume PTFE coated tips (Tecan, Crailsheim, Germany). Sample volume reduction and fewer repeat dispenses showed a 3.4% increase in relative standard deviation of measured water drop masses (data is shown in Supplementary Figure C1). This was considered an acceptable error as the total relative standard deviation remained below 5% while reducing sample volume a six-fold and experimental time to 3.5 hours, instead of ~9.5 hours, for a complete 96-well microtiter plate. Samples were transferred and measured in round, clear bottom 96-well microtiter plates (Greiner Bio-One GmbH, Frickenhausen, Germany). Plates were sealed with Duck Brand HD Clear sealing tape (ShurTech Brands, Avon, OH, USA) to prevent evaporation. Scalpel slits were made in the sealing tape prior to measurements. Ultrapure water, purified with a PURELAB Ultra (ELGA LabWare, Bucks, UK), was used as a reference solution with a surface tension of 72.62 mN/m[197]. Data processing and evaluation was preformed using an in-house developed MATLAB script (version R2017b, MathWorks, Natick, MA, USA). Obtained protein solution surface tensions were normalized with the corresponding buffer surface tension.

5.2.5 Dynamic light scattering (DLS)

DLS measurements were performed with a Zetasizer Nano ZSP (Malvern Instruments Ltd, Malvern, United Kingdom) using a ZEN2112 quartz cuvette (Hellma GmbH & Co. KG, Muellheim, Germany). Each sample was measured in duplicate, where each measurement contained two runs with 15 sub runs. Cuvettes were washed once with ddH2O and twice with 40 µL of the appropriate buffer before each measurement. A sample volume of 40 µL was used. Zetasizer software used the default distribution analysis to obtain a diffusion coefficient distribution from the correlogram. An in-house developed MATLAB script (version R2017b, MathWorks, Natick, MA, USA) was used to calculate the radius distribution from the obtained diffusion distribution and correct for sample viscosity[61]. The radius distributions were used to obtain the apparent hydrodynamic radius ($R_{H\ App}$) of lysozyme by extracting intensity peak between 0.5 nm and 4 nm.

5.2.6 Mixed-mode measurement, phase analysis light scattering (M3-PALS)

Electrophoretic mobility was measured with the Zetasizer Nano ZSP (Malvern Instruments Ltd, Malvern, United Kingdom). Folded capillary cells (DTS1070, Malvern Instruments Ltd, Malvern United Kingdom) were filled with the corresponding buffer and 20 µL of the samples was pipetted into the bottom of the cell with a 200 µL round 0.5 mm thick Corning Costar gel-loading pipet tip (Corning Incorporated, Corning NY, USA) to employ the diffusion barrier technique. Each sample was measured twice at 25 °C, where each measurement consisted of 120 seconds equilibration time and two runs with a maximum of 15 sub runs. The applied voltage was set to 60 mV and the automatic measurement mode was selected. A reflective index of 1.45 and absorption of 0.01 was used. The dispersant was set equal to water. An in-house developed MATLAB script (version R2017b, MathWorks, Natick, MA, USA) was used to extract the average electrophoretic mobility and calculate the average zeta potential. The average zeta potential was calculated with the corresponding sample solution viscosity[61], a dielectric constant of 78.54 and Smoluchowski's approximation of 1.5[198]. Due to the varying conductivity of the samples, only the average electrophoretic mobility was extracted. An electrophoretic mobility distribution can only be obtained in the automated mode when the sample solution conductivity is below 5 mS/cm, which was not the case for all samples.

5.2.7 Static light scattering (SLS) and intrinsic fluorescence (IF)

SLS and IF were used to determine the aggregation onset temperature (T_{Agg}) and melting temperature (T_M), respectively, using an Optim2 (Avacta Analytics, Yorkshire, UK). A temperature range from 20 °C to 90 °C with a 1 °C per minute step gradient and a temperature hold time of 60 seconds was set. The UV 266 nm and blue 473 nm laser attenuation was set to filter 4 and filter 1, respectively. The samples were loaded into a micro-cuvette array (MCA) in three-fold, with a sample volume of 9 µL. Each MCA also contained a 2 g/L lysozyme solution (dissolved in water) as a reference sample to monitor the measurement quality. The peak position of the barycentric mean of fluorescence as a function of temperature was extracted from the IF measurement. The maximum gradient of the slope, which defines $T_{M,}$ was found using an in-house developed MATLAB script (version R2017b, MathWorks, Natick, MA, USA). Light scattering counts at 473 nm as a function of temperature were obtained from SLS measurements. An in-house MATLAB script (version R2017b, MathWorks, Natick, MA, USA) was used to determine the start of the intensity gradient, which defines T_{Agg}. The start of the intensity gradient was defined based on a linear fit of at least three data points for which intensity increased consecutively between the minimum and maximum intensity and for which the first data point showed a minimum of 10% intensity increase from the starting intensity. The identified data points were fitted to a linear equation to extract the *x*-intercept (T_{Agg}) at the starting intensity.

5.2.8 Fourier transform infrared (FTIR) spectroscopy

FTIR spectroscopy was used to determine changes in protein secondary structure. A Nicolet iS5 with an iD7 ATR detector (Thermo Fischer Scientific, Waltham, MA, USA) was used. Sample and blank absorbance were scanned 150 times with a spectral resolution of 2 cm^{-1} from wavenumber 3500 to 1000 cm^{-1}. Each blank was measured once and each protein sample was measured twice, using a volume of 5 μL. A background spectrum was recorded with 254 scans. An in-house developed MATLAB script (version R2017b, MathWorks, Natick, MA, USA) was used for data evaluation. The average protein sample single beam spectrum was calculated using the duplicate single beam spectral data. The transmittance spectrum of the average protein sample and blank sample were obtained via normalization using the background single beam spectrum. The transmittance spectrum was converted to an absorbance spectrum and the blank spectrum was subtracted from the corresponding protein sample spectrum. The subtracted spectrum was vector normalized (i.e., standard normal variate normalization)[199], the second derivative was calculated and smoothed with Savitsky-Golay smoothing using a 3rd order polynomial and a window length of 33. Within the amide I range (1600 to 1700 cm^{-1}) the absolute area under 1648 ± 2 cm^{-1}, 1656 ± 2 cm^{-1}, and 1667 ± 1 cm^{-1} was extracted using the *trapz* function available in MATLAB version 2017b. These areas were selected based on a data-dependent wavenumber absorption variation analysis.

5.2.9 Empirical phase diagram construction

Each experimental protein property was normalized between zero and one. Before visualization, internal correlation between all experimental protein properties was evaluated using the Pearson correlation coefficient with a cut-off value of 0.750 and -0.750 for positive and negative internal correlation respectively. The selected experimental protein properties were used to build an EPD. The EPD construction method used is described in literature[78,91]. In brief, singular value decomposition (SVD) was used to reduce dataset dimensionality to three dimensions. The three-dimensional (3D) data was clustered to identify formulation conditions that display similar experimental protein properties. The optimal number of clusters was determined by iterating 100 times over the *evalcluster* function available in MATLAB version R2017b (MathWorks, Natick, MA, USA). For each iteration, an optimal cluster number between 5 and 10 was selected using the *k*-means cluster algorithm with a silhouette criterion based on squared Euclidean distance metric. The final cluster number was selected as the mode optimal cluster number. The 3D SVD data was clustered with the k-means clustering function (*kmeans*, available in MATLAB version 2017b), using the optimal cluster number, a maximum of 1000 iterations, and randomly chosen initial cluster centroid positions. A RGB color for each data point was calculated by normalization of (*x*,*y*,*z*)-values between zero and one. The average cluster RGB color value was defined as the mean RBG color based on each data point within the cluster. With R (version 1.0.136, using *ggplot2* and *fmsb* library) the 3D color data was visualized. The mean cluster color was plotted against all solution conditions (pH, salt, ionic strength and protein concentration). A radar plot was constructed for each cluster to represent the median value of empirical protein properties, as well as the median absolute deviation to represent distribution of empirical protein properties within the cluster.

5.3 Results and Discussion

5.3.1 Data processing

Extraction of multiple empirical protein properties may lead to internal correlation between features due to overrepresentation of a single system property within the dataset. To prevent this, internal correlation between all empirical protein properties was evaluated with the Pearson correlation coefficient. The obtained correlation coefficient matrix is shown in Supplementary Table C1. The set threshold of a Pearson correlation coefficient of 0.750 and -0.750 for positive and negative linear dependency, respectively, was not reached. Data dimension reduction and clustering was therefore preformed with all features: 3 FTIR region areas (1648 ± 2.0 cm^{-1}; random coil, 1656 ± 2.0 cm^{-1}; α-helix, 1667 ± 1.0 cm^{-1}; β-turn), apparent hydrodynamic radius of lysozyme ($R_{H\ App}$), melting temperature (T_M), aggregation onset temperature (T_{Agg}), normalized surface tension (γ_N), and mean zeta potential (ζ-potential). An overview of the used empirical protein properties, corresponding short descriptions, and value range within the dataset are shown in Table 5.1. Before these empirical protein properties could be used to cluster formulation conditions, dataset dimensionality was reduced with SVD. After data dimension reduction an energy value of 96.8% was obtained. This implies an information loss of 3.2%. This percentage of information loss falls within the general rule of thumb for SVD, where a 10% loss is considered the maximum[147]. A number of six formulation clusters was determined to be optimal with the obtained 3D data set.

Table 5.1: List of symbols and description of empirical protein properties used to compile the empirical protein property diagram, including the absolute value range.

Symbol	Min	Max	Description
$R_{H\ App}$	1.0	3.2	Apparent hydrodynamic radius of lysozyme [nm]
ζ	-14.1	9.5	Zeta potential charge [mV]
γ_N	0.99	1.21	Normalized surface tension of the protein solution [-]
T_{Agg}	20	90	Aggregation onset temperature [°C]
T_M	61	82	Melting temperature [°C]
β-turn	1.9	24.3	Area under 1666 – 1668 cm^{-1} $[AU/(cm^{-1})^2] \cdot 10^5$
α-helix	34.5	134.5	Area under 1654 – 1658 cm^{-1} $[AU/(cm^{-1})^2] \cdot 10^5$
Coil	9.9	66.3	Area under 1646 – 1650 cm^{-1} $[AU/(cm^{-1})^2] \cdot 10^5$

5.3.2 Empirical protein property diagram (EPPD)

Figure 5.1a shows the EPPD and six radar charts based on empirical protein property data. Each radar chart represents a cluster of formulations that resulted in similar empirical protein properties. The median value of each empirical protein property for each formulation cluster is represented by the radar chart using a colored surface. Value distribution within each formulation cluster is represented by the median absolute deviation (MAD), shown as a dashed line in the radar chart. Exact median and MAD values of each identified cluster can be found in Supplementary Table C2. Cluster colors and characters were used to visualize formulations in the EPPD below the radar charts. Grid columns refer to formulation pH value (pH 3.0-9.0), where the top grid row refers to formulations with ammonium sulfate ($NH_4(SO_4)_2$) and the bottom grid row refers to formulations with sodium chloride (NaCl). Individual diagrams show lysozyme concentration (25-125 g/L) on the y-axis and ionic strength (0-275 mM) of the respective salt type on the x-axis. Figure 5.1b shows previously published data that was used to construct a multidimensional protein phase diagram (MPPD)[180]. The MPPD is based on experimental long-term stability data obtained for similar formulations stored in duplicate at 20 °C for 40 days. Image-based data obtained during storage, describing aggregate dimensions and time-dependent aggregation features, was extracted and visualized with the EPD method. Table 5.2 lists the extracted image-based features, corresponding short descriptions, and the obtained value range.

Table 5.2: List of symbols and description of image features obtained from long-term microbatch experiments used to compile the multidimensional protein phase diagram, including the absolute value range. Data obtained from[180].

Symbol	Min	Max	Description
L_C	0	192	Absolute crystal size. Defined as average length of four crystals [μm]
$Đ_{L:W}$	0	3.1	Diversity in crystal shape. Defined as inter quartile range of four crystal axial ratios [-]
Δt_P	0	520	Growth time precipitate [h]
t_P	0	324	Onset time precipitate [h]
D_P	0	920	Absolute precipitation size. Defined as diameter of precipitate [μm]
Δt_C	0	959	Growth duration [h]
t_C	0	696	Onset time crystal [h]
n_C	0	100	Number of crystals. Scored between 0 and 100, where 100 is a well filled with crystals

The observed phase behavior morphology based on extracted image-based features is stated above the radar charts. Results of Figure 5.1b will only be briefly discussed in this study as the data was solely used to investigate the correlation between empirical protein properties obtained directly after formulation preparation and observed physical stability of identical formulations after 40 days of storage at 20 °C. In short, Figure 5.1b shows physically stable formulations as part of cluster I and instable formulations were identified as cluster II, III, IV, and V. Increasing supersaturation was assigned from cluster II to cluster V based on the crystal size, crystal amount, and crystal growth time. Formulations with sodium chloride showed salting-out behavior for all pH values, while formulations with ammonium sulfate showed salting-in behavior for pH 7.0 and pH 9.0.

Correlating short-term empirical protein properties to long-term physical stability via EPDs starts with the comparison of identified formulation clusters. A uniform identification of stable formulations (cluster I) during 40 days storage at 20 °C for all pH 3.0 and pH 5.0 ammonium sulfate formulations is shown in Figure 5.1b, the MPPD. Figure 5.1a, the EPPD, shows four clusters for similar formulations based on empirical protein properties, namely cluster A, B, C, and D. Identification of multiple EPPD clusters for stable formulations is due to a higher resolution and diversity of the obtained empirical data. The difference between EPPD clusters A, B, C, and D is mainly defined by $R_{H\,App}$, but also by secondary structure (α-helix region area and random coil region area), T_{Agg}, γ_N, and ζ-potential. The identification of four different formulation clusters that correspond to physical stability over time emphasizes the multidimensionality of protein phase behavior. Formulations with sodium chloride were also grouped in multiple EPPD clusters at pH 3.0

and pH 5.0, while the MPPD shows crystal formation (cluster IV) in addition to stable formulations (cluster I). At pH 3.0, the crystallized formulation with 275 mM sodium chloride and 125 g/L lysozyme was identified as EPPD cluster E. This corresponds to the identification of cluster E for other crystallized formulations. However, formulations with sodium chloride at pH 5.0 that crystallized at 125 g/L and 100 g/L lysozyme with 275 mM sodium chloride were identified as EPPD cluster A and cluster B. This is not in agreement with other formulations, where cluster A and cluster B formulations remained stable over time. These observations indicate there is at least a certain degree of positive correlation between the EPPD and MPPD for this dataset. This is also demonstrated by EPPD cluster C and cluster D formulations. Cluster C and cluster D were identified at pH 5.0 and pH 7.0 for both salt types. All ammonium sulfate and sodium chloride formulation at pH 5.0 identified as cluster C or cluster D remained stable during 40 days of storage at 20 °C, while at pH 7.0 both salt types showed physical instability for cluster C and cluster D formulations. Similar discrepancies are seen for EPPD cluster E and cluster F. Formulations at pH 9.0 for both salt types show a correlation between increasing supersaturation (MPPD transformation from cluster IV to V) and EPPD cluster E and cluster F. A comparable trend is seen for formulations at pH 7.0 with sodium chloride but not for formulations at pH 7.0 with ammonium sulfate. At pH 7.0 with ammonium sulfate, cluster E and cluster F were identified as stable formulations.

Initial MPPD and EPPD evaluation showed a partial correlation between observed phase behavior after 40 days of storage at 20 °C and the empirical protein properties measured directly after formulation preparation. To discuss this correlation as a function of formulation conditions, a stability percentage for each EPPD cluster was calculated. The stability percentage in this context is defined as the percentage of stable formulations after storage within each EPPD cluster. A decrease in stability percentage is seen from cluster A (95.3%) and cluster B (96.4%) to cluster F (8.0%). Cluster C, D and E have a stability percentage of 74.3%, 62.5% and 18.7%, respectively. Small increments of ionic strength showed EPPD cluster transformations for both salt types. For example, a cluster transition from cluster A to cluster B can be seen for formulations at pH 3.0 for increasing sodium chloride. The stability percentage for these clusters is similar, but the ionic strength increase from 50 mM to 175 mM by sodium chloride causes a lower colloidal stability (T_{Agg}), a decrease in ζ-potential, and a larger $R_{H\ App}$. Salt type dependent cluster transformations were also observed. For example, the previously mentioned transformation of cluster A to cluster B by sodium chloride at pH 3.0 is also seen for ammonium sulfate formulations at pH 3.0, but was already present between 25 mM and 50 mM ionic strength. This indicates that ionic strength increases at pH 3.0 by sodium chloride has a similar effect on lysozyme as ionic strength increase by ammonium sulfate, but the evaluated protein properties were more sensitive to the latter.

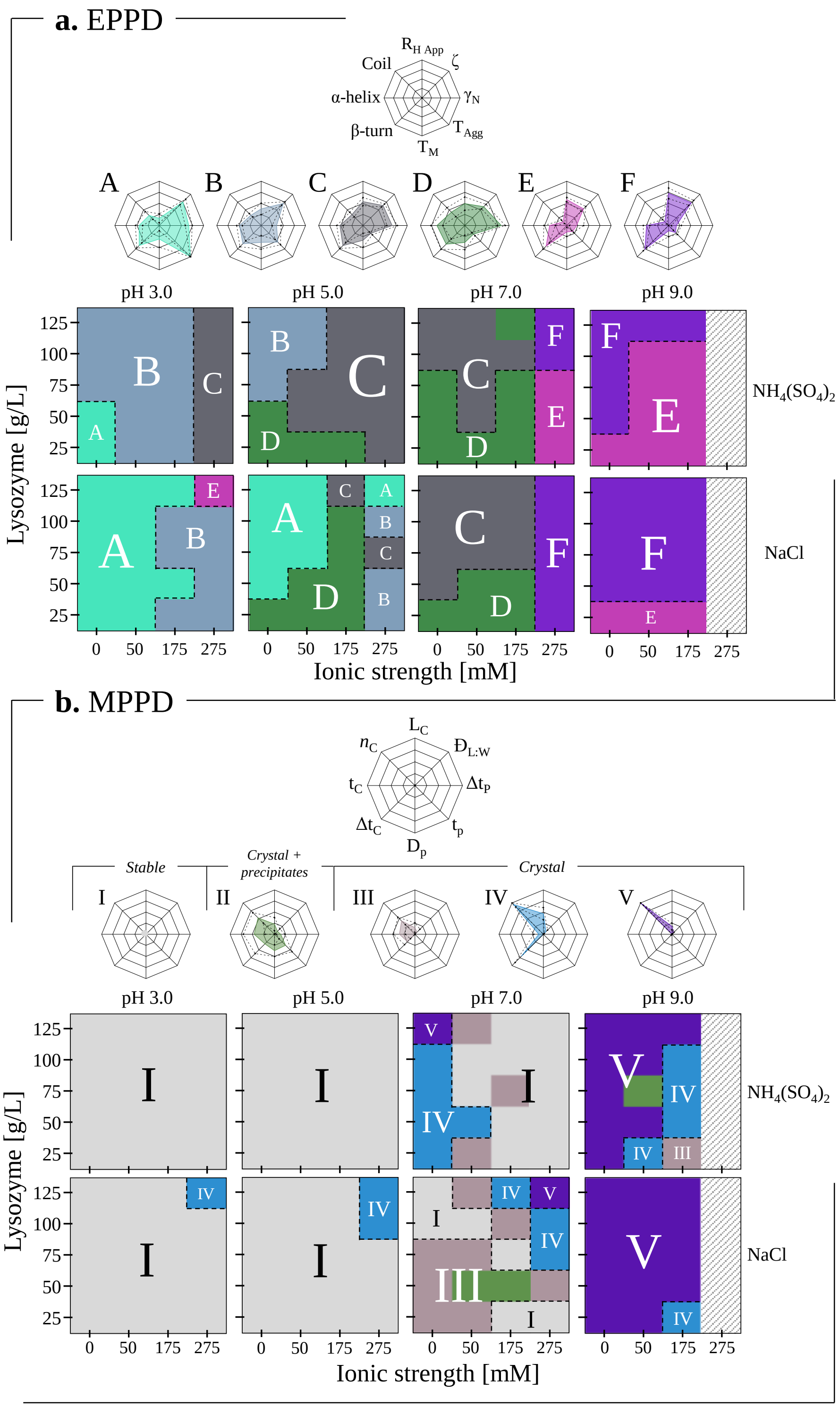


Figure 5.1: (a) Empirical protein property diagram (EPPD) and (b) multidimensional protein phase diagram (MPPD) for varying lysozyme concentrations (*y*-axis), ionic strength (*x*-axis), formulation pH value (grid column), and salt type (grid row). The MPPD is adjusted from data presented in[180]. Clusters are indicated with a cluster color and character within each diagram. Dashed lines are used to guide the eye between adjacent clusters within the diagrams. A legend radar chart is given to indicate the position of the properties compiled in the EPPD and MPPD radar charts. The colored surface of the radar charts shows the normalized median value for each property within the cluster and the dotted line indicates the median absolute deviation for each property.

Cluster transformations for increasing lysozyme concentration were identified as well. Formulations at pH 7.0 in the presence of 275 mM ammonium sulfate showed such a transformation, where cluster E transforms in cluster F for increasing lysozyme concentration. This illustrates how a higher protein concentration can influence properties such as secondary structure. Cluster transformations are visible for each separate grid diagram but formulation pH dominates cluster transformations. This is indicated by the identification of similar EPPD clusters for each salt type for each formulation pH value, which is a result of the relatively large pH range used in this study.

Growing supersaturation zones were identified for increasing formulation pH value, as shown in the MPPD by increased identification of cluster V. Supersaturation increase was attributed to the loss of positive charge, as the formulation pH moves towards lysozyme's isoelectric point (pI) of ~11.3[148]. Formulation pH affects amino acids residue protonation and can thereby diminishing the repulsive electrostatic forces closer to the protein's pI [26,34]. This is in accordance with the decreasing stability percentage seen in the EPPD clusters, shown by the transformation from cluster A to cluster F. This trend was independent of the salt type. Lysozyme's $R_{H\ App}$ also showed a trend from formulation cluster A to cluster F, as seen for stability percentage, where it increased from 1.4 ± 0.2 nm (cluster A) to 2.6 ± 0.3 nm (cluster F). Note that the mentioned empirical protein property values are median ± MAD, which represents the central tendency and distribution within each formulation cluster. Lysozyme's $R_{H\ App}$ dependence on changing formulation pH is in agreement with literature, where a $R_{H\ App}$ increase above pH 6.0 was measured using dielectric spectroscopy[200,201]. It was stated that lysozyme's $R_{H\ App}$ increased from approximately 1.8 nm at pH 4 to approximately 2.6 nm at pH 10, which was interpreted as an index of aggregation due to loss of positive charge along lysozyme's surface as the formulation pH shifts towards its pI[200]. Loss of electrostatic repulsive forces may be identified upon further inspection of $R_{H\ App}$ as a function of lysozyme concentration using the data from individual formulations[202]. In Supplementary Figure C2, four examples are shown where a shift from repulsive to attractive protein-protein interactions as a function of pH and ionic strength can be observed. This indicates that an increasing $R_{H\ App}$ resulted from diminishing electrostatic repulsive forces which allows for attractive protein-protein interactions. This is also reflected by a decrease in colloidal stability, represented by T_{Agg}, from cluster A (88.5 ± 0.9 °C) to cluster F (35.2 ± 4.6 °C). Nevertheless, a similar stability percentage was found for cluster A and cluster B formulations. This shows that the corresponding formulations remained physically stable during the long-term experiment despite short-term observed changes in $R_{H\ App}$ (from 1.4 ± 0.2 nm to 1.8 ± 0.3 nm, respectively) and T_{Agg} (from 88.5 ± 0.9 °C to 56.7 ± 5.9 °C, respectively). A further increase of $R_{H\ App}$ (2.2 ± 0.2

nm) and corresponding decrease in T_{Agg} (37.4 ± 3.2 °C) for cluster C formulations resulted in decreased stability percentage (74.3%). This indicates a threshold for the short-term parameters which results in critically low colloidal stability causing an increased aggregation propensity.

The loss of electrostatic repulsion is also partially confirmed by the measured surface charge. In the EPPD, lysozyme surface charge is represented by the ζ-potential. The median ζ-potential decreased from cluster A (3.8 ± 2.1 mV) to cluster E (-1.5 ± 1.8 mV), but cluster F formulations showed a ζ-potential of 3.1 ± 1.9 mV while displaying the largest $R_{H\ App}$ (2.6 ± 0.3 nm). Lysozyme's secondary structure can be used to obtain a more detailed understanding of $R_{H\ App}$ and ζ-potential. FTIR peak regions were used in this study to represent secondary structural changes. The FTIR peak regions have been empirically assigned to secondary motifs, where a decrease in FTIR spectral region area indicates a loss of the corresponding secondary structure [64,203]. The EPPD shows similar median values for FTIR region areas in formulation cluster A, B, and C. This suggests that there was no secondary structural change due to corresponding changes in formulation conditions. Cluster D formulations showed secondary structure changes and a further decrease in stability percentage (62.5%). Secondary structure changes were quantified by a maximum median random coil region of 34 ± 8.0 AU/(cm^{-1})^2 and a maximum median α-helix region area of 97.3 ± 13.6 AU/(cm^{-1})^2. These structural properties were accompanied by the dataset maximum γ_N, with a median value of 1.17 ± 0.02, a $R_{H\ App}$ of 2.1 ± 0.3 nm, and a ζ-potential of 0.2 ± 1.7 mV. A γ_N of 1.17 reflects a 17% increase of surface tension after adding lysozyme to the formulation in question. Presumably, surface tension increased as a result of water molecules entering the altered lysozyme structure [204,205]. Cluster D formulations also resulted in a low colloidal stability, represented by a median T_{Agg} of 39.0 ± 3.5 °C, but a conformational stability (T_M of 68.9 ± 3.5 C) comparable to cluster A, B, and C. The obtained median T_M for cluster A, B, C, and D formulations (~67 °C) is relatively lower compared to reported literature values (approximately between 68 – 82 °C) [206–208]. This may be due to the use of a multicomponent buffer system [208]. In addition, it should be noted that for cluster C to cluster F the obtained T_{Agg} lies ~30 °C below T_M. This may indicate that thermal aggregation mainly consists of native lysozyme structures, or that aggregation already starts when only small amount of unfolded structures was present [209].

In contrast to cluster D, cluster E formulations resulted in a dataset minimum for the median random coil region area (16.4 ± 3.0 AU/(cm^{-1})^2) and median α-helix area (73.1 ± 12.8 AU/(cm^{-1})^2). Secondary structure loss presumably contributed to the relatively low median T_{Agg} (33.1 ± 1.9 °C) obtained for cluster E formulations. Next to low colloidal stability, secondary structure changes may also be the reason for the observed γ_N decrease (median value of 1.04 ± 0.03). A decrease in surface tension due to protein unfolding was previously observed for whey protein isolate dispersions. The observed decrease in surface tension was attributed to an increase in structure flexibility that led to an increased air-water interface adsorption[210]. This suggests that the observed γ_N decrease may represent (partial) structural unfolding under cluster E formulation conditions. Cluster F formulations resulted in a dataset maximum for the median β-turn region area (18.9 ± 1.1 $AU/(cm^{-1})^2$), while all other clusters show a comparable median β-turn region area (~16 $AU/(cm^{-1})^2$). The median random coil FTIR peak region obtained in cluster F formulations was comparable to cluster E, while the median α-helix FTIR peak region was comparable to cluster A, B, and C. Compared to cluster E, cluster F displayed a similar γ_N (1.04 ± 0.02), T_M (63.2 ± 2.6 °C), and T_{Agg} (35.2 ± 4.6 °C). Considering the combination of secondary structure changes and corresponding effects on γ_N, colloidal stability, and conformation stability, it is assumed that lysozyme aggregated directly after formulation preparation under the corresponding conditions.

The evaluated formulation conditions resulted in different protein structural conformations. Formulations part of cluster A, B, and C did not show secondary structural change, based on the similarities in FTIR peak region areas. A stability percentage decrease was observed from cluster A to cluster C, which was presumably a result of decreased repulsive electrostatic forces causing increased protein-protein interaction. Cluster D formulations resulted in a secondary structure change which showed similar conformational stability as the structures found for cluster A, B, and C. On the contrary, cluster E and F resulted in (partially) unfolded structures, which was reflected by secondary structure changes and surface tension decrease. Structural unfolding is regarded as the main reason for the stability percentage drop of cluster E and cluster F (<19%) compared to the other EPPD clusters (>62%). Coupling back to the correlation between the EPPD and the MPPD, supersaturation was identified in the MPPD by a large amount of fast growing, small crystals after 40 days of storage. Crystal growth cessation is proposed to be an effect of improper implementation of lysozyme molecules or surface poisoning[155]. With EPPD data showing lysozyme's unfolding for supersaturated formulations, surface poisoning by incorporation of unfolded molecules is considered the reason for crystal growth cessation.

In general, it is expected that near the protein's pI solubility increases for low ionic strength and decreases for higher ionic strength [211]. However, sodium chloride formulations show solely increasing physical instability for increasing ionic strength and increasing formulation pH. This phase behavior has been reported by other studies as well[150,152,212]. The lack of salting-in behavior for lysozyme formulations with sodium chloride was attributed to screening of electrostatic repulsive forces[150]. The EPPD depicts that sodium chloride formulations at pH 9.0 showed no influence of ionic strength, represented by the uniform identification of cluster E and cluster F for different ionic strengths. This observation is also in accordance with work by Retailleau et al., where it was hypothesized that chloride ion adsorption caused a pI shift to approximately pH 9.5[150]. A recent molecular dynamics (MD) simulation study attributed an apparent unfolded state to dominant interactions with the sodium ion compared to interactions with the chloride ion[213]. Despite the lack of direct evidence of specific protein-ion interactions in the EPPD, observations obtained from the reported MD simulations correspond to changes in the secondary structure for increasing sodium chloride ionic strength seen in the EPPD.

Interactions at the basis of salting-in behavior, as observed for ammonium sulfate formulations, remains speculative as well. Salting-in effects primarily occur because of interactions between the protein and salt ions[153]. Ammonium sulfate formulations show lysozyme unfolding for pH 7.0 at higher ionic strength which resulted in physical stability over time, while formulations at lower ionic strength containing more stable structures showed crystallization. Salting-in behavior observed for ammonium sulfate formulations at pH 7.0 corresponds to work regarding the effect of anions on salting-in and salting-out behavior, where an inverse Hofmeister series was identified for positively charged lysozyme under relatively low ionic strength (<300 mM) conditions[42]. The underlying anion interactions are dependent on the positive protein charge, and therefore it can be assumed that salting-in becomes less pronounced towards lysozyme’s pI. This may cause the observed decrease in physical stability for ammonium sulfate formulations for increasing pH value, as seen when moving from pH 5.0 to pH 7.0 and from pH 7.0 to pH 9.0 for similar ionic strength. For ammonium sulfate formulations at pH 9.0, structural unfolding is seen for all formulations but it is more pronounced for lower ionic strength. Decreased supersaturation at pH 9.0 formulations for increasing ionic strength by ammonium sulfate may be due to ammonium ions adsorption to hydrophobic amino acid side-chains. This was demonstrated to cause aggregation deceleration in a thermal unfolding study of lysozyme [192].

The use of the empirical phase diagram method as visualization technique resulted in a comprehensive overview of multidimensional data, which allowed for an uncomplicated investigation of the correlation between long-term phase behavior and short-term empirical protein properties. The transition from long-term stable formulations to long-term instable formulations was partially represented, but short-term empirical protein properties were unable to fully capture the observed phase behavior. Inclusion of additional empirical properties could lead to a closer match between EPPD formulation clusters and MPPD phase behavior clusters. Possible additional empirical properties to extend the multidimensional dataset could include quantification of protein-protein interaction strength (e.g., rheological data[122,187,214]), identification of multimeric protein species, or quantification of protein-salt ion interactions. Based on the presented dataset it is evident that there is not a single straightforward combination of analytical techniques to evaluate long-term physical stability using short-term analytics. The required combination of analytical techniques cannot be determined beforehand without prior knowledge of the aggregation pathway. An advantage of the data evaluation workflow presented in this study is its applicability for other proteins, different formulation conditions, and various analytical techniques. Further investigation of different experimental designs in combination with the presented data evaluation workflow can expand understanding of underlying cause and consequence regarding long-term protein phase behavior, which is required to move towards knowledge-based formulation screenings and phase behavior prediction.

5.4 Conclusion

This study employed the EPD method to correlate long-term physical stability of 150 unique lysozyme formulations to empirical protein properties that can be determined right after formulation preparation. This was done to overcome commonly seen protein phase behavior evaluation limitations involving dataset completeness and size, sample adjustment prior to analysis, and subsequent data visualization and interpretation. For this purpose, six different analytical techniques were used to determine the effects of 150 unique formulation conditions on the lysozyme's colloidal and conformational stability, secondary structure, as well as surface charge and hydrophobicity. The EPD method allowed for the representation of both the long-term and the short-term dataset in a single figure. This resulted in a systematic and comprehensive visualization and interpretation of MPPD data in relation to EPPD data for all 150 unique formulations. A correlation between short-term and long-term data was found based on increased formulation cluster supersaturation and decreased formulation cluster stability percentage. The decrease of long-term storage stability was found mainly a result of loss of repulsive electrostatic interaction and loss of secondary structure. It was shown that changing formulation conditions leaves a different fingerprint in terms of structural properties, colloidal stability, conformation stability, and surface properties. Both physically stable and instable EPPD clusters showed varying protein property sets, which emphasizes the multidimensionality of protein properties determining protein phase behavior. Biopharmaceutical formulation screening studies can benefit from the presented multidimensional data evaluation workflow as it allows for a comprehensive overview and uncomplicated interpretation of large datasets. There are no limitations regarding screening targets, analytical techniques, or conditions that can be evaluated with this method. For protein phase behavior studies, trends observed throughout multidimensional datasets can provide detailed insight, but it can also provide targets for phase behavior optimization through combined information on aggregation kinetics and empirical protein properties. Detailed insight and optimization targets can guide knowledge-based formulation screening design and aid short-term predictive parameters development for long-term protein phase behavior.

5.5 Acknowledgements

This work was financially supported by the BE-Basic Foundation (www.be-basic.org), under project FS2.003. The authors would like to thank academic and industrial partners for scientific discussions during the development of this work. The authors also appreciate the support of Jana Zabel and Susanna Suhm during the experimental work

.

6

Redesigning food protein formulations with empirical phase diagrams: A case study on glycerol-poor and glycerol-free formulations

Marieke E. Klijn[1] and Jürgen Hubuch[1]

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

Abstract

Redesigning existing food protein formulations is necessary in situations where food authorities propose dose adjustments or removal of currently employed additives. Redesigning formulations involves evaluating substitute additives to obtain similar long-term physical stability as the original formulation. Such formulation screening experiments benefit from comprehensive data visualization, understanding the effects of substitute additives on long-term physical stability, and identification of short-term optimization targets. This work employs empirical phase diagrams to reach these benefits by combining multidimensional long-term protein physical stability data with short-term empirical protein properties. A case study was performed where multidimensional protein phase diagrams (1152 formulations) allowed for identification of stabilizing effects as a result of pH, methionine, sugars, salt, and minimized glycerol content. Corresponding empirical protein property diagrams (144 formulations) resulted in the identification of normalized surface tension as a short-term empirical protein property to reach long-term physical stability presumably similar to the original product, namely via preferential hydration. Additionally, changes in pH and salt were identified as environmental optimization targets to reach stability via repulsive electrostatic forces. This case study shows the applicability of the empirical phase diagram method to rationally perform formulation redesign screenings, while simultaneously expanding knowledge on protein long-term physical stability.

6.1 Introduction

Food additives are used to enhance and ensure product quality aspects such as sensory, microbial, enzymatic, and long-term physical stability[215,216]. The safety of food additives for these purposes is assessed by the European Food Safety Authority (EFSA). Re-evaluations occur when changes, such as new scientific findings, come to light[217]. In the period from 2013 to 2018, over 130 reports were produced by the EFSA concerning re-evaluation of food additives[218]. Re-evaluation mainly results in confirmation of the existing regulations[219–221] or dose adjustments[222–224]. If re-evaluation results in a regulatory change such as a dose adjustment, it may be necessary to (partially) remove the food additive in question from the current formulation. When an existing product formulation needs to be redesigned, the new formulation should meet product quality aspects similar to the original formulation. An example of a recent dose adjustment is glycerol[223]. The re-evaluation report states that side effects, such as headaches and nausea, can be induced from a dose of 125 mg/kg body weight per hour. It was noted that this dose is easily reached in infants and toddlers upon consumption of a 330 mL flavored drink. Glycerol is a food additive whose properties include, but are not limited to, enhancing microbial and long-term physical stability[225–227]. Complete, or even partial removal of glycerol from an existing product formulation requires the addition of one or more substitute additives to obtain a similar microbial and long-term physical stability compared to the original formulation.

This case study investigates glycerol-poor (75 g/L instead of 1050 g/L) and glycerol-free protein formulations with respect to long-term physical protein stability. In this case study, long-term physical instability is defined by the formation of visible, insoluble aggregates, such as crystals or amorphous precipitates. Protein-protein, protein-solvent, and protein-additive interactions determine long-term physical stability[228–230]. These interactions are governed by protein properties such as protein structure, surface charge, conformational stability, and colloidal stability[21,191,231]. In turn, these protein properties are influenced by formulation additives, environmental conditions, and the protein itself[104,229,232]. Glycerol, like other polyols, is known to enhance a protein's conformational stability via preferential hydration[189,227,233]. Preferential hydration is used to describe the depletion of additives, such as glycerol, from the protein surface. The exclusion of glycerol is thermodynamically unfavorable, as it increases the protein's chemical potential. This leads to a protein surface area minimization in order to reduce unfavorable interactions between the protein and the solvent. Thus, the native folded protein state is thermodynamically favorable over unfolded state, which results in a higher conformational stability[227]. In addition to preferential hydration, it has been demonstrated that in some cases glycerol can decrease attractive interactions between protein molecules via preferential interaction with hydrophobic patches on the protein surface[234,235]. Preferential hydration of protein molecules can also be caused by sugars[48]. It has been shown that a variety of sugars can increase protein conformational stability[236–238]. Sugar effects are dependent on the sugar type (e.g., monosaccharide or disaccharide)[48,236] and sugar concentration[236,238]. For these reasons, two monosaccharides (fructose, glucose) and two disaccharides (lactose and sucrose) were considered in this case study as glycerol substitutes at three different concentrations (30, 60, and 80 g/L).

Long-term physical stability can decrease via protein oxidation[239,240]. Non-site-specific oxidation, such as oxidation by the presence of oxidants, is dependent on the exposure of amino acids to the environment. Oxidation can be limited or prevented by addition of an antioxidant[240]. For example, free methionine molecules can act as sacrificial agent that will be oxidized instead of the product[241,242]. In the presented case study, approximately 15% of the total solvent accessible amino acids residues of the investigated protein are made up from amino acids prone to oxidation, namely cysteine, methionine, tryptophan, histidine, and tyrosine[240]. To investigate the effectiveness of methionine to improve long-term physical stability in glycerol-poor and glycerol-free formulations, methionine was tested at two different concentrations (1.45 g/L and 9.50 g/L) in this case study. The methionine concentration of 1.45 g/L was used as it is comparable to the concentration of the original formulation. To determine the potential beneficial effect of an increased methionine concentration, a formulation with 9.50 g/L methionine was evaluated.

Other environmental conditions, such as formulation pH and salt, can affect physical stability as well[34,229]. Formulation pH determines protein charge, which plays an important role in physical stability as it influences both conformational stability and colloidal stability[21,191]. For this case study it was chosen to stay above the protein isoelectric point (pI) and relatively close to the pH of the original formulation. In total three pH values (pH 5.0, 5.5, and 6.0) were included in the formulation search space. Salts can influence long-term physical stability via various mechanisms such as preferential exclusion, preferential interaction with the protein surface, and screening of repulsive electrostatic forces[239]. The mechanism at action depends on the formulation conditions, salt type, and salt concentration[211,243,244]. Therefore, this case study included two salts (sodium chloride and potassium chloride) at four different concentration ratios (100:0, 60:37, 40:55, 0:90 g/L, sodium chloride to potassium chloride, respectively).

As mentioned before, glycerol is also used to increase microbial stability. Sodium lactate can be used as a substitute for glycerol to enhance the microbial stability aspect[245]. Even though sodium lactate is typically used to ensure microbial stability, but not protein stability, it was still added to the formulation search space in this case study. This was done because a substitute additive that enhances one aspect of formulation stability, such as microbial stability, may adversely influence another, such as protein stability[246,247].

The number of considered substitutes for glycerol-poor and glycerol-free formulations illustrates the multidimensionality and magnitude of such screening experiments. This results in a considerable experimental workload and a large amount of data which complicates data evaluation. To reduce the experimental workload, a high-throughput storage setup was employed to monitor long-term physical stability via automated imaging, where formulations were stored for 30 days at 20 °C[115]. This experimental setup results in a multidimensional output by combining final physical stability data after storage and aggregation kinetics during storage [180]. Formulation conditions that lead to long-term physical instability can be found with the final physical stability data, while the aggregation kinetics allows for the identification of potential formulation optimization targets by quantifying the degree of instability as a function of the applied conditions. Processing and evaluation of such multidimensional data was facilitated by employing the empirical phase diagram (EPD) method[78,89] to construct a multidimensional protein phase diagram (MPPD) [180].

Faster, smaller, and efficient screening methods can reduce experimental workload, but the necessary storage time remains equally long. The desire for accelerated screenings resulted in the search for long-term physical stability predictors from short-term empirical protein properties[82,121,193,195]. Screenings for new formulations can also benefit from short-term predictors as these can minimize time and efforts to reach original formulation quality, as well as provide insight on the responsible interactions. Therefore, short-term empirical properties (apparent hydrodynamic radius of the protein, mean apparent hydrodynamic radius of high weight species, protein surface hydrophobicity, conformational stability, and colloidal stability) were experimentally determined directly after formulation preparation to investigate the correlation between the original formulation and new formulations, as well as the corresponding long-term physical stability. The obtained multi-source empirical protein property dataset encounters issues concerning data evaluation as well. These issues were resolved with a systematic and data-dependent workflow, which also employs the EPD method for visualization. This approach combines all multi-source data into one single empirical protein property diagram (EPPD)[248].

This case study presents an MPPD screening dataset including 1152 formulations to investigate the influence of 4 sugars at 3 different concentrations, 3 pH values, 2 salts at 4 different ratios, 2 methionine concentrations, and sodium lactate, on long-term protein stability in glycerol-poor and glycerol-free formulations. A corresponding EPPD was constructed with a subset of 144 formulations. This subset of formulations was selected based on physical stability transitions in the MPPD. The aim of this case study was to identify new formulations which have similar long-term stability compared to the original formulation. It was investigated whether short-term empirical properties are similar between long-term stable redesigned formulations and the original formulation, as it is assumed that similar short-term properties lead to similar long-term behavior. An additional aim was to use the MPPD and EPPD approach to identify short-term empirical properties that are not similar to the original formulation, but still display similar physical stability after long-term storage. This was done to create a better understanding of underlying interactions that determine long-term physical stability. Thus, this study combined multidimensional long-term physical stability data with multi-source empirical protein property data to rationally approach screening for new formulations by means of a case study with glycerol-free and glycerol-poor formulations.

6.2 Material and Methods

6.2.1 Buffer preparation

Buffer solution with a pH value of 5.0, 5.5, and 6.0 were made with 0.1 mol/L citric acid (Merck KGaA, Darmstadt, DE) and 0.1 mol/L sodium acetate (Merck KGaA, Darmstadt, DE) solutions. Buffer pH was adjusted with a 0.05 pH unit accuracy using a five-point calibrated pH-meter (HI-3220, Hanna Instruments, Woonsocket, RI) equipped with a SenTix 62 pH electrode (Xylmen Inc., White Plains, NY, USA) using either 0.1 mol/L citric acid or 0.1 mol/L sodium acetate as titrant.

For each pH, 1.45 g/L and 9.50 g/L L-methionine (Alfa Aesar, Haverhill, MA, USA) buffer solutions were made. This corresponds to roughly 0.1% (w/w) and 1.0% (w/w) of L-methionine, where % (w/w) refers to the weight percentage of the respective compound per total weight of the formulation. For each pH and methionine concentration, a buffer was made with 100 g/L sodium lactate (Sigma-Aldrich, Saint Louis, MO, USA), corresponding to roughly 8.50% (w/w).

For each of the 12 obtained buffer solutions (that is 3 pH values, with 1.45 g/L or 9.50 g/L, and with or without 100 g/L sodium lactate) four different sugars were added with a concentration of 30 g/L, 60 g/L, and 80 g/L (corresponding to roughly 2% (w/w), 4% (w/w), and 6% (w/w), respectively). The employed sugars were D-fructose (Merck KGaA, Darmstadt, DE), D(+)-glucose monohydrate (Merck KGaA, Darmstadt, DE), sucrose (Thermo Fisher Scientific, Waltham, MA, USA) and lactose monohydrate (Merck KGaA, Darmstadt, DE). In addition, four different mixtures of sodium chloride (NaCl) (Merck KGaA, Darmstadt, DE) and potassium chloride (KCl) (Alfa Aesar, Haverhill, MA, USA) were made. Mixing ratios of NaCl:KCl were as following: (1) 100 g/L:0 g/L, (2) 60 g/L:37 g/L, (3) 40 g/L:50 g/L, (4) 0 g/L:90 g/L. The ratios correspond to roughly (1) 7.5:0.0, (2) 5.0:2.5, (3) 2.5:5.0, and (4) 0.0:7.5% (w/w). The maximum salt and sugar concentrations were determined based on the solubility limit. The systematic dilution was chosen to capture the effect of different salt and sugar concentrations on long-term physical protein stability. All the obtained mixtures were also prepared with the additions of 75 g/L glycerol (Merck KGaA, Darmstadt, DE), which corresponds to roughly 5% (w/w). Prior to use, all buffer solutions were filtered over a 0.2 μm cellulose acetate filter (Sartorius, Göttingen, Germany) and the pH value was checked, and adjusted when necessary.

6.2.2 Protein solution preparation

Freeze-dried protein, an enzyme (30-45 kDa) referred to as protein I, was kindly provided by DSM Biotechnology Center (Delft, NL). For protein phase diagram preparation, a 12 g/L protein I solution was prepared for each pH value, in combination with each L-methionine concentration and each sodium lactate concentration. For analytical measurements a concentration of 3 g/L protein I was prepared in the identical solution as mentioned for the protein phase diagram. Lyophilized protein I was dissolved and filtered using a 13 mm 0.2 μm Supor® pre-syringe filter (Pall Corporation, New York, NY, USA). The resulting concentration of protein I was determined with a Nanodrop 2000c UV-Vis spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA).

6.2.3 Long-term storage

Protein crystallization plates were prepared with a Tecan liquid handling (LiHa) station (Tecan, Maennedorf, CH). Sugar and salt stock solutions were prepared in 1.5 mL Eppendorf tubes (Eppendorf, Hamburg, DE) and mixed with the LiHa to obtain appropriate mixing ratios. Protein stock of 12 g/L was mixed 1:1 with the appropriate buffer, with and without 150 g/L glycerol. Salt and sugar solutions were mixed with the diluted protein solutions using a 1:1 mixing ratio to obtain 3 g/L protein I in a final volume of 24 μL in a 96-well crystallization plate (Swissci, Neuheum, CH). The long-term storage experiment was carried out according to the method developed in our lab[115] with the following adaptation: protein solutions were stored in duplicate for 30 days at 20°C in a Rock Imager 54 (Formulatrix, Bedford, MA, USA). The employed imaging method was similar to previous work[180]. Protein phase behavior was evaluated after storage based on aggregation surface coverage in the well, aggregation onset time, and aggregation growth time. In this case study, aggregation could be in the form of crystals or precipitates. For crystalline aggregation the crystal length and crystal width were extracted. The image-based features where processed with MATLAB (version 2017b, Mathworks, Natick, MA, USA) and served as data for the multidimensional protein phase diagram construction. All further operations performed with MATLAB mentioned in this work also used version 2017b and in-house developed scripts.

6.2.4 Short-term empirical protein properties

Five short-term empirical protein properties (apparent hydrodynamic radius of protein I and high weight species, aggregation onset temperature, melting temperature, and normalized surface tension) were obtained with four different analytical techniques (dynamic light scattering, static light scattering, intrinsic fluorescence, stalagmometry). All applied operational settings and corresponding data analysis have been described elsewhere[248], unless stated otherwise.

Dynamic light scattering was used to determine the apparent hydrodynamic radius of protein I and high weight species, using a Zetasizer Nano ZSP (Malvern Instruments Ltd, Malvern, United Kingdom). The intensity peak between 1 nm and 10 nm was used to extract the apparent hydrodynamic radius of protein I. The mean apparent hydrodynamic radius of high weight species was based on the mean intensity peaks above 10 nm. Before extraction of hydrodynamic radii, all radius distributions were corrected with the corresponding bulk viscosity. Buffer viscosity was determined with duplicate samples (200 µL sample volume) at 25 °C using a density sensor (Integrated Sensing Systems, Inc., Ypsilanti, MI, USA) and pure water as a reference viscosity value. Viscosity was determined for all buffers at pH 6.0 with 1.45 g/L methionine. The viscosity of buffers with a methionine concentration of 9.50 g/L was not determined, as it was established that buffers containing 9.50 g/L methionine returned similar viscosity values compared to buffers containing 1.45 g/L methionine (data not shown). Obtained viscosity values for all buffers at pH 6.0 with 1.45 g/L methionine can be found in Supplementary Table D1.

Static light scattering and intrinsic fluorescence were used to determine the aggregation onset temperature and melting temperature, using an OPTIM2 (Avacta Analytics, Yorkshire, UK). These short-term empirical protein properties were used to represent colloidal and conformational stability, respectively. Device settings and data extraction protocols were similar to those presented in previous work[248].

Normalized surface tension was determined with a fully automated LiHa station-based high-throughput stalagmometer[74]. This short-term empirical protein property was used to represent apparent surface hydrophobicity. The experimental procedure was carried out as presented in previous work[248] but with a sample volume of 120 µL.

6.2.5 Data handling

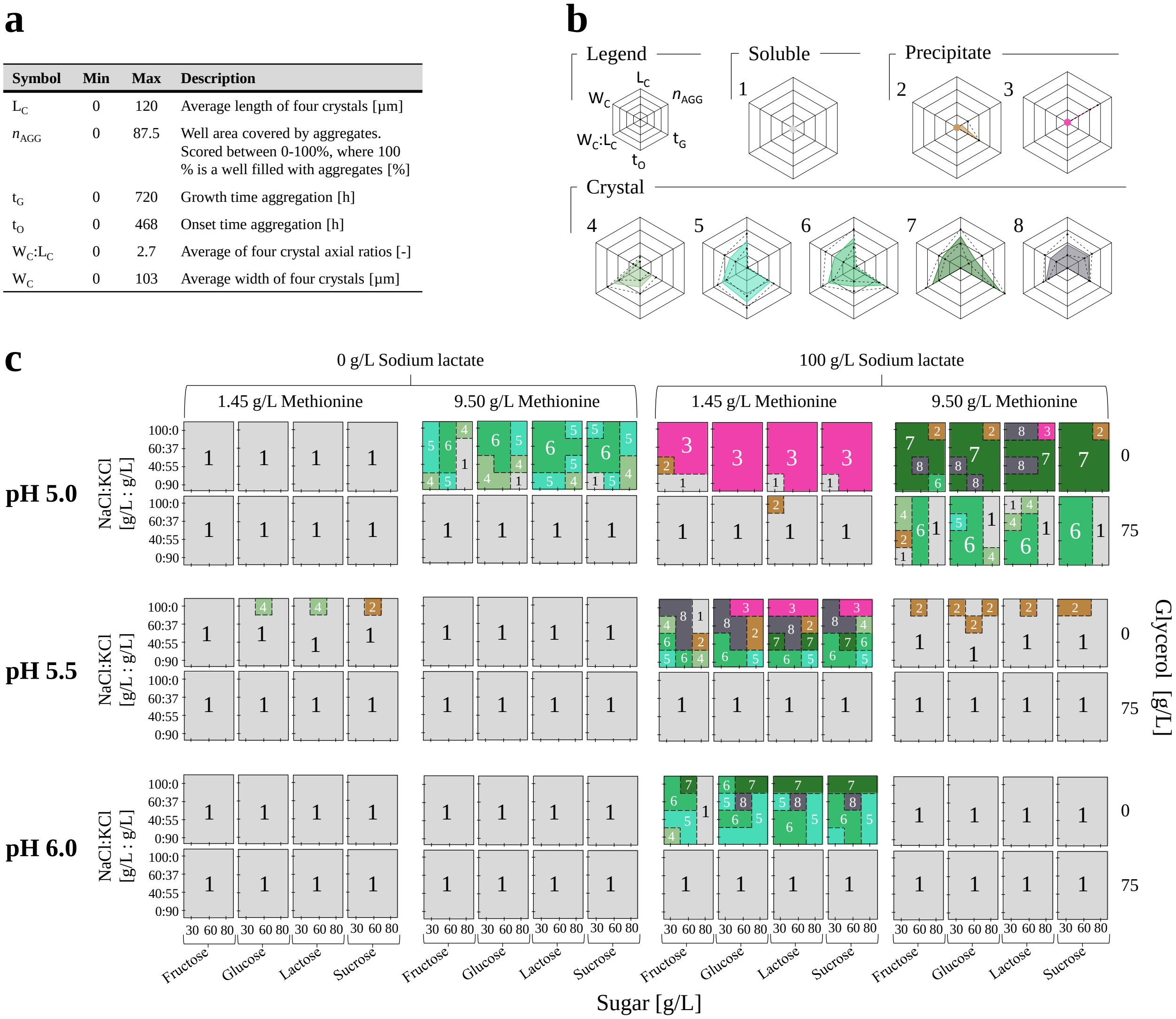
Data (pre)processing and visualization was performed with MATLAB and R (version 1.0.136), respectively. The multidimensional protein phase diagram was constructed as described in previous work[180], selecting the optimal number of clusters between five and eight clusters. The empirical protein property diagram was constructed as described in[248], selecting the optimal cluster number between three and ten clusters.

6.3 Results and Discussion

6.3.1 Multidimensional protein phase diagram (MPPD)

Figure 6.1 depicts the extracted image features for all 1152 conditions monitored during 30 days of storage at 20 °C. The absolute value range for each image feature and corresponding description is shown in Figure 6.1a. Data processing resulted in a 3D dataset with an energy value of 95.5%. This means there was a 4.5% information loss due to singular value decomposition (SVD) dimension reduction. This information loss is considered acceptable as it falls within the general rule, where 10% information loss is considered the maximum[147]. Based on this 3D dataset, the optimal number of formulation clusters was determined to be eight. For each of the eight clusters, image feature median values were calculated and depicted as a colored surface in the corresponding radar charts in Figure 6.1b. The median absolute deviation (MAD) of the image features was calculated for each cluster as well, shown as a dotted line in the radar charts in Figure 6.1b. A representative image for each cluster can be found in Supplementary Figure D1 and an overview of median ± MAD image feature values per MPPD cluster can be found in Supplementary Table D2.

The MPPD allows for identification of different supersaturation zones based on aggregation kinetics and aggregation dimension data, as was shown in a previous study using hen egg-white lysozyme[180]. Cluster 1 represents undersaturated formulations that showed long-term physical stability, as all extracted image features (shown in Figure 6.1b) are equal to zero when no precipitates or crystals are observed. Cluster 2 and 3 correspond to relatively high supersaturation, reflected by precipitate formation. This was identified by a crystal size (L_C) and crystal width (W_C) equal to zero in Figure 6.1b. Both cluster 2 and 3 represent formulations that precipitated directly after preparation, indicated by an onset time (t_O) of 0 h.



Symbol	Min	Max	Description
L_C	0	120	Average length of four crystals [µm]
n_{AGG}	0	87.5	Well area covered by aggregates. Scored between 0-100%, where 100 % is a well filled with aggregates [%]
t_G	0	720	Growth time aggregation [h]
t_O	0	468	Onset time aggregation [h]
W_C:L_C	0	2.7	Average of four crystal axial ratios [-]
W_C	0	103	Average width of four crystals [µm]

Figure 6.1: (a) Symbols, description, and absolute value range of image features. (b) Cluster radar charts with a legend to indicate the position of image features. The normalized median value of each image feature is represented with a colored surface. The dotted line represents the median absolute deviation within each cluster for each image feature. (c) Multidimensional protein phase diagram (MPPD) for changing sodium lactate content (major grid columns), pH values (major grid rows), methionine concentrations (minor grid columns), glycerol-poor or glycerol-free conditions (minor grid rows), sodium chloride (NaCl) to potassium chloride (KCl) ratios (y-axis), and sugar types and concentration (x-axis). The eight identified clusters are visualized in the MPPD using the mean cluster color and cluster number similar to the radar charts in (b). MPPD cluster regions are highlighted with a dashed line to guide the eye.

Lower supersaturation was found for cluster 2 compared to cluster 3, based on the amount of aggregation (n_{AGG} of 9.3 ± 9.3 % versus 52.5 ± 7.4 %, respectively) and growth time (t_G of 361 ± 2 h versus 2 ± 1 h, respectively). Note that all mentioned cluster values consist of the median ± MAD for representation of the distribution within a formulation cluster. Differences in supersaturation can also be identified for crystallized formulations, which were found for cluster 4, 5, 6, 7, and 8. A combination of increased nucleation rate (i.e., a larger n_{AGG}), earlier crystal onset time (t_O) and decreased growth time (t_G) indicates increasing supersaturation. These properties were identified for MPPD cluster 8, where 41.2 ± 7.4 % of the well was covered by relatively small crystals which formed at 0 h and grew for 361 ± 10 h. Contrarily, MPPD cluster 4 shows a relatively low nucleation rate (0.5 ± 0 %) and little crystal growth (onset at 174 ± 62 h and growth of 198 ± 62 h), which indicates low supersaturation. Based on the clusters found in the MPPD, the order of crystal clusters regarding level of supersaturation is proposed to be the following: 8>7>6>5>4, where cluster 8 represents the highest supersaturation level. Transformations to lower supersaturation identify optimization targets to reach physical stability with minor formulation adjustments. For example, glycerol-free formulations at pH 5.0 with 9.50 g/L methionine and 0 g/L sodium lactate show a transformation to lower supersaturation (MPPD cluster 6 to 5 to 4) for increasing KCl concentration. This indicates that similar formulations with a higher KCl concentration or lower ionic strength are likely remain physically stable over time. Another example of such a cluster transformation is seen for increasing formulation pH. Glycerol-free formulations at pH 5.5, with 1.45 g/L methionine and 100 g/L sodium lactate, were mostly identified as high supersaturation formulations (cluster 8 and 6). Similar formulations at pH 6.0 showed a shift towards MPPD cluster 5 identification. This indicates that a higher formulation pH is likely to result in long-term physical stability, even though long-term physical stability was not observed in the evaluated formulation condition range.

6.3.1.1 Stability percentage

Solutions depicted in Figure 6.1c present 77% physically stable new formulations as a result of different formulation conditions. To discuss the effects of each formulation variable on long-term physical stability, a percentage was calculated per formulation variable representing formulations that remained physically stable during long-term storage (i.e., formulations part of MPPD cluster 1). This stability percentage is listed in Table 6.1. This percentage was calculated with the number of formulations that showed an MPPD cluster transformation upon changing the respective variable. This means that the listed percentage is not a percentage of all 1152 formulations, but a percentage of formulations that were affected by the respective variable. For example, 96 formulations with 9.50 g/L methionine at pH 6.0 remained physically stable independent of the addition of sodium lactate. This means sodium lactate did not affect these 96 formulations. Similar formulations, but with 1.45 g/L methionine, showed cluster transitions for 44 out of 96 formulations. The formulations that showed a cluster transformation were considered affected by sodium lactate. Formulation selection took into account any type of cluster transformation as an effect of the respective formulation variable. For example, this selection resulted in a total of 220 formulations that were affected by sodium lactate. An overview of percentages for all MPPD clusters per formulation variable and the number of formulations taken into account per formulation variable can be found in Supplementary Table D3.

Table 6.1: List of stability percentages per formulation variable. The stability percentage is calculated based on the number of formulations that showed a cluster transformation upon changing the respective variable in the multidimensional protein phase diagram.

Variable	Unit	Value	Stable [%]
Sodium lactate	[g/L]	0	79
		100	0
Glycerol	[g/L]	0	0
		75	86
pH	-	5.0	5
		5.5	68
		6.0	75
Methionine	[g/L]	1.45	37
		9.50	40
Salt ratio	NaCl:KCl [g/L:g/L]	100:0	6
		60:37	20
		40:55	20
		0:90	31
Sugar type	-	Fructose	37
		Glucose	14
		Lactose	19
		Sucrose	20
Sugar concentration	[g/L]	30	19
		60	10
		80	38

6.3.1.2 Glycerol

Glycerol showed the largest physical stability increase, reflected by a stability percentage of 86% upon addition of 75 g/L glycerol listed in Table 6.1. This indicates that partial removal of glycerol (75 g/L instead of 1050 g/L) still resulted in long-term physical stability. Protein aggregation inhibition upon glycerol addition corresponds to previous research[235,249,250]. This may be an effect of preferential hydration resulting in higher conformational stability[227] or a reduction in attractive protein-protein interactions due to preferential interaction with hydrophobic patches on the protein surface[234,235].

6.3.1.3 pH

Increasing the formulation pH value resulted in an increase of physical stability as well, represented by stability percentages of 5% (pH 5.0), 68% (pH 5.5), and 75% (pH 6.0). Increased physical stability of formulations for pH values further away from the pI of protein I correlates well with general protein aggregation theory. Electrostatic repulsive forces become stronger as a protein becomes more charged which results in higher colloidal stability [21].

6.3.1.4 Sodium lactate

Figure 6.1 shows that addition of 100 g/L sodium lactate resulted in 0% stable formulations. It has been observed that sodium lactate decreases protein conformational stability via surface tension decrease due to formation of lactoyl lactic acid[251]. However, the concentration of 100 g/L sodium lactate (~8.5% (w/w)) which was used in this case study is lower than the range for which this effect was observed (>20% (w/v)). According to the same study, concentrations below the range for which lactoyl lactic was formed, were found to increase conformational stability. This observation does not correspond to data published in another study, where millimolar addition of lactic acid, which is present when sodium lactate is dissolved in an aqueous solution, resulted in protein denaturation [252]. Lactic acid was also investigated for its effectiveness of whey protein gelation, a process that requires protein denaturation[253,254]. Despite the divergent information of sodium lactate effects on protein denaturation, our results, as presented in Figure 6.1c, confirm long-term physical instability upon sodium lactate addition. Based on the previously mentioned protein denaturation studies, this effect is assumed to be a result of a destabilized conformation of protein I. This would also support the stabilizing effect of glycerol, as glycerol can increase conformational stability via preferential hydration.

6.3.1.5 Methionine

A 3% stability percentage difference between the two tested methionine concentrations reflects a weak influence of an increase in methionine concentration on long-term physical stability. However, this number is does not fully represent the trends seen in Figure 6.1c as increased long-term physical stability is shown for both concentrations. For pH 5.5 and pH 6.0, an increase in long-term physical stability is seen upon addition of 9.50 g/L methionine, while for pH 5.0 formulations a decrease was observed for the same methionine concentration. Increased long-term physical stability for increasing methionine concentration at pH 5.5 and pH 6.0 was expected. Free methionine can act as sacrificial agent and has shown to increase physical stability[255–257]. This contradicts the decrease in physical stability for pH 5.0 formulations. Decreasing methionine oxidation rate can be excluded because methionine oxidation rates were found stable between pH 2 and pH 8[258].

It has been demonstrated that solvent accessibility can control oxidation rates, but this would not influence free methionine[258,259]. It has been stated that antioxidant additives can also influence local dynamics and solvent accessibility of reactive groups, which may lead to instability[23]. However, to the best of our knowledge, no prior studies have described the loss of physical stability for increasing methionine concentration under comparable conditions. For other amino acids which are often used to enhance long-term stability, such as arginine, different optimal concentrations and destabilizing effects have been reported for two IgG1 monoclonal antibodies[260]. Nevertheless, the nature, use, and mechanism of stabilization of arginine is not comparable to methionine[261]. At pH 5.0, the formulation pH is closest to the pI of protein I. It can be speculated that methionine influences the solubility or conformational stability independent of its antioxidant effects, due to a close to neutral protein net charge. It can be speculated that methionine influences the colloidal or conformation stability, independently of its antioxidant effects, due to a close to neutral protein net charge. A neutral protein charge allows for enhanced protein-protein interaction due to the reduction of electrostatic repulsive forces. Protein-protein interactions may be further promoted by methionine as a result of preferential exclusion from the protein surface. This is just speculative as the molecular mechanism behind the observed decrease in long-term physical stability for higher methionine concentrations at pH 5.0 remains unknown based on the data obtained in this case study.

6.3.1.6 Salt

Different salt compositions were also evaluated in the formulation search space. The highest physical stability percentage (31%) was found for formulations with 90 g/L potassium. Mixed salt ratios (60:37 and 40:55 g/L) of sodium chloride and potassium chloride resulted in a physical stability percentage of 20% and formulations with 100 g/L NaCl resulted in a physical stability percentage of 6%. The effect of salt on physical stability is often related to its position in the Hofmeister series[262]. Under the evaluated formulations, it is assumed that protein I carries a net negative charge as the solution pH lies above its pI. A negative net protein charged combined with salt concentrations >1.0 mol/L means a direct Hofmeister series applies[263]. Salts used in this case study differ in cation (potassium and sodium) while the anion (chloride) remains constant. The direct Hofmeister series shows that potassium is slightly more kosmotrope. This indicates potassium is slightly more prone to promote conformational stability and salting-out compared to sodium. However, this is not reflected by a higher long-term stability percentage for potassium (31%) compared to sodium (6%). It should be noted that the Hofmeister series positions of potassium and sodium are relatively close, which means other factors influencing the observed stability difference should be considered as well. Based on the used salt concentrations it can be calculated that the employed ratios resulted in different formulation ionic strength. Gram per liter ratios, ordered from high to low NaCl content, correspond to approximately 1.7 mol/L, 1.5 mol/L, 1.4 mol/L, and 1.2 mol/L ionic strength. Increasing ionic strength is known to increase aggregation propensity by screening electrostatic repulsion, as it lowers a protein's colloidal stability[239]. Screening effects support the decrease in stability percentage for increasing formulation ionic strength. In addition, such screening effects diminish close to the pI, which supports lack of influence of the salt ratio on physical stability seen in Figure 6.1c for glycerol-free formulations at pH 5.0 with sodium lactate.

6.3.1.7 Sugar

Evaluating the influence of sugar type resulted in the highest stability percentage for fructose (37%) and lowest for glucose (14%), while lactose and sucrose resulted in a similar stability percentage (~20%). Sugar concentrations resulted in a minimum stability percentage for 60 g/L (10%), and the highest stability percentage was found for 80 g/L (38%). As previously stated, sugars can increase conformational stability due to preferential hydration of protein molecules[48], where disaccharides were found more effective[236]. In this work, the higher effectiveness of disaccharides cannot be confirmed due to the use of similar concentrations of monosaccharides (fructose and glucose) and disaccharides (lactose and sucrose). From a molar perspective, roughly twice as many fructose and glucose (0.16 mol/L and 0.44 mol/L, respectively) molecules are present in the formulations compared to lactose and sucrose (0.08 mol/L and 0.23 mol/L, respectively). The relatively high stability percentage for fructose concurs with other research, where fructose was compared to glucose and sucrose[236]. Increasing stability percentage for increasing sugar concentration was also in accordance with previous research, where it was reported that a minimum amount of sugar is needed to induce preferential hydration effects[238].

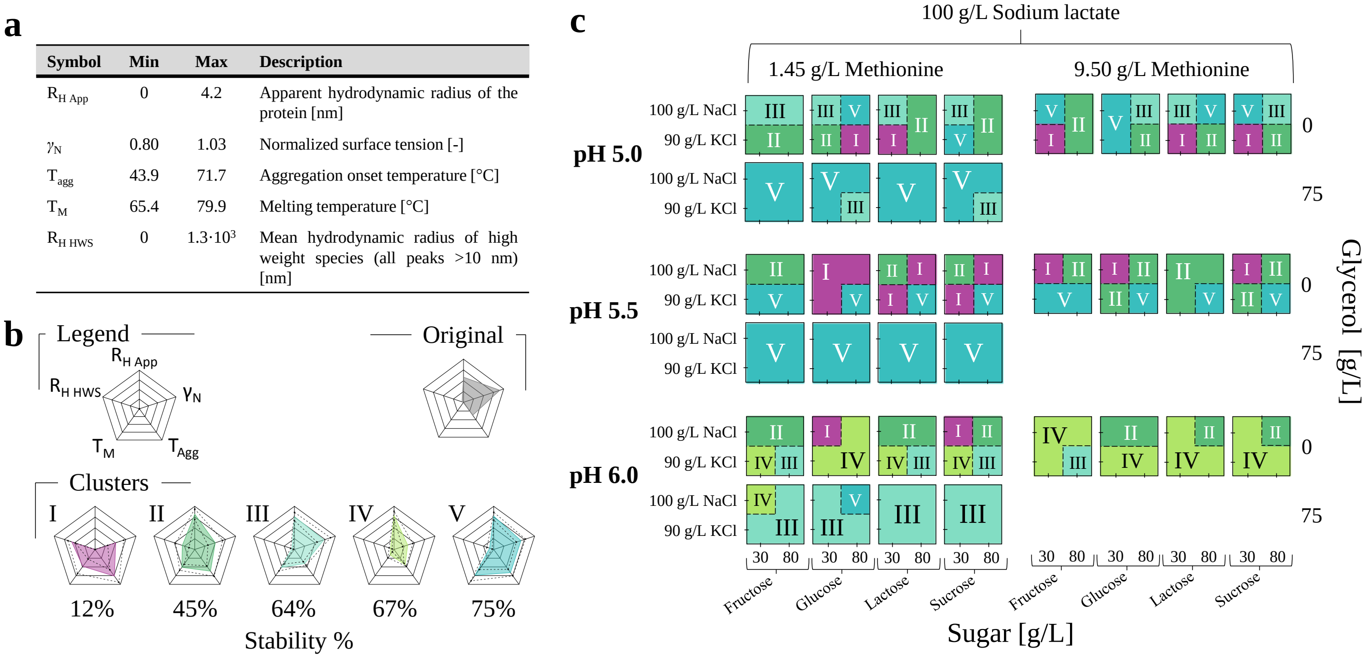
6.3.2 Empirical protein property diagram

Moving towards faster identification and possibly prediction of long-term physically stable formulations requires short-term predictive parameters which can be obtained right after formulation preparation. In this study, the applicability of such short-term empirical protein properties for the use of screening for new formulations was investigated from two perspectives. The first perspective is based on the original formulation, which is known to remain physically stable during long-term storage. Therefore, it is desired for glycerol-poor and glycerol-free formulations to display similar short-term empirical protein properties as the original formulation, because similar properties are thought to result in similar long-term physical stability. The second perspective relies on the multidimensionality of physical stability. Different mechanisms may lead to long-term physical stability, which can be a result of different protein properties or different combinations of these properties.

This means that glycerol-poor and glycerol-free formulations displaying different empirical properties than the original formulation may still show long-term physical stability. Multiple analytical techniques were used to generate information on the apparent hydrodynamic radius of protein I ($R_{H\ App}$), the mean apparent hydrodynamic radius of high weight species ($R_{H\ HWS}$), normalized surface tension (γ_N), melting temperature (T_M) and aggregation temperature (T_{Agg}) for a subset of 144 formulations right after formulation preparation.

The selected formulations contained 100 g/L sodium lactate for all three pH values. Fructose, glucose, lactose, and sucrose were tested at 30 g/L and 80 g/L in combination with either 100 g/L sodium chloride or 90 g/L potassium chloride. These formulations were tested glycerol-free in combination with 1.45 g/L and 9.50 g/L methionine. Glycerol-poor formulations were also tested, but only in combination with 1.45 g/L methionine. Figure 6.2a lists the empirical properties, a short description, and the corresponding value range. Data dimension reduction of empirical protein property data resulted in a 3.5% information loss. Clustering of the 3D dataset resulted in an optimum of five clusters.

Figure 6.2b shows a legend, five EPPD cluster radar charts (Roman numerals I to V), and a radar chart displaying empirical protein properties obtained for the original formulation. The radar charts representing cluster I to V depict the median value of each empirical property as a colored surface and the empirical property cluster MAD as a dotted line. An overview of median ± MAD values for each empirical property per EPPD cluster can be found in Supplementary Table D4. Radar charts in Figure 6.2b can be used to identify glycerol-free and glycerol-poor formulations that show empirical properties similar to the original formulation. In addition, a stability percentage for each formulation cluster is shown below the corresponding radar chart. The stability percentage is defined as the formulation percentage within each respective EPPD cluster that showed long-term physical stability (i.e., formulations that were also part of MPPD cluster 1). For example, cluster II has a stability percentage of 45% which means that 45% of the formulations part of cluster II remained physically stable during the long-term storage experiment. Further elucidation of the MPPD cluster content of each EPPD cluster can be found in Supplementary Table D5. Visualization of the clusters for each considered formulation variable is shown in Figure 6.2c, the EPPD.



a

Symbol	Min	Max	Description
$R_{H\ App}$	0	4.2	Apparent hydrodynamic radius of the protein [nm]
γ_N	0.80	1.03	Normalized surface tension [-]
T_{agg}	43.9	71.7	Aggregation onset temperature [°C]
T_M	65.4	79.9	Melting temperature [°C]
$R_{H\ HWS}$	0	$1.3{\cdot}10^3$	Mean hydrodynamic radius of high weight species (all peaks >10 nm) [nm]

Figure 6.2: (a) Symbols, description, and absolute value range of empirical protein properties (EPPs). (b) Cluster radar charts with a legend to indicate the position of each EPP. The normalized median value of each EPP is represented with a colored surface. The dotted line represents the median absolute deviation within each cluster for each EPP. The EPPs obtained for the original formulation are shown in the *Orignal* radar chart. (c) Empirical protein property diagram (EPPD) for 100 g/L sodium lactate (major grid column), varying pH values (major grid rows), different methionine concentrations (minor grid columns), glycerol-poor or glycerol-free conditions (minor grid rows), 100 g/L sodium chloride (NaCl) or 90 g/L potassium chloride (KCl) (y-axis), and different sugar types at 30 g/L or 80 g/L (x-axis). The five identified clusters are visualized in the EPPD using the mean cluster color and cluster Roman numeral similar to the radar charts in (b). EPPD cluster regions are highlighted with a dashed line to guide the eye.

The EPPD shows a near uniform identification of cluster V for glycerol-poor formulations at pH 5.0 and pH 5.5, while at pH 6.0 glycerol-poor formulations show an almost uniform identification of cluster III. Both EPPD clusters show a relatively high stability percentage (V = 75% and III = 65%), indicating that the measured empirical protein properties relate to long-term physical stability. Cluster III and cluster V show a relatively high value for γ_N (0.97 ± 0.04 and 0.95 ± 0.03, respectively) and a similar $R_{H\ App}$ of ~3.2 nm. The $R_{H\ App}$ of the original formulation is smaller (2.4 nm) but γ_N is comparable (1.01). In this case study, $R_{H\ App}$ was assumed to differ between new formulations and the original formulation as a result of inter-particle interactions because the measurements were corrected for the bulk viscosity[61,264]. A relative smaller $R_{H\ App}$ for the original formulation, compared to new formulations, was attributed to interaction between glycerol and the protein surface[234,235]. The higher glycerol content in the original formulation was assumed to result in a greater loss of attractive protein-protein interactions. The γ_N is comparable between cluster III, cluster V, and the original formulation, where a γ_N of ~1.00 indicates there is no effect on the formulation surface tension upon addition of protein I. As previously mentioned, glycerol influences physical stability via preferential hydration or preferential interaction with hydrophobic patches on the protein surface. Both effects would minimize the apparent protein surface hydrophobicity and thereby minimize the surface tension change upon addition of protein I[74,265]. Some glycerol-free formulations have also been identified as being part of cluster III or V, which means these formulations result in similar short-term empirical protein properties without the addition of glycerol. Such glycerol-free formulation can be found at pH 5.5 with 90 g/L potassium chloride, where each formulation with 80 g/L sugar was identified as cluster V.

Cluster V and cluster III showed comparable values for y_N and $R_{H\ App}$, but differences were observed for T_M (76.1 ± 2.3 °C to 72.9 ± 1.4 °C, respectively) and T_{Agg} (62.3 ± 2.6 °C to 55.1 ± 1.5 °C, respectively). The decrease in conformational and colloidal stability was caused by an increased formulation pH. This effect can also be observed for glycerol-free formulations, represented by cluster IV at pH 6.0. Cluster IV is characterized by the lowest T_M (67.7 ± 1.2 °C) and T_{Agg} (56.1 ± 1.1 °C), when compared to all other EPPD clusters. Nevertheless, T_M and T_{Agg} of cluster IV were still comparable to the original cluster (T_M = 65.5 °C and T_{Agg} = 54.2 °C), which means that colloidal and conformational stability were comparable. Different from the original formulation, cluster IV formulations show a relatively low γ_N of 0.88 ± 0.03. The relatively low γ_N was considered a consequence of glycerol's absence and confirms that apparent protein surface hydrophobicity was minimized upon addition of glycerol in glycerol-poor formulations. Despite the differences in γ_N and T_M, cluster IV and cluster III have comparable stability percentages (67% and 65%, respectively). Cluster III's long-term physical stability was attributed to either the prevention of denaturation and subsequent aggregation via preferential hydration or reduction of attractive protein-protein interaction by hydrophobic interaction between the protein surface and glycerol, while cluster IV formulations are assumed to remain physically stable due to repulsive electrostatic forces obtained by the increased positive net charge of protein I[239]. These repulsive forces diminished for increasing ionic strength (from 90 g/L potassium chloride to 100 g/L sodium chloride) and for lower formulation pH values. This is reflected by the increased identification of cluster I and cluster II for these formulation condition changes. Cluster II and cluster I show a decreasing stability percentage of 45% and 6%, respectively. Based on the large $R_{H\ HWS}$ (796 ± 120 nm) and lack of $R_{H\ App}$, it can be concluded that cluster I formulation conditions caused immediate aggregation. Cluster II formulations resulted in a $R_{H\ App}$ of 3.4 ± 0.5 nm and a $R_{H\ HWS}$ of 396 ± 213 nm. This indicates that aggregation was present immediately after formulation preparation, as seen in cluster I formulations, but to a lesser extent. The effect of methionine can be observed by cluster I and cluster II as well. For glycerol-free formulations at pH 5.5, the addition of methionine causes a decrease in cluster I identification frequency. This indicates that aggregation tendency decreases for a higher methionine concentration. A similar cluster transformation can be observed for glycerol-free formulations at pH 6.0. At pH 5.0 a decrease in long-term physical stability was observed for 9.50 g/L methionine in the MPPD. A corresponding increase in aggregation propensity (i.e., a transformation from cluster II to cluster I) is not unambiguously reflected by the EPPD at pH 5.0 between the two methionine concentrations.

Cluster V and III represent glycerol-poor and glycerol-free formulations that showed a γ_N comparable to the original formulation, in combination with a stability percentage of 75% and 65%. This indicates that screening new formulations for a γ_N close to 1.00 should result in physical long-term stability which is reached via the same mechanism as the original formulation. However, none of these new formulations displayed a similar $R_{H\ App}$ value. Presumably, the decrease of inter-particle attraction or increased conformational stability, as seen for the original formulation, could not be reached with the relatively low glycerol and sugar concentrations used in this case study. Formulation adjustments to optimize and match this property of the original formulation might result in higher stability percentages. Glycerol-free formulations part of cluster IV at pH 6.0 showed a relatively high stability percentage as well (67%). These formulations displayed a lower γ_N, but T_M and T_{Agg} values comparable to the original formulation. Based on the corresponding formulation conditions, it was assumed that long-term physical stability for cluster IV is obtained via repulsive electrostatic forces. Further characterization of new formulations by including additional short-term analytical techniques might confirm these results and resolve open questions about the observed effects on long-term physical stability as a function of the evaluated formulation conditions. Such analytical techniques could include quantification of the secondary structure, strength of protein-protein interactions, or surface charge measurements. Nevertheless, this case study demonstrates the application of short-term empirical data to rationally screen for new formulations based on short-term empirical data of a stable original formulation. The use of MPPDs and EPPDs for such an application was not shown before. In addition, this case study underlines the ability to obtain an increased understanding of observed long-term physical stability due to correlation of MPPD data to EPPD data. This allowed for the identification of possible alternative long-term protein stabilization routes compared to the original formulation, as well as environmental conditions that can be used as a target for optimization of the system under investigation.

6.4 Conclusion

The presented work applied a combination of multidimensional long-term physical stability data (1152 formulations) and multi-source short-term empirical protein property data (144 formulations) to redesign a protein food formulation containing 1050 g/L glycerol. The empirical phase diagram method was applied to present and analyze the multidimensional data. The obtained results were employed to identify redesigned formulations that resulted in similar long-term stability, but with minimized glycerol content. Long-term stability of redesigned formulations was found for two instances. In the first instance, redesigned formulations showed a similar short-term normalized surface tension compared to the original formulations. This short-term property was found for all glycerol-poor (75 g/L) formulations, and for glycerol-free formulations at pH 5.5 and pH 6.0, containing 90 g/L potassium chloride and 80 g/L sugar. This was observed for all sugar types. The comparable short-term empirical property profile and corresponding long-term stability of these redesigned formulations indicated a similar stabilization pathway as the original formulation. For the second instance, glycerol-poor formulations at pH 6.0 in combination with 90 g/L potassium chloride showed a relatively high long-term stability percentage (67%) as well. This was attributed to the increasing net protein charge as a result of a formulation pH value further away from its isoelectric point, thereby inducing repulsive electrostatic forces. This indicated that the combination of the MPPD and EPPD identified a different stabilization pathway compared to the original formulation.

The case study illustrated the potential of the multidimensional data visualization and analysis methods to rationally design screening experiments for new formulations of an existing product. In addition, straightforward identification of underlying short-term empirical protein properties provided a more detailed insight to long-term protein physical stability. In a broader perspective, the applied method can also be used to screen formulations for other product quality aspects that are sensitive to formulation additives, such as enzymatic stability studies. Further development of multidimensional data analysis might lead to knowledge-based long-term screening experiments, which can include predictive cluster classification models using short-term empirical protein properties as input.

6.5 Acknowledgements

This work was financially supported by the BE-Basic Foundation (www.be-basic.org), under project FS2.003. The authors want to thank academic and industrial partners for scientific discussions during the development of this work. Additionally, the authors would like to thank Birgit Roser, Kristina Schleining, and Susanna Suhm for their contribution to the experimental work.

7

High-throughput computational pipeline for 3-D structure preparation and in silico protein surface property screening: A case study on HBcAg dimer structures

Marieke E. Klijn[1†], Philipp Vormittag[1†], Nicolai Bluthardt[1], and Jürgen Hubuch[1]

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

[†] Contributed equally

Abstract

Knowledge-based experimental design can aid biopharmaceutical high-throughput screening (HTS) experiments needed to identify critical manufacturability parameters. Prior knowledge can be obtained via computational methods such as protein property extraction from 3-D protein structures. This study presents a high-throughput 3-D structure preparation and refinement pipeline that supports structure screenings with an automated and data-dependent workflow. As a case study, three chimeric virus-like particle (VLP) building blocks, hepatitis B core antigen (HBcAg) dimers, were constructed. Molecular dynamics (MD) refinement quality, speed, stability, and correlation to zeta potential data was evaluated using different MD simulation settings. Settings included 2 force fields (YASARA2 and AMBER03) and 2 pKa computation methods (YASARA and H++). MD simulations contained a data-dependent termination via identification of a 2 ns Window of Stability, which was also used for robust descriptor extraction. MD simulation with YASARA2, independent of pKa computation method, was found to be most stable and computationally efficient. These settings resulted in a fast refinement (6.6 – 37.5 hours), a good structure quality (-1.17 - -1.13) and a strong linear dependence between dimer surface charge and complete chimeric HBcAg VLP zeta potential. These results indicate the computational pipeline's applicability for early-stage candidate assessment and design optimization of HTS manufacturability or formulability experiments.

7.1 Introduction

Virus-like particles (VLPs) are macromolecular assemblages, which in their simplest form consist of multiple copies of one viral structural protein[266]. Their particulate and highly repetitive structure invokes an immune response similar to that of native viruses, but VLPs are incapable of reproduction as viral nucleic acids are lacking[266,267]. VLPs can therefore provide immunization against the virus they were derived from, as was done for hepatitis B virus (HBV; Engerix B, Recombivax)[268] and human papilloma virus (HPV: Cervarix, Gardasil)[269]. Immunization unrelated to the native virus can be achieved with chimeric VLPs, which are VLPs containing a foreign antigenic epitope. These antigenic epitopes can be inserted into a capsid forming protein at either the N-terminus, C-terminus, or major immunodominant region (MIR)[270]. This insertion aims to trigger an immune response, adjuvanted by the particulate and repetitive VLP structure[271]. Chimeric VLPs are increasingly used in preclinical and clinical studies[272]. An example of a chimeric VLP that received positive opinion of the European Medical Agency is a malaria vaccine based on a HBV surface antigen VLP with an inserted segment of the *Plasmodium falciparum* circumsporozoite protein[273]. Another platform for chimeric antigen display is the HBV core antigen (HBcAg) protein. Chimeric HBcAg VLPs with foreign and self-epitopes have been shown to induce strong B cell responses, a characteristic that can be used to develop VLPs for the treatment of cancer[274–276].

Chimeric VLP development involves screening large numbers of candidate epitope insertions[277]. During screenings, chimeric VLPs are evaluated based on immunogenicity, structure stability, and assembly-competence[267,278]. For example, fewer than 50% of inserted peptides in the HBcAg platform resulted in a properly assembled and soluble VLP [279]. Structural stability and solubility are not only desired in the final formulation to ensure product efficacy, quality, and safety, but also throughout downstream processing to ensure manufacturability[278,280,281]. During manufacturing, VLPs are exposed to different environmental conditions such as changes in pH, ionic strength, and temperature. These conditions influence physicochemical properties of VLPs, which in turn determine critical evaluation parameters such as the structural stability and assembly-competence[282]. High-throughput screening (HTS) experiments allow for workload reduction in virus and VLP studies to determine optimal processing[283,284] and formulation[285] parameters. HTS design for VLP studies can be further optimized by search space minimization and manufacturability assessment using prior knowledge of physicochemical properties obtained computationally from 3-D protein structures[278,286,287]. Physicochemical properties that are most important for virus particles include electrostatic surface charge[288–290]. Research on bacteriophage MS2 showed correlation between experimentally determined virus surface charge using zeta potential measurements and computationally calculated protein charge[291]. Moreover, experimentally determined protein zeta potential showed stronger correlation with calculated protein charge using only capsid surface atoms compared to protein charge calculated using all MS2 capsid atoms. Other research showed that calculated protein charge using the surface of a single MS2 capsid protein was in agreement with theoretically determined protein surface electrostatic potential of the entire MS2 capsid[292]. Ionizable groups of a protein determine protein properties such as surface charge, structure, and stability[293]. Therefore, both 3-D structure preparation and in silico determination of surface charge require an estimation of the pKa of titratable groups. Fast and fairly accurate pKa estimation methods have been developed, such as methods to monitor pKa shifts during an MD simulation[294] or to process a large number of structures parallelized in a short time[295].

Candidate chimeric VLP 3-D structures have to be available for computational physicochemical property extraction. As it would be impractical to produce all candidates and experimentally determine their 3-D structures, an in silico 3-D structure preparation approach is needed. This approach would require an automated and high-throughput framework to support screening a large number of chimeric VLPs to minimize manual effort. These requirements can be met with homology modeling, also known as comparative modeling. With this method, unknown 3-D protein structures are created based on known template structures[296]. Homology modeling can be performed using several approaches[297,298], but all resulting 3-D structures remain only an estimation of reality. Further model refinement is needed to meet structure quality requirements and should therefore include a molecular dynamics (MD) simulation step[99]. Structure refinement requires the selection of a force field. The choice depends on the application and it can be notoriously difficult to identify the best-performing force field for a particular application. Novel self-parameterizing knowledge-based force fields, such as YASARA2, have been developed to improve the calculation of torsional angles and have shown to be useful and accurate for the physical correction of proteins by energy minimization[299]. Several authors have analyzed the performance of different open-source force fields by comparing in silico structural data to NMR experimental data[300–302]. In general, modern force fields perform reasonably accurate and reproducible for MD simulation of proteins[303].

For VLPs, in silico experiments have most frequently been applied to study capsid stability using complete VLP capsid 3-D structures. All-atom MD simulations of complete capsids are as challenging as they are computationally expensive and can only be done using relatively short in silico timescales. Reported simulations reach <10 ns per day on super-computers[304–306] or 30 ns/day when using constrained bond-lengths[307]. However, modeling

VLP structural transitions (e.g., self-assembly, capsid disintegration) requires a much larger timescale (µs or ms)[308]. Compared to all-atom MD simulations, computational expense has been reduced to reach these relatively large timescales using coarse-grained[309–311] or multi-scale[312–316] models in various capsid studies. Supercomputers, such as the Blue Waters supercomputer with 128000 cores, were used and a simulation duration of several days for a single capsid was reported[306]. In silico candidate screening would require an equal amount of simulations as available chimeric VLP candidates. Depending on the application, this could involve screening of hundreds of chimeric VLP candidates. In this case, simulation time would increase to a timespan of a year, even with the use of a supercomputer. Time requirement, super computer availability, and respective expertise hamper the implementation of these methods in computational high-throughput candidate screenings. Simulation simplification, by using only a single capsid protein or capsid building block models[317,318], aids in resolving these limitations. Monomers and pentamers were compared to an entire VLP 3-D capsid model to evaluate the applicability to immunogenicity prediction[314]. Joshi and coworkers showed that the immunogenicity predictor (epitope flexibility) was dependent on the complete capsid construct and thus a complete VLP capsid 3-D model was required to capture this effect. This requirement is not expected for the evaluation of surface charge as it has been shown that MS2 capsid protein surface charge descriptors have a high correlation to experimental zeta potential data of the entire structure[291,292]. In addition, this case study used chimeric HBcAg structures that differ only in the epitope located on the outer VLP surface. Therefore, the influence of dimer contact area on possible zeta potential changes observed for entire chimeric HBcAg VLP structures was considered to be minimal. Thus, surface charge after 3-D structure preparation of HBcAg dimers was evaluated based on its correlation to experimental zeta potential obtained for entire HBcAg VLP structures. Monomers were not considered as model simplification, since only dimers or larger assemblies (i.e., capsids) are present under physiological conditions[319].

This study presents a computationally inexpensive, high-throughput, and entry-level pipeline to obtain 3-D protein structures. Time and computational effort were minimized by automated homology modeling including novel, data-dependent, and stepwise MD simulation for homology model refinement. Refinement termination was determined data-dependently via identification of a 2 ns Window of Stability (WoS) consisting of 1000 structural snapshots. The WoS was used to calculate the median structure quality and median surface charge based on all 1000 structural snapshots to account for MD simulation fluctuations. As a case study, three chimeric HBcAg dimer structures were processed under similar environmental conditions, each with a unique antigenic epitope insert. Homology model construction and subsequent refinement performance was evaluated based on simulation quality, speed, and stability. The median surface charge was used to investigate the application of the prepared structures for surface property extraction. This was evaluated based on the correlation between in silico calculated surface charge extracted from chimeric HBcAg dimers and experimental zeta potential obtained with complete chimeric HBcAg VLPs. To identify performance sensitivity, MD simulations using 2 different force fields (YASARA2 and AMBER03) and 2 high-throughput methods for pKa value computation (H++ and YASARA) were compared. The presented case study of three chimeric HBcAg dimers was performed to show the potential of the proposed high-throughput and automated structure preparation pipeline to explore computationally determined physicochemical protein surface properties.

7.2 Material and Methods

7.2.1 Sample preparation

Recombinant chimeric HBcAg constructs used in this study (referred to as VLP A, VLP B, and VLP C irrespective of being a HBcAg dimer or VLP) were modified in the MIR to display foreign epitopes on the VLP surface. Constructs were expressed and purified according to the production protocol generously provided by BioNTech Protein Therapeutics GmbH (Mainz, DE). Purified and assembled VLPs were stored at -20 °C and dialyzed into a 50 mM Tris (Merck KGaA, Darmstadt, DE) buffer at pH 7.2 containing 100 mM NaCl (Merck KGaA, Darmstadt, DE) for analysis. Buffer was prepared with ultrapure water (PURELAB Ultra, ELGA LabWater, Lane End, UK) and filtered through a 0.20 µm pore size Supor® filter (Pall, Port Washington, NY, USA). Samples were brought to room temperature and filtered through a 0.20 µm polyethersulfone (PES) filter (VWR International, Radnor, PA, USA) before measurements. Required VLP sample concentrations were obtained using Vivaspin® 20 filters with a 30 kDa pore rating (Sartorius, Goettingen, DE). VLP concentration was determined with a NanoDrop2000c UV-Vis spectrophotometer (Thermo Fischer Scientific, Waltham, MA, USA). The E1% (280 nm) extinction coefficient was calculated by the online Swiss Institute of Bioinformatics ProtParam tool (https://web.expasy.org/protparam.html) based on the primary structure of the HBcAg monomer[320].

7.2.2 Zeta potential

Electrophoretic mobility measurements were performed with the Zetasizer Nano ZSP (Malvern Instruments Ltd., Malvern, UK). Folded disposable capillary cells (DTS1070, Malvern Instruments Ltd., Malvern, UK) were filled with the appropriate buffer and 50 µL of a 1 g/L VLP sample. VLP samples were inserted by employing the diffusion barrier technique[321] using a 200 µL round, 0.5 mm thick Corning Costar gel-loading tip (Corning Inc., Corning NY, USA). Six replicates were measured at 25 °C in automatic mode, where each measurement consisted of 120 seconds equilibrium time and five runs with a maximum of 15 sub runs. The applied voltage was set to 60 mV and the dispersant was set to water. A material refractive index of 1.45 and absorption of 0.001 AU was used. The average zeta potential was calculated by Zetasizer Software (version 7.12, Malvern Instruments Ltd., Malvern, UK) with the measured average electrophoretic mobility, a viscosity of 0.8872 mPas, a dielectric constant of 78.54, and Smoluchowski's approximation of 1.5[198]. For each VLP sample, outlier detection was performed with MATLAB (version 2017b, MathWorks, Natick, MA, USA), using the interquartile range rule with a whisker length of 0.75[322], followed by median zeta potential calculation.

7.2.3 Computational methods

Figure 7.1 depicts the computational pipeline used to compute surface property information from dimer chimeric HBcAg structures. Required input is a template 3-D structure, the target sequences, and experimental conditions (i.e., oligostate, pH, and salt concentration). 3-D structure curation and MD scene preparation, described in section 7.2.3.1 (p. 100), were performed fully automated by employing an in-house developed MATLAB script (version 2017b, MathWorks, Natick, MA, USA).

All depicted steps in section Curation and Preparation in Figure 7.1 were an automated operation of either MATLAB, YASARA (version 16.9.23, YASARA Biosciences GmbH, Vienna, AT), Modeller (version 9.18, University of California, San Francisco, CA, USA)[323], H++ (Virginia Tech, Blacksburg, VA, USA, biophysics.cs.vt.edu) or Python (version 2.7.13, Python Software Foundation, Wilmington, DE, USA) sub scripts. These steps resulted in prepared scenes for MD simulation of each VLP construct. MD simulation of the prepared scene is described in section 7.2.3.2 (p.102) and extraction of VLP surface properties is described in section 7.2.3.3 (p.103). The 3-D structure quality was monitored throughout the workflow with the quality Z-score. This is the mean value of the WHAT IF parameters Packing1, PhiPsi and Backbone[299,324]. Quality parameters were calculated using the YASARA2 force field in a TIP3P water filled cubic cell[325], with walls extended 10 Å from the 3-D structure.

7.2.3.1 Structure and scene preparation

The three HBcAg structures used in this study were based on C-terminally truncated and histidine(His)-tagged HBcAg, which were modified in the MIR. All experimental structures have an identical C-terminus. Therefore, it was assumed that the His-tag would not have a significant impact on the relative assessment of 3-D structural biophysical parameters. To avoid homology modeling of the His-tag, the C-termini of the input target sequences matched the template structure C-terminus.

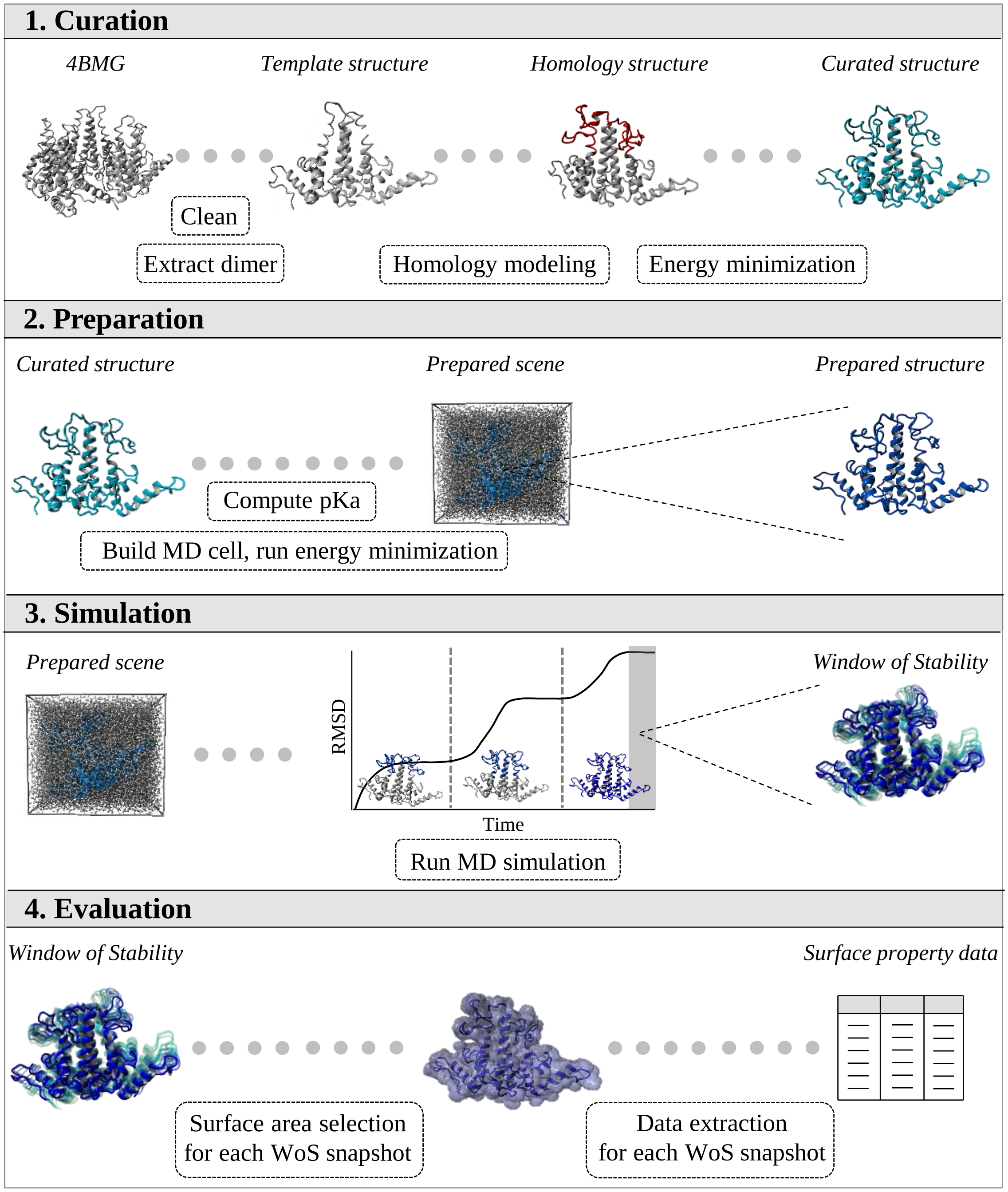


Figure 7.1: Computational pipeline for high-throughput homology model surface property data extraction. Four stages are depicted: (1) Curation: epitope insertion using homology modeling (Modeller), followed by an energy minimization run (YASARA); (2) Preparation: computed pKa values (H++) are assigned, followed by an energy minimization in a simulation cell (YASARA); (3) Simulation: 3-step data-dependent molecular dynamics (MD) simulation (YASARA) terminated by identification of a 2 ns window of stability (WoS); (4) Evaluation: surface area selection and extraction of surface property data for each snapshot in the WoS (YASARA).

The 3-D crystal structure of C-terminally truncated (1-149) hexameric HBc Y132A was obtained from the online research collaborator for structural bioinformatics protein data bank (RCSB PDB, www.rcsb.org), under PDB ID 4BMG with a resolution of 3 Å[326,327]. All non-protein molecules were removed and the hydrogen bonding network was optimized with YASARA [328]. The multimeric state was corrected to obtain a dimeric 3-D structure, which resulted in the template structure shown in Figure 7.1. Subsequently, homology modeling was performed to adjust the template structure to the target sequence using Modeller. The automodel function constructed five homology models, where gap initiation and extension penalties for sequence alignment were set to -600 and -400, respectively. Obtained homology models were superposed in YASARA and their atom coordinates averaged (referred to as homology structure). The hydrogen network was optimized and an energy minimization was run with the averaged structure at experimental pH and using the AMBER99 force field[329]. After steepest descent minimization, the procedure continued by simulating annealing using 2 fs time steps. Atom velocities scaled down by 0.9 every 10th step until the energy improved by less than 0.05 kJ/mol per atom during 200 steps. The resulting structure is referred to as the curated structure in Figure 7.1. The curated structure was uploaded to the H++ webserver using a Python web scraping algorithm (selenium library) to compute pKa values[295]. The external and internal dielectric constant were set to 80 and 10, respectively, and salinity and pH were set equal to experimental conditions (i.e., 0.1 molar salinity and pH 7.2). Obtained pKa values and the resulting 3-D structure were automatically downloaded and used to build an MD simulation cell. Additionally, to investigate the effect of H++ computed pKa values, pKa values computed by YASARA were used instead of H++[294]. The simulation cell contained the prepared 3-D structure, which included computed pKa values as well as (de)protonated termini based on the experimental pH and computed pKa values. Cell walls were built at a distance of 10 Å from the refined 3-D structure. After simulation cell construction, a neutralization run was performed. TIP3P water molecules[325] were added to the simulation cell (water density was set to 0.997) as well as salt ions (set to experimental conditions).The final step of MD scene preparation was an energy minimization using identical settings as described before. This resulted in the prepared MD scene depicted in Figure 7.1.

7.2.3.2 Molecular dynamics

Prepared MD scenes with H++ pKa values were simulated using the YASARA2 or the AMBER03 force field[330], and with YASARA pKa values using YASARA2[299,331], with a cutoff of 7.86 Å[332] and long range Coulomb interactions using the particle mesh Ewald method[333]. Temperature was controlled by rescaling velocities using a modified Berendsen Thermostat[332,334]. Hardware consisted of two Windows 10 computers with an Intel i7-6700 CPU and a GeForce GTX 1080 GPU. Results of the second computer are shown in Supplementary Material Figure E2, Figure E3, Figure E4, and Figure E5. Intramolecular forces were calculated every 2 fs (1 fs for AMBER03) and intermolecular, non-bonded Van der Waals, and electrostatic forces every 4 fs (2 fs for AMBER03) to improve performance and subsequently scaled by 2[335]. MD scene snapshots were saved every 2 ps and superposed on the prepared structure to calculate a root-mean-square deviation (RMSD) of atom coordinates. The simulation was automatically performed in three RMSD-controlled steps. In step 1, only the epitope and five adjacent amino acids were simulated. All other amino acid atom positions were constrained. In step 2, 18 additional amino acids towards the N-terminus and ten amino acids towards the C-terminus (i.e., the dimer spike consisting of two alpha-helical hairpins) were simulated without position constraints. Other amino acids were simulated with free side chain atoms but fixed backbone atom positions. In step 3, all atom positions were unconstrained. All H-bonds were constrained during step 1 and step 2 using the linear constraint solver (LINCS) algorithm[336]. In step 3, all H-bond constraints were removed after 0.2 ns and the time steps for intermolecular forces and intramolecular forces were reduced to 2 fs and 1 fs, respectively. The simulation advanced to the next step when the moving average (window: 0.15 ns, sampling rate: 10 ps) RMSD change was below a set threshold of 0.75 Å/ns for 0.1 ns. A penalty of 0.02 ns was used if the rate of RMSD change was above the threshold. Step 3 was terminated based on the RMSD coefficient of variance (CV) in a window of the last 2 ns of simulation. MD simulation was terminated when the window CV fell below 2.5%, using a sampling rate of 2 ps. The snapshots of the obtained window of stability (WoS) were used for the calculation of quality and descriptors. Simulations that did not reach a WoS within 30 ns were manually stopped.

7.2.3.3 Data processing

The homology structure and all MD snapshots of the WoS obtained with H++ or YASARA pKa values and YASARA2 or AMBER03 force field were analyzed based on their solvent accessible surface area (SASA). Structure SASA was calculated by finding all points a 1.4 Å water probe's oxygen nucleus can reach while rolling over the protein surface approximated by YASARA's numeric algorithm. Contribution of the intra-dimer surface was excluded. Molecular parameters were automatically extracted using similar settings as in the MD simulation. Surface charge was calculated for all atoms contributing to the SASA and the resulting surface charge was divided by the total SASA. This was done to exclude size effects that can occur between different epitope insertions. In silico zeta potential values were obtained via linear transformation of surface charge data. Linear transformation included normalization of in silico data between 0 and 1 and transformation using the minimum and maximum of the experimental data, as shown by Equation 7.1.

$$y_{Transform} = [\tilde{y}_{norm} \cdot (y_{max} - y_{min})] + y_{min} \qquad (7.1)$$

Normalized in silico data is indicated as $\tilde{y}_{norm}$, experimental minimum and maximum data are represented by y_{max} and y_{min}, respectively. Descriptors derived from each snapshot in the WoS are reported as medians and corresponding median absolute deviation (MAD). Correlation between linear transformed in silico data and experimental data was evaluated based on the Pearson correlation coefficient (PCC). PCC was calculated with the *corrcoef* function available in MATLAB. The error between in silico and experimental data was evaluated with the mean squared error (MSE), obtained with Equation 7.2.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 \tag{7.2}$$

where *n* is the sample size, y_{i} experimental data, and $\tilde{y}_{\mathrm{i}}$ in silico generated data.

7.3 Results and Discussion

7.3.1 Quality

Figure 7.2 shows an overview of structural quality Z-scores during curation, preparation, and simulation of each chimeric HBcAg dimer. The structural quality Z-score is an average of three parameters: (1) 3-D direction-dependent packing normality, (2) position normality of residues and secondary structural motifs in the Ramachandran plot, and (3) backbone conformation normality[299]. A value below -2 is considered to represent a poor structure and Z-scores close to or above zero indicate more reliable structures. Separate parameter values can be found in Supplementary Material Figure E1.

Quality Z-score differences were observed throughout the structure preparation workflow and between different identified windows of stability. The template structure quality Z-score (-1.18, gray dashed line) increased after homology modeling with 0.12 and 0.16 for VLP B and VLP C, respectively. VLP A showed a 0.03 quality Z-score decrease compared to the template structure. Underlying parameters showed that VLP A's backbone conformation quality decreased roughly 1.5 times more than the other constructs. This is attributed to the amount of additional atoms included in the homology model. VLP A contains 17% additional atoms compared to the template structure, while VLP B and VLP C contain 13% and 11% additional atoms, respectively. The other two underlying quality parameters (packing normality and Ramachandran plot position normality) show a similar trend between VLP constructs when comparing the template and homology structure (data shown in Supplementary Material Figure E1).

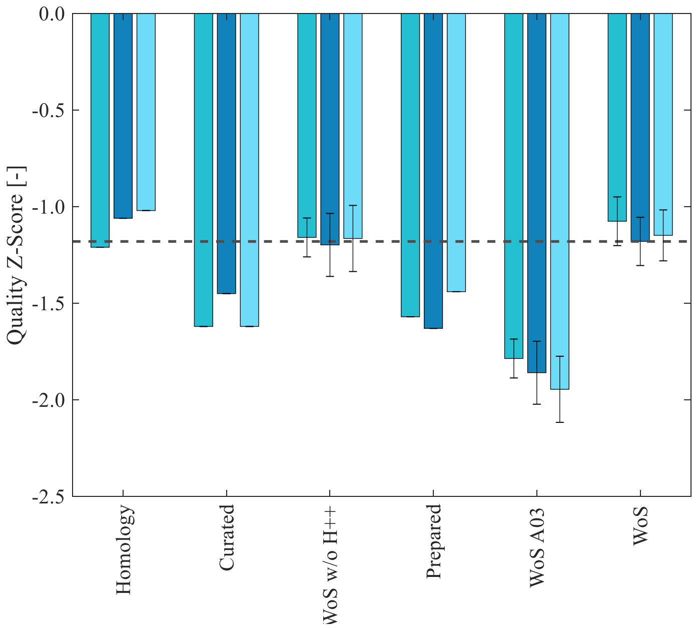


Figure 7.2: Overview of quality Z-scores for the template, homology structure, curated structure, window of stability (WoS) without H++ and the YASARA2 force field ("WoS w/o H++"), the prepared structure, WoS obtained with H++ and the AMBER03 force field ("WoS A03"), and WoS obtained with H++ and the YASARA2 force field ("WoS"). The quality Z-score is an average value of the WHAT IF quality factors 3-D packing (QUACHK), Ramachandran Z-score (RAMCHK) and backbone conformation (BBCCHK)[299]. A median value and median absolute deviation as error bar is shown for the WoS quality Z-scores. The gray dashed line represents the quality Z-score of the template structure.

The observed quality improvement of homology structures VLP B and VLP C, which is dominated by Ramachandran position normality parameter improvement, might be an effect of the restraint-based homology modeling and knowledge-based loop modeling used by Modeller[298,337]. Quality Z-scores of curated and prepared structures were between -1.44 and -1.62, which is between 22% and 37% lower compared to the template structure. Both structures are evaluated after energy minimization at experimental pH, where prepared structures included H++ computed pKa values and the curated structures did not. Energy minimization is used to remove global errors in 3-D structures, such as steric clashes. However, optimization of global and local structural quality with an energy minimization run is not trivial. Energy minimization may result in lower quality structures because global errors are removed but local errors accumulate[299,338]. This may explain quality decrease of curated and prepared structures, when compared to the template and homology structures. A similar decrease in quality Z-score after energy minimization with an AMBER99 force field has been reported before [299]. Structural issues present in curated and prepared structures were resolved by running an MD simulation with the YASARA2 force field, independent of the used pKa computation method. Mean quality Z-scores of all VLP constructs for MD simulation WoS without H++ (-1.17) and MD simulation WoS with H++ (-1.13) were comparable to the template. This shows there is no quality loss after completing the proposed structure preparation pipeline with the YASARA2 force field. Additionally, the coefficients of quality Z-score variance of 4.72% and 1.79% for the MD simulation with and without H++ pKa values, respectively, reflected that there is no quality influence of the inserted epitope length. However, a decrease in quality is seen for the WoS obtained with the MD simulation using the AMBER03 force field (WoS A03), represented by a mean quality Z-score of -1.86 considering all VLP constructs. This corresponds to observations previously reported about diverse structure quality values obtained with different force fields [339]. Quality Z-scores for intermediate structures and final MD simulation WoS showed that chimeric HBcAg dimer structure quality in this dataset was mostly influenced by the force field and an MD simulation, independent of the used pKa computation method.

7.3.2 MD simulation

All chimeric HBcAg homology models were refined with MD simulations. This was done because MD simulations correct structural errors present in homology models[99]. An MD simulation results in a change of atom coordinates, which is measured by the RMSD of those atom coordinates. Structure refinement is achieved upon stabilization of atom positions, referred to as the equilibrium state. This state is identified by a plateau of the RMSD value over simulation time. Plateau identification is frequently done subjectively based on visual inspection of RMSD plots. This approach is not recommended as it was shown to be biased in a survey among researchers in the field[340]. To avoid subjective plateau identification, this study employed automated equilibrium state determination based on the average RMSD slope or CV. Automated determination was used within a 3-step MD simulation. In each step, a growing part of the chimeric HBcAg dimer structure was refined until an equilibrium was identified. Separate refinement of structure parts was used to reduce simulation time in addition to automated identification of the equilibrium state. The simulation was terminated when equilibrium was reached for the full chimeric HBcAg dimer structure. This state is referred to as the WoS, which was defined as a 2 ns simulation window where the RMSD CV in step 3 was below 2.5%. The 3-step MD simulation was specifically implemented for the HBcAg dimer structure, as sequences differ only in the MIR. For other applications, (i.e., formulation condition screening of a single protein or a diverse protein dataset) a 3-step MD simulation may not be necessary and a WoS could be determined in one simulation step.

Figure 7.3 shows the progress of 3-step MD simulations with the YASARA2 force field and H++ computed pKa values for all three VLPs. The atom coordinate RMSD was calculated every 0.002 ns by superposing a simulation snapshot on the prepared structure.

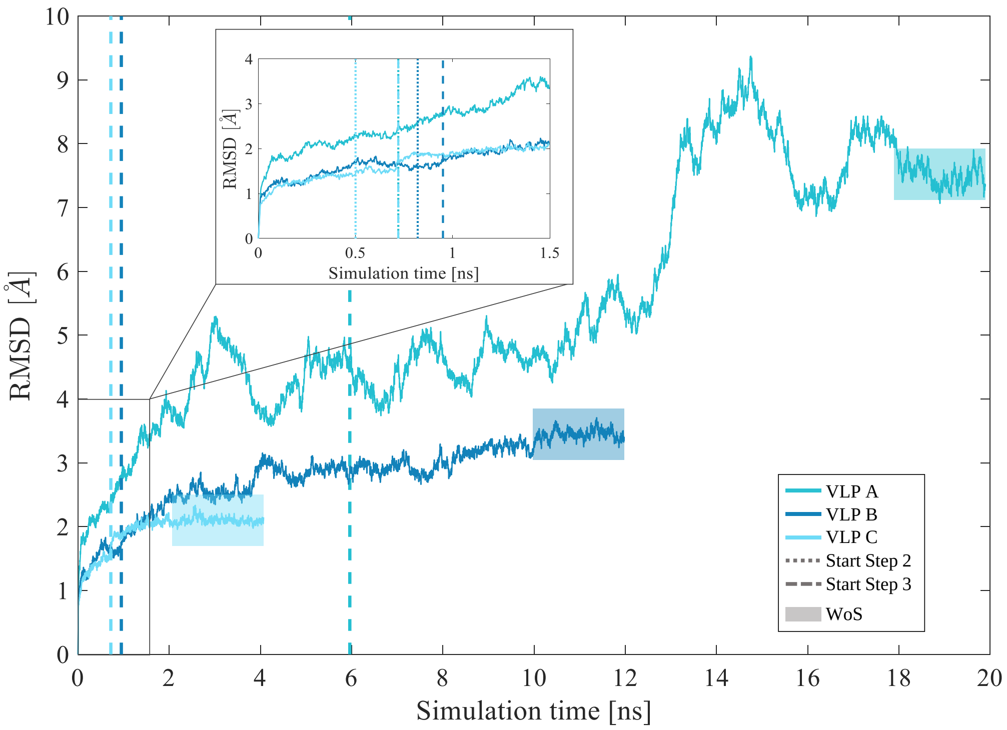


Figure 7.3: Progress of molecular dynamics (MD) simulations for VLP A, B, and C presented by root-mean-square deviation (RMSD) of atom coordinates (Å) over simulation time (ns). Three different simulation steps are separated by vertical lines, where vertical lines indicate simulation transition points. From 0 ns to dotted line: simulation of epitope and five adjacent amino acids; from dotted to dashed line: simulation of Hepatitis B core antigen (HBcAg) dimer spike; from dashed line to the end of simulation: full dimer simulation. The highlighted area is defined as the 2 ns window of stability (WoS).

Overall simulation time ranged from 4.0 ns to 19.9 ns and the absolute RMSD increased to 2.10 ± 0.04 Å to 7.52 ± 0.15 Å during MD simulation. The in silico time span difference between structures to reach the WoS is in agreement with other work, where structure stability was achieved earlier, later, or not at all, depending on the protein[99]. VLP C showed the lowest RMSD increase (2.1 Å ± 0.04 in the WoS) and shortest simulation time (6.6 h; in silico: 4.01 ns). VLP A resulted in the largest RMSD increase (7.52 ± 0.15 Å in the WoS) and longest simulation time (37.5 h; in silico: 19.89 ns). Simulation time increased from VLP C to VLP B to VLP A, which corresponds to the number of inserted atoms of 11%, 13%, and 17%, respectively. Step 1, which simulates the inserted epitope and five adjacent amino acids, showed 32.1% to 69.2% of the total RMSD change. This is a relatively large percentage considering step 1 accounted for 3.6% to 12.5% of the total simulation time. The epitope was not part of the template 4BMG crystal structure and therefore it was inserted with homology modeling. Homology models typically have errors in the secondary structure and atomic packing which should be resolved during MD simulation[99]. This is presumably one factor contributing to the relatively large RMSD change observed in step 1, which only refined the inserted epitope and five adjacent amino acids. Another factor that can influence the observed RMSD profile of the epitope is its flexible design. It was stated that epitope flexibility allows for efficient presentation to the immune system[341], but increased structure flexibility can also result in larger RMSD change during MD simulation. Other parts of the HBcAg dimer are less flexible. Therefore, only small deviations in atom coordinates of the less flexible and conserved region of chimeric HBcAg (i.e., the molecule base and lower part of the spike) were observed when comparing MD simulation steps. This is also illustrated by Figure 7.4, where the RMSD per residue number is shown. Figure 7.4 shows that regions around the epitope have higher RMSD values than other regions.

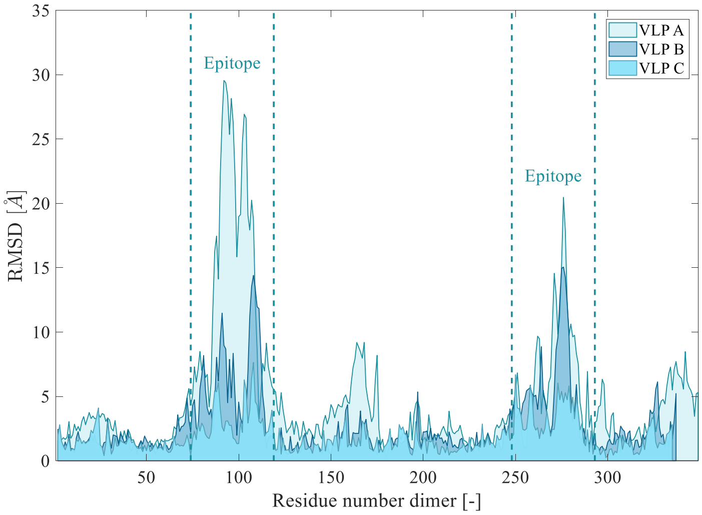


Figure 7.4: Local structural changes during molecular dynamics (MD) simulation represented by root-mean-square deviation (RMSD) of atom coordinates (Å) over residue number (-). Initial structures were compared with last MD simulation snapshots of VLP A, B, and C, respectively, with the YASARA2 force field and H++ computed pKa values. Vertical lines mark the inserted epitope exemplarily for VLP A.

Simulation speed improved due to bond and regional atom constraints and due to an increased time step for force calculation in the first two steps of the simulation. On average, step 1 was 72% (21.26 ns/day) and step 2 was 69% (20.82 ns/day) faster compared to step 3 without constraints and with a smaller time step (12.32 ns/day). This supports the expected simulation speed improvement by employing a data-dependent 3-step method. This corresponds to the previous statement that simulation design should be adjusted to the application and starting structure to obtain optimal speed and stability output. With the used simulation approach, the 2 ns WoS of three chimeric HBcAg dimers were created on a Windows 10 computer with an Intel i7-6700 CPU and a GeForce GTX 1080 GPU in 66.0 h of computational time using the YASARA2 force field and H++ computed pKas. Simulations with H++ pKa values and YASARA2 as force field were also run on another computer containing similar hardware to evaluate reproducibility. No significant difference in simulation outcome was found, including calculation of quality and surface charge.

More detailed information on reproducibility can be found in Supplementary Material Figure E2 to Figure E5.

Two additional simulations were performed, the first to evaluate the effect of different pKa value computation methods, and the second to compare MD simulation with YASARA2 to a standard force field for protein simulations, AMBER03. Figure 7.5a shows the progress of MD simulations using the YASARA2 force field and with YASARA computed pKas (without H++; w/o H++) and Figure 7.5b shows MD simulations with the AMBER03 force field with H++ computed pKas (A03).

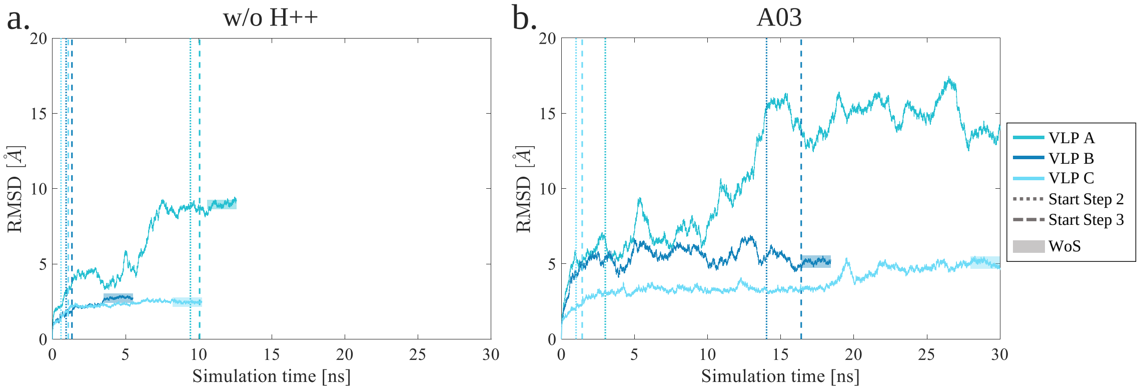
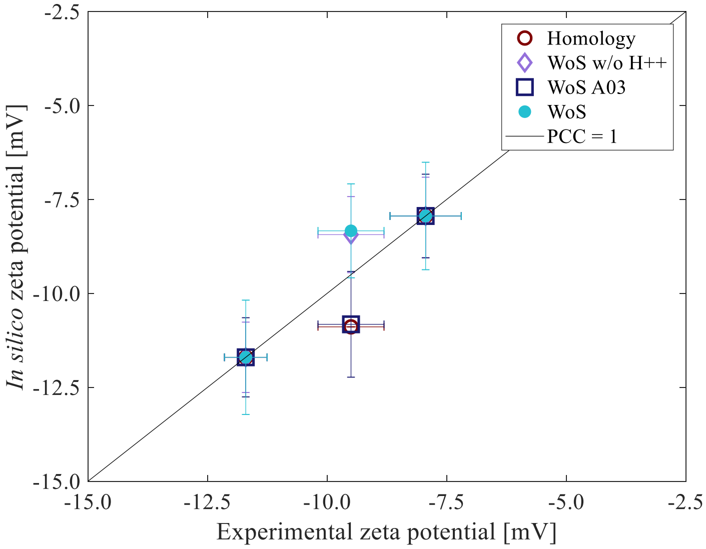


Figure 7.5: Progress of molecular dynamics (MD) simulation for VLP A, B, and C presented by root-mean-square deviation (RMSD) of atom coordinates (Å) over simulation time (ns) for (a) MD simulation without H++ with YASARA2 as force field ("w/o H++") and (b) MD simulation with H++ and AMBER03 as force field ("A03"). Three different simulation steps are separated by vertical lines, where vertical lines indicate simulation transition points. From 0 ns to dotted line: simulation of epitope and five adjacent amino acids; from dotted to dashed line: simulation of Hepatitis B core antigen (HBcAg) dimer spike; from dashed line to the end of simulation: full dimer simulation. The highlighted area is defined as the 2 ns window of stability (WoS).

During MD simulations w/o H++, RMSD increased by 2.46 ± 0.05 Å to 8.95 ± 0.17 Å in 5.5 ns to 12.6 ns corresponding to 11.0 h to 30.5 h of computational time. The total computational time of 59.6 h for MD simulations without H++ computed pKa values was comparable to 66.0 h for MD simulations with H++ pKa values. This shows that the pKa calculation method did not have a significant influence on MD simulation performance. MD simulations with AMBER03 resulted in RMSD values of 5.10 ± 0.16 Å to 13.66 ± 0.25 Å. MD simulation took 18.42 ns to 30.0 ns which corresponds to a total computational time of 156 h. For A03, the MD time step had to be reduced to 1 fs for intramolecular and to 2 fs for intermolecular forces to avoid simulation failure. Structure instability also prevented the transition to MD simulation step 3 for VLP A, which is elucidated by a fluctuating RMSD curve in Figure 5. Furthermore, VLP C did not reach a WoS within 30 ns. Both results indicated that using AMBER03 resulted in less stable simulations compared to simulations with YASARA2. Simulations with H++ or YASARA computed pKa values using the YASARA2 force field have shown the best performance based on simulation time, simulation stability, and overall completion of the 3-step MD simulation method. This indicates that MD simulations evaluated in this study benefitted from the empirical data that is embodied in a force field containing knowledge-based potentials[299,342]. Evaluation of this method based on other (refined) force fields and other software platforms would give more detailed insight into simulation performance.

7.3.3 Zeta potential

Zeta potential was experimentally determined for all three HBcAg VLP constructs and compared to in silico determined total surface charge based on the HBcAg dimer structures. This was done to determine the applicability of the prepared structures for computational surface property extraction. It was assumed that the observed zeta potential differences that occur due to the changes on the outer surface of the entire VLP structure are captured by the dimeric HBcAg structure[343]. The obtained in silico surface charge extracted from the homology model and three different WoS, for each of the three chimeric HBcAg dimer structures, are shown in Figure 7.6.



Structure	**MSE**	**PCC**
Homology	0.64	0.915
WoS w/o H++	0.38	0.954
WoS A03	0.58	0.922
WoS	0.45	0.946

Figure 7.6: In silico computed zeta potential (mV) plotted against experimentally determined zeta potential (mV). Symbols represent in silico data based on the homology structure ("Homology", red open circle), window of stability (WoS) obtained without H++ and with YASARA2 ("WoS w/o H++", purple diamond), WoS obtained with H++ and AMBER03 ("WoS A03", purple square), and WoS obtained with H++ and YASARA2 ("WoS", blue filled circle). The diagonal line represents theoretical data with a Pearson correlation coefficient of 1 (PCC = 1). X-axis error bars represent the median absolute deviation (MAD) of experimental data and y-axis error bars represent MAD for in silico data points. For each in silico data series the PCC and mean squared error (MSE) are calculated (n = 3) and listed.

Linear transformation of in silico data was applied to obtain comparable scales and different MD simulation refinement settings were used to determine the effects on in silico generated data and the respective correlation to experimentally determined zeta potential. Linear transformation resulted in ranking three VLPs according to their zeta potential. Figure 7.6 shows that zeta potentials of complete chimeric VLPs were ranked correctly by all dimer structures, which causes overlaying symbols at [-11.70, -11.70] and [-7.94, -7.94]. The main difference is seen for VLP C, which has an experimental zeta potential of -9.50 ± 0.69 mV. This data point was used to evaluate the influence of pKa value computation method and force field selection on in silico surface charge calculations. The evaluation parameters, PCC and MSE, are listed for each data series in Figure 7.6. A PCC value above 0.900 indicates a strong linear dependency with experimental data[344]. This was seen for all evaluated data series because of the limited dataset size, but small differences were observed for VLP C's surface charge. WoS simulated without H++ pKa values and WoS with H++ pKa values showed the highest PCC, with values of 0.954 and 0.946, respectively. Transformed VLP C surface charges for WoS w/o H++ (-8.44 ± 1.18) and WoS with H++ (-8.33 ± 1.43) were also comparable, which resulted in a 0.07 MSE difference in favor of WoS w/o H++. The WoS transformed surface charge distribution, represented by the MAD, shows an overlap between these two values. This indicates there is no significant influence of the used pKa value computation methods in correlation to experimental data. This result was reproducible (data shown in Supplementary Material Figure E2 to Figure E5). Transformed surface charges based on the homology structure (-10.89) and WoS A03 (-10.82 ± 1.11) showed a weaker correlation than the WoS previously discussed. This is shown by MSE values of 0.64 and 0.58, respectively. Linear dependency is also weaker compared to the other two WoS, where the homology structure showed a PCC of 0.915 and WoS obtained with AMBER03 showed a PCC of 0.922. As mentioned during the discussion of the MD simulations, VLP A did not complete step 2 and VLP C did not reach a WoS when the AMBER03 force field was used during MD simulation. Presumably this also caused the decreased correlation to experimental data. This leads to the conclusion that for this case study the largest positive effect was obtained with the YASARA2 force field, regardless of the used pKa values, when evaluating the correlation between in silico HBcAg dimer surface charge and complete chimeric VLP zeta potential. The observed force field effect should be confirmed with a larger dataset. Nevertheless, results indicate that surface properties extracted from structures obtained with the presented pipeline can represent experimental behavior. It should be noted that the applicability of chimeric dimer 3-D structure surface charge to quantitatively predict complete chimeric VLP zeta potential lies outside the scope of this case study, and should be investigated using a more diverse sample space.

All evaluated WoS show a relatively large coefficient of variation (10% – 16%) regarding the in silico zeta potential, which means there is a significant variation in protein surface property value within the WoS. For example, VLP A simulated with H++ pKa values and the YASARA2 force field resulted a maximum in silico zeta potential of -5.74 mV and minimum of -11.07 mV within its 2 ns WoS. This emphasizes cautiousness regarding the use of a single MD simulation snapshot because a snapshot can theoretically take any random value within the WoS. The use of a single snapshot can decrease correlation accuracy and thereby reduce the reliability of computational protein structure-based models. Therefore, a robust central tendency describing statistic which is less sensitive for outliers, such as the median[345], is considered appropriate for the extraction of protein surface property information within a WoS. The presented computational pipeline did not only show the potential of a high-throughput approach for 3-D structure preparation, but also how a WoS can provide an objective MD simulation termination to reduce computational effort and a robust descriptor extraction platform. The approach could be used for other proteins, such as antibodies, and other prediction targets, such as assembly competence, solubility, or surface hydrophobicity. A variety of proteins and other prediction targets should be investigated to determine the full potential of the proposed computational 3-D structure preparation pipeline.

7.4 Conclusion

A computationally inexpensive, fully automated, and data-dependent pipeline for high-throughput 3-D protein structure preparation and refinement was constructed and evaluated using a case study of three chimeric HBcAg dimers. Structure quality, computational speed, simulation stability, and zeta potential correlation have been evaluated for three different simulation settings. This was done by homology modeling and subsequent structure refinement with 2 different force fields (YASARA2 or AMBER03) and 2 different pKa values (H++ or YASARA computed pKa values). All evaluation parameters showed to be mainly influenced by the choice of force field, where YASARA2 showed a more stable performance than AMBER03. YASARA2 simulations using either pKa computation method resulted in comparable average quality Z-score (-1.17 and -1.13). All three chimeric HBcAg dimer structures, modelled and refined with YASARA2, were obtained within 59.6 to 66.0 hours (in silico time of ~4 ns to ~20 ns per structure) on a powerful yet ordinary desktop computer. These simulation times were ~2.4 times shorter than simulations using the AMBER03 force field. Computational efficiency was achieved by designing a 3-step MD simulation refinement complementary to the structures in question. This design resulted in simulating 31.2% to 69.2% of the total RMSD change in 3.6% to 12.5% of the simulation time. In addition, homology model refinement included a data-dependent simulation termination based on a 2 ns window of stability, which was also used for robust surface property descriptor extraction. Validity of the calculated surface property was exemplarily evaluated by correlating in silico determined surface charge, based on the chimeric HBcAg dimer structures, to experimental zeta potential of the entire VLP structure. The use of dimers instead of entire VLP structures contributed to the relative short simulation time, while a high correlation (PCC of ~0.950) to experimental zeta potential was maintained. The case study showed promising results for high-throughput in silico surface property screening, but its full potential should be further explored with a larger dataset. The simple, standardized, and automated framework allows for the implementation of the computational pipeline in manufacturability and formulability screening studies for early candidate assessment.

7.5 Acknowledgements

This work was financially supported by the BE-Basic Foundation (www.be-basic.org), under project FS2.003 and received funding from Deutsche Forschungsgemeinschaft (DFG) in the frame of SPP 1934, project number 273937032. The authors want to thank academic and industrial partners for scientific discussions during the development of this work, especially Thorsten Klamp from BioNTech Protein Therapeutics GmbH. The authors also want to thank Angela Valentíc for assisting with experimental work.

8 Conclusion

This thesis focused on the design and implementation of computational methods applicable for the analysis of long-term protein phase behavior. The employed computational methods include unsupervised and supervised machine learning approaches, with an emphasis on data-dependent automation of what were previously manual procedures. Unsupervised multidimensional data visualization was applied for a comprehensive and complete representation of data obtained from long-term protein phase behavior experiments. The resulting MPPDs allowed for the identification of subtle kinetic and morphological changes as a function of applied environmental conditions. A supervised image recognition algorithm was implemented to reduce the manual effort required for image-based feature extraction, a necessary step to construct MPPDs. To improve image recognition accuracy for automated evaluation of long-term protein phase behavior studies, the combination of three light sources (visible, cross polarized, and UV light) and kinetic features was investigated. The additional information returned by cross polarized light, UV light, and kinetic data resulted in a 17.3 percent point increase in balanced accuracy compared to the use of only end point visible light images. Subsequent connection of the supervised image recognition algorithm to unsupervised multidimensional MPPD construction resulted in an automated workflow that mines raw images to classify protein phase diagrams and constructs MPPDs.

Unsupervised multidimensional data visualization was also applied to investigate the correlation between empirical properties obtained directly after formulation preparation and long-term protein phase behavior. This resulted in the construction of an EPPD, a figure containing short-term data from six different analytical techniques. The EPPD showed partial correlation to long-term protein phase behavior represented by the MPPD. This indicated the applicability of short-term empirical data for the design of rational and knowledge-based long-term stability screenings. The developed workflow was subsequently applied to an industry case study to identify long-term stable glycerol-poor and glycerol-free food protein formulations. This led to the identification of apparent protein surface hydrophobicity and electrostatic repulsive forces as product-specific targets to enhance stability.

To construct predictive models for long-term protein phase behavior based on in silico generated protein properties by means of MD simulations, an automated pipeline for high-throughput 3-D structure preparation and refinement was developed and evaluated. Due to its computational inexpensiveness and data-dependent framework, the largest structural errors were refined within only 3.6-12.5% of the total computational time, leading to an automated refinement within 6.6-37.5 hours. A case study was performed with a relatively small set of three dimeric VLP capsid protein structures, which indicated the advantages of the presented automated structure preparation pipeline for large scale screening purposes.

In conclusion, this thesis explored and applied computational methods to automatically extract data from protein phase behavior experiments and visualize such multidimensional dataset for straightforward interpretation. The obtained information was used to assess the correlation between long-term protein phase behavior and short-term multidimensional empirical data sets. In addition, a contribution was made for the in silico generation of protein properties by developing of a high-throughput 3-D protein structure preparation pipeline. The efforts reported in thesis and in literature, in combination with the challenges that still need to be resolved, shape an outline of the computer-supported and knowledge-based infrastructure required to rationally and systematically develop long-term stable protein formulations in a shorter time frame.

9 Outlook

The correlation between short-term empirical properties and long-term protein phase behavior was shown retrospectively in this thesis, as long-term protein phase behavior experiments were performed for the verification of the found short-term empirical profiles. This time-consuming step will still be required when other proteins are investigated due to the unique stability behavior of different proteins, or when different additives are used which may lead to different stabilizing pathways. Thus, an internal standard representing the protein and its corresponding potential stabilizing pathways is currently necessary in order to verify the correlation between short-term empirical data and long-term phase behavior. A standard-free and indisputable verification of the predictive capacity of short-term measurable or computable properties is the most challenging task to complete in order to eliminate the long experimental time required for the demonstration of protein-product shelf life. Without a reliable correlation between short-term effects and results obtained with the established long-term methods, short-term protocols cannot be accepted by regulatory agencies as a stand-alone product assessment. To verify the correlation, advanced, well-defined, physically realistic in silico simulations are required, in addition to short-term empirical measurements and machine learning approaches trained on previous protein-product formulations results. Key to the construction of such an infrastructure is the collaboration between the biotechnology industry and academia. Over the past decades a large amount of data for successful long-term formulation has been generated. In addition, there is even more data available on the many failed attempts, which should not be considered unimportant. Currently, this large amount data acquired on successes, and even more failures, in long-term protein phase behavior studies is not shared between industry and academia. One of the first steps to solve this mutual knowledge gap is mining these important sources.

The presented short-term characterization and visualization of protein phase behavior under a variety of environmental conditions is not only applicable to study long-term protein phase behavior for effective formulation development. In fact, environmental conditions are continuously changed during downstream processing (DSP) in order to obtain a pure product. In addition, fluctuations in the environment during upstream processing (USP) influence product stability due to variations in system properties such as impurity content. A correlation between USP and DSP environmental conditions, short-term empirical properties, and protein stability would be interesting to investigate. This may lead to machine learning approaches that do not only select process parameters optimized towards yield and purity, as is currently done for USP and DSP, but also incorporate protein stability as an optimization target.

As mentioned above, the prediction of protein stability will undoubtedly include in silico generated protein properties. However, standardized guidelines and quality control protocols for the development of MD simulation workflows to reliably extract the desired in silico protein property data are currently missing. The know-how of computational workflows should be more widely discussed in order to obtain higher quality studies which focus on the generation and exploration of these in silico generated protein properties. The guidelines should include, but are not limited to, the required number of protein structures in training sets, an assessment on a representative variety of biophysical protein properties in training sets, standardized protocols for probing MD simulation settings, and comparable evaluation parameters. Once a well-defined infrastructure is realized, it would be desired to generate an in silico protein property diagram to complement the EPPD and MPPD. In addition, the generation of predictive models for analytical techniques, such as for M3-PALS to obtain the zeta potential or the stalagmometric method to obtain the apparent surface hydrophobicity, would be of interest to move towards an in silico predicted EPPD.

In general, machine learning is increasingly applied in the biotechnological field. Currently, so-called black box algorithms, such as artificial neural networks, are used to identify data patterns in dimensions experimenters cannot comprehend or interpret. However, the obtained patterns should be translated to biologically relevant and transparent information, which will allow experimenters to move away from black box approaches. In turn, pattern interpretation allows for data utilization to control and predict current biological uncertainties. Inherent to the use of machine learning approaches for pattern recognition or data utilization, is the need for consistent data acquisition in order to retrieve reliable information. Experimental deviations should be minimized via reproducible, high-throughput, automated, and systematic approaches. In addition, limitations of applied algorithms should be carefully evaluated and discussed in order to prevent an under- or overestimation of its applicability. To reach the desired understanding and control, future research requires a continuing exploration and expansion of the symbiotic relationship between experimenters, robotics, sensors, and data handling algorithms in all stages of biotechnological product development.

Bibliography

1. Cohen SN, Chang ACY, Boyer HW, Helling RB. Construction of biologically functional bacterial plasmids in vitro. *PNAS*. 1973;70(11):3240-3244. doi:https://doi.org/10.1073/pnas.70.11.3240
2. Sheridan C. Illumina claims $1000 genome win. *Nature Biotechnology*. 2014;32(2):111-112. doi:10.1038/nbt0214-111
3. Mardis ER. Anticipating the $1000 genome. *Genome Biology*. 2006;7(7):112. doi:10.1186/gb-2006-7-7-112
4. Tyers M, Mann M. From genomics to proteomics. *Nature*. 2003;422(6928):193-197. doi:10.1038/nature01510
5. Tavassoly I, Goldfarb J, Iyengar R. Systems biology primer: The basic methods and approaches. *Essays In Biochemistry*. 2018;62(4):487-500. doi:10.1042/ebc20180003
6. Redwan E. Cumulative updating of approved biopharmaceuticals. *Human Antibodies*. 2018;16(3-4):137-158. doi:10.3233/hab-2007-163-408
7. Mittler R, Blumwald E. Genetic engineering for modern agriculture: Challenges and perspectives. *Annual Review of Plant Biology*. 2010;61(1):443-462. doi:10.1146/annurev-arplant-042809-112116
8. Purnick PEM, Weiss R. The second wave of synthetic biology: From modules to systems. *Nature Reviews Molecular Cell Biology*. 2009;10(6):410-422. doi:10.1038/nrm2698
9. Endy D. Foundations for engineering biology. *Nature*. 2005;438(7067):449-453. doi:10.1038/nature04342
10. Lee JM. *Biochemical Engineering*. Englewood Sliffs, NJ: Prentice Hall; 1992. doi:10.1016/s0958-1669(02)00306-3
11. Nayar R, Mosharraf M. Effective approaches to formulation development of biopharmaceuticals. In: Jameel F, Hershenson S, eds. *Formulation and Process Development Strategies for Manufacturing Biopharmaceuticals*. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2010:309-328.
12. Weiss WF, Young TM, Roberts CJ. Principles, approaches, and challenges for predicting protein aggregation rates and shelf life. *Journal of pharmaceutical sciences*. 2009;98(4):1246-1277. doi:10.1002/jps.21521
13. Walsh G. Biopharmaceutical benchmarks 2010. *Nature Biotechnology*. 2010;28(9):917. doi:10.1038/nbt0910-917.
14. Clemente M, Corigliano MG, Pariani SA, Sánchez-López EF, Sander VA, Ramos-Duarte VA. Plant serine protease inhibitors: Biotechnology application in agriculture and molecular farming. *International journal of molecular sciences*. 2019;20(6):1345. doi:10.3390/ijms20061345
15. Trono D. Recombinant enzymes in the food and pharamceutical industries. In: Singh RS, Singhania RR, Pandey A, Larroche C, eds. *Advances in Enzyme Technology*. Amsterdam: Elsevier B.V.;

2019:349-387. doi:10.1016/B978-0-444-64114-4.00013-3

16. Demain AL, Vaishnav P. Production of recombinant proteins by microbes and higher organisms. *Biotechnology advances*. 2009;27(3):333-345. doi:10.1016/j.biotechadv.2009.01.008

17. Stability testing of biotechnological/biological products Q5C. In: *Internation Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use*. ; 1995.

18. Chi EY, Krishnan S, Randolph TW, Carpenter JF. Physical stability of proteins in aqueous solution: Mechanism and driving forces in nonnative protein aggregation. *Pharmaceutical Research*. 2003;20(9):1325-1336. doi:10.1023/A:1025771421906

19. Chang BS, Yueng B. Physical stability of protein pharmaceuticals. In: Jameel F, Hershenson S, eds. *Formulation and Process Development Strategies for Manufacturing Biopharmaceuticals*. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2010:69-104.

20. Chaudhuri R, Cheng Y, Middaugh CR, Volkin DB. High-throughput biophysical analysis of protein therapeutics to examine interrelationships between aggregate formation and conformational stability. *The AAPS Journal*. 2014;16(1):48-64. doi:10.1208/s12248-013-9539-6

21. Wang W, Roberts CJ. Protein aggregation – Mechanisms, detection, and control. *International Journal of Pharmaceutics*. 2018;550(1-2):251-268. doi:10.1016/j.ijpharm.2018.08.043

22. Cromwell MEM, Hilario E, Jacobson F. Protein aggregation and bioprocessing. *The AAPS journal*. 2006;8(3):E572-E579. doi:10.1208/aapsj080366

23. Manning MC, Liu J, Li T, Holcomb RE. Rational design of liquid formulations of proteins. *Advances in Protein Chemistry and Structural Biology*. 2018;112:1-59. doi:10.1016/bs.apcsb.2018.01.005

24. Mardis ER. The $1000 genome , the $100000 analysis? *Genome Medicine*. 2010;2(11):84. doi:10.1186/gm205

25. Philo JS, Arakawa T. Mechanisms of protein aggregation. *Current Pharmaceutical Biotechnology*. 2009;10(4):348-351. doi:10.2174/138920109788488932

26. Wang W, Nema S, Teagarden D. Protein aggregation - Pathways and influencing factors. *International Journal of Pharmaceutics*. 2010;390(2):89-99. doi:10.1016/j.ijpharm.2010.02.025

27. Leckband D, Sivasankar S. Forces controlling protein interactions: Theory and experiment. *Colloids and Surfaces B: Biointerfaces*. 1999;14(1-4):83-97. doi:10.1016/S0927-7765(99)00027-2

28. De Young LR, Fink AL, Dill KA. Aggregation of globular proteins. *Accounts of Chemical Research*. 1993;26(12):614-620. doi:10.1021/ar00036a002

29. Philo J. A critical review of methods for size characterization of non-particular protein aggregates. *Current Pharmaceutical Biotechnology*. 2009;10(4):359-372. doi:10.2174/138920109788488815

30. Hawe A, Wiggenhorn M, van der Weert M, Garbe JHO, Mahler H-C, Jiskoot W. Forced degradation of therapeutic proteins. *Journal of pharmaceutical sciences*. 2008;101(3):895-913.

doi:10.1002/jps.22812

31. Chang BS, Kendrick BS, Carpenter JF. Surface-induced denaturation of proteins during freezing and its inhibition by surfactants. *Journal of Pharmaceutical Sciences*. 1996;85(12):1325-1330. doi:10.1021/js960080y

32. Franks F. Freeze-Drying: A Combination of physics, chemistry, engineering and economics. *Japanese Journal of Freezing and Drying*. 1992;38:5-16. doi:10.20585/touketsukansokaishi.38.0_5

33. Privalov PL. Cold denaturation of protein. *Critical Reviews in Biochemistry and Molecular Biology*. 1990;25(4):281-306. doi:10.3109/10409239009090612

34. Scopes RK. Protein purification: Principles and practise. In: Cantor CR, ed. *Protein Purification: Principles and Practise*. Third. New York: Springer science + Business Media New York; 2013.

35. Dill KA. Dominant forces in protein folding. *Biochemistry*. 1990;29(31):7133-7155. doi:10.1021/bi00483a001

36. Giger K, Vanam RP, Seyrek E, Dubin PL. Suppression of insulin aggregation by heparin. *Biomacromolecules*. 2008;9(9):2338-2344. doi:10.1021/bm8002557

37. Majhi PR, Ganta RR, Vanam RP, Seyrek E, Giger K, Dubin PL. Electrostatically driven protein aggregation: Beta-lactoglobulin at low ionic strength. *Langmuir*. 2006;22(22):9150-9159. doi:10.1021/la053528w

38. Arakawa T, Timasheff SN. Mechanism of protein salting in and salting out by divalent cation salts: Balance between hydration and salt binding. *Biochemistry*. 1984;23(25):5912-5923. doi:10.1021/bi00320a004

39. Hofmeister F. Zur Lehre von der Werkung der Salze. *Archiv fuer experimentelle Pathologie und Pharmakologie*. 1888;24(4-5):247-260.

40. Omta AW, Kropman MF, Woutersen S, Bakker HJ. Negligible effect of ions on the hydrogen-bond structure in liquid water. *Science*. 2003;347(5631):347-350. doi:10.1126/science.1084801

41. Boström M, Tavares FW, Finet S, Skouri-Panet F, Tardieu A, Ninham BW. Why forces between proteins follow different Hofmeister series for pH above and below pI. *Biophysical Chemistry*. 2005;117(3):217-224. doi:10.1016/j.bpc.2005.05.010

42. Zhang Y, Cremer PS. The inverse and direct Hofmeister series for lysozyme. *PNAS*. 2009;106(36):15243-15253.

43. Schwierz N, Horinek D, Sivan U, Netz RR. Reversed Hofmeister series—The rule rather than the exception. *Current Opinion in Colloid and Interface Science*. 2016;23:10-18. doi:10.1016/j.cocis.2016.04.003

44. Saluja A, Kalonia DS. Nature and consequences of protein-protein interactions in high protein concentration solutions. *International Journal of Pharmaceutics*. 2008;358(1-2):1-15. doi:10.1016/j.ijpharm.2008.03.041

45. Hamada H, Arakawa T, Shiraki K. Effect of additives on protein aggregation. *Current*

pharmaceutical biotechnology. 2009;10(4):400-407. doi:10.2174/138920109788488941

46. Kamerzell TJ, Esfandiary R, Joshi SB, Middaugh CR, Volkin DB. Protein-excipient interactions: Mechanisms and biophysical characterization applied to protein formulation development. *Advanced Drug Delivery Reviews*. 2011;63(13):1118-1159. doi:10.1016/j.addr.2011.07.006

47. Goldberg DS, Bishop SM, Shah AU, Sathish HA. Formulation development of therapeutic monoclonal antibodies using high-troughput fluorescence and static light scattering techniques: Role fo conformational and colloidal stability. *Journal of pharmaceutical sciences*. 2011;100(4):1306-1315. doi:10.1002/jps.22371

48. Arakawa T, Timasheff SN. Stabilization of protein structure by sugars. *Biochemistry*. 1982;21(25):6536-6544. doi:10.1021/bi00268a033

49. Mollmann SH, Elofsson U, Bukrinsky JT, Frokjaer S. Displacement of adsorbed insulin by Tween 80 monitored using total internal reflection fluorescence and ellipsometry. *Pharmaceutical Research*. 2005;22(11):1931-1941. doi:10.1007/s11095-005-7249-1

50. Rupp B. Origin and use of crystallization phase diagrams. *Acta Crystallographica Section F Structural Biology Communications*. 2015;71(3):247-260. doi:10.1107/s2053230x1500374x

51. Asherie N. Protein crystallization and phase diagrams. *Methods*. 2004;34(3):266-272. doi:10.1016/j.ymeth.2004.03.028

52. Krauss IR, Merlino A, Vergara A, Sica F. An overview of biological macromolecule crystallization. *International Journal of Molecular Sciences*. 2013;14(6):11643-11691. doi:10.3390/ijms140611643

53. Arakawa T, Timasheff SN. Theory of protein solubility. *Methods in enzymology*. 1985;114:49-77. doi:10.1016/0076-6879(85)14005-X

54. Luft JR, Wolfley JR, Snell EH. What's in a drop? Correlating observations and outcomes to guide macromolecular crystallization experiments. *Crystal Growth and Design*. 2011;11(3):651-663. doi:10.1021/cg1013945

55. Bhattacharjee S. DLS and zeta potential - What they are and what they are not? *Journal of Controlled Release*. 2016;235:337-351. doi:10.1016/j.jconrel.2016.06.017

56. Filipe V, Hawe A, Carpenter JF, Jiskoot W. Analytical approaches to assess the degradation of therapeutic proteins. *TrAC Trends in Analytical Chemistry*. 2013;49:118-125. doi:10.1016/j.trac.2013.05.005

57. Lorber B. Analytical light scattering methods in molecular and structural biology: Experimental aspects and results. In: *International Conference on Pattern Recognition Applications and Methods*. SCITEPRESS; 2018:663-668. http://arxiv.org/abs/1810.00611.

58. Barnett CE. Some applications of wave-length turbimetry in the infrared. *The journal of Physical Chemistry*. 1942;46(1):69-75.

59. Introduction to OPTIM. :1-11. http://wolfson.huji.ac.il/purification/PDF/Protein

Characterization/AVACTA_Optim_Brochure.pdf.

60. Zetasizer Nano User Manual. 2013.

61. Lorber B, Fischer F, Bailly M, Roy H, Kern D. Protein analysis by dynamic light scattering: Methods and techniques for students. *Biochemistry and Molecular Biology Education*. 2012;40(6):372-382. doi:10.1002/bmb.20644

62. Miyazawa T, Blout ER. The infrared spectra of polypeptides in various conformations: Amide I and II bands. *Journal of the American Chemical Society*. 1961;83(3):712-719. doi:10.1021/ja01464a042

63. Krimm S, Bandekart J. Vibrational Spectroscopy and Conformation. *Advances in protein chemistry*. 1986;38:48109. doi:10.1016/S0065-3233(08)60528-8

64. Dong A, Huang P, Caughey WS. Protein secondary structures in water from second-derivative Amide I infrared spectra. *Biochemistry*. 1990;29(13):3303-3308. doi:10.1021/bi00465a022

65. Nicolet FT-IR User's Guide. 2004.

66. Eftink MR. Fluorescence techniques for studying protein structure. In: Suelter CH, ed. *Methods of Biochemical Analysis*. Vol 35. New York, NY: John Wiley & Sons, Inc.; 1991:127-205. http://www.ncbi.nlm.nih.gov/pubmed/2002770.

67. Royer CA. Approaches to teaching fluorescence spectroscopy. *Biophysical Journal*. 1995;68(3):1191-1195. doi:10.1016/S0006-3495(95)80295-X

68. Matheus S, Mahler HC, Friess W. A critical evaluation of Tm(FTIR) measurements of high-concentration IgG1 antibody formulations as a formulation development tool. *Pharmaceutical Research*. 2006;23(7):1617-1627. doi:10.1007/s11095-006-0283-9

69. Filoti DI, Shire SJ, Yadav S, Laue TM. Comparative study of analytical techniques for determining protein charge. *Journal of Pharmaceutical Sciences*. 2015;104(7):2123-2131. doi:10.1002/jps.24454

70. Hawe A, Sutter M, Jiskoot W. Extrinsic fluorescent dyes as tools for protein characterization. *Pharmaceutical Research*. 2008;25(7):1487-1499. doi:10.1007/s11095-007-9516-9

71. Delgado A V., González-Caballero F, Hunter RJ, Koopal LK, Lyklema J. Measurement and interpretation of electrokinetic phenomena. *Journal of Colloid and Interface Science*. 2007;309(2):194-224. doi:10.1016/j.jcis.2006.12.075

72. Connah MT, Kaszuba M, Morfesis A. High resolution zeta potential measurements: Analysis of multi-component mixtures. *Journal of Dispersion Science and Technology*. 2002;23(5):663-669. doi:10.1081/DIS-120015369

73. Minor M, Van Der Linde AJ, Van Leeuwen HP, Lyklema J. Dynamic aspects of electrophoresis and electroosmosis: A new fast method for measuring particle mobilities. *Journal of Colloid and Interface Science*. 1997;189(2):370-375. doi:10.1006/jcis.1997.4844

74. Amrhein S, Bauer KC, Galm L, Hubbuch J. Non-invasive high throughput approach for protein

hydrophobicity determination based on surface tension. *Biotechnology and Bioengineering*. 2015;112(12):2485-2494. doi:10.1002/bit.25677

75. Tate T. On the magnitude of a drop of liquid formed under different circumstances. *The Philosophical Magazine*. 1864;27(March):176-179.

76. Amrhein S, Suhm S, Hubbuch J. Surface tension determination by means of liquid handling stations. *Engineering in Life Sciences*. 2016;16:532-537. doi:10.1002/elsc.201500179

77. Baumgartner K, Großhans S, Schütz J, Suhm S, Hubbuch J. Prediction of salt effects on protein phase behavior by HIC retention and thermal stability. *Journal of Pharmaceutical and Biomedical Analysis*. 2016;128:216-225. doi:10.1016/j.jpba.2016.04.040

78. Maddux NR, Joshi SB, Volkin DB, Ralston JP, Middaugh CR. Multidimensional methods for the formulation of biopharmaceuticals and vaccines. *Journal of Pharmaceutical Sciences*. 2011;100(10):4171-4197. doi:10.1002/jps.22618

79. Mele K, Li R, Fazio J, Newman J. Quantifying the quality of the experiments used to grow protein crystals: the iQC suite. *Journal of Applied Crystallography*. 2014;47:1097-1106. doi:10.1107/S1600576714009728

80. Lekamge BMT, Sowmya A, Newman J. Prediction of protein X-ray crystallisation trail images time-courses. In: *Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods*. SCITEPRESS; 2017:663-668. doi:10.5220/0006246506630668

81. Kissmann J, Joshi SB, Haynes JR, Dokken L, Richardson C, Middaugh CR. H1N1 Influenze virus-like particles: Physical degradation pathways and identification of stabilizers. *Journal of pharmaceutical sciences*. 2011;100(2):634-645. doi:10.1002/jps.22304

82. Galm L, Amrhein S, Hubbuch J. Predictive approach for protein aggregation: Correlation of protein surface characteristics and conformational flexibility to protein aggregation propensity. *Biotechnology and Bioengineering*. 2017;114:1170–1183.

83. Oliveira AL. Biotechnology, big data and artificial intelligence. *Biotechnology Journal*. 2019:1800613. doi:10.1002/biot.201800613

84. García S, Luengo J, Herrera F. *Data Preprocessing in Data Mining*. Vol 72. (Kacprzyk J, Jain LC, eds.). New York, NY: Springer; 2015. doi:10.1007/978-3-319-10247-4

85. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics*. 2007;23(19):2507-2517. doi:10.1093/bioinformatics/btm344

86. Liu H, Motoda H. *Computational Methods of Feature Selection*. (Liu H, Motoda H, eds.). Boca Raton, FL: Taylor & Francis Group; 2007. doi:10.1007/s11042-018-6083-5

87. Guyon I, Elisseeff A. An introduction to variable and feature selection. *Journal of Machine Learning Research*. 2011;3:1157-1182. doi:10.1016/j.aca.2011.07.027

88. Xyntarakis M, Antoniou C. Data science and data visualization. In: Antoniou C, Dimitriou L, Pereire F, eds. *Mobility Patterns, Big Data and Transport Analytics*. Amsterdam: Elsevier Inc.;

2019:107-144. doi:10.1016/b978-0-12-812970-8.00006-3

89. Kueltzo LA, Ersoy B, Ralston JP, Middaugh CR. Derivative absorbance spectroscopy and protein phase diagrams as tools for comprehensive protein characterization: A bGCSF case study. *Journal of Pharmaceutical Sciences*. 2003;92(9):1805-1820. doi:10.1002/jps.10439

90. Joshi SB, Bhambhani A, Zeng Y, Middaugh CR. An empirical phase diagram - High-throughput screening approach to the characterization and formulation of biopharmaceuticals. In: Jameel F, Hershenson S, eds. *Formulation and Process Development Strategies for Manufacturing Biopharmaceuticals*. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2010:173-205.

91. Kim JH, Iyer V, Joshi SB, Volkin DB, Middaugh CR. Improved data visualization techniques for analyzing macromolecule structural changes. *Protein Science*. 2012;21:1540-1553. doi:10.1002/pro.2144

92. Tarca AL, Carey VJ, Chen X wen, Romero R, Drăghici S. Machine learning and its applications to biology. *PLoS computational biology*. 2007;3(6):e116. doi:10.1371/journal.pcbi.0030116

93. Marsland S. *Machine Learning: An Algorithmic Perspective*. (Herbich R, Graepel T, eds.). Boca Raton, FL: Taylor & Francis Group; 2009. http://blogs.msdn.com/b/msdntaiwan/archive/2015/03/05/machine-learning.aspx.

94. Chen X, Wang M, Zhang H. The use of classification trees for bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. 2011;1(1):55-63. doi:10.1002/widm.14

95. Angermueller C, Pärnamaa T, Parts L, Stegle O. Deep learning for computational biology. *Molecular systems biology*. 2016;12(7):878. doi:10.15252/msb.20156651

96. Zhao X-M, Li X, Chen L, Aihara K. Protein classification with imbalanced data. *Proteins: Structure, Function and Bioinformatics*. 2008;70(4):1125-1132. doi:10.1002/prot.21870

97. Michiels S, Koscielny S, Hill C. Prediction of cancer outcome with microarrys: A multiple random validation strategy. *The Lancet*. 2005;365(9458):488-492. doi:10.1016/S0140-6736(05)17866-0

98. Wilson J. Automated evaluation of crystallisation experiments. *Crystallography Reviews*. 2004;10(1):73-84. doi:10.1080/08893110410001664837

99. Fan HAO, Mark AE. Refinement of homology-based protein structures by molecular dynamics simulation techniques. *Protein Science*. 2004;13(1):211-220. doi:10.1110/ps.03381404.normally

100. Tropsha A. Best practices for QSAR model development, validation, and exploitation. *Molecular Informatics*. 2010;29(6-7):476-488. doi:10.1002/minf.201000061

101. Roy K, Supratik K, Das RN. Statistical methods in QSAR/QSPR. In: *A Primer on QSAR/QSPR Modeling*. Cham, Switzerland: Springer; 2015:37-59. doi:10.1007/978-3-319-17281-1

102. Piazza R. Interactions and phase transitions in protein solutions. *Current Opinion in Colloid and Interface Science*. 2000;5(1-2):38-43. doi:10.1016/S1359-0294(00)00034-0

103. McPherson A. Introduction to protein crystallization. *Methods*. 2004;34(3):254-265.

doi:10.1016/j.ymeth.2004.03.019

104. Mahler H, Friess W, Grauschopf U, Kiese S. Protein aggregation: pathways, induction factors and analysis. *Journal of the American Pharmaceutical Association*. 2009;98(9):2909-2934. doi:10.1002/jps.21566

105. Obrezanova O, Arnell A, Gómez de la Cuesta R, et al. Aggregation risk prediction for antibodies and its application to biotherapeutic development. *mAbs*. 2016;7(2):352-363. doi:10.1080/19420862.2015.1007828

106. Agrawal NJ, Kumar S, Wang X, Helk B, Singh SK, Trout BL. Aggregation in protein-based biotherapeutics: Computational studies and tools toiIdentify aggregation-prone regions. *Journal of Pharmaceutical Sciences*. 2011;100(12):5081-5095. doi:10.1002/jps

107. Huettmann H, Zich S, Berkemeyer M, Buchinger W. Design of industrial crystallization of interferon gamma: Phase diagrams and solubility curves. *Chemical Engineering Science*. 2015;126:341-348. doi:10.1016/j.ces.2014.12.018

108. Ng JD, Baird JK, Coates L, Garcia-ruiz JM, Hodge TA, Huang S. Large-volume protein crystal growth for neutron macromolecular crystallography. *Acta Crystallographica Section F Structural Biology Communications*. 2015;71(4):358-370. doi:10.1107/S2053230X15005348

109. Calero G, Cohen AE, Luft JR, Newman J, Snell EH. Identifying, studying and making good use of macromolecular crystals. *Acta Crystallographica Section F Structural Biology Communications*. 2014;70(8):993-1008. doi:10.1107/S2053230X14016574

110. Durbin SD, Feher G. Crystal growth studies of lysozyme as a model for protein crystallization. *Journal of Crystal Growth*. 1986;76:583-592.

111. Judge RA, Jacobs RS, Frazier T, Snell EH, Pusey ML. The effect of temperature and solution pH on the nucleation of tetragonal lysozyme crystals. *Biophysical Journal*. 1999;77(3):1585-1593. doi:10.1016/S0006-3495(99)77006-2

112. Moretti JJ, Sandler SI, Lenhoff AM. Phase equilibria in the lysozyme – ammonium sulfate – water system. *Biotechnology and Bioengineering*. 2000;70(5):498-506.

113. Smeller L. Pressure - temperature phase diagrams of biomolecules. *Biochimica et Biophysica Acta*. 2002;11(29):11-29.

114. Dumetz C, Chockla AM, Kaler EW, Lenhoff AM. Protein phase behavior in aqueous solutions: Crystallization, liquid-liquid phase separation, gels, and aggregates. *Biophysical Journal*. 2008;94(2):570-583. doi:10.1529/biophysj.107.116152

115. Baumgartner K, Galm L, Nötzold J, et al. Determination of protein phase diagrams by microbatch experiments: Exploring the influence of precipitants and pH. *International Journal of Pharmaceutics*. 2015;479(1):28-40. doi:10.1016/j.ijpharm.2014.12.027

116. Watanabe EO, Popova E, Alves E, Maurer G, Alcântara P De, Filho P. Phase equilibria for salt-induced lysozyme precipitation : Effect of salt type and temperature. *Fluid Phase Equilibria*.

2009;281:32-39. doi:10.1016/j.fluid.2009.03.021

117. Ahamed T, Esteban BN, Ottens M, et al. Phase behavior of an intact monoclonal antibody. *Biophysical journal*. 2007;93(2):610-619. doi:10.1529/biophysj.106.098293

118. Velev OD, Kaler EW, Lenhoff AM. Protein interactions in solution characterized by light and neutron scattering: Comparison of lysozyme and chymotrypsinogen. *Biophysical Journal*. 1998;75(6):2682-2697. doi:10.1016/S0006-3495(98)77713-6

119. Galm L, Morgenstern J, Hubbuch J. Manipulation of lysozyme phase behavior by additives as function of conformational stability. *International Journal of Pharmaceutics*. 2015;494(1):370-380. doi:10.1016/j.ijpharm.2015.08.045

120. Morgenstern J, Baumann P, Brunner C, Hubbuch J. Effect of PEG molecular weight and PEGylation degree on the physical stability of PEGylated lysozyme. *International Journal of Pharmaceutics*. 2017;519(1-2):408-417. doi:10.1016/j.ijpharm.2017.01.040

121. Bauer KC, Göbel M, Schwab M-L, Schermeyer M-T, Hubbuch J. Concentration-dependent changes in apparent diffusion coefficients as indicator for colloidal stability of protein solutions. *International Journal of Pharmaceutics*. 2016;511(1):276-287. doi:10.1016/j.ijpharm.2016.07.007

122. Schermeyer M-T, Sigloch H, Bauer KC, Oelschlaeger C, Hubbuch J. Squeeze flow rheometry as a novel tool for the characterization of highly concentrated protein solutions. *Biotechnology and Bioengineering*. 2016;113(3):576-587. doi:10.1002/bit.25834

123. Zeelen JP. Interpretation of crystallization drop results. In: Bergfors TM, ed. *Protein Crystallization - Techniques, Strategies, and Tips. A Laboratory Manual*. 2nd ed. International University Line; 2009:175-194.

124. Dumetz AC, Chockla AM, Kaler EW, Lenhoff AM. Effects of pH on protein – protein interactions and implications for protein phase behavior. *Biochimica et Biophysica Acta*. 2008;1784(4):600-610. doi:10.1016/j.bbapap.2007.12.016

125. Lu J, Wang X-J, Ching C-B. Effect of additives on lysozyme and chymotrypsinogen A. *Crystal growth & Design*. 2003;3(1):84-87. doi:10.1021/cg0200412

126. McPherson A, Cudney B. Searching for silver bullets: An alternative strategy for crystallizing macromolecules. *Journal of Structural Biology*. 2006;156(3):387-406. doi:10.1016/j.jsb.2006.09.006

127. Lorber B, Jenner G, Giege R. Effect of high hydrostatic pressure on nucleation and growth of protein crystals. *Journal of Crystal Growth*. 1996;158(1-2):103-117. doi:10.1016/0022-0248(95)00399-1

128. Forsythe E, Pusey ML. The effects of temperature and NaCl concentration on tetragonal lysozyme face growth rates. *Journal of Crystal Growth*. 1994;139(1-2):89-94. doi:10.1016/0022-0248(94)90032-9

129. Forsythe EL, Snell EH, Pusey ML, Drive C, Es B. Crystallization of chicken egg-white lysozyme

from ammonium sulfate. *Acta Crystallographica Section D*. 1997;53(6):795-797. doi:10.1107/S0907444997006896

130. Forsythe EL. Growth of (101) faces of tetragonal lysozyme crystals: measured growth-rate trends research papers. *Acta Crystallographica Section D*. 1999;55(5):1005-1011. doi:10.1107/S0907444999002899

131. Cheng Y, Lobo RF, Sandler SI, Lenhoff AM. Kinetics and equilibria of lysozyme precipitation and crystallization in concentrated ammonium sulfate solutions. *Biotechnology and Bioengineering*. 2006;94(1):177-188. doi:10.1002/bit

132. Barnett G V, Razinkov VI, Kerwin BA, et al. Specific-ion effects on the aggregation mechanisms and protein-protein interactions for anti-streptavidin immunoglobulin gamma-1. *The Journal of Physical Chemistry B*. 2015;119(18):5793-5804. doi:10.1021/acs.jpcb.5b01881

133. Judge RA, Forsythe EL, Pusey ML. Growth rate dispersion in protein crystal growth. *Crystal growth*. 2010;10(7):3164-3168. doi:10.1021/cg1002989

134. Liu Y, Wang X, Ching CB. Toward further understanding of lysozyme crystallization: Phase diagram, protein - protein interaction, nucleation kinetics, and growth kinetics. *Crystal growth & Design*. 2010;10(2):548-558. doi:10.1021/cg900919w

135. Falkner JC, Al-somali AM, Jamison JA, et al. Generation of size-controlled, submicrometer protein crystals. *Chemistry of Materials*. 2005;13(6):2679-2686. doi:10.1021/cm047924w

136. Ataka M. The growth of large single crystals of lysozyme. *Biopolymers,*. 1986;25(1):337-350.

137. Fiddis RW, Longman RA, Calvert PD. Crystal growth kinetics of globular proteins. *Journal of the Chemical Society, Faraday Transactions 1*. 1979;75:2753-2761. doi:10.1039/F19797502753

138. Forsythe EL, Snell EH, Malone CC, Pusey ML. Crystallization of chicken egg white lysozyme from assorted sulfate salts. *Journal of Crystal Growth*. 1999;196(2-4):332-343. doi:10.1016/S0022-0248(98)00843-4

139. Schermeyer M-T, Wöll AK, Kokke B, Eppink M, Hubbuch J. Characterization of highly concentrated antibody solution - A toolbox for the description of protein long-term solution stability. *mAbs*. 2017;9(7):1169-1185. doi:10.1080/19420862.2017.1338222

140. Rakel N, Baum M, Hubbuch J. Moving through three-dimensional phase diagrams of monoclonal antibodies. *Biotechnology and Bioengineering*. 2014;30(5):1103-1113. doi:10.1002/btpr.1947

141. Newman J, Xu J, Willis MC. Initial evaluations of the reproducibility of vapor-diffusion crystallization research papers. *Acta Crystallographica Section D*. 2007;63(7):826-832. doi:10.1107/S0907444907025784

142. Carter CW, Yin W. Quantitative analysis in the characterization and optimization of protein crystal growth. *Acta Crystallographica Section D*. 1994;50(4):572-590. doi:10.1107/S0907444994001228

143. Hui R, Edwards A. High-throughput protein crystallization. *Journal of Structural Biology*. 2003;142(1):154-161. doi:10.1016/S1047-8477(03)00046-7

144. Luft JR, Wolfley J, Jurisica I, Glasgow J, Fortier S, DeTitta GT. Macromolecular crystallization in a high throughput laboratory - the search phase. *Journal of Crystal Growth*. 2001;232(1-4):591-595.

145. Kröner F, Hubbuch J. Systematic generation of buffer systems for pH gradient ion exchange chromatography and their application. *Journal of Chromatography A*. 2013;1285:78-87. doi:10.1016/j.chroma.2013.02.017

146. Desbois S, Seabrook SA, Newman J. Some practical guidelines for UV imaging in the protein crystallization laboratory. *Acta Crystallographica Section F: Structural Biology and Crystallization Communications*. 2013;69(2):201-208. doi:10.1107/S1744309112048634

147. Leskovec J. Singluar-value decomposition. In: Leskovec J, Rajaraman A, Ullman JD, eds. *Mining of Massive Datasets*. 2nd ed. New York: Cambridge University Press; 2014:424.

148. Wetter LR, Deutsch HF. Immunological studies on egg white proteins: iv. immunochemical and physical studies of lysozyme. *Journal of Biological Chemistry*. 1951;192:237-242. https://pdfs.semanticscholar.org/1866/20dc210e0d52c6ec629c1690d9ef083a571b.pdf.

149. Green AA. Studies in the physical chemistry of proteins: The solubility of hemoglobin in solutions of chlorides and sulfates of varying concentration. *Journal of Biological Chemistry*. 1931;95(1):47-66. http://www.jbc.org/content/95/1/47.short.

150. Retailleau P, Ries-kautt M, Ducruix A. No salting-in of lysozyme chloride observed at low ionic strength over a large range of pH. *Biophysical Journal*. 1997;73(4):2156-2163. doi:10.1016/S0006-3495(97)78246-8

151. Schwierz N, Horinek D, Netz RR. Anionic and cationic Hofmeister effects on hydrophobic and hydrophilic surfaces. *Langmuir*. 2013;29(8):2602-2614. doi:10.1021/la303924e

152. Riès-Kautt MM, Ducruix AF. Relative effectiveness of vaious ions on solubility and crystal growth of lysozyme. *Journal of Biological Chemistry*. 1989;264(2):745-748.

153. Collins KD. Ions from the Hofmeister series and osmolytes: effects on proteins in solution and in the crystallization process. *Methods*. 2004;34(3):300-311. doi:10.1016/j.ymeth.2004.03.021

154. Okur HI, Hladilkova J, Rembert KB, et al. Beyond the Hofmeister series: Ion-specific effects on proteins and their biological functions. *Journal of Physical Chemistry*. 2017;121(9):1997-2014. doi:10.1021/acs.jpcb.6b10797

155. Durbin SD, Feher G. Protein crystallization. *Annual review of physical chemistry*. 1996;47(1):171-204. doi:10.1146/annurev.physchem.47.1.171

156. Howard SB, Twigg PJ, Baird JK, Meehan EJ. The solubility of hen egg-white lysozyme. *Journal of Crystal Growth*. 1988;90(1-3):94-104. doi:10.1016/0022-0248(88)90303-X

157. Burke MW, Leardi R, Judge RA, Pusey ML. Quantifying main trends in lysozyme nucleation: The effect of precipitant concentration, supersaturation, and impurities. *Crystal growth & Design*. 2001;1(4):333-337. doi:10.1021/cg0155088

158. McPherson A, Cudney B. Optimization of crystallization conditions for biological macromolecules. *Acta Crystallographica Section F Structural Biology Communications*. 2014;70(11):1445-1467. doi:10.1107/S2053230X14019670

159. Lin Y Bin, Zhu DW, Wang T, et al. An extensive study of protein phase diagram modification: Increasing macromolecular crystallizability by temperature screening. *Crystal Growth and Design*. 2008;8(12):4277-4283. doi:10.1021/cg800698p

160. Rakel N, Galm L, Bauer KC, Hubbuch J. Influence of macromolecular precipitants on phase behavior of monoclonal antibodies. *Biotechnology Progress*. 2015;31(1):145-153. doi:10.1002/btpr.2027

161. Berry IM, Dym O, Esnouf RM, et al. SPINE high-throughput crystallization, crystal imaging and recognition techniques: Current state, performance analysis, new technologies and future aspects. *Acta Crystallographica Section D: Biological Crystallography*. 2006;62(10):1137-1149. doi:10.1107/S090744490602943X

162. Wilson J. Automated classification of images from crystallisation experiments. In: Perner P, ed. *Advances in Data Mining. Applications in Medicine, Web Mining, Marketing, Image and Signal Mining. ICDM 2006. Lecture Notes in Computer Science*. Vol 4065. Springer Berlin / Heidelberg; 2006:459-473. doi:10.1007/11790853_36

163. Watts D, Cowtan K, Wilson J. Automated classification of crystallization experiments using wavelets and statistical texture characterization techniques. *Journal of Applied Crystallography*. 2008;41(1):8-17. doi:10.1107/s0021889807049308

164. Pusey ML, Aygün RS. Robotic image acquisition. In: Dress A, Linial M, Troyanskaya O, Vingron M, eds. *Data Analytics for Protein Crystallization*. Vol 25. Springer; 2017:57-81. doi:10.1007/978-3-319-58937-4

165. Pusey ML, Aygün RS. *Data Analytics for Protein Crystallization*. Vol 25. (Dress A, Linial M, Troyanskaya O, Vingron M, eds.). Springer; 2017. doi:10.1007/978-3-319-58937-4

166. Sigdel M, Pusey ML, Aygun RS. Real-time protein crystallization image acquisition and classification system. *Crystal Growth & Design*. 2013;13(7):2728-2736. doi:10.1021/cg3016029.Real-Time

167. Bruno AE, Charbonneau P, Newman J, et al. Classification of crystallization outcomes using deep convolutional neural networks. *PLoS ONE*. 2018;13(6):e0198883. doi:10.1371/journal.pone.0198883

168. Sigdel M, Dinc I, Sigdel MS, Dinc S, Pusey ML, Aygun RS. Feature analysis for classification of trace fluorescent labeled protein crystallization images. *BioData Mining*. 2017;10(1):14. doi:10.1186/s13040-017-0133-9

169. Yann ML-J, Tang Y. Learning deep convolutional neural networks for X-ray protein crystllization image analysis. In: *30th AAAI Conference on Artificial Intelligence*. ; 2016:1373-1379.

170. Hung J, Collins J, Weldetsion M, et al. Protein crystallization image classification with Elastic Net. In: *Medical Imaging 2014: Image Processing*. Vol 9034. International Society for Optics and Photonics; 2014:90341X. doi:10.1117/12.2043882

171. Cumbaa CA, Jurisica I. Protein crystallization analysis on the World Community Grid. *Journal of Structural and Functional Genomics*. 2010;11(1):61-69. doi:10.1007/s10969-009-9076-9

172. Buchala S, Wilson JC. Improved classification of crystallization images using data fusion and multiple classifiers. *Acta Crystallographica Section D*. 2008;64(8):823-833. doi:10.1107/S0907444908014273

173. Forsythe E, Achari A, Pusey ML. Trace fluorescent labeling for high-throughput crystallography. *Acta Crystallographica Section D: Biological Crystallography*. 2006;62(3):339-346. doi:10.1107/S0907444906000813

174. Pusey M, Barcena J, Morris M, Singhal A, Yuan Q, Ng J. Trace fluorescent labeling for protein crystallization. *Acta Crystallographica Section:F Structural Biology Communications*. 2015;71(7):806-814. doi:10.1107/S2053230X15008626

175. Kissick DJ, Wanapun D, Simpson GJ. Second-order nonlinear optical imaging of chiral crystals. *Annual Review of Analytical Chemistry*. 2011;4(1):419-437. doi:10.1146/annurev.anchem.111808.073722

176. Judge RA, Swift K, González C. An ultraviolet fluorescence-based method for identifying and distinguishing protein crystals. *Acta Crystallographica Section D: Biological Crystallography*. 2005;61(1):60-66. doi:10.1107/S0907444904026538

177. Dierks K, Meyer A, Oberthür D, Rapp G, Einspahr H, Betzel C. Efficient UV detection of protein crystals enabled by fluorescence excitation at wavelengths longer than 300 nm. *Acta Crystallographica Section F: Structural Biology and Crystallization Communications*. 2010;66(4):478-484. doi:10.1107/S1744309110007153

178. Madden JT, Dewalt EL, Simpson GJ. Two-photon excited UV fluorescence for protein crystal detection. *Acta Crystallographica Section D: Biological Crystallography*. 2011;67(10):839-846. doi:10.1107/S0907444911028253

179. Mele K, Lekamge BMT, Fazio VJ, Newman J. Using time courses to enrich the information obtained from images of crystallization trials. *Crystal Growth and Design*. 2014;14(1):261-269. doi:10.1021/cg4014569

180. Klijn ME, Hubbuch J. Application of empirical phase diagrams for multidimensional data visualization of high-throughput microbatch crystallization experiments. *Journal of Pharmaceutical Sciences*. 2018;107(8):2063-2069. doi:10.1016/j.xphs.2018.04.018

181. Uppuluri A. GLCM texture features. 2008. https://de.mathworks.com/matlabcentral/fileexchange/22187-glcm-texture-features.

182. Garcia V, Mollineda RA, Sanchez JS. Index of balanced accuracy: A performance measure for

skewed class distributions. In: Araujo H, Mendonca AM, Pinho AJ, Torres IM, eds. *Pattern Recognition and Image Analysis*. Springer Berlin / Heidelberg; 2009:441-448.

183. Kerwin BA, Remmele RLJ. Protect from light: photodegradation and protein biologics. *Journal of Pharmaceutical Sciences*. 2007;96(9):1468-1479. doi:10.1002/jps.20815

184. Bajaj S, Singla D, Sakhuja N. Stability testing of pharmaceutical products. *Journal of Applied Pharmaceutical Science*. 2012;02(03):129-138. doi:10.7324/JAPS.2012.2322

185. Roberts CJ, Das TK, Sahin E. Predicting solution aggregation rates for therapeutic proteins: Approaches and challenges. *International Journal of Pharmaceutics*. 2011;418(2):318-333. doi:10.1016/j.ijpharm.2011.03.064

186. Curtis R a., Lue L. A molecular approach to bioseparations: Protein-protein and protein-salt interactions. *Chemical Engineering Science*. 2006;61(3):907-923. doi:10.1016/j.ces.2005.04.007

187. Chari R, Jerath K, Badkar A V, Kalonia DS. Long- and short-range electrostatic interactions affect the rheology of highly concentrated antibody solutions. *Pharmaceutical Research*. 2009;26(12):2607-2618. doi:10.1007/s11095-009-9975-2

188. Kumar V, Dixit N, Zhou LL, Fraunhofer W. Impact of short range hydrophobic interactions and long range electrostatic forces on the aggregation kinetics of a monoclonal antibody and a dual-variable domain immunoglobulin at low and high concentrations. *International journal of pharmaceutics*. 2011;421(1):82-93. doi:10.1016/j.ijpharm.2011.09.017

189. Nicoud L, Cohrs N, Arosio P, Norrant E, Morbidelli M. Effect of polyol sugars on the stabilization of monoclonal antibodies. *Biophysical Chemistry*. 2015;197:40-46. doi:10.1016/j.bpc.2014.12.003

190. Minton AP. Recent applications of light scattering measurement in the biological and biopharmaceutical sciences. *Analytical Biochemistry*. 2016;501:4-22. doi:10.1016/j.ab.2016.02.007

191. Kramer RM, Shende VR, Motl N, Pace CN, Scholtz JM. Toward a molecular understanding of protein solubility: Increased negative surface charge correlates with increased solubility. *Biophysical Journal*. 2012;102(8):1907-1915. doi:10.1016/j.bpj.2012.01.060

192. Hirano A, Hamada H, Okubo T, Noguchi T, Higashibata H, Shiraki K. Correlation between thermal aggregation and stability of lysozyme with salts described by molar surface tension increment: An exceptional propensity of ammonium salts as aggregation suppressor. *Protein Journal*. 2007;26(6):423-433. doi:10.1007/s10930-007-9082-3

193. Maddux NR, Iyer V, Cheng W, et al. High throughput prediction of the long-term stability of pharmaceutical macromolecules from short-term multi-instrument spectroscopic data. *Pharmaceutical Biotechnology*. 2014;103(3):828-839. doi:10.1002/jps.23849

194. Lewis EN, Qi W, Kidder LH, Amin S, Kenyon SM, Blake S. Combined dynamic light scattering and raman spectroscopy approach for characterizing the aggregation of therapeutic proteins. *Molecules*. 2014;19(12):20888-20905. doi:10.3390/molecules191220888

195. Thiagarajan G, Semple A, James JK, Cheung JK, Shameem M. A comparison of biophysical

characterization techniques in predicting monoclonal antibody stability. *mAbs*. 2016;8(6):1088-1097. doi:10.1080/19420862.2016.1189048

196. Bauer KC, Göbel M, Schwab ML, Schermeyer MT, Hubbuch J. Concentration-dependent changes in apparent diffusion coefficients as indicator for colloidal stability of protein solutions. *International Journal of Pharmaceutics*. 2016;511(1):276-287. doi:10.1016/j.ijpharm.2016.07.007

197. Richards TW, Coombs LB. The surface tensions of water, methyl, ethyl and isobutyl alcohols, ethyl butyrate, benzene and toluene. *Journal of the American Chemical Society*. 1915;37(7):1656-1676. doi:10.1021/ja02172a002

198. Smoluchowski M von. Handbuch der Elektrizität und des Magnetismus. In: Greatz I, ed. *Band II*. Leipzig, Germany: Barth-Verlag; 1921:366.

199. Gautam R, Vanga S, Ariese F, Umapathy S. Review of multidimensional data processing approaches for Raman and infrared spectroscopy. *EPJ Techniques and Instrumentation*. 2015;2(1):8. doi:10.1140/epjti/s40485-015-0018-6

200. Bonincontro A, De Francesco A, Onori G. Influence of pH on lysozyme conformation revealed by dielectric spectroscopy. *Colloids and Surfaces B: Biointerfaces*. 1998;12(1):1-5. doi:10.1016/S0927-7765(98)00048-4

201. Bonincontro A, Risuleo G. Dielectric spectroscopy as a probe for the investigation of conformational properties of proteins. *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy*. 2003;59(12):2677-2684. doi:10.1016/S1386-1425(03)00085-4

202. Neergaard MS, Kalonia DS, Parshad H, Nielsen AD, Møller EH, Weert M Van De. Viscosity of high concentration protein formulations of monoclonal antibodies of the IgG1 and IgG4 subclass – Prediction of viscosity through protein – protein interaction measurements. *European Journal of Pharmaceutical Sciences*. 2013;49(3):400-410. doi:10.1016/j.ejps.2013.04.019

203. Kong J, Yu S. Fourier transform infrared spectroscopic analysis of protein secondary structures. *Acta Biochimica et Biophysica Sinica*. 2007;39(8):549-559. doi:10.1111/j.1745-7270.2007.00320.x

204. Groot CCM, Bakker HJ. Proteins take up water before unfolding. *Journal of Physical Chemistry Letters*. 2016;7(10):1800-1804. doi:10.1021/acs.jpclett.6b00708

205. Bychkova VE, Semisotnov GV, Balobanov VA, Finkelstein AV. The molten globule concept: 45 years later. *Biochemistry (Moscow)*. 2018;83(1):S22-S47. doi:10.1134/S0006297918140043

206. Blumlein A, McManus JJ. Reversible and non-reversible thermal denaturation of lysozyme with varying pH at low ionic strength. *Biochimica et Biophysica Acta - Proteins and Proteomics*. 2013;1834(10):2064-2070. doi:10.1016/j.bbapap.2013.06.001

207. Singh S, Singh J. Effect of polyols on the conformational stability and biological activity of a model protein lysozyme. *AAPS PharmSciTech*. 2003;4(3):101-109. doi:10.1208/pt040342

208. James S, McManus JJ. Thermal and solution stability of lysozyme in the presence of sucrose, glucose, and trehalose. *Journal of Physical Chemistry B*. 2012;116(34):10182-10188.

doi:10.1021/jp303898g

209. Robinson MJ, Matejtschuk P, Bristow AF, Dalby PA. Tm-Values and unfolded fraction can predict aggregation rates for granulocyte colony stimulating factor variant formulations but not under predominantly native conditions. *Molecular Pharmaceutics*. 2018;15(1):256-267. doi:10.1021/acs.molpharmaceut.7b00876

210. Tomczyńska-Mleko M, Kamysz E, Sikorska E, et al. Changes of secondary structure and surface tension of whey protein isolate dispersions upon pH and temperature. *Czech Journal of Food Science*. 2014;32(1):82-89. doi:10.17221/326/2012-CJFS

211. Zhang J. Protein-protein interactions in salt solutions. In: Cai W, ed. *Protein-Protein Interactions - Computational and Experimental Tools*. Rijeka, Croatia: INTECH Open Access Publisher; 2012:359-376. doi:10.5772/38056

212. Shih Y, Prausnitz JM, Blanch HW. Some characteristics of protein precipitation by salts. *Biotechnology and Bioengineering*. 1992;40(10):1155-1164. doi:10.1002/bit.260401004

213. Du H, Liu Z, Jennings R, Qian X. The effects of salt ions on the dynamics and thermodynamics of lysozyme unfolding. *Separation Science and Technology (Philadelphia)*. 2017;52(2):320-331. doi:10.1080/01496395.2016.1229336

214. Saluja A, Badkar A V, Zeng DL, Nema S, Kalonia DS. Application of high-frequency rheology measurements for analyzing protein – protein interactions in high protein concentration solutions using a model monoclonal antibody (IgG2). *Journal of Pharmaceutical Sciences*. 2006;95(9):1967-1983. doi:10.1002/jps.20663

215. Iyer P V., Ananthanarayan L. Enzyme stability and stabilization - Aqueous and non-aqueous environment. *Process Biochemistry*. 2008;43(10):1019-1032. doi:10.1016/j.procbio.2008.06.004

216. Carocho M, Barreiro MF, Morales P, Ferreira ICFR. Adding molecules to food, pros and cons: A review on synthetic and natural food additives. *Comprehensive Reviews in Food Science and Food Safety*. 2014;13(4):377-399. doi:10.1111/1541-4337.12065

217. *EU Commision Regulation No 257/2010*. European Union; 2010:19-27.

218. EFSA online library. http://efsa.onlinelibrary.wiley.com. Accessed November 22, 2018.

219. Mortensen A, Aguilar F, Crebelli R, et al. Re-evaluation of ammonium phosphatides (E 442) as a food additive. *EFSA Journal*. 2017;15(2):1-32. doi:10.2903/j.efsa.2017.4669

220. Younes M, Aggett P, Aguilar F, et al. Re-evaluation of stannous chlroide (E 512) as food additive. *EFSA Journal*. 2018;16(7):1-38. doi:10.2903/j.efsa.2018.5371

221. Mortensen A, Aguilar F, Crebelli R, et al. Re-evaluation of tara gum (E 417) as a food additive. *EFSA Journal*. 2017;15(2):1-37. doi:10.2903/j.efsa.2017.4669

222. Mortensen A, Aguilar F, Crebelli R, et al. Re-evaluation of glutamic acid (E 620), sodium glutamate (E 621), potassium glutamate (E 622), calcium glutamate (E 623), ammonium glutamate (E 624) and magnesium glutamate (E 625) as food additives. *EFSA Journal*. 2017;15(7):1-90.

doi:10.2903/j.efsa.2017.4910

223. Younes M, Aggett P, Aguilar F, et al. Re-evaluation of glycerol (E 422) as a food additive. *EFSA Journal*. 2018;16(1):1-64. doi:10.2903/j.efsa.2018.5088

224. Aguilar F, Crebelli R, Di Domenico A, et al. Re-evaluation of sucrose acetate isobutyrate (E 444) as a food additive. *EFSA Journal*. 2016;14(5):1-39. doi:10.2903/j.efsa.2016.4489

225. Silva C, Martins M, Jing S, Fu J, Cavaco-Paulo A. Practical insights on enzyme stabilization. *Critical Reviews in Biotechnology*. 2018;38(3):335-350. doi:10.1080/07388551.2017.1355294

226. Chirife J, Favetto GJ. Some physico-chemical basis of food preservation by combined methods. *Food Research International*. 1992;25(5):389-396. doi:10.1016/0963-9969(92)90158-2

227. Gekko K, Timasheff SN. Mechanism of protein stabilization by glycerol: Preferential hydration in glycerol-water mixtures. *Biochemistry*. 1981;20(16):4667-4676. doi:10.1021/bi00519a023

228. Kamerzell TJ, Esfandiary R, Joshi SB, Middaugh CR, Volkin DB. Protein – excipient interactions: Mechanisms and biophysical characterization applied to protein formulation development. *Advanced Drug Delivery Reviews*. 2011;63(13):1118-1159. doi:10.1016/j.addr.2011.07.006

229. Wang W, Nema S, Teagarden D. Protein aggregation - Pathways and influencing factors. *International Journal of Pharmaceutics*. 2010;390(2):89-99. doi:10.1016/j.ijpharm.2010.02.025

230. Manning MC, Liu J, Li T, Holcomb RE. *Rational Design of Liquid Formulations of Proteins*. Vol 112. 1st ed. (Donev R, ed.). Cambridge, MA, USA: Academic Press; 2018. doi:10.1016/bs.apcsb.2018.01.005

231. Goldberg DS, Bishop SM, Shah AU, Sathish HA. Formulation development of therapeutic monoclonal antibodies using high-throughput fluorescence and static light scattering techniques: Role of conformational and colloidal stability. *Journal of Pharmaceutical and Biomedical Analysis*. 2011;100(4):1306-1315. doi:10.1002/jps

232. Chi EY, Krishnan S, Randolph TW, Carpenter JF. Physical stability of proteins in aqueous Solution: Mechanism and driving forces in nonnative protein aggregation. *Pharmaceutical Research*. 2003;20(9):1325-1336. doi:10.1023/A:1025771421906

233. Kumar V, Chari R, Sharma VK, Kalonia DS. Modulation of the thermodynamic stability of proteins by polyols: Significance of polyol hydrophobicity and impact on the chemical potential of water. *International Journal of Pharmaceutics*. 2011;413(1-2):19-28. doi:10.1016/j.ijpharm.2011.04.011

234. Vagenende V, Yap MGS, Trout BL. Mechanisms of protein stabilization and prevention of protein aggregation by glycerol. *Biochemistry*. 2009;48(46):11084-11096. doi:10.1021/bi900649t

235. Abbas SA, Sharma VK, Patapoff TW, Kalonia DS. Characterization of antibody-polyol interactions by static light scattering: Implications for physical stability of protein formulations. *International Journal of Pharmaceutics*. 2013;448(2):382-389. doi:10.1016/j.ijpharm.2013.03.058

236. Ajito S, Iwase H, Takata SI, Hirai M. Sugar-mediated stabilization of protein against chemical or

thermal denaturation. *Journal of Physical Chemistry B*. 2018;122(37):8685-8697. doi:10.1021/acs.jpcb.8b06572

237. Lee JC, Timasheff SN. The stabilization of proteins by sucrose. *The Journal of Biological Chemistry*. 1981;256(14):7193-7201. doi:10.1016/S0006-3495(85)83932-1

238. Spiegel T. Whey protein aggregation under shear conditions - effects of lactose and heating temperature on aggregate size and structure. *International Journal of Food Science and Technology*. 1999;34(5-6):523-531. doi:10.1046/j.1365-2621.1999.00309.x

239. Wang W. Instability, stabilization, and formulation of liquid protein pharmaceuticals. *International Journal of Pharmaceutics*. 1999;185(2):129-188. doi:10.1016/S0378-5173(99)00152-0

240. Li S, Schöneich C, Borchardt RT. Chemical instability of protein pharmaceuticals: Mechanisms of oxidation and strategies for stabilization. *Biotechnology and Bioengineering*. 1995;48(5):490-500. doi:10.1002/bit.260480511

241. Cleland JL, Powel MF, Shire SJ. The development of stable protein formulations: A close look at protein aggregation, deamidation, and oxidation. *Critical reviews in therapeutic drug carrier systems*. 1993;10(4):307-377. doi:10.4319/lo.2013.58.2.0489

242. Levine RL, Mosoni L, Berlett BS, Stadtman ER. Methionine residues as endogenous antioxidants in proteins. *Proceedings of the National Academy of Sciences*. 1996;93(26):15036-15040. doi:10.1073/pnas.93.26.15036

243. Zhou HX. Interactions of macromolecules with salt ions: An electrostatic theory for the Hofmeister effect. *Proteins: Structure, Function and Genetics*. 2005;61(1):69-78. doi:10.1002/prot.20500

244. Zhang Y, Cremer PS. The inverse and direct Hofmeister series for lysozyme. *Proceedings of the National Academy of Sciences*. 2009;106(36):15249-15253. doi:10.1073/pnas.0907616106

245. Shelef LA. Antimicrobial effects of lactates: A review. *Journal of Food Protection*. 1994;57(5):445-450. doi:10.4315/0362-028X-57.5.445

246. Bis RL, Mallela KMG. Antimicrobial preservatives induce aggregation of interferon alpha-2a: The order in which preservatives induce protein aggregation is independent of the protein. *International Journal of Pharmaceutics*. 2014;472(1-2):356-361. doi:10.1016/j.ijpharm.2014.06.044

247. Hutchings RL, Singh SM, Cabello-Villegas J, Mallela KMG. Effect of antimicrobial preservatives on partial protein unfolding and aggregation. *Journal of pharmaceutical sciences*. 2013;102(2):365-376. doi:10.1002/jps.23362.Effect

248. Klijn ME, Hubbuch J. Correlating multidimensional short-term empirical protein properties to long-term physical stability data via empirical phase diagrams. *International Journal of Pharmaceutics*. 2019;560:166-176. https://doi.org/10.1016/j.ijpharm.2019.02.006.

249. Khan MV, Ishtikhar M, Rabbani G, Zaman M, Abdelhameed AS, Khan RH. Polyols (glycerol and ethylene glycol) mediated amorphous aggregate inhibition and secondary structure restoration of metalloproteinase-conalbumin (ovotransferrin). *International Journal of Biological*

Macromolecules. 2017;94:290-300. doi:10.1016/j.ijbiomac.2016.10.023

250. Liu W, Bratko D, Prausnitz JM, Blanch HW. Effect of alcohols on aqueous lysozyme-lysozyme interactions from static light-scattering measurements. *Biophysical Chemistry*. 2004;107(3):289-298. doi:10.1016/j.bpc.2003.09.012

251. MacDonald GA, Lanier TC, Swaisgood HE, Hamann DD. Mechanism for stabilization of fish actomyosin by sodium lactate. *Journal of Agricultural and Food Chemistry*. 1996;44(1):106-112. doi:10.1021/jf940698y

252. Tang H-M, Ou W-B, Zhou H-M. Effects of lactic acid and NaCl on creatine kinase from rabbit muscle. *International Journal of Biochemistry and Cell Biology*. 2001;33(11):1064-1070. doi:10.1016/S1357-2725(01)00079-6

253. Resch JJ, Daubert CR, Foegeding EA. The effects of acidulant type on the rheological properties of beta-lactoglobulin gels and powders derived from these gels. *Food Hydrocolloids*. 2005;19(5):851-860. doi:10.1016/j.foodhyd.2004.10.034

254. Farías ME, Pilosof AMR. The influence of acid type on self-assembly, rheological and textural properties of caseinomacropeptide. *International Dairy Journal*. 2016;55:17-25. doi:10.1016/j.idairyj.2015.11.003

255. Folzer E, Diepold K, Bomans K, et al. Selective oxidation of methionine and tryptophan residues in a therapeutic IgG1 molecule. *Journal of Pharmaceutical Sciences*. 2015;104(9):2824-2831. doi:10.1002/jps.24509

256. Ji JA, Zhang B, Cheng W, Wang YJ. Methionine, tryptophan, and histidine oxidation in a model protein, PTH: Mechanisms and stabilization. *Journal of pharmaceutical sciences*. 2009;98(12):4485-4500. doi:10.1002/jps

257. Lam XM, Yang JY, Cleland JL. Antioxidants for prevention of methionine oxidation in recombinant monoclonal antibody HER2. *Journal of Pharmaceutical Sciences*. 1997;86(11):1250-1255. doi:10.1021/js970143s

258. Chu JW, Yin J, Brooks BR, et al. A comprehensive picture of non-site specific oxidation of methionine residues by peroxides in protein pharmaceuticals. *Journal of Pharmaceutical Sciences*. 2004;93(12):3096-3102. doi:10.1002/jps.20207

259. Manning MC, Chou DK, Murphy BM, Payne RW, Katayama DS. Stability of protein pharmaceuticals: An update. *Pharmaceutical Research*. 2010;27(4):544-575. doi:10.1007/s11095-009-0045-6

260. Thakkar S V., Joshi SB, Jones ME, et al. Excipients differentially influence the conformational stability and pretransition dynamics of two IgG1 monoclonal antibodies. *Journal of Pharmaceutical Sciences*. 2012;101(9):3062-3077. doi:10.1002/jps.23187

261. Baynes BM, Wang DIC, Trout BL. Role of arginine in the stabilization of proteins against aggregation. *Biochemistry*. 2005;44(12):4919-4925. doi:10.1021/bi047528r

262. Zhang Y, Cremer PS. Chemistry of Hofmeister anions and osmolytes. *Annual Review of Physical Chemistry*. 2010;61(1):63-83. doi:10.1146/annurev.physchem.59.032607.093635

263. Schwierz N, Horinek D, Sivan U, Netz RR. Reversed Hofmeister series - The rule rather than the exception. *Current Opinion in Colloid and Interface Science*. 2016;23:10-18. doi:10.1016/j.cocis.2016.04.003

264. Lomakin A, Teplow DB, Benedek GB. Quasielastic light scattering for protein assembly studies. In: Sigurdsson E, ed. *Amyloid Proteins*. Vol 299. Totowa, NJ, USA: Humana Press Inc.; 2005:153-174.

265. Bull HB, Breese K. Surface tension of amino acid solutions: A hydrophobicity scale of the amino acids residues. *Archives of Biochemistry and Biophysics*. 1974;161(2):665-670. doi:10.1016/0003-9861(74)90352-X

266. Kushnir N, Streatfield SJ, Yusibov V. Virus-like particles as a highly efficient vaccine platform: Diversity of targets and production systems and advances in clinical development. *Vaccine*. 2012;31(1):58-83. doi:10.1016/j.vaccine.2012.10.083

267. Chackerian B. Virus-like particles: Flexible platforms for vaccine development. *Expert Reviews of Vaccines*. 2007;6(3):381-390. doi:10.1586/14760584.6.3.381

268. McAleer WJ, Buynak EB, Maigetter RZ, Wampler DE, Miller WJ, Hilleman MR. Human hepatitis B vaccine from recombinant yeast. *Nature*. 1984;307(5947):178-180.

269. Bryan JT, Buckland B, Hammond J, Jansen KU. Prevention of cervical cancer: Journey to develop the first human papillomavirus virus-like particle vaccine and the next generation vaccine. *Current Opinion in Chemical Biology*. 2016;32:34-47. doi:10.1016/j.cbpa.2016.03.001

270. Pumpens P, Grens E. HBV core particles as a carrier for B cell/T cell epitopes. *Intervirology*. 2001;44(2-3):98-114. doi:10.1159/000050037

271. Kratz PA, Böttcher B, Nassal M. Native display of complete foreign protein domains on the surface of hepatitis B virus capsids. *PNAS*. 1999;96(5):1915-1920. doi:10.1073/pnas.96.5.1915

272. Mohsen MO, Zha L, Cabral-Miranda G, Bachmann MF. Major findings and recent advances in virus-like particle (VLP)-based vaccines. *Seminars in Immunology*. 2017;34:123-132. doi:10.1016/j.smim.2017.08.014

273. Nielsen CM, Vekemans J, Lievens M, Kester KE, Regules JA, Ockenhouse CF. RTS ,S malaria vaccine efficacy and immunogenicity during Plasmodium falciparum challenge is associated with HLA genotype. *Vaccine*. 2018;36(12):1637-1642. doi:10.1016/j.vaccine.2018.01.069

274. Milich DR, Sallberg M, Maruyama T. The humoral immune response in acute and chronic hepatitis B virus infection. *Springer Seminars in immunopathology*. 1995;17(2-3):149-166. doi:https://doi.org/10.1007/BF00196163

275. Fehr T, Skrastine D, Pumpens P, Zinkernagel RM. T cell-independent type I antibody response against B cell epitopes expressed repetitively on recombinant virus particles. *Proceedings of the*

National Academy of Sciences of the United States of America. 1998;95(16):9477-9481. doi:10.1073/pnas.95.16.9477

276. Klamp T, Schumacher J, Huber G, et al. Highly specific auto-antibodies again claudin-18 isoform 2 induced by a chimeric HBcAg virus-like particle vaccine kill tumor cells and inhibit the growth of lung metastases. *Cancer research*. 2011;71(12):516-527. doi:10.1158/0008-5472.CAN-10-2292

277. Pumpens P, Ultich RG, Sasnauskas K, Kazaks A, Ose V, Grens E. Construction of novel vaccines on the basis of the virus-like particles: Hepatitis B virus proteins as vaccine carriers. In: Khudaykov YE, ed. *Medicinal Protein Engineering*. Boca Raton, FL, USA: CRC Press Taylor & Francis Group; 2008:205-248.

278. Ding X, Liu D, Booth G, Gao W, Lu Y. Virus-like particle engineering: From rational design to versatile applications. *Biotechnology Journal*. 2018;13(5):1-7. doi:10.1002/biot.201700324

279. Jegerlehner A, Tissot A, Lechner F, et al. A molecular assembly system that renders antigens of choice highly repetitive for induction of protective B cell responses. *Vaccine*. 2002;20(25-26):3104-3112. doi:10.1016/S0264-410X(02)00266-9

280. Buckland BC. The process development challenge for a new vaccine. *Nature medicine*. 2005;11(4):16-19. doi:10.1038/nm1218

281. Vicente T, Roldão A, Peixoto C, Carrondo MJT, Alves PM. Large-scale production and purification of VLP-based vaccines. *Journal of Invertebrate Pathology*. 2011;107:S42-S48. doi:10.1016/j.jip.2011.05.004

282. Priddy TS, Middaugh CR. Stabilization and formulation of vaccines. In: Wen EP, Ellis R, Pujar NS, eds. *Vaccines Development and Manufacturing*. 1st ed. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2015:237-261.

283. Ladd Effio C, Hubbuch J. Next generation vaccines and vectors: Designing downstream processes for recombinant protein-based virus-like particles. *Biotechnology Journal*. 2015;10(5):715-727. doi:10.1002/biot.201400392

284. Hämmerling F, Ladd Effio C, Andris S, Kittelmann J, Hubbuch J. Investigation and prediction of protein precipitation by polyethylene glycol using quantitative structure-activity relationship models. *Journal of Biotechnology*. 2017;241:87-97. doi:10.1016/j.jbiotec.2016.11.014

285. Hämmerling F, Lorenz-Cristea O, Baumann P, Hubbuch J. Strategy for assessment of the colloidal and biological stability of H1N1 influenza A viruses. *International Journal of Pharmaceutics*. 2017;517(1-2):80-87. doi:10.1016/j.ijpharm.2016.11.058

286. Lua LHL, Connors NK, Sainsbury F, Chuan YP, Wibowo N, Middelberg APJ. Bioengineering virus-like particles as vaccines. *Biotechnology and Bioengineering*. 2014;111(3):425-440. doi:10.1002/bit.25159

287. Vicente T, Mota JPB, Peixoto C, Alves PM, Carrondo MJT. Rational design and optimization of downstream processes of virus particles for biopharmaceutical applications: Current advances.

Biotechnology Advances. 2011;29(6):869-878. doi:10.1016/j.biotechadv.2011.07.004

288. Mellado MCM, Mena JA, Lopes A, et al. Impact of physicochemical parameters on in vitro assembly and disassembly kinetics of recombinant triple-layered rotavirus-like particles. *Biotechnology and Bioengineering*. 2009;104(4):674-686. doi:10.1002/bit.22430

289. Schijven JF, Hassanizadeh SM. Removal of viruses by soil passage: Overview of modeling, processes, and parameters. *Environmental Science and Technology*. 2010;30(1):49-127. doi:10.1080/10643380091184174

290. Ghanem N, Kiesel B, Kallies R, Harms H, Chatzinotas A, Wick LY. Marine phages as tracers: Effects of size, morphology, and physicochemical surface properties on transport in a porous medium. *Environmental Science and Technology*. 2016;50(23):12816-12824. doi:10.1021/acs.est.6b04236

291. Penrod SL, Olson TM, Grant SB. Deposition kinetics of two viruses in packed beds of quartz granular media. *Langmuir*. 1996;7463(18):5576-5587. doi:10.1021/la950884d

292. Lošdosfer Božič A, Podgornik R. pH Dependence of charge multipole moments in proteins. *Biophysical journal*. 2017;113(7):1454-1465. doi:10.1016/j.bpj.2017.08.017

293. Johnston MA, Søndergaard CR, Nielsen JE. Integrated prediction of the effect of mutations on multiple protein characteristics. *Proteins: Structure, Function, and Bioinformatics*. 2011;79(1):165-178. doi:10.1002/prot.22870

294. Krieger E, Nielsen JE, Spronk CAEM, Vriend G. Fast empirical pKa prediction by Ewald summation. *Journal of Molecular Graphics and Modelling*. 2006;25(4):481-486. doi:10.1016/j.jmgm.2006.02.009

295. Anandakrishnan R, Aguilar B, Onufriev A V. H++ 3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic Acids Research*. 2012;40(W1):537-541. doi:10.1093/nar/gks375

296. Martí-Renom MA, Stuart AC, Fiser A, Sánchez R, Melo F, Sali A. Comparative protein structure modeling of genes and genomes. *Annual Review of Biophysics and Biomolecular Structure*. 2000;29(1):291-325. doi:10.1201/9780203911327.ch7

297. Venselaar H, Joosten RP, Vroling B, et al. Homology modelling and spectroscopy, a never-ending love story. *European Biophysics Journal*. 2010;39(4):551-563. doi:10.1007/s00249-009-0531-0

298. Forster MJ. Molecular modelling in structural biology. *Micron*. 2002;33(4):365-384. doi:10.1016/S0968-4328(01)00035-X

299. Krieger E, Joo K, Lee J, et al. Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: Four approaches that performed well in CASP8. *Proteins: Structure, Function, and Bioinformatics*. 2009;77(S9):114-122. doi:10.1002/prot.22570

300. Best RB, Buchete NV, Hummer G. Are current molecular dynamics force fields too helical? *Biophysical Journal*. 2008;95(1):7-9. doi:10.1529/biophysj.108.132696

301. Lange OF, Van Der Spoel D, De Groot BL. Scrutinizing molecular mechanics force fields on the submicrosecond timescale with NMR Data. *Biophysical Journal*. 2010;99(2):647-655. doi:10.1016/j.bpj.2010.04.062

302. Beauchamp KA, Lin YS, Das R, Pande VS. Are protein force fields getting better? A systematic benchmark on 524 diverse NMR measurements. *Journal of Chemical Theory and Computation*. 2012;8(4):1409-1414. doi:10.1021/ct2007814

303. Martín-García F, Papaleo E, Gomez-Puertas P, Boomsma W, Lindorff-Larsen K. Comparing molecular dynamics force fields in the essential subspace. *PLoS ONE*. 2015;10(3):1-16. doi:10.1371/journal.pone.0121114

304. Freddolino PL, Arkhipov AS, Larson SB, Mcpherson A, Schulten K. Molecular dynamics simulations of the complete satellite tobacco mosaic virus. *Structure*. 2006;14(3):437-449. doi:10.1016/j.str.2005.11.014

305. Roberts JA, Kuiper MJ, Thorley BR, Smooker PM, Hung A. Investigation of a predicted N-terminal amphipathic a-helix using atomistic molecular dynamics simulation of a complete prototype poliovirus virion. *Journal of Molecular Graphics and Modelling*. 2012;38:165-173. doi:10.1016/j.jmgm.2012.06.009

306. Zhao G, Perilla JR, Yufenyuy EL, et al. Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics. *Nature*. 2013;497(7451):643-646. doi:10.1038/nature12162.Mature

307. Larsson DSD, Liljas L, Spoel D Van Der. Virus capsid dissolution studied by microsecond molecular dynamics simulations. *PLoS Computational Biology*. 2012;8(5):1-8. doi:10.1371/journal.pcbi.1002502

308. Mansour AA, Sereda Y V, Yang J, Ortoleva PJ. Prospective on multiscale simulation of virus-like particles: Application to computer-aided vaccine design. *Vaccine*. 2015;33(44):5890-5896. doi:10.1016/j.vaccine.2015.05.099

309. Arkhipov A, Freddolino PL, Schulten K. Stability and dynamics of virus capsids described by coarse-grained modeling. *Structure*. 2006;14(12):1767-1777. doi:10.1016/j.str.2006.10.003

310. Reddy T, Shorthouse D, Parton DL, et al. Nothing to sneeze at: A dynamic and integrative computational model of an influenza A virion. *Structure*. 2015;23(3):584-597. doi:10.1016/j.str.2014.12.019

311. Reddy T, Sansom MSP, Reddy T, Sansom MSP. The role of the membrane in the structure and biophysical robustness of the dengue virion. *Structure*. 2016;24(3):375-382. doi:10.1016/j.str.2015.12.011

312. Ayton GS, Voth GA. Multiscale computer simulation of the immature HIV-1 virion. *Biophysical Journal*. 2010;99(9):2757-2765. doi:10.1016/j.bpj.2010.08.018

313. Cheluvaraja S, Ortoleva P. Thermal nanostructure: An order parameter multiscale ensemble

approach. *Journal of Chemical Physics*. 2010;132(7):1-9. doi:10.1063/1.3316793

314. Joshi H, Cheluvaraja S, Somogyi E, Brown DR, Ortoleva P. A molecular dynamics study of loop fluctuation in human papillomavirus type 16 virus-like particles: A possible indicator of immunogenicity. *Vaccine*. 2011;29(51):9423-9430. doi:10.1016/j.vaccine.2011.10.039

315. Miao Y, Johnson JE, Ortoleva PJ. All-atom multiscale simulation of cowpea chlorotic mottle virus capsid swelling. *Journal of Physical Chemistry*. 2011;114(34):11181-11195. doi:10.1021/jp102314e.All-atom

316. Machado MR, González HC, Pantano S. MD simulations of viruslike particles with Supra CG solvation afforadble to desktop computers. *Journal of Chemical Theory and Computation*. 2017;13(10):5106-5116. doi:10.1021/acs.jctc.7b00659

317. Zhang L, Tang R, Bai S, et al. Molecular energetics in the capsomere of virus-like particle revealed by molecular dynamics simulations. *The Journal of Physical Chemistry B*. 2013;117(18):5411-5421. doi:10.1021/jp311170w

318. Lua LHL, Fan Y, Chang C, Connors NK, Middelberg APJ. Synthetic biology design to display an 18 kDa rotavirus large antigen on a modular virus-like particle. *Vaccines*. 2015;33(44):5937-5944. doi:10.1016/j.vaccine.2015.09.017

319. Zlotnick A, Tan Z, Selzer L. One protein, at least three structures, and many functions. *Structure*. 2013;21(1):6-8. doi:10.1016/j.str.2012.12.003.One

320. Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A. ExPASy : the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Research*. 2003;31(13):3784-3788. doi:10.1093/nar/gkg563

321. Corbett JCW, Connah MT, Mattison K. Advances in the measurement of protein mobility using laser Doppler electrophoresis - the diffusion barrier technique. *Electrophoresis*. 2011;32(14):1787-1794. doi:10.1002/elps.201100108

322. Moore DS, McCabe GP, Craig BA. Describing distributions with numbers. In: Burke S, Scanlan-Rohrer A, eds. *Introduction to the Practice of Statistics*. 6th ed. New York, NY, USA: W.H. Freeman and Company; 2009.

323. Fiser A, Šali A. Modeller: Generation and refinement of homology-based protein structure models. *Method in Enzymology*. 2003;374:461-491. doi:10.1016/S0076-6879(03)74020-8

324. Vriend G. WHAT IF: A molecular modeling and drug design program. *Journal of molecular graphics*. 1990;8(1):52-56. doi:10.1016/0263-7855(90)80070-V

325. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*. 1983;79(2):926-935. doi:10.1063/1.445869

326. Berman HM, Westbrook J, Feng Z, et al. The protein data bank. *Nucleic Acids Research*. 2000;28(1):235-242. doi:10.1093/nar/28.1.235

327. Alexander CG, Jürgens MC, Shepherd DA, Freund SM V, Ashcroft AE. Thermodynamic origins of protein folding, allostery, and capsid formation in the human hepatitis B virus core protein. *PNAS*. 2013;110(30):2782-2791. doi:10.1073/pnas.1308846110

328. Krieger E, Dunbrack RL, Hooft RW, Krieger B. Assignment of protonation states in protein and ligands: Combining pKa prediction with hydrogen bonding network optimization. In: Baron R, ed. *Computational Drug Discovery and Design*. Vol 819. New York, NY, USA: Springer; 2012:405-421. doi:10.1007/978-1-61779-465-0_25

329. Wang J, Cieplak P, Kollman PA. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *Journal of Computational Chemistry*. 2000;21(12):1049-1074. doi:10.1002/1096-987X(200009)21:12<1049::AID-JCC3>3.0.CO;2-F

330. Duan Y, Wu C, Chowdhury S, et al. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *Journal of Computational Chemistry*. 2003;24(16):1999-2012. doi:10.1002/jcc.10349

331. Krieger E, Vriend G. New ways to boost molecular dynamic simulations. *Journal of computational chemistry*. 2015;36(13):996-1007. doi:10.1002/jcc.23899

332. Krieger E, Darden T, Nabuurs SB, Finkelstein A, Vriend G. Making optimal use of empirical energy functions: Force-field parameterization in crystal space. *Proteins: Structure, Function, and Bioinformatics*. 2004;57(5):678-683. doi:10.1002/prot.20251

333. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LG. A smooth particle mesh Ewald method A smooth particle mesh Ewald method. *The Journal of Chemical Physics*. 1995;103(19):8577-8593. doi:10.1063/1.470117

334. Berendsen HJC, Postma JPM, Gunsteren WF Van, Dinola A, Haak JR. Molecular dynamics with coupling to an external bath Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*. 1984;81(8):3684-3690. doi:10.1063/1.448118

335. Grubmüller H, Tavan P. Multiple time step algorithms for molecules dynamics simulations of proteins: How good are they? *Journal of computational chemistry*. 1998;19(13):1534-1552. doi:10.1002/(SICI)1096-987X(199810)19:13<1534::AID-JCC10>3.0.CO;2-I

336. Hess B, Bekker H, Berendsen HJC, Fraaije JGEM. LINCS : A linear constraint solver for molecular simulations. *Journal of computational chemistry*. 1997;18(12):1463-1472. doi:10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H

337. Krieger E, Nabuurs SB, Vriend G. Homology modeling. In: Bourne PE, Weissig H, eds. *Methods of Biochemical Analysis*. Hoboken, NJ, USA: Wiley-Liss, Inc; 2003:509-524.

338. Xu D, Zhang Y. Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. *Biophysical journal*. 2011;101(10):2525-2534. doi:10.1016/j.bpj.2011.10.024

339. Spronk CAEM, Linge JP, Hilbers CW, Vuister GW. Improving the quality of protein structures derived by NMR spectroscopy. *Journal of Biomolecular NMR*. 2002;22(3):281-289. doi:10.1023/A:1014971029663

340. Knapp B, Frantal S, Cibena M, Schreiner W, Bauer P. Is an intuitive convergence definition of molecular dynamics simulations solely based on the root mean square deviation possible? *Journal of Computational biology*. 2011;18(8):997-1005. doi:10.1089/cmb.2010.0237

341. Schumacher J, Bacic T, Staritzbichler R, et al. Enhanced stability of a chimeric hepatitis B core antigen virus‐like‐particle (HBcAg‐VLP) by a C‐terminal linker‐hexahistidine‐peptide. *Journal of Nanobiotechnology*. 2018;16(1):39. doi:10.1186/s12951-018-0363-0

342. Sippl MJ. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge based prediction of local structures in globular proteins. *Journal of Molecular Biology*. 1990;213(4):859-883. doi:10.1016/S0022-2836(05)80269-4

343. Lošdorfer Božič A, Siber A, Podgornik R. How simple can a model of an empty viral capsid be? Charge distributions in viral capsids. *Journal of biological physics*. 2012;38(4):657-671. doi:10.1007/s10867-012-9278-4

344. Rodgers JL, Nicewander WA. Thirteen ways to look at the correlation coefficient. *The American Statistician*. 1988;42(1):59-66. doi:10.1080/00031305.1988.10475524

345. Andersen A. Important Background. In: Knight V, ed. *Modern Methods for Robust Regression*. Thousand Oaks: Sage Publications, Inc.; 2008. doi:10.4135/9781412985109

Appendix A

Supplementary Material Chapter 3

A.1 Internal feature correlation

Strong internal correlation between image features is not desired, as two (or more) correlated features may over represent a phase behavior property compared to a phase behavior property represented by one feature. The Pearson correlation coefficient is used as a measure of internal correlation strength. A threshold of 0.850 for either positive or negative correlation is used in this work. Supplementary Table A1 shows the results for each of the extracted image features. With the set threshold the precipitation diameter and the precipitation intensity are strongly correlated, indicated in red. This led to the removal of the precipitation intensity from the feature dataset.

Table A1: Pearson correlation coefficient matrix between all extracted image features. Red highlighted numbers indicate a violation of the set threshold (-0.850 < x < 0.850).

		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	Length crystal									
2	Number of crystals	0.826								
3	Crystal onset time	0.334	0.289							
4	Crystal growth time	0.802	0.625	0.475						
5	Precipitation diameter	0.319	0.327	0.266	0.212					
6	Precipitation intensity	0.297	0.319	0.240	0.207	0.952				
7	Precipitation onset time	0.269	0.295	0.429	0.218	0.794	0.752			
8	Precipitation growth time	0.136	0.168	0.223	0.089	0.528	0.536	0.581		
9	IQR ratio crystal length-width	0.531	0.478	0.441	0.489	0.184	0.184	0.253	0.108	

A.2 Feature reproducibility

Each stored condition was made in duplicate. To show the level of reproducibility of the phase behavior, an overview of 6 image features for each of the duplicate conditions for lysozyme at pH 9 in the presence of ammonium sulfate is shown in Supplementary Figure A1.

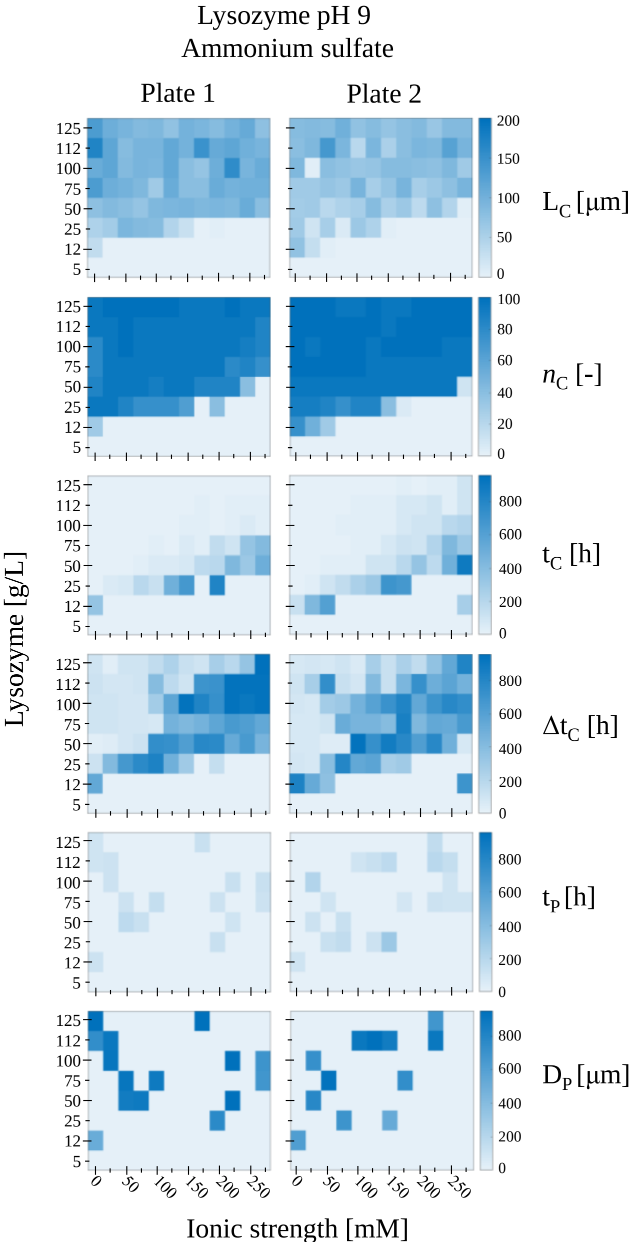


Figure A1: Untreated extracted image features for replicates plates containing lysozyme at pH 9 in the presence of ammonium sulfate. The crystal length (L_C) in µm, number of crystals (n_C), crystal onset time (t_C) in hours, crystal growth time (Δt_C) in hours, precipitation onset time (t_P) in hours and precipitation diameter (D_P) in µm are colored according to their value. White indicates the minimum value and blue indicates the maximal value.

A.3 Visible and UV light images

UV light imaging is used to distinguish between protein and non-protein crystallization and/or precipitation based on UV light signal. In Supplementary Figure A2 example images for visible and UV light are shown for 5 different observations.

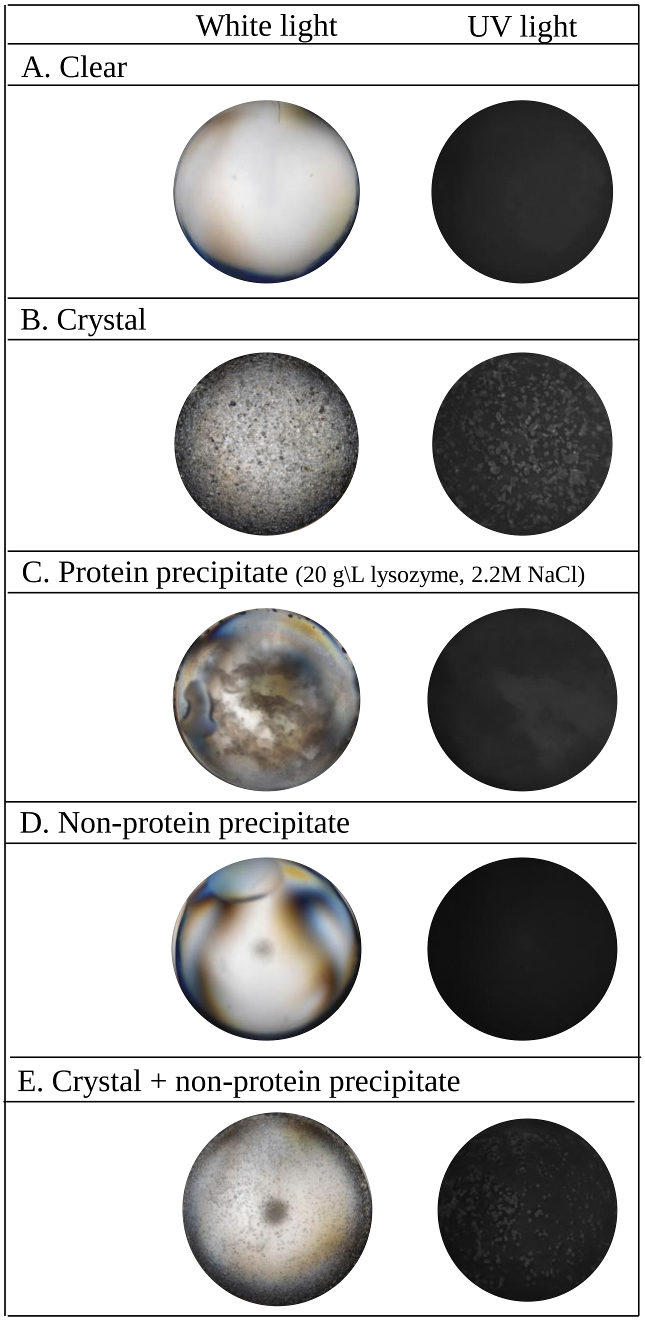

White light
UV light
A. Clear
B. Crystal
C. Protein precipitate (20 g\L lysozyme, 2.2M NaCl)
D. Non-protein precipitate
E. Crystal + non-protein precipitate

Figure A2: Examples of visible light (first column) and UV light (second column) images for (A) clear solution, (B) crystallized protein, (C) precipitated protein, (D) non-protein precipitation, and (E) crystallized protein in the presence of non-protein precipitation.

Appendix B

Supplementary Material Chapter 4

B.1 Example class images

An overview of example images can be found in Figure B1.

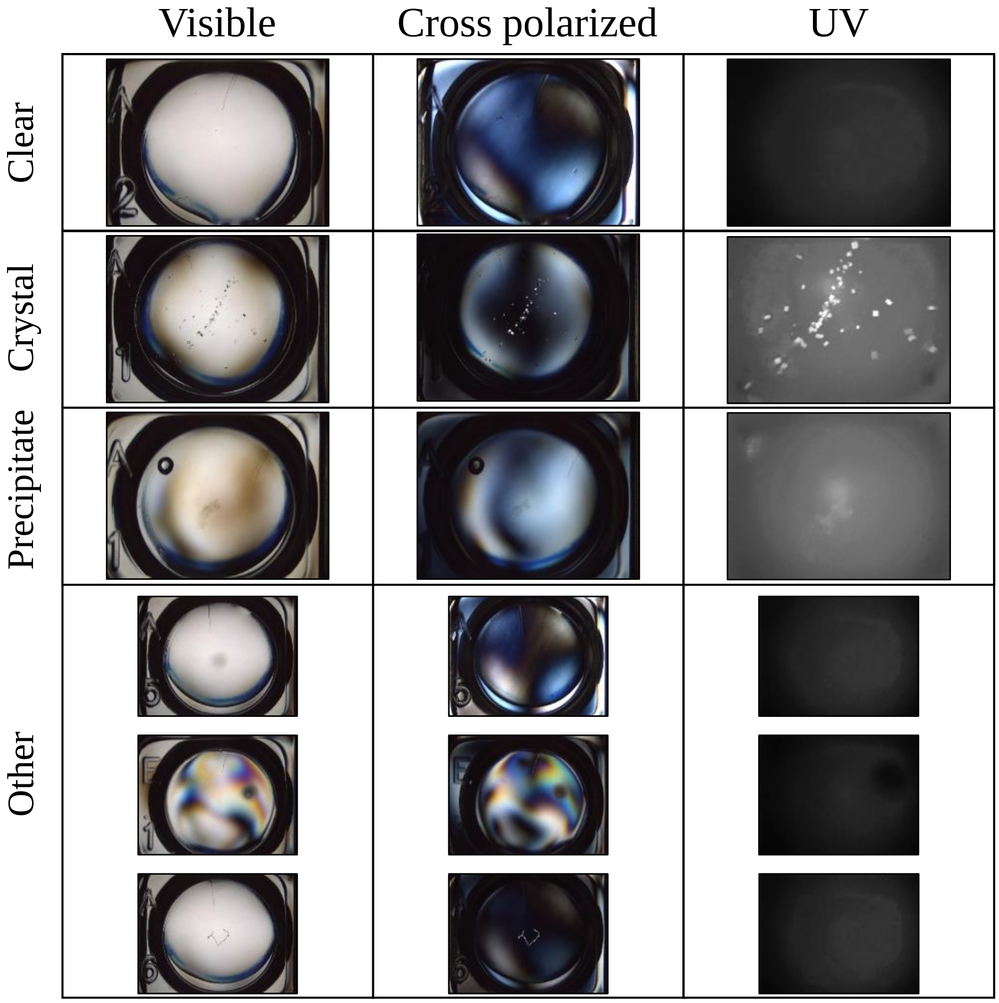


Figure B1: Examples of visible light (first column), cross polarized light (second column), and UV light (third column) images for the employed classification classes “clear”, “crystal”, “precipitate”, and “other”. For the class “other”, three examples are shown.

Figure B1 shows examples of obtained images for each class. The class “other” shows three different examples, where the first example shows light precipitation, which does not light up in the corresponding UV image. The second example shows microbial instability and the third example shows the presence of a hair/dust particle in the formulation well.

B.2 Image schedule

The employed image schedules are listed in Table B1.

Table B1: Detailed overview of the imaging schedule used per storage time, for each light source.

14 days		**30 days**	
Visible light	Cross polarized and UV light	Visible light	Cross polarized and UV light
Day 0 to 1: Every 2 hours Day 1 to 2: Every 4 hours Day 2 to 6: Every 6 hours Day 6 to 14: Every 24 hours	Hours: 0 and 4 Day: 1, 4, 7, 10, and 14	Day 0 to 2: Every 2 hours Day 2 to 6: Every 6 hours Day 6 to 30: Every 24 hours	Hours: 0 and 22 Days: 5, 10, 20, and 30

The imaging schedules in Table B1 resulted in a total of 57 images per well for 14 days of storage and 67 images per well for 30 days of storage.

B.3 Image masking

Images obtained with visible light and cross polarized light were cropped with black pixels (also known as a mask) to remove irrelevant data from the images. The well walls and part of the plate were considered to be irrelevant. Masking was also used for visible light and cross polarized light images after edge detection. The second mask was used to remove sharp lines a mask leaves in an image, which are often recognized as edges. An example of the first and second mask is shown in Figure B2, for all light sources.

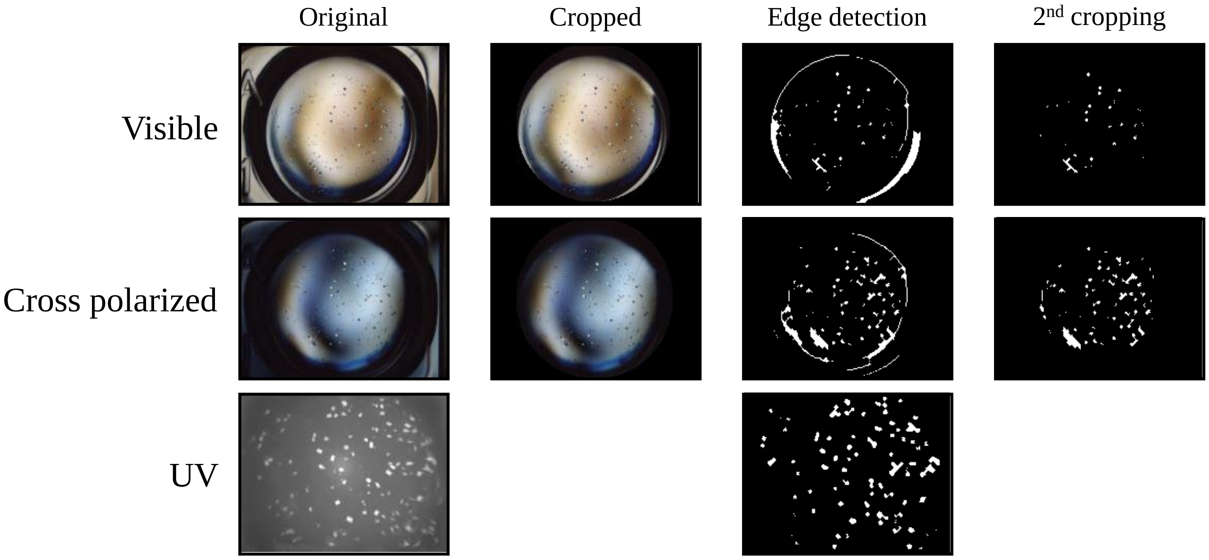


Figure B2: Example images during image feature extraction. First column, "Original": original image for each light source (visible, cross polarized, and UV light); second column, "Cropped": cropped visible light and cross polarized light images; third column, "Edge detection": edge detection results for all light images using the image in the previous column; fourth column, "2nd cropping": removal of the mask wall for visible light and cross polarized light images.

Figure B2 shows different images throughout the image extraction workflow, for each light source. The original image shows that visible light and cross polarized light capture the liquid formulation and a part of the plate well. UV light images are obtained with a larger zoom, 7x instead of 2.5x. Therefore, a close up of what is seen in visible light and cross polarized light images is visible in UV light images. This is also the reason why cropping is not needed for UV light images. The cropped version of visible light and cross polarized light images is seen in the second column, “Cropped”. The cropped image is used for edge detection, the result is shown in the third column (“Edge detection”). Edge detection with visible light and cross polarized light images resulted in a mask edge as well. This was resolved by masking the result of edge detection once more. The fourth column (“2nd cropping”) shows the final image.

B.4 Overview all image features

Table B2 lists all image features that were extracted for each light source image.

Table B2: Overview of all image features extracted for each light source.

Histogram features	**Description**
ER	Entropy of the red color level image
EG	Entropy of the green color level image
EB	Entropy of the blue color level image
EGray	Entropy of the gray image
MR	Mean pixel level of the red color level image
MG	Mean pixel level of the green color level
MB	Mean pixel level of the blue color level
MGray	Mean pixel level of the gray image
VR	Variance of the image histogram for the red color level
VG	Variance of the image histogram for the green color level
VB	Variance of the image histogram for the blue color level
VGray	Variance of the image histogram for the gray image
SDR	Standard deviation of the image histogram for the red color level
SDG	Standard deviation of the image histogram for the green color level
SDB	Standard deviation of the image histogram for the blue color level
SDGray	Standard deviation of the image histogram for the gray image
SkR	Skewness of the image histogram for the red color level
SkG	Skewness of the image histogram for the green color level
SkB	Skewness of the image histogram for the blue color level
SkGray	Skewness of the image histogram for the gray image
KtR	Kurtosis of the image histogram for the red color level
KtG	Kurtosis of the image histogram for the green color level
KtB	Kurtosis of the image histogram for the blue color level
KtGray	Kurtosis of the image histogram for the gray image
GLMC features	**Description**
autoc	Autocorrelation
contr	Contrast
corm	Correlation (MATLAB)
corrp	Correlation
cprom	Cluster prominence
cshad	Cluster shade
dissi	Dissimilarity
energy	Energy
entro	Entropy
homom	Homogeneity (MATLAB)
homop	Homogeneity
maxpr	Maximum probability
sosvh	Sum of squares (variance)
savgg	Sum of average
svarh	Sum of variance
senth	Sum of entropy
dvarh	Difference variance
denth	Difference entropy
inf1h	Information measure of correlation1
infh2	Information measure of correlation2

indnc idmnc	Inverse difference normalized Inverse difference moment normalized
Blob features	**Description**
BlobArea BlobAmount	The total area of all identified blobs The total number of identified blobs
Features over time	**Description**
Intensity Intensity difference	Total sum of pixel intensity of the gray scale image, for each time point Total sum of pixel intensity of the obtained difference image t_0 - t_x, for each time point

B.5 Overview of multidimensional protein phase diagram features

Table B3 lists all image features that were extracted to construct the multidimensional protein phase diagrams.

Table B3: Image-based features extracted from visible, cross polarized, and UV light images for the automated construction of a multidimensional protein phase diagram.

Blob	**Description**	**Extraction class**
Aggregation area	Sum of the area of all blobs	Crystal/Precipitate
Crystal area	Mean area of all blobs	Crystal
Crystal count	Total number of blobs	Crystal
Crystal length	Mean blob major axis length	Crystal
Intensity	**Description**	**Extraction class**
Intensity difference	Total pixel intensity of the difference between the last and first image (tend – t0)	Crystal/Precipitate
Start intensity	Total pixel intensity of the first image	Crystal/Precipitate
Time-dependent	**Description**	**Extraction class**
Growth time	Time point at which a plateau in the intensity change was identified	Crystal/Precipitate

The extraction class is indicated to show for which predicted class the features were extracted.

B.6 Number of trees determination

The number of trees that are needed to construct a random forest classification model was determined by investigating the effect of the number of trees on the Out-of-Bag classification error. This was investigated with the Vis+CP+UV+Time feature set and the results are shown in Figure B3.

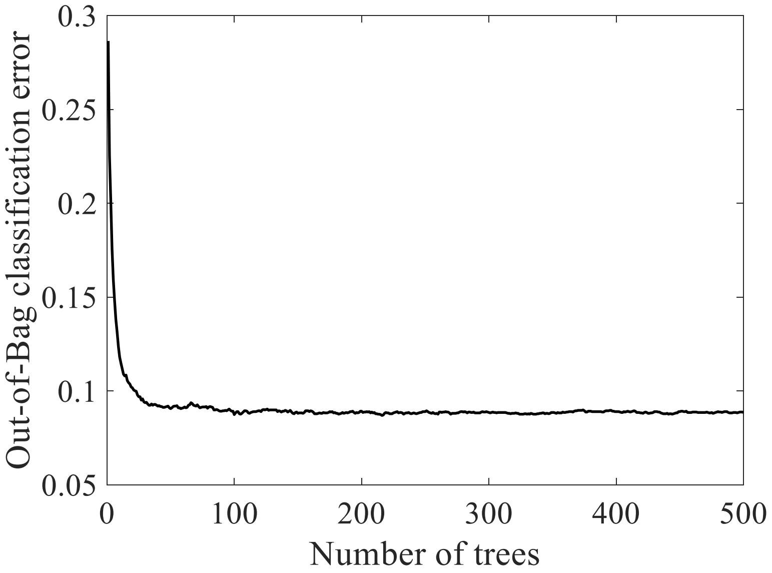


Figure B3: Out-of-Bag classification error over the number of trees.

Based on the results shown in Figure B3, the number of trees was set to 100.

B.7 Selected features per feature set

Image feature selection was performed for each fold and each image feature set during internal 10-fold cross validation. Between different folds for the same image feature set, similar features were selected. The selected features are listed in Table B4 per feature set.

Table B4: Overview of features selected during 10-fold cross validation for all evaluated feature sets.

	Vis	**Vis+Time**	**Vis+CP**	**Vis+UV**	**Vis+CP+UV**	**Vis+CP+UV+Time**
1	EB_Vis	EG_Vis	ER_Vis	ER_Vis	ER_Vis	ER_Vis
2	MR_Vis	MR_Vis	EG_Vis	EG_Vis	EG_Vis	EG_Vis
3	VR_Vis	VR_Vis	EB_Vis	EB_Vis	EB_Vis	EB_Vis
4	KtR_Vis	AmountBlobs_Vis	MR_Vis	BlobArea_Vis	VR_Vis	VR_Vis
5	BlobArea_Vis	contr_Vis	VR_Vis	AmountBlobs_Vis	VB_Vis	BlobArea_Vis
6	AmountBlobs_Vis	energ_Vis	VB_Vis	contr_Vis	BlobArea_Vis	AmountBlobs_Vis
7	contr_Vis	inf2h_Vis	KtR_Vis	dissi_Vis	AmountBlobs_Vis	contr_Vis
8	energ_Vis	EndInt_Vis	BlobArea_Vis	inf2h_Vis	contr_Vis	corrm_Vis
9	inf2h_Vis	EndInt_CP	AmountBlobs_Vis	ER_UV	corrm_Vis	dissi_Vis
10		EndInt_UV	contr_Vis	MR_UV	cprom_Vis	energ_Vis
11			energ_Vis	SkR_UV	cshad_Vis	entro_Vis
12			inf2h_Vis	KtR_UV	dissi_Vis	inf2h_Vis
13			EB_CP	BlobArea_UV	entro_Vis	EndInt_Vis
14			VR_CP	AmountBlobs_UV	inf2h_Vis	ER_CP
15			VB_CP	contr_UV	AmountBlobs_CP	AmountBlobs_CP
16			SkB_CP	corrm_UV	ER_UV	EndInt_CP
17			BlobArea_CP	cshad_UV	MR_UV	ER_UV
18			AmountBlobs_CP		SkR_UV	MR_UV
19			corrm_CP		KtR_UV	VR_UV
20					BlobArea_UV	SkR_UV
21					AmountBlobs_UV	KtR_UV
22					contr_UV	BlobArea_UV
23					corrm_UV	AmountBlobs_UV
24					cshad_UV	contr_UV
25					energ_UV	corrm_UV
26					inf1h_UV	cshad_UV
27					inf2h_UV	inf2h_UV
28						EndInt_UV

B.8 Evaluation parameters per class per feature set

Recall, precision, accuracy, and balanced accuracy were obtained for each class separately during 10-fold cross validation. Table B5 lists all evaluation parameters per class for each feature set.

Table B5: Overview of all evaluation parameters (recall, precision, accuracy, and balanced accuracy) for all evaluated feature sets, shown per classification class. The values are given as mean ± standard deviation, obtained by each fold during 10-fold cross validation.

Feature set: Vis				
	Recall [%]	**Precision [%]**	**Accuracy[%]**	**Balanced accuracy [%]**
Clear	95.0 ± 1.0	81.0 ± 1.5	82.8 ± 1.5	88.0 ± 0.9
Precipitate	63.4 ± 12.2	81.7 ± 7.7	97.8 ± 0.4	72.5 ± 5.8
Crystal	64.1 ± 4.9	85.6 ± 3.5	89.1 ± 1.1	74.8 ± 2.8
Non-protein	29.5 ± 8.8	53.9 ± 9.5	91.0 ± 1.0	41.7 ± 7.8
Feature set: Vis + Time				
	Recall [%]	**Precision [%]**	**Accuracy[%]**	**Balanced accuracy [%]**
Clear	95.5 ± 1.8	83.2 ± 1.1	84.9 ± 1.2	89.3 ± 1.0
Precipitate	64.3 ± 11.7	83.9 ± 7.3	98.0 ± 0.6	74.1 ± 8.4
Crystal	67.1 ± 3.6	88.7 ± 3.6	90.3 ± 1.0	77.9 ± 2.5
Non-protein	45.6 ± 4.2	65.5 ± 9.9	92.5 ± 1.1	55.5 ± 6.0
Feature set: Vis + CP				
	Recall [%]	**Precision [%]**	**Accuracy[%]**	**Balanced accuracy [%]**
Clear	96.3 ± 1.0	81.8 ± 1.1	84.1 ± 1.3	89.0 ± 0.9
Precipitate	61.3 ± 10.4	85.6 ± 7.6	97.9 ± 0.5	73.4 ± 7.0
Crystal	68.9 ± 3.7	90.1 ± 2.7	90.9 ± 1.0	79.5 ± 2.4
Non-protein	27.7 ± 7.5	54.2 ± 10.1	91.0 ± 1.1	41.0 ± 8.0
Feature set: Vis + UV				
	Recall [%]	**Precision [%]**	**Accuracy[%]**	**Balanced accuracy [%]**
Clear	97.1 ± 0.8	87.9 ± 1.5	89.7 ± 1.2	92.5 ± 0.8
Precipitate	82.8 ± 14.2	94.2 ± 5.8	99.1 ± 0.7	88.5 ± 8.8
Crystal	86.5 ± 2.8	97.2 ± 1.5	96.3 ± 0.7	91.9 ± 1.5
Non-protein	42.6 ± 7.7	69.3 ± 8.7	92.9 ± 1.0	56.0 ± 7.7
Feature set: Vis + CP + UV				
	Recall [%]	**Precision [%]**	**Accuracy[%]**	**Balanced accuracy [%]**
Clear	97.1 ± 0.9	88.0 ± 1.2	89.8 ± 1.0	92.6 ± 0.7
Precipitate	84.2 ± 12.7	94.5 ± 5.5	99.1 ± 0.6	89.4 ± 7.5
Crystal	87.1 ± 2.2	96.6 ± 1.9	96.3 ± 0.7	91.8 ± 1.6
Non-protein	41.3 ± 6.1	70.1 ± 7.9	92.8 ± 0.9	55.7 ± 6.4

B.9 Correlation coefficient matrix

The first step during feature selection was filtering based on internal correlation with the Pearson correlation coefficient. A correlation coefficient matrix is often used to visualize the correlation coefficient between variables in a dataset. The correlation coefficient matrix consisting of all features extracted for the Vis+CP+UV+Time dataset is shown in Figure B4.

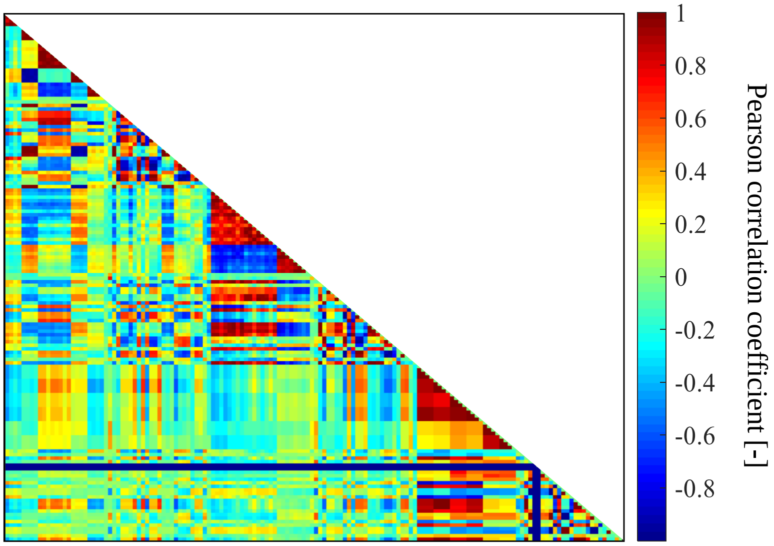


Figure B4: Correlation coefficient matrix, based on the Pearson correlation coefficient, for all 150 extracted image features. The color bar indicates the Pearson correlation coefficient value, where +1 (red) indicates a strong positive linear dependency and -1 (blue) indicates a strong negative linear dependency between variables.

All image features with a linear dependency higher than +0.950 or lower than -0.950 have been removed from the dataset during feature selection.

B.10 Feature importance

The second step during feature selection was selecting features based on their relative importance to the classification problem in question. The feature importance per evaluated feature is shown in Figure B5.

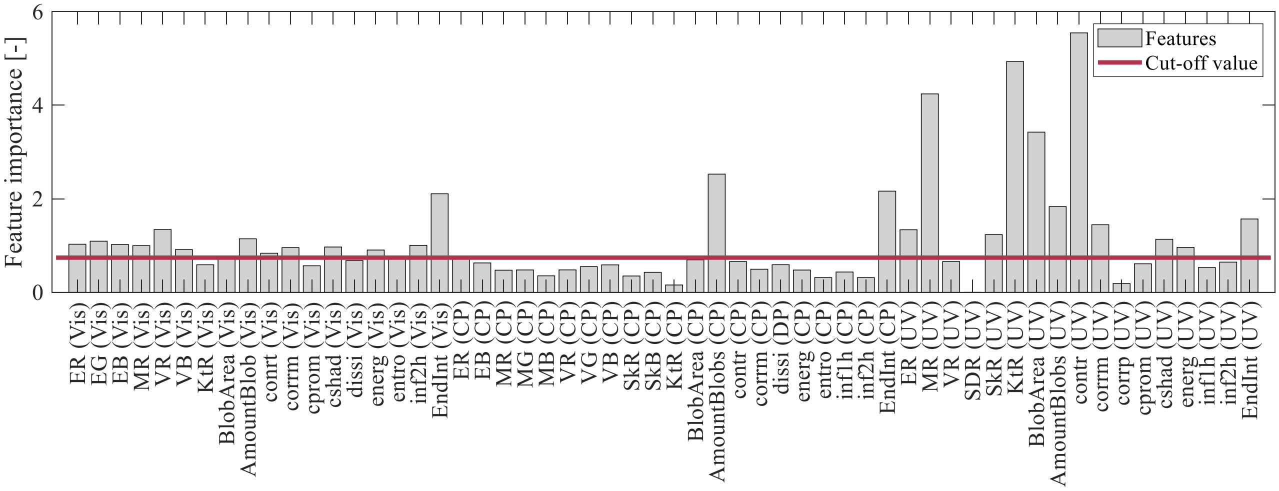


Figure B5: Feature importance of all features that remained after feature removal based on internal correlation. The red line indicates the cut-off value at the 50th percentile of the feature importance.

All image features depicted in Figure B5 with a feature importance below the 50th percentile value were removed during feature selection.

B.11 Confusion matrix Vis+CP+UV+Time

Table B6 lists all confusion matrices obtained during 10-fold cross validation using the Vis+CP+UV+Time feature set.

Table B6: List of all confusion matrices obtained during 10-fold cross validation with the Vis+CP+UV+Time feature set.

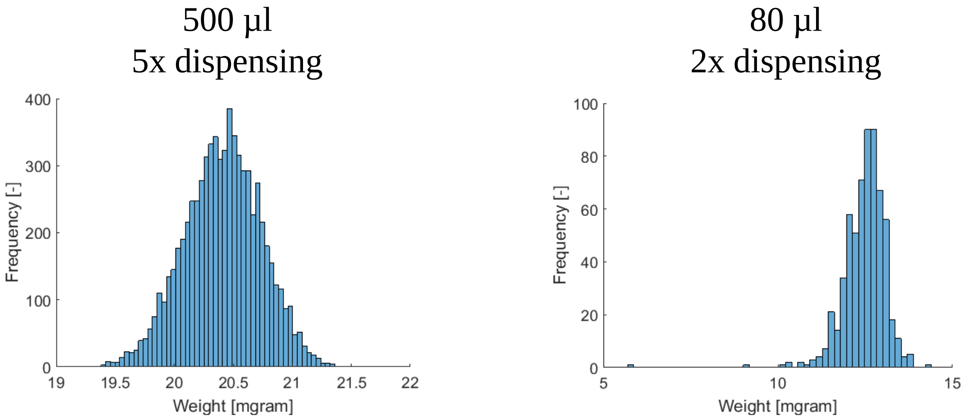
	Fold #1	**Predicted class**			
		Clear	Precipitate	Crystal	Non-protein
True class	Clear	290	0	2	4
	Precipitate	1	19	0	0
	Crystal	14	0	95	1
	Non-protein	12	0	0	32
	Fold #2	**Predicted class**			
		Clear	Precipitate	Crystal	Non-protein
True class	Clear	289	0	2	6
	Precipitate	3	16	1	0
	Crystal	9	0	101	0
	Non-protein	11	0	0	33
	Fold #3	**Predicted class**			
		Clear	Precipitate	Crystal	Non-protein
True class	Clear	291	0	1	5
	Precipitate	2	18	0	0
	Crystal	12	2	95	1
	Non-protein	15	0	0	29
	Fold #4	**Predicted class**			
		Clear	Precipitate	Crystal	Non-protein
True class	Clear	285	2	2	8
	Precipitate	2	17	0	1
	Crystal	14	1	94	1
	Non-protein	16	0	1	27
	Fold #5	**Predicted class**			
		Clear	Precipitate	Crystal	Non-protein
True class	Clear	289	1	2	5
	Precipitate	1	16	2	0
	Crystal	17	1	89	3
	Non-protein	21	0	0	24
	Fold #6	**Predicted class**			
		Clear	Precipitate	Crystal	Non-protein
True class	Clear	292	0	1	4
	Precipitate	4	14	1	0
	Crystal	15	0	93	2
	Non-protein	14	0	2	28
	Fold #7	**Predicted class**			
		Clear	Precipitate	Crystal	Non-protein
True class	Clear	285	0	3	9
	Precipitate	1	18	0	0
	Crystal	13	1	94	2
	Non-protein	17	0	0	27
	Fold #8	**Predicted class**			
		Clear	Precipitate	Crystal	Non-protein
True class	Clear	288	0	1	8
	Precipitate	1	18	0	0
	Crystal	16	0	94	0
	Non-protein	14	0	0	30
	Fold #9	**Predicted class**			
		Clear	Precipitate	Crystal	Non-protein
True class	Clear	282	0	3	12
	Precipitate	1	17	2	0
	Crystal	9	2	97	1
	Non-protein	10	0	0	34
	Fold #10	**Predicted class**			
		Clear	Precipitate	Crystal	Non-protein
True class	Clear	284	0	2	10
	Precipitate	6	10	4	0
	Crystal	9	1	98	2
	Non-protein	13	0	0	31

Appendix C

Supplementary Material Chapter 5

C.1 Stalagmometer

To decrease experimental time and sample volume, the high-throughput stalagmometer setup was adjusted to a lower sample volume and fewer repeat dispenses. The results of an evaluation run with water is shown in Figure C1.


500 µl
5x dispensing

80 µl
2x dispensing

Mean [gram]	0.0204
Standard dev [gram]	$3.22 \cdot 10^{-4}$
Relative st.dev.	1.58%
Number of drops	7079

Mean [gram]	0.0125
Standard dev [gram]	$6.18 \cdot 10^{-4}$
Relative st.dev.	4.95%
Number of drops	613

Figure C1: Difference between the (a) original and (b) adjusted experimental setup for stalagmometer measurements. The standard deviation (standard dev) and relative standard deviation (relative st. dev.) are used as evaluation parameters. In addition, the mean weight and number of drops is listed.

Due to a lower dispense volume, the drop mass decreased from 20.4 milligram to 12.5 milligram. The number of drops decreased from 7079 to 613 due to the decrease in sample volume and number of repeat dispenses. Less dispensed drops resulted in a 2.5-fold shorter experimental time. The adjustments resulted in 2-fold increase of standard deviation ($3.22 \cdot 10^{-4}$ to $6.18 \cdot 10^{-4}$ gram) and a 3.4% higher relative standard deviation.

C.2 Internal correlation

Strong internal correlation between empirical protein properties is not desired, as two (or more) correlated properties may over represent a single system property. The Pearson correlation coefficient is often used as measure for internal correlation strength. A threshold of 0.750 for either positive or negative correlation is used in this study. Table C1 shows the results for each of the extracted empirical protein properties. No properties were removed based on the set threshold.

Table C1: Pearson correlation coefficient matrix for all empirical protein properties.

	$R_{H\ App}$	**ζ**	**γ**	**T_{Agg}**	**T_M**	**β-turn**	**α-helix**	**Coil**
$R_{H\ App}$	1							
ζ	-0.345	1						
γ	-0.263	-0.287	1					
T_{Agg}	-0.717	0.351	0.145	1				
T_M	-0.496	-0.020	0.265	0.255	1			
β-turn	0.185	0.122	-0.058	-0.168	-0.393	1		
α-helix	0.138	-0.284	0.385	-0.150	0.173	-0.069	1	
Coil	-0.226	-0.232	0.392	0.182	0.470	-0.438	0.568	1

C.3 Cluster information

An overview of the exact median and median absolute deviation values for the clusters presented in the empirical protein property diagram are shown in Table C2.

Table C2: Overview of median ± median absolute deviation of the empirical protein properties per cluster identified in the empirical protein property diagram.

	A	B	C	D	E	F
$R_{H\ App}$ [nm]	1.4 ± 0.2	1.8 ± 0.3	2.2 ± 0.2	2.1 ± 0.3	2.3 ± 0.1	2.6 ± 0.3
ζ [mV]	3.8 ± 2.1	2.6 ± 1.3	1.1 ± 1.1	0.2 ± 1.7	-1.5 ± 1.8	3.1 ± 1.9
γ [-]	1.14 ± 0.02	1.07 ± 0.02	1.14 ± 0.03	1.17 ± 0.02	1.04 ± 0.03	1.04 ± 0.02
T_{Agg} [°C]	88.5 ± 0.9	56.7 ± 5.9	37.4 ± 3.2	39.0 ± 3.5	33.1 ± 1.9	35.2 ± 4.6
T_M [°C]	67.2 ± 4.0	68.8 ± 3.3	67.8 ± 3.4	68.9 ± 3.5	63.7 ± 2.2	63.2 ± 2.6
β-turn $[AU/(cm^{-1})^2] \cdot 10^5$	16.2 ± 1.8	15.0 ± 2.1	16.2 ± 2.0	15.2 ± 3.7	16.6 ± 2.3	18.9 ± 1.1
α-helix $[AU/(cm^{-1})^2] \cdot 10^5$	80.0 ± 8.1	83.2 ± 6.3	86.5 ± 7.1	97.3 ± 13.6	73.1 ± 12.8	84.2 ± 7.4
Coil $[AU/(cm^{-1})^2] \cdot 10^5$	28.1 ± 6.9	28.4 ± 2.9	29.7 ± 4.0	34.0 ± 8.0	16.4 ± 3.0	16.5 ± 3.9

C.4 Dynamic light scattering data

Exemplary plots of the apparent hydrodynamic radius ($R_{H\ App}$) as a function of lysozyme concentration are shown in Figure C2.

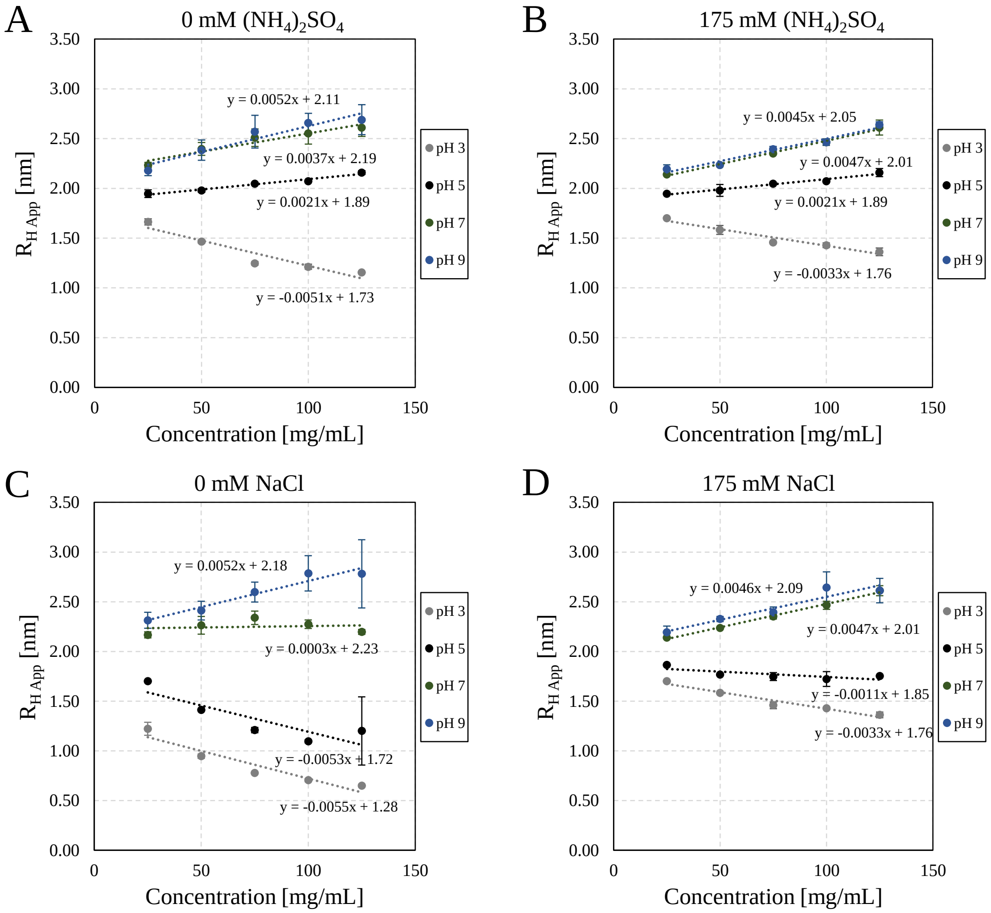


Figure C2: Apparent hydrodynamic radius ($R_{H\ App}$) plotted against lysozyme concentration for (a) 0 mM ammonium sulfate ($(NH_4)_2SO_4$); (b) 175 mM $(NH_4)_2SO_4$; (c) 0 mM sodium chloride (NaCl); (d) 175 mM NaCl at pH 3 (gray), pH 5 (black), pH 7 (green), and pH 9 (blue). The equation for a linear fit is given for each data series and the error bars represent the standard deviation.

Figure C2 shows the $R_{H\ App}$ plotted against lysozyme concentration for ammonium sulfate (Figure C2a and Figure C2b) and sodium chloride (Figure C2c and Figure C2d) at two different ionic strengths (0 and 175 mM) four pH values (pH 3, 5, 7, and 9). Each subplot shows a negative slope for pH 3 that increases for increasing formulation pH. Increasing ionic strength also increases the slope, but to a lesser extent compared to formulation pH. A negative slope for increasing lysozyme concentrations indicates repulsive electrostatic interactions between the protein molecules, while a positive slope represents attractive interactions.

Appendix D

Supplementary Material Chapter 6

D.1 Viscosity

Viscosity measurements were performed to correct the intensity-size distribution plots obtained with dynamic light scattering. The viscosity was measured by dissolving the additives in a pH 6.0 buffer containing 1.45 g/L methionine. An overview of the results is shown in Table D1.

D.2 Image examples

Figure D1 shows a representative example for each cluster identified in the multidimensional protein phase diagram (MPPD).

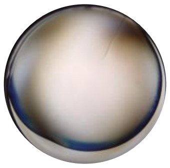
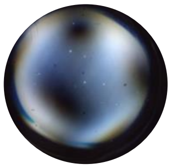
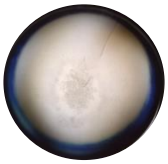
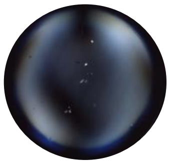
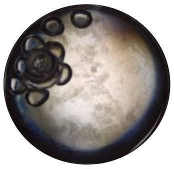
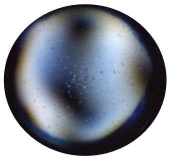
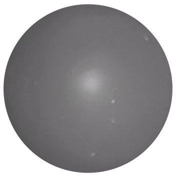
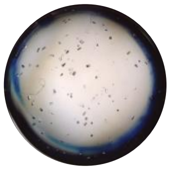
Cluster 1 	**Cluster 5** (Cross polarized light to highlight crystals) 
Cluster 2 	**Cluster 6** (Cross polarized light to highlight crystals) 
Cluster 3 	**Cluster 7** (Cross polarized light to highlight crystals) 
Cluster 4 (UV-light to highlight crystals) 	**Cluster 8** 

Figure D1: Representative example images for each identified multidimensional protein phase diagram cluster.

Table D1: Overview of mean viscosity values (mean η in kg/m·s) obtained for the listed formulation conditions, including the coefficient of variation (CV in %).

Sugar type	Sugar [g/L]	NaCl [g/L]	KCl [g/L]	Glycerol [g/L]	Mean η [kg/m·s]	CV [%]
Fructose	30	100	0	0	0.0017	0.31
Fructose	80	100	0	0	0.0019	0.31
Fructose	30	100	0	5	0.0022	0.26
Fructose	80	100	0	5	0.0026	0.23
Fructose	30	0	90	0	0.0020	0.41
Fructose	80	0	90	0	0.0018	0.20
Fructose	30	0	90	5	0.0024	0.23
Fructose	80	0	90	5	0.0023	0.30
Sucrose	30	100	0	0	0.0019	0.35
Sucrose	80	100	0	0	0.0019	0.33
Sucrose	30	100	0	5	0.0024	0.30
Sucrose	80	100	0	5	0.0025	0.20
Sucrose	30	0	90	0	0.0020	0.49
Sucrose	80	0	90	0	0.0017	0.31
Sucrose	30	0	90	5	0.0019	0.32
Sucrose	80	0	90	5	0.0022	0.31
Glucose	30	100	0	0	0.0017	0.35
Glucose	80	100	0	0	0.0018	0.27
Glucose	30	100	0	5	0.0022	0.24
Glucose	80	100	0	5	0.0023	0.28
Glucose	30	0	90	0	0.0016	0.28
Glucose	80	0	90	0	0.0023	0.31
Glucose	30	0	90	5	0.0020	0.34
Glucose	80	0	90	5	0.0023	0.29
Lactose	30	100	0	0	0.0017	0.29
Lactose	80	100	0	0	0.0021	0.30
Lactose	30	100	0	5	0.0023	0.29
Lactose	80	100	0	5	0.0028	0.72
Lactose	30	0	90	0	0.0016	0.41
Lactose	80	0	90	0	0.0019	0.31
Lactose	30	0	90	5	0.0021	0.30
Lactose	80	0	90	5	0.0023	0.31

D.3 MPPD cluster information

Each cluster identified within the MPPD consists of six image features. Each feature has a median and median absolute deviation within such a cluster. An overview of these values is shown in Table D2, per MPPD cluster.

Table D2: Overview of median ± median absolute deviation values for all image-based features per multidimensional protein phase diagram cluster

	L_C [µm]	W_C [µm]	W_C:L_C [-]	t_0 [hours]	t_G[hours]	n_{AGG} [%]
Cluster 1	0 ± 0	0 ± 0	0.0 ± 0.0	0 ± 0	0 ± 0	0 ± 0
Cluster 2	0 ± 0	0 ± 0	0.0 ± 0.0	0 ± 0	361 ±2	9.3 ± 9.3
Cluster 3	0 ± 0	0 ± 0	0.0 ± 0.0	0 ± 0	2 ± 1	52.5 ± 7.4
Cluster 4	21 ± 5	13 ± 3	1.6 ± 0.3	174 ± 62	198 ± 62	0.5 ± 0.0
Cluster 5	64 ± 17	39 ± 13	1.5 ± 0.3	312 ± 53	384 ± 53	1.5 ± 0.7
Cluster 6	71 ± 21	45 ± 21	1.6 ± 0.3	168 ± 44	480 ± 71	3.0 ± 3.0
Cluster 7	75 ± 17	43 ± 7	1.7 ± 0.4	0 ± 0	648 ± 107	28.8 ± 13.0
Cluster 8	55 ± 25	39 ± 14	1.3 ± 0.2	0 ± 0	361 ± 10	41.2 ± 7.4

A stability percentage per formulation condition was calculated based on the formulation for which the respective formulation condition showed an influence. The stability percentage represents the percentage of formulations that became stable after adjustment of the respective formulation condition. However, more cluster transformation were observed which did not result in physical stability. To quantify these transformations, a percentile contribution of all MPPD clusters was calculated per formulation conditions. These values are listed in Table S3. In addition, the number of formulations that were influence by the respective formulation condition is listed in Table D3 as well.

Table D3: Overview of all formulation percentages, based on the formulations where a change in physical stability or morphology was observed upon changing the respective variable.

Variable	Description	Cluster 1 [%]	Cluster 2 [%]	Cluster 3 [%]	Cluster 4 [%]	Cluster 5 [%]	Cluster 6 [%]	Cluster 7 [%]	Cluster 8 [%]	Total nr.
Sodium lactate	With 0 g/L sodium lactate	79	0	0	5	6	9	0	0	220
	With 100 g/L sodium lactate	0	8	23	4	10	23	22	11	220
Glycerol	With 0 g/L glycerol	0	7	21	6	15	20	21	10	234
	With 75 g/L glycerol	86	1	0	2	0	10	0	0	234
pH	5.0	5	3	25	8	9	26	21	4	174
	5.5	68	7	4	3	2	6	2	8	174
	6.0	75	0	0	1	10	8	5	2	174
Methionine	1.45 g/L	37	3	22	3	10	12	6	8	215
	9.50 g/L	40	5	0	7	7	21	17	3	215
Salt ratio	100 g/L NaCl, 0 g/L KCl	6	17	20	8	8	14	18	9	65
	60 g/L NaCl, 37 g/L KCl	20	5	8	6	14	18	14	15	65
	40 g/L NaCl, 55 g/L KCl	20	6	6	5	11	29	12	11	65
	0 g/L NaCl, 90 g/L KCl	31	0	0	12	23	22	11	2	65
Sugar type	Fructose	37	7	3	9	11	17	9	7	75
	Glucose	14	8	12	8	14	24	11	9	75
	Lactose	19	7	11	5	13	23	12	11	75
	Sucrose	20	5	9	4	12	28	16	5	75
Sugar concentration	30 g/L	19	5	5	8	12	30	12	9	100
	60 g/L	10	5	10	4	7	36	13	15	100
	80 g/L	38	10	11	8	18	3	11	0	100

D.4 Empirical protein property diagram cluster information

Each cluster identified within the empirical protein property diagram (EPPD) consists of 5 empirical protein properties. Each empirical protein property has a median and median absolute deviation within such a cluster. An overview of these values is shown in Table D4, per EPPD cluster.

Table D4: Overview of median ± median absolute deviation values for all empirical protein properties per empirical protein property diagram cluster. Empirical properties obtained with the original formulation are also listed.

	$R_{H\ App}$ [nm]	$R_{H\ HWS}$ [nm]	T_M [°C]	T_{Agg} [°C]	γ_N [-]
Cluster I	0.0 ± 0.0	743 ± 120	72.4 ± 3.0	65.1 ± 4.9	0.92 ± 0.04
Cluster II	3.4 ± 0.5	396 ± 213	72.8 ± 1.2	61.1 ± 4.1	0.92 ± 0.04
Cluster III	3.2 ± 0.4	78 ± 44	72.9 ± 1.4	55.1 ± 1.5	0.97 ± 0.04
Cluster IV	3.3 ± 0.4	110 ± 47	67.7 ± 1.2	56.1 ± 1.1	0.88 ± 0.03
Cluster V	3.2 ± 0.4	131 ± 82	76.1 ± 2.3	62.3 ± 2.6	0.95 ± 0.03
Original	2.4	47	65.5	54.2	1.01

A stability percentage was calculated per EPPD cluster. This was based on the percentage of formulations that were part of MPPD cluster 1, which represented physical stability. An overview of the composition of each EPPD cluster in terms of all MPPD cluster is shown in Table D5.

Table D5: Overview of the composition of empirical protein property diagram clusters shown as the formulation percentage per MPPD cluster

	Cluster I	Cluster II	Cluster III	Cluster IV	Cluster V
Cluster 1	12%	45%	64%	67%	75%
Cluster 2	12%	6%	7%	0%	2%
Cluster 3	24%	18%	18%	0%	4%
Cluster 4	0%	0%	0%	6%	2%
Cluster 5	0%	0%	7%	17%	8%
Cluster 6	24%	6%	0%	6%	0%
Cluster 7	24%	18%	0%	6%	8%
Cluster 8	6%	6%	4%	0%	0%

Appendix E

Supplementary Material Chapter 7

E.1 Quality parameters

Quality Z-score for each intermediate structure and WoS obtained in the proposed structure preparation pipeline is defined as the mean value of three separate WHAT IF parameter. The separate values are shown in Figure E1 for each intermediate structure and WoS for each VLP construct.

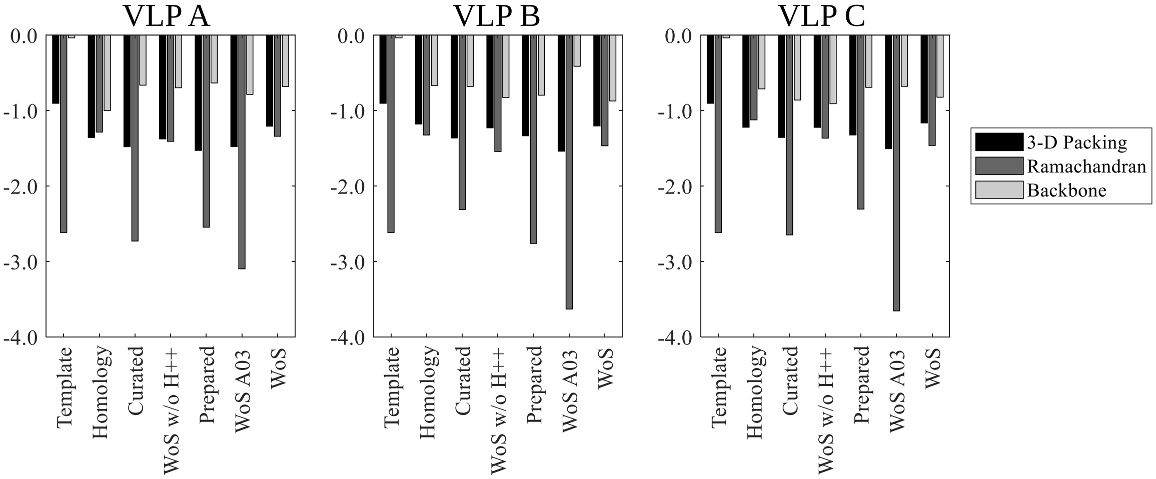


Figure E1: Overview of WHAT IF quality factors for the template, homology structure, curated structure, window of stability (WoS) without H++ and the YASARA2 force field ("WoS w/o H++"), the prepared structure, WoS obtained with H++ and the AMBER03 force field ("WoS A03"), and WoS obtained with H++ and the YASARA2 force field ("WoS"). WHAT IF quality factors 3-D packing (QUACHK, black), Ramachandran Z-score (RAMCHK, dark gray) and backbone conformation (BBCCHK, light grey)[299].

Figure E1 shows 3 quality parameters for each VLP construct for each intermediate structure and WoS obtained with the proposed 3-D structure preparation workflow. The backbone parameter shows a decrease from the template to the homology model for each VLP structure. The backbone and 3-D packing quality parameter values for all VLP constructs and obtained structures are smaller than the template. The Ramachandran quality parameter shows fluctuation between intermediate structures and MD simulation WoS. The fluctuations are similar between VLP constructs. The lowest Ramachandran quality parameter is found for WoS A03, followed by the curated, prepared and template structure. The homology structure, WoS w/o H++, and WoS show an increase of the Ramachandran quality parameter.

E.2 Reproducibility

To determine the reproducibility of the proposed protein 3-D structure preparation pipeline, all VLP constructs were simulated on two different computers using H++ computed pKa values and the YASARA2 force field. The hardware setup of the second computer was similar, using a Windows 10 computer with an Intel i7-6700 CPU and a GeForce GTX 1080 GPU. Reproducibility is evaluated based on obtained structure quality parameters, RMSD course during MD simulation, and correlation between the subsequent extracted surface charge descriptor and experimental zeta potential data. Figure E2 shows the quality Z-score plot for all intermediate structures and WoS obtained with the proposed structure preparation pipeline. All data is similar to the data presented in the main research article, except the WoS which was obtained using a different computer.

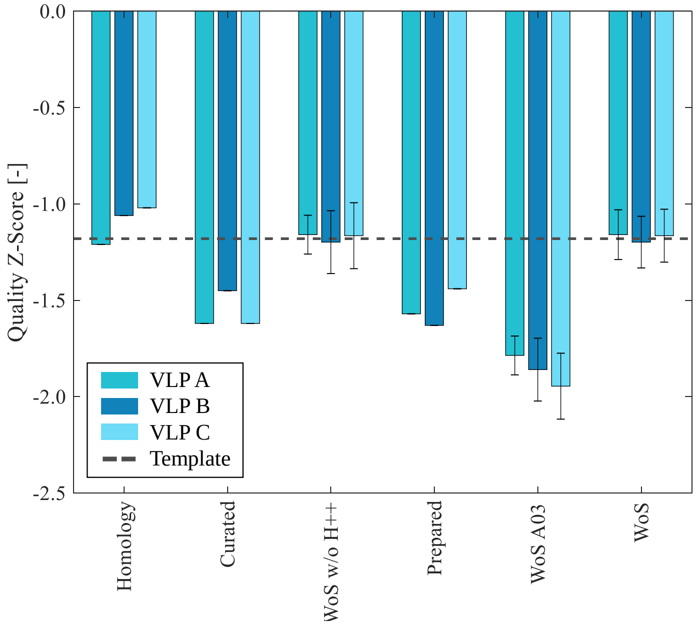


Figure E2: Overview of quality Z-scores for the template, homology structure, curated structure, window of stability (WoS) without H++ and the YASARA2 force field ("WoS w/o H++"), the prepared structure, WoS obtained with H++ and the AMBER03 force field ("WoS A03"), and WoS obtained with H++ and the YASARA2 force field ("WoS") on the second computer. The quality Z-score is an average value of the WHAT IF quality factors 3-D packing (QUACHK), Ramachandran Z-score (RAMCHK) and backbone conformation (BBCCHK)) [299]. A median value and median absolute deviation as error bar is shown for the WoS quality Z-scores. A dashed line is used to guide the eye between the different quality Z-scores.

Figure E2 shows a quality Z-score of -1.16 ± 0.13, -1.20 ± 0.13, and -1.16 ± 0.14 for VLP A, B, and C, respectively. These quality Z-scores have a mean difference of 0.07 compared to the quality Z-scored obtained with the first computer. An overview of the separate quality parameter obtained for the VLP constructs simulated with the second computer are shown in Figure E3.

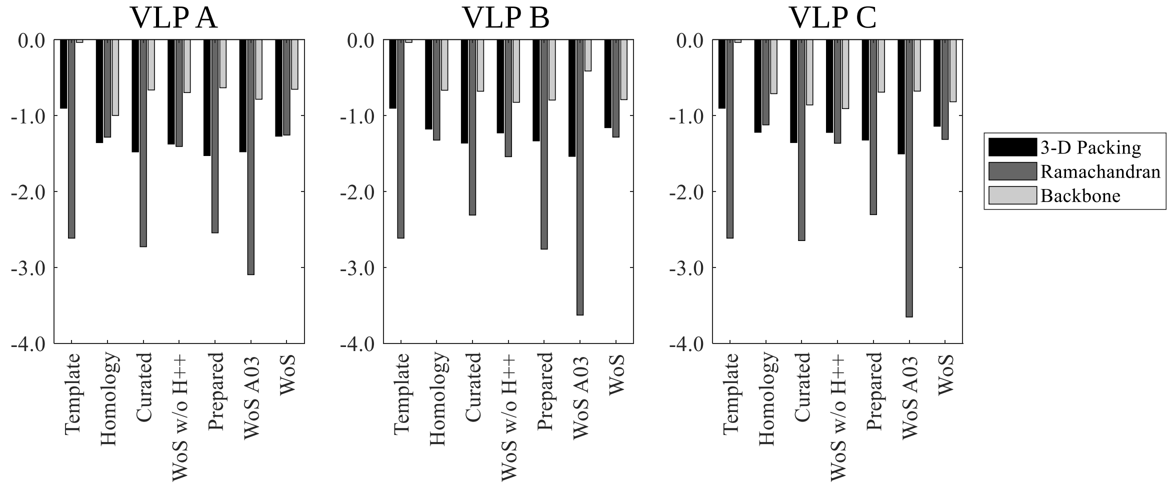


Figure E3: Overview of WHAT IF quality factors for the template, homology structure, curated structure, window of stability (WoS) without H++ and the YASARA2 force field (“WoS w/o H++”), the prepared structure, WoS obtained with H++ and the AMBER03 force field (“WoS A03”), and WoS obtained with H++ and the YASARA2 force field (“WoS”) on the second computer. WHAT IF quality factors 3-D packing (QUACHK, black), Ramachandran Z-score (RAMCHK, dark gray) and backbone conformation (BBCCHK, light grey)[299].

Figure E3 shows 3 separate WHAT IF quality parameters. A mean difference of 0.06, 0.10, and 0.06 was calculated using all VLP constructs simulated on the second computer in values for 3-D packing normality, Ramachandran plot position normality, and the backbone conformation, respectively. This indicates that quality was not influenced by simulation of identical constructs on another computer. The course of the MD simulation, represented by the change of atom coordinates over time was monitored for the simulations with the second computer as well. The obtained data is shown in Figure E4.

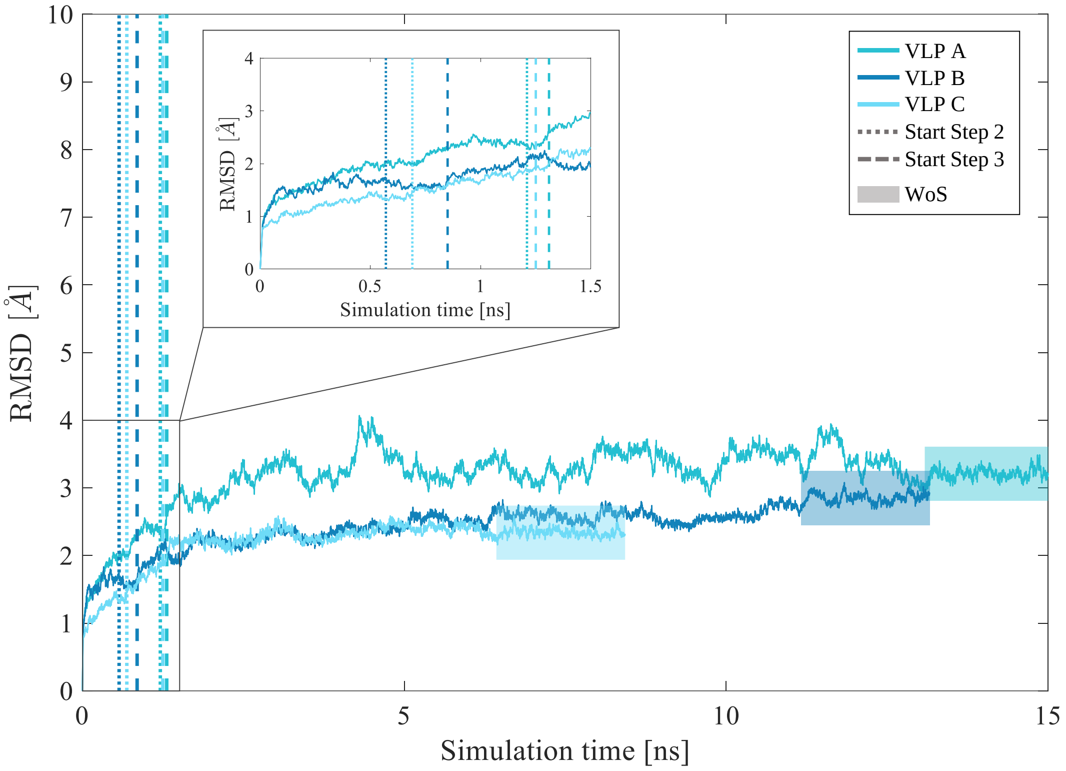
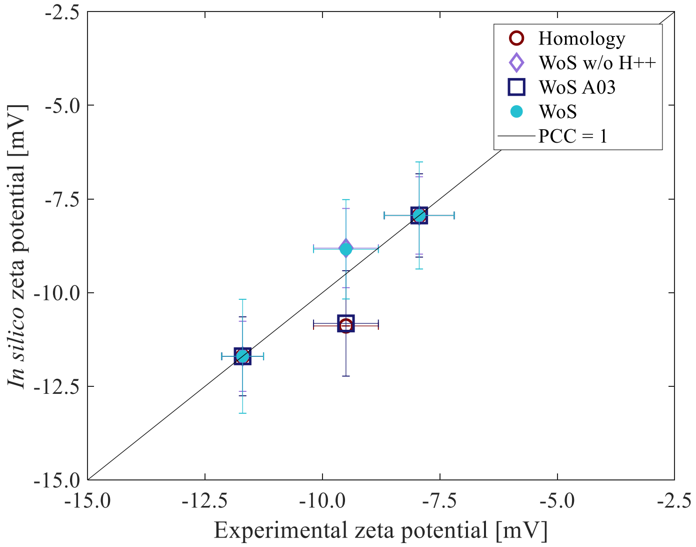


Figure E4: Reproducibility run of MD simulations for VLP A, B, and C presented by root-mean-square deviation (RMSD) of atom coordinates (Å) over simulation time (n), using a second Windows 10 computer with an Intel i7-6700 CPU and a GeForce GTX 1080 GPU. Three different simulation steps are separated by vertical lines, where vertical lines indicate simulation transition points. From 0 ns to dotted line: simulation of epitope and five adjacent amino acids; from dotted to dashed line: simulation of Hepatitis B core antigen (HBcAg) dimer spike; from dashed line to the end of simulation: full dimer simulation. The highlighted area is defined as the 2 ns window of stability (WoS).

Figure E4 shows the MD simulation course of three VLP constructs when simulated with the second computer. VLP A reached the WoS after 15.08 ns instead of 19.89 ns seen in the main research article. VLP B and VLP C reached the WoS later compared to the first computer, with a difference of 1.19 ns and 4.36 ns, respectively. The simulation time is still in accordance with the length of epitope insertion, where VLP A contains the largest insert. The maximum RMSD reached for each VLP construct is different compared to the RMSD shown in the main research article. VLP A, B, and C have a median WoS RMSD of 3.21 ± 0.06 Å, 2.86 ± 0.06 Å, and 2.33 ± 0.05 Å, respectively, in the simulation on the second computer. This should be compared to the median WoS RMSD of 7.52 ± 0.15 Å, 3.45 ± 0.07 Å, and 2.09 ± 0.04 Å on the first computer.

The influence of a different simulation course is also evaluated based on the prediction of complete HBcAg VLP zeta potential. The results are shown in Figure E5.



Structure	MSE	PCC
Homology	0.64	0.780
WoS w/o H++	0.16	0.998
WoS A03	0.58	0.791
WoS	0.14	0.997

Figure E5: In silico computed zeta potential (mV) plotted against experimentally determined zeta potential (mV). Symbols represent in silico data based on the homology structure ("Homology", red open circle), window of stability (WoS) obtained without H++ and with YASARA2 ("WoS w/o H++", purple diamond), WoS obtained with H++ and AMBER03 ("WoS A03", purple square), and WoS obtained with H++ and YASARA2 ("WoS", blue filled circle). The diagonal line represents theoretical data with a Pearson correlation coefficient of 1 (PCC = 1). X-axis error bars represent the median absolute deviation (MAD) of experimental data and y-axis error bars represent MAD for in silico data points. For each in silico data series the PCC and mean squared error (MSE) are calculated (n = 3) and listed.

Figure E5 shows that the WoS obtained with the second computer resulted in a in silico zeta potential of -8.84 ± 1.14 mV for VLP C, which shows a lower MSE (0.14) compared to the first computer (MSE = 0.45). The PCC also increases from 0.946 to 0.981. In silico zeta potential obtained with the second computer simulation also overlays with in silico zeta potential obtained without H++ pKa values. This supports the initial observation, where the used pKa values do not influence the surface charge description.