In silico AND EXPERIMENTAL APPROACHES TO DETERMINING THE AGGREGATION PROPENSITY OF BIOPHARMACEUTICAL PRODUCTS

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

der Fakultät für Chemieingenieurwesen und Verfahrenstechnik des
Karlsruher Instituts für Technologie (KIT)
vorgelegte

genehmigte

DISSERTATION

von
Dipl.-Ing. Frank Hämmerling
aus Ludwigsburg

Referent: Prof. Dr. Jürgen Hubbuch
Korreferent: Prof. Dr. rer. nat. Christoph Syldatk
Tag der mündlichen Prüfung: 28.03.2017

Danksagung

Während der letzten Jahre haben mich viele Menschen begleitet und unterstützt, denen ich an dieser Stelle von ganzem Herzen danken möchte.

Ich danke meinem Doktorvater Prof. Dr. Jürgen Hubbuch für die Möglichkeit, meine Doktorarbeit an seinem Lehrstuhl anzufertigen, für die vielen fachlichen Diskussionen, sein entgegengebrachtes Vertrauen, seine professionelle Expertise und die exzellente Ausstattung des Labors.

Prof. Dr. rer. nat. Christoph Syldatk danke ich für die freundliche Übernahme des Zweitgutachtens.

Meinen Kollegen vom MAB danke ich dafür, dass sie nicht einfach nur Kollegen waren, sondern auch zu Freunden wurden. Danke für die schöne Zeit bei der Arbeit, bei den Kaffeerunden und bei den vielen gemeinsamen Freizeitaktivitäten. Danke auch an die „erste Generation“ der Doktoranden am MAB, ganz besonders an Natalie Rakel und Patrick Diederich. Durch eure ausgezeichnete Betreuung während meiner Studien- bzw. Diplomarbeit am MAB habt ihr maßgeblich dazu beigetragen, dass ich diesen Weg eingeschlagen habe.

Danke an Pascal Baumann und Kai Baumgartner, die während den Höhen und Tiefen der vergangen Jahre zu besten Freunden für mich wurden.

Vielen lieben Dank an Josefine Morgenstern, Katharina Bauer, Christopher Ladd Effio, Carsten Radtke, Marie-Therese Schermeyer, Lara Galm, Cathrin Dürr und Sven Amrhein. Es war mir eine große Freude, mit euch zusammen zu arbeiten und auch viel Zeit außerhalb der Arbeit mit euch zu verbringen.

Meinen Bürokollegen aus dem „alten“ und „neuen“ EBI möchte ich für die stets angenehme Atmosphäre im Büro, die vielen fachlichen Diskussionen und die amüsanten Gespräche über eine Vielzahl von Themen weit abseits von Proteinen, Chromatographie & Co. danken.

Danke an meine fleißigen studentischen Helfer Philipp Holz, Markus Ziegler, Moritz Ebeler, Sebastian Andris, Fabian Görlich, Matthias Forschner, Oliver Lorenz-Cristea und Jan-Tobias Weggen, die mich durch ihre Abschlussarbeiten und Arbeiten im Labor sehr unterstützt haben.

Margret Meixner und Marion Krenz, die bei Fragen zur Bürokratie und häufig auch darüber hinaus immer eine große Hilfe waren, danke ich besonders.

Thomas Lebe vom Institut für Mechanische Verfahrenstechnik und Mechanik (MVM) möchte ich für die Hilfe beim Anfertigen der TEM-Aufnahmen danken.

Ein ganz besonderer Dank meinen Freunden von daheim, die immer einen schönen und für mich wichtigen Ausgleich zu Karlsruhe dargestellt haben.

Zu guter Letzt möchte ich meinen Eltern und meinem Zwillingsbruder Timo danken. Durch eure Unterstützung in jeglicher Hinsicht während des Studiums und der Promotionszeit habt ihr mir diesen Weg ermöglicht und mir zu jeder Zeit einen enormen Rückhalt gegeben. Danke für alles!

„Nicht wie der Wind weht,
sondern wie wir die Segel setzen,
darauf kommt es an.“
-unbekannter Verfasser-

Abstract

Biopharmaceutical products, such as monoclonal antibodies and vaccines, have significantly improved the treatment and prevention of various diseases in the last decade. Aggregation of these products is on the one hand often exploited during the purification processes. On the other hand, aggregation also leads to product loss during manufacturing and storage and potential safety concerns due to immunogenic reactions after administration in patients can arise. Hence, knowledge about the current phase state, the aggregation propensity, and the influence of changing environmental conditions is necessary to control aggregation of biopharmaceuticals.

The assessment of the aggregation propensity of biopharmaceutical products is still mainly based on heuristic approaches incorporating high sample material and time consumption, which is in most cases very limited during early stage process development. *In silico* methods help to overcome these drawbacks by drastically reducing the experimental effort. Furthermore, *in silico* methods help to generate a deeper understanding of the respective processes itself. This aspect is more and more moving into the spotlight during biopharmaceutical purification process development, as regulatory authorities are increasingly demanding a deeper process understanding, forwarding the quality by design guideline. The first two projects, that are presented in the subsequent sections, address the implementation of *in silico* methods during manufacturing of biopharmaceuticals. For vaccines that often comprise inactivated viruses, these *in silico* approaches are still hampered down to the present day, as they require enormous computational power due to the high complexity of such molecules. Nevertheless, there is a lack of fast, high-throughput compatible approaches to assess the colloidal and biological stability of this class of products. The third and fourth part of this thesis address this issue and present experimental approaches for the determination of surface properties of virus particles influencing their stability.

The first part of this thesis is focused on the quantitative structure-activity relationship (QSAR) modeling of diffusion coefficients of proteins. QSAR modeling is an *in silico* method that correlates the properties of the molecule's structure with its experimental behavior. It can additionally be used for predicting the experimental behavior of new entities, as well as gaining insights into the underlying mechanisms. The diffusion coefficient of proteins can be used as a measure for protein-protein interactions. To be able to capture these protein-protein interactions, diffusion coefficients have to be determined experimentally until now. In this part of the thesis, a QSAR model for the diffusion coefficient was generated. Therefore, the diffusion coefficients of six proteins at different pH values and sodium chloride concentrations were determined experimentally. The protein 3D structures of these proteins were obtained from a protein data bank and adapted to the respective experimental conditions. Based on these protein 3D structures, molecular descriptors accounting for structure properties, electrostatics, and hydrophobicity were calculated *in silico* and related to the experimentally determined diffusion coefficients by partial least squares regressions. As a result, a QSAR model for predicting diffusion coefficients sensitive to protein type, pH value, and sodium chloride concentration with a coefficient of determination R^2 of 0.9 was generated. The predictive capabilities were evaluated with an external test set and the predicted diffusion coefficients showed a coefficient of determination R^2 of 0.91. The model has demonstrated the potential to

predict the diffusion coefficients of proteins *in silico* and, hence, enables the possibility to capture protein-protein interactions. Furthermore, the model was able to give a more detailed picture of the protein properties influencing the diffusion coefficient and the acting protein-protein interactions. As up to now, available crude models for the estimation of diffusion coefficients only accounted for the proteins' molecular weight.

In the second part of this thesis, the methodology of QSAR was applied to model the precipitation of proteins with polyethylene glycol (PEG). Precipitation of proteins with PEG is considered to be an effective purification method for proteins, particularly as an alternative for costly chromatography processes. Additionally, the precipitation of proteins with PEG can be applied as a predictive tool to assess the long-term colloidal stability of protein formulations. Due to a lack of understanding of the underlying mechanisms, process development for these precipitation steps, however, still is mainly based on heuristic approaches and high-throughput experimentation. First reported models only account for the hydrodynamic radius of the proteins and PEG, and are not able to predict the complete precipitation curves. This deficiency was addressed in this project by modeling two parameters, namely the discontinuity point m^* and the β-value, that completely describe the precipitation curve of a protein. m^* depicts the PEG concentration at which protein solubility equals the protein concentration initially set, and the β-value the slope of the precipitation curve in the region where precipitation occurs. The generated QSAR models for m^* and β are sensitive to protein type, pH, and ionic strength and exhibit a good correlation between observed and predicted data with a coefficient of determination R^2 of 0.9 and, hence, are able to predict complete precipitation curves for proteins. It was found that m^* is mainly influenced by molecular structure properties and electrostatics, while β is mainly determined by electrostatics and hydrophobic properties. Model validation was performed by the application to an external test set of proteins that were not included in the generation of the models. The validation resulted in accurate predictions for two of the three investigated conditions. A deviation was observed for proteins with a molecular weight below 25 kDa. The presented project is the first reported approach enabling the *in silico* prediction of complete precipitation curves for proteins. The models help to accelerate process development for purification and formulation of biopharmaceuticals following the tenet of quality by design.

The third part of this thesis addresses the colloidal and biological stability of H1N1 influenza A viruses. Current influenza vaccines are mostly formulated as liquids, which requires a continuous cold chain to maintain the stability of the antigens. To overcome this dependency and to make optimized vaccines available that exhibit an increased stability at ambient temperatures, the influence of manifold parameters on the colloidal and biological stability has to be systematically investigated and understood. In this part of the work, phase diagrams of H1N1 influenza A viruses were generated in the microliter scale, using an automated liquid handling station for a large set of initial H1N1 and sodium chloride concentrations at different pH values. After incubation for 40 days at 20°C, the supernatant in each well with the respective conditions was evaluated for H1N1 mass recovery as a measure for colloidal stability, as well as for the remaining hemagglutination activity. These results serve as a basis for the subsequent part of this project, where a toolbox for the rapid assessment of the colloidal and biological stability was developed. This toolbox comprised the precipitation of H1N1 with polyethylene glycol as a predictive tool for the colloidal stability, and a combination of surface hydrophobicity determination, zeta potential measurements, as well as and Fourier transform infrared (FT-IR) spectroscopy as a predictive toolbox for the estimation of biological stability. The highest H1N1 mass recoveries were obtained at pH 6, the lowest ones at pH 4.5. It was found that there is a significantly lower H1N1 mass recovery for sodium chloride concentrations below 100 mM, and that recovery increases with increasing sodium chloride content. The highest values of remaining HA activity were determined at pH 9 and considerably lower relative remaining hemagglutination activities were observed for systems with low initial H1N1 concentrations. The precipitation of H1N1 with polyethylene glycol has proven its potential to replace time-consuming phase diagrams and to be a fast, high-throughput compatible predictive method to assess the colloidal stability of H1N1 virus particles. Combining surface hydrophobicity and zeta potential measurements and FT-IR sprectroscopy, it was possible to detect conformational changes in the surface proteins of the virus particles leading to a decrease in the hemagglutination activity. The combination of these methodologies depicts a powerful toolbox for the development of influenza vaccines with a preserved colloidal and biological stability and enables the rapid development of vaccine formulations that are stable at ambient temperatures.

During the last part of this thesis, the influence of the production system on the surface properties of H1N1 influenza A viruses was investigated. Influenza A/Puerto Rico/8/34 H1N1 (A/PR) viruses cultivated either in adherent or suspension Madin Darby canine kidney (MDCK) cells show a different aggregation behavior. In a first step, the differences in the aggregation behavior were revealed by the particle size distributions obtained from differential centrifugal sedimentation and dynamic light scattering measurements. Virus particles produced in adherent MDCK cells exhibit a higher aggregation tendency under low-salt conditions compared to those derived from suspension cell culture. In a second step, all surface characteristics of the virus particles, that might cause the deviations in aggregation behavior were investigated. The zeta potential, surface hydrophobicity, *N*-glycosylation fingerprints of the major A/PR surface antigen hemagglutinin, and lipid composition were determined for the two virus samples produced in adherent or suspension MDCK cells. It was found that the virus particles produced in adherent cells have a more negative zeta potential and a significantly lower surface hydrophobicity compared to those produced in suspension cells. The lipid composition of both virus particle samples was found to be fairly identical. Differences were also revealed in the *N*-glycosylation fingerprints of the hemagglutinin surface protein. The hemagglutinin of the virus particles derived from the adherent MDCK cells comprise longer *N*-glycans, which is also the explanation for the lower surface hydrophobicity as well as the higher aggregation propensity. The longer *N*-glycans probably weaken the electrostatic interactions by steric hindrance. This work demonstrates the severe influence of the production system on the surface properties of both virus particles and the importance of carefully selecting an appropriate production system. Thereby, the aggregation tendency of the virus particles can be drastically reduced and, hence, product loss minimized and critical quality attributes can be guaranteed.

In summary, the methods presented in this doctoral thesis represent powerful tools including both, *in silico* and high-throughput compatible methods, for predicting the aggregation propensity of biopharmaceutical products. The first-mentioned methods enable the *in silico* formulation and downstream process development for biopharmaceutical proteins and, thus, follow the demand of regulatory authorities to pursue the quality by design guideline. The latter experimental methods enable the rapid development of optimized vaccines with an increased stability during production and formulation at ambient temperatures.

Zusammenfassung

Arzneimittel auf Basis von biopharmazeutischen Herstellungsverfahren, wie zum Beispiel monoklonale Antikörper und Impfstoffe, haben die Behandlung und Prävention einer Vielzahl unterschiedlicher Erkrankungen im vergangenen Jahrzehnt erheblich verbessert. Die Aggregation dieser Produkte wird einerseits häufig während der Aufreinigungsprozesse ausgenutzt, andererseits führt die Aggregation zu einem Produktverlust während der Herstellung und Lagerung und zu potentiellen Sicherheitsrisiken durch immunogene Reaktionen nach der Verabreichung. Deshalb ist die Kenntnis über den gegenwärtigen Phasenzustand, die Aggregationsneigung und den Einfluss sich verändernder Umgebungsbedinungen notwenig, um die Aggregation von Biopharmazeutika zu kontrollieren.

Die Einschätzung der Aggregationsneigung biopharmazeutischer Produkte basiert noch immer hauptsächlich auf heuristischen Ansätzen, die zeitaufwendig sind und mit einem hohen Verbrauch an Probenmaterial einhergehen, das gerade in der frühen Prozessentwicklung sehr limitiert ist. *In silico*-Methoden helfen dabei, diese Nachteile zu überwinden, in dem sie den experimentellen Aufwand erheblich reduzieren. *In silico*-Methoden ermöglichen es außerdem, ein tieferes Verständnis über die Prozesse an sich zu erzuegen. Im Zuge des "Quality-by-Design"-Konzepts, das von den Zulassungsbehörden zunehmend verlangt wird, rückt dieser Aspekt während der Entwicklung von Aufreinigungsverfahren biopharmazeutischer Produkte zunehmend in den Fokus. Die ersten beiden Projekte, die in folgenden Abschnitten vorgestellt werden, befassen sich mit der Implementierung von *in silico*-Methoden während der Herstellung von Biopharmazeutika. Für Impfstoffe, die oftmals inaktivierte Viren beinhalten, ist der Einsatz dieser *in silico*-Methoden bis heute stark erschwert, da diese durch die sehr hohe Komplexität dieser Moleküle eine enorm hohe Rechenleistung benötigen. Dessen ungeachtet gibt es innerhalb dieser Produktklasse einen Mangel an hochdurchsatzfähigen Konzepten zur Beurteilung der kolloidalen und biologischen Stabilität. Der dritte und vierte Teil dieser Arbeit befasst sich mit dieser Thematik und stellt experimentelle Ansätze zur Bestimmung der Oberflächeneigenschaften von Viruspartikeln, die deren Stabilität beeinflussen, vor.

Der erste Teil dieser Arbeit beschäftigt sich mit der Modellierung von Diffusionskoeffizienten von Proteinen mittels quantitativer Struktur-Wirkungs-Beziehungen (*engl.* quantitative structure-activity relationship (QSAR)). Die Modellierung mithilfe von QSAR ist eine *in silico*-Methode, die strukturelle Eigenschaften von Molekülen mit deren experimentellem Verhalten korreliert. Zusätzlich kann diese Methode dazu verwendet werden, um das experimentelle Verhalten von neuen Substanzen vorherzusagen und Verständnis über die zugrunde liegenden Mechanismen zu generieren. Der Diffusionskoeffizient von Proteinen kann als Maß für Protein-Protein-Wechselwirkungen verwendet werden. Zur Bestimmung dieser Protein-Protein-Wechselwirkungen müssen die Diffusionskoeffizienten bis heute experimentell ermittelt werden. In diesem Teil der Arbeit wurde ein QSAR-Modell für den Diffusionskoeffizienten entwickelt. Hierfür wurden die Diffusionskoeffizienten von sechs Proteinen bei unterschiedlichen pH-Werten und Natriumchlorid-Konzentrationen experimentell bestimmt. Die 3D-Proteinstrukturen dieser Proteine wurden von einer Proteindatenbank bezogen und an die entsprechenden experimentellen Bedingungen angepasst. Ausgehend von diesen 3D-Strukturen wurden *in silico* molekulare Deskriptoren berechnet, die die strukturellen Eigenschaften, die Elektrostatik und Hydrophobizität beschreiben und mittels Partial Least Squares Regression mit den experimentell bestimmten Dif-

fusionskoeffizienten in Beziehung gebracht. Dadurch wurde ein QSAR-Modell mit einem Bestimmtheitsmaß (R^2) von 0,9 zur Vorhersage von Diffusionskoeffizienten erstellt, das die Art des Proteins, den pH-Wert und die Natriumchlorid-Konzentration berücksichtigt. Die prädiktiven Fähigkeiten des Modells wurden mithilfe eines externen Testsets beurteilt, dabei zeigten die Diffusionskoeffizienten ein Bestimmtheitsmaß R^2 von 0,91. Das Modell hat gezeigt, dass der Diffusionkoeffizient von Proteinen *in silico* vorhergesagt werden kann und ermöglicht dadurch die Bestimmung von Protein-Protein-Wechselwirkungen. Weiterhin konnte das Modell ein detaillierteres Verständnis über die Proteineigenschaften liefern, die den Diffusionskoeffizienten beeinflussen und über die vorherrschenden Protein-Protein-Wechselwirkungen Auskunft geben. Bisherige Modelle konnten lediglich das Molekulargewicht der Proteine berücksichtigen.

Im zweiten Teil dieser Arbeit wurde die Methodik von QSAR angewandt, um die Präzipitation von Proteinen mittels Polyethylenglykol (PEG) zu modellieren. Die Präzipitation von Proteinen mittels PEG wird als eine effektive Aufreinigungsmethode für Proteine angesehen, insbesondere auch als eine Alternative für kostenintensive chromatographische Verfahren. Zusätzlich kann die Präzipitation von Proteinen mittels PEG als ein prädiktives Instrument zur Beurteilung der Langzeitstabilität von Proteinformulierungen verwendet werden. Durch das fehlende Verständnis von den zugrunde liegenden Mechanismen, basiert die Prozessentwicklung dieser Präzipitationsschritte jedoch noch immer auf heuristischen Ansätzen und Hochdurchsatz-Experimenten (*engl.* high-throughput experimentation (HTE)). Erste veröffentlichte Modelle berücksichtigen ausschließlich den hydrodynamischen Radius der Proteine sowie des PEG und sind nicht dazu in der Lage, die vollständige Präzipitationskurve vorherzusagen. Dieses Defizit wurde in diesem Projekt durch die Modellierung zweier Parameter, dem Diskontinuitätspunkt m^* und dem β-Wert, die die vollständige Präzipitationskurve eines Proteins beschreiben, aufgegriffen. m^* stellt diejenige PEG-Konzentration dar, bei der die Löslichkeit des Proteins gleich der anfänglich eingestellten Proteinkonzentration ist, der β-Wert beschreibt die Steigung der Präzipitationskurve im Bereich mit auftretender Präzipitation. Die generierten QSAR-Modelle für m^* und β berücksichtigen die Art des Proteins, den pH-Wert sowie die Ionenstärke und weisen eine hohe Korrelation zwischen den experimentellen und vorhergesagten Daten mit einem Bestimmtheitsmaß von 0,9 auf. Daher können diese Modelle zur Vorhersage von vollständigen Präzipitationskurven verwendet werden. Es zeigte sich, dass m^* hauptsächlich von den strukturellen Eigenschaften der Proteine und der Elektrostatik beeinflusst wird, während der Wert von β überwiegend durch die Elektrostatik und hydrophoben Eigenschaften bestimmt wird. Die Validierung beider Modelle erfolgte durch ein externes Testset an Proteinen, das von der Modellbildung ausgeschlossen wurde. Die Validierung ergab präzise Vorhersagen für zwei der drei untersuchten Bedingungen. Eine Abweichung wurde für Proteine mit einem Molekluargewicht von weniger als 25 kDa beobachtet. Das vorgestellte Projekt ist der erste berichtete Ansatz der es ermöglicht, die vollständige Fällungskurve von Proteinen *in silico* vorherzusagen. Die Modelle helfen dabei, die Prozessentwicklung für die Aufreinigung und Formulierung von Biopharmazeutika zu beschleunigen und den Grundsatz des "Quality-by-Design"-Konzepts zu verfolgen.

Der dritte Teil dieser Arbeit widmet sich der kolloidalen und biologischen Stabilität von Influenza-A-Viren. Gegenwärtige Influenzaimpfstoffe sind meist als Flüssigkeiten formuliert, was für den Erhalt der Stabilität der Antigene eine durchgehende Kühlkette erforderlich macht. Um diese Abhängigkeit zu überwinden und optimierte Impfstoffe mit einer erhöhten Stabilität bei Raumtemperatur verfügbar zu machen, muss der Einfluss verschiedener Parameter auf die kolloidale und biologische Stabilität systematisch untersucht und verstanden werden. In diesem Teil der Arbeit wurden Phasendiagramme von H1N1 Influenza-A-Viren im Mikrolitermaßstab mithilfe einer robotergestützten Pipettierplattform für eine Vielzahl unterschiedlicher Ausgangskonzentrationen von H1N1 und Natriumchlorid bei verschiedenen pH-Werten angefertigt. Nach einer Inkubationszeit von 40 Tagen bei 20°C wurde der Überstand in jedem einzelnen System bei den entsprechenden Bedingungen ausgewertet. Hierbei wurde sowohl die Wiederfindungsrate von H1N1 als Maß für die kolloidale Stabilität, als auch die verbleibende Hämagglutinations-Aktivität untersucht. Diese Ergebnisse dienen als Basis für den zweiten Teil dieses Projekts, bei dem eine Toolbox für die rasche Beurteilung der kolloidalen und biologischen Stabilität entwickelt wurde. Diese Toolbox beinhaltet die Präzipitation von H1N1 mittels Polyethylenglykol als prädiktives Tool für die kolloidale Stabiliät, sowie eine Kombination aus der Bestimmung der Oberflächenhydrophobizität, der Messung des Zeta-Potentials und der Fourier-Transformierten-Infrarotspektroskopie (*engl.* Fourier transform infrared (FT-IR) spectroscopy) als prädiktive Toolbox für die Abschätzung der biologischen Stabilität. Die höchsten H1N1-Wiederfindungsraten wurden bei pH 6 ermittelt, die geringsten bei einem pH-Wert von 4,5. Es wurde festgestellt, dass die Wiederfindungsraten von H1N1 bei Natriumchlorid-Konzentrationen unter 100 mM erheblich abnehmen und dass die Wiederfindung mit zunehmendem Natriumchlorid-Gehalt zunimmt. Die höchsten Werte für die verbleibende Hämagglutinations-Aktivität wurden bei pH 9 erzielt. Für Systeme mit einer niedrigen Ausgangskonzentration von H1N1 wurden deutlich niedrigere relative verbleibende Hämagglutinations-Aktivitäten bestimmt. Die Präzipitation von H1N1 mittels Polyethylenglykol hat das Potential bewiesen, zeitaufwendige Phasendiagramme zu ersetzen und als schnelles, hochdurchsatzfähiges prädiktives Verfahren die kolloidale Stabilität von H1N1-Viruspartikeln zu beurteilen. Durch die Kombination aus der Bestimmung der Oberflächenhydrophobizität und des Zeta-Potentials und der FT-IR-Spektroskopie war es möglich, konformative Änderungen innerhalb der Oberflächenproteine der Viruspartikel zu detektieren, die zu einer Abnahme der Hämagglutinations-Aktivität führen. Die Verknüpfung all dieser Methodiken stellt eine leistungsfähige Toolbox für die schnelle Entwicklung von Formulierungen von Influenzaimpfstoffen mit einer erhaltenen kolloidalen und biologischen Stabilität bei Raumtemperatur dar.

Im letzten Teil dieser Arbeit wurde der Einfluss des Produktionssystems auf die Oberflächeneigenschaften von Influenza A-H1N1-Viren untersucht. Influenza-A/Puerto Rico/8/34 H1N1 (A/PR)-Viren, die entweder in adhärent oder in Suspension wachsenden Madin Darby Canine Kidney (MDCK)-Zellen kultiviert wurden, weisen ein unterschiedliches Aggregationsverhalten auf. In einem ersten Schritt wurden die Unterschiede im Aggregationsverhalten durch Dichtegradientenzentrifugation und dynamische Lichtstreuung aufgezeigt. Viruspartikel, die in adhärenten MDCK-Zellen hergestellt wurden, weisen im Vergleich zu den Partikeln aus der Suspensionskultur, eine höhere Aggregationstendenz unter Niedrigsalzbedingungen auf. In einem zweiten Schritt wurden alle Oberflächeneigenschaften der Viruspartikel untersucht, die für diese Unterschiede im Aggregationsverhalten verantwortlich sein können. Das Zeta-Potential, die Oberflächenhydrophobizität, das *N*-Glykosylierungsmuster des bedeutendsten A/PR Oberflächenantigens Hämagglutinin und die Lipidzusammensetzung der beiden in adhärenten oder in Suspensions-MDCK-Zellen hergestellten Virusproben wurde bestimmt. Es zeigte sich, dass die Viruspartikel die in den adhärenten Zellen hergestellt wurden im Vergleich zu denen in Suspensionkultur hergestellten, ein negativeres Zeta-Potential und eine deutlich geringere Oberflächenhydrophobizität besitzen. Die Lipidzusammensetzung der beiden Virusproben war annähernd identisch. Weitere Unterschiede zeigten sich im N-Glykosylierungsmuster des Oberflächenproteins Hämagglutinin. Das Hämagglutinin der Viruspartikel aus den adhärenten Zellen besteht aus längeren N-Glykanen, was sowohl die Erklärung für die geringere Oberflächenhydrophobizität als auch für die höhere Aggregationsneigung darstellt. Es wird angenommen, dass die längeren N-Glykane die elektrostatischen Wechselwirkungen durch eine sterische Hinderung herabsetzen. Diese Arbeit zeigt den starken Einfluss des Produktionssystems auf die Oberflächeneigenschaften der beiden Viruspartikel und die große Bedeutung der sorgfältigen Auswahl des geeigneten Produktionssystems. Dadurch kann die Aggregationsneigung der Viruspartikel erheblich reduziert und somit der Produktverlust verringert werden, sowie kritische Qualitätsmerkmale erreicht werden.

Insgesamt stellen die in dieser Dissertation vorgestellten Methoden ein leistungsfähiges Werkzeug zur Vorhersage der Aggregationsneigung biopharmazeutischer Produkte dar, die sowohl *in silico* als auch hochdurchsatzfähige Methoden beinhalten. Die zuerst beschriebenen Methoden ermöglichen die *in silico* Entwicklung von Formulierungen und Aufreinigungsprozessen für biopharmazeutische Proteine und ermöglichen dadurch das Verfolgen der von den Zulassungsbehörden geforderten "Quality-by-Design"-Richtlinie. Die zuletzt beschriebenen experimentellen Methoden ermöglichen die schnelle Entwicklung von optimierten Impfstoffen mit einer verbesserten Stabilität bei Raumtemperatur während der Produktion und Formulierung.

Contents

1 Introduction

1.1 Biopharmaceutical Products

Biopharmaceutical products have proven a huge clinical benefit in the treatment and prevention of manifold diseases during the last decade. They are used in the therapy of a wide range of medical indications from cancer and inflammatory diseases to hormone and enzyme replacement therapies. Biopharmaceuticals are cells, proteins or nucleic acid based pharmaceutical substances used for therapeutic or *in vivo* diagnostic purposes, which are produced by means other than direct extraction from natural (non-engineered) biological sources (Walsh [2013]). Such substances depict the fastest growing segment of the global pharmaceutical industry with sales in the U.S. exceeding US$ 54 billion in the year 2011. 32% of new drugs approved by the FDA in 2012 were protein therapeutics (Aggarwal [2012], Love *et al.* [2013]). Vaccines and monoclonal antibody-based compounds account for the majority of these products, but the range of products also includes cells for cell therapy, blood products, and nucleic acids (Walsh [2013]). Figure 1 shows an overview of the main biopharmaceutical products.

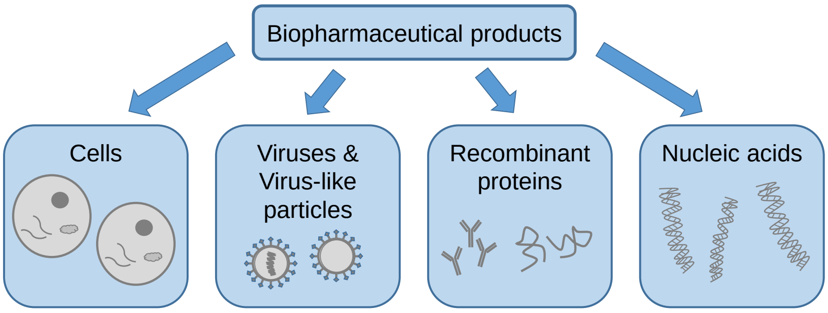


Figure 1: Overview of the main biopharmaceutical products. Typical size ranges are 1 - 100 μm for cells, 20 - 250 nm for viruses and virus-like-particles, and 1 - 10 nm for recombinant proteins.

Within the range of biopharmaceutical products, vaccines are gaining increasing attention. Vaccines are biological preparations that improve immunity to a particular disease. They typically contain an agent that resembles a disease-causing microorganism or surface epitope, and is often made from weakened or killed forms of the microbe, its toxins or one of its surface proteins. As the agent is recognized as foreign, it induces an immune response in the patient, is destroyed, and 'remembered' for later encounters (World Health Organization [2016]). Even though vaccines currently only account for 2 - 3% of total sales of the global pharmaceutical market, they show considerable annual growth rates of 10 - 15% compared to 5 - 7% for common pharmaceuticals. The vaccine market increased its value almost fivefold from US$ 5 billion in 2000 to US$ 24 billion in 2013 and is supposed to rise to US$ 100 billion by the year 2025. By having more than 120 new products in the development pipeline, vaccines are becoming the growth drivers of pharmaceutical industry (Kaddar [2013]). The vaccines with the highest sales in 2012 are the pneumococcal 13-valent conjugate vaccine Prevnar 13® (US$ 3.7 billion), Gardasil® for protection against human papillomavirus (US$ 1.9 billion), and PENTAct-HIB that protects against diphteria, pertussis, tetanus, polio, and haemophilus influenza type B

(US$ 1.5 billion). The influenza vaccine Fluzone$^{®}$ manufactured by GlaxoSmithKline is on the fifth position with sales topping US$ 1.1 billion (Philippidis [2013]). Unlike most other pharmaceutical products, almost all vaccines require a complex cold chain management to address the issue of stability that is mandatory to provide patients with safe formulations. Chen and Zehrung postulated that vaccines should be safe, efficacious, affordable, and manufacturable at low costs (Chen and Zehrung [2013]), but also current limitations like increasing the thermostability to overcome cold chain issues should be addressed during development and production (Pujar *et al.* [2014]). Table 1 lists the most common available types of vaccines and a brief description of the mode of action.

Table 1: Overview of vaccine types that are used to induce an immune response in the patients according to the National Institute of Allergy and Infectious Diseases [2012].

Type of vaccine	Description
Live, attenuated vaccine	Contain a version of the living microbe that has been weakened in the lab so that it cannot cause the disease. They elicit strong cellular and antibody responses and often confer lifelong immunity with only one or two administered doses.
Inactivated vaccine	The disease-causing microbe is inactivated by chemicals, heat, or radiation. This type of vaccine shows an enhanced stability and safety compared to vaccines employing living organisms, but stimulate a weaker immune response.
Subunit vaccine	Instead of the entire disease-causing organism, subunit vaccines include only the antigens that best stimulate the immune system. Therefore, the chance of adverse reactions is lower.
Toxoid vaccine	Toxoid vaccines are used when a bacterial toxin is the main cause of illness. The toxin is inactivated by treatment with formalin.
Conjugate vaccine	Many harmful bacteria possess an outer coating of polysaccharides that disguise the antigens and thereby prevent the recognition by the immune system. Conjugate vaccines contain the antigen or toxoid linked to polysaccharides that helps the immune system to react to polysaccharide coatings.
DNA vaccine	The genes for a microbe's antigen are introduced into the organism of the patient and some cells will take up the DNA. The DNA then instructs those cells to produce the antigen molecules and to secrete them. The body's own cells produce and provide the antigens for the immune response.
Recombinant vector vaccine	Recombinant vaccines are very similar to DNA vaccines, but they use an attenuated virus or bacterium as a carrier to introduce the DNA to the cells of the body.

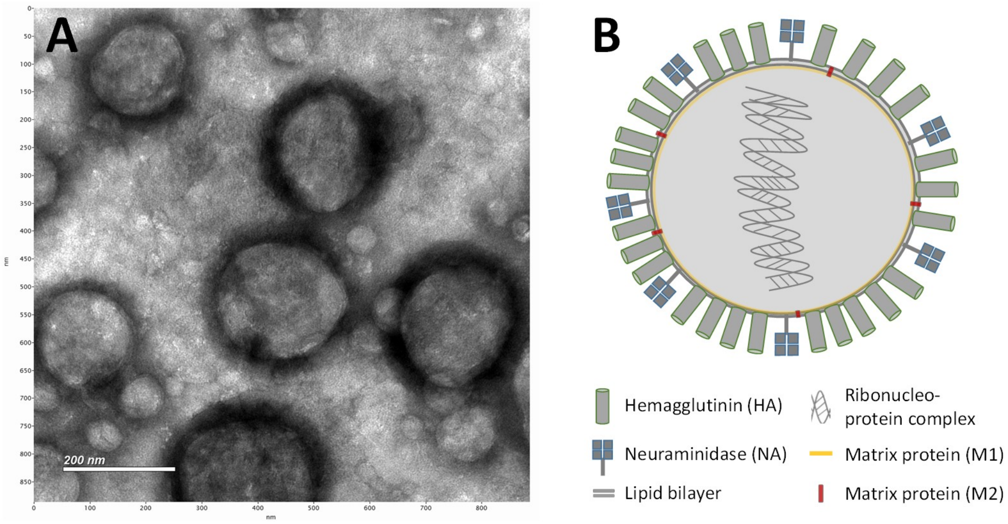


Figure 2: (A): TEM micrograph of negatively stained pandemic influenza A/Jena/5258/2009 (H1N1) virus particles with a diameter of approximately 200 nm. (B): Schematic representation of an influenza A virus particle.

Most influenza vaccines belong, with a few exceptions, to the group of inactivated vaccines. Influenza viruses are part of the family of *Orthomyxoviridae* and annually infect 5 - 10% of adults worldwide. This results in 3 to 5 million cases of severe illness and up to 500,000 deaths every year (Lamb and Krug [2001], World Health Organization [2014]). There are three types of influenza (A, B, and C), classified on basis of antigenic differences in their matrix and nucleoproteins. These three types of viruses also differ with respect to host range, variability of the surface glycoproteins, genome organization, and morphology. Influenza A viruses are responsible for pandemic outbreaks of influenza and for most of the annual flu epidemics. They also show the potential to cause worldwide pandemics by genetic changes, host changes, and introduction of a virus with a novel surface protein subtype that is new to human populations (Neumann *et al.* [2009], Taubenberger and Kash [2010]). Influenza A viruses are further charaterized according to their subtype of surface glycoproteins, namely hemagglutinin (HA) and neuraminidase (NA), embedded in a host cell-derived lipid membrane. So far, 16 subtypes of HA and 9 of NA have been found (Amorij *et al.* [2008]). The hemagglutinin surface protein is a glycosylated viral membrane protein, which is protruding in a spike-like form from the virus particle (VP) surface. It is responsible for both, attachment of the virus to sialic acid-containing receptors on the host cell surface, and fusion of the viral and target endosomal membrane (Taubenberger and Kash [2010]). HA has a molecular weight of approximately 225 kDa and consits of three identical monomers, each with a molecular weight of 75 kDa. Each monomer by itself consists of the polypeptides HA1 (∼50 kDa) and HA2 (∼25 kDa), which are linked by two disulfide bonds. The neuraminidase is a tetrameric glycoprotein (∼240 kDa) consisting of a hydrophobic stick and a globular head (Amorij *et al.* [2008]). The ratio of HA to NA is approximately four to one. A small number of matrix ion channels (M2) traverse the lipid envelope, with a ratio of M2 to HA of about one to 10^1-10^2 (Bouvier and Palese [2008]). Figure 2 shows a TEM micrograph of pandemic influenza A virus particles (A) and a schematic representation of the virus particle itself.

As HA represents approximately 35% of the total VP protein (Fields *et al.* [2001]) and the surface of influenza viruses mainly consists of the two surface proteins HA and NA, it is therefore assumed, that protein-protein interactions play a prominent role in the stability of the viruses and in their aggregation properties.

1.2 Downstream Process Development for Biopharmaceutical Products

Safe vaccines require the standard production chain of biopharmaceutical products, including production, purification and formaulation/ storage. Downstream processing refers to the recovery and purification of biosynthetic products, especially biopharmaceutical products, from natural sources. The development of downstream processes is based on three different approaches: heuristic, experimental, and model-based approaches. While heuristic approaches are based on expert knowledge, rules of thumb, and the application of platform processes, the experimental approach uses methods such as high-throughput experimentation (HTE) and statistical methods such as 'Design of Experiments' (DoE). HTE is one of the hot topics in pharmaceutical research and has become a standard tool in industry and academia. It describes the methodology for a large number of parallelized, miniaturized, and automated experiments in pharmaceutical research (Kelley *et al.* [2008], Łącki [2014]). The model-based approach applies empirical or mechanistic models to simulate experiments *in silico* (Baumann and Hubbuch [2016]).

1.3 *In Silico* Methods in Downstream Processing of Biomolecules

These computer-aided approaches belong to the model-based approaches and increasingly become the focus of attention as they can drastically decrease the number of experiments and follow the tenet and demands of the 'Quality by Design' (QbD) approach stated by regulatory authorities. The aim of QbD during production of biopharmaceuticals is to provide scientifically grounded processes, a risk-based evaluation of manufacturing performance that enables the most suitable choices for process parameters for robust and flexible operation (Chhatre *et al.* [2011], Mhatre and Rathore [2008]). Quantitative structure-activity relationship (QSAR) is a hybrid approach and depicts a combination of the experimental and model-based approach (Baumann and Hubbuch [2016]). The methods of molecular dynamics (MD) simulations and QSAR were used in this thesis and are presented in the subsequent chapters. MD simulations were used during the preparation of protein 3D structures and their adaption to the respective environmental conditions for QSAR modeling.

1.3.1 Molecular Dynamics Simulation

Molecular dynamics simulations have the capability of providing molecular and atomistic insights into mechanisms, kinetics, and chemical processes. They enable a deeper understanding of the fundamental principles and are an important complement to experimental results (Schaller *et al.* [2015]). MD simulations require a realistic description of the underlying physical system and its molecular interactions. The explicit model of interactions, including a mathematical model and the parameters that are required for that model, is referred to as the 'force field' or the 'interaction potential'. MD simulations have been used for a large number of applications in the field of biopharmaceutical science recently. These include prediction of recombinant protein expression, characterization of peptid hydrophobicity, and prediction of retention during chromatography (Amrhein *et al.* [2014], Dismer and Hubbuch [2010], Oelmeier *et al.* [2012], Schaller *et al.* [2015]).

1.3.2 Quantitative Structure-Activity Relationship Modeling

Quantitative structure-activity relationships (QSARs) are mathematical models that attempt to relate the structure-derived features of a molecule to its biological or physicochemical activity. QSAR works on the assumption, that structurally similar compounds have similar activities and, therefore, the models have predictive abilities (Dehmer *et al.* [2012]). For the generation of a QSAR model for a particular activity, experiments for a large set of proteins under various conditions are performed to obtain the respective activity. Subsequently, a MD simulation experiment is performed in order to adapt the protein 3D structure to the environmental conditions. Based on this three-dimensional structure of the proteins, molecular descriptors accounting for size, shape, electrostatic, and hydrophobic properties are calculated. The complete data set is split into a training set and a test set. The training set is used for the generation of the model, the test set is excluded from model generation and reserved for model validation. For the actual generation of the QSAR model, where the most influencing molecular descriptors are related to the activity, multi-variate data analysis techniques are applied. After validation of the QSAR model it can then be applied for prediction of the behavior of the molecules or process conditions (Baumann and Hubbuch [2016], Dehmer *et al.* [2012], Hanke and Ottens [2014]). Figure 3 illustrates the complete procedure of QSAR modeling.

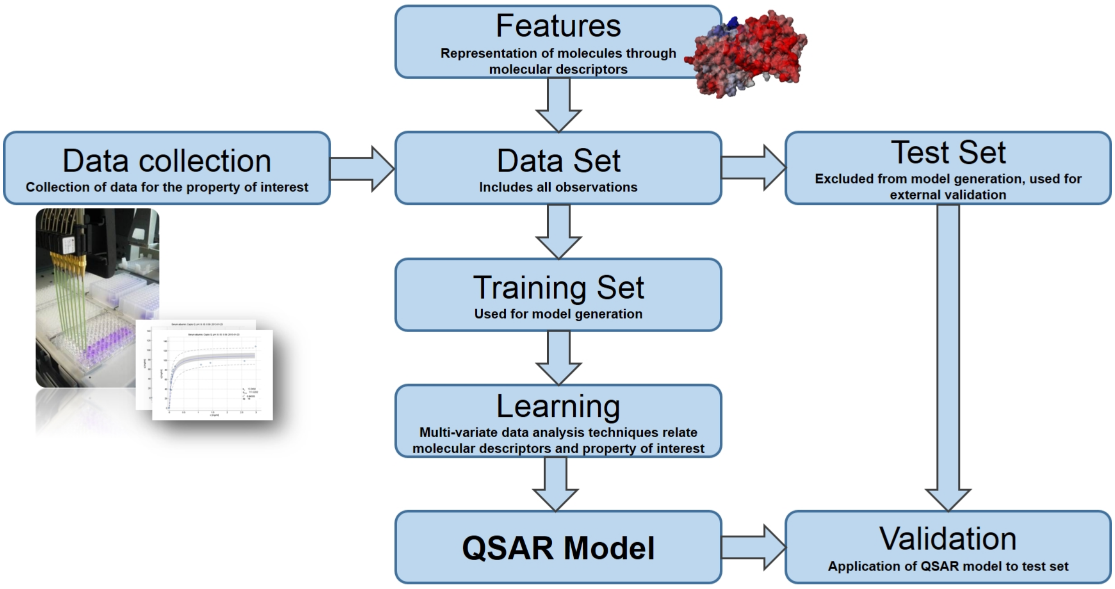

Features
Representation of molecules through molecular descriptors
Data collection
Collection of data for the property of interest
Data Set
Includes all observations
Test Set
Excluded from model generation, used for external validation
Training Set
Used for model generation
Learning
Multi-variate data analysis techniques relate molecular descriptors and property of interest
QSAR Model
Validation
Application of QSAR model to test set

Figure 3: Workflow for the generation of a QSAR model according to (Yee and Wei [2012]).

A useful accompaniment of QSAR modeling is the gain of an enhanced understanding of the mechanisms driving the investigated activities by the evaluation of the importance of each descriptor to the model. QSAR modeling therefore often considered as semi-mechanistic modeling technique. In the field of downstream process development for biopharmaceutical products QSAR models have been successfully used to describe and predict the binding and retention of proteins during chromatography. For ion-exchange chromatography, QSAR models have been reported by Mazza *et al.* [2001, 2002], Ladiwala *et al.* [2005], and Dismer and Hubbuch [2010]. Chung *et al.* [2010] generated QSAR models to predict the retention times of proteins during mixed-mode chromatography and elucidated the factors influencing protein retention. Besides the prediction of retention times, Yang *et al.* [2007a] applied the QSAR methodology to describe and design mixed-mode ligands with the help of molecular descriptors in order to acquire insight into the important physicochemical properties required for protein binding under high-salt conditions. The retention of proteins during hydrophobic interaction chromatography and the influence of the ligand and backbone chemistry of the resin were published by Ladiwala *et al.* [2006]. All of these publications are based on systems containing only one single protein. The first QSAR model for a complex feedstock was reported by Buyel *et al.* [2013] where QSAR was applied for the chromatographic depletion of tobacco host cell proteins. More than 100 host cell proteins were identified by mass spectrometry, their 3D structures were reconstructed from X-ray crystallography, molecular descriptors calculated and used for the generation of QSAR models predicting retention times. The application of QSAR for very large biomolecules, such as viruses, is severely hindered because of extremely high computational times necessary. For a molecular dynamics simulation, the computational power is scaling with the number of atoms simulated within the simulation box. If a system with twice as many atoms is simulated, the calculation would require between two to four times as much computational power (Freddolino *et al.* [2006], Segall *et al.* [2015]).

1.4 Stability of Biopharmaceutical Products

As mentioned in section 1.1, most vaccines require a continous cold chain to maintain the stability of the antigens. Generally, the stability of a pharmaceutical product may be defined as the capability of a particular formulation in a specific container system, to remain within its physical, chemical, microbiological, therapeutic, and toxicological specifications (Bokser and O'Donnell [2006]). Especially for biopharmaceutical products, Manning *et al.* [2010] defined two general types of instabilities: chemical and physical instabilities. Chemical instabilities comprise processes that generate or break chemical covalent bonds, resulting in new entities. Such reactions include deamidation, oxidation, or hydrolysis of single amino acids or of the entire protein molecule. Physical instabilities include denaturation, aggregation, precipitation, and adsorption of proteins and trigger a change of the physical state of the molecule without any changes in the chemical composition. The natural three-dimensional or tertiary structure of the protein is designated as the native state of a protein. Denaturation is referred to a loss of this native structure. Denaturation can be provoked by exposure of proteins to thermal stress, both heat and cold, the addition of chemical agents, especially chaotropic salts from the Hofmeister series (Hofmeister [1888]), or by exposure to unfavorable pH values or high pressure.

Aggregation is often triggered by denaturation of the protein. Aggregation describes the assembly of monomers to protein multimers. In this work, the term 'aggegates' is referred to as a summary of species of higher molecular weight, such as oligomers or multimers, instead of the desired defined species (e.g., a monomer) (Mahler *et al.* [2009]). According to Mahler et al., particularly for proteins, aggregates can be classified by the following categories:

- the type of bond: noncovalent aggregates (bound by weak electrostatic forces) versus covalent aggregates (e.g., caused by disulfide bridges)
- by reversibility: reversible versus irreversible aggregates
- by size: small soluble aggregates (oligomers) such as dimers, trimers etc. versus large ($\geq$decamer) oligomers versus aggregates in the diameter range up to 1 μm or insoluble particles with larger diameters
- by protein conformation: aggregates with predominantly native structure versus aggregates with predominantly nonnative structure

Aggregation is a major challenge during the production, purification, and formulation of biopharmaceuticals (Shire *et al.* [2004]). Aggregates are regarded as critical for the product quality and as a potential safety concern due to the increased immunogenicity. The presence of small aggregates may lead to an immunogenic reaction, whereas large aggregates may cause adverse events upon administration (Cromwell *et al.* [2006], van Beers and Bardor [2012], Wang [2015]). Nonnative aggregation is particularly problematic because it is encountered routinely during refolding, purification, sterilization, shipping, and storage processes (Chi *et al.* [2003]). Because the environmental conditions are subject to frequent changes during these operations, the stability of a biopharmaceutical product is greatly influenced by a number of environmental conditions, such as pH value, type and concentration of added salt, redox potential, temperature, and the presence of stabilizing excipients (Brandau *et al.* [2003], Priddy *et al.* [2014]).

1.5 Factors Affecting Aggregation of Biopharmaceutical Products

As discussed in the previous section, the stability of biopharmaceutical products is a function of various parameters. In this thesis, stability investigations of proteins as well as influenza viruses were subject of research. As the two proteins HA and NA form the surface peplomer of influenza A viruses (Fields *et al.* [2001], Kapoor and Dhama [2014]), it is therefore assumed that protein-protein interactions also determine the interactions between the virus particles and, thus, their aggregation behavior. In the next sections the four parameters with the highest influence on the aggregation of biopharmaceutical products, namely the temperature, the pH value, as well as the concentration and type of salt are discussed in detail.

1.5.1 Influence of Temperature on Protein Stability

The folding ('conformation') of a protein is essential for its biologic function. The thermodynamic stability of the native protein conformation is only marginal, the stability is about 5 - 20 kcal/mol of free energy enhanced to the unfolded, biologically inactive conformations under physiologic conditions. The small net conformational stability results from a balance between large stabilizing and large destabilizing forces. Relatively small changes of external variables (e.g., temperature) might destabilize the strucuture of the protein, i.e., induce its unfolding (Chi *et al.* [2003]). Unfolding of the protein structure leads to an exposure of hydrophobic groups, that were buried in the hydrophobic core of the protein, and induce aggregation. Besides influencing the conformational stability, temperature also strongly affects the colloidal stability of proteins. Thermal kinetic energy of molecules, atoms, and subatomic particles and temperature are directly correlated by the Stokes-Einstein equation (Equation 1). The Stokes-Einstein equation describes the diffusion coefficient of a species (D) in relation to its temperature (T):

$$D = \frac{k_B T}{4\pi r_h \eta_S}. \tag{1}$$

In this equation r_h depicts the hydrodynamic radius of the solute, η_S the viscosity of the surrounding solution, and $k_B T$ the thermal kinetic energy. A higher temperature, resulting in a higher value of D, leads to a higher collision frequency, as well as to a higher probability of collisions with enough energy to overcome activation energies and therefore to an increased aggregation rate (Chi *et al.* [2003]). As a consequence, many biopharmaceutical products, and particularly vaccines, require a continous cold chain. Elevated temperatures usually result in higher aggregation rates, but this statement is not generally valid. Lin *et al.* [2008] reported a contrary effect of temperature where some proteins, e.g., lysozyme, showed a higher solubility at higher temperatures. It is thus protein species dependent, whether the protein solution can be stored at room temperature or needs to be cooled.

1.5.2 Influence of pH Value on Protein Stability

The pH value of the surrounding aqueous solution strongly influences the type and distribution of surface charges on the protein surface and, hence, affects intramolecular folding and protein-protein interactions. If the pH exceeds the pK_a value of the respective amino acid side chain, the titrable group is deprotonated. Under conditions with a pH below the respective pK_a, the titrable group is protonated. The sum of all positive and negative charges of the protein depicts its net charge. The pH value where the surface net charge of the protein is zero is referred to as the isoelectric point (pI). Nevertheless, there are still charged surface patches present at the pI, but protein solubility is theoretically minimal at the pI due to minimal electrostatic repulsion. At pH values away from the pI, the protein is strongly charged and long-range repulsive electrostatic interactions occur that have a stabilizing effect on colloidal stability of protein solutions. But the strong charge of the protein might also cause intramolecular charge repulsion that might induce conformational changes within the tertiary protein structure. Hydrophobic amino acid side chains that are buried in the core of the proteins might get exposed due to this change in tertiary structure, to the surface of the molecules and, thus, lead to aggregation (Wang *et al.* [2010]). Additionally, the pH value of a solution is affected by the temperature.

1.5.3 Influence of Salts on Protein Stability

Besides the temperature and the pH value, salts have complex effects on physical stability of proteins, e.g., by modifying conformational stability, equilibrium solubility (salting-in and salting-out effect), and rate of formation of nonnative aggregates (Chi *et al.* [2003]). Ions also reduce long-range electrostatic interactions by shielding charges as they bind to or interact with proteins. The effect of salts on protein solutions can be caused by both, the type and the concentration of salt added. The overall effect of ionic strength on protein aggregation is strongly dependent on the protein species. If neutralization of protein surface charges is beneficial for protein folding and stability, the reduction of such interactions would destabilize the protein, partially expose hydrophobic patches due to strong intramolecular charge repulsion, and lead to an increased aggregation propensity (Wang *et al.* [2010]). Additionally the effect of ionic strength can be dependent on more parameters, e.g., the pH of the solution that mainly determines the type of interactions present (Saluja *et al.* [2007b]) and glycosylation state of a protein, as glycans may weaken the electrostatic interactions through steric hindrance (Høiberg-Nielsen *et al.* [2006]). Other consequences resulting from the addition of salt to a protein solution are the 'salting-in' and 'salting-out' effect. If the added salt ions preferentially bind to proteins (the so-called chaotropic ions or water structure breaker), they increase the protein's net charge and, hence, the solubility. This effect is also referred to as salting-in effect. For example when adding sodium chloride (NaCl), the protein solubility as a function of the NaCl concentration depicts a bell-shaped course and maximum solubility was observed at NaCl concentrations up to 2.0 - 2.5 M NaCl (Collins [2004]). At higher concentrations, chaotropic salts might also decrease the intramolecular stability of the protein, leading to partial unfolding and, thus, promoting aggregation. By contrast, ions that are polar and strongly hydrated are defined as kosmotropic ions (or water structure maker). They retract water from the protein surface and thereby expose hydrophobic surface patches. By encouraging the protein to minimize its solvent accessible surface area, they decrease the solubility. This effect is denoted as salting-out effect (Arakawa and Timasheff [1984], Brandau *et al.* [2003], Curtis *et al.* [1998]). For particles with semipermeable membranes, such as bacteria and enveloped influenza viruses, high ionic strengths might additionally induce membrane lysis (Priddy *et al.* [2014]).

1.6 Stability Assessment of Biopharmaceutical Products

As the effects on stability mentioned above are highly complex and interconnected, there is a major need for fast and straightforward tools for stability assessment of biopharmaceutical products.

1.6.1 Colloidal Stability of Proteins and Viruses

The self-association of biopharmaceutical products, such as proteins and viruses, is a frequently observed type of colloidal instability of these molecules or particles in solution. Aggregates are regarded as critical for the product quality and as a potential safety concern due to the increased immunogenicity (van Beers and Bardor [2012]). Many models have been used to describe and calculate the colloidal stability of a protein solution. However, many of these simple colloidal models reveal limitations, as they have no clear physical meaning and, thus, are not able to provide a link between the solution variables (e.g., ionic strength) and the potential parameters.

1.6.1.1 Description of Colloidal Stability

The Derjaguin-Landau-Verway-Overbeck (DLVO) model, which includes long-range electrostatic interactions as well as short-range attractive van der Waals interactions, is widely used to describe the thermodynamics and kinetics of colloidal stability of protein solutions (Li *et al.* [2008]). Figure 4 displays the DLVO theory according to De Young et al. (De Young *et al.* [1993]).

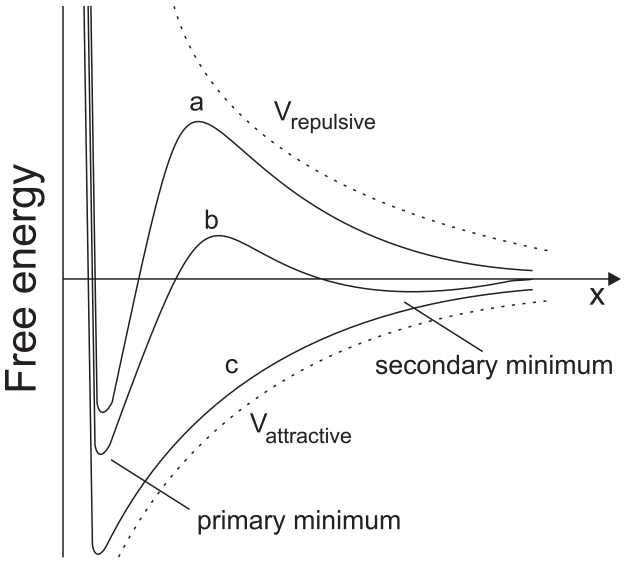


Figure 4: Summary of DLVO theory according to De Young et al. (De Young *et al.* [1993]): Attractive van der Waals ($V_{attractive}$) and repulsive electrostatic forces ($V_{repulsive}$) are plotted as a function of the distance between two particles x. The sum of these two forces at (a) low, (b) medium, and (c) high salt concentrations determines whether the solution will be colloidal stable or will aggregate in the primary or secondary minimum.

The DLVO theory accounts for steric and electrostatic repulsions and van der Waals attractions between particles in solution. Under low-salt conditions (curve a), there is a strong electrostatic repulsion between the particles, resulting in a large free energy barrier to aggregation. These solutions are referred to as kinetically stabilized (Russel *et al.* [1989]). Under high-salt conditions (curve c), the electrostatic charges on the particle surfaces are shielded from each other and attractive van der Waals forces dominate, resulting in a strong attraction of two particles in a deep primary minimum of free energy. According to the DLVO theory, the rate of aggregation is high under high salt concentrations, as there is no barrier of free energy. Due to the strongly pronounced primary minimum of free energy, this aggregation is often designated as irreversible. Under medium salt concentrations (curve b), electrostatic and van der Waals forces are more balanced. In this case, small amounts of added salt can decrease the barrier height, and hence cause a large increase in the aggregation rate. For moderate barrier heights, the soluble protein species and aggregates can coexist at equilibrium. A secondary minimum of free energy energy exists at low and medium salt concentrations. Aggregation in the secondary minimum is often called reversible, since the barrier of free energy is small and there is an equilibrium between aggregated and dissolved phases (De Young *et al.* [1993]). Under dilute conditions, the colloidal stability is mainly determined by the long-range repulsive electrostatic interactions (Saluja *et al.* [2007a,b]), while for highly concentrated protein solutions several short-range interactions are dominating, due to the short distance between the molecules in solution (Saluja and Kalonia [2008]).

1.6.1.2 Experimental Evaluation of Colloidal Stability

The colloidal stability of biopharmaceutical products can be displayed through phase diagrams. Phase diagrams are usually created with the help of an automated liquid-handling station in a microbatch format. They provide information about the phase state of a biomolecule under the investigated conditions. A schematic protein phase diagram is displayed in Figure 5. It provides information about the phase state of the protein as a function of initial protein and precipitant concentration. The solubility line divides the phase diagram into two sections, namely, the undersaturated and supersaturated zone. Under conditions in the undersaturated zone, the protein remains soluble. The supersaturated zone depicts those conditions where the solubility limit is exceeded and crystallization, gel or skin formation, phase separation, or even precipitation occurs. Crystallization can be either induced by heterogeneous nucleation in the metastable zone or spontaneously in the labile zone (Ahamed *et al.* [2007], Asherie [2004], Baumgartner *et al.* [2015]).

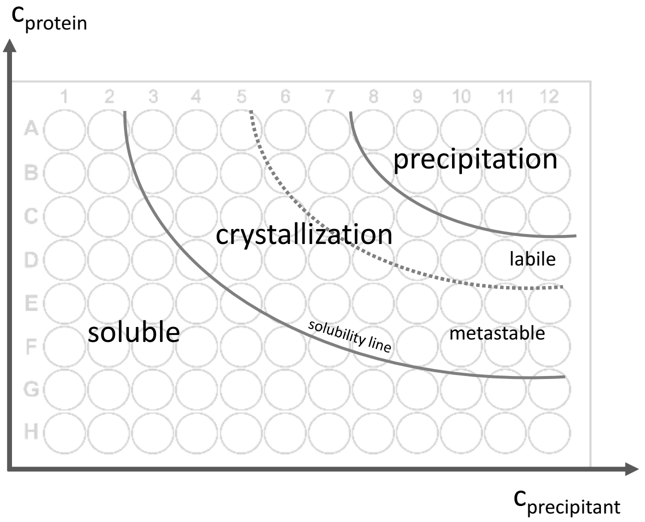


Figure 5: Schematic drawing of a protein phase diagram. It depicts the protein phase behavior under varying initial protein and precipitant concentration. All conditions below the solubility line are undersaturated and the protein stays soluble. At conditions above the solubility line, the solubility limit is exceeded, and crystallization and precipitation, respectively, occurs.

Crystallization as well as precipitation of biopharmaceutical products are widely used techniques during purification and formulation of these molecules. Protein crystals are highly pure and reveal a huge long-term stability. Precipitation of proteins and viruses is a frequent alternative purification process and may be achieved by adding sufficient concentrations of precipitants, e.g. salts, organic solvents, or organic polymers, or also through varying the pH, temperature, or concentration of the solution. Protein precipitation using an organic polymer, such as polyethylene glycol (PEG), has been widely used during purification of biomolecules including monoclonal antibodies (mAbs), viruses, and virus-like particles (Juckles [1971], Oelmeier *et al.* [2013], Tsoka *et al.* [2000]). PEG offers some advantages over other precipitants, as it is inert, non-flammable, non-toxic, uncharged, and relatively unexpensive (Janson [2011], Sim *et al.* [2012a]). Matheus *et al.* [2009] demonstrated that the native secondary structure and activity of a mAb were preserved after precipitation by PEG4000 and subsequent re-dissolution of the precipitate. This means that the precipitation with PEG presumably does not induce conformational changes of proteins and the biological activity is maintained.

There are two theories applied to describe the PEG-induced precipitation in a mechanistic way, namely, the theory of attractive depletion (Asakura and Oosawa [1958], Odijk [2009]) and the theory of excluded volume (Iverius and Laurent [1967], Polson [1977]). The attractive depletion theory assumes that the PEG's center of mass is excluded from the vicinity of the protein surface due to its size and structure and, hence, creates a 'depletion zone'. When two neighboring protein molecules get sufficiently close to each other, the depletion zones overlap and an additional volume is recovered for the polymer. This results in an increasing entropy and a decreasing free energy which leads to a thermodynamically driven aggregation of protein molecules (Lee *et al.* [2012], Tardieu *et al.* [2002]). This mechanism is illustrated in Figure 6.

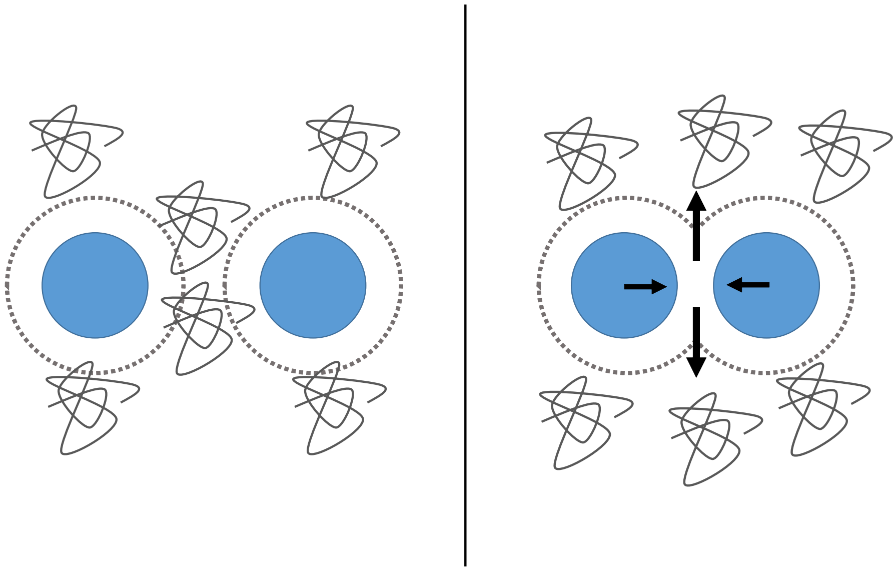


Figure 6: Schematic representation of the PEG induced attractive depletion in a protein solution. The PEG molecules (grey) are excluded from the vicinity of the proteins (blue). Additional volume for the polymer is recovered by the overlap of the depletion zones of two protein molecules.

The excluded volume theory, by contrast, is based on the assumption of protein molecules being sterically excluded from the volume of PEG molecules, which means that the concentration of proteins gets highly increased in the remaining volume of the solution. Aggregation and precipitation of proteins occurs when the solubility limit of the protein is exceeded (Atha and Ingham [1981], Knevelman *et al.* [2009]). For proteins the solubility S decreases exponentially with increasing concentration of PEG (c_{PEG}) according to Equation 2.

$$log\ S = log\ S_0 - \beta \cdot c_{PEG}. \tag{2}$$

In this equation S_0 represents the apparent intrinsic protein solubility in the absence of PEG, β the slope of the precipitation curve in the region where precipitation occurs. The threshold PEG concentration, at which protein solubility equals the protein concentration initially set, is referred to as m^*. All parameters are derived from the Cohn equation that describes the salting-out effect of salts on proteins and can be applied analogously to precipitation curves with PEG (Cohn [1925], Przybycien and Bailey [1989], Sim *et al.* [2012b]).

1.6.1.3 Prediction of Colloidal Stability

There are several parameters for the measurement of the tendency of protein-protein self-association, such as the second osmotic virial coefficient B_{22} (Ahamed *et al.* [2007]), the diffusion coefficient D (Saluja *et al.* [2007b]), the conformational flexilbility of the protein structure (Galm *et al.* [2016]), the ratio of the rheological parameters G' and G'' (Schermeyer *et al.* [2016]), or the precipitation with polyethylene glycol (Gibson *et al.* [2011]). The B_{22} quantifies the thermodynamic non-ideality of the diluted protein solution

and characterizes solute-solute interactions and was found to correlate well with protein solubility in dilute (George *et al.* [1997], Valente *et al.* [2005]). It can be determined experimentally by static light scattering or self-interaction chromatography. The value of B_{22} reflects the magnitude of the deviation from ideality and its algebraic sign reflects the nature of this deviation. A positive algebraic sign indicates predominately repulsive interactions, whereas a negative algebraic sign reflects predominantly attractive interactions (Neal *et al.* [1998]).
As an alternative methodology for capturing protein-protein interactions, the diffusion coefficient can be employed. The diffusion coefficient D can also be applied to capture protein-protein interactions and hence be used as an indicator of the aggregation propensity. The value of D is described by the Stokes-Einstein equation (Equation 1). For a non-ideal solution, intermolecular interactions have an additional impact on the diffusion coefficient. Thus, the diffusion coefficient is expanded by a term representing protein-protein interactions:

$$D = D_0 \cdot (1 + k_D \cdot c_{prot}). \tag{3}$$

Here, D_0 depicts the diffusion coefficient of the protein at infinite dilution and k_D the diffusion interaction parameter, summarizing all protein-protein interactions (Kuehner *et al.* [1997], Mahler *et al.* [2009]). In general, a decrease of the apparent diffusion coefficient is an indicator for predominating attractive interactions, whereas an increase represents predominating repulsive interactions (Muschol and Rosenberger [1995]).
Galm et al. (Galm *et al.* [2016]) applied MD simulations to identify highly flexible protein regions which could be associated to less regular secondary structure elements and random coiled and terminal regions in particular. Conformational flexibility of the entire protein structure and protein surface hydrophobicity could be correlated to differing aggregation propensities among the studied proteins and could be applied for the prediction of protein phase behavior in aqueous solution without precipitants.
Schermeyer et al. (Schermeyer *et al.* [2016]) used oscillatory frequency sweep measurements of samples to determine the rheological parameters G' and G'' of a protein solution. For lysozyme, the ratio of these both parameters was correlated with the phase behavior of the same samples obtained from phase diagrams and both showed a good correlation.
Gibson et al. (Gibson *et al.* [2011]) reported an additional method to investigate the solubility of a protein solution and rank different formulation conditions in terms of pH and buffer ions. They therefor exposed a monoclonal antibody to a variety of buffer conditions to an increasing concentration of PEG and determined the remaining amount of protein in the supernatant. By comparison of the weight% PEG in solution, that is necessary to decrease the initial protein concentration by 50% (PEG$_{midpt}$) the protein solubility under the respective conditions could be assessed and compared. The further the PEG$_{midpt}$ value moves to higher PEG concentrations, the more colloidal stable the protein is.

1.6.2 Conformational Stability of Proteins and Viruses

Most proteins fold a specific globular conformation that is essential for their biologic functions. Conformational instabilities, this means changes in the native protein 3D structure, also favor the aggregation of proteins and viruses. Aggregates are most commonly formed from the interaction of partially unfolded species that still contain significant native-like structure (Fink [1998]). Aggregation is very likely for partially unfolded protein monomers as in most cases partial unfolding increases their hydrophobicity. This type of aggregation results in an loss of the protein's native state and, thus, leads to non-native aggregation (Chi *et al.* [2003]). Whether proteins or the surface proteins of viruses, respectively, are still in a native state or already partially unfolded can be determined by Fourier transform infrared (FT-IR) spectroscopy.

1.6.3 Biological Stability of Proteins and Viruses

The term of biological stability is often referred to as the preservation of the biological function of a protein or a virus. The most frequent measures of vaccine stability remain biological assays that demonstrate maintenance of sufficient immunogenicity to produce protective immunological responses in humans and animals. These methods are typically very time consuming, of low accuracy and precision, and provide no insight into mechanisms of destabilization. To overcome these limitations, high-resolution analytical methods are necessary that are able to detect small changes in complex macromolecular systems of multiple components (Brandau *et al.* [2003]). The biological stability of influenza viruses is assessed through the hemagglutination (HA) assay. The hemagglutinin surface protein of influenza virus particles is capable of binding to specific glycosylation patterns (N-acetylneuraminic acid-containing proteins) on avian and mammalian erythrocytes. If the influenza virus is present in a sufficient concentration, there is an agglutination reaction and cross-linking between the erythrocytes and the influenza viruses occurs and the sedimentation behavior of erythrocytes changes from point-like to carpet-like sedimentation. The HA assay does not necessarily indicate the presence of viable virus, but also the presence of degraded or inactivated virus particles (Kalbfuss *et al.* [2008], Killian [2008]).

2 Research Proposal

This research work is part of the project 'Optimization of an industrial process for the production of cell-culture-based seasonal and pandemic influenza vaccines' funded by the German Federal Ministry of Education and Research (BMBF). The entire research consortium focuses on the optimization of an alternative industrial process for the biotechnological production of seasonal and pandemic influenza vaccines using cell cultures as an alternative to traditional egg-based processes (see Figure 7). The focus is set on the implementation of novel innovative materials and methods during purification of the influenza viruses. These novel materials and methods should be applied during the development of integrated and intensified processes in terms of robustness, recovery, purity, reproducibility, and safety, and also for gaining a deeper understanding of the parameters influencing the aggregation behavior of influenza viruses and to establish continuous purification processes.

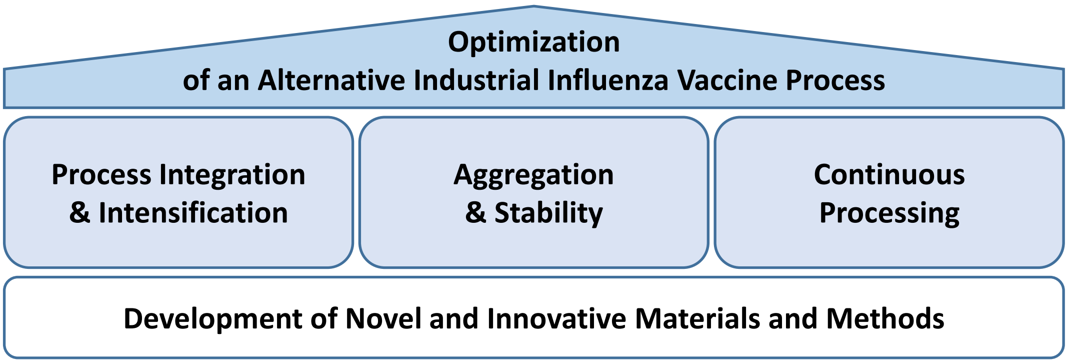

Optimization
of an Alternative Industrial Influenza Vaccine Process
Process Integration & Intensification
Aggregation & Stability
Continuous Processing
Development of Novel and Innovative Materials and Methods

Figure 7: Overview of the project 'Optimization of an industrial process for the production of cell-culture-based seasonal and pandemic influenza vaccines'.

This doctoral thesis focuses on the work package dealing with aggregation of influenza viruses during manufacturing, formulation, and storage. Aggregation of biopharmaceutical products, such as viruses and proteins, can occur at all stages during the production process. The aggregation propensity is severely influenced by a large number of environmental parameters that are subject to change in every step of manufacturing. As the surface of influenza viruses mainly consists of the surface proteins hemagglutinin and neuraminidase, it is assumed that the aggregation can be tackled from the perspective of proteins.

The first part of this work focuses on the challenge to generate models to assess and to predict the aggregation propensity of proteins through the methodology of quantitative structure-activity relationship (QSAR) modeling. QSAR modeling is employed as an *in silico* method to predict parameters that capture the tendency for aggregation of proteins that have to be determined experimentally so far. For proteins, the methodology of QSAR has not yet been applied to describe and to predict other parameters besides those for the binding of proteins during chromatography processes. This work aims to expand the application of the QSAR methodology to the novel field of protein aggregation, as it has proven its potential in the field of chromatography.

During a first project, a QSAR model for the prediction of a parameter that has proven the ability to capture protein-protein interactions will be developed. Thus, this model can be used as an indicator for the aggregation propensity. During a second project in this first part of the thesis, QSAR modeling is applied to enable the prediction of precipitation of proteins with with an organic polymer. Until today, the development of such precipitation processes is mainly based on heuristic and experimental approaches. This work enables to design an alternative purification step for proteins *in silico*, comprising precipitation with polyethylene glycol. Additionally, the aggregation propensity under different buffer conditions of proteins can be derived from the results of precipitation experiments. In the second part of this work, the stability of influenza viruses is experimentally assessed by applying high-throughput compatible methods and the surface properties of the virus particles leading to aggregation are revealed. As a basis for the systematic investigation of the influence of environmental parameters on the colloidal and biological stability of influenza viruses in solution, phase diagrams in microliter scale are generated and evaluated for the stability of the viruses under the respective environmental conditions. On the basis of these phase diagrams, a toolbox consisting of measurements of surface properties, and precipitation experiments will be proposed as a predictive tool for the colloidal and biological stability of influenza viruses under the respective conditions. In a second project of this part of the work, the influence of the production system on the surface properties and, thus, on the aggregation tendency of influenza viruses is subject of research. Influenza viruses produced in adherent and suspension Madin Darby canine kidney cells show differences in the aggregation behavior. An analytical toolbox is applied to obtain information about all surface characteristics, differences in lipid composition of the membrane, and glycosylation of the hemagglutinin surface protein. The experiments reveal the parameters that are responsible for the observed differences in aggregation behavior.

3 Publications & Manuscripts

1. **Influence of Structure Properties on Protein-Protein Interactions - QSAR Modeling of Changes in Diffusion Coefficients**

   Katharina Christin Bauer[‡], Frank Hämmerling[‡], Jörg Kittelmann, Cathrin Dürr, Fabian Görlich, Jürgen Hubbuch
   ([‡]: Contributed equally to this work)

   Biotechnology and Bioengineering 114 (2017) 821–831
   doi: 10.1002/bit.26210

   This article presents a quantitative structure-activity relationship model to predict the diffusion coefficients of proteins *in silico* sensitive to protein type, pH, and ionic strength. As the diffusion coefficient can be used as a measure for protein-protein interactions, this work allows to decide whether overall attractive or repulsive interactions are present. Hence, this work allows to predict the aggregation propensity of proteins.

2. **Investigation and Prediction of Protein Precipitation by Polyethylene Glycol Using Quantitative Structure-Activity Relationship Models**

   Frank Hämmerling, Christopher Ladd Effio, Sebastian Andris, Jörg Kittelmann, Jürgen Hubbuch

   Journal of Biotechnology 241 (2017) 87–97
   doi: 10.1016/j.jbiotec.2016.11.014

   This paper presents a novel *in silico* approach for predicting the complete precipitation curves for the precipitation of proteins with polyethylene glycol, using quantitative structure-activity relationship models. Models were genereated for two parameters, namely the discontinuity point m^* and the β-value, that fully describe the precipitation curve. This work enables to design precipitation steps for proteins as a purification method *in silico*. Furthermore, the value of m^* can be used as an indicator for long-term colloidal stability of biopharmaceutical products.

3. **Strategy for Assessment of the Colloidal and Biological Stability of H1N1 Influenza A Viruses**

   Frank Hämmerling, Oliver Lorenz-Cristea, Pascal Baumann, Jürgen Hubbuch

   International Journal of Pharmaceutics 517 (2017) 80–87
   doi: 10.1016/j.ijpharm.2016.11.058

   This study investigates the influence of initial H1N1 concentration, pH, and sodium chloride concentration on the colloidal and biological stability of H1N1 influenza A viruses at ambient temperatures to overcome the dependency of a continuous cold chain for vaccines. A toolbox was developed to rapidly assess both stabilities using high-throughput compatible methods. This methodology depicts a powerful tool for the development of optimized influenza vaccines with an increased stability at ambient temperatures.

4. **Influence of the Production System on the Surface Properties of Influenza A Virus Particles**

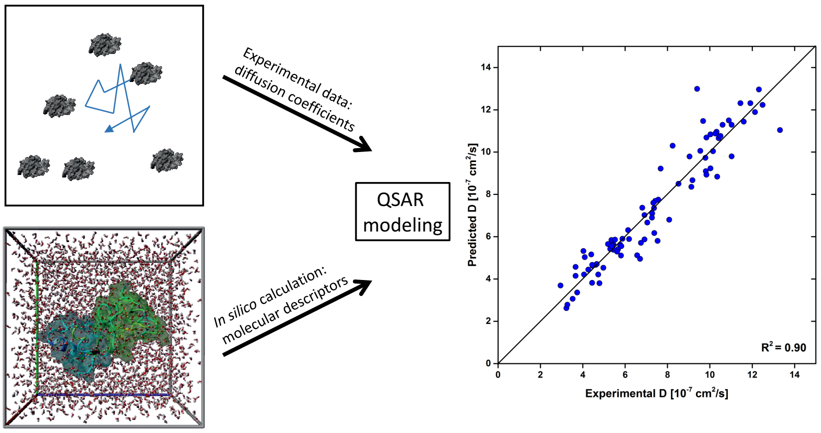
   Frank Hämmerling[‡], Michael M. Pieler[‡], René Hennig, Anja Serve, Erdmann Rapp, Michael W. Wolff, Udo Reichl, Jürgen Hubbuch
   ([‡]: Contributed equally to this work)

   Manuscript submitted to Engineering in Life Sciences

   This article presents the investigation of the influence of the production system on the surface properties of influenza A/Puerto Rico/8/34 H1N1 viruses. Virus particles produced in adherent MDCK cells show a significantly higher aggregation propensity as those produced in suspension MDCK cells. It was found that both virus particle samples have considerable different surface properties, resulting from a different glycosylation of the hemagglutinin surface protein. This work reveals the importance of carefully selecting the production system during the production of biopharmaceutical products to avoid aggregation and, thus, to meet critical quality attributes.

Influence of structure properties on protein-protein interactions - QSAR modeling of changes in diffusion coefficients

Katharina Christin Bauer[‡], Frank Hämmerling[‡], Jörg Kittelmann, Cathrin Dürr, Fabian Görlich and Jürgen Hubbuch*



Institute of Engineering in Life Sciences, Section IV: Biomolecular Separation Engineering, Karlsruhe Institute of Technology, Engler-Bunte-Ring 3, 76131 Karlsruhe, Germany

[‡] : *These authors contributed equally to this work.*
* : *Corresponding author; email address: juergen.hubbuch@kit.edu*

Biotechnology and Bioengineering

Volume 114, (2017), 821-831

Abstract

Information about protein-protein interactions provides valuable knowledge about the phase behavior of protein solutions during the biopharmaceutical production process. Up to date it is possible to capture their overall impact by an experimentally determined potential of mean force. For the description of this potential, the second virial coefficient B_{22}, the diffusion interaction parameter k_D, the storage modulus G', or the diffusion coefficient D is applied. *In silico* methods do not only have the potential to predict these parameters, but also to provide deeper understanding of the molecular origin of the protein-protein interactions by correlating the data to the protein's three-dimensional structure. This methodology furthermore allows a lower sample consumption and less experimental effort. Of all *in silico* methods, QSAR modeling, which correlates the properties of the molecule's structure with the experimental behavior, seems to be particularly suitable for this purpose. To verify this, the study reported here dealt with the determination of a QSAR model for the diffusion coefficient of proteins. This model consisted of diffusion coefficients for six different model proteins at various pH values and NaCl concentrations. The generated QSAR model showed a good correlation between experimental and predicted data with a coefficient of determination $R^2 = 0.9$ and a good predictability for an external test set with $R^2 = 0.91$. The information about the properties affecting protein-protein interactions present in solution was in agreement with experiment and theory. Furthermore, the model was able to give a more detailed picture of the protein properties influencing the diffusion coefficient and the acting protein-protein interactions.

Keywords: Quantitative Structure-Activity Relationship, PDB, Electrostatic Interactions, Hydrophobic Interactions, Protein Size, Protein Shape

1 Introduction

Protein-protein interactions govern the phase behavior, or more precisely, physical properties such as solubility or viscosity of a biopharmaceutical protein solution. Already small changes in these properties can affect the outcome of each process step until the final product is obtained. A decrease in solubility, for example, can provoke aggregation, whereas an increase of viscosity can inhibit processability. In both of these cases, product loss can be the consequence [1, 2]. To predict or prevent these changes, protein as well as protein-solvent interactions have to be understood. On a molecular level, protein-protein interactions are based on the protein's configuration as well as on its surface patches with their specific properties, meaning its electrostatic charge and hydrophobicity. Depending on the solution conditions, these specific surface patches change and interact differently with their surrounding [3]. Electrostatic interactions can act attractively or repulsively over long-range distance. At short-range distance additional interactions can have an impact. These interactions are attractive van der Waals and hydrophobic interactions as well as repulsive hydration forces [4, 5, 6]. Yet researchers are able to experimentally determine an overall potential of all these acting forces, called the potential of mean force [7]. The potential of mean force can be derived from one physical solution property and its deviation from ideality. This deviation is usually represented by parameters, such as the second virial coefficient B_{22} [8, 9] or the diffusion interaction parameter k_D [10] for dilute solutions, the storage modulus G' [11] for highly concentrated solutions, or the mutual diffusion coefficient D [12, 13, 14] for dilute, represented by k_D, as well as highly concentrated protein solutions. Using this approach, scientists can capture the overall change in interactions, but they cannot correlate them to their origin on the protein surface [7].

Computational methods, so-called *in silico* methods, which use the protein structure as basic information, have the potential to fill this gap by correlating the three-dimensional molecule structure to the overall potential gained in experiments. A highly suitable approach is to use quantitative structure-activity relationship (QSAR). The principal aim of this method is to predict experimental properties of a compound based on the molecular structure. QSARs work on the assumption that structurally similar compounds have similar activities and therefore have predictive abilities [15]. QSARs still are mainly applied for small molecules during the development of bioactive compounds [16]. During the last two decades, QSAR models were successfully used to describe and to predict the experimental behavior of proteins and complex biopharmaceutical products during ion-exchange [16, 17], mixed-mode [18, 19] and hydrophobic interaction [20] chromatography. Buyel et al. [21] used QSAR to predict the chromatographic separation of tobacco host cell proteins out of a complex feedstock.

This work aimed at extending use of QSAR modeling for proteins from chromatography to stability and processability of protein solutions during downstream processing and storage. For this purpose, the capability of QSAR modeling to predict protein-protein interactions from protein structure properties was examined. Furthermore, the ability to create a deeper understanding of the mechanisms affecting protein-protein interactions was considered. For the investigation of protein-protein interactions, the apparent diffusion coefficients of six different globular proteins, namely, α-lactalbumin, lysozyme, β-lactoglobulin, ovalbumin, BSA, and glucose oxidase, with a concentration of 10 mg/mL at varying pH values and NaCl concentrations were determined. These data were used to build a QSAR model. Apart from the predictive capacity of this QSAR model, its information about the protein-protein interactions having an impact on the value of the apparent diffusion coefficient was evaluated.

2 Materials and Methods

In this section the materials and methods for building a QSAR model to describe and predict the diffusion coefficient of different proteins at various pH values and NaCl concentrations are explained. It covers the preparation of the buffers as well as protein solutions, the determination of the diffusion coefficient by dynamic light scattering, and the QSAR modeling.

2.1 Buffers and Protein Solutions

Buffer stock solutions with and without NaCl were prepared for pH 3, 5, 7, and 9. The buffer components were citric acid (Merck KGaA, Darmstadt, Germany) and sodium citrate (Sigma-Aldrich, St. Louis, MO, USA) for pH 3, acetic acid (Merck KGaA) and sodium acetate (Sigma-Aldrich, St. Louis, MO, USA) for pH 5, MOPSO (AppliChem GmbH, Darmstadt, Germany) for pH 7, and BisTris (Sigma-Aldrich) for pH 9. Without addition of NaCl, each buffer stock solution had an ionic strength of 100 mM. For the stock solutions with NaCl, 2.5 M NaCl (Merck KGaA) were weighed in with the rest of the components. The pH was controlled using a five-point calibrated pH meter (HI-3220, Hanna® Instruments, Woonsocket, RI, USA) equipped with a SenTix® 62 pH electrode (Xylem Inc., White Plains, NY, USA) and corrected by titration of hydrochloric acid or sodium hydroxide with an accuracy of ±0.05. Both chemicals were purchased from Merck (Darmstadt, Germany). Each buffer was filtrated with a 0.22 μm cellulose acetate membrane (Sartorius AG, Göttingen, Germany). The buffers were used at constant pH 24 h after preparation. Lysozyme from chicken egg-white was purchased from Hampton Research (Aliso Viejo, CA, USA). α-lactalbumin from bovine milk, β-lactoglobulin from bovine milk, ovalbumin, bovine serum albumin (BSA), and glucose oxidase were purchased from Sigma-Aldrich. Each protein was weighed in and diluted with the buffer stock solution without salt at the respective pH. The protein solutions were filtered through 0.22 μm syringe filters with cellulose acetate membrane (VWR, Radnor, PA, USA). By centrifugation with Vivaspin® centrifugal concentrators (Sartorius AG) with polyethersulfone membrane, the solutions were desalted until 99.9 % of the solution were exchanged and then concentrated. Protein concentration was determined photometrically with a NanoDrop™ 2000c UV-Vis spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA). The respective extinction coefficients were $E^{1\%}$(280 nm) = 20.01 L g^{-1} cm^{-1} for α-lactalbumin, $E^{1\%}$(280 nm) = 22.00 L g^{-1} cm^{-1} for lysozyme, $E^{1\%}$(280 nm) = 7.65 L g^{-1} cm^{-1} for β-lactoglobulin, $E^{1\%}$(280 nm) = 6.90 L g^{-1} cm^{-1} for ovalbumin, $E^{1\%}$(280 nm) = 5.72 L g^{-1} cm^{-1} for BSA, and $E^{1\%}$(280 nm) = 16.07 L g^{-1} cm^{-1} for glucose oxidase. The samples of 10 mg/mL at different pH values and NaCl concentrations were prepared by mixing the protein stock solution with the buffer stock solutions with or without NaCl of the respective pH.

2.2 Dynamic Light Scattering (DLS)

Dynamic light scattering (DLS) measurements are based on the interference of the scattered light by diffusing particles in solution. This method is mainly used to determine the size and size distribution of these diffusing particles based on the Stokes-Einstein equation:

$$D = \frac{k_B T}{4\pi r_h \eta_S}. \tag{1}$$

In this equation for the ideal dilute state the diffusion coefficient D of a scattering particle depends on its hydrodynamic radius r_h, the viscosity of the surrounding solution η_S and the thermal energy $k_B T$. For a non-ideal solution, such as protein solutions, intermolecular interactions have an additional impact on the diffusion coefficient. For this purpose the diffusion coefficient is expanded by a term representing protein-protein interactions:

$$D = D_0(1 + k_D \cdot c_{prot}). \tag{2}$$

In this equation D_0 is the diffusion coefficient of the protein at infinite dilution and k_D the diffusion interaction parameter summarizing all protein-protein interactions [22, 23].

2.3 Principle of Determining Changes in Interactions by DLS

As described in the previous section the principle of determining changes in interactions by dynamic light scattering is based on the changes of the determined diffusion coefficient due to protein-protein interactions. In general, a decrease in the apparent diffusion coefficient can be interpreted as predominating attractive interactions, an increase suggests predominating negative interactions in solution [13]. For the purpose of our work, we determined diffusion coefficients at a constant concentration of 10 mg/mL for different proteins, namely, α-lactalbumin, lysozyme, β-lactoglobulin, ovalbumin, BSA, and glucose oxidase at pH 3, 5, 7, and 9 and NaCl concentrations between 0 and 1.82 M. By the changes of the diffusion coefficient depending on the respective condition, changes of present interactions in solution were determined.

To exclude that observed changes in diffusion coefficient D are solely the effect of a perturbation on the diffusion coefficient at infinite dilution D_0, this parameter was calculated and determined experimentally for selected proteins and conditions. $D_{0,calc}$ was determined with the correlation that relates D to the molecular weight published by Young et al. [24]. The experimentally determined $D_{0,exp}$ was extrapolated to infinite dilute protein concentration from diffusion coefficients determined at several protein concentrations according to Saluja et al. [11]. At each pH the respective values of D_0 for all investigated NaCl concentrations were averaged and the standard deviation was calculated.

2.4 DLS Measurements

Dynamic light scattering (DLS) measurements of the protein solutions were conducted in triplicate with the high-throughput compatible Wyatt Technology DynaPro™ Plate Reader (Wyatt Technology Corporation, Santa Barbara, CA, USA). For each measurement, the sample volume of 30 μL was pipetted into one well of a Corning® Low Volume

384 Well Microplate NBS™ (Corning Incorporated, Tewksbury, MA, USA) and covered by silicon oil WACKER® AK 20 (Wacker Chemie AG, Munich, Germany) to prevent evaporation. Each measurement consisted of 10 acquisitions for 5 s at 23 °C. The apparent diffusion coefficient of the respective protein was determined by the distributional result of the DYNAMICS® Software Version 7.1.7.16 (Wyatt Technology Corporation) and averaged over the three measured wells of the same sample.

2.5 QSAR Modeling

2.5.1 Protein Structure Preparation

According to protein name and organism, the UniProtID for all proteins was obtained from UniProt [25]. All PDB files were downloaded from the RCSB Protein Data Bank [26]. The specific IDs can be found in Table 1.

Table 1: PDB ID, pI, and molecular weight of the proteins used in this study.

Protein Name	PDB ID	pI	Molecular weight [kDa]
α-lactalbumin	1F6S	4.5	14.2
Lysozyme	1LYZ	11.0	14.9
β-lactoglobulin	2AKQ	4.9	18.4
Ovalbumin	1OVA	4.5	44.3
BSA	3V03	4.9	66.4
Glucose oxidase*	1CF3	4.2	160.0

*Dimer created with SWISS-MODEL (http://swissmodel.expasy.org)

In Yasara [27], a software for visualization, modeling of molecules, and molecular dynamics simulations, a protein structure reflecting the conditions in solution was generated. Therefor the structures were checked for completeness and, if necessary, missing residues or intramolecular disulfide bonds were added manually. The hydrogen bonding network was optimized and an energy minimization experiment was conducted using the Amber03 force field [28]. Heteroatoms were separated from the protein structure and the protonation of amino acids was executed in H++ [29] according to the respective pH value and ionic strength. After protonation of amino acid residues, the heteroatoms were inserted again. Using the Amber03 force field another energy minimization and molecular dynamics (MD) simulation experiment were performed. The 10 ps MD simulation experiment was carried out at 298 K, the size of the simulation box was extended 10 $\AA$ on every side of the protein, periodic boundaries were chosen and snapshots were taken every 1 ps and averaged afterwards. This averaged structure was then used for the calculation of molecular descriptors. Glucose oxidase, which exists as a dimer under the studied conditions, was assembled by two monomers with the help of SWISS-MODEL [30].

2.5.2 Calculation of Molecular Descriptors

The 'mantoQSAR' software developed in-house was used for the calculation of molecular descriptors based on the averaged PDB structure after the MD simulation. It accounts for molecular structure, electrostatic and hydrophobic properties of the proteins at distinct pH values and ionic strengths. The group of molecular structure properties descriptors include all descriptors derived from geometric data of proteins, such as protein size, number of amino acids, protein shape and others. The hydrophobic properties are calculated using the hydropathy score published by Kyte and Doolittle [31]. For a detailed breakdown of each of these properties, four different types of descriptors are defined:

1. Full molecule descriptors: This set of descriptors comprises the complete molecule and calculates properties for the overall molecule's structure.
2. Plane descriptors: A number of 120 planes is tangentially approached to the protein molecule's surface until a set distance of 5 $\mathring{A}$ between the protein and the plane. This distance is adapted from previous work published by Dismer et al. [32] and Lang et al. [33]. For this study a set of 120 plane orientations, randomly distributed along the protein surface, was chosen and respective descriptors calculated for each orientation.
3. Patch descriptors: Patch descriptors only account for a part of the molecule and only calculate the values for the selected part ("patch") of the molecule. The size of the protein surface patch considered for calculation of the patch descriptors was derived from the calculated planes: based on the orientation of the planes, solvent-accessible protein surface area within a distance below 20 $\mathring{A}$ was taken into account for calculation of molecular descriptors and thus only parts of the molecule are represented.
4. Shell descriptors: The calculated descriptor values obtained from all 120 plane orientations are summed up to gain a 'shell projection', representing the properties at a distance of 5 $\mathring{A}$ around the molecule.

2.5.3 Multi-variate Data Analysis & Modeling

Partial least squares regression (PLSR) was used for QSAR modeling of the diffusion coefficient D. For this purpose, the complete data set with 94 observations and the associated descriptor values from mantoQSAR was split into a training and a test set. The training set containing 84 observations was used to build a QSAR model. This resulting model was then applied to the test set containing 10 observations. The experiments for the test set were randomly chosen, considering that the observations were located within the borders of the PLSR score scatter plot. During the first step, all 251 molecular descriptors were used to calculate an initial crude model with the training set data. Descriptors with a significant influence on protein diffusion coefficients were chosen according to the value of the variable influence on the projection (VIP). The VIP is a parameter that summarizes the importance of the X-variables to the X- and Y-models. Descriptors with a VIP value > 1 are deemed to contribute strongly to the resulting PLSR model [34]. Based on the 68 selected descriptors of the first crude model with a VIP value > 1, a final PLS model was created and then applied to predict the diffusion coefficients of the training set. This model had its own new VIP values, whose interpretation allowed for the generation of a mechanistic understanding. To exclude a random correlation of the selected molecular descriptors and the diffusion coefficients, a response permutation (Y-scrambling) with the final QSAR model was performed. The X-dataset, including the descriptors, was left intact, while the Y-dataset, including the observations, was randomly re-ordered 100 times. For each of the 100 Y-permutations, the data were PLSR-modeled. The correlation of X- and Y-data was assessed through the resulting coefficients of determination R^2 and the predictive capabilities of the respective model through the value of Q^2 [35, 36].

3 Results

This section presents the results for the diffusion coefficients as well as the QSAR model for the different proteins at various pH values and NaCl concentrations. The results for the QSAR model cover the training and the test. To underline the correlation between the surface properties of the proteins, captured by 68 descriptors, and the diffusion coefficient, the permutation plot is depicted.

3.1 Diffusion Coefficients

To examine protein-protein interactions, the diffusion coefficient was determined. Figure 1 displays the diffusion coefficients as well as calculated and experimentally determined diffusion coefficients $D_{0,calc}$ and $D_{0,exp}$ of α-lactalbumin and lysozyme at selected conditions.

$D_{0,calc}$ had a value of $10.1 \cdot 10^{-7}$ cm^2/s for α-lactalbumin and $9.9 \cdot 10^{-7}$ cm^2/s for lysozyme. The values for $D_{0,exp}$ varied depending on protein and pH value. For α-lactalbumin these were 12.6, 13.2, and $12.9 \cdot 10^{-7}$ cm^2/s for pH 5, 7, and 9, for lysozyme 11.5, 11.7, and $11.1 \cdot 10^{-7}$ cm^2/s for pH 3, 5, and 7. The standard deviation for all these values was below $0.6 \cdot 10^{-7}$ cm^2/s. The determined diffusion coefficients for α-lactalbumin and lysozyme varied dependent on protein type, pH and NaCl concentration. These values are also displayed in Figure 2 that shows the apparent diffusion coefficient of the studied proteins, namely, α-lactalbumin, lysozyme, β-lactoglobulin, ovalbumin, BSA, and glucose oxidase, with a concentration of 10 mg/mL at pH 3, 5, 7, and 9 and NaCl concentrations between 0 and 1.82 M. In these experiments the diffusion coefficient decreased with increasing NaCl concentration at constant pH. Furthermore, the diffusion coefficient varied depending on the pH value. In all experiments the standard deviation was below 1.47 $\cdot 10^{-7}$ cm^2/s . This maximum value was determined for ovalbumin at pH 5 with 1.82 M NaCl.

By looking at the results individually the diffusion coefficient of α-lactalbumin at 0 M NaCl had a value around $11 \cdot 10^{-7}$ cm^2/s for pH 5, 7 and 9. The values for pH 3 were neglected, because the protein formed a molten globule state [37]. These changes in tertiary and quaternary structure can not be described by *in silico* simulation experiments. For high NaCl concentrations, this protein precipitated at all studied pH values. For this reason, no diffusion coefficients were determined. For lysozyme, the values of D at 0 M NaCl were within the same range as for α-lactalbumin, but precipitation could only be observed at pH 3. For β-lactoglobulin, the values at 0 M NaCl were lower and varied with pH. The highest value was measured at pH 7, followed by pH 3 and pH 9. Precipitation occurred for high NaCl concentrations at pH 3. At pH 5, β-lactoglobulin was not soluble, which is why no values were obtained. Ovalbumin at 0 M NaCl showed diffusion

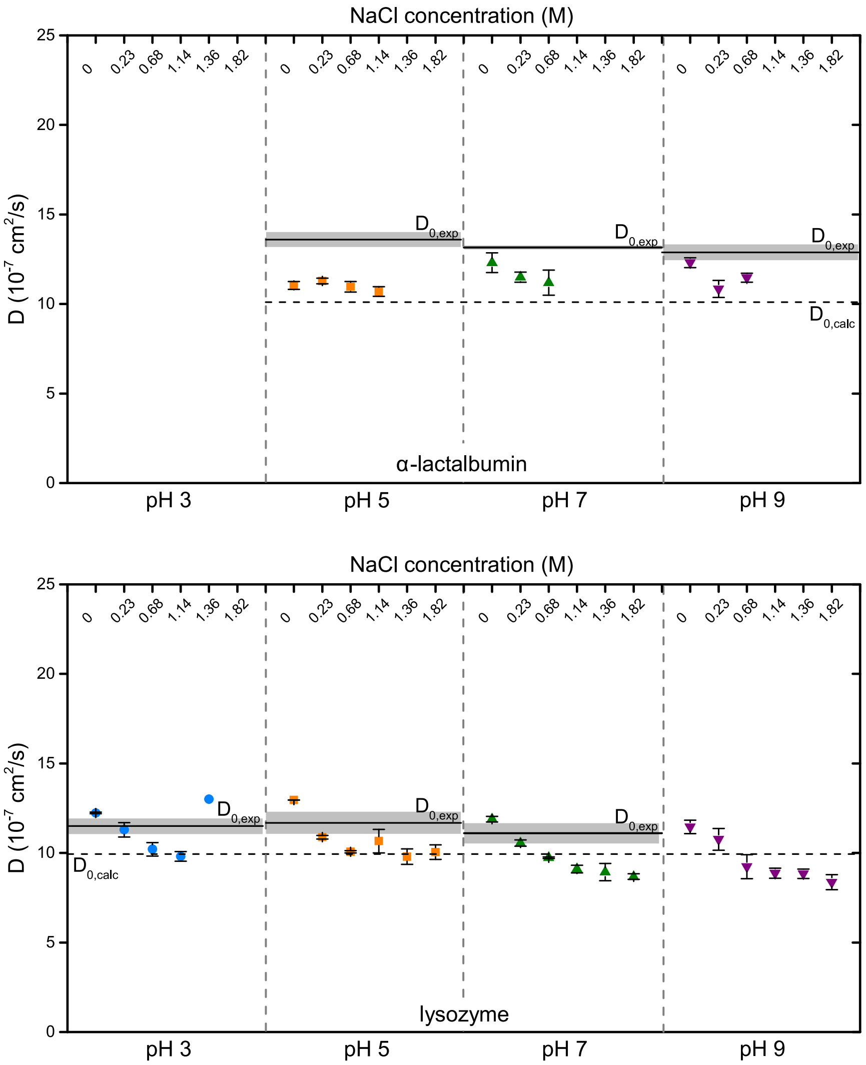


Figure 1: Diffusion coefficients at infinite dilution D_0 and a protein concentration of 10 mg/mL of α-lactalbumin and lysozyme at selected conditions. The solid line represents the experimentally determined $D_{0,exp}$ and the standard deviation colored in grey, and the dashed line the calculated $D_{0,calc}$.

coefficients around 6.6·10^{-7} cm^2/s with a maximum value of 7.6·10^{-7} cm^2/s for pH 7. In comparison to β-lactoglobulin, these values were lower. For pH 5, the value of the diffusion coefficients depending on NaCl concentration was nearly constant. The values for pH 3 had to be neglected for the same reason as for α-lactalbumin. The molten globule state of ovalbumin under this condition [38] could not be modeled by Yasara. For BSA, no strong influence of pH and NaCl concentration could be detected, with the exception of pH 3. Under this condition, precipitation could be observed at high NaCl concentrations. Almost the same behavior was found for glucose oxidase. For this protein, no data is shown for pH 3, because precipitation occurred directly upon addition of NaCl [39].

3.2 QSAR Modeling

A QSAR model was built as described in section 2.5.3 with the molecular descriptors and the experimentally determined diffusion coefficients from section 3.1. The best re-

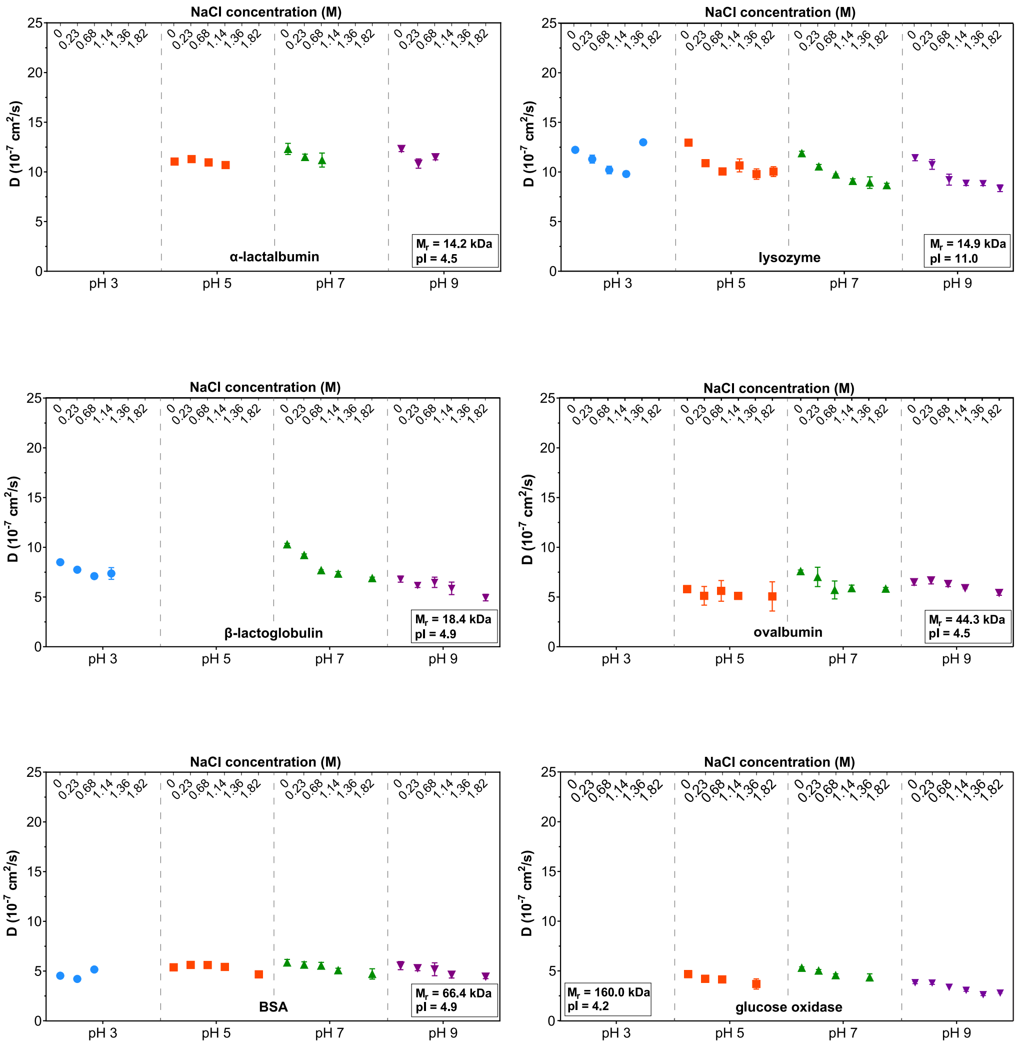


Figure 2: Diffusion coefficients of α-lactalbumin, lysozyme, β-lactoglobulin, ovalbumin, BSA, and glucose oxidase at 10 mg/mL for NaCl concentrations between 0 and 1.82 M and pH 3, 5, 7, and 9.

sulting model contained 68 molecular descriptors and consisted of four PLS components. Comparison of experimental and predicted data of the training set is shown in Figure 3 with a coefficient of determination R^2 of 0.91 and a predictability Q^2 of 0.88. The R^2 value is considered as a measure for the strength of the association between the observed and predicted observations, while the cross validation square correlation coefficient Q^2 is a measure for the predictability of the model. The root mean square error of cross-validation (RMSECV) was 0.98 $\cdot 10^{-7}$ cm^2/s.

This model was used for the prediction of the diffusion coefficients from the external test set, consisting of ten experiments that had been excluded from the training set. Figure 4 shows the experimental and the predicted data for these 10 conditions with a coefficient of determination R^2 of 0.91.

For an additional assessment of the statistical significance of the predictive power, a response permutation (Y-scrambling) was performed. Randomization of Y-data while keeping the X-data intact results in the generation of 100 "scrambled" models, each with a respective R^2 and Q^2 that are displayed in Figure 5. Both values for the scrambled

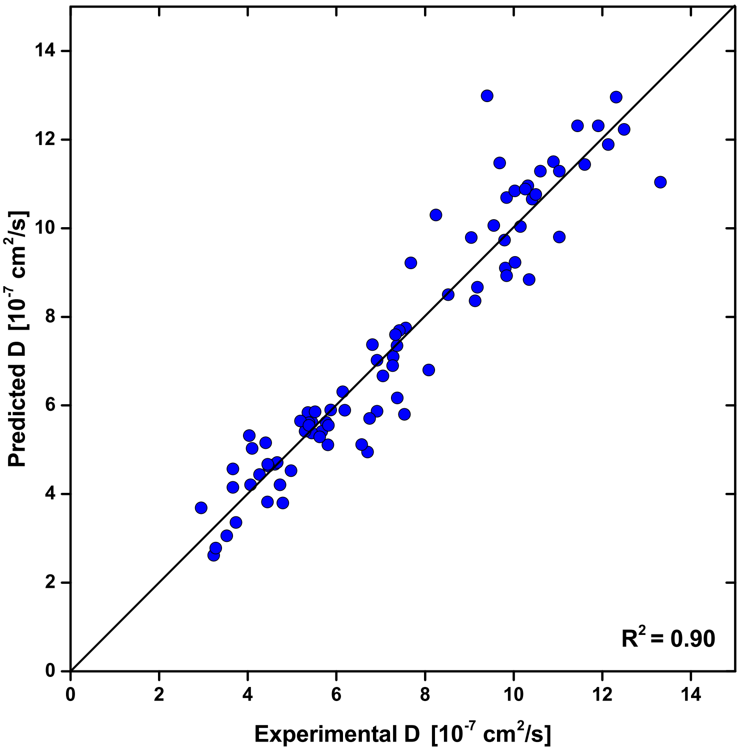


Figure 3: QSAR model of the training set: Experimental vs. predicted values of the diffusion coefficient.

models were compared with the values of the real model. All values for R^2 and Q^2 are lower for the scrambled models.

In order to evaluate the descriptors with the highest impact on the model, a VIP plot was created. It shows the VIP values and the respective regression coefficient for each descriptor in Figure 6. Descriptors with a VIP > 1.0 are considered to have a strong influence on the target figure. Descriptors with values below 1.0 have a minor impact [34]. The sign of the regression coefficient indicates the direction of the influence. Descriptors with a positive regression coefficient are proportional to the value of the diffusion coefficient, negative regression coefficients are inversely proportional [40].

The three descriptors with the highest VIP value were found to represent the electrostatic surface potential (ESP), the total surface area of the protein, and the solvent-accessible surface area of the protein patch with the lowest hydrophobicity. The 20 descriptors with VIP values > 1.0 are listed and explained in Table 2.

4 Discussion

As mentioned in the Introduction, several parameters can be used to describe protein-protein interactions in solution. For this study, the diffusion coefficient was selected to directly correlate a physical solution property to protein structure properties without further manipulation of data. To avoid an additional uncertainty that downgrades the quality of the QSAR model, the diffusion interaction parameter k_D was not considered as an alternative. The use of this parameter would require concentration-dependent linearity of the diffusion coefficient. Especially at conditions where additional short-

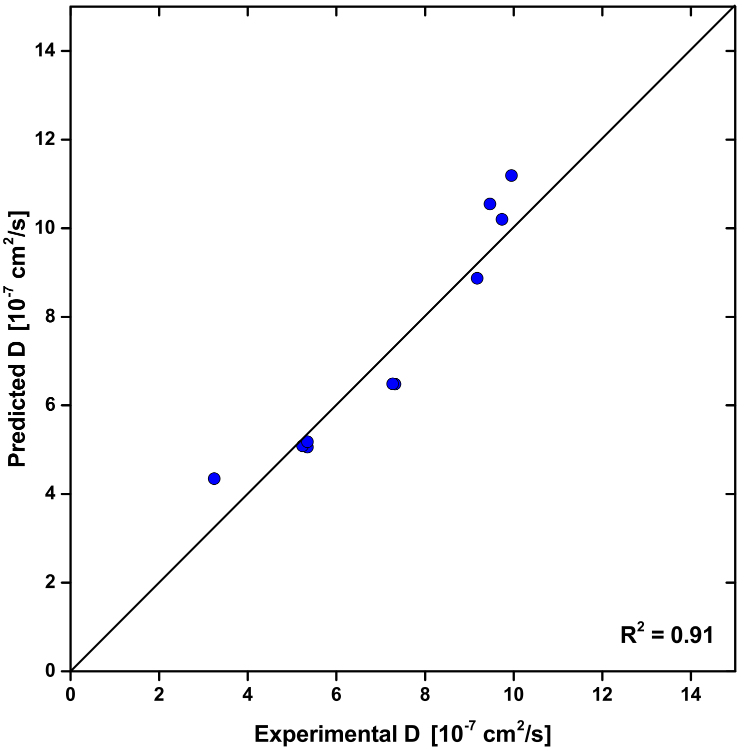


Figure 4: External validation of the QSAR model with the training set: Experimental vs. predicted values of the diffusion coefficient.

range interactions have an impact, such as for high protein or salt concentrations, this state of ideal dilution and, thus, concentration-dependent linearity cannot be commonly assumed [13, 41, 14].

4.1 Protein-protein Interactions Obtained by Determination of Diffusion Coefficients

When looking at the diffusion coefficients and the impact of protein-protein interactions, all parameters that can have an impact on these values need to be considered. According to the Stokes-Einstein relation, the diffusion coefficient depends on the hydrodynamic radius of the protein, the temperature, and the viscosity of the solvent (Equation 1). Whereas temperature and viscosity of the solvent were constant in this study, the hydrodynamic radius, which depends on the shape and size of the protein, could have an impact. As we only used globular proteins, the shape was supposed to have a negligible impact when interpreting the investigated data. In this study, the proteins with a higher molecular weight showed a lower diffusion coefficient compared to those with a lower molecular weight, following the Stokes-Einstein equation.

For protein solutions, apart from these influencing parameters for the ideal state reflected by Stokes-Einstein, interactions have to be taken into account. For this purpose this diffusion coefficient for the ideal state is complemented by a virial expansion resulting in Equation 2 where D_0 is the diffusion coefficient of one particle in solution at infinite dilution. This parameter exclusively is a function of particle size, shape, and the surrounding solvent [42, 23]. D_0 is fairly constant for one protein under the conditions investigated in

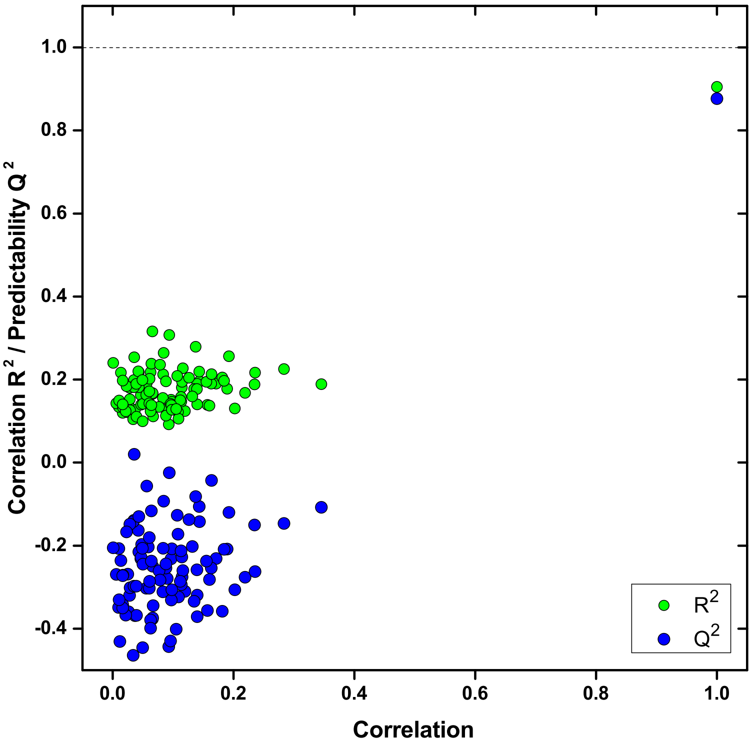


Figure 5: Permutation plot for the randomized Y-vector displaying the respective correlation R^2 and predictability Q^2.

this study. The values for D at 10 mg/mL differ significantly from D_0 and its perturbation. Therefor the observed differences in the diffusion coefficients displayed in Figure 1 are due to changes in the diffusion interaction parameter k_D. These interactions varied when changing the pH or adding NaCl. In theory, variation in pH changes the electrostatic charge distribution on the protein surface by protonation or deprotonation of amino acid side chains. Far from isoelectric point (pI), at dilute state, electrostatic interactions predominate. Due to their long-range repulsive nature, these interactions prevail over short-range interactions and cause an increase of the diffusion coefficient. Nevertheless, short-range interactions are present and also influence the diffusion coefficients of the proteins. These interactions include attractive van der Waals and hydrophobic interactions as well as repulsive hydration forces [5]. Close to the pI, the electrostatic net charge of a protein is close to zero. Here, attractive short-range interactions have an increasing impact. The overall potential of these forces can cause an attraction of the proteins, which is reflected by a lower diffusion coefficient [6, 11]. For the experimental data of this study, this theoretical decrease of repulsive interaction towards the pI was observed at 0 M NaCl for α-lactalbumin from pH 7 to pH 5, lysozyme from pH 5 to pH 7, and for ovalbumin, BSA, and glucose oxidase from pH 7 to pH 5. In contrast, pH values far from the pI deviated from this theory. The values for the diffusion coefficient for lysozyme at pH 3 and for β-lactoglobulin, ovalbumin, BSA, and glucose oxidase at pH 9 did not further increase, which indicates an increase in attractive interactions under this conditions. One reason for this increase far from the pI could be the strong deprotonation or protonation of the protein surface, which promotes an increase in hydrophobic surface area [43].

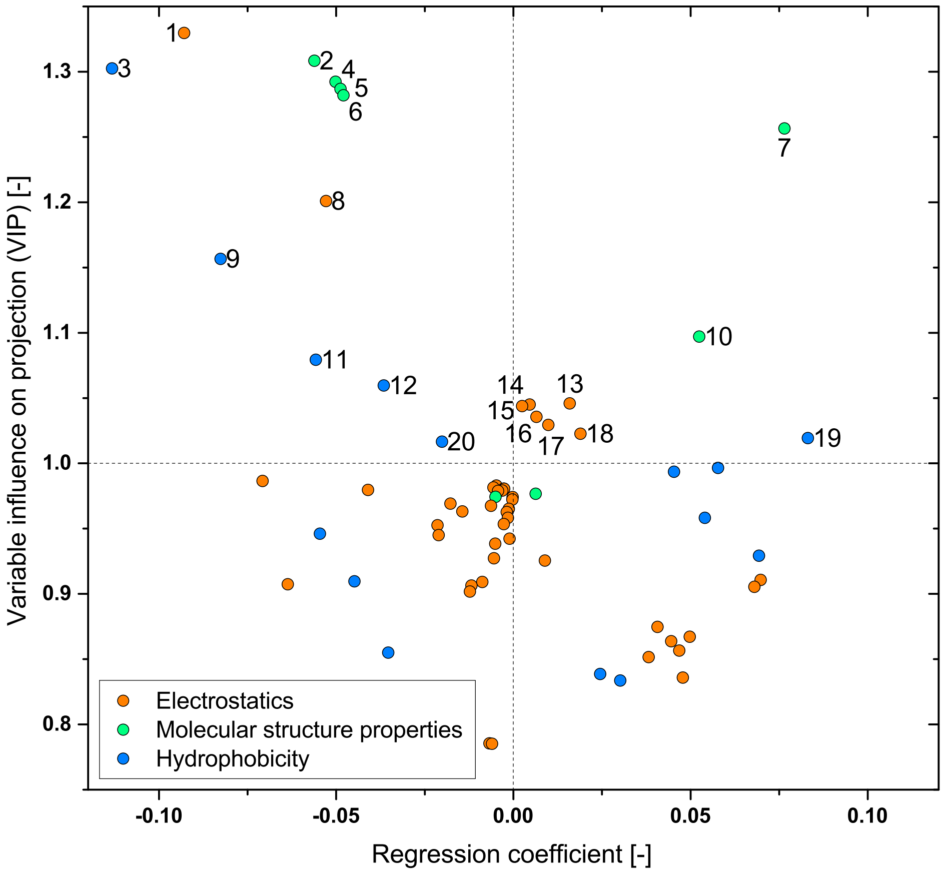


Figure 6: VIP values and regression coefficients for all 68 descriptors of the final QSAR model. The 20 descriptors with a VIP value > 1 are numbered and described in Table 2.

In contrast to the changes in pH, variation of NaCl concentration causes electrostatic shielding of the charged surface patches. As a result, electrostatic interactions decrease and the impact of short-range interactions, such as hydrophobic interactions, are promoted [44]. This effect is reflected by a decrease in the value of the diffusion coefficient with increasing NaCl concentration [45]. For the diffusion coefficients determined in this study, this observation could be seen for all proteins at constant pH. Precipitation caused by strong attractive interactions [46] occurred for α-lactalbumin at pH 9 and for lysozyme as well as BSA at pH 3.
In summary, the results for the diffusion coefficient in this study provided valuable information about the interactions in solution for each protein and its respective condition. With their variety in proteins, their size, pH values, and NaCl concentrations, the data seemed suitable for building a sound QSAR model.

4.2 Evaluation of QSAR Modeling

For the description of the diffusion coefficient by QSAR modeling, the calculated molecular descriptors are considered to take into account all protein properties as well as changes in pH and ionic strength. For this study, a set of 68 descriptors represented the molecular properties, which determined the value of the diffusion coefficient of the respective protein. For the training set, the results for the predicted values of the diffusion coefficient compared with the experimentally determined values were taken from Figure 3. With a coefficient of determination R^2 of 0.90, prediction agreed well with the experimental data. Predictability Q^2 was 0.88 and determined by internal cross-validation. Still, the quality of the model could be improved any further by decreasing the experimental error. Predicted values for two conditions deviated in model response compared to experimental data. Those were lysozyme at pH 3, 1.46 M NaCl and α-lactalbumin at pH 5, 0.1 M NaCl. This could be due to unstable conditions, caused by approximation to the solubility line or the pI.

Besides internal validation, within the training set, the final QSAR model was also applied and validated with an external test set including 10 observations (Figure 4). The results of this external validation indicate that the predicted values for the diffusion coefficient were in good agreement with the experimental data. The high value of $R^2 = 0.91$ for the test set suggests that the resulting QSAR model has a high predictive ability. This means that the QSAR model also allows for the accurate prediction of diffusion coefficients under new conditions excluded the training set.

For an additional assessment of the statistical significance of the predictive power, a response permutation (Y-scrambling) was performed (Figure 5). It can be seen clearly that all values for R^2 and Q^2 are significantly lower for the scrambled models. This reflects a clear statistical significance of the estimated predictive power of the QSAR model and its validity. A random correlation between the descriptors and the experimental data can therefor be excluded.

4.3 Influence of Protein Structure Properties on Protein-protein Interactions

The resulting QSAR model did not only allow the prediction of the diffusion coefficient, but also provided mechanistic insight into the properties influencing the diffusion coefficient. In this study the impact of the molecular size and shape as well as protein-protein interactions were captured by the 68 molecular descriptors of the QSAR model. The importance of each descriptor to the model can be evaluated by the VIP value. Descriptors with a VIP > 1.0 are considered to have a strong influence on the target figure. Descriptors with values below 1.0 have a minor impact [34]. Figure 6 shows the VIP value and the regression coefficient for each descriptor. The sign of the regression coefficient indicates the direction of the influence. Descriptors with a positive regression coefficient are proportional to the value of the diffusion coefficient, negative regression coefficients are inversely proportional [40].

Using this model for the diffusion coefficient, descriptors with information about electrostatic surface and molecular structure properties showed the highest VIP values. Five of seven descriptors with a VIP > 1.25 were connected to protein structure properties, the remaining ones to electrostatics. This strong influence of molecular structure properties was also obvious from the experimental data. This is in accordance with the Stokes-Einstein equation and contributes to D_0, the diffusion coefficient at infinite dilution included in the virial expansion of D. In the Stokes-Einstein equation the parameter for molecular structure properties is represented by the hydrodynamic radius r_h. Its inversely proportional impact was also captured by the model through negative regression

Table 2: Descriptors with a VIP value > 1.0 included in the final QSAR model and their descriptions.

No.	Descriptor	Definition
1	sumSurfA_ShellEsp	Sum of ESP of surface points projected on a shell around the molecule with a distance of 4.2 $\mathring{A}$
2	totalSurf	Surface area of the protein in $\mathring{A}^2$
3	totalSurf_PatchHydLow	Solvent-accessible surface area of the protein surface patch with the lowest hydrophobicity value in $\mathring{A}^2$
4	nAtom	Number of atoms of the protein
5	mass	Molecular weight of the molecule
6	nAAcid	Chain length of the protein
7	shapeMin	Value for the sphericity of the protein: (minimum distance between mass center and protein surface)/(mean distance between mass center and protein surface)
8	totalSurf_PatchEspLow	Solvent-accessible surface area of the protein surface patch with the lowest ESP value in $\mathring{A}^2$
9	totalSurf_PatchHydHigh	Solvent-accessible surface area of the protein surface patch with the highest hydrophobicity value in $\mathring{A}^2$
10	shapeMax	Value for the sphericity of the protein: (maximum distance between mass center and protein surface)/(mean distance between mass center and protein surface)
11	binAbs_SurfHyd_3	Number of points with low hydrophobicity on the protein surface
12	nPos_SurfHyd	Number of hydrophobic surface points on the protein surface
13	relPos_SurfEsp	Ratio of positively charged surface points on the protein surface
14	relPos_PatchEspHigh	Ratio of positively charged surface points on the protein patch with the highest ESP value
15	sumSurf_PatchEspLow	Sum of ESP on the protein patch with the lowest ESP value
16	sumNeg_PatchEspLow	Sum of negative charge on the surface patch with the lowest ESP value
17	nPos_ShellEsp	Number of positively charged surface points projected on a shell around the molecule with a distance of 4.2 $\mathring{A}$
18	relPos_PatchEspLow	Ratio of positively charged surface points on the protein surface patch with the lowest ESP value
19	mean_PatchHydHigh	Mean hydrophobicity on the protein surface patch with the highest hydrophobicity
20	sumPos_SurfHyd	Sum of points with positive hydropathy score on the protein surface

coefficients for the descriptors 2, 4, 5, and 6, as is displayed in Table 2. In contrast to this, descriptor 7 had a positive regression coefficient, although it belonged to the same set. The reason is the missing correlation to r_h. The descriptor describes the influence of the molecule's shape on the diffusion coefficient. The more spherical the molecule, the higher is the value for this descriptor, which results in a higher diffusion coefficient. This correlation is in agreement with theory, because the non-spherical shape of a molecule increases the friction coefficient and, thus, results in a decrease of the diffusion coefficient [47]. Although this study was conducted with globular proteins only, it is obvious that this model was sensitive to changes in molecular shape. By looking at the values of this descriptor for the proteins used in this study, it can be seen that BSA and glucose oxidase deviate most strongly from a spherical shape. Besides descriptors for molecular structure properties, descriptors representing protein-protein interactions accounted for VIP values > 1.25. These protein-protein interactions are captured by the diffusion interaction parameter k_D included in the virial expansion of D. Descriptor 1, which had the highest VIP value in this model, represented the strong influence of electrostatic surface potential and, thus, the important impact of electrostatic interactions. Under the screened conditions, these long-range interactions revealed a strong influence for many conditions investigated in this study. Descriptor 3 represents the surface area of the protein surface patch with the lowest hydrophobicity. This property further underlines the strong influence of electrostatics under the studied conditions, because a large area with low hydrophobicity results in a mainly electrostatic effect.

In contrast to the descriptors mentioned above, descriptors 8 to 20 with a VIP value between 1.0 and 1.25 captured electrostatic, but also short-range interactions, e.g. hydrophobic properties. For this group, no clear correlation with the experimental data could be made. Nevertheless, the main effects of the descriptors describing the same property could be pointed out. For the descriptors describing hydrophobic properties, a negative regression coefficient was determined. This is in accordance with theory, because hydrophobic interactions always have an attractive character and, hence, result in a decrease of the diffusion coefficient [5]. For the protein concentration used in this study, however, the VIP > 1 for these descriptors was remarkable. It showed that although electrostatic interactions dominate over short-range interactions under dilute conditions, the latter contribute to the value of the diffusion coefficient. According to theory, this most likely occurs at conditions close to the pI of the proteins or at high NaCl concentrations causing charge shielding effects and therefor promoting short-range interactions, such as hydrophobic interactions [44, 48]. Another more unlikely reason could be that for the studied proteins with a high molecular weight, namely, ovalbumin, BSA, and glucose oxidase, a protein concentration of 10 mg/mL exceeded the dilute state. According to the findings of Kumar et al., this would promote the impact of hydrophobic interactions [49]. By exemplarily taking a closer look at descriptor 9, this assumption was maintained: The impact of the descriptor was particularly important to proteins with high molecular weight (data not shown).

For the descriptors describing electrostatic properties (descriptors 13-18), slightly positive regression coefficients were found. In contrast to descriptors 1 and 8, they have an influence in opposite direction. Additionally, it is remarkable that four of these descriptors were related to positively charged surface points, although mainly proteins with negative net charge under the studied conditions were used in this work. These contrasts underline the complexity of the electrostatic impact on protein-protein interactions in solution. Electrostatic interactions can be influenced by a multitude of parameters [50]. In the presented model, these were the pH value, ionic strength through addition of NaCl, and surface charge of the protein. For the description of their synergetic effects on the impact of electrostatics, a variety of descriptors is necessary. This also includes oppositely directed descriptors, as can be seen for descriptors 1 and 8 with a negative regression coefficient, which probably capture the influence of negative charge, and descriptors 13 to 18 with positive values, which capture the influence of positive charge. These descriptors with positive regression coefficient values might be considered as a compensation of strong negative influence of the descriptors 1 and 8.

Among the descriptors with VIP values between 1.0 and 1.25, one descriptor capturing molecular structure properties could be found. This descriptor again underlines the strong impact of protein shape on the value of D, which was already observed for descriptor 7.

Taking all observations together, the diffusion coefficient is a result of various properties depending on the protein's structure. The size of the molecule and electrostatic interactions were found to be the properties with the main impact for this study. Further interactions that play a role for the overall potential could be identified. For other experimental setups, e.g. for concentrated protein solutions, differing results of QSAR modeling due to changes in underlying interactions could be expected. In this study, it was also shown that there is a complex relationship between the acting forces, which can also influence each other. Thus, QSAR modeling does not only enable the prediction of protein-protein interactions by determination of the diffusion coefficient, but also provides insights into the fundamental understanding of the properties influencing this parameter.

5 Conclusions and Outlook

Determination of the diffusion coefficient by QSAR modeling did not only reveal the predictive capacity of this method, but also its ability to improve mechanistic understanding on a molecular basis. The diffusion coefficients determined in this study showed clear correlations to the protein-protein interactions in solution. The QSAR model based on these results and the three-dimensional structure properties of the proteins was able to determine and predict these values with a coefficient of determination R^2 of 0.9 and a predictability Q^2 of 0.88. In accordance with the experimental data, it described the strong impact of the protein size. Regarding protein-protein interactions, which experimentally can only be captured by an overall potential, the VIP value for each descriptor of the final model agreed with theory and reflected the predominant impact of electrostatic interactions under the studied dilute conditions. It also provided deeper insight, as it accounted for the shape and additional short-range interactions of the molecules, such as hydrophobic forces. With this promising results, QSAR modeling cannot only be used to gain more information with less sample consumption and working effort, but also improves mechanistic understanding of various parameters in biotherapeutics associated with the protein's three-dimensional structure.

So far, QSAR has only been used to describe and predict the chromatographic behavior of large biomolecules during purification processes. This work is the first application of QSAR beyond chromatography and the results demonstrate the potential of this methodology for future applications in the field of protein phase behavior and understanding the underlying mechanisms and interactions. Future work in this field could focus on the generation of QSAR models for other parameters reflecting protein-protein interactions, such as the second virial coefficient B_{22}, the diffusion interaction parameter k_D, or the storage modulus G'. Additionally, the implementation of non-globular proteins and the generation of advanced models sensitive to protein concentration could be topic of further research. This option mentioned last might enable to overcome the drawback that QSAR models have only been valid for the respective protein concentration so far.

6 Acknowledgments

This research work was supported by Lonza Biologics PLC and is also part of the project 'Protein aggregation during production of modern biopharmaceuticals' (0315342B) funded by the German Federal Ministry of Education and Research (BMBF). We thank Maike Schröder for proofreading this manuscript.

References

[1] R. A. Lewus, P. A. Darcy, A. M. Lenhoff, S. I. Sandler, Interactions and phase behavior of a monoclonal antibody, Biotechnology Progress 27 (1) (2011) 280–289. doi:10.1002/btpr.536.

[2] S. J. Shire, Z. Shahrokh, J. Liu, Challenges in the development of high protein concentration formulations., Journal of pharmaceutical sciences 93 (6) (2004) 1390–402. doi:10.1002/jps.20079.

[3] Y.-C. Cheng, C. L. Bianco, S. I. Sandler, A. M. Lenhoff, Salting-Out of Lysozyme and Ovalbumin from Mixtures: Predicting Precipitation Performance from Protein-Protein Interactions, Industrial & Engineering Chemistry Research 47 (15) (2008) 5203–5213. doi:10.1021/ie071462p.

[4] C. J. van Oss, Long-range and short-range mechanisms of hydrophobic attraction and hydrophilic repulsion in specific and aspecific interactions, Journal of Molecular Recognition 16 (4) (2003) 177–190. doi:10.1002/jmr.618.

[5] Y. Liang, N. Hilal, P. Langston, V. Starov, Interaction forces between colloidal particles in liquid: Theory and experiment, Advances in Colloid and Interface Science 134-135 (2007) 151–166. doi:10.1016/j.cis.2007.04.003.

[6] D. J. A. Crommelin, R. D. Sindelar, B. Meibohm, Pharmaceutical Biotechnology: Fundamentals and Applications, SpringerLink : Bücher, Springer New York, 2013.

[7] A. Saluja, D. S. Kalonia, Nature and consequences of proteinprotein interactions in high protein concentration solutions, International Journal of Pharmaceutics 358 (1-2) (2008) 1–15. doi:10.1016/j.ijpharm.2008.03.041.

[8] A. George, W. W. Wilson, Predicting protein crystallization from a dilute solution property, Acta Crystallographica Section D Biological Crystallography 50 (4) (1994) 361–365. doi:10.1107/S0907444994001216.

[9] T. Ahamed, B. N. A. Esteban, M. Ottens, G. W. K. van Dedem, L. A. M. van der Wielen, M. A. T. Bisschops, A. Lee, C. Pham, J. Thömmes, Phase behavior of an intact monoclonal antibody., Biophysical journal 93 (2) (2007) 610–9. doi:10.1529/biophysj.106.098293.

[10] B. D. Connolly, C. Petry, S. Yadav, B. Demeule, N. Ciaccio, J. M. R. Moore, S. J. Shire, Y. R. Gokarn, Weak interactions govern the viscosity of concentrated antibody solutions: high-throughput analysis using the diffusion interaction parameter., Biophysical journal 103 (1) (2012) 69–78. doi:10.1016/j.bpj.2012.04.047.

[11] A. Saluja, A. V. Badkar, D. L. Zeng, S. Nema, D. S. Kalonia, Ultrasonic storage modulus as a novel parameter for analyzing protein-protein interactions in high protein concentration solutions: correlation with static and dynamic light scattering measurements., Biophysical journal 92 (1) (2007) 234–44. doi:10.1529/biophysj.106.095174.

[12] J. Zhang, X. Y. Liu, Effect of proteinprotein interactions on protein aggregation kinetics, The Journal of Chemical Physics 119 (20) (2003) 10972. doi:10.1063/1.1622380.

[13] M. Muschol, F. Rosenberger, Interactions in undersaturated and supersaturated lysozyme solutions: Static and dynamic light scattering results, The Journal of Chemical Physics 103 (24) (1995) 10424. doi:10.1063/1.469891.

[14] K. C. Bauer, M. Göbel, M.-L. Schwab, M.-T. Schermeyer, J. Hubbuch, Concentration-dependent changes in apparent diffusion coefficients as indicator for colloidal stability of protein solutions, Int. J. Pharm. (Amsterdam, Neth.). 511 (1) (2016) 276–287. doi:10.1016/j.ijpharm.2016.07.007.

[15] M. Dehmer, K. Varmuza, D. Bonchev, Statistical Modelling of Molecular Descriptors in QSAR/QSPR, Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, 2012. doi:10.1002/9783527645121.

[16] C. B. Mazza, N. Sukumar, C. M. Breneman, S. M. Cramer, Prediction of protein retention in ion-exchange systems using molecular descriptors obtained from crystal structure., Analytical chemistry 73 (22) (2001) 5457–61.

[17] C. B. Mazza, C. E. Whitehead, C. M. Breneman, S. M. Cramer, Predictive quantitative structure retention relationship models for ion-exchange chromatography, Chromatographia 56 (3-4) (2002) 147–152. doi:10.1007/BF02493203.

[18] T. Yang, C. M. Breneman, S. M. Cramer, Investigation of multi-modal high-salt binding ion-exchange chromatography using quantitative structure-property relationship modeling, Journal of Chromatography A 1175 (1) (2007) 96–105. doi:10.1016/j.chroma.2007.10.037.

[19] W. K. Chung, Y. Hou, M. Holstein, A. Freed, G. I. Makhatadze, S. M. Cramer, Investigation of protein binding affinity in multimodal chromatographic systems using a homologous protein library, Journal of Chromatography A 1217 (2) (2010) 191–198. doi:10.1016/j.chroma.2009.08.005.

[20] A. Ladiwala, F. Xia, Q. Luo, C. M. Breneman, S. M. Cramer, Investigation of protein retention and selectivity in HIC systems using quantitative structure retention relationship models., Biotechnology and bioengineering 93 (5) (2006) 836–50. doi:10.1002/bit.20771.

[21] J. Buyel, J. Woo, S. Cramer, R. Fischer, The use of quantitative structureactivity relationship models to develop optimized processes for the removal of tobacco host cell proteins during biopharmaceutical production, Journal of Chromatography A 1322 (2013) 18–28. doi:10.1016/j.chroma.2013.10.076.

[22] D. E. Kuehner, C. Heyer, C. Rämsch, U. M. Fornefeld, H. W. Blanch, J. M. Prausnitz, Interactions of lysozyme in concentrated electrolyte solutions from dynamic light-scattering measurements., Biophysical journal 73 (6) (1997) 3211–24. doi:10.1016/S0006-3495(97)78346-2.

[23] C. Lehermayr, H.-C. Mahler, K. Mäder, S. Fischer, Assessment of Net Charge and ProteinProtein Interactions of Different Monoclonal Antibodies, Journal of Pharmaceutical Sciences 100 (7) (2011) 2551–2562. doi:10.1002/jps.22506.

[24] M. E. Young, P. A. Carroad, R. L. Bell, Estimation of diffusion coefficients of proteins, Biotechnology and Bioengineering 22 (5) (1980) 947–955. doi:10.1002/bit.260220504.

[25] The Uniprot Consortium, UniProt: a hub for protein information, Nucleic Acids Research 43 (D1) (2015) D204–D212. doi:10.1093/nar/gku989.

[26] H. M. Berman, The Protein Data Bank, Nucleic Acids Research 28 (1) (2000) 235–242. doi:10.1093/nar/28.1.235.

[27] E. Krieger, G. Koraimann, G. Vriend, Increasing the precision of comparative models with YASARA NOVA-a self-parameterizing force field, Proteins: Structure, Function, and Bioinformatics 47 (3) (2002) 393–402. doi:10.1002/prot.10104.

[28] Y. Duan, C. Wu, S. S. Chowdhury, M. C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, J. Caldwell, J. Wang, P. Kollman, A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations., Journal of computational chemistry 24 (16) (2003) 1999–2012. doi:10.1002/jcc.10349.

[29] R. Anandakrishnan, B. Aguilar, A. V. Onufriev, H++ 3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations, Nucleic Acids Research 40 (W1) (2012) W537–W541. doi:10.1093/nar/gks375.

[30] M. Biasini, S. Bienert, A. Waterhouse, K. Arnold, G. Studer, T. Schmidt, F. Kiefer, T. G. Cassarino, M. Bertoni, L. Bordoli, T. Schwede, SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information, Nucleic Acids Research 42 (W1) (2014) W252–W258. doi:10.1093/nar/gku340.

[31] J. Kyte, R. F. Doolittle, A simple method for displaying the hydropathic character of a protein, Journal of Molecular Biology 157 (1) (1982) 105–132. doi:10.1016/0022-2836(82)90515-0.

[32] F. Dismer, J. Hubbuch, 3D structure-based protein retention prediction for ion-exchange chromatography, J. Chromatogr. A 1217 (8) (2010) 1343–1353. doi:10.1016/j.chroma.2009.12.061.

[33] K. M. H. Lang, J. Kittelmann, C. Dürr, A. Osberghaus, J. Hubbuch, A comprehensive molecular dynamics approach to protein retention modeling in ion exchange chromatography, J. Chromatogr. A 1381 (2015) 184–193. doi:10.1016/j.chroma.2015.01.018.

[34] L. Eriksson, T. Byrne, E. Johansson, J. Trygg, C. Wikström, Multi- and Megavariate Data Analysis: Part I: Basic Principles and Applications, MKS Umetrics AB, 2013.

[35] L. Eriksson, J. Jaworska, A. P. Worth, M. T. Cronin, R. M. McDowell, P. Gramatica, Methods for Reliability and Uncertainty Assessment and for Applicability Evaluations of Classification- and Regression-Based QSARs, Environmental Health Perspectives 111 (10) (2003) 1361–1375. doi:10.1289/ehp.5758.

[36] A. Tropsha, P. Gramatica, V. Gombar, The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models, QSAR & Combinatorial Science 22 (1) (2003) 69–77. doi:10.1002/qsar.200390007.

[37] E. A. Permyakov, L. J. Berliner, α-Lactalbumin: structure and function, FEBS Letters 473 (3) (2000) 269–274. doi:10.1016/S0014-5793(00)01546-5.

[38] E. Tatsumi, M. Hirose, Highly ordered molten globule-like state of ovalbumin at acidic pH: native-like fragmentation by protease and selective modification of Cys367 with dithiodipyridine., Journal of biochemistry 122 (1997) 300–308.

[39] K. Baumgartner, L. Galm, J. Nötzold, H. Sigloch, J. Morgenstern, K. Schleining, S. Suhm, S. A. Oelmeier, J. Hubbuch, Determination of protein phase diagrams by microbatch experiments: Exploring the influence of precipitants and pH, International Journal of Pharmaceutics 479 (1) (2015) 28–40. doi:10.1016/j.ijpharm.2014.12.027.

[40] W. Kessler, Multivariate Datenanalyse, Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, 2007. doi:10.1002/9783527610037.

[41] O. D. Velev, E. W. Kaler, A. M. Lenhoff, Protein interactions in solution characterized by light and neutron scattering: comparison of lysozyme and chymotrypsinogen., Biophys. J. 75 (6) (1998) 2682–97. doi:10.1016/S0006-3495(98)77713-6.

[42] B. U. Felderhof, Diffusion of interacting Brownian particles, Journal of Physics A: Mathematical and General 11 (5) (1978) 929–937. doi:10.1088/0305-4470/11/5/022.

[43] D. Guo, C. T. Mant, A. K. Taneja, J. Parker, R. S. Rodges, Prediction of peptide retention times in reversed-phase high-performance liquid chromatography I. Determination of retention coefficients of amino acid residues of model synthetic peptides, Journal of Chromatography A 359 (8) (1986) 499–518. arXiv:1011.1669, doi:10.1016/0021-9673(86)80102-9.

[44] R. A. Curtis, C. Steinbrecher, M. Heinemann, H. W. Blanch, J. M. Prausnitz, Hydrophobic forces between protein molecules in aqueous solutions of concentrated electrolyte., Biophysical chemistry 98 (3) (2002) 249–65.

[45] A. S. Parmar, M. Muschol, Hydration and Hydrodynamic Interactions of Lysozyme: Effects of Chaotropic versus Kosmotropic Ions, Biophysical Journal 97 (2) (2009) 590–598. doi:10.1016/j.bpj.2009.04.045.

[46] A. C. Dumetz, Protein Interactions and Phase Behavior in Aqueous Solutions: Effects of Salt, Polymer, and Organic Additives, University of Delaware, 2007.

[47] M. B. Jackson, Molecular and Cellular Biophysics, Cambridge University Press, 2006.

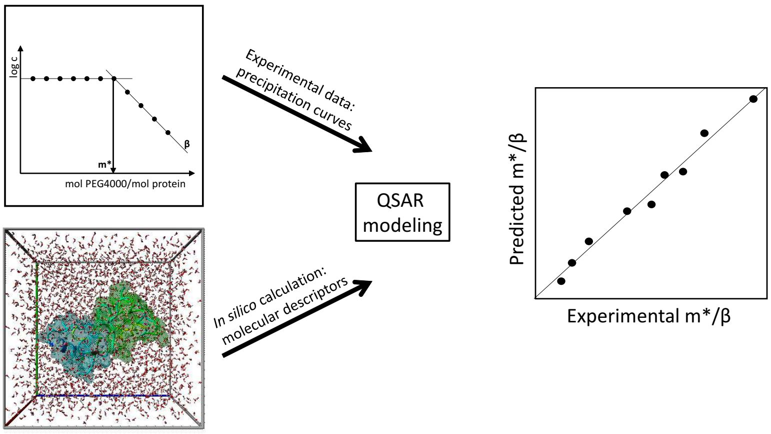
[48] E. Y. Chi, S. Krishnan, T. W. Randolph, J. F. Carpenter, Physical stability of proteins in aqueous solution: mechanism and driving forces in nonnative protein aggregation., Pharmaceutical research 20 (9) (2003) 1325–36.

[49] V. Kumar, N. Dixit, L. Zhou, W. Fraunhofer, 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 421 (1) (2011) 82–93. doi:10.1016/j.ijpharm.2011.09.017.

[50] M. S. Wisz, H. W. Hellinga, An empirical model for electrostatic interactions in proteins incorporating multiple geometry-dependent dielectric constants, Proteins: Structure, Function and Genetics 51 (3) (2003) 360–377. doi:10.1002/prot.10332.

Investigation and Prediction of Protein Precipitation by Polyethylene Glycol Using Quantitative Structure-Activity Relationship Models

Frank Hämmerling, Christopher Ladd Effio, Sebastian Andris, Jörg Kittelmann and Jürgen Hubbuch*


log c
m*
β
mol PEG4000/mol protein
Experimental data: precipitation curves
In silico calculation: molecular descriptors
QSAR modeling
Predicted m*/β
Experimental m*/β

Institute of Engineering in Life Sciences, Section IV: Biomolecular Separation Engineering, Karlsruhe Institute of Technology, Engler-Bunte-Ring 3, 76131 Karlsruhe, Germany

* *: Corresponding author; email address: juergen.hubbuch@kit.edu*

Journal of Biotechnology

Volume 241, (2017), 87-97

Abstract

Precipitation of proteins is considered to be an effective purification method for proteins and has proven its potential to replace costly chromatography processes. Besides salts and polyelectrolytes, polymers, such as polyethylene glycol (PEG), are commonly used for precipitation applications under mild conditions. Process development,however, for protein precipitation steps still is based mainly on heuristic approaches and high-throughput experimentation due to a lack of understanding of the underlying mechanisms. In this work we apply quantitative structure-activity relationships (QSARs) to model two parameters, the discontinuity point m^* and the β-value, that describe the complete precipitation curve of a protein under defined conditions. The generated QSAR models are sensitive to the protein type, pH, and ionic strength. It was found that the discontinuity point m^* is mainly dependent on protein molecular structure properties and electrostatic surface properties, whereas the β-value is influenced by the variance in electrostatics and hydrophobicity on the protein surface. The models for m^* and the β-value exhibit a good correlation between observed and predicted data with a coefficient of determination of $R^2 \geq 0.90$ and, hence, are able to accurately predict precipitation curves for proteins. The predictive capabilities were demonstrated for a set of combinations of protein type, pH, and ionic strength not included in the generation of the models and good agreement between predicted and experimental data was achieved.

Keywords: Polyethylene Glycol, Precipitation, Quantitative Structure-Activity Relationship (QSAR), Semi-mechanistic modeling, Monoclonal antibody

1 Introduction

Biopharmaceuticals, such as monoclonal antibodies (mAbs), have gained a leading role in the treatment of various diseases, e.g. cancer, multiple sclerosis or rheumatoid arthritis. Downstream processing of these products is crucial to receiving highly pure molecules for administration in patients. Roque et al. [1] postulated that 50 % - 80 % of manufacturing costs arise from downstream processing of a monoclonal antibody. Most of the presently used purification processes for biopharmaceutical products comprise chromatography steps, as chromatography still is the workhorse in downstream processing [2]. These unit operations are associated with high costs for chromatographic media and long cycle times. Chromatographic processes also suffer from limited scalability. For this reason, an increasing number of alternative bioseparation operations, such as aqueous two-phase extraction (ATPE), membrane filtration, crystallization, and precipitation are gaining increasing attention [3, 4, 5, 6, 7, 8]. Of these technologies, precipitation is considered very promising to overcome the challenges in scalability and cost reduction during downstream processing [2, 9].

To gain a deeper process understanding, computational methods and mechanistic modeling are increasingly moved into the spotlight of downstream process development in order to meet the demands of the quality by design approach stated by regulatory authorities [10, 11, 12]. One powerful tool among these *in silico* methods is quantitative structure-activity relationship (QSAR), where structural descriptors based on the protein 3D structures are used in combination with multi-variate data analysis tools to relate protein properties to experimental behavior. The purpose of QSAR is to gain an understanding of the underlying mechanisms and to build predictive models that can be applied to new compounds that were not included within the generation of the models. QSAR for proteins was applied successfully to describe and predict retention during several chromatography operations with different modes of interaction [13, 14, 15, 16, 17].

Precipitation of proteins with salts or polymers, such as polyethylene glycol (PEG), is already being applied as an alternative to traditional chromatography steps for the capturing or intermediate purification of biopharmaceuticals [18, 7]. Matheus et al. [19] demonstrated that the native secondary structure and activity of a mAb were preserved after precipitation by PEG4000 and subsequent re-dissolution of the precipitate. It was shown that precipitation of proteins as a purification step can be scaled up to the intermediate and pilot scales in a 100 L stirred tank reactor [20, 18]. Other advantageous attributes of polyethylene glycol as a precipitation agent are its inert nature, the relatively low costs for material and laboratory equipment, and its non-toxic, non-corrosive, non-flammable properties as well as its low vapor pressure [21].

Currently, two theories are applied to describe the mechanism of PEG-induced precipitation in a mechanistic way, namely, the theory of attractive depletion [22, 23] and the theory of excluded volume [24, 25]. The attractive depletion theory assumes that the PEG's center of mass is excluded from the vicinity of the protein surface due to its size and structure and, hence, creates a 'depletion zone'. When two neighboring protein molecules get sufficiently close to each other, the depletion zones overlap and an additional volume is recovered for the polymer. This results in an increasing entropy and a decreasing free energy which leads to a thermodynamically driven aggregation of protein molecules [26, 27]. The excluded volume theory, by contrast, is based on the assumption of protein molecules being sterically excluded from the volume of PEG molecules, which means that they get highly concentrated in the remaining volume of the solution. Aggregation of proteins occurs when the solubility limit is exceeded [28].

Besides the two theories mentioned above, the development of protein precipitation processes still is highly empirical and dependent on a vast number of parameters, such as precipitant type, temperature, pH, and ionic strength [29]. Juckles [30] and Ingham [31] demonstrated that precipitation of proteins by PEG does not only depend on ionic composition, temperature, and initial protein concentration, but also strongly dependent on the pH value. At pH values close to the isoelectric point (pI) of the protein, precipitation occurred at lower PEG concentrations than at pH values far from the nominal pI. Different studies revealed the hydrodynamic radius of PEG $r_{h,PEG}$ and protein $r_{h,prot}$ to be the main parameters influencing the precipitation efficiency [30, 28, 23, 32]. Polson et al. investigated the influence of PEG molecular mass on protein precipitation and found that the efficiency of protein precipitation increases with the molecular mass of the linear PEG polymer [33]. There are three characteristic parameters that describe the precipitation curve of proteins: The apparent intrinsic protein solubility in the absence of PEG (S_0), the slope of the precipitation curve in the region where precipitation occurs (β-value), and the PEG concentration at which protein solubility equals the protein concentration initially set (m^*). All parameters are derived from the Cohn equation that describes the salting-out effect of salts on proteins and can be applied analogously to precipitation curves with PEG [34, 35, 32]. A schematic precipitation curve is shown in Figure 1.

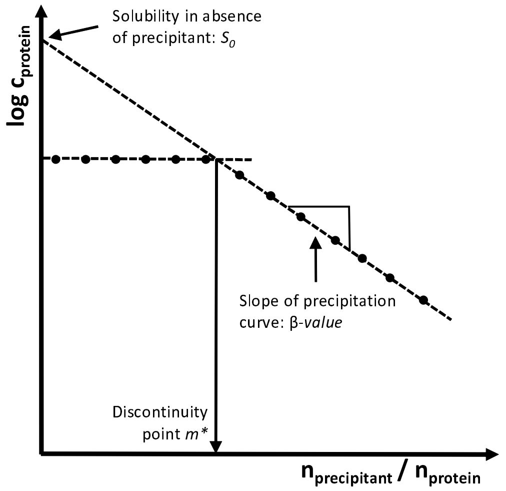


Figure 1: Schematic precipitation curve for proteins. The protein concentration in the supernatant is plotted over the molar ratio of precipitant and protein. S_0 depicts the maximum solubility in the absence of precipitant, m^* the maximum protein solubility at a distinct precipitant-to-protein ratio, and β the slope of the precipitation curve in the second segment.

To describe the complete precipitation curve of a protein, it is necessary to determine at least two of those three parameters. The effect of protein size on the slope of the precipitation curve (β-value) was revealed by Juckles et al. There is a linear correlation between the Stokes radius a of a protein and the β-value, meaning that smaller proteins show smaller values for β. β was found to be proportional to $a^{1.14}$. Changes in solution conditions were found to have a limited effect on the slope [30]. Sim et al. proposed a predictive model for calculating β, including the hydrodynamic radius of PEG $r_{h,PEG}$ and the protein $r_{h,prot}$ and two empirical coefficients γ and δ:

$$\beta = (\gamma \cdot r_{h,PEG}^{0.211} + \delta) r_{h,prot} \tag{1}$$

In this equation the first term $(\gamma \cdot r_{h,PEG}^{0.211} \cdot r_{h,prot})$ describes the depletion of protein by PEG, while the second term $(\delta \cdot r_{h,prot})$ accounts for the intrinsic excluded volume of protein and the depletion of PEG by the protein [32]. According to this equation, precipitation curves exhibit a steeper slope for proteins with increasing $r_{h,prot}$ while using the same PEG. Equation 1 was previously used to account for the impact of environmental conditions, such as temperature, pH, and ionic strength considering that $r_{h,PEG}$ and $r_{h,prot}$ are determined under the respective conditions [32]. Nevertheless, the predictive capabilities of this approach are limited. Only for larger proteins with a molecular mass >25 kDa can reasonable estimations be made [32]. To obtain the complete solubility curve when applying these models for the prediction of β, the apparent intrinsic protein solubility in the absence of PEG S_0 or m^* for a distinct precipitant concentration still have to be determined experimentally.

The presented approaches to describing protein precipitation with polyethylene glycol do not yet take into account protein surface properties, such as electrostatics and hydrophobicity, and are mostly based on the hydrodynamic radius of PEG and the protein. As described, there are first predictive approaches to estimating the β-value, but it is not yet possible to predict the complete precipitation curve of a protein.

In this work we apply the methodology of QSAR to generate models for both precipitation curve parameters m^* and β. These models are then applied for the prediction of complete protein precipitation curves for an external test set of proteins under different process conditions. The importance of each molecular descriptor on the models is evaluated to gain an enhanced understanding of the mechanisms influencing precipitation. As a first step, a set of multiple precipitation curves for a set of nine proteins was generated at different pH values and ionic strengths with PEG4000 as a precipitant. Based on the experimental data, QSAR models were generated for each of the two precipitation curve parameters, namely, the discontinuity point m^* and the slope of the precipitation curve β. These models were applied to gain mechanistic understanding of protein properties and surface characteristics influencing the two parameters and to predict precipitation curves. The predictive capabilities were evaluated by applying the model to an external test set of three combinations of protein type, pH, and ionic strength that were not included in the generation of the models, for which the precipitation curves were calculated *in silico.*

2 Materials and Methods

2.1 Disposables, Chemicals, and Buffers

2.1.1 Disposables

All precipitation experiments were carried out in 350 μL polypropylene flat bottom microplates (Greiner Bio-One, Kremsmünster, Austria). For spectrophotometric measurements, Greiner UV-STAR® microplates (Greiner Bio-One) were used.

2.1.2 Chemicals

The buffer substances used were formic acid (pH 4.0), acetic acid (pH 5.0) (both Merck KGaA, Darmstadt, Germany), MES (pH 6.0), MOPSO (pH 7.0), TAPS (pH 8.0 + 9.0), and CAPS (pH 10.0) (all AppliChem GmbH, Darmstadt, Germany). Polyethylene glycol with an average molecular mass of 4000 g/mol was purchased from Merck Millipore (Darmstadt, Germany). The proteins α-chymotrypsinogen A, α-lactalbumin, avidin, BSA, concanavalin A, glucose oxidase, hemoglobin, HSA, and ovalbumin were obtained from Sigma-Aldrich (St. Louis, MO, USA), the monoclonal antibody mAb1 was generously supplied by Synthon Biopharmaceuticals BV (Nijmegen, The Netherlands).

2.1.3 Buffers

All buffers were prepared with a concentration of 50 mM using ultrapure water. pH was controlled using the five-point calibrated pH meter HI-3220 (Hanna® Instruments, Woonsocket, RI, USA) equipped with a SenTix® 62 pH electrode (Xylem Inc., White Plains, NY, USA) and corrected by titration with hydrochloric acid or sodium hydroxide (both Merck KGaA, Darmstadt, Germany) with an accuracy of ± 0.05 pH units. The buffers were filtered with a 0.22 μm Supor® PES membrane (Pall GmbH, Dreieich, Germany), the 40 %(m/m) PEG4000 stock solution was filtered with a 1.2 μm cellulose acetate membrane (Sartorius AG, Göttingen, Germany). The density of the 40 %(m/m) PEG4000 solution $\rho_{40\%(m/m)PEG4000}$ was 1.067 g/mL, as was determined with a pycnometer (Brand GmbH & Co. KG, Wertheim, Germany). The relative standard deviation of density determination was 0.4 %.

2.1.4 Preparation of Protein Stock Solutions

A 5 mg/mL stock solution of each protein listed in Section 2.1.2 was prepared in the respective buffer. Table 1 displays all proteins used in this study, the corresponding Protein Data Bank (PDB) IDs, and properties.

Table 1: PDB ID, isoelectric point (pI), molecular mass (MM), and extinction coefficient of the proteins used in this study.

Protein	PDB ID	pI	MM	E1%$_{280nm}$
			[kDa]	[L g^{-1} cm^{-1}]
α-chymotrypsinogen A	2CGA	9.0	25.7	19.8
α-lactalbumin	1F6S	4.5	14.2	19.7
Avidin	1VYO	10.0	66	16.4
BSA	3V03	4.9	66	6.1
Concanavalin A	1CVN	5.5	104	12.7
Glucose oxidase	1CF3	4.2	160	15.1
Hemoglobin	1G09	6.8	64.5	7.7
HSA	1H9Z	4.7	66.5	4.9
Ovalbumin	1OVA	4.5	44.3	7.4
mAb1	n/a	8.5	148.7	14.7

Protein concentration was determined spectrophotometrically at 280 nm with a Nanodrop$^{\text{TM}}$ 2000c UV-Vis spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA). The theoretical extinction coefficient of each protein was calculated based on the amino acid sequence obtained from the UniProt database [36] with the ExPASy ProtParam tool from the Swiss Institute of Bioinformatics [37]. The molecular mass was adopted from the manufacturer's specifications or, if not stated, also calculated with the ExPASY ProtParam tool based on the primary protein structure. All protein solutions were filtered using 0.2 μm syringe filters with PES membranes (VWR, Radnor, PR, USA).

2.2 Automated Generation of Protein Precipitation Curves

A Tecan Freedom Evo 200 (Tecan GmbH, Crailsheim, Germany) was used for automated liquid handling. The platform was equipped with an 8-tips liquid handling arm, a TeShake orbital shaker, an Infinite® 200 UV-Vis spectrophotometer for absorption measurements (all Tecan GmbH, Crailsheim, Germany), and a Rotanta 46RSC centrifuge (Hettich GmbH & Co.KG, Tuttlingen, Germany). The precipitation procedure on the automated liquid handling station was described earlier by Oelmeier et al. [38]. All experiments were carried out at 25 °C, which was controlled by air conditioning. Systems with a total volume of 300 μL, including a protein concentration of 1 mg/mL and varying PEG4000 concentrations between 0 %(m/m) and 33 %(m/m), were prepared on the liquid handling station. After addition of the protein, the systems were first incubated at room temperature on the orbital shaker at 1000 rpm for 30 min and subsequently incubated for another 30 min without shaking. To analyze the remaining protein in the supernatant, the microplate was centrifuged for 30 min at 3470$\times g$ (4000 rpm) for separation of the protein precipitate. Then, 100 μL of supernatant were transferred into a UV-STAR® microplate and diluted with 100 μL of buffer using the liquid-level detection function of the liquid handling arm. Subsequently, the absorption of diluted supernatant was measured spectrophotometrically at wavelengths of 280 nm for determination of protein concentration, 410 nm for detection of precipitate carry-over, and 900 nm and 990 nm for the evaluation of the filling level. All experiments were performed in triplicates.

2.3 Evaluation of Precipitation Curves

The precipitation curves obtained show the logarithmized protein concentration in the supernatant as a function of the molar ratio between employed PEG4000 and the protein. The curves can be divided into two segments: In the first segment the protein concentration in the supernatant is at a constant level, it starts to decrease in the second segment. Due to logarithmization of the ordinate, this decrease is linear. Both parts of the curve were fitted linearly and the intersection point of both lines was calculated as the discontinuity point m^*. The slope of the linear regression in the second segment corresponds to the β-value (Figure 1). Automated evaluation of experimental data was performed with MATLAB R2015a (The MathWorks, Inc., Natick, MA, USA). The discontinuity point m^* was determined according to the method published by Hachem et al. [39], i.e. by calculation of the slopes between a subsequent triplet of data points. If both calculated slopes fell below a threshold value, the second data point of this triplet was determined to be m^* and remaining data points on both sides of m^* were assigned to the two segments and fitted linearly. For better comparability of all experimentally determined precipitation curves, the molar ratio of PEG4000 and protein was plotted logarithmically.

The mechanisms of protein precipitation are based on thermodynamics, where molar parameters are used for the description of reactions and kinetics. For this, a molar perspective is required for gaining a mechanistic understanding based on thermodynamics. We used a novel approach to describing the concentration of precipitant. Whereas previous publications used mass percent [%(m/m)] to represent PEG concentration, we calculated the molar ratio between PEG and protein [mol PEG/mol protein]. Hence, in this novel molar approach the m^* value represents the ratio of PEG4000 molecules and protein that is necessary to trigger precipitation of protein. Due to differences in molecular mass, the number of protein molecules in the systems varies within the set of proteins used in this study by a factor of 5.8 for a constant protein concentration given in [mg/mL]. The 40 %(m/m) PEG4000 stock solution equals a molar concentration of 0.107 mol/L PEG4000. To increase the PEG4000 concentration by 1 %(m/m) in the systems with a volume of 300 μL, 7.1 μL of 40 %(m/m) PEG4000 were added. The m^* values of the novel approach using the molar ratio can be converted easily into [%(m/m) PEG4000] using the protein's molecular mass.

2.4 QSAR Modeling

2.4.1 Preparation of Protein 3D Structures and Calculation of Molecular Descriptors

The UniProtID for all proteins was obtained from UniProt [36]. All PDB files were downloaded from the RCSB Protein Data Bank [40]. The specific IDs are illustrated in Table 1. In YASARA [41], a software for visualization, modeling of molecules, and molecular dynamics simulations, a protein structure reflecting the conditions in solution was generated. The structures were checked for completeness and, if necessary, missing residues or intramolecular disulfide bonds were added manually. The hydrogen bonding network was optimized and an energy minimization experiment was conducted with settings adapted from Lang et al. [42] and using the Amber03 force field [43]. Heteroatoms were separated from the protein structure and the protonation of amino acids was executed in H++ [44] according to the respective pH value and ionic strength of the buffer. After protonation of amino acid residues, the heteroatoms were reinserted into the protein 3D structure. Using the Amber03 force field, another energy minimization and

molecular dynamics (MD) simulation experiment was performed. The 10 ps MD simulation experiment was carried out at 298 K. The size of the simulation box was extended by 10 $\AA$ on every side of the protein, periodic boundaries were set, and snapshots were taken every 1 ps and averaged afterwards. This averaged structure was then used for the calculation of molecular descriptors. Glucose oxidase, which exists as a dimer under the studied conditions, was assembled by two monomers with the help of SWISS-MODEL [45]. Molecular descriptors were calculated with an in-house developed software, which allows the calculation of molecular descriptors accounting for protein molecular structure properties, electrostatics, and hydrophobicity. Descriptors accounting for hydrophobicity are based on the hydropathy scale published by Kyte and Doolittle [46]. Four different types of projections can be chosen:
Whole molecule descriptors: These descriptors take the complete protein molecule into account.
Plane descriptors: The calculated values for the protein properties are projected onto a plane that is tangentially approached towards the molecule's surface with a set distance of 5 $\AA$ between the protein and the plane. This distance is adapted from previous work published by Dismer et al. [47] and Lang et al. [42]. For this study, a set of 120 plane orientations, randomly distributed along the protein surface, was chosen and the respective descriptors were calculated for each orientation.
Patch descriptors: These descriptors describe a part of the protein molecule. The size of the protein surface patch considered for calculation of the patch descriptors was derived from the calculated planes: Based on the orientation of the planes, a solvent-accessible protein surface area within a distance below 20 $\AA$ was taken into account for calculation of molecular descriptors. Hence, only parts of the molecule are represented.
Shell descriptors: The calculated descriptor values obtained from all 120 plane orientations are summed up to gain a 'shell projection' representing the properties at a distance of 5 $\AA$ around the molecule.

2.4.2 Multi-variate Data Analysis

Table 2 displays all 17 combinations of protein, pH, and ionic strength that were included in the training set and the three conditions that were excluded from the generation of the models and used as external test set.

Table 2: Overview of proteins and experimental conditions included in the training and test set for QSAR modeling.

Protein	pH [-]	ionic strength [mM]
Training set		
α-chymotrypsinogen A	8.0	12
	9.0	39
α-lactalbumin	4.0	12
	6.0	20
BSA	4.0	33
	5.0	33
	6.0	20
Concanavalin A	6.0	20
Glucose oxidase	4.0	33
Hemoglobin	7.0	27
	8.0	12
HSA	4.0	33
	5.0	33
Ovalbumin	4.0	33
	5.0	33
mAb1	8.0	12
	9.0	39
Test set		
α-lactalbumin	5.0	33
Avidin	10.0	12
mAb1	10.0	12

A set of 132 molecular descriptors in total was calculated with the in-house software for each condition. Unit variance scaling was selected to scale the descriptor values appropriately. A Partial Least-Squares Regression (PLSR) was performed with SIMCA 13.0.3 (MKS Instruments AB, Umeå, Sweden) to calculate a first model to describe the target variables m^* and β, respectively. Descriptors with a variable influence on projection (VIP) value of $\geq$ 1.0 were selected from the first model for building a second final model including descriptors of significant influence only. To exclude a random correlation of the X-dataset consisting of the molecular descriptors with the experimental data, a Y-randomization with 100 permutations was performed. The X-dataset consisting of the descriptors was left intact, while the Y-dataset consisting of the observations was randomly re-ordered and the data were PSLR-modeled subsequently. The Y-randomization plot displays the correlation coefficient of the original Y-variable and the permuted Y-variable versus the R^2 and Q^2 of the Y-randomized models [48]. For assessing the statistical significance of the parent QSAR model, the Y-randomization plots were evaluated with the method presented by Eriksson et al. [49] and Kiralj [50]. Regression lines were fitted among the 'scrambled' R^2 and Q^2 values. The intercepts of these regression lines can be used as a measure of statistical significance. It is recommended that maximum values for the intercepts should not exceed 0.3 for R^2 and 0.05 for Q^2. The complete workflow of the presented study is illustrated in Figure 2.

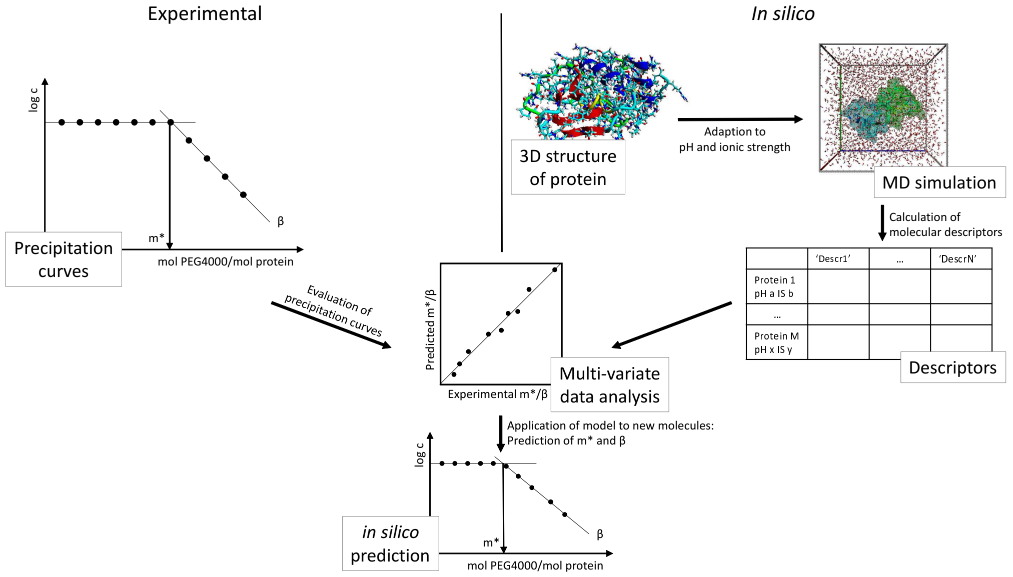


Figure 2: Schematic overview of work flow for QSAR modeling of protein precipitation by polyethylene glycol. Experimental data are related with molecular descriptors derived from the protein 3D structure by multi-variate data analysis. The created models are then used for *in silico* prediction of precipitation curves.

3 Results and Discussion

3.1 Evaluation of Precipitation Curves

The experimental data and the fitted precipitation curves of all experiments are displayed in Figure 3.

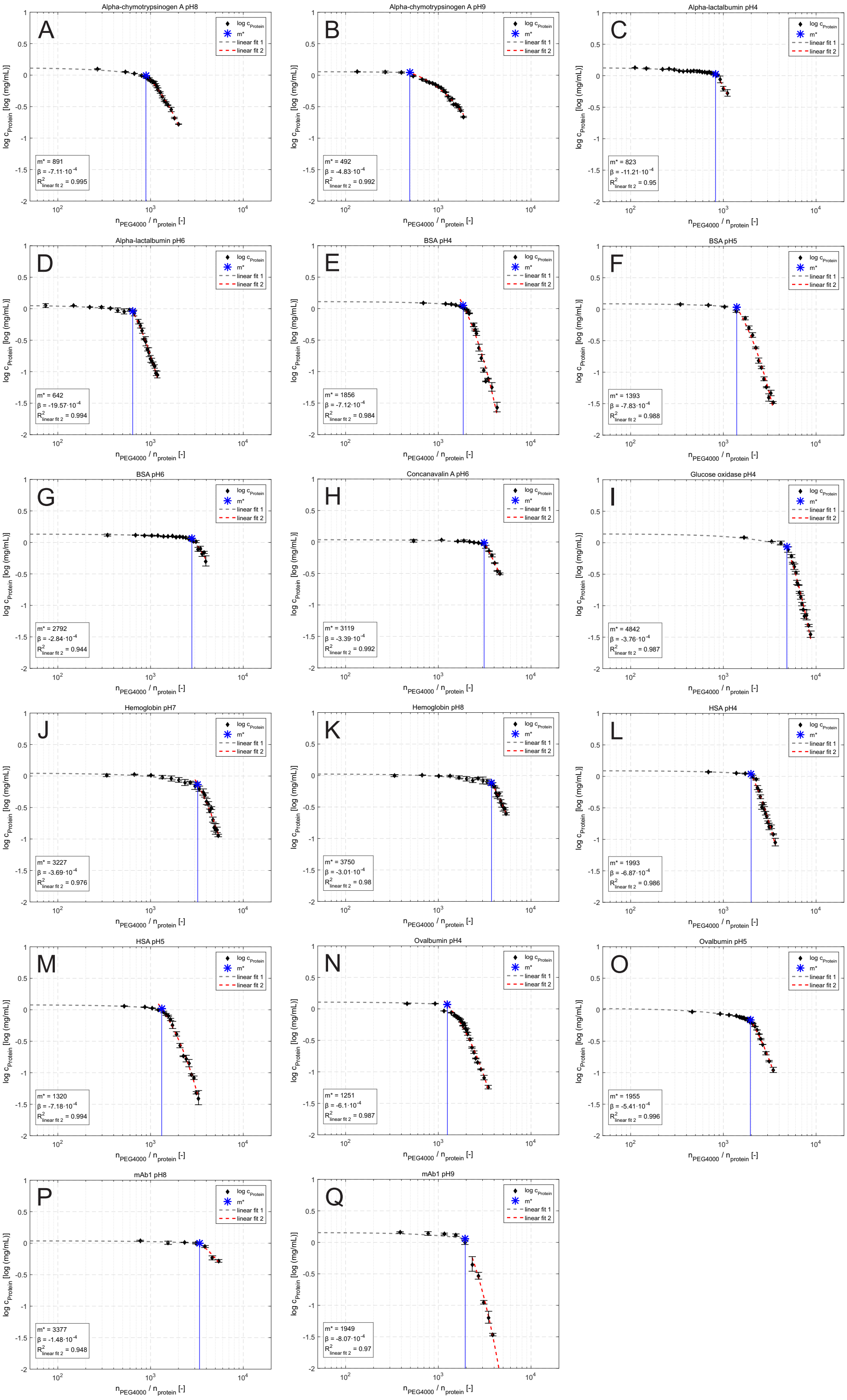


Figure 3: Experimentally determined solubility curves as a basis of QSAR model generation: α-chymotrypsinogen A (A-B), α-lactalbumin (C-D), BSA (E-G), concanavalin A (H), glucose oxidase (I), hemoglobin (J-K), HSA (L-M), ovalbumin (N-O), and mAb1 (P-Q).

m^* values for the investigated proteins were in the molar ratio range from 492 to 4842 mol PEG4000/mol protein. All β-values were linearly fitted with $R^2 \geq 0.94$ and determined to be in the range from -19.57 to -1.48 $\cdot 10^{-4}$ log (mg/mL). Proteins with a comparatively low molecular mass, e.g. α-lactalbumin, showed smaller values for m^* (823 mol PEG4000/mol protein at pH 4.0) compared to those with a high molecular mass, as for example glucose oxidase (4842 mol PEG4000/mol protein at pH 4.0).

Evaluating the precipitation curves for BSA at pH 4.0, pH 5.0, and pH 6.0 yielded values for m^* of 1856, 1393, and 2792 mol PEG4000/mol protein were observed (Fig. 3E-G). With a pI of 4.9, the propensity for precipitation was higher at a pH value close to the pI, while at pH 4.0 and pH 6.0, a higher molar PEG4000 to protein ratio was necessary to induce precipitation. These differences in m^*-values for the same protein under different solution conditions demonstrate that the sensitivity of proteins to polyethylene glycol is not only determined by the size of the molecule, but also by further inter molecular interactions, such as electrostatic and hydrophobic forces. This observation can also be made for the β-value, where the slopes of the precipitation curves for the same protein under different pH values showed fluctuating values. For BSA, β-values of -7.12, -7.83, and -2.84 $\cdot 10^{-4}$ log (mg/mL) were determined for pH 4.0, 5.0, and 6.0, respectively. The influence of pH on the value of β is obvious from the fact that the precipitation curves at pH values close to the isoelectric point of BSA show steeper slopes in the second segment. This supports the assumption that the β-value is not only affected by the molecule's size, but also by various protein properties and that the value of β increases with increasing protein net charge.

3.2 Evaluation of QSAR Modeling for Precipitation Curve Parameters

QSAR modeling was performed as described in Section 2.4.2 for both precipitation curve parameters, namely, the discontinuity point m^* and the β-value, i.e. the slope of the linear fit in the second segment. The models were interpreted for gaining information about the parameters and the interactions that influence protein precipitation by polyethylene glycol. An additional approach to generating a combined model for both parameters m^* and β resulted in a much lower model quality compared to the two separate models.

3.2.1 QSAR Model for Discontinuity Point m^*

The resulting QSAR model for m^* consisted of two latent variables and 33 molecular descriptors. Figure 4A displays the experimentally determined values for m^* compared to the values predicted by the model.

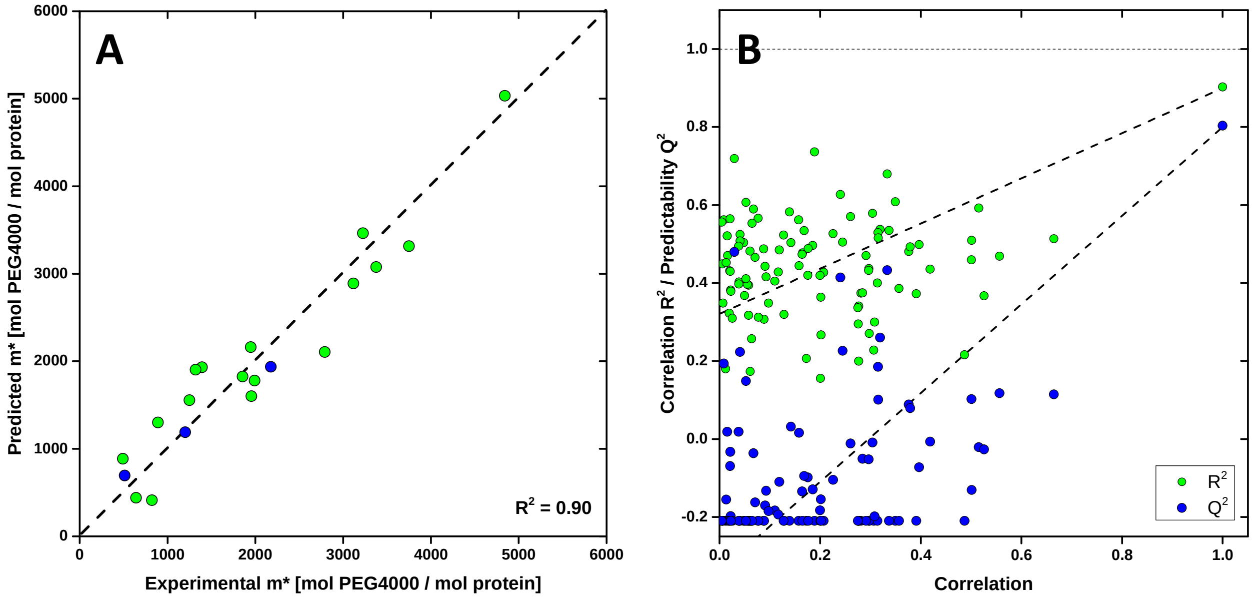


Figure 4: (A) QSAR model for the discontinuity point m^*: Experimental vs. predicted values of the training set (green) and test set (blue); (B) Permutation plot for the randomized Y-vector displaying the respective correlation R^2 and predictability Q^2.

The coefficient of determination between observed and predicted data (R^2) was 0.90 and the predictability (Q^2) was calculated as described by Tropsha [48] and Kiralj [50] and was found to be 0.80. The root mean square error of cross-validation (RMSECV) was 551.3 mol PEG4000/mol protein. The RMSECV is in the range between 11% - 112% of the observed values and especially pronounced for the proteins with a low molecular mass. This model was applied to an external test set of three conditions and the respective values of m^* were predicted. To exclude the possibility of a random correlation between determined m^*-values and molecular descriptors, a Y-randomization with 100 permutations was performed. The R^2 and Q^2 values for each of the Y-randomized models are shown in Figure 4B.

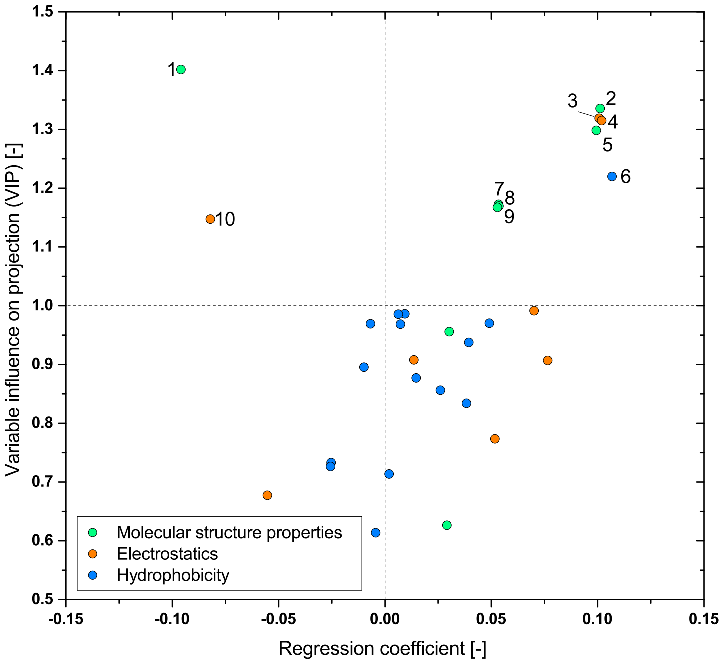


Figure 5: VIP values and regression coefficients for all 32 descriptors of the final QSAR model for the discontinuity point m^*. The 10 descriptors with a VIP value > 1.0 are numbered and described in Table 3.

Compared to the real model, all values are lower for the scrambled models. The regression line of the 'scrambled' R^2 values depicts an intercept of 0.32, while the value for Q^2 is -0.34. This indicates a meaningful selection of molecular descriptors and statistical significance of the parent QSAR model [49].

In order to obtain mechanistic understanding of the protein properties that mainly account for the value of m^*, the variable influence on the projection (VIP) was plotted over the regression coefficient for the 33 molecular descriptors (Figure 5). The VIP summarizes the importance of each molecular descriptor to the X- and Y-models. Descriptors with a VIP > 1.0 make a major contribution to the resulting PLSR model [51]. The algebraic sign of the regression coefficient indicates the direction of the influence, descriptors with a positive algebraic sign are proportional to the value of m^* and vice versa [51]. Table 3 lists all descriptors of the final QSAR model for m^* with a VIP > 1.0.

Table 3: Descriptors with a VIP value > 1.0 included in the final QSAR model for discontinuity point m^* and their description.

No.	Descriptor	Definition
1	shapeMin	Value for the sphericity of the protein: (minimum distance between mass center and protein surface)/(mean distance between mass center and protein surface)
2	dens	Density of the protein
3	sumSurfA_ShellEsp	Sum of ESP of surface points projected on a shell around the molecule with a distance of 5 $\AA$
4	totalSurf_PatchEsp	Solvent-accessible surface area of protein in $\AA^2$ on the patch with the highest ESP value
5	totalSurfA_Shell	Solvent-accessible surface area of a shell around the molecule with a distance of 5 $\AA$
6	totalSurf_PatchHyd	Solvent-accessible surface area of the protein surface patch with the highest hydrophobicity value in $\AA^2$
7	nAAcid	Chain length of the protein
8	nAtom	Number of atoms of the protein
9	mass	Molecular mass of the molecule
10	devA_PlaneEsp	(maximum ESP value - minimum ESP value)/mean value of ESP on the plane with the highest ESP value

According to the model, the shape of the protein has the strongest influence on m^* with a VIP of 1.4. This 'shapeMin' descriptor is calculated by dividing the surface point closest to the molecule center by the average surface point distance to molecule center. This means that for globular proteins, the value of this descriptor is close to 1, while it assumes smaller values for longitudinally shaped biomolecules. The negative regression coefficient indicates that the more the protein shape is spherical, the less PEG per protein is needed to initiate protein precipitation. This is in agreement with the regression coefficient of the 'dens' descriptor, which reflects the density of a protein. The higher the density of the protein is, the smaller is the surface area to volume ratio. A comparably strong impact on the value of m^* is exerted by the overall electrostatic potential projected on a shell around the protein molecule. This can be attributed to higher repulsive electrostatic interactions that prevent attractive protein interactions. Descriptor 6 represents the surface area of the protein patch with the highest hydrophobicity value and also shows a positive regression coefficient. This seems to be inconsistent with theory, as more hydrophobic molecules are said to encounter higher attractive protein interactions and, hence, lower m^*-values. This discrepancy can be explained by the assumption that these descriptors represent a counterbalance for other descriptors that were overestimated by the QSAR model or that these descriptors are embedded in a more complex network and cannot only be regarded alone. The cluster of descriptors 7 - 9 with similar VIP values and positive regression coefficients all represent the size of the protein. They reveal that a higher PEG4000 to protein-ratio is necessary to precipitate large molecules compared to molecules with smaller dimensions.

3.2.2 QSAR Model for the β-value

The final QSAR model for β was built with two latent variables and included 38 molecular descriptors. Figure 6A displays the experimentally determined values for β compared to the values predicted by the model.

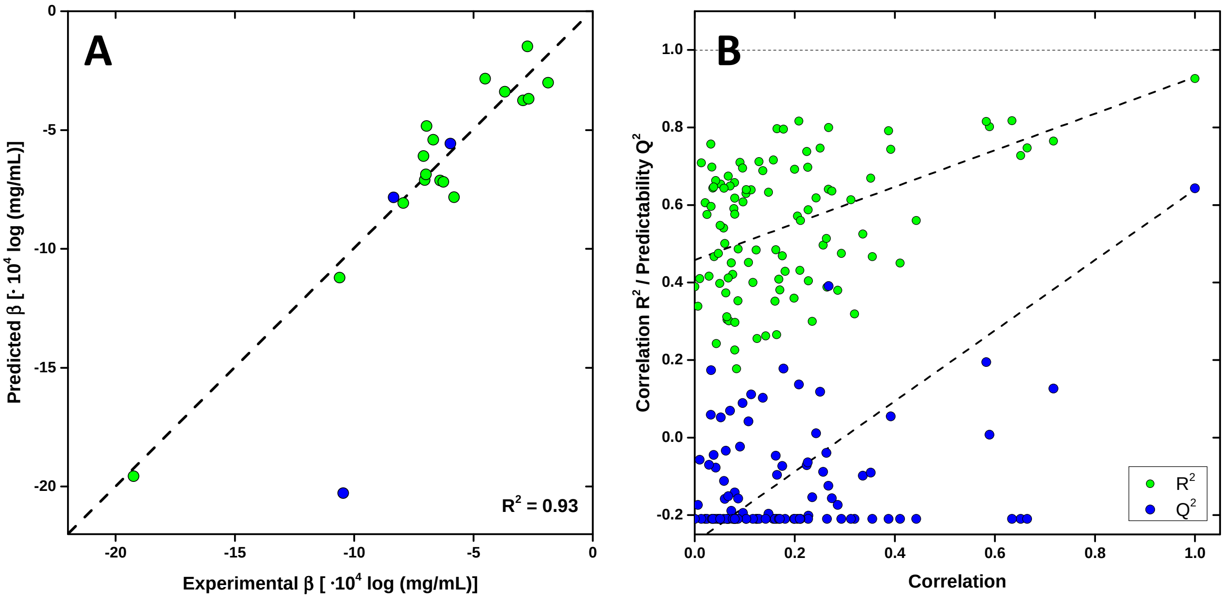


Figure 6: (A) QSAR model for the β-value: Experimental vs. predicted values of the training set (green) and test set (blue); (B) Permutation plot for the randomized Y-vector displaying the respective correlation R^2 and predictability Q^2.

With a coefficient of determination R^2 of 0.93, the correlation between experimental and predicted data is quite promising, while the predictability Q^2 has a value of 0.64. This indicates that there is a degree of perturbation in the data. We assume that this discrepancy between R^2 and Q^2 for β is due to the complexity of this parameter, that is influenced by the formation of clusters and therefor more complicated to model. An RMSECV of 2.61 $\cdot 10^{-4}$ $log\ (mg/mL)$ supports this assumption. The β-values that show the highest deviation between experimental and predicted data in the QSAR model are those for small proteins (here, α-lactalbumin and α-chymotrypsinogen). The highly negative β-value for α-lactalbumin of -19.57 $\cdot 10^{-4}$ $log\ (mg/mL)$ at pH 6.0 is subject to the highest prediction error of -8.6 $\cdot 10^{-4}$ $log\ (mg/mL)$. This observed deviation for small proteins was also reported in earlier publications and can be explained either by the repulsive Coulomb potential that becomes more pronounced for small proteins or by an interpenetration of PEG and small proteins, resulting in a change of osmotic pressure [32, 52]. The Y-permutation performed to exclude a random correlation between experimentally determined β-values and the molecular descriptors was performed in analogy with that for the m^* model in the previous section. The results are depicted in Figure 6B.

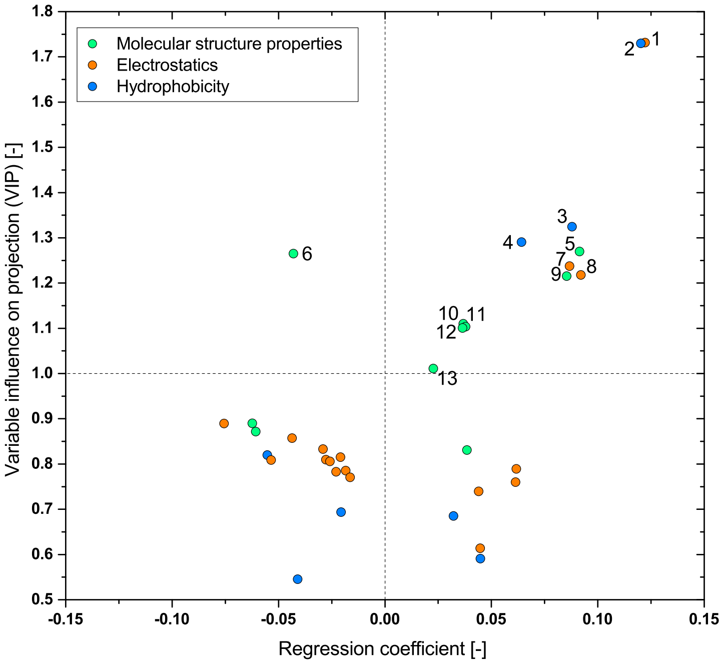


Figure 7: VIP values and regression coefficients for all 38 descriptors of the final QSAR model for the β-value. The 13 descriptors with a VIP value $>$ 1.0 are numbered and described in Table 4.

Again, the values for R^2 and Q^2 decrease with decreasing correlation to the 'unscrambled' Y-vector, but to a lower degree compared to the model for m^*. The plotted regression lines depict an intercept of 0.45 for R^2 and -0.27 for Q^2. While the intercept for Q^2 meets the recommended maximum value of 0.05, the latter is exceeded for R^2. Nevertheless, we still expect the parent QSAR model to be statistically significant due to the high number of performed permutations, which is known to moderately increase the number of chance correlations that result in a high R^2 [53].

QSAR modeling provides insight into the molecular properties that influence the value of β, although the higher perturbation of the data only allows for a more general reflection. As described in the Materials and Methods section, a novel approach based on the molar ratio of PEG and protein was used to describe the precipitation curves in this work. Consequently, the impact of the molecular mass of the proteins in the QSAR model for β is supposed to be considerably lower compared to the traditional approach that uses mass concentrations of the precipitant. Figure 7 shows the VIP values and the regression coefficients for all descriptors included in the final QSAR model for β. Table 4 lists the descriptors with a VIP $>$ 1.0 and their descriptions.

Table 4: Descriptors with a VIP value > 1.0 included in the final QSAR model for β-value and their description.

No.	Descriptor	Definition
1	devA_SurfEsp	(maximum ESP value - minimum ESP value)/mean value of ESP of the entire molecule
2	devA_PatchHyd	(maximum hydrophobicity value - minimum hydrophobicity value)/mean value of hydrophobicity on the patch with the highest hydrophobicity value
3	totalSurf_PatchHyd	Solvent-accessible surface area of the protein surface patch with the highest hydrophobicity value in $\mathring{A}^2$
4	devB_SurfHyd	(maximum hydrophobicity value - minimum hydrophobicity value)/median value of hydrophobicity of the entire molecule
5	dens	Density of the protein
6	shapeMin	Value for the sphericity of the protein: (minimum distance between mass center and protein surface)/(mean distance between mass center and protein surface)
7	sumSurfA_ShellEsp	Sum of ESP of surface points projected on a shell around the molecule with a distance of 5 $\mathring{A}$
8	devB_SurfEsp	(maximum ESP value - minimum ESP value)/median value of ESP on the protein surface
9	totalSurfA_Shell	Solvent-accessible surface area of a shell around the molecule with a distance of 5 $\mathring{A}$
10	nAtom	Number of atoms of the protein
11	nAAcid	Chain length of the protein
12	mass	Molecular mass of the molecule
13	totalSurf	Surface area of the protein $\mathring{A}^2$

A positive regression coefficient of a molecular descriptor means that an increasing value of the descriptor is accompanied by an increasing absolute value of β (the slope of the precipitation curve flattens). In contrast to m^*, where descriptors for protein molecular structure properties were the parameters with the highest VIP values, they play a minor role in the case of β. Here, the variance of electrostatic surface potential on the overall protein surface exhibits the highest VIP value and, hence, has the strongest influence on β. A comparable VIP value is observed for the variance of hydrophobicity on the protein patch with the highest hydrophobicity value of the molecule. Both descriptors have a positive regression coefficient, which indicates a shallower slope of the precipitation curve, if the variance of electrostatic surface potential (ESP) of the entire molecule or hydrophobicity on the described patch increases. Descriptors 5 and 6 relate the protein's molecular structure properties to the value of β. The higher the density of the protein (descriptor 5) is, the shallower is the slope, and the higher the sphericity of the molecule (descriptor 6), the steeper is the slope of the curve. Descriptors 9 - 13 are directly related to the molecular mass or describe the surface area of the protein and show positive regression coefficients. Thus, the slope of the precipitation curve flattens when protein size increases.

Previous publications based on the mass concentration of precipitant only revealed the hydrodynamic radius of the protein as the main parameter to model β [30, 32], and steeper slopes of the precipitation curves were observed with increasing protein size [32, 23, 28]. Due to the molar ratio approach in this study, the molecular mass is supposed to have a minor impact in this QSAR model, as the protein concentrations determined in the supernatant are related to the number of PEG and protein molecules in solution rather than to the mass of both substances in solution. When generating the QSAR model for β with the traditional approach using [$\%(m/m)$ PEG4000] (results are shown in the Supplementary Material), the strong impact of the molecular mass (and, hence, of the hydrodynamic radius) of the protein can be seen clearly among the molecular descriptors with the highest VIP values and negative regression coefficients. For comparison, the β-values calculated according to Equation 1 reported by Sim et al. [32] are illustrated together with the values predicted by the QSAR model (see Supplementary Figure 1).

3.2.3 Application of QSAR Models for *in silico* Prediction of Precipitation Curves

The two QSAR models were used to predict the m^* and β-values for an external test set of three conditions that were excluded from the generation of the models, namely, α-lactalbumin at pH 5.0, avidin at pH 10.0, and mAb1 at pH 10.0. For this purpose, molecular descriptors calculated based on their 3D structures were employed. The predicted values (shown as blue dots in Figures 4 and 6) were then used to calculate the complete precipitation curves. These *in silico* generated precipitation curves are illustrated and compared with the experimental data in Figure 8.

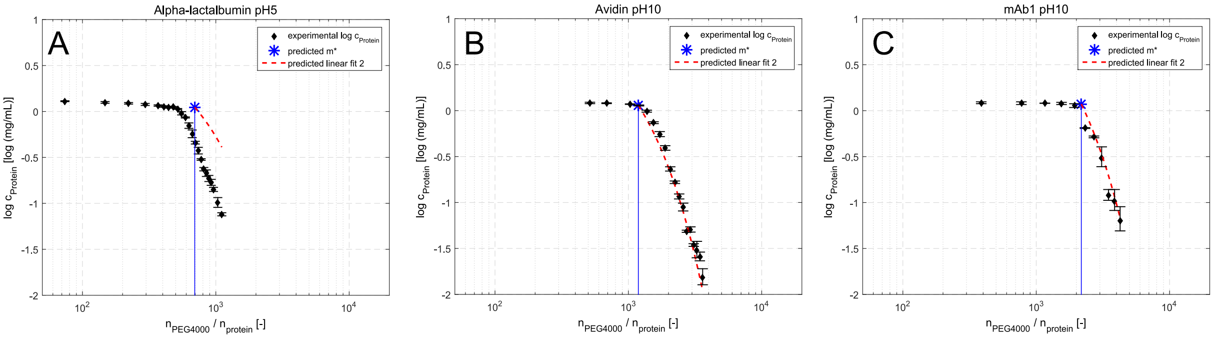


Figure 8: Predicted solubility curves for alpha-lactalbumin pH 5.0 (A), avidin pH 10.0 (B), and mAb1 pH 10.0 (C). m^* and β were predicted with the respective QSAR model, experimental data are shown for comparison with *in silico* generated data.

Very good agreement was achieved between the predicted and experimentally determined precipitation curves for avidin and mAb1. The predicted values for m^* were 695 (experimental: 514) for α-lactalbumin, 1188 (experimental: 1202) for avidin, and 2178 (experimental: 1938) for mAb1. For β, the QSAR model predicted values of -10.46 $\cdot 10^{-4}$ $log\ (mg/mL)$ (experimental: -20.28) for α-lactalbumin, -8.35 $\cdot 10^{-4}$ $log\ (mg/mL)$ (experimental: -7.84) for avidin, and -5.97 $\cdot 10^{-4}$ $log\ (mg/mL)$ (experimental: -5.57) for mAb1. The discrepancy between the observed values and the values predicted by both models in case of small proteins like α-lactalbumin was discussed earlier in Section 3.2.2. It is caused by the Coulomb potential that is more pronounced for small proteins or the interpenetration of PEG and small proteins as discussed earlier. Both models proved to be valid for generating *in silico* precipitation curves for proteins of different size and shape. The QSAR approach, thus, allows for obtaining a deeper mechanistic process understanding in accordance with the quality by design guideline.

4 Conclusion

To the best of our knowledge, we were the first to successfully apply QSAR modeling in the field of protein precipitation and expanded this methodology to an alternative protein purification technique other than chromatography. The results allow for obtaining a semi-mechanistic understanding of protein precipitation by polyethylene glycol and, hence, will help to implement this technology in biopharmaceutical industry to support the quality by design approach. A QSAR model for each precipitation curve parameter m^* and β based on molecular descriptors obtained from protein 3D structures was introduced for protein precipitation by PEG4000. The models were generated with a data set obtained from precipitation experiments using nine different proteins, including one mAb, at varying pH values and ionic strengths. For both parameters, a $R^2 \geq 0.90$ was obtained, which reflects a good correlation between observed and predicted data. The predictability Q^2 with a value of 0.8 was good for m^*, but moderate for β with a value of 0.63. Both models provided valuable insights into the structural properties of proteins that account for differences in both parameters. It was found that the protein molecular structure properties and electrostatic surface characteristics have the main impact on the value of the discontinuity point m^*. In case of β, variance in electrostatic surface potential and hydrophobicity were found to be the main properties of the molecule influencing the slope of the precipitation curve.

The generated models were applied to an external test set of three combinations of protein type, pH, and ionic strength that were excluded from the generation of both models and the entire precipitation curves were calculated *in silico*. For two of the three conditions, these predictions were accurate, while a deviation was observed for α-lactalbumin, which might have been caused by the model's perturbation for small molecules. To the best of our knowledge, this is the first publication of a method enabling the prediction of complete precipitation curves for proteins by polyethylene glycol.

The presented method can accelerate process development for purification and formulation of biopharmaceuticals following the tenet of quality by design. Future work should address the integration of additional molecular descriptors for polymers into the QSAR models as well as the introduction of parameters considering interactions in protein mixtures.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This research work is part of the project 'Optimization of an industrial process for the production of cell-culture-based seasonal and pandemic influenza vaccines' (grant number 0315640B) funded by the German Federal Ministry of Education and Research (BMBF) and was supported by Lonza Biologics PLC. We thank Cathrin Dürr for preparing the PDB files, Oliver Lorenz-Cristea and Jan Tobias Weggen for performing part of the experimental work, and Pascal Baumann and Maike Schröder for proofreading this manuscript.

References

[1] A. C. A. Roque, C. R. Lowe, M. A. Taipa, Antibodies and genetically engineered related molecules: production and purification., Biotechnology Progress 20 (3) (2004) 639–54. doi:10.1021/bp030070k.

[2] U. Gottschalk, K. Brorson, A. A. Shukla, The need for innovation in biomanufacturing, Nature Biotechnology 30 (6) (2012) 489–492. doi:10.1038/nbt.2263.

[3] T. M. Przybycien, N. S. Pujar, L. M. Steele, Alternative bioseparation operations: life beyond packed-bed chromatography, Current Opinion in Biotechnology 15 (5) (2004) 469–478. doi:10.1016/j.copbio.2004.08.008.

[4] S. Sommerfeld, J. Strube, Challenges in biotechnology production-generic processes and process optimization for monoclonal antibodies, Chemical Engineering and Processing: Process Intensification 44 (10) (2005) 1123–1137. doi:10.1016/j.cep.2005.03.006.

[5] D. Low, R. O'Leary, N. S. Pujar, Future of antibody purification, Journal of Chromatography B 848 (1) (2007) 48–63. doi:10.1016/j.jchromb.2006.10.033.

[6] S. M. Cramer, M. A. Holstein, Downstream bioprocessing: recent advances and future promise, Current Opinion in Chemical Engineering 1 (1) (2011) 27–37. doi:10.1016/j.coche.2011.08.008.

[7] S. A. Oelmeier, C. Ladd-Effio, J. Hubbuch, Alternative separation steps for monoclonal antibody purification: Combination of centrifugal partitioning chromatography and precipitation, Journal of Chromatography A 1319 (2013) 118–126. doi:10.1016/j.chroma.2013.10.043.

[8] N. Hammerschmidt, A. Tscheliessnig, R. Sommer, B. Helk, A. Jungbauer, Economics of recombinant antibody production processes at various scales: Industry-standard compared to continuous precipitation, Biotechnology Journal 9 (6) (2014) 766–775. doi:10.1002/biot.201300480.

[9] J. Thömmes, M. Etzel, Alternatives to Chromatographic Separations, Biotechnology Progress 23 (1) (2007) 42–45. doi:10.1021/bp0603661.

[10] ICQ Quality Implementation Working Group, ICH-Endorsed Guide for ICH Q8/Q9/Q10 Implementation, 2011, Tech. Rep. December (2011).

[11] S. Chhatre, S. S. Farid, J. Coffman, P. Bird, A. R. Newcombe, N. J. Titchener-Hooker, How implementation of Quality by Design and advances in Biochemical Engineering are enabling efficient bioprocess development and manufacture, Journal of Chemical Technology & Biotechnology 86 (9) (2011) 1125–1129. doi:10.1002/jctb.2628.

[12] P. Baumann, J. Hubbuch, Downstream process development strategies for effective bioprocesses: Trends, progress, and combinatorial approaches, Engineering in Life Sciences (2016) 1–29doi:10.1002/elsc.201600033.

[13] C. B. Mazza, N. Sukumar, C. M. Breneman, S. M. Cramer, Prediction of protein retention in ion-exchange systems using molecular descriptors obtained from crystal structure., Analytical chemistry 73 (22) (2001) 5457–61.

[14] C. B. Mazza, C. E. Whitehead, C. M. Breneman, S. M. Cramer, Predictive quantitative structure retention relationship models for ion-exchange chromatography, Chromatographia 56 (3-4) (2002) 147–152. doi:10.1007/BF02493203.

[15] A. Ladiwala, F. Xia, Q. Luo, C. M. Breneman, S. M. Cramer, Investigation of protein retention and selectivity in HIC systems using quantitative structure retention relationship models., Biotechnology and Bioengineering 93 (5) (2006) 836–50. doi:10.1002/bit.20771.

[16] T. Yang, C. M. Breneman, S. M. Cramer, Investigation of multi-modal high-salt binding ion-exchange chromatography using quantitative structure-property relationship modeling, Journal of Chromatography A 1175 (1) (2007) 96–105. doi:10.1016/j.chroma.2007.10.037.

[17] J. Buyel, J. Woo, S. Cramer, R. Fischer, The use of quantitative structure-activity relationship models to develop optimized processes for the removal of tobacco host cell proteins during biopharmaceutical production, Journal of Chromatography A 1322 (2013) 18–28. doi:10.1016/j.chroma.2013.10.076.

[18] S. Tsoka, O. Ciniawskyj, O. Thomas, N. Titchener-Hooker, M. Hoare, Selective Flocculation and Precipitation for the Improvement of Virus-Like Particle Recovery from Yeast Homogenate, Biotechnology Progress 16 (4) (2000) 661–667. doi:10.1021/bp0000407.

[19] S. Matheus, W. Friess, D. Schwartz, H.-C. Mahler, Liquid high concentration IgG1 antibody formulations by precipitation, Journal of Pharmaceutical Sciences 98 (9) (2009) 3043–3057. doi:10.1002/jps.21526.

[20] M. Kuczewski, E. Schirmer, B. Lain, G. Zarbis-Papastoitsis, A single-use purification process for the production of a monoclonal antibody produced in a PER.C6 human cell line, Biotechnology Journal 6 (1) (2011) 56–65. doi:10.1002/biot.201000292.

[21] S.-L. Sim, T. He, A. Tscheliessnig, M. Mueller, R. B. Tan, A. Jungbauer, Branched polyethylene glycol for protein precipitation, Biotechnology and Bioengineering 109 (3) (2012) 736–746. doi:10.1002/bit.24343.

[22] S. Asakura, F. Oosawa, Interaction between particles suspended in solutions of macromolecules, Journal of Polymer Science 33 (126) (1958) 183–192. doi:10.1002/pol.1958.1203312618.

[23] T. Odijk, Depletion Theory and the Precipitation of Protein by Polymer, The Journal of Physical Chemistry B 113 (12) (2009) 3941–3946. arXiv:0807.4997, doi:10.1021/jp806722j.

[24] P. Iverius, T. Laurent, Precipitation of some plasma proteins by the addition of dextran or polyethylene glycol, Biochimica et Biophysica Acta (BBA) - Protein Structure 133 (2) (1967) 371–373. doi:10.1016/0005-2795(67)90079-7.

[25] A. Polson, A Theory for the Displacement of Proteins and Viruses with Polyethylene Glycol, Preparative Biochemistry 7 (2) (1977) 129–154. doi:10.1080/00327487708061631.

[26] A. Tardieu, F. Bonneté, S. Finet, D. Vivarès, Understanding salt or PEG induced attractive interactions to crystallize biological macromolecules, Acta Crystallographica Section D Biological Crystallography 58 (10) (2002) 1549–1553. doi:10.1107/S0907444902014439.

[27] J. Lee, H. T. Gan, S. M. A. Latiff, C. Chuah, W. Y. Lee, Y.-S. Yang, B. Loo, S. K. Ng, P. Gagnon, Principles and applications of steric exclusion chromatography, Journal of Chromatography A 1270 (2012) 162–170. doi:10.1016/j.chroma.2012.10.062.

[28] D. H. Atha, K. C. Ingham, Mechanism of precipitation of proteins by polyethylene glycols. Analysis in terms of excluded volume., The Journal of Biological Chemistry 256 (23) (1981) 12108–17.

[29] C. Knevelman, J. Davies, L. Allen, N. J. Titchener-Hooker, High-throughput screening techniques for rapid PEG-based precipitation of IgG4 mAb from clarified cell culture supernatant, Biotechnology Progress 26 (3) (2009) 697–705. doi:10.1002/btpr.357.

[30] I. Juckles, Fractionation of proteins and viruses with polyethylene glycol, Biochimica et Biophysica Acta (BBA) - Protein Structure 229 (3) (1971) 535–546. doi:10.1016/0005-2795(71)90269-8.

[31] K. C. Ingham, Precipitation of proteins with polyethylene glycol: Characterization of albumin, Archives of Biochemistry and Biophysics 186 (1) (1978) 106–113. doi:10.1016/0003-9861(78)90469-1.

[32] S.-L. Sim, T. He, A. Tscheliessnig, M. Mueller, R. B. Tan, A. Jungbauer, Protein precipitation by polyethylene glycol: A generalized model based on hydrodynamic radius, Journal of Biotechnology 157 (2) (2012) 315–319. doi:10.1016/j.jbiotec.2011.09.028.

[33] A. Polson, G. Potgieter, J. Largier, G. Mears, F. Joubert, The fractionation of protein mixtures by linear polymers of high molecular weight, Biochimica et Biophysica Acta (BBA) - General Subjects 82 (3) (1964) 463–475. doi:10.1016/0304-4165(64)90438-6.

[34] E. J. Cohn, The Physical Chemistry of the Proteins, Physiological Reviews 5 (3) (1925) 349–437.

[35] T. M. Przybycien, J. E. Bailey, Solubility-activity relationships in the inorganic salt-induced precipitation of α-chymotrypsin, Enzyme and Microbial Technology 11 (5) (1989) 264–276. doi:10.1016/0141-0229(89)90041-0.

[36] The Uniprot Consortium, UniProt: a hub for protein information, Nucleic Acids Research 43 (D1) (2015) D204–D212. doi:10.1093/nar/gku989.

[37] E. Gasteiger, C. Hoogland, A. Gattiker, S. Duvaud, M. R. Wilkins, R. D. Appel, A. Bairoch, Protein Identification and Analysis Tools on the ExPASy Server, in: The Proteomics Protocols Handbook, Humana Press, Totowa, NJ, 2005, pp. 571–607. doi:10.1385/1-59259-890-0:571.

[38] S. A. Oelmeier, C. Ladd Effio, J. Hubbuch, High throughput screening based selection of phases for aqueous two-phase system-centrifugal partitioning chromatography of monoclonal antibodies, Journal of Chromatography A 1252 (2012) 104–114. doi:10.1016/j.chroma.2012.06.075.

[39] F. Hachem, B. Andrews, J. Asenjo, Hydrophobic partitioning of proteins in aqueous two-phase systems, Enzyme and Microbial Technology 19 (7) (1996) 507–517. doi:10.1016/S0141-0229(96)80002-D.

[40] H. M. Berman, The Protein Data Bank, Nucleic Acids Research 28 (1) (2000) 235–242. doi:10.1093/nar/28.1.235.

[41] E. Krieger, G. Koraimann, G. Vriend, Increasing the precision of comparative models with YASARA NOVA-a self-parameterizing force field, Proteins: Structure, Function, and Bioinformatics 47 (3) (2002) 393–402. doi:10.1002/prot.10104.

[42] K. M. Lang, J. Kittelmann, C. Dürr, A. Osberghaus, J. Hubbuch, A comprehensive molecular dynamics approach to protein retention modeling in ion exchange chromatography, Journal of Chromatography A 1381 (2015) 184–193. doi:10.1016/j.chroma.2015.01.018.

[43] Y. Duan, C. Wu, S. Chowdhury, M. C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, J. Caldwell, J. Wang, P. Kollman, A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations, Journal of Computational Chemistry 24 (16) (2003) 1999–2012. doi:10.1002/jcc.10349.

[44] R. Anandakrishnan, B. Aguilar, A. V. Onufriev, H++ 3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations, Nucleic Acids Research 40 (W1) (2012) W537–W541. doi:10.1093/nar/gks375.

[45] M. Biasini, S. Bienert, A. Waterhouse, K. Arnold, G. Studer, T. Schmidt, F. Kiefer, T. G. Cassarino, M. Bertoni, L. Bordoli, T. Schwede, SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information, Nucleic Acids Research 42 (W1) (2014) W252–W258. doi:10.1093/nar/gku340.

[46] J. Kyte, R. F. Doolittle, A simple method for displaying the hydropathic character of a protein, Journal of Molecular Biology 157 (1) (1982) 105–132. doi:10.1016/0022-2836(82)90515-0.

[47] F. Dismer, J. Hubbuch, 3D structure-based protein retention prediction for ion-exchange chromatography, Journal of Chromatography A 1217 (8) (2010) 1343–1353. doi:10.1016/j.chroma.2009.12.061.

[48] A. Tropsha, P. Gramatica, V. Gombar, The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models, QSAR & Combinatorial Science 22 (1) (2003) 69–77. doi:10.1002/qsar.200390007.

[49] L. Eriksson, J. Jaworska, A. P. Worth, M. T. Cronin, R. M. McDowell, P. Gramatica, Methods for Reliability and Uncertainty Assessment and for Applicability Evaluations of Classification- and Regression-Based QSARs, Environmental Health Perspectives 111 (10) (2003) 1361–1375. doi:10.1289/ehp.5758.

[50] R. Kiralj, M. M. C. Ferreira, Basic validation procedures for regression models in QSAR and QSPR studies: theory and application, Journal of the Brazilian Chemical Society 20 (4) (2009) 770–787. doi:10.1590/S0103-50532009000400021.

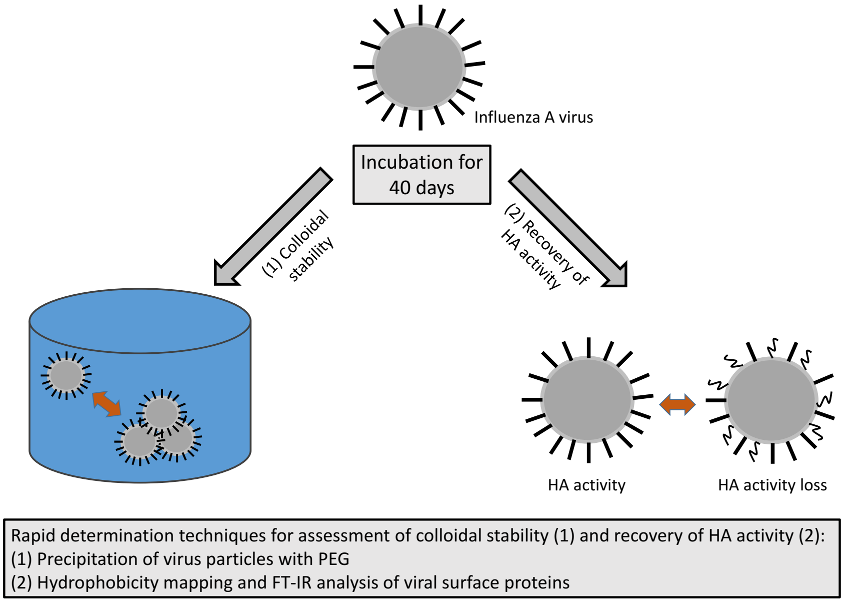
[51] W. Kessler, Multivariate Datenanalyse, Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, 2007. doi:10.1002/9783527610037.

[52] J. Li, R. Rajagopalan, J. Jiang, Polymer-induced phase separation and crystallization in immunoglobulin G solutions., The Journal of Chemical Physics 128 (20) (2008) 205105. doi:10.1063/1.2919565.

[53] C. Rücker, G. Rücker, M. Meringer, y-Randomization and Its Variants in QSPR/QSAR, Journal of Chemical Information and Modeling 47 (6) (2007) 2345–2357. doi:10.1021/ci700157b.

Strategy for Assessment of the Colloidal and Biological Stability of H1N1 Influenza A Viruses

Frank Hämmerling, Oliver Lorenz-Cristea, Pascal Baumann and Jürgen Hubbuch*



Rapid determination techniques for assessment of colloidal stability (1) and recovery of HA activity (2):
(1) Precipitation of virus particles with PEG
(2) Hydrophobicity mapping and FT-IR analysis of viral surface proteins

Institute of Engineering in Life Sciences, Section IV: Biomolecular Separation Engineering, Karlsruhe Institute of Technology, Engler-Bunte-Ring 3, 76131 Karlsruhe, Germany

* : *Corresponding author; email address: juergen.hubbuch@kit.edu*

International Journal of Pharmaceutics

Volume 517, (2017), 80-87

Abstract

Current influenza vaccines are mostly formulated as liquids which requires a continuous cold chain to maintain the stability of the antigen. For development of vaccines with an increased stability at ambient temperatures, manifold parameters and their influences on the colloidal stability and activity of the antigen have to be understood. This work presents a strategy to examine both, the colloidal stability and the remaining biological activity of H1N1 influenza viruses under various conditions after an incubation of 40 days. H1N1 phase diagrams were generated for several pH values and different initial H1N1 and NaCl concentrations. It was shown that the highest H1N1 recoveries were obtained for pH 6 and that moderate amounts of NaCl are favorable for increased recoveries. In contrast to colloidal stability, the highest remaining HA activity was observed at pH 9. The electrostatic and hydrophobic surface properties of H1N1 were investigated to reveal the mechanisms accounting for the decrease in stability. Secondly, the capability of virus precipitation by polyethylene glycol in combination with determination of surface hydrophobicity was proven to be useful as a predictive tool to rank stability under different conditions. This methodology enables the rapid assessment of aggregation propensity of H1N1 formulations and the influence on the activity of the virus particles and might become a standard tool during the development of vaccine formulations.

Keywords: pandemic influenza virus, stability, surface properties, formulation, virus phase diagram

1 Introduction

According to the WHO, seasonal influenza has an estimated annual attack rate of 5% - 10% in adults worldwide which results in 3 to 5 million cases of severe illness and up to 500,000 deaths every year [1]. Influenza A viruses not only cause seasonal epidemics, they also show the potential to cause worldwide pandemics by genetic changes, host changes, and introduction of a virus with a novel surface protein subtype that is new to human populations [2, 3]. Vaccination is recognized as the most effective strategy to prevent and control the spread of influenza. Currently, there are two types of influenza vaccine formulations on the market. They are either formulated as liquids or lyophilized in a solid state [4]. Lyophilized or dry-state influenza vaccine formulations, as presented by Anamur et al. [5], Amorij et al. [6], and Garmise et al. [7], are attended by an increased stability, but also accompanied by several drawbacks. These include an exposure to a variety of environmental stresses [8], destabilizing effects during reconstitution to the liquid phase prior to administration [9], the need of appropriate excipients that may also interact with the antigen, and additional expenses for development and process costs for lyophilization steps [10]. The majority of the current influenza vaccines are provided as liquid formulations which are known to be temperature-sensitive and require a continuous cold chain. To maintain the antigens' activity, this cold chain is mandatory during distribution and storage of vaccines until administration [6, 11]. Temperature affects both the conformational and colloidal stability of viral proteins. Elevated temperatures can provoke changes in the folding of proteins whereby hydrophobic amino acids that were buried in the hydrophobic core of the protein get exposed to the surface of the molecule. As a consequence, attractive hydrophobic interactions, that enhance aggregation, emerge [12, 13]. The diffusion of molecules in solution is directly related to the absolute temperature as described by the Stokes-Einstein equation [14]. Hence, elevated temperatures increase the diffusion of molecules which results in a higher propensity for aggregation [15, 13]. According to a study published in 2001, vaccines worth US$ 6-31 million were wasted in the U.S., predominantly due to a discontinuous cold chain [16]. As a robust cold chain is difficult to accomplish especially in developing countries [17, 18] and the contribution of the cold chain incorporates approximately 80% of the costs of vaccination programs in developing countries [19], there is an urgent need to reduce the dependency on this factor. Therefore, the parameters and mechanisms that influence and reduce the formulation stability of vaccine formulations at moderate temperatures have to be assessed systematically and rapid methods for the determination of long-term stability need to be developed. Stable vaccine formulations maintain the antigens' native biological activity and immunogenicity until its administration [20]. Being more general, the stability of a pharmaceutical product may be defined as the capability of a particular formulation in a specific container system, to remain within its physical, chemical, microbiological, therapeutic, and toxicological specifications [21]. The stability of a vaccine is influenced by a number of environmental conditions being mainly the pH value, type and concentration of added salt, the redox potential, the temperature, and the presence of different stabilizing excipients [20, 9]. In aqueous environment, the antigens are subject to physical and chemical degradation such as aggregation, denaturation, conformational changes, and consequently the loss of activity [6, 9]. Virus particles in solution interact with each other. In the case of influenza viruses, the surface proteins account for virus-

virus interactions. The two glycoproteins hemagglutinin (HA) and neuraminidase (NA) form the peplomer [22] and therefore determine the interactions between the viruses in terms of protein-protein interactions.

The net charge of a biomolecule is strongly dependent on the pH value of the solution. The resulting electrostatic interactions have a repulsive character and therefore incorporate a stabilizing effect on protein solution. Changes of the pH value strongly influence the protonation state of amino acids and amino acid side chains, respectively. These changes in charge distribution can influence the tertiary structure of the protein and its folding which might be accompanied by decreased colloidal stability or a loss in activity [13]. Burke et al. [23] investigated the influence of the pH on the activity of several vaccines. They report a significant loss of activity for an influenza A vaccine below pH 7 and above pH 10, while the activity was fairly maintained in the range of pH 7 and pH 10. Miller [24] observed a complete drop in the infectivity of A/PR/8 influenza viruses within one hour at pH 3, pH 4, and pH 5 at room temperature. Furthermore, precipitation of viruses was observed at pH 4 and pH 5. Under conditions between pH 6 and pH 9, the viruses exhibited an increased stability at pH 6 and pH 7. For these conditions, about 10% of activity remained after 10 days, and after 30 days, the complete activity was lost. A loss in activity was also observed when storing the virus at 4 °C in a diluted state with a concentration below 0.1 mg/mL, while the activity remained constant for the corresponding virus stock solution of 2 mg/mL. At pH values far from the isoelectric point (pI), proteins are charged strongly. The increased charge repulsion within the protein molecule destabilizes the conformation and leads to a pH-induced unfolding [15, 25, 26].

Besides pH, ions in a protein solution have complex effects on the aggregation of proteins. They have the potential to bind or to interact electrostatically with the protein molecule [13]. The addition of salt ions to a protein solution causes a shielding of repulsive and thus stabilizing long-range electrostatic interactions. Another consequence resulting from the addition of salt is the 'salting in' and 'salting out' effect. If the added ions preferentially bind to proteins ('chaotropic ions'), the protein's net charge increases as well as its solubility. This effect is referred to as salting in effect. With NaCl, the solubility shows a bell-shaped behavior and maximum solubility was observed at NaCl concentrations up to 2.0 - 2.5 M NaCl. The chaotropic effects of certain ions might also decrease the intramolecular stability at high concentrations. For the salting out effect, per contrast, polar and strongly hydrated ions ('kosmotropic ions') retract water from the protein surface and, thus, expose hydrophobic surface patches and thereby decrease the solubility by encouraging the protein to minimize its solvent accessible surface area [27, 9, 28, 29, 30]. For particles with semipermeable membranes such as bacteria and enveloped influenza viruses, high ionic strengths might additionally induce lysis of the membrane [20, 9].

In summary, these parameters were shown to have strong effects on virus stability. Hence, there is a need for fast and effective tools to monitor and predict virus stability. In this work, we use automated high-throughput-compatible methods, only requiring a very low sample volume, to investigate the influence of several parameters, namely the initial H1N1 concentration, pH value, and ionic strength on the colloidal and biological stability of inactivated H1N1 influenza viruses. After an incubation of 40 days at 20 °C, the recovery of H1N1 and the remaining HA activity in the supernatant are evaluated. Secondly, the electrostatic and hydrophobic surface characteristics of the H1N1 influenza viruses are determined to gain an understanding of the mechanisms leading to aggregation of the virus particles. Furthermore, the precipitation of H1N1 by polyethylene glycol is assessed as a rapid methodology to determine the aggregation propensity under the investigated conditions.

2 Materials and Methods

2.1 H1N1 Influenza A Virus

The pandemic influenza A/Jena/5258/2009 (H1N1) virus feedstream was generously provided by IDT Biologika GmbH (Dessau-Roßlau, Germany). Influenza viruses were cultivated in Madin-Darby Bovine Kidney cells and harvested two days post infection. Subsequently, the virus particles were inactivated with β-propiolactone and concentrated 20-fold.

2.1.1 Purification of H1N1

H1N1 was purified by size exclusion chromatography (SEC) using a Toyopearl® HW-65S resin (Tosoh Bioscience GmbH, Griesheim, Germany) and anion-exchange chromatography (AEX) operated in flow-through mode using a Capto Q resin (GE Healthcare, Uppsala, Sweden). 20 mM Tris buffer at pH 7.5 containing 500 mM sodium chloride was used for both chromatography steps. Both chromatography processes were performed with an ÄKTApurifier system from GE Healthcare (Uppsala, Sweden) which was controlled by UNICORN 5.31. Bed volumes were 150 mL for the SEC column and 12 mL for the AEX column. SEC was operated with a flow rate of 3 mL/min and AEX was performed with a flow rate of 2.5 mL/min. A sample volume of 50 mL was loaded onto the columns for SEC and AEX. Elution of contaminants during AEX was achieved by changing the eluent to 20 mM Tris buffer at pH 7.5 containing 1500 mM NaCl and a stripping step with 1 M sodium hydroxide. Purified H1N1 was concentrated with a prototype Sartocon® Slice 200 Hydrosart® membrane with a 300 kDa cut-off (Sartorius Stedim Biotech, Göttingen, Germany) using a Cogent® μScale tangential flow system (Millipore Corporation, Billerica, MA, USA) to a final HA concentration of 12,000 HAU/100 μL. Throughout concentration, transmembrane pressure was set to 0.3 bar at a feed flow rate of 50 mL/min. Final purified and concentrated H1N1 sample had an UV absorption at 280 nm of 8.19 AU as determined with the NanoDrop2000c spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA). During purification, a depletion of >99% of host cell proteins and DNA was achieved. Aliquots of the final sample were stored at -80 °C. For exchanging the buffer prior to the experiments, PD MiniTrap G-25 columns (GE Healthcare, Uppsala, Sweden) were used, following the spin protocol.

2.2 Hemagglutination Assay

The hemagglutination assay is a rapid and simple method that can be used to determine levels of influenza virus present in a sample. The hemagglutinin protein on the surface of the influenza virus particles is capable of binding to N-acetylneuraminic acid-containing proteins on avian and mammalian erythrocytes. When the influenza virus is present in a sufficient concentration, there is an agglutination reaction and the erythrocytes link together to form a diffuse lattice. Otherwise, point sedimentation of erythrocytes occurs [31, 32]. Erythrocytes from chicken blood, stabilized in Alsever's solution, were purchased from preclinics GmbH (Potsdam, Germany). Sedimentation behavior was evaluated by absorbance measurements at 700 nm with an Infinite® 200 UV-Vis spectrophotometer (Tecan GmbH, Crailsheim, Germany). The complete protocols for the preparation of the erythrocyte solution and execution of the hemagglutination assay were conducted as described in detail by Kalbfuss et al. [33].

2.3 Automated Generation of H1N1 Phase Diagrams

The colloidal stability of H1N1 was investigated through phase diagrams after an incubation of 40 days. All buffers for these experiments were prepared with a concentration of 20 mM using acetic acid for pH 4.5 (Merck KGaA, Darmstadt, Germany), MES for pH 6, Tris for pH 7.5 (Merck KGaA, Darmstadt, Germany), and TAPS for pH 9 (AppliChem GmbH, Darmstadt, Germany) in ultra-pure water. Sodium chloride purchased from Merck KGaA (Darmstadt, Germany) was added to adjust the ionic strength. The pH was controlled using a pH meter HI-3220 (Hanna® Instruments, Woonsocket, RI, USA) equipped with a SenTix® 62 pH electrode (Xylem Inc., White Plains, NY, USA) and corrected by titration with hydrochloric acid or sodium hydroxide, respectively (both Merck KGaA, Darmstadt, Germany). Sodium azide (Merck KGaA, Darmstadt, Germany) was added to a final concentration of 0.02%(w/v) to the buffers in the phase diagrams to inhibit microbial growth. Phase diagrams for H1N1 were generated according to the method described by Baumgartner et al. [34] in a 24 μL microbatch format using the automated liquid handling station Tecan Freedom Evo 100 (Tecan GmbH, Crailsheim, Germany). Briefly, H1N1 viruses were transferred into the respective buffer including 200 mM NaCl and diluted to seven concentrations between 12,000 and 4,800 HAU/100 μL of virus sample. Based on a high salt buffer containing 2.5 M NaCl and a low salt buffer containing 0 M NaCl, a set of twelve buffer compositions containing NaCl in a range of 0 M and 2.5 M was prepared. 12 μL of each of the H1N1 diluted sample and the diluted NaCl buffer were transferred into a MRC Under Oil 96 Well Crystallization Plate (Swissci AG, Neuheim, Switzerland) and sealed with HDclear™ sealing tape (ShurTech Brands, Avon, OH, USA) to prevent evaporation. Plates were incubated at 20 °C in the Rock Imager 54 (Formulatrix, Waltham, MA, USA) for 40 days. This device was used as an automated system for periodical imaging of the crystallization plates and determining phase transitions. After incubation, the H1N1 concentration in the supernatant was determined spectrophotometrically with a NanoDrop2000c (Thermo Fisher Scientific, Waltham, MA, USA) UV-Vis spectrophotometer. For this purpose, the UV absorption at 280 nm of the initial sample was set in relation to the measured absorption in the supernatant after 40 days. The extinction coefficient of H1N1 was kept constant for initial samples and samples after 40 days of incubation, bearing in mind that it might be subject to slight variations due to conformational changes of the virus. Additionally, the hemagglutinin activity in the supernatant was determined with the hemagglutination assay for selected wells (Section 2.2).

2.4 Zeta Potential Measurements

The determination of zeta potential was performed to capture the surface charge and the resulting electrostatic interactions of H1N1 at the investigated pH values. The experiments were carried out with the Zetasizer Nano ZSP (Malvern Instruments Ltd, Malvern, UK) by measuring the electrophoretic mobility of the virus particles. The experiments were performed in triplicates in the respective buffer with addition of 100 mM NaCl. A voltage of 25 V was applied throughout the measurement.

2.5 Hydrophobicity Determination of H1N1

Hydrophobic interactions are one of the crucial parameters leading to aggregation of biomolecules. Hence, the knowledge of the surface hydrophobicity of biomolecules is mandatory to assess and control the influence of hydrophobic interactions during all stages of a purification and formulation process. The determination of the hydrophobicity of the H1N1 virus particles was performed according to the stalagmometric method published by Amrhein et al. [35, 36]. This methodology enables to calculate the surface tension of samples based on the mass of a drop: The sample is purged very slowly through a vertical capillary while drops grow up to a specific maximum volume and fall onto an analytical balance. By comparison with the mass of a drop of a reference solution with known surface tension (e.g. ultrapure water), the surface tension of the sample can be calculated. As more hydrophobic molecules exhibit a higher tendency to attach to the air-water interface, they decrease the surface tension of the sample which results in a lower mass of a drop. The determination of surface tension was conducted with the automated liquid handling station Tecan Freedom Evo 100 (Tecan GmbH, Crailsheim, Germany) and the high-precision analytical balance WXTS205DU (Mettler Toledo, Greifensee, Switzerland).

2.6 Fourier Transform Infrared (FT-IR) Spectroscopy

H1N1 solutions in the respective buffers of Section 2.3 containing 500 mM NaCl were scanned in absorbance mode with 400 scans using a BioATRCell II flow cell (Bruker Corporation, Billerica, MA, USA). Spectra were recorded from 4000 cm^{-1} to 900 cm^{-1} with a resolution of 2 cm^{-1}. Background spectra of the respective pure buffers were measured with the identical settings and subtracted from the sample spectra. The sample volume was set to 50 μL and all experiments were conducted as duplicates. Data was smoothed using the Savitzky-Golay filter, and the second derivative spectra were calculated using MATLAB R2015a (The MathWorks, Inc., Natick, MA, USA).

2.7 Generation of H1N1 Precipitation Curves

H1N1 influenza virus particles were precipitated using polyethylene glycol (PEG) with an average molecular weight of 600 g/mol (Merck KGaA, Darmstadt, Germany) as precipitant. The precipitation experiments were carried out on a Tecan Freedom Evo 200 (Tecan GmbH, Crailsheim, Germany) liquid handling platform. A 70%(w/w) stock solution of PEG600 with a density of 1.101 g/mL was used for the precipitation experiments. Purified H1N1 was re-buffered in the respective buffer of Section 2.3 containing 500 mM NaCl. The absorption at 280 nm was evaluated after buffer exchange to ensure that the H1N1 concentration was constant in all precipitation experiments. 30 μL of re-buffered H1N1 sample was added to every system with a total volume of 150 μL. The systems contained between 0 and 28%(w/w) PEG600. After incubation and separation of the precipitate by centrifugation (30 min at 4000 rpm), 75 μL of the supernatant was diluted with 50 μL buffer and analyzed for remaining H1N1 spectrophotometrically through absorption at 280 nm. The precipitation curves display the logarithmized protein concentration in the supernatant as a function of the PEG concentration in %(w/w). The m^* value depicts the PEG concentration, at which protein solubility equals the protein concentration initially set. This value was utilized for the evaluation of the precipitation curves in this work.

3 Results and Discussion

3.1 Evaluation of Phase Diagrams

The supernatant of the phase diagrams was evaluated for the determination of recovery of H1N1 and for selected wells of the remaining HA activity.

3.1.1 Mass Recovery of H1N1

Phase diagrams were prepared to evaluate H1N1 phase behavior and stability after an incubation of 40 days at 20 °C. The supernatant was measured by means of UV absorption at 280 nm and HA activity. Figure 1 displays the remaining H1N1 concentration in the supernatant determined by absorption measurements, as a function of the respective initial H1N1 and NaCl concentration. Contour lines depict a change of mass recovery by 10% and were added to guide the eye of the reader.

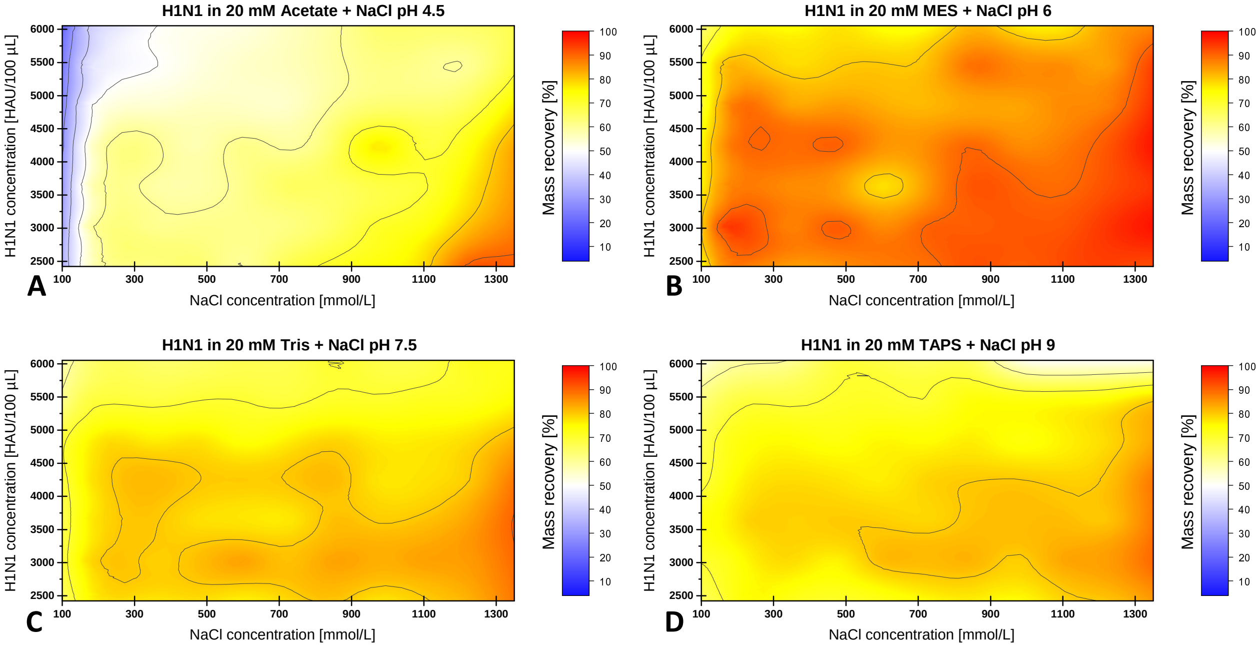


Figure 1: Phase diagrams of H1N1 using sodium chloride as precipitant at pH 4.5 (A), pH 6 (B), pH 7.5 (C), and pH 9 (D). The phase diagrams depict the mass recovery of H1N1 after an incubation of 40 days as a function of the respective initial H1N1 and NaCl concentration.

The highest average mass recovery among the investigated conditions was obtained for pH 6. The average H1N1 mass recovery was lower for pH 7.5 and pH 9 and again significantly lower for pH 4.5. For all phase diagrams, a higher H1N1 mass recovery was obtained with increasing NaCl concentrations due to the salting in effect of NaCl that was shown to increase solubility up to a concentration of 2.0 - 2.5 M [27]. A strong decrease of H1N1 in the supernatant for NaCl concentrations ≤100 mM NaCl was observed for all investigated pH values. For all pH values the H1N1 recoveries were higher under conditions with low initial H1N1 concentrations. In contrast, lower recoveries were determined under conditions with a high initial H1N1 concentration, this means that the solubility limit of H1N1 was exceeded under the latter conditions. A locally higher mass recovery was observed for all pH values at a NaCl concentration in a range of 300 mM over a wide range of initial HA concentrations. The maximum remaining H1N1 concentration ≥99% was determined for the system at pH 6 with an initial HA concentration of 3027 HAU/100 μL and 163 mM NaCl and the lowest one for the system at pH 4.5 with an initial HA concentration of 6054 HAU/100 μL including 100 mM NaCl.

The decreased H1N1 mass recovery for sodium chloride concentrations below 100 mM is caused by the pronounced hydrophobic character of the hemagglutinin surface protein [37]. Scopes [38] reported that proteins with a substantial hydrophobic amino acid content at their surface generally exhibit a low solubility under low-salt conditions, due to the tendency of these proteins to minimize unfavorable interactions between the aqueous solvent and exposed hydrophobic side chains. The low recoveries of H1N1 at pH 4.5 can be explained through the determined zeta potential which is depicted in Figure 2. The zeta potential was determined to be -2.9 mV at pH 4.5 and showed decreasing values with increasing pH values. For pH 9, a zeta potential of -10.7 mV was obtained.

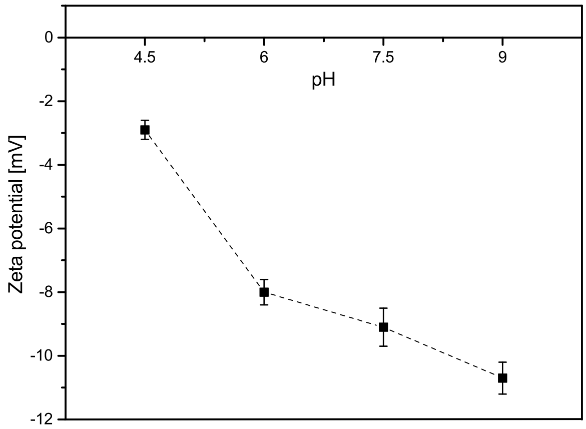


Figure 2: Zeta potential of H1N1 virus particles determined at different pH values.

According to these results, the isoelectric point of H1N1 is determined to be below pH 4.5. Consequently, at pH 4.5, the net charge of the virus particles exhibits the lowest value among the investigated pH values. Due to the lack of repulsive electrostatic interactions at pH 4.5, hydrophobic interactions between the surface proteins prevail which leads to aggregation of H1N1. With increasing pH values, the net charge of H1N1 increases and repulsive electrostatic interactions occur. This leads to increasing H1N1 recoveries at pH 6. At pH 7.5 and pH 9, under conditions where the virus particles are strongly negatively charged, the recovery decreases. This diminished H1N1 recovery is probably caused by a strong intramolecular charge repulsion that initiates structural changes within the surface proteins. Hydrophobic amino acid side chains that were buried in the core of the proteins get exposed to the surface of the molecules. To verify this assumption, the hydrophobicity of H1N1 was determined.

The hydrophobic surface properties of H1N1 were evaluated by measurements of the surface tension increment of the virus particles at pH 4.5, pH 6, pH 7.5, and pH 9. The normalized surface tension profiles are shown in Figure 3A.

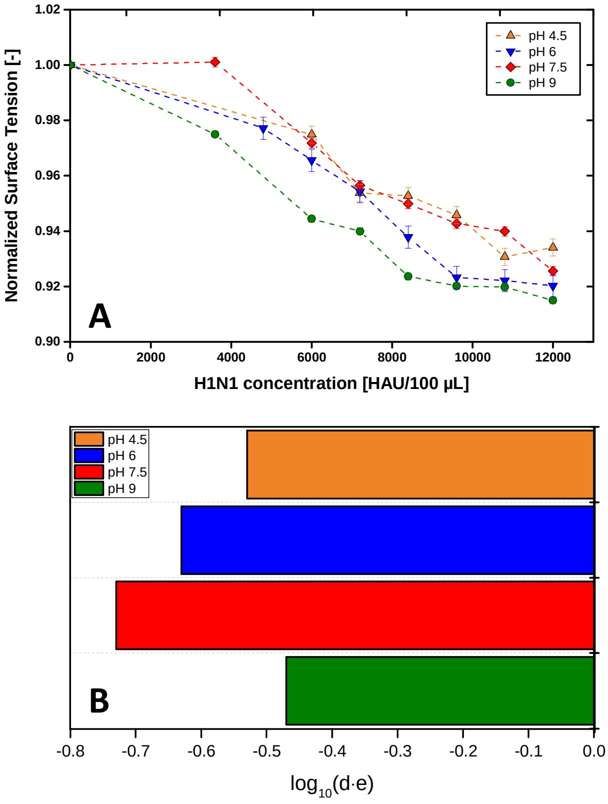


Figure 3: Surface tension profiles of H1N1 influenza virus particles at different pH values. Surface tension was normalized to the respective buffer (mean±STDEV (A). (B): Ranking according to H1N1 hydrophobic character was performed according to [36]: Hydrophobicity decreases with more negative values of $log_{10}(d \cdot e)$.

The profiles were fitted according to Amrhein et al. [36] with Equation 1, which was derived from a modified Langmuir adsorption isotherm. This has been shown to be appropriate, as the determined changes in surface tension arise from the adsorption of biomolecules to the air-water interface which leads to a decrease in surface tension. In this equation, c, d, and e represent the fitting parameters, and x is the HA activity in HAU/100 μL. The coefficient of correlation R^2 was determined to be $\geq$0.97 for pH 6, pH 7.5 and pH 9 and 0.93 for pH 4.5.

$$\gamma_{norm} = 1 - \frac{d \cdot e \cdot (x + c)}{1 + e \cdot (x + c)} \tag{1}$$

To obtain a ranking of hydrophobicity, the $log_{10}(d \cdot e)$ was calculated and is illustrated in Figure 3B. The higher the value of this hydrophobicity ranking, the more hydrophobic is the surface of the sample. H1N1 influenza viruses exhibit the highest hydrophobicity at pH 9. At pH 4.5, the hydrophobicity is in the same range but slightly decreased. Compared to these two pH values, H1N1 was found to be less hydrophobic at pH 6 and pH 7.5.

The determined hydrophobicities underline the increased hydrophobic character of H1N1 at pH 9. More hydrophobic character is triggered by an increased intramolecular charge repulsion within the surface proteins that are strongly negatively charged under this condition. The high charge density destabilizes the folded protein conformation so that hydrophobic amino acid side chains are partly exposed to the protein's surface which results in an increase of hydrophobicity [15, 13, 39]. At pH 4.5, in proximity to the isoelectric point of H1N1, the net charge of the virus particle is close to 0 and protein structure stabilizing electrostatic effects are reduced. This condition is also close to the pK_a value of glutamic acid ($pK_a = 4.2$) which is attended by a change in surface charge distribution of the surface proteins and is expected to have a large effect on the hydrophobicity of polypeptides [40]. Hughson [41] reported irreversible changes in the structure of hemagglutinin at pH values below pH 5-6 and Korte et al. [37] reported for HA the development of hydrophobic properties for acidic conditions. H1N1 depicts a more hydrophilic character at pH 6 and pH 7.5 which leads to the conclusion that these pH values are favorable for an increased colloidal stability of H1N1.

3.1.2 Recovery of HA Activity

Table 1 illustrates the recovered HA activities of selected conditions under the investigated pH values after 40 days.

Table 1: Remaining HA activity of selected systems of the phase diagrams after an incubation of 40 days.

pH	**NaCl**	**Initial HA activity**	**HA activity +40 days**	**Recovered HA activity**	**Recovered HA activity/mass recovery**
	[mmol/L]	[HAU/100 μL]	[HAU/100 μL]	[%]	[%]
4.5	475	5449	0	0	n/a
	1225	5449	0	0	n/a
	1225	2422	0	0	n/a
6	475	5449	371	6	6
	1225	5449	1317	24	29
	1225	2422	503	21	22
7.5	475	5449	4755	72	102
	1225	5449	4935	91	126
	1225	2422	1711	71	86
9	475	5449	6074	91	136
	1225	5449	5670	104	139
	1225	2422	1647	68	85

For systems with an initial HA activity of 5449 HAU/100 μL, the highest remaining HA activities among the investigated systems with values of ~100% were obtained for pH 9 for moderate and high NaCl concentrations. For pH 7.5, the recovered activities were reduced with values around 90%. Significantly lower recoveries of HA activity were observed at acidic pH values. For pH 6, the value decreased to 24% for 1225 mM NaCl and to 6% for 475 mM NaCl, respectively. No remaining HA activity was determined for all conditions at pH 4.5. Systems with an initial HA activity of 2422 HAU/100 μL resulted in lower recovered activities compared to the systems with higher initial concentration while keeping the NaCl concentration constant.

The complete loss in HA activity for all investigated systems at pH 4.5 results from the structural changes of hemagglutinin under acidic conditions [41]. The decrease of activity at pH 6 and a fairly constant activity in a range between pH 7 and pH 9 was also reported by Burke et al. [23]. The ratio between recovered HA activities and mass recovery of H1N1 can in reality of course never reach values >100%. Although the hemagglutination assay is known to come along with a comparatively high standard deviation in a range of 20% - 30%, it still is the gold standard for the determination of active influenza virus titers [33]. The comparatively low accuracy of the HA assay was also observed by Vajda et al. [42] and Kalbfuss et al. [33] and might also be influenced by the buffer composition or the aging of erythrocytes. In this work the accuracy of the HA assay was evaluated through a five-fold determination of the purified and concentrated stock solution, which resulted in a standard deviation of 7% (data not shown). In this context, the determined recovered HA activities after 40 days reflect the conserved trends within the experimental set-up and study presented here. The values >100% determined under pH 7.5 and 9 are probably due to salt contributions and dilution effects influencing the HA assay. The significantly lower recovery of HA activity under conditions with a low initial HA activity, as seen for the rows 2 and 3 of the respective pH values (Table 1), is a result of the dilution-induced destabilization that is still subject to further research [43]. This decrease in HA activity for diluted conditions is more distinct at pH 7.5 and pH 9 than at pH 6. Per contrast, the overall H1N1 recovery is higher under these diluted conditions.

3.1.3 Stability of H1N1 Surface Proteins

Changes in the second derivative of the Fourier transform infrared spectroscopy (FT-IR) spectra indicate conformational changes of the investigated samples. Figure 4 displays the second derivative spectra from Fourier transform infrared spectroscopy of H1N1 at pH 4.5, 6, 7.5, and 9.

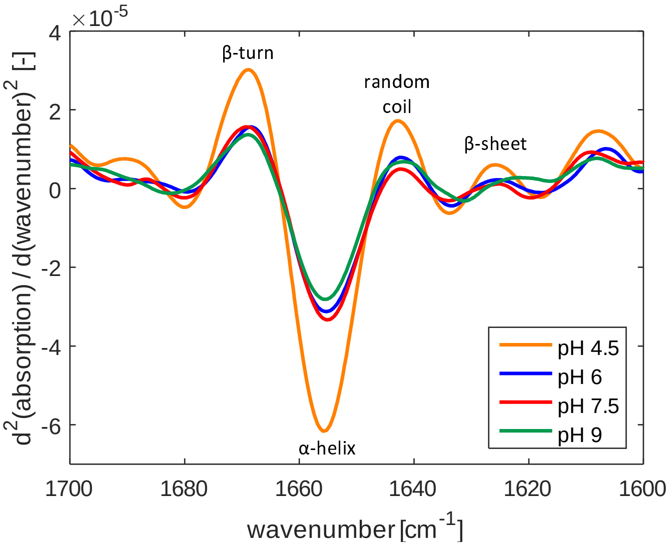


Figure 4: Second derivative spectra from Fourier transform infrared (FT-IR) spectroscopy measurements of H1N1 at pH 4.5 (orange), pH 6 (blue), pH 7.5 (red), and pH 9 (green).

α-helical structures are detected at wavenumbers around 1650 cm^{-1}, β-sheets can be identified at 1630 cm^{-1}, and random coil structures are measured at wavenumbers around 1640 cm^{-1} [44]. While the second derivative spectra for pH 6, pH 7.5, and pH 9 depict an almost identical course, significant differences were obtained for pH 4.5 resulting from structural changes of hemagglutinin under these conditions [41]. Among the native folded states, H1N1 exhibits the highest content of α-helical structures at pH 7.5. A slight dissipation and simultaneous increase of random coil structures was observed for pH 6 and pH 9. These changes especially pronounced for pH 9 might be an indication for the proposed structural changes at pH 9 due to strong intramolecular repulsive electrostatic interactions.

3.2 Precipitation of H1N1 by Polyethylene Glycol as Predictive Tool for Colloidal Stability

H1N1 influenza viruses were precipitated by PEG600 at pH 4.5, pH 6, pH 7.5, and pH 9 for investigating the capability of rapid precipitation experiments as an indicator for colloidal stability of H1N1 as an alternative to phase diagrams. The precipitation curves obtained are displayed in Figure 5.

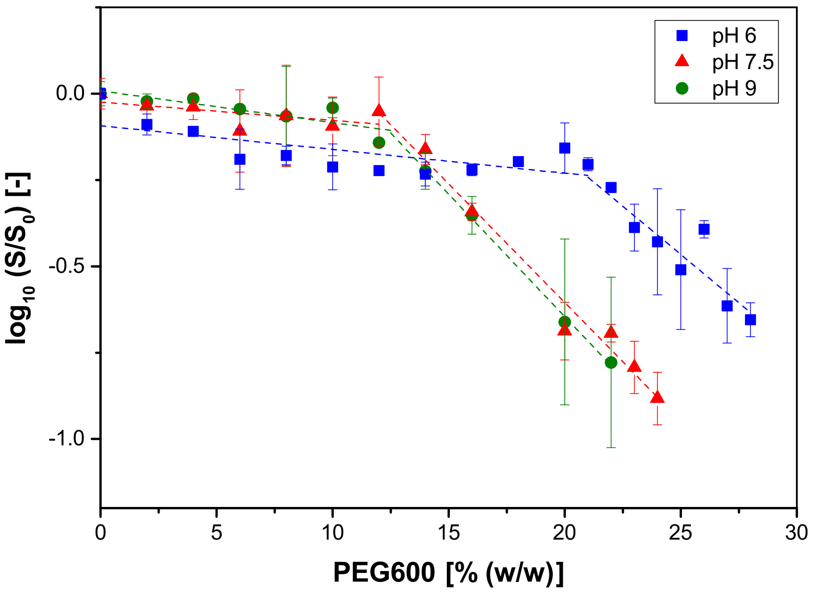


Figure 5: Precipitation curves of H1N1 with PEG600 as precipitant at pH 6 (blue), pH 7.5 (red), and pH 9 (green). For pH 4.5, immediate precipitation occurred by addition of PEG600 and no experimental data could be obtained.

For pH 4.5, precipitation occurred spontaneously after addition of PEG600 for all investigated PEG600 concentrations (no data could be shown due to instant precipitation). The curves for pH 7.5 and pH 9 follow a similar course with the discontinuity point m^* being determined at $12\%(w/w)$ PEG600 for pH 7.5 and $12.5\%(w/w)$ PEG600 for pH 9. At pH 6, a higher PEG600 concentration is necessary to initiate precipitation of virus particles. Under this condition, H1N1 solubility equals the concentration initially set at $21\%(w/w)$ PEG600. The slopes of the linear fits in the second segment of the curves exhibit similar values for all investigated pH values.

It has been demonstrated, that the precipitation of proteins and monoclonal antibodies by polyethylene glycol (PEG) can be applied as a method to predict their solubility [45, 46]. Because of the inert nature of PEG, protein precipitation occurs via an excluded volume effect [47]. Li et al. [46] demonstrated for equally concentrated solutions with different mAbs that a more soluble mAb requires higher PEG concentration to precipitate. Therefore, parameters such as the midpoint of PEG precipitation or the minimum % PEG needed for initiating protein precipitation can be used as indicator for the relative apparent solubility of the protein. This precipitation method has the potential to assess and compare relative protein solubilities for different solution conditions [48, 49, 50, 45, 46]. For the precipitation experiments of H1N1 conducted in this study, a clear correlation of colloidal stability obtained from the phase diagrams and stability against PEG600 was found. For pH 4.5, the lowest H1N1 recovery, thus the highest loss of virus particles through aggregation, was observed. For this pH value, precipitation occurred directly after addition of the lowest PEG600 concentration, resulting in a value for $m^* \leq 2\%(w/w)$ PEG600. For pH 7.5 and pH 9, comparable concentrations of $12\%(w/w)$ and $12.5\%(w/w)$ PEG600 were necessary to initiate precipitation of H1N1. The determined H1N1 recoveries in the phase diagrams for both these pH values were also in the same magnitude. For pH 6, significant higher recoveries attended by an increased colloidal stability were obtained. For this pH, a considerably higher PEG600 concentration ($m^* = 21\%(w/w)$ PEG600) was required to initiate precipitation of H1N1. Hence, the clear correlation between the precipitation experiments and the H1N1 concentration in the supernatant determined in the phase diagrams indicate the applicability of this method to predicting the long-term aggregation propensity of H1N1.

4 Conclusion

For the development of stable influenza vaccine formulations, the understanding of parameters influencing the aggregation propensity and the activity of viruses is crucial. In this work, the colloidal stability and remaining HA activity of H1N1 influenza virus particles was determined under different experimental conditions after 40 days. For the colloidal stability, the lowest recoveries were obtained for pH 4.5, at conditions close to the pI of H1N1. The highest H1N1 recoveries were obtained for pH 6. It was found that there is a significantly lower H1N1 recovery for NaCl concentrations below 100 mM and that recovery increases with increasing NaCl content. The recoveries were generally higher for lower initial H1N1 concentrations. The higher aggregation propensity for pH 9 was caused by a high intramolecular electrostatic repulsion of the surface proteins which leads to conformational changes and the exposure of hydrophobic amino acids to the surface of the proteins. The measurements of zeta potential, hydrophobicity, and FT-IR spectra underline these findings. For the conformational stability of H1N1, in contrast to the colloidal stability, the highest remaining HA activities were obtained for pH 9. At pH 4.5, the complete HA activity was lost due to conformational changes of hemagglutinin at acidic pH values. Results of FT-IR spectroscopy and determination hydrophobicity also revealed these changes in protein structure. The remaining HA activities were considerably lower for systems with a low initial H1N1 concentration.

The precipitation of H1N1 by PEG600 was proven to be a suitable, fast, and high-throughput-compatible method for the prediction of colloidal stability of H1N1 virus particles. This method shows the potential to replace time-consuming phase diagrams. The highest PEG600 concentration to initiate precipitation of H1N1 was neccessary for pH 6, where the lowest aggregation propensity was determined with the phase diagrams. Further results of the precipitaion experiments correlated well for the other pH values.

Combining all presented methods for investigation of the stability of H1N1 influenza viruses, a potential strategy to develop more stable vaccine formulations includes PEG precipitation experiments to define conditions where optimal values of overall H1N1 recovery are maintained. The high-throughput-compatible stalagmometric method for determination of the hydrophobicity of virus particles offers a reliable method to determine conformational changes of the surface proteins and the resulting loss in HA activities as a consequence. The combination of both methodologies depicts a powerful tool for the development of formulations with a preserved colloidal and conformational stability and thereby facilitates the rapid development of stable and safe vaccine formulations.

Acknowledgments

This research work is part of the project 'Optimization of an industrial process for the production of cell-culture-based seasonal and pandemic influenza vaccines' (grant number

0315640B) funded by the German Federal Ministry of Education and Research (BMBF). We thank Dr. Boris Hundt, Sabine Stefaniak, and Sebastian Grundke from IDT Biologika GmbH (Dessau-Roßlau, Germany) for providing the H1N1 feedstream and the kind support and Heidemarie Knierim from KIT Language Services for proofreading this manuscript. We gratefully acknowledge the support from Dr. Sven Amrhein for surface tension measurements.

Conflicts of Interests

The authors have declared no conflict of interest.

References

[1] World Health Organization, Influenza (Seasonal), Fact sheet No 211 (2014).
URL `http://www.who.int/mediacentre/factsheets/fs211/en/`

[2] G. Neumann, T. Noda, Y. Kawaoka, Emergence and pandemic potential of swine-origin H1N1 influenza virus, Nature 459 (7249) (2009) 931–939. doi:10.1038/nature08157.

[3] J. K. Taubenberger, J. C. Kash, Influenza Virus Evolution, Host Adaptation, and Pandemic Formation, Cell Host & Microbe 7 (6) (2010) 440–451. doi:10.1016/j.chom.2010.05.009.

[4] O. S. Kumru, S. B. Joshi, D. E. Smith, C. R. Middaugh, T. Prusik, D. B. Volkin, Vaccine instability in the cold chain: Mechanisms, analysis and formulation strategies, Biologicals 42 (5) (2014) 237–259. doi:10.1016/j.biologicals.2014.05.007.

[5] C. Anamur, G. Winter, J. Engert, Stability of collapse lyophilized influenza vaccine formulations, International Journal of Pharmaceutics 483 (1-2) (2015) 131–141. doi:10.1016/j.ijpharm.2015.01.053.

[6] J.-P. Amorij, A. Huckriede, J. Wilschut, H. W. Frijlink, W. L. J. Hinrichs, Development of Stable Influenza Vaccine Powder Formulations: Challenges and Possibilities, Pharmaceutical Research 25 (6) (2008) 1256–1273. doi:10.1007/s11095-008-9559-6.

[7] R. J. Garmise, K. Mar, T. M. Crowder, C. R. Hwang, M. Ferriter, J. Huang, J. a. Mikszta, V. J. Sullivan, A. J. Hickey, Formulation of a dry powder influenza vaccine for nasal delivery, AAPS PharmSciTech 7 (1) (2006) E131–E137. doi:10.1208/pt070119.

[8] N. K. Jain, N. Sahni, O. S. Kumru, S. B. Joshi, D. B. Volkin, C. Russell Middaugh, Formulation and stabilization of recombinant protein based virus-like particle vaccines, Advanced Drug Delivery Reviews 93 (2015) 42–55. doi:10.1016/j.addr.2014.10.023.

[9] D. T. Brandau, L. S. Jones, C. M. Wiethoff, J. Rexroad, C. Middaugh, Thermal Stability of Vaccines, Journal of Pharmaceutical Sciences 92 (2) (2003) 218–231. doi:10.1002/jps.10296.

[10] A. Wasserman, R. Sarpal, B. R. Phillips, E. P Wen, R. Ellis, N. S Pujar, Lyophilization in Vaccine Processes, Vaccine Development and Manufacturing (2014) 263–285.

[11] World Health Organization, Controlled Temperature Chain (CTC) (2016).
URL `http://www.who.int/biologicals/areas/vaccines/controlledtemperaturechain/en/`

[12] H.-C. Mahler, W. Friess, U. Grauschopf, S. Kiese, Protein aggregation: Pathways, induction factors and analysis, Journal of Pharmaceutical Sciences 98 (9) (2009) 2909–2934. arXiv:z0024, doi:10.1002/jps.21566.

[13] W. Wang, S. Nema, D. Teagarden, Protein aggregationPathways and influencing factors, International Journal of Pharmaceutics 390 (2) (2010) 89–99. doi:10.1016/j.ijpharm.2010.02.025.

[14] M. E. Young, P. A. Carroad, R. L. Bell, Estimation of diffusion coefficients of proteins, Biotechnology and Bioengineering 22 (5) (1980) 947–955. doi:10.1002/bit.260220504.

[15] E. Y. Chi, S. Krishnan, T. W. Randolph, J. F. Carpenter, Physical Stability of Proteins in Aqueous Solution: Mechanism and Driving Forces in Nonnative Protein Aggregation, Pharmaceutical Research 20 (9) (2003) 1325–1336. doi:10.1023/A:1025771421906.

[16] S. Setia, H. Mainzer, M. L. Washington, G. Coil, R. Snyder, B. G. Weniger, Frequency and causes of vaccine wastage, Vaccine 20 (7-8) (2002) 1148–1156. doi:10.1016/S0264-410X(01)00433-9.

[17] C. Nelson, P. Froes, A. M. V. Dyck, J. Chavarría, E. Boda, A. Coca, G. Crespo, H. Lima, Monitoring temperatures in the vaccine cold chain in Bolivia, Vaccine 25 (3) (2007) 433–437. doi:10.1016/j.vaccine.2006.08.017.

[18] S. Techathawat, P. Varinsathien, A. Rasdjarmrearnsook, P. Tharmaphornpilas, Exposure to heat and freezing in the vaccine cold chain in Thailand, Vaccine 25 (7) (2007) 1328–1333. doi:10.1016/j.vaccine.2006.09.092.

[19] X. Chen, G. J. Fernando, M. L. Crichton, C. Flaim, S. R. Yukiko, E. J. Fairmaid, H. J. Corbett, C. A. Primiero, A. B. Ansaldo, I. H. Frazer, L. E. Brown, M. A. Kendall, Improving the reach of vaccines to low-resource regions, with a needle-free vaccine delivery device and long-term thermostabilization, Journal of Controlled Release 152 (3) (2011) 349–355. doi:10.1016/j.jconrel.2011.02.026.

[20] T. S. Priddy, C. R. Middaugh, E. P Wen, R. Ellis, N. S Pujar, Stabilization and Formulation of Vaccines, Vaccine Development and Manufacturing (2014) 237–261.

[21] A. D. Bokser, P. B. O'Donnell, Remington: The science and practice of pharmacy, in: L. A. Felton (Ed.), Remington Essentials of pharmaceutics, 21st Edition, Pharmaceutical Press, London, 2006, Ch. 4, pp. 37–49.

[22] S. Kapoor, K. Dhama, Insight into Influenza Viruses of Animals and Humans, Springer International Publishing, Cham, 2014. doi:10.1007/978-3-319-05512-1.

[23] C. J. Burke, T.-A. Hsu, D. B. Volkin, Formulation, stability, and delivery of live attenuated vaccines for human use, Critical Reviews in Therapeutic Drug Carrier Systems 16 (1).

[24] G. L. Miller, Influence of pH and of certain other conditons on the stability of the infectivity and red cell agglutinating activity fo influenza virus, Journal of Experimental Medicine 80 (6) (1944) 507–520.

[25] E. A. Permyakov, L. J. Berliner, α-Lactalbumin: structure and function, FEBS Letters 473 (3) (2000) 269–274. doi:10.1016/S0014-5793(00)01546-5.

[26] A. L. Fink, L. J. Calciano, Y. Goto, T. Kurotsu, D. R. Palleros, Classification of Acid Denaturation of Proteins: Intermediates and Unfolded States, Biochemistry 33 (41) (1994) 12504–12511. doi:10.1021/bi00207a018.

[27] K. Collins, Ions from the Hofmeister series and osmolytes: effects on proteins in solution and in the crystallization process, Methods 34 (3) (2004) 300–311. doi:10.1016/j.ymeth.2004.03.021.

[28] R. A. Curtis, J. M. Prausnitz, H. W. Blanch, Protein-protein and protein-salt interactions in aqueous protein solutions containing concentrated electrolytes, Biotechnology and Bioengineering 57 (1) (1998) 11–21. doi:10.1002/(SICI)1097-0290(19980105)57:1¡11::AID-BIT2¿3.0.CO;2-Y.

[29] T. Arakawa, S. N. Timasheff, Mechanism of protein salting in and salting out by divalent cation salts: balance between hydration and salt binding, Biochemistry 23 (25) (1984) 5912–5923. doi:10.1021/bi00320a004.

[30] F. Hofmeister, Zur Lehre von der Wirkung der Salze, Archiv für Experimentelle Pathologie und Pharmakologie 25 (1) (1888) 1–30. doi:10.1007/BF01838161.

[31] M. L. Killian, Hemagglutination Assay for the Avian Influenza Virus, in: Avian Influenza Virus, Humana Press, Totowa, NJ, 2008, pp. 47–52.

[32] R. Webster, N. Cox, K. Stohr, WHO Manual on Animal Influenza Diagnosis and Surveillance, WHO/CDS/CDR/2002.5 Rev. 1.

[33] B. Kalbfuss, A. Knöchlein, T. Kröber, U. Reichl, Monitoring influenza virus content in vaccine production: Precise assays for the quantitation of hemagglutination and neuraminidase activity, Biologicals 36 (3) (2008) 145–161. doi:10.1016/j.biologicals.2007.10.002.

[34] K. Baumgartner, L. Galm, J. Nötzold, H. Sigloch, J. Morgenstern, K. Schleining, S. Suhm, S. A. Oelmeier, J. Hubbuch, Determination of protein phase diagrams by microbatch experiments: Exploring the influence of precipitants and pH, International Journal of Pharmaceutics 479 (1) (2015) 28–40. doi:10.1016/j.ijpharm.2014.12.027.

[35] S. Amrhein, S. Suhm, J. Hubbuch, Surface tension determination by means of liquid handling stations, Engineering in Life Sciences 16 (6) (2016) 532–537. doi:10.1002/elsc.201500179.

[36] S. Amrhein, K. C. Bauer, L. Galm, J. Hubbuch, Non-invasive high throughput approach for protein hydrophobicity determination based on surface tension, Biotechnology and Bioengineering 112 (12) (2015) 2485–2494. doi:10.1002/bit.25677.

[37] T. Korte, K. Ludwig, A. Herrmann, ph-Dependent hydrophobicity profile of hemagglutinin of influenza virus and its possible relevance in virus fusion, Bioscience Reports 12 (5) (1992) 397–406. doi:10.1007/BF01121503.

[38] R. K. Scopes, Separation by Precipitation, Springer New York, New York, NY, 1994, pp. 71–101. doi:10.1007/978-1-4757-2333-5_4.
URL `http://link.springer.com/10.1007/978-1-4757-2333-5_4`

[39] K. A. Dill, Dominant forces in protein folding, Biochemistry 29 (31) (1990) 7133–7155. arXiv:arXiv:1011.1669v3, doi:10.1021/bi00483a001.

[40] D. Guo, C. T. Mant, A. K. Taneja, J. Parker, R. S. Rodges, Prediction of peptide retention times in reversed-phase high-performance liquid chromatography I. Determination of retention coefficients of amino acid residues of model synthetic peptides, Journal of Chromatography A 359 (C) (1986) 499–518. arXiv:1011.1669, doi:10.1016/0021-9673(86)80102-9.

[41] F. M. Hughson, Structural characterization of viral fusion proteins, Current Biology 5 (3) (1995) 265–274. doi:10.1016/S0960-9822(95)00057-1.

[42] J. Vajda, D. Weber, S. Stefaniak, B. Hundt, T. Rathfelder, E. Müller, Mono- and polyprotic buffer systems in anion exchange chromatography of influenza virus particles, Journal of Chromatography A 1448 (2016) 73–80. doi:10.1016/j.chroma.2016.04.047.

[43] H.-J. Choi, C. F. Ebersbacher, M.-C. Kim, S.-M. Kang, C. D. Montemagno, A Mechanistic Study on the Destabilization of Whole Inactivated Influenza Virus Vaccine in Gastric Environment, PLoS ONE 8 (6) (2013) e66316. doi:10.1371/journal.pone.0066316.

[44] E. Goormaghtigh, V. Cabiaux, J.-M. Ruysschaert, Determination of Soluble and Membrane Protein Structure by Fourier Transform Infrared Spectroscopy, Springer US, Boston, MA, 1994, pp. 405–450. doi:10.1007/978-1-4615-1863-1_10.

[45] T. J. Gibson, K. Mccarty, I. J. McFadyen, E. Cash, P. Dalmonte, K. D. Hinds, A. A. Dinerman, J. C. Alvarez, D. B. Volkin, Application of a High-Throughput Screening Procedure with PEG-Induced Precipitation to Compare Relative Protein Solubility During Formulation Development with IgG1 Monoclonal Antibodies, Journal of Pharmaceutical Sciences 100 (3) (2011) 1009–1021. doi:10.1002/jps.22350.

[46] L. Li, A. Kantor, N. Warne, Application of a PEG precipitation method for solubility screening: A tool for developing high protein concentration formulations, Protein Science 22 (8) (2013) 1118–1123. doi:10.1002/pro.2289.

[47] D. H. Atha, K. C. Ingham, Mechanism of precipitation of proteins by polyethylene glycols. Analysis in terms of excluded volume., The Journal of Biological Chemistry 256 (23) (1981) 12108–17.

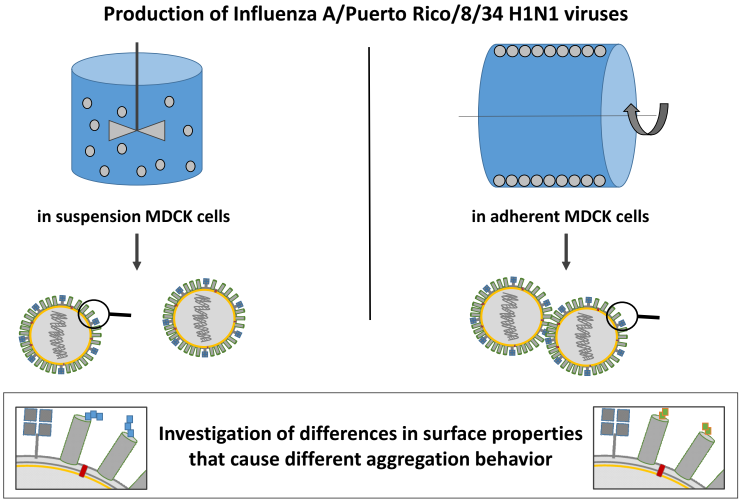
[48] I. Juckles, Fractionation of proteins and viruses with polyethylene glycol, Biochimica et Biophysica Acta (BBA) - Protein Structure 229 (3) (1971) 535–546. doi:10.1016/0005-2795(71)90269-8.

[49] C. R. Middaugh, W. A. Tisel, R. N. Haire, A. Rosenberg, Determination of the apparent thermodynamic activities of saturated protein solutions, Journal of Biological Chemistry 254 (2) (1979) 367–370.

[50] M. Bončina, J. Reščič, V. Vlachy, Solubility of Lysozyme in Polyethylene Glycol-Electrolyte Mixtures: The Depletion Interaction and Ion-Specific Effects, Biophysical Journal 95 (3) (2008) 1285–1294. doi:10.1529/biophysj.108.128694.

Influence of the Production System on the Surface Properties of Influenza A Virus Particles

Frank Hämmerling[1‡], Michael M. Pieler[2‡], René Hennig[2,3], Anja Serve[2], Erdmann Rapp[2,3], Michael W. Wolff[2,4], Udo Reichl[2,5] and Jürgen Hubbuch[1*]


Production of Influenza A/Puerto Rico/8/34 H1N1 viruses
in suspension MDCK cells
in adherent MDCK cells
Investigation of differences in surface properties that cause different aggregation behavior

[1] : *Karlsruhe Institute of Technology, Institute of Engineering in Life Sciences, Section IV: Biomolecular Separation Engineering, Fritz-Haber-Weg 2, 76131 Karlsruhe, Germany*
[2] : *Max Planck Institute for Dynamics of Complex Technical Systems, Bioprocess Engineering, Sandtorstraße 1, 39106 Magdeburg, Germany*
[3] : *glyXera GmbH, Leipziger Straße 44, 39120 Magdeburg, Germany*
[4] : *University of Applied Sciences Mittelhessen, Institute of Bioprocess Engineering and Pharmaceutical Technology, Wiesenstraße 14, 35390 Gießen, Germany*
[5] : *Otto von Guericke University Magdeburg, Chair of Bioprocess Engineering, Universitätsplatz 2, 39106 Magdeburg, Germany*

‡ : *These authors contributed equally to this work.*
* : *Corresponding author; email address: juergen.hubbuch@kit.edu*

Manuscript submitted to Engineering in Life Sciences

Abstract

In this study, influenza A/Puerto Rico/8/34 H1N1 virus particles (VP) produced in adherent and suspension Madin Darby canine kidney cells were investigated with a broad analytical toolbox to obtain more information on the VP surface properties potentially affecting their aggregation behavior. First, differences in aggregation behavior were revealed by VP size distributions obtained via differential centrifugal sedimentation confirmed by dynamic light scattering. The VP produced in adherent cells showed increased levels of aggregation in 20 mM NaCl 10 mM Tris-HCl pH 7.4 buffer. This included the formation of multimers (dimers up to pentamers), whereas VP produced in suspension cells displayed no tendency towards aggregate formation. To investigate the cause of these differences in aggregation behavior, the VP samples were compared based on their zeta potential, their surface hydrophobicity, their lipid composition, and the N-glycosylation of their major VP surface protein hemagglutinin. The zeta potential and the hydrophobicity of the VP produced in the adherent cells was significantly decreased compared to the VP produced in the suspension cells. The lipid composition of both VP systems was approximately identical. The hemagglutinin of the VP produced in adherent cells included more of the larger N-glycans, whereas the VP produced in suspension cells included more of the smaller N-glycans. These results indicate that differences in the glycosylation of viral surface proteins should be monitored to characterize VP hydrophobicity and aggregation behavior, and to avoid aggregate formation and product losses in virus purification processes for vaccines and gene therapy.

Keywords: Influenza, Virus Particles, Aggregation, Glycosylation, Lipidomics

1 Introduction

Aggregation of virus particles (VP) affects a wide range of different processes ranging from VP quantification [1, 2] and inactivation [3, 4, 5] to downstream processing of virus-based biopharmaceuticals [6]. However, the underlying causes of VP aggregation are in general not known due to the elusive architecture of the VP, which can consist of various proteins, carbohydrates, lipids, and nucleic acids with diverse physicochemical properties. Besides the buffer conditions, e.g., pH value, ionic strength, osmolarity, and the presence of excipients [2, 3], the structural characteristics of VP surface proteins, e.g., mutations or glycosylation strongly influence the aggregation propensity [7]. It has already been shown that the N-glycosylation of the influenza A virus surface protein hemagglutinin (HA) depends strongly on the host cell line [8, 9, 10, 11], which might change the physicochemical properties of the whole VP. In a previous study [12] we investigated ion effects on particle size distributions and aggregation of influenza A virus (A/Puerto Rico/8/34 H1N1, A/PR) VP produced in adherent and suspension Madin Darby canine kidney (MDCK) cells. However, the experimental setup used did not allow for a more comprehensive investigation of the underlying causes for the differences in aggregation behavior. In a next step, we established an analytical toolbox to further characterize differences in VP systems and to increase our knowledge on VP aggregation. As an application, we investigate both of the A/PR samples from animal cell culture. First, the particle size distributions (PSD) of the VP were measured by differential centrifugation sedimentation (DCS) in a buffer known to induce aggregation, i.e., 20 mM NaCl 10 mM Tris-HCl pH 7.4, and a buffer not affecting the aggregation status, i.e., 60 mM NaCl 10 mM Tris-HCl pH 7.4 [12]. Then, the obtained PSD were confirmed by dynamic light scattering (DLS). In a second step, surface characteristics of both VP samples were evaluated to establish a link between the VP aggregation behavior and specific surface features. For that, the zeta potential was obtained by measuring the electrophoretic mobility of the VP, and the VP surface hydrophobicity by stalagmometry [13, 14]. In addition, the N-glycosylation of the major A/PR surface protein hemagglutinin (HA) was analyzed by multiplexed capillary gel electrophoresis with laser induced fluorescence detection (xCGE-LIF) [15]. The HA is a highly glycosylated and antigenic viral membrane protein, which is protruding as a trimeric spike from the VP surface (see Figure 1). As the HA represents approximately 35% of the total VP protein content [16] and the ratio of HA to the two other membrane proteins neuraminidase (NA) and M2 are approximately 4:1 and 10:1 - 100:1, respectively [17], it potentially plays the key role in the aggregation behavior. Furthermore, the lipid compositions of the VP were investigated by mass spectrometry (MS). Finally, the results were discussed and put in context of previously published data on proteins, synthetic particles, and VP. Ultimately, this work aims to cast light on differences in viral surface properties and, hence, causes of virus aggregation. Therefore, it should represent a starting point for future studies to uncover the diverse nature of VP systems relevant in vaccine manufacturing and gene therapy.

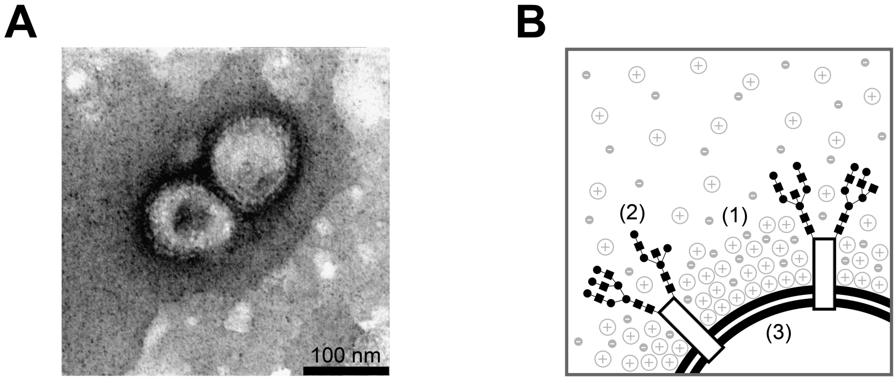


Figure 1: (A): Transmission electron microscopy (TEM) pictures of negatively stained enveloped influenza A/Puerto Rico/8/34 H1N1 virus particles (VP). (B): Schematic representation of an influenza A VP including the investigated features in the scope of this work: charge distribution on the VP responsible for the zeta potential (1), the *N*-glycosylation of the VP membrane-bound hemagglutinin (2) and the lipid-bilayer membrane of the enveloped VP (3).

2 Materials and Methods

2.1 Hemagglutination Activity Analytics

The hemagglutination assay was based on Kalbfuss et al. and the obtained hemagglutination activities are reported in HA activity units per mL (HAU/mL, n = 1) [18].

2.2 Influenza A virus particle production and sample preparation

Influenza A/Puerto Rico/8/34 H1N1 VP were produced in suspension MDCK (MDCK$_{SUS2}$) and adherent MDCK cell lines (MDCK$_{ADH}$, ECACC No. 84121903) according to Pieler et al. [12]. Briefly, MDCK$_{SUS2}$ cells, termed A/PR$_{SUS}$ thereafter, were produced in a 5 L bioreactor and VP derived from MDCK$_{ADH}$ cells, termed A/PR$_{ADH}$ thereafter, in 850 cm^2 roller bottles, both with a multiplicity of infection (MOI) of 10^{-4}.
For DCS, DLS, stalagmometry, and xCGE-LIF the harvested VP were clarified by depth filtration, chemically inactivated by β-propiolactone, and concentrated by tangential flow filtration (TFF). Retentates were 0.1 μm filtered to remove unwanted aggregates and particulate impurities, 0.5 mL aliquoted, frozen at -80 °C, and thawed to obtain monodisperse VP samples with approximately 10^5 HAU/mL [19, 20, 21].
The VP samples for lipidomics were produced with the same virus stock and cell lines mentioned before with slightly different infection conditions (MOI of 0.025, $2{\cdot}10^{-6}$ units trypsin per cell (# 27250-018, Gibco$^{\text{TM}}$, Thermo Fisher Scientific, Waltham, USA)). VP production with MDCK$_{SUS}$ was carried out in 250 mL baffled flasks. The harvested VP broths were centrifuged at 4000 g for 35 min to remove cell and cell debris, followed by a 10000 g centrifugation step for 45 min to remove smaller cell fragments and compartments with an Avanti J-20XP centrifuge (rotor JA-14, Beckman Coulter, Brea, USA). At each step, the supernatant was transferred into a new centrifuge tube and the pellet

was discarded. The resulting supernatant was overlayed with a 20% sucrose solution before the VP sample was concentrated by ultracentrifugation at $25{\cdot}10^3$ revolutions per minute (RPM) for 2 h at 4 °C (Optima$^{\text{TM}}$ LE-80 K, rotor SW-28, Beckman Coulter, Brea, USA). The resulting VP pellet was resuspended, loaded on a 30 - 60% sucrose gradient and centrifuged again at $25{\cdot}10^3$ RPM for 3 h at 4 °C. The banded VP were collected, diluted approximately 1:30 and pelleted again by centrifugation at $25{\cdot}10^3$ RPM for 1.5 h at 4 °C. Finally, the pelleted VP were resuspended in 150 mM ammonium bicarbonate buffer, heat inactivated (80 °C for 3 min) and stored at -80 °C.

2.3 Virus Particle Sample Dialysis

For the DCS, DLS, and stalagmometry analytics the prepared VP samples were dialyzed to 20 mM NaCl 10 mM Tris-HCl pH 7.4 buffer and 60 mM NaCl 10 mM Tris-HCl pH 7.4 buffer. For that, 500 μL TFF-concentrated VP sample were transferred into a 14 kDa molecular weight cut-off cellulose dialysis membrane (# 0653.1, Carl Roth GmbH & Co. KG, Karlsruhe, Germany) and dialyzed under stirring overnight at room temperature against 500 mL buffer, containing 0.05% NaN_3 to avoid microbial growth.

2.4 Virus Particle Size Distribution Measurement by Differential Centrifugal Sedimentation

The DCS measurements were carried out according to Pieler et al. [12]. Briefly, a CPS DC24000 UHR disc centrifuge was used at 24000 RPM with a 4 to 16% (w/v) sucrose gradient in the respective buffer. For the PSD measurements, 100 μL of the dialyzed VP samples were injected. The PSD are visualized as normalized weight in % over the apparent hydrodynamic diameter in nm (n = 1).

2.5 Z-average Value Measurement of the Virus Particles by Dynamic Light Scattering

The z-average value that is derived from the intensity weighted mean hydrodynamic diameter of the VP samples was determined by DLS using a Zetasizer Nano ZSP (Malvern Instruments Ltd, Malvern, UK). The z-average value in nm was compared to the DCS data to confirm the aggregation status of the VP samples. For the measurements, 45 μL of the dialyzed VP samples were dispensed in a ZEN2112 quartz cuvette (Hellma® GmbH & Co. KG, Müllheim, Germany) and measured with the Non-Invasive Back Scatter optics (NIBS®) at 25 °C (n = 3).

2.6 Zeta Potential Measurement of the Virus Particles

The zeta potential of the VP was determined by electrophoretic mobility measurements using a Zetasizer Nano ZSP (Malvern Instruments Ltd, Malvern, UK) with a voltage of 25 V. Therefore, 500 μL VP sample dialyzed to 60 mM NaCl 10 mM Tris-HCl pH 7.4 were measured (n = 3). Data acquisition and evaluation was performed with the Zetasizer software version 7.11 (Malvern Instruments Ltd, Malvern, UK). The zeta potentials are reported in mV.

2.7 Hydrophobicity Measurements of the Virus Particles

The hydrophobicity of the VP was determined by stalagmometry previously described by Amrhein et al. [13, 14]. For the measurements, both VP samples were dialyzed to 20 mM NaCl 10 mM Tris-HCl pH 7.4 and 60 mM NaCl 10 mM Tris-HCl pH 7.4, respectively. Then, the dialyzed VP samples were diluted with fresh dialysis buffer to ten different concentrations to decrease the HA activity by 10% in every dilution step. After the dilution, the VP samples were purged with very slow flow-rate of 5 μL/s through a vertical capillary from the top to the bottom with an automated liquid handling system (Tecan Freedom EVO 100, Tecan GmbH, Crailsheim, Germany). Due to the applied flow, a drop will grow up to a specific maximum volume at the capillary end, which depends on the surface tension of the suspended VP and the suspension buffer itself. Exceeding the maximum volume, the drop detaches from the capillary end and falls onto the analytical balance WXTS205DU (Mettler Toledo, Greifensee, Switzerland) located below. The mass of the drop is compared to the mass of a drop of the pure reference solution, e.g. ultrapure water or buffer, and thereby the surface tension can be calculated. The obtained surface tension correlates with the hydrophobicity, as more hydrophobic molecules tend to attach to the air-water interface and thereby decrease the surface tension and the drop size.

2.8 Hemagglutinin *N*-Glycan Analysis of Influenza A Virus Particles

Influenza A virus HA *N*-glycan analysis was performed as described by Hennig et al. [15]. Briefly, virus proteins of purified VP with a HA activity of 3500 HAU/mL were separated by one-dimensional sodium dodecylsulfate polyacrylamide gel electrophoresis (1D-SDS-PAGE). Protein bands of the HA_0 monomer were excised from the SDS-PAGE and *N*-glycans were released from the protein backbone by in-gel deglycosylation using PNGase F [15]. Released *N*-glycans were extracted from gel bands, labeled with the fluorescent dye aminopyrene trisulfonic acid (APTS) and analyzed by xCGE-LIF. Thereby, APTS labeled *N*-glycans were separated inside a polymer filled capillary by their charge, size, and shape, where small *N*-glycans migrate faster through the capillary than the larger *N*-glycans [22]. After laser assisted excitation of the APTS-labeled *N*-glycans, the emitted signal was recorded in an electropherogram. By normalizing the migration time of the labeled *N*-glycans to an internal size standard, a so-called *N*-glycan fingerprint with normalized migration time units (MTU") was generated, creating a highly reproducible measurement. The signal intensity of the *N*-glycan fingerprints was normalized by dividing the absolute intensity by the sum of the intensity of all *N*-glycan fingerprint peaks.

2.9 Lipid analysis of virus particles

-Internal standard lipid mixture

The internal standard lipid mixture contained 20 pmol diacylglycerol (DAG) 17:0−17:0, 24 pmol phosphatidic acid (PA) 17:0−17:0, 52 pmol phosphatidylethanolamine (PE) 17:0−17:0, 7.5 pmol phosphatidylglycerol (PG) 17:0−17:0, 43 pmol phosphatidylserine (PS) 17:0−17:0, 40 pmol phosphatidylcholine (PC) 18:3−18:3, 54 pmol phosphatidylinositol (PI) 17:0−17:0, 10 pmol ceramide (Cer) 18:0;3/18:0, 40 pmol sphingomyelin (SM)

18:1;2/17:0, 53 pmol monosialodihexosylganglioside (GM3) bovine mixture, 66 pmol forssman glycolipid (For) sheep extract, 20 pmol galactosylceramide (GalCer) 18:1;2/12:0, 20 pmol lactosylceramide (LacCer) 18:1;2/12:0, and 50 pmol Cholesterol (Chol)-d7.

-Lipidextraction, virus particle sample preparation, and mass spectrometry based analysis

Samples were spiked with 10 μL of the internal standard lipid mixture and dissolved in 200 μL 150 mM ammonium bicarbonate buffer. After extraction with 1 mL of 10:1 (v:v) chloroform:methanol for 2 h, the lower organic phase was collected, and the aqueous phase was re-extracted with 1 mL of 2:1 (v:v) chloroform:methanol for 1 h. The lower organic phase was collected and the organic solvent was evaporated in a vacuum desiccator at 4 °C. Shotgun lipidomic analysis by MS was carried out according to Herzog et al. [23]. Briefly, lipid extracts were dissolved in 100 μL of 1:2 (v:v) chloroform:methanol. For MS analysis of PC, ether linked PC (PC O-), SM, DAG, cholesterol ester (CE), triacylglycerol (TAG), and the glycolipid For, 10 μL of the lipid extract was mixed with 13 μL 13 mM ammonium acetate in propanol and subjected to shotgun lipidomics analysis by MS in a LTQ-Orbitrap (Thermo Fisher Scientific, positive-ion mode, Fourier transform MS with $R_{m/z=400}$ = 100000) equipped with a robotic nanoflow ion source TriVersa NanoMate (Advion Biosciences, Ithaca, NY, USA). For MS analysis of PA, PS, PE, ether linked PE (PE O-), PI, Cer, hexosylceramide (HexCer), GM3, sulfatide (Sulf) and PG, 10 μL of the lipid extract was mixed with 10 μL 0.1% methylamine before being measured (negative-ion mode, Fourier transform MS with $R_{m/z=400}$ = 100000). Cholesterol was quantified after chemical acetylation as described elsewhere [24]. Automated processing of acquired mass spectra and identification and quantification of detected molecular lipid species were performed with the Lipid Profiler software (MDS Sciex) [25] and the LipidXplorer software (in-house software, Max Planck Institute of Molecular Cell Biology and Genetics (MPI-CBG)) [26].

3 Results and Discussion

3.1 Virus Particle Aggregation Behavior

The aggregation status of A/PR$_{SUS}$ and A/PR$_{ADH}$ VP after dialysis was measured by DCS to obtain particle size distributions. A/PR$_{SUS}$ showed no aggregation after being dialyzed to 20 mM NaCl 10 mM Tris-HCl pH 7.4 whereas A/PR$_{ADH}$ was highly aggregated (see Figure 2). Furthermore, A/PR$_{SUS}$ and A/PR$_{ADH}$ showed no aggregation after being dialyzed to 60 mM NaCl 10 mM Tris-HCl pH 7.4 (see Figure 2). This aggregation behavior was confirmed by the z-average values obtained by DLS shown in Table reftTable1. There, A/PR$_{ADH}$ dialyzed to 20 mM NaCl 10 mM Tris-HCl pH 7.4 shows an increased z-average value, indicating aggregation, whereas A/PR$_{ADH}$ in 60 mM NaCl 10 mM Tris-HCl pH 7.4 and A/PR$_{SUS}$ in both buffers show similar z-average values, indicating no aggregation. The observed aggregation differences were further investigated by zeta potential measurements, stalagmometry, xCGE-LIF-based glycoanalytics, and lipid analytics, to cast light on the underlying factors as addressed in the next sections.

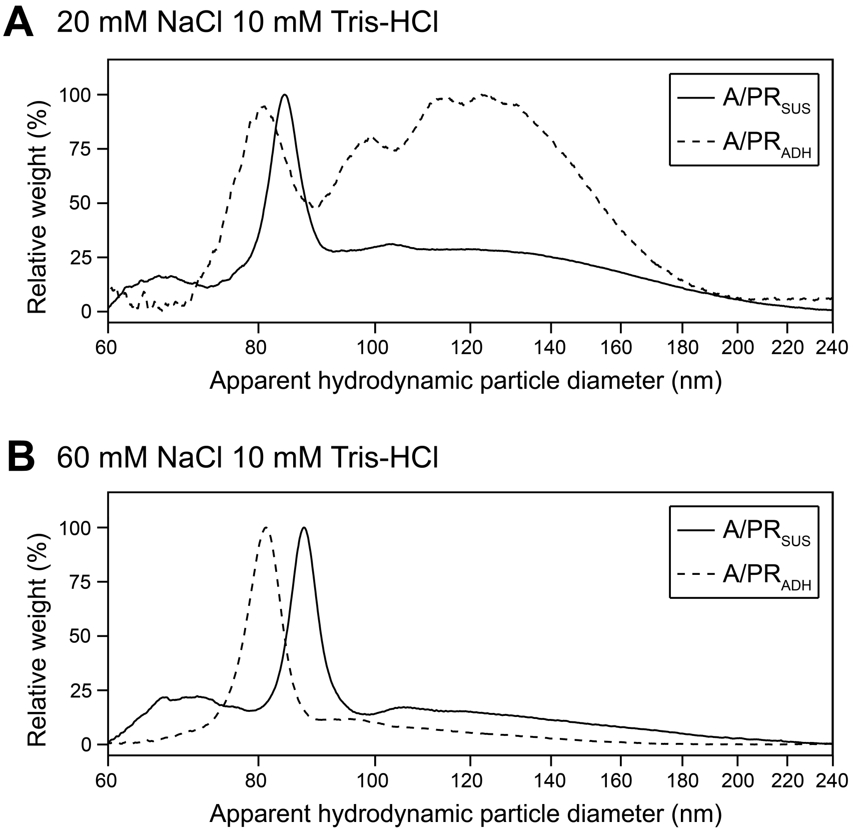


Figure 2: Particle size distributions of suspension (A/PR$_{SUS}$) and adherent cell culture-derived influenza A virus particles (A/PR$_{ADH}$) dialyzed against (A) 20 mM NaCl 10 mM Tris-HCl pH 7.4 and (B) 60 mM NaCl 10 mM Tris-HCl pH 7.4 (n = 1). A/PR$_{ADH}$ shows in the 20 mM NaCl 10 mM Tris-HCl buffer aggregation in the size range of 90 to 200 nm, whereas A/PR$_{SUS}$ shows no aggregation. Both samples show no aggregation in the 60 mM NaCl 10 mM Tris-HCl buffer.

Table 1: Z-average diameter of suspension (A/PR$_{SUS}$) and adherent cell culture-derived influenza A virus particles (A/PR$_{ADH}$) dialyzed against 20 mM NaCl 10 mM Tris-HCl pH 7.4 and 60 mM NaCl 10 mM Tris-HCl pH 7.4. Z-average diameter is shown as mean± STD (n = 3).

	z-average diameter in 20 mM NaCl 10 mM Tris-HCl pH 7.4 [nm]	**z-average diameter in 60 mM NaCl 10 mM Tris-HCl pH 7.4** [nm]
A/PR$_{SUS}$	148±2	142±0
A/PR$_{ADH}$	229±9	125±1

3.2 Virus Particle Surface Characteristics

-Zeta potential

With -8.6±0.7 mV (mean±STD), the zeta potential of A/PR$_{SUS}$S was significantly different from A/PR$_{ADH}$ (-11.2±0.6 mV; p= 0.003, paired t-test) in 60 mM NaCl 10 mM Tris-HCl pH 7.4. The lower zeta potential of A/PR$_{ADH}$ could potentially trigger higher electrostatic repulsive forces between suspended VP compared to the A/PR$_{SUS}$ VP. These higher repulsive interactions could prevent aggregation of A/PR$_{ADH}$ VP [27], which was not the case. We refrained from measuring the zeta potential in 20 mM NaCl 10 mM Tris-HCl pH 7.4 due known aggregation under this condition.

-Hydrophobicity

Hydrophobic surface characteristics are, besides electrostatic interactions, one of the crucial parameters influencing the aggregation behavior of biomolecules [28]. Figure 3 displays the determined surface tensions of both VP samples in dependency of the HA activity. All surface tensions were normalized based on the respective pure buffer.

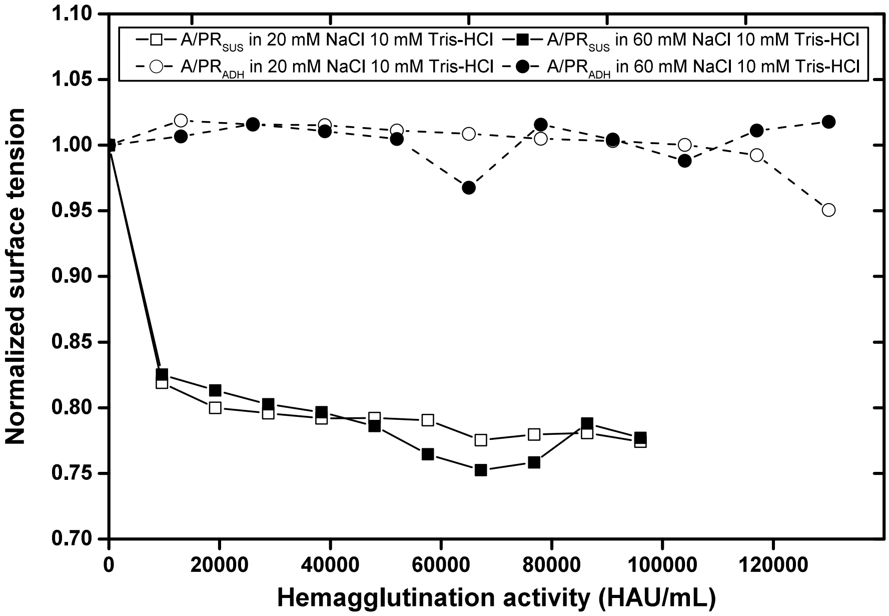


Figure 3: Particle size distributions of suspension (A/PR$_{SUS}$) and adherent cell culture-derived influenza A virus particles (A/PR$_{ADH}$) dialyzed against (A) 20 mM NaCl 1 0 mM Tris-HCl pH 7.4 and (B) 60 mM NaCl 10 mM Tris-HCl pH 7.4 (n = 1). A/PR$_{ADH}$ shows in the 20 mM NaCl 10 mM Tris-HCl buffer aggregation in the size range of 90 to 200 nm, whereas A/PR$_{SUS}$ shows no aggregation. Both samples show no aggregation in the 60 mM NaCl 10 mM Tris-HCl buffer.

The A/PR$_{SUS}$ sample showed a surface tension decrease by 20%, starting from low HA activities at 1000 HAU/mL, whereas no significant decrease of the surface tension was observed for A/PR$_{ADH}$ sample over the entire HA activity range. Comparing the normalized surface tension profiles of both VP samples reveals significant differences in the hydrophobic character: a considerably higher surface hydrophobicity (lower surface tension) is determined for A/PR$_{SUS}$ compared to A/PR$_{ADH}$.

-Viral hemagglutinin *N*-glycan fingerprint

The *N*-glycan fingerprints of both samples are shown in Figure 4. Overall, the A/PR$_{SUS}$ and A/PR$_{ADH}$ samples show a similar *N*-glycan fingerprint but with different relative abundances of the individual *N*-glycan peaks. While the A/PR$_{SUS}$ *N*-glycan fingerprint shows higher signal intensities for peaks at lower MTU", the A/PR$_{ADH}$ *N*-glycan fingerprint shows higher signal intensities for peaks at higher MTU". Since small *N*-glycans are migrating faster through the capillary than the larger *N*-glycans, the HA of A/PR$_{SUS}$ VP has more of the smaller *N*-glycan structures, whereas the HA of A/PR$_{ADH}$ VP has more of the larger structures.

Considering that carbohydrates, respectively *N*-glycans, are highly hydrophilic molecules, this finding of the *N*-glycan analysis is consistent with the results of the surface hydrophobicity determinations of Figure 3. A/PR$_{ADH}$ VP contain larger *N*-glycan structures, which makes them more hydrophilic than A/PR$_{SUS}$ VP with smaller N-glycan structures. Accordingly, the average size of the N-glycans attached the HA seems to greatly affect the aggregation behavior. Interestingly, however, the hydrophobic A/PR$_{SUS}$ particles seem to have a lower tendency towards aggregation in the screened buffers (Figure 2A) than the hydrophilic A/PR$_{ADH}$ particles.

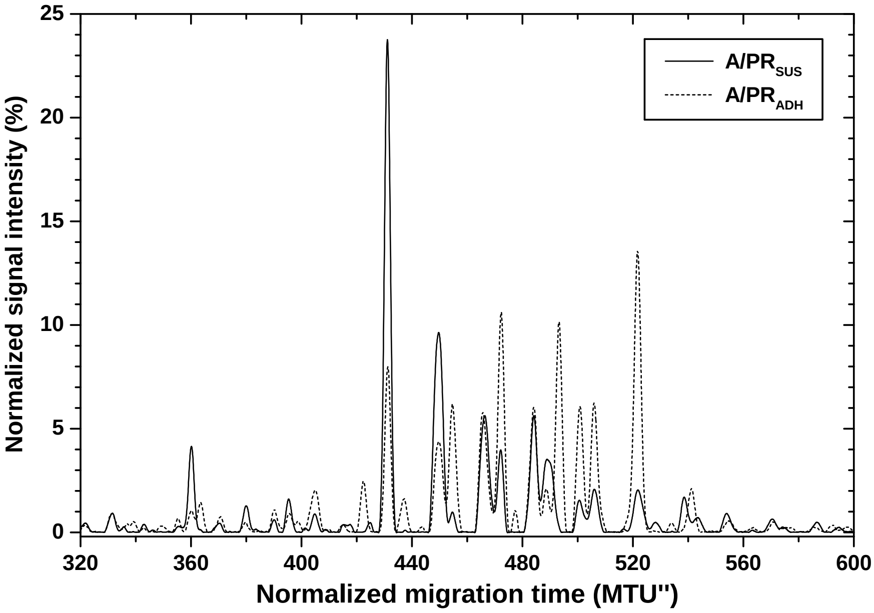


Figure 4: Overlay of the N-glycosylation fingerprints of the major surface protein hemagglutinin (HA) of suspension (A/PR$_{SUS}$) and adherent cell culture-derived influenza A virus particles (A/PR$_{ADH}$). Normalized total signal intensity in % is plotted over the normalized migration time in MTU".

This is not in agreement with previous work on colloidal particle systems, where particles with increased hydrophilicities appeared to be more stable [29]. It has to be taken into account, however, that both systems have a very low critical stabilization concentration (CSC) and a high number of stabilizing ionic species [12], which are characteristics of a highly hydrophilic particle system [29]. An explanation for the higher aggregation propensity of the A/PR$_{ADH}$ VP could be that the larger glycan structures on A/PR$_{ADH}$ shield stabilizing electrostatic interactions arising from the charged VP surface proteins or that glycan-glycan interactions promote VP aggregation. As the influenza N-glycans contain no charged carbohydrates, electrostatic interactions between the glycans can be excluded.

-Virus particle lipid composition

Analysis of the lipid composition of A/PR$_{ADH}$ and A/PR$_{SUS}$ VP revealed only few significant differences (Figure 5). High molar percentages without significant differences were found for the glycerophospholipids PS, which is negatively charged at pH 7.4, PEO, and PC. The major sphingolipid in both VP samples is SM in slightly different quantities, followed by For and GM3 with similar quantities (both <2 mol%). The amount of cholesterol in the VP is, compared to the amount in the host cell membrane (data not shown), more than two times enriched and comprises up to 41% of the total lipids in A/PR$_{ADH}$ and A/PR$_{SUS}$.

Overall, due to these small differences in the VP lipid composition, the VP lipid bilayer membrane is very likely not a major contributor to the observed aggregation behavior differences of A/PR$_{ADH}$ and A/PR$_{SUS}$ VP. Furthermore, it can be assumed that (glycosylated) influenza A membrane proteins, e.g., HA and NA, shield the lipid bilayer membrane from the environment and, therefore, the differences in lipid composition have no great influence on the aggregation behavior.

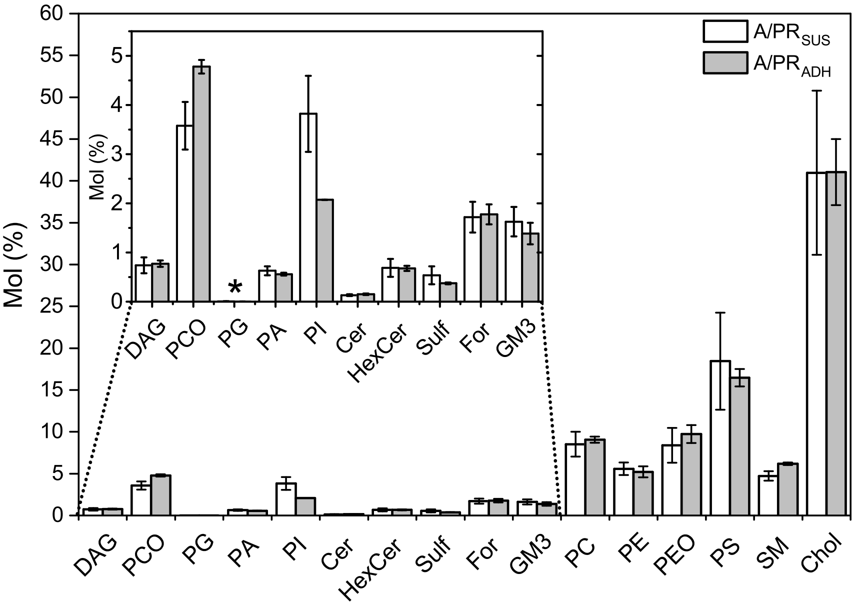


Figure 5: Lipid composition of suspension (A/PR$_{SUS}$) and adherent cell culture-derived influenza A virus particles (A/PR$_{ADH}$, n=3, mean±STD). A/PR samples were lipid extracted and their lipidome was determined. Bars show molar percentage of total lipids. Abbreviations: DAG: diacyl glycerol; PC: phosphatidylcholine; PCO: ether linked PC; PG: phosphatidylglycerol; PA: phosphatidic acid; PI: phosphatidylinositol; CER: ceramide; HexCer: hexosylceramide; Sulf: sulfated; For: Forssmann glycolipid; GM3: monosialodihexosylgangliosid; PE: phosphatidylethanolamine; PEO: ether linked PE; PS: phosphatidylserine; SM: sphingomyelin; Chol: cholesterol. * <0.01 mol% for A/PR$_{SUS}$ and A/PR$_{ADH}$.

4 Conclusion

In this work two influenza A VP samples, i.e., A/PR$_{ADH}$ produced in MDCK$_{ADH}$ and A/PR$_{SUS}$ produced in MDCK$_{SUS2}$, were compared based on their aggregation behavior in low-salt buffers, and their zeta potential, surface hydrophobicity, lipid composition as well as N-glycosylation of the major surface protein HA.

A/PR$_{ADH}$ showed a higher aggregation propensity in low-salt buffers when compared to A/PR$_{SUS}$ with aggregates up to pentamers. The investigation of the surface properties revealed a slightly but significantly lower zeta potential and a lower hydrophobicity (higher hydrophilicity) for A/PR$_{ADH}$ compared to A/PR$_{SUS}$. The HA of A/PR$_{ADH}$ had more of the larger non-charged N-glycans, whereas A/PR$_{SUS}$ had more smaller non-charged N-glycans. Based on this, a high surface hydrophobicity in combination with smaller N-glycans on the major surface protein HA seems to be linked to a higher colloidal stability in low-salt buffers. Therefore, N-glycosylation differences are very likely responsible for the observed aggregation behavior. Nevertheless, further detailed investigation on the VP membrane components, in particular of the second most common surface protein NA, are

needed to support or reject this hypothesis.
In summary, with the analytical toolbox outlined in this work we were able to cast more light on the underlying causes of influenza A VP aggregation. We were able to identify features that could be linked to the VP aggregation behavior, i.e., *N*-glycosylation of viral membrane proteins, but a causal relationship needs to be still investigated.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This research work is part of the project ”Optimization of an industrial process for the production of cell-culture-based seasonal and pandemic influenza vaccines” (grant number 0315640B and 0315640C) funded by the German Federal Ministry of Education and Research (BMBF). René Hennig and Erdmann Rapp acknowledge support by European Union (EC) under the project ”HighGlycan” (grant number 278535) and ”HTP-GlycoMet” (grant number 324400). Erdmann Rapp acknowledges support by the Germany Federal Ministry of Education and Research (BMBF) under the project ”Die Golgi Glykan Fabrik 2.0” (grant identifier 031A557). We gratefully acknowledge the support from Claudia Best, Ilona Behrendt, and Yvonne Genzel for the production of the virus material from the Max Planck Institute for Dynamics of Complex Technical System, and Dr. Sven Amrhein for the support with the surface tension measurements from the Karlsruhe Institute of Technology. We gratefully acknowledge the support from Dr. Julio Sampaio from the Max Planck Institute of Molecular Cell Biology and Genetics (MPI-CBG) with the quantitative lipidome analysis.

References

[1] R. Dunlap, E. Brown, D. Barry, Determination of the viral particle content of influenza vaccines by electron microscopy, Journal of Biological Standardization 3 (3) (1975) 281–289. doi:10.1016/0092-1157(75)90032-3.

[2] G. K. Hirst, M. W. Pons, Mechanism of influenza recombination, Virology 56 (2) (1973) 620–631. doi:10.1016/0042-6822(73)90063-9.

[3] D. C. Young, D. G. Sharp, Poliovirus aggregates and their survival in water., Applied and environmental microbiology 33 (1) (1977) 168–77.

[4] M. J. Mattle, B. Crouzy, M. Brennecke, K. R. Wigginton, P. Perona, T. Kohn, Impact of Virus Aggregation on Inactivation by Peracetic Acid and Implications for Other Disinfectants, Environmental Science & Technology 45 (18) (2011) 7710–7717. doi:10.1021/es201633s.

[5] C. Wallis, J. L. Melnick, Virus aggregation as the cause of the non-neutralizable persistent fraction., Journal of virology 1 (3) (1967) 478–88.

[6] J. O. Konz, A. L. Lee, J. A. Lewis, S. L. Sagar, Development of a Purification Process for Adenovirus: Controlling Virus Aggregation to Improve the Clearance of Host Cell DNA, Biotechnology Progress 21 (2) (2008) 466–472. doi:10.1021/bp049644r.

[7] R. J. Solá, K. Griebenow, Effects of glycosylation on the stability of protein pharmaceuticals, Journal of Pharmaceutical Sciences 98 (4) (2009) 1223–1245. arXiv:z0024, doi:10.1002/jps.21504.

[8] J. Schwarzer, E. Rapp, U. Reichl, N -glycan analysis by CGE-LIF: Profiling influenza A virus hemagglutinin N -glycosylation during vaccine production, ELECTROPHORESIS 29 (20) (2008) 4203–4214. doi:10.1002/elps.200800042.
URL `http://doi.wiley.com/10.1002/elps.200800042`

[9] J. Schwarzer, E. Rapp, R. Hennig, Y. Genzel, I. Jordan, V. Sandig, U. Reichl, Glycan analysis in cell culture-based influenza vaccine production: Influence of host cell line and virus strain on the glycosylation pattern of viral hemagglutinin, Vaccine 27 (32) (2009) 4325–4336. doi:10.1016/j.vaccine.2009.04.076.
URL `http://linkinghub.elsevier.com/retrieve/pii/S0264410X09006422`

[10] J. V. Rödig, E. Rapp, J. Bohne, M. Kampe, H. Kaffka, A. Bock, Y. Genzel, U. Reichl, Impact of cultivation conditions on N -glycosylation of influenza virus a hemagglutinin produced in MDCK cell culture, Biotechnology and Bioengineering 110 (6) (2013) 1691–1703. doi:10.1002/bit.24834.

[11] I. T. Schulze, Effects of Glycosylation on the Properties and Functions of Influenza Virus Hemagglutinin, The Journal of Infectious Diseases 176 (s1) (1997) S24–S28. doi:10.1086/514170.

[12] M. M. Pieler, A. Heyse, M. W. Wolff, U. Reichl, Specific ion effects on the particle size distributions of cell culture-derived influenza A virus particles within the Hofmeister series, Engineering in Life Sciences (2016) 1–29doi:10.1002/elsc.201600153.

[13] S. Amrhein, K. C. Bauer, L. Galm, J. Hubbuch, Non-invasive high throughput approach for protein hydrophobicity determination based on surface tension, Biotechnology and Bioengineering 112 (12) (2015) 2485–2494. doi:10.1002/bit.25677.

[14] S. Amrhein, S. Suhm, J. Hubbuch, Surface tension determination by means of liquid handling stations, Engineering in Life Sciences 16 (6) (2016) 532–537. doi:10.1002/elsc.201500179.

[15] R. Hennig, E. Rapp, R. Kottler, S. Cajic, M. Borowiak, U. Reichl, N-Glycosylation Fingerprinting of Viral Glycoproteins by xCGE-LIF, in: B. Lepenies (Ed.), Carbohydrate-Based Vaccines: Methods and Protocols, Methods in Molecular Biology, Springer New York, New York, NY, 2015, Ch. 8, pp. 123–143. doi:10.1007/978-1-4939-2874-3.

[16] B. Fields, D. M. Knipe, P. . Howley, Fields Virology, 4th Edition, Lippincott Williams & Wilkins, Philadelphia, 2001.

[17] N. M. Bouvier, P. Palese, The biology of influenza viruses, Vaccine 26 (SUPPL. 4) (2008) D49–D53. arXiv:NIHMS150003, doi:10.1016/j.vaccine.2008.07.039.

[18] B. Kalbfuss, A. Knöchlein, T. Kröber, U. Reichl, Monitoring influenza virus content in vaccine production: Precise assays for the quantitation of hemagglutination and neuraminidase activity, Biologicals 36 (3) (2008) 145–161. doi:10.1016/j.biologicals.2007.10.002.

[19] V. Lohr, Y. Genzel, I. Behrendt, K. Scharfenberg, U. Reichl, A new MDCK suspension line cultivated in a fully defined medium in stirred-tank and wave bioreactor, Vaccine 28 (38) (2010) 6256–6264. doi:10.1016/j.vaccine.2010.07.004.

[20] S. Kluge, D. Benndorf, Y. Genzel, K. Scharfenberg, E. Rapp, U. Reichl, Monitoring changes in proteome during stepwise adaptation of a MDCK cell line from adherence to growth in suspension, Vaccine 33 (35) (2015) 4269–4280. doi:10.1016/j.vaccine.2015.02.077.

[21] B. Peschel, S. Frentzel, T. Laske, Y. Genzel, U. Reichl, Comparison of influenza virus yields and apoptosis-induction in an adherent and a suspension MDCK cell line, Vaccine 31 (48) (2013) 5693–5699. doi:10.1016/j.vaccine.2013.09.051.

[22] R. Hennig, S. Cajic, M. Borowiak, M. Hoffmann, R. Kottler, U. Reichl, E. Rapp, Towards personalized diagnostics via longitudinal study of the human plasma N-glycome, Biochimica et Biophysica Acta (BBA) - General Subjects 1860 (8) (2016) 1728–1738. doi:10.1016/j.bbagen.2016.03.035.

[23] R. Herzog, D. Schwudke, K. Schuhmann, J. L. Sampaio, S. R. Bornstein, M. Schroeder, A. Shevchenko, A novel informatics concept for high-throughput shotgun lipidomics based on the molecular fragmentation query language, Genome Biology 12 (1) (2011) R8. doi:10.1186/gb-2011-12-1-r8.

[24] G. Liebisch, M. Binder, R. Schifferer, T. Langmann, B. Schulz, G. Schmitz, High throughput quantification of cholesterol and cholesteryl ester by electrospray ionization tandem mass spectrometry (ESI-MS/MS), Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids 1761 (1) (2006) 121–128. doi:10.1016/j.bbalip.2005.12.007.

[25] C. S. Ejsing, E. Duchoslav, J. Sampaio, K. Simons, R. Bonner, C. Thiele, K. Ekroos, A. Shevchenko, Automated Identification and Quantification of Glycerophospholipid Molecular Species by Multiple Precursor Ion Scanning, Analytical Chemistry 78 (17) (2006) 6202–6214. doi:10.1021/ac060545x.

[26] R. Herzog, D. Schwudke, A. Shevchenko, LipidXplorer: Software for Quantitative Shotgun Lipidomics Compatible with Multiple Mass Spectrometry Platforms, in: Current Protocols in Bioinformatics, John Wiley & Sons, Inc., Hoboken, NJ, USA, 2013, pp. 14.12.1–14.12.30. doi:10.1002/0471250953.bi1412s43.

[27] E. Y. Chi, S. Krishnan, T. W. Randolph, J. F. Carpenter, Physical Stability of Proteins in Aqueous Solution: Mechanism and Driving Forces in Non-native Protein Aggregation, Pharmaceutical Research 20 (9) (2003) 1325–1336. doi:10.1023/A:1025771421906.

[28] W. Wang, S. Nema, D. Teagarden, Protein aggregationPathways and influencing factors, International Journal of Pharmaceutics 390 (2) (2010) 89–99. doi:10.1016/j.ijpharm.2010.02.025.

[29] T. López-León, M. J. Santander-Ortega, J. L. Ortega-Vinuesa, D. Bastos-González, Hofmeister Effects in Colloidal Systems: Influence of the Surface Nature, The Journal of Physical Chemistry C 112 (41) (2008) 16060–16069. doi:10.1021/jp803796a.

4 Conclusion & Outlook

This PhD thesis tackles the aggregation of biopharmaceutical products during manufacturing, formulation, and storage. For this purpose, *in silico* as well as experimental methods were applied. In this context the following fields were approached:

- *In silico* methods:
  - QSAR modeling of the diffusion coefficient of proteins as a measure for protein-protein interactions
  - QSAR modeling of protein precipitation by polyethylene glycol for purification and prediction of aggregation propensity
- Experimental methods:
  - Assessment of the stability of influenza A viruses and the influence of parameters affecting their stability
  - Development of fast high-throughput compatible tools for estimating the stability of influenza A viruses

The application of the QSAR methodology as an *in silico* method enabled to successfully model the diffusion coefficients of proteins and the precipitation of proteins by polyethylene glycol. The calculated molecular descriptors accounted for structural properties, electrostatics, and the hydrophobicity of the protein molecules and the generated QSAR models were sensitive to the type of the protein, pH value, and ionic strength. QSAR modeling of protein diffusion coefficients allowed to gain a deeper understanding of the protein properties affecting the value of the diffusion coefficient as well as the protein-protein interactions present in solution. The application of the generated model to an external test set of proteins resulted in accurate predictions of the diffusion coefficients. Hence, this model allows the estimation of the colloidal stability of proteins, as changes in the diffusion coefficient can be applied as a measure for protein-protein interactions. Until now, for the capturing of these interactions, the diffusion coefficients have to be determined experimentally. QSAR modeling of protein precipitation by polyethylene glycol enabled to gain a deeper understanding of the underlying mechanisms. So far, the development of these precipitation steps is mainly based on heuristic and experimental approaches due to the lack of mechanistic understanding. The predictive capabilities were also assessed by the application to an external test set. The model has proven its potential to accurately predict the complete precipitation curves for proteins, but revealed deficiencies for proteins with a molecular weight below 25 kDa.

The methodology of QSAR was in this work for the first time successfully applied in a field other than chromatography during the process development and manufacturing of pharmaceutical proteins. The application of QSAR enabled a deeper understanding of the modeled processes and, thus, follows the tenet of the quality by design approach, that is increasingly demanded by regulatory authorities. Additionally, the implementation of *in silico* methods during process development of biopharmaceutical products has the potential to significantly reduce the sample consumption and experimental time. As computational power is constantly increasing, further work in this field of research could focus on the application of the QSAR methodology to biopharmaceutical products with a high complexity, such as viruses or virus-like particles, also enabling *in silico* process development for this class of products.

For influenza viruses, the colloidal and biological stability was systematically evaluated by phase diagrams generated with an automated liquid handling station in the microliter scale for a huge number of varying environmental conditions. Based on these results, a toolbox for the rapid assessment of virus stability was developed. For the colloidal stability, the precipitation of viruses with polyethylene glycol has proven to be a suitable, fast, and high-throughput compatible method to predicting the aggregation of viruses under the respective conditions. For the biological stability, the combination of the determination of the zeta potential and surface hydrophobicity and FT-IR analytics revealed changes in the structure of the surface proteins of the virus particles. These conformational changes are accompanied by a decrease in the hemagglutination activity. The combination of these methodologies depicts a powerful toolbox for the development of influenza vaccine formulations with a preserved colloidal and conformational stability at ambient temperature and thereby facilitates the rapid development of stable and safe vaccine formulations.

The influence of the production system on the surface characteristics of influenza virus particles was also investigated in this work. It could be shown that the surface properties reveal significant differences, whether the influenza viruses particles were produced in adherent or suspension MDCK cells. This leads to an increased aggregation propensity for the virus particles derived from the adherent cells compared to the virus particles derived from the suspension culture under low-salt conditions. Both virus particle samples were compared based on their zeta potential, hydrophobicity, lipid composition, and N-glycosylation fingerprints. Results indicated a slightly lower zeta potential and a lower hydrophobicity for the virus particles produced in adherent MDCK cells. The lipid composition of the membrane was fairly identical for both virus particle samples. Furthermore, it was found that there are more bigger non-charged N-glycans for the hemagglutinin of virus particles produced in adherent MDCK cells, whereas the virus particles derived from suspension culture had more smaller non-charged N-glycans. Based on this, a high surface hydrophobicity in combination with smaller N-glycans on the major surface protein hemagglutinin seems to be linked to a higher colloidal stability in low-salt buffers. Therefore, N-glycosylation differences are very likely responsible for the observed differences in aggregation behavior. Nevertheless, further detailed investigation on the virus particle membrane components are needed to support or reject this hypothesis.

5 Abbreviations

Abbreviation	Definition
AEX	Anion-exchange chromatography
A/PR	Influenza A/Puerto Rico/8/34
APTS	Aminopyrene trisulfonic acid
ATPE	Aqueous two-phase extraction
B_{22}	Second osmotic virial coefficient
β	Slope of precipitation curve
BisTris	2,2-Bis(hydroxymethyl)-2,2',2"nitrilotriethanol
BMBF	German Federal Ministry of Education and Research
BSA	Bovine serum albumin
CAPS	3-(Cyclohexylamino)-1-propanesulfonic acid
CE	Cholesterol ester
Cer	Ceramide
Chol	Cholesterol
CSC	Critical stabilization concentration
D	Diffusion coefficient
D_0	Diffusion coefficient at infinite dilution
DAG	Diacylglycerol
DCS	Differential centrifugation sedimentation
δ	Empirical coefficient
DLS	dynamic light scattering
DLVO	Deryagin-Landau-Verwey-Overbeek
DoE	Design of experiments
ECACC	European Collection of Authenticated Cell Cultures
ESP	Electrostatic surface potential
η_S	Viscosity of surrounding solution
For	Forssman glycolipid
FT-IR	Fourier transform infrared
GalCer	Galactosylceramide
γ	Empirical coefficient
GM3	Monosialodihexosylganglioside
HA	Hemagglutinin
HexCer	Hexosylceramide
HIC	Hydrophobic interaction chromatography
HSA	Human serum albumin
HTE	High-throughput experimentation
k_b	Boltzmann constant
k_D	Diffusion interaction parameter
LacCer	Lactosylceramide
m^*	Discontinuity point
mAb	monoclonal antibody
MDCK	Madin Darby canine kidney
MD	Molecular dynamics
MES	2-(N-morpholino)ethanesulfonic acid

MM	Molecular mass
MOI	Multiplicity of infection
MOPSO	3-Morpholino-2-hydroxypropanesulfonic acid
MTU"	Normalized migration time units
MS	Mass spectrometry
NA	Neuraminidase
NaCl	Sodium chloride
PA	Phosphatidic acid
PC	Phosphatidyl choline
PC O-	Ether linked PC
PDB	Protein data bank
PE	Phosphatidylethanolamine
PE O-	Ether linked PE
PEG	Polyethylene glycol
PG	Phosphatidylglycerol
PI	Phosphatidylinostitol
pK_a	Acid dissociation constant
PLSR	Partial least squares regression
pI	Isoelectric point
PS	Phosphatidylserine
PSD	Particle size distribution
Q^2	Predictability
QbD	Quality by Design
QSAR	Quantitative structure-activity relationship
R^2	Coefficient of determination
ρ	Density
r_h	Hydrodynamic radius
RMSECV	Root mean square error of cross-validation
RPM	Revolutions per minute
S_0	Apparent intrinsic protein solubility in the absence of precipitant
SEC	Size exclusion chromatography
SM	Sphingomyelin
STD	Standard deviation
STDEV	Standard deviation
Sulf	Sulfatide
T	Temperature
TAG	Triacylglycerol
TAPS	N-Tris(hydroxymethyl)methyl-3-aminopropanesulfonic acid
TEM	Transmission electron microscopy
TFF	Tangential flow filtration
UV	Ultraviolet
VIP	Variable influence on the projection
VP	Virus particle
xCGE-LIF	Multiplexed capillary gel electrophoresis with laser induced fluorescence detection

References

AGGARWAL, S. R. (2012). *What's fueling the biotech engine-2011 to 2012.* Nature Biotechnology, 30(12):1191.

AHAMED, T., ESTEBAN, B. N., OTTENS, M., VAN DEDEM, G. W., VAN DER WIELEN, L. A., BISSCHOPS, M. A., LEE, A., PHAM, C. and THÖMMES, J. (2007). *Phase Behavior of an Intact Monoclonal Antibody.* Biophysical Journal, 93(2):610.

AMORIJ, J.-P., HUCKRIEDE, A., WILSCHUT, J., FRIJLINK, H. W. and HINRICHS, W. L. J. (2008). *Development of Stable Influenza Vaccine Powder Formulations: Challenges and Possibilities.* Pharmaceutical Research, 25(6):1256.

AMRHEIN, S., OELMEIER, S. A., DISMER, F. and HUBBUCH, J. (2014). *Molecular Dynamics Simulations Approach for the Characterization of Peptides with Respect to Hydrophobicity.* The Journal of Physical Chemistry B, 118(7):1707.

ANDREWS, L., RALSTON, S., BLOMME, E. and BARNHART, K. (2015). *A snapshot of biologic drug development: Challenges and opportunities.* Human & Experimental Toxicology, 34(12):1279.

ARAKAWA, T. and TIMASHEFF, S. N. (1984). *Mechanism of protein salting in and salting out by divalent cation salts: balance between hydration and salt binding.* Biochemistry, 23(25):5912.

ASAKURA, S. and OOSAWA, F. (1958). *Interaction between particles suspended in solutions of macromolecules.* Journal of Polymer Science, 33(126):183.

ASHERIE, N. (2004). *Protein crystallization and phase diagrams.* Methods, 34(3):266.

ATHA, D. H. and INGHAM, K. C. (1981). *Mechanism of precipitation of proteins by polyethylene glycols. Analysis in terms of excluded volume.* The Journal of Biological Chemistry, 256(23):12108.

BAUMANN, P. and HUBBUCH, J. (2016). *Downstream process development strategies for effective bioprocesses: Trends, progress, and combinatorial approaches.* Engineering in Life Sciences, pages 1–17.

BAUMGARTNER, K., GALM, L., NÖTZOLD, J., SIGLOCH, H., MORGENSTERN, J., SCHLEINING, K., SUHM, S., OELMEIER, S. A. and HUBBUCH, J. (2015). *Determination of protein phase diagrams by microbatch experiments: Exploring the influence of precipitants and pH.* International Journal of Pharmaceutics, 479(1):28.

BOKSER, A. D. and O'DONNELL, P. B. (2006). *Remington: The science and practice of pharmacy.* In FELTON, L. A. (editor), *Remington Essentials of pharmaceutics*, chapter 4, pages 37–49. Pharmaceutical Press, London, 21st edition.

BOUVIER, N. M. and PALESE, P. (2008). *The biology of influenza viruses.* Vaccine, 26(SUPPL. 4):D49.

BRANDAU, D. T., JONES, L. S., WIETHOFF, C. M., REXROAD, J. and MIDDAUGH, C. (2003). *Thermal Stability of Vaccines.* Journal of Pharmaceutical Sciences, 92(2):218.

BRIGHAM, K. L. (1997). *Gene Therapy for Diseases of the Lung.* CRC Press, Nashville, 104th edition.

BUYEL, J., WOO, J., CRAMER, S. and FISCHER, R. (2013). *The use of quantitative structure–activity relationship models to develop optimized processes for the removal of tobacco host cell proteins during biopharmaceutical production.* Journal of Chromatography A, 1322:18.

CHAUDHURI, R., CHENG, Y., MIDDAUGH, C. R. and VOLKIN, D. B. (2014). *High-throughput biophysical analysis of protein therapeutics to examine interrelationships between aggregate formation and conformational stability.* The AAPS journal, 16(1):48.

CHEN, D. and ZEHRUNG, D. (2013). *Desirable Attributes of Vaccines for Deployment in Low-Resource Settings.* Journal of Pharmaceutical Sciences, 102(1):29.

CHHATRE, S., FARID, S. S., COFFMAN, J., BIRD, P., NEWCOMBE, A. R. and TITCHENER-HOOKER, N. J. (2011). *How implementation of Quality by Design and advances in Biochemical Engineering are enabling efficient bioprocess development and manufacture.* Journal of Chemical Technology & Biotechnology, 86(9):1125.

CHI, E. Y., KRISHNAN, S., RANDOLPH, T. W. and CARPENTER, J. F. (2003). *Physical Stability of Proteins in Aqueous Solution: Mechanism and Driving Forces in Nonnative Protein Aggregation.* Pharmaceutical Research, 20(9):1325.

CHUNG, W. K., HOU, Y., HOLSTEIN, M., FREED, A., MAKHATADZE, G. I. and CRAMER, S. M. (2010). *Investigation of protein binding affinity in multimodal chromatographic systems using a homologous protein library.* Journal of Chromatography A, 1217(2):191.

COHN, E. J. (1925). *The Physical Chemistry of the Proteins.* Physiological Reviews, 5(3):349.

COLLINS, K. (2004). *Ions from the Hofmeister series and osmolytes: effects on proteins in solution and in the crystallization process.* Methods, 34(3):300.

CROMWELL, M. E. M., HILARIO, E. and JACOBSON, F. (2006). *Protein aggregation and bioprocessing.* The AAPS Journal, 8(3):E572.

CURTIS, R. A., PRAUSNITZ, J. M. and BLANCH, H. W. (1998). *Protein-protein and protein-salt interactions in aqueous protein solutions containing concentrated electrolytes.* Biotechnology and Bioengineering, 57(1):11.

DE YOUNG, L. R., FINK, A. L. and DILL, K. A. (1993). *Aggregation of globular proteins.* Accounts of Chemical Research, 26(12):614.

DEHMER, M., VARMUZA, K. and BONCHEV, D. (2012). *Statistical Modelling of Molecular Descriptors in QSAR/QSPR.* Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany.

DISMER, F. and HUBBUCH, J. (2010). *3D structure-based protein retention prediction for ion-exchange chromatography.* Journal of Chromatography A, 1217(8):1343.

FIELDS, B., KNIPE, D. M. and HOWLEY, P. . (2001). *Fields Virology.* Lippincott Williams & Wilkins, Philadelphia, 4th edition.

FINK, A. L. (1998). *Protein aggregation: folding aggregates, inclusion bodies and amyloid.* Folding and Design, 3(1):R9.

FREDDOLINO, P. L., ARKHIPOV, A. S., LARSON, S. B., MCPHERSON, A. and SCHULTEN, K. (2006). *Molecular Dynamics Simulations of the Complete Satellite Tobacco Mosaic Virus.* Structure, 14(3):437.

GALM, L., AMRHEIN, S. and HUBBUCH, J. (2016). *Predictive approach for protein aggregation: Correlation of protein surface characteristics and conformational flexibility to protein aggregation propensity.* Biotechnology and Bioengineering, pages n/a–n/a.

GEORGE, A., CHIANG, Y., GUO, B., ARABSHAHI, A., CAI, Z. and WILSON, W. W. (1997). *Second virial coefficient as predictor in protein crystal growth.* Methods in Enzymology, 276:100.

GIBSON, T. J., MCCARTY, K., MCFADYEN, I. J., CASH, E., DALMONTE, P., HINDS, K. D., DINERMAN, A. A., ALVAREZ, J. C. and VOLKIN, D. B. (2011). *Application of a High-Throughput Screening Procedure with PEG-Induced Precipitation to Compare Relative Protein Solubility During Formulation Development with IgG1 Monoclonal Antibodies.* Journal of Pharmaceutical Sciences, 100(3):1009.

HANKE, A. T. and OTTENS, M. (2014). *Purifying biopharmaceuticals: knowledge-based chromatographic process development.* Trends in Biotechnology, 32(4):210.

HOFMEISTER, F. (1888). *Zur Lehre von der Wirkung der Salze.* Archiv für Experimentelle Pathologie und Pharmakologie, 25(1):1.

HØIBERG-NIELSEN, R., FUGLSANG, C. C., ARLETH, L. and WESTH, P. (2006). *Interrelationships of Glycosylation and Aggregation Kinetics for Peniophora lycii Phytase †.* Biochemistry, 45(15):5057.

ICQ QUALITY IMPLEMENTATION WORKING GROUP (2011). *ICH-Endorsed Guide for ICH Q8/Q9/Q10 Implementation, 2011.* Technical Report December.

IVERIUS, P. and LAURENT, T. (1967). *Precipitation of some plasma proteins by the addition of dextran or polyethylene glycol.* Biochimica et Biophysica Acta (BBA) - Protein Structure, 133(2):371.

JANSON, J.-C. (editor) (2011). *Protein Purification.* Methods of Biochemical Analysis. John Wiley & Sons, Inc., Hoboken, NJ, USA.

JIANG, L., GAO, Y., MAO, F., LIU, Z. and LAI, L. (2002). *Potential of mean force for protein-protein interaction studies.* Proteins: Structure, Function, and Genetics, 46(2):190.

Juckles, I. (1971). *Fractionation of proteins and viruses with polyethylene glycol.* Biochimica et Biophysica Acta (BBA) - Protein Structure, 229(3):535.

Kaddar, M. (2013). *Global Vaccine Market Features and Trends.* World Health Organisation, pages 1–38.

Kalbfuss, B., Knöchlein, A., Kröber, T. and Reichl, U. (2008). *Monitoring influenza virus content in vaccine production: Precise assays for the quantitation of hemagglutination and neuraminidase activity.* Biologicals, 36(3):145.

Kapoor, S. and Dhama, K. (2014). *Insight into Influenza Viruses of Animals and Humans.* Springer International Publishing, Cham.

Kelley, B. D., Switzer, M., Bastek, P., Kramarczyk, J. F., Molnar, K., Yu, T. and Coffman, J. (2008). *High-throughput screening of chromatographic separations: IV. Ion-exchange.* Biotechnology and Bioengineering, 100(5):950.

Killian, M. L. (2008). *Hemagglutination Assay for the Avian Influenza Virus.* In *Avian Influenza Virus*, pages 47–52. Humana Press, Totowa, NJ.

Knevelman, C., Davies, J., Allen, L. and Titchener-Hooker, N. J. (2009). *High-throughput screening techniques for rapid PEG-based precipitation of IgG4 mAb from clarified cell culture supernatant.* Biotechnology Progress, 26(3):697.

Kuehner, D. E., Heyer, C., Rämsch, C., Fornefeld, U. M., Blanch, H. W. and Prausnitz, J. M. (1997). *Interactions of lysozyme in concentrated electrolyte solutions from dynamic light-scattering measurements.* Biophysical journal, 73(6):3211.

Kumar, V., Dixit, N., Zhou, L. L. and Fraunhofer, W. (2011). *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, 421(1):82.

Łącki, K. M. (2014). *High throughput process development in biomanufacturing.* Current Opinion in Chemical Engineering, 6:25.

Ladiwala, A., Rege, K., Breneman, C. M. and Cramer, S. M. (2005). *A priori prediction of adsorption isotherm parameters and chromatographic behavior in ion-exchange systems.* Proceedings of the National Academy of Sciences of the United States of America, 102(33):11710.

Ladiwala, A., Xia, F., Luo, Q., Breneman, C. M. and Cramer, S. M. (2006). *Investigation of protein retention and selectivity in HIC systems using quantitative structure retention relationship models.* Biotechnology and bioengineering, 93(5):836.

Lamb, R. and Krug, R. (2001). *Orthomyxoviridae: the viruses and their replication.* In Knipe, D. M. and Howley, P. M. (editors), *Fields VirologyVirology*, pages 1487–1531. Lippincott Williams & Wilkins, Philadelphia, 4 edition.

Lee, J., Gan, H. T., Latiff, S. M. A., Chuah, C., Lee, W. Y., Yang, Y.-S., Loo, B., Ng, S. K. and Gagnon, P. (2012). *Principles and applications of steric exclusion chromatography.* Journal of Chromatography A, 1270:162.

Lehermayr, C., Mahler, H.-C., Mäder, K. and Fischer, S. (2011). *Assessment of Net Charge and Protein–Protein Interactions of Different Monoclonal Antibodies.* Journal of Pharmaceutical Sciences, 100(7):2551.

Li, J., Rajagopalan, R. and Jiang, J. (2008). *Polymer-induced phase separation and crystallization in immunoglobulin G solutions.* The Journal of Chemical Physics, 128(20):205105.

Lin, Y.-B., Zhu, D.-W., Wang, T., Song, J., Zou, Y.-S., Zhang, Y.-L. and Lin, S.-X. (2008). *An Extensive Study of Protein Phase Diagram Modification: Increasing Macromolecular Crystallizability by Temperature Screening* †. Crystal Growth & Design, 8(12):4277.

Love, J. C., Love, K. R. and Barone, P. W. (2013). *Enabling global access to high-quality biopharmaceuticals.* Current Opinion in Chemical Engineering, 2(4):383.

Mahler, H.-C., Friess, W., Grauschopf, U. and Kiese, S. (2009). *Protein aggregation: Pathways, induction factors and analysis.* Journal of Pharmaceutical Sciences, 98(9):2909.

Manning, M. C., Chou, D. K., Murphy, B. M., Payne, R. W. and Katayama, D. S. (2010). *Stability of Protein Pharmaceuticals: An Update.* Pharmaceutical Research, 27(4):544.

Matheus, S., Friess, W., Schwartz, D. and Mahler, H.-C. (2009). *Liquid high concentration IgG1 antibody formulations by precipitation.* Journal of Pharmaceutical Sciences, 98(9):3043.

Mazza, C. B., Sukumar, N., Breneman, C. M. and Cramer, S. M. (2001). *Prediction of protein retention in ion-exchange systems using molecular descriptors obtained from crystal structure.* Analytical chemistry, 73(22):5457.

Mazza, C. B., Whitehead, C. E., Breneman, C. M. and Cramer, S. M. (2002). *Predictive quantitative structure retention relationship models for ion-exchange chromatography.* Chromatographia, 56(3-4):147.

Mhatre, R. and Rathore, A. S. (2008). *Quality by Design: An Overview of the Basic Concepts.* In *Quality by Design for Biopharmaceuticals*, pages 1–8. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Muschol, M. and Rosenberger, F. (1995). *Interactions in undersaturated and supersaturated lysozyme solutions: Static and dynamic light scattering results.* The Journal of Chemical Physics, 103(24):10424.

National Institute of Allergy and Infectious Diseases (2012). *https://www.niaid.nih.gov/research/vaccine-types.*

Neal, B., Asthagiri, D. and Lenhoff, A. (1998). *Molecular Origins of Osmotic Second Virial Coefficients of Proteins.* Biophysical Journal, 75(5):2469.

Neumann, G., Noda, T. and Kawaoka, Y. (2009). *Emergence and pandemic potential of swine-origin H1N1 influenza virus.* Nature, 459(7249):931.

Odijk, T. (2009). *Depletion Theory and the Precipitation of Protein by Polymer.* The Journal of Physical Chemistry B, 113(12):3941.

Oelmeier, S. A., Dismer, F. and Hubbuch, J. (2012). *Molecular dynamics simulations on aqueous two-phase systems - Single PEG-molecules in solution.* BMC Biophysics, 5(1):14.

Oelmeier, S. A., Ladd-Effio, C. and Hubbuch, J. (2013). *Alternative separation steps for monoclonal antibody purification: Combination of centrifugal partitioning chromatography and precipitation.* Journal of Chromatography A, 1319:118.

Ojala, F., Degerman, M., Hansen, T. B., Broberg Hansen, E. and Nilsson, B. (2014). *Prediction of IgG1 aggregation in solution.* Biotechnology Journal, 9(6):800.

Philippidis, A. (2013). *Genetic Engineering and Biotechnology News.*

Polson, A. (1977). *A Theory for the Displacement of Proteins and Viruses with Polyethylene Glycol.* Preparative Biochemistry, 7(2):129.

Priddy, T. S., Middaugh, C. R., P Wen, E., Ellis, R. and S Pujar, N. (2014). *Stabilization and Formulation of Vaccines.* Vaccine Development and Manufacturing, pages 237–261.

Przybycien, T. M. and Bailey, J. E. (1989). *Solubility-activity relationships in the inorganic salt-induced precipitation of α-chymotrypsin.* Enzyme and Microbial Technology, 11(5):264.

Pujar, N. S., Sagar, S. L. and Lee, A. L. (2014). *History of Vaccine Process Development*, pages 1–24. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Russel, W. B., Saville, D. A. and Schowalter, W. R. (1989). *Colloidal dispersions.* Cambridge University Press.

Saluja, A., Badkar, A. V., Zeng, D. L. and Kalonia, D. S. (2007a). *Ultrasonic rheology of a monoclonal antibody (IgG2) solution: Implications for physical stability of proteins in high concentration formulations.* Journal of Pharmaceutical Sciences, 96(12):3181.

Saluja, A., Badkar, A. V., Zeng, D. L., Nema, S. and Kalonia, D. S. (2007b). *Ultrasonic storage modulus as a novel parameter for analyzing protein-protein interactions in high protein concentration solutions: correlation with static and dynamic light scattering measurements.* Biophysical journal, 92(1):234.

SALUJA, A. and KALONIA, D. S. (2008). *Nature and consequences of protein–protein interactions in high protein concentration solutions.* International Journal of Pharmaceutics, 358(1-2):1.

SCHALLER, A., CONNORS, N. K., OELMEIER, S. A., HUBBUCH, J. and MIDDELBERG, A. P. J. (2015). *Predicting recombinant protein expression experiments using molecular dynamics simulation.* Chemical Engineering Science, 121:340.

SCHERMEYER, M.-T., SIGLOCH, H., BAUER, K. C., OELSCHLAEGER, C. and HUBBUCH, J. (2016). *Squeeze flow rheometry as a novel tool for the characterization of highly concentrated protein solutions.* Biotechnology and Bioengineering, 113(3):576.

SEGALL, R. S., COOK, J. S. and ZHANG, Q. (2015). *Research and Applications in Global Supercomputing.* IGI Global, Hershey PA, 1st edition.

SHIRE, S. J., SHAHROKH, Z. and LIU, J. (2004). *Challenges in the development of high protein concentration formulations.* Journal of pharmaceutical sciences, 93(6):1390.

SIM, S.-L., HE, T., TSCHELIESSNIG, A., MUELLER, M., TAN, R. B. and JUNGBAUER, A. (2012a). *Branched polyethylene glycol for protein precipitation.* Biotechnology and Bioengineering, 109(3):736.

SIM, S.-L., HE, T., TSCHELIESSNIG, A., MUELLER, M., TAN, R. B. and JUNGBAUER, A. (2012b). *Protein precipitation by polyethylene glycol: A generalized model based on hydrodynamic radius.* Journal of Biotechnology, 157(2):315.

TARDIEU, A., BONNETÉ, F., FINET, S. and VIVARÈS, D. (2002). *Understanding salt or PEG induced attractive interactions to crystallize biological macromolecules.* Acta Crystallographica Section D Biological Crystallography, 58(10):1549.

TAUBENBERGER, J. K. and KASH, J. C. (2010). *Influenza Virus Evolution, Host Adaptation, and Pandemic Formation.* Cell Host & Microbe, 7(6):440.

TAUFER, M., GANESAN, N. and PATEL, S. (2013). *GPU-Enabled Macromolecular Simulation: Challenges and Opportunities.* Computing in Science & Engineering, 15(1):56.

TSOKA, S., CINIAWSKYJ, O., THOMAS, O., TITCHENER-HOOKER, N. and HOARE, M. (2000). *Selective Flocculation and Precipitation for the Improvement of Virus-Like Particle Recovery from Yeast Homogenate.* Biotechnology Progress, 16(4):661.

VALENTE, J. J., PAYNE, R. W., MANNING, M. C., WILSON, W. W. and HENRY, C. S. (2005). *Colloidal behavior of proteins: effects of the second virial coefficient on solubility, crystallization and aggregation of proteins in aqueous solution.* Current pharmaceutical biotechnology, 6(6):427.

VAN BEERS, M. M. C. and BARDOR, M. (2012). *Minimizing immunogenicity of biopharmaceuticals by controlling critical quality attributes of proteins.* Biotechnology Journal, 7(12):1473.

WALSH, G. (2013). *Biopharmaceuticals: biochemistry and biotechnology.* John Wiley & Sons.

WALSH, G. (2014). *Biopharmaceutical benchmarks 2014.* Nature biotechnology, 32(7):992.

WANG, W. (2015). *Advanced protein formulations.* Protein Science, 24(7):1031.

WANG, W., NEMA, S. and TEAGARDEN, D. (2010). *Protein aggregation—Pathways and influencing factors.* International Journal of Pharmaceutics, 390(2):89.

WANG, W., SINGH, S. K., LI, N., TOLER, M. R., KING, K. R. and NEMA, S. (2012). *Immunogenicity of protein aggregates—Concerns and realities.* International Journal of Pharmaceutics, 431(1-2):1.

WORLD HEALTH ORGANIZATION (2014). *Influenza (Seasonal), Fact sheet No 211.*

WORLD HEALTH ORGANIZATION (2016). *http://www.who.int/topics/vaccines.*

WRIGHT, P. and WEBSTER, R. (2001). *Orthomyxoviruses.* In KNIPE, D. M. and HOWLEY, P. M. (editors), *Fields Virology*, pages 1533–1579. Lippincott Williams & Wilkins, Philadelphia, 4th edition.

YADAV, S., LIU, J., SCHERER, T. M., GOKARN, Y., DEMEULE, B., KANAI, S., ANDYA, J. D. and SHIRE, S. J. (2013). *Assessment and significance of protein–protein interactions during development of protein biopharmaceuticals.* Biophysical Reviews, 5(2):121.

YANG, T., BRENEMAN, C. M. and CRAMER, S. M. (2007a). *Investigation of multimodal high-salt binding ion-exchange chromatography using quantitative structure-property relationship modeling.* Journal of Chromatography A, 1175(1):96.

YANG, T., SUNDLING, M. C., FREED, A. S., BRENEMAN, C. M. and CRAMER, S. M. (2007b). *Prediction of pH-dependent chromatographic behavior in ion-exchange systems.* Analytical chemistry, 79(23):8927.

YEE, L. C. and WEI, Y. C. (2012). *Current Modeling Methods Used in QSAR/QSPR*, pages 1–31. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany.