Data-Driven Process Development for Virus-Like Particles

Implementation of Process Analytical Technology, Molecular Modeling, and Machine Learning

zur Erlangung des akademischen Grades eines

DOKTORS DER INGENIEURWISSENSCHAFTEN (DR.-ING.)

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

genehmigte

DISSERTATION

von

Philipp Vormittag, M.Sc.

aus Heilbronn

Erstgutachter: Prof. Dr. Jürgen Hubbuch

Zweitgutachter: Associate Prof. Dr. Marcel Ottens, PDEng

Tag der mündlichen Prüfung: 05.10.2020

Acknowledgement

In the last four years that I have worked on this PhD thesis, several people have contributed both scientifically and personally in many different ways to this result and the path thereto. I am very grateful for every discussion, advice, and entertainment that I could enjoy in these years and, although it is impossible to mention everyone, I would like to name the most significant contributors.

First, I would like to thank Prof. Jürgen Hubbuch for being my PhD supervisor and a great mentor. Jürgen always guards the back of his PhD students, which I experienced and greatly appreciated in several situations. I would like to stress how lucky I feel, that Jürgen let me go to so many different international conferences.

I am grateful to Prof. Marcel Ottens for taking the time to review my thesis although, during the coronavirus pandemic, so much time and effort was and is required to be put into the transformation of teaching.

Much of this research was made possible through a rich amount of data and material generously provided by BioNTech Protein Therapeutics GmbH. I would like to address special thanks to Thorsten Klamp, who gave valuable scientific input and took care of the prompt approval of conference papers, when I repeatedly closely approached the deadlines for abstract submissions.

I am both honored and grateful to have been part of the MAB, which, since my first encounter in 2012, has proved and stayed the institute of the KIT (or even the world?) with the most appealing character to me. This is due to great scientific minds, who are eager to learn, helpful collaborators, energetic party people, and cake-enthusiasts. I am proud that we established the amazing Snablehouse Parties together and am confident they will rise and shine again once we get back to normality. I have enjoyed all seminars, lab courses, cooperation, chats, and conferences. NOLA was legendary!

I have learned so much from my collaborators Marieke Klijn, Matthias Rüdt, and Nils Hillebrandt in the projects and can confidently say that this thesis would have been very different if we had not worked together. I would also like to thank Nicolai Bluthardt for his contributions to the manuscripts and our shared duty as champagne receptionists. Adrian Sanden provided great

input in our discussions on data science and we shared the doubts and challenges that arose towards the end of our PhD. Susanna Suhm's ability for swift action proved to be priceless when new lab material was needed on very short notice and she always had an ear for complaints of all kinds (especially those regarding the failure of people to comply with laboratory rules, which is a dear topic to her).

My students Daniel Büchler, Nils Hillebrandt, Jonathan Seidel, Angela Valentic, Christopher Berg, Carla Segovia, Christina Wegner, and Annabelle Dietrich were invaluable both in the lab as well as by contributing to ideas and to a good time.

I learned quite a lot from my ever helpful office mates Basti and Steffen with whom I shared room 201 (experimental office forever!) most of the time during my PhD. I highly enjoyed our almost every day politics talk and our scientific discussions, resulting in some inspiration!

I would like to thank my friends and family for taking my mind off of work whenever they had the chance to. My brothers Michael and Fabian, who proofread parts of my thesis and listened to me practicing conference talks. My girlfriend Leonie who shared all ups and downs that I lived through with this thesis and supported me until the day of the PhD defense, which meant a lot to me. The rest of my family, especially my grandma, and all my friends for their continuous interest in my progress. Last but not least, I would like to mention my parents, Melinda and Franz, who supported me during the whole of my 24 years of education – and not only in my professional endeavors – for which I am truly grateful.

Philipp Vormittag

Karlsruhe, 12th October 2020

"They say a little knowledge is a dangerous thing,
but it is not one half so bad as a lot of ignorance"

Terry Pratchett

Abstract

During the 20th century, life expectancy increased dramatically. From a medical standpoint, the major contributors to this success were widespread improvement of hygiene and the introduction of vaccination programs. Vaccines were the first systematically developed biological products to be applied as medical compounds and therefore paved the way towards modern pharmaceutical biotechnology. After insulin and human growth hormone, one of the earliest biotechnologically produced pharmaceutical products was a recombinant vaccine, in particular a recombinant hepatitis B surface antigen virus-like particle (VLP). VLPs lack viral infectious nucleic acids but resemble the virus they are derived from, thus inducing an immune response. While this Hepatitis B vaccine is still in use today, the application of VLPs diversified greatly as seen from numerous pre-clinical and clinical studies. VLPs are investigated as potential vaccines against infectious diseases, immunological disorders, or cancer. Their strong immunogenicity is harnessed for the display of foreign antigenic epitopes on the VLP, resulting in chimeric VLPs (cVLPs). As such, they have been shown to induce immune responses against cancer cells, overcoming the natural immunological self-tolerance towards cancer antigens. This being said, their high potential comes with challenges, for example associated with their molecular design and the production process. The aim of the molecular design is to create immunogenic and stable VLP candidates. However, the process to find viable VLP candidates is typically empirical, bringing along challenges such as a low solubility after expression in recombinant hosts or a lack of VLP immunogenicity. The VLP production process lacks tailored purification methods, resulting in lower productivities as compared to more established biopharmaceutical products, such as monoclonal antibodies. Additionally, VLP processing comes with the need to design VLP-specific process steps, such as the dis- and reassembly of the particles. Tackling these challenges would benefit from data-driven approaches, such as process analytical technology (PAT), molecular modeling, and machine learning. These would enhance process and product understanding, reduce experimental effort, and enable efficient monitoring and control of the processes.

Therefore, the goal of this thesis was to find answers to several of these challenges by implementing data-driven approaches to accompany the development of tailored process steps. In the first part of this thesis, VLPs and their production processes are reviewed, the advantages of the implementation

of PAT are described, the challenges associated with their molecular design are elucidated, and the opportunities of the application of machine learning to VLP development and processing are pointed out.

The second part of this thesis describes five studies, addressing various challenges associated with VLP design and bioprocessing. The first study (Chapter 3) focuses on a unique VLP-specific process step. For improved stability, homogeneity, and immunogenicity, VLPs have to be dis- and reassembled. Starting from a high pH solution containing disassembled VLPs, reassembly is achieved by increasing ionic strength and lowering the pH. Most laboratory-scale processes utilize dialysis for this buffer exchange, while cross-flow filtration (CFF) for buffer exchange is more scalable, reduces the buffer consumption, and improves the yield. Compared to dialysis, CFF requires more technical knowledge and knowledge of the VLP reassembly progress during the process. A comprehensive monitoring strategy would therefore be highly beneficial to implement (near-) real-time control of the VLP reassembly process by CFF. In this first study, a set-up was developed to monitor VLP reassembly by CFF with an on-line measurement loop comprising two different spectroscopic sensors. A potential control strategy for the VLP assembly process was seen in monitoring static and dynamic light scattering. The maximum of the static light scattering signal coincided with the maximum VLP concentration. This information is valuable, since after the VLP peak concentration, a degradation phase was observed, which has to be omitted to optimize VLP yield and purity. Analysis of the second derivative ultraviolet and visible (UV/Vis) spectra proved to be a useful orthogonal method to monitor VLP assembly, especially with the so-called a/b-ratio. The a/b-ratio, which changed over the course of the processes, describes the solvatization of tyrosine. The observation of the change in the a/b ratio is consistent with the fact that tyrosine 132 is embedded in a hydrophobic pocket after assembly. Additionally, a partial least squares regression model based on the recorded UV/Vis spectra estimated VLP concentrations, with the potential to be applied as a (near) real-time model. The established monitoring strategy was used to investigate optimal process conditions for three chimeric hepatitis B core antigen (HBcAg) constructs. This resulted in different process times to reach the maximum VLP concentration. The cVLP with strongest negative zeta potential assembled the latest, probably due to repulsive electrostatic forces, demanding higher ionic strength buffers for reassembly.

The importance of the zeta potential for VLP processing was part of the motivation for the second study (Chapter 4). Zeta potential and other process-relevant biophysical parameters can only be measured when the molecules are produced experimentally in sufficient quantities. It would therefore be desirable to predict these properties, thus saving resources. It was already shown that surface properties can be derived from three-dimensional (3-D) structures. However, 3-D structures of novel molecules are not available and their experimental creation is lengthy and laborious. An alternative is computational 3-D structure generation based on template modeling and molecular dynamics (MD) simulations. This *in silico* workflow typically requires significant user interaction, expert knowledge to design and steer the simulations, and much computational power. To overcome these limitations, a robust and automated 3-D structure generation workflow was established in this study. The workflow is data dependent, minimizes user interaction, and reduces required computational resources. The input to the developed workflow was an amino acid sequence and a structure template. The template was automatically downloaded from a protein structure database, cleaned, and the structure was homology modeled, followed by an energy minimization. A data-dependent 3-step MD simulation refined the structure, where a continuously increasing region of the molecule was simulated, until, finally, the entire molecule was simulated freely. The 3-step MD simulation approach was a major contributor to a reduction in required computational resources by first simulating structurally particularly uncertain areas of the molecule separately. Often, MD simulations are terminated after a fixed simulation time. In this study, the developed data-dependent simulation control terminated the simulations, when a Window of Stability (WoS) of 2 ns was reached, defined by the root mean square deviation (RMSD) of atom coordinates. This ensured that the MD simulation fluctuations were comparable between all simulated constructs in said WoS at the end of the simulation. The workflow resulted in reasonable simulation times (6.6-37.5 h) and high overall structural quality for the three chimeric HBcAg dimer structures. To demonstrate the applicability of the method, a case study was conducted in which the *in silico* surface charge of HBcAg dimers was correlated to the experimental zeta potential of entire capsids, showing high linear correlation. The extraction of the surface charge from the WoS was more robust than from a single simulation snapshot, underpinning the usefulness of the developed approach.

The third study (Chapter 5) addresses the problem that VLPs are often processed with technologies originally developed for products that are smaller

in size. This often results in processing limitations, such as low binding capacity of the chromatography resins used in the downstream process. Therefore, a new purification strategy was developed, integrating three different size-selective methods, as they seemed promising for selective separation of VLPs from impurities. The methods were precipitation/re-dissolution, CFF, and size-exclusion chromatography (SEC). Three process variants were designed and examined, where the best consisted of precipitation, wash, and re-dissolution on a CFF unit, followed by purification by a multimodal SEC column. This process showed the highest purity and a high yield and productivity. The developed processes were comparable or superior to literature processes. Further, monitoring and fractionation of the permeate stream allowed to identify product-containing fractions for selective pooling. Thus, product concentration and purity can be adjusted.

One of the major problems in cVLP molecular design is that candidates are often insoluble upon expression. The process to identify insoluble VLP constructs is typically empirical and thus time-consuming and resource-intensive. This challenge can be met by a model that predicts cVLP solubility. In Chapter 6, a soft ensemble vote classifier (sEVC) was developed as a machine learning tool to predict cVLP solubility, based on 568 different amino acid sequences and 91 different hydrophobicity scales. The ensemble model unifies the prediction of individual classifiers, which were one-level decision trees. The decision trees were trained with a hydrophobicity feature based on one hydrophobicity scale each. Stratified training set sampling and feature selection benefitted the model construction. Best models showed a Matthew's correlation coefficient (MCC) of $>.6$, which is comparable or superior to literature solubility model statistical values. Additionally, feature selection allowed to identify characteristic features of the investigated solubility problem, pointing out the importance of different amino acids for cVLP solubility. The analysis suggested that arginine might have an important role in recruiting VLP subunits during capsid assembly.

The last study was built on the model and results of Chapter 6, with the aim to optimize prediction outcomes and to extract more hidden information from the data. Systematic misclassification was observed in the previous study. This was addressed with an optimization algorithm adjusting the prediction of the model, when these systematic misclassifications were observed in the training set. A second optimization strategy synthesized and optimized hydrophobicity scales specifically for the presented cVLP solubility problem. Hereby, the

importance of tryptophan as a possible disruptor of protein folding was suggested based on the data. The best model created with the developed optimization workflows resulted in an external test set MCC of .77 (accuracy of .88) and is therefore significantly better than the non-optimized model and literature solubility models. Finally, the sEVC framework was evaluated in a case study to predict ammonium sulfate concentrations, as required for VLP precipitation (applied in Chapter 5). Therefore, the model was redesigned to function as a regression tool. It was evaluated with data of the precipitation of ten cVLPs by ammonium sulfate. The linear fit showed a promising correlation with an R^2 of .69.

In summary, an array of methods has been developed, from both a process development and computational development point of view, which may pave the way towards a VLP platform process. The integration of data-driven approaches, such as PAT, 3-D structure modeling, and machine learning can benefit both the performance and the understanding of VLP processing in the biopharmaceutical industry.

Zusammenfassung

Im Laufe des 20. Jahrhunderts stieg die Lebenserwartung deutlich an. Aus medizinischer Sicht trugen vor allem die umfassende Verbesserung der Hygiene und die Einführung von Impfprogrammen zu diesem Erfolg bei. Impfstoffe waren die ersten biologischen Produkte, die systematisch als medizinische Präparate eingesetzt wurden, und ebneten damit den Weg zur modernen pharmazeutischen Biotechnologie. Nach Insulin und menschlichem Wachstumshormon war eines der frühesten biotechnologisch hergestellten pharmazeutischen Produkte ein rekombinanter Impfstoff, im Speziellen ein virusähnliches Partikel (*virus-like particle*, VLP) auf Basis von rekombinantem Hepatitis-B-Oberflächenantigen. VLPs beinhalten keine infektiösen viralen Nukleinsäuren und sie ähneln dem Virus, von dem sie abgeleitet sind, wodurch sie eine Immunantwort induzieren können. Obwohl dieser Hepatitis-B-Impfstoff gegenwärtig noch verwendet wird, ist die heutige Anwendung von VLPs sehr unterschiedlich, wie aus zahlreichen präklinischen und klinischen Studien hervorgeht. VLPs werden als mögliche Impfstoffe gegen Infektionskrankheiten, immunologische Erkrankungen oder Krebs untersucht. Ihre starke Immunogenität wird für die Präsentierung von fremdantigenen Epitopen auf den VLPs genutzt, was sie zu chimären VLPs (*chimeric virus-like particles*, cVLPs) macht. Als solche induzieren sie nachweislich Immunantworten gegen Krebszellen und überwinden die natürliche immunologische Selbsttoleranz gegenüber Krebsantigenen. Allerdings ist ihr hohes Potenzial mit Herausforderungen verbunden, beispielsweise im Zusammenhang mit ihrem molekularen Design und dem Produktionsprozess. Das Ziel des molekularen Designs ist die Entwicklung immunogener und stabiler VLP-Kandidaten. Der Prozess, um geeignete VLP-Kandidaten zu finden, ist jedoch typischerweise empirisch und bringt Herausforderungen wie eine geringe Löslichkeit nach der Expression in rekombinanten Wirten oder unzureichende VLP-Immunogenität mit sich. Dem VLP-Produktionsprozess mangelt es an maßgeschneiderten Aufreinigungsmethoden, was im Vergleich zu etablierten biopharmazeutischen Produkten, wie z.B. monoklonalen Antikörpern, zu einer geringeren Produktivität führt. Hinzu kommt, dass bei der VLP-Prozessierung VLP-spezifische Prozessschritte, wie z.B. die Zerlegung und Reassemblierung der Partikel, entworfen werden müssen. Die Bewältigung dieser Herausforderungen würde von datengestützten Ansätzen wie der prozessanalytischen Technologie (*process analytical technology*, PAT), der

molekularen Modellierung und dem maschinellen Lernen profitieren. Diese würden das Prozess- und Produktverständnis verbessern, den experimentellen Aufwand reduzieren und eine effiziente Überwachung und Steuerung der Prozesse ermöglichen.

Daher war es Ziel dieser Arbeit, Antworten auf mehrere dieser Herausforderungen zu finden, indem datengestützte Ansätze implementiert wurden, um die Entwicklung maßgeschneiderter Prozessschritte zu begleiten. Im ersten Teil dieser Arbeit werden VLPs und ihre Produktionsprozesse besprochen, die Vorteile der Implementierung von PAT beschreiben, die Herausforderungen im Zusammenhang mit ihrem molekularen Design beleuchtet und die Möglichkeiten der Anwendung des maschinellen Lernens bei der VLP-Entwicklung und -Prozessierung aufgezeigt.

Der zweite Teil dieser Arbeit beschreibt fünf Studien, die darauf abzielen, Antworten auf einige der mit dem VLP-Design und der biotechnologischen Verfahrenstechnik verbundenen Herausforderungen zu finden. Die erste Studie (Kapitel 3) befasst sich mit einem besonderen VLP-spezifischen Prozessschritt. Für eine verbesserte Stabilität, Homogenität und Immunogenität müssen VLPs zerlegt und wieder reassembliert werden. Ausgehend von einer Hoch-pH-Lösung, die zerlegte VLPs enthält, wird die Reassemblierung durch die Erhöhung der Ionenstärke und die Senkung des pH-Wertes erreicht. Die meisten Prozesse im Labormaßstab nutzen die Dialyse für diesen Pufferaustausch, während die Querstromfiltration (*cross-flow filtration*, CFF) für den Pufferaustausch besser skalierbar ist, den Pufferverbrauch reduziert und die Ausbeute verbessert. Im Vergleich zur Dialyse erfordert die CFF mehr technisches Wissen und Kenntnisse über den VLP-Reassemblierungssfortschritt während des Prozesses. Eine umfassende Überwachungsstrategie wäre daher sehr vorteilhaft, um eine (Beinahe-) Echtzeit-Kontrolle des VLP-Reassemblierungsprozesses durch CFF zu implementieren. In dieser ersten Studie wird ein Aufbau zur Überwachung der VLP-Reassemblierung durch CFF mittels einer Online-Messschleife mit zwei verschiedenen spektroskopischen Sensoren beschrieben. Eine mögliche Kontrollstrategie für den VLP-Assemblierungsprozess wurde in der Überwachung der statischen und dynamischen Lichtstreuung gesehen. Das Maximum des statischen Streulichtsignals fiel mit der maximalen VLP-Konzentration zusammen. Diese Information ist sehr wertvoll, da nach diesem VLP-Konzentrationsmaximum eine Degradationsphase beobachtet wurde, die vermieden werden sollte, um Ausbeute und Reinheit der VLPs zu optimieren.

Die Analyse der zweiten Ableitung der ultravioletten und sichtbaren (*ultraviolet and visible*, UV/Vis) Spektren erwies sich als praktikable orthogonale Methode zur Überwachung der VLP-Assemblierung, insbesondere mit dem sogenannten a/b-Verhältnis. Das a/b-Verhältnis, welches sich im Zeitverlauf der Prozesse änderte, beschreibt die Solvatisierung von Tyrosin. Die Beobachtung der Veränderung des a/b-Verhältnisses deckt sich mit der Tatsache, dass Tyrosin 132 nach der Assemblierung in einer hydrophoben Tasche eingebettet wird. Zusätzlich konnte ein Modell der Regression der partiellen kleinsten Quadrate (*partial least squares*), das auf den aufgezeichneten UV/Vis-Spektren basiert, die VLP-Konzentrationen abschätzen mit dem Potential, als (Beinahe-) Echtzeitmodell angewendet zu werden. Die etablierte Überwachungsstragie wurde genutzt um optimale Prozessbedingungen für drei chimäre *hepatitis B core antigen* (HBcAg)-Konstrukte zu ermitteln. Dies resultierte in unterschiedlichen Prozesszeiten, um die maximale VLP-Konzentration zu erreichen. Das cVLP mit dem stärksten negativen Zetapotential assemblierte am spätesten, wahrscheinlich aufgrund abstoßender elektrostatischer Kräfte. Es erfordert daher Puffer mit höheren Ionenstärken für die Reassemblierung.

Die Bedeutung des Zetapotenzials für die VLP-Prozessierung war Teil der Motivation für die zweite Studie (Kapitel 4). Das Zetapotential und andere biophysikalische Parameter können nur gemessen werden, wenn Material experimentell in ausreichenden Mengen produziert wurde. Es wäre daher wünschenswert, diese Parameter vorherzusagen, um Ressourcen zu sparen. Es wurde bereits gezeigt, dass Oberflächeneigenschaften aus dreidimensionalen (3-D) Strukturen abgeleitet werden können. 3-D-Strukturen neuartiger Moleküle sind jedoch nicht verfügbar und ihre experimentelle Erzeugung ist langwierig und mühsam. Eine Alternative ist die rechnergestützte 3-D-Strukturerzeugung mit Template-Modellierung und Molekulardynamik-Simulationen (MD). Dieser *in silico* Arbeitsablauf erfordert üblicherweise signifikante Benutzerinteraktion, Expertenwissen, um die Simulationen zu designen und zu steuern, und viel Rechenleistung. Um diese Limitationen zu überwinden, wurde in dieser Studie ein robuster und automatisierter Arbeitsablauf zur Erzeugung von 3-D Strukturen etabliert. Der Arbeitsablauf ist datenabhängig, minimiert Benutzerinteraktion und reduziert die benötigte Rechenleistung. Die Eingabe in den entwickelten Arbeitsablauf war eine Aminosäuresequenz und eine Strukturvorlage. Die Vorlage wurde automatisch von einer Proteinstrukturdatenbank heruntergeladen, bereinigt und die Struktur wurde Homologie-modelliert, gefolgt von einer Energieminimierung. Eine

datenabhängige dreistufige MD-Simulation verfeinerte die Struktur, wobei ein kontinuierlich zunehmender Bereich des Moleküls simuliert wurde, bis schließlich das gesamte Molekül frei simuliert wurde. Der dreistufige MD-Simulationsansatz lieferte hierbei einen großen Beitrag zur Reduktion der benötigten Rechenleistung, in dem strukturell besonders unsichere Bereiche des Moleküls zunächst gesondert simuliert wurden. Oft werden MD-Simulationen nach einer bestimmten Simulationszeit beendet. In dieser Studie beendete die entwickelte datenabhängige Simulationskontrolle die Simulationen, wenn ein Stabilitätsfenster (*Window of Stability,* WoS) von 2 ns erreicht wurde, definiert durch die Wurzel der mittleren quadratischen Abweichung (*root mean square deviation*, RMSD) der Atomkoordinaten. Dies stellte sicher, dass die Fluktuationen der MD-Simulation zwischen allen simulierten Konstrukten innerhalb des genannten WoS am Ende der Simulation vergleichbar waren. Der Arbeitsablauf führte zu angemessenen Simulationszeiten (6,6-37,5 h) und einer hohen Gesamtstrukturqualität für die drei chimären HBcAg-Dimere. Um die Anwendbarkeit der Methode zu demonstrieren, wurde eine Fallstudie durchgeführt, in der die *in silico* Oberflächenladung von HBcAg-Dimeren mit dem experimentellen Zeta-Potential ganzer Kapside korreliert wurde, was eine hohe lineare Korrelation zeigte. Die Extraktion der Oberflächenladung aus dem WoS war robuster als aus einem einzelnen Simulationsschnappschuss, was die Nützlichkeit des entwickelten Ansatzes unterstreicht.

Die dritte Studie (Kapitel 5) befasst sich mit dem Problem, dass VLPs häufig mit Technologien prozessiert werden, die ursprünglich für kleinere Produkte entwickelt wurden. Dies führt oft zu Prozesslimitationen wie geringe Bindekapazitäten von Chromatographieharzen, die im *downstream process* verwendet werden. Daher wurde eine neue Aufreinigungsstrategie entwickelt, die drei verschiedene größenselektive Methoden integriert, da sie für die selektive Abtrennung von VLPs von Verunreinigungen vielversprechend erschienen. Die Methoden waren Fällung/Rücklösung, CFF und Größenausschlusschromatographie (*size exclusion chromatography*, SEC). Es wurden drei Verfahrensvarianten entwickelt und untersucht, wobei die beste aus Fällung, Waschen und Rücklösung auf einer CFF-Einheit, gefolgt von einer Reinigung durch eine multimodale SEC-Säule bestand. Dieses Verfahren zeigte die höchste Reinheit sowie eine hohe Ausbeute und Produktivität. Die entwickelten Verfahren waren den in der Literatur beschriebenen Verfahren vergleichbar oder überlegen. Die Überwachung und Fraktionierung des Permeatstroms ermöglichte es zudem, produkthaltige Fraktionen für das

selektive Vereinigen zu identifizieren. Auf diese Weise können Produktkonzentration- und Reinheit eingestellt werden.

Eines der Hauptprobleme beim Molekulardesign von cVLPs ist, dass die Kandidaten bei der Expression oft unlöslich sind. Der Prozess zur Identifizierung unlöslicher VLP-Konstrukte ist typischerweise empirisch und deshalb Zeit- und Ressourcenintensiv. Diese Herausforderung kann mit einem Modell bewältigt werden, welches die Löslichkeit von cVLPs vorhersagt. In Kapitel 6 wurde ein Soft Ensemble Vote Classifier (sEVC) als Werkzeug auf Basis von maschinellem Lernen zur Vorhersage der cVLP-Löslichkeit entwickelt, basierend auf 568 verschiedenen Aminosäuresequenzen und 91 verschiedenen Hydrophobizitäts-Skalen. Das Ensemble-Modell aggregiert die Vorhersage der einzelnen Klassifikatoren, bei denen es sich um einstufige Entscheidungsbäume handelt. Diese wurden jeweils mit einem Hydrophobizitäts-Merkmal auf der Grundlage einer Hydrophobizitäts-Skala trainiert. Stratifizierte Trainingssatzprobenahme und Merkmalsauswahl kamen der Modellbildung zugute. Die besten Modelle wiesen einen Matthew-Korrelationskoeffizienten (*Matthew's correlation coefficient,* MCC) von >0,6 auf, der mit den statistischen Größen von Löslichkeitsmodellen aus der Literatur vergleichbar oder diesen überlegen ist. Zusätzlich ermöglichte die Merkmalsauswahl (*feature selection*) die Identifizierung charakteristischer Eigenschaften (*features*) des untersuchten cVLP-Löslichkeitsproblems, wobei die Bedeutung verschiedener Aminosäuren für die cVLP-Löslichkeit hervorgehoben wurde. Die Analyse legte nahe, dass Arginin eine wichtige Rolle bei der Rekrutierung von VLP-Untereinheiten während der Kapsidassemblierung spielen könnte.

Die letzte Studie baute auf dem Modell und den Ergebnissen von Kapitel 6 auf, mit dem Ziel, die Vorhersageergebnisse zu optimieren und mehr versteckte Informationen aus den Daten zu extrahieren. In der vorherigen Studie wurde eine systematische Fehlklassifikation beobachtet. Dies wurde mit einem Optimierungsalgorithmus angegangen, der die Vorhersage des Modells anpasste, wenn diese systematischen Fehlklassifikationen im Trainingsdatensatz beobachtet wurden. Eine zweite Optimierungsstrategie synthetisierte und optimierte Hydrophobizitäts-Skalen spezifisch für das vorgestellte cVLP-Löslichkeitsproblem. Dabei wurde die Bedeutung von Tryptophan als möglicher Disruptor der Proteinfaltung anhand der Daten vorgeschlagen. Das beste Modell, das mit den entwickelten Optimierungsworkflows erstellt wurde, zeigte einen MCC von 0,77

(Korrektklassifikationsrate von 0,88) in Bezug auf das externe Test-Set. Schließlich wurde das sEVC-Framework in einer Fallstudie evaluiert, um Ammoniumsulfatkonzentrationen vorherzusagen, wie sie für die VLP-Fällung erforderlich sind (wie auch in Kapitel 5 angewandt). Daher wurde das Modell so umgestaltet, dass es als Regressionswerkzeug fungiert. Es wurde mit Daten der Ammoniumsulfat-induzierten Fällung von zehn cVLPs bewertet. Die lineare Regression zeigte eine vielversprechende Korrelation mit einem R^2 von 0,69.

Zusammenfassend lässt sich sagen, dass sowohl von dem Standpunkt der Prozessentwicklung als auch von der computergestützen Entwicklung aus eine Reihe von Methoden entwickelt wurde, die den Weg zu einem VLP-Plattformprozess ebnen könnten. Die Integration von datengesteuerten Ansätzen wie PAT, 3-D-Strukturmodellierung und maschinelles Lernen kann sowohl der Effizienz als auch dem Verständnis der VLP-Prozessierung in der biopharmazeutischen Industrie zugutekommen.

Table of Contents

1

Introduction

The 20th century has seen the most substantial number of medical revolutions in the human history, bringing along fundamental improvements in the health of everyone – from children to the elderly. Whereas in 1900 infectious diseases such as smallpox, measles, diphtheria, and pertussis were among the leading causes of death, their incidence has been radically reduced or even eliminated throughout the 20th century, as was the case with smallpox in 1979 (Centers for Disease Control and Prevention, 1999; Fenner et al., 1988). The most valuable contribution to this development was accomplished by widespread improvement of hygiene and through the introduction of vaccination programs, especially by reducing child mortality (Janeway, Murphy, Travers, & Walport, 2008; McGovern & Canning, 2015). The origin of vaccine technology is often referred to as Edward Jenner's experiments with immunization against smallpox by infection with the, for humans, less harmful cow pox (Gross & Sepkowitz, 1998). Almost 100 years later, Louis Pasteur's work laid the fundament for modern vaccinology by describing the idea of attenuation for vaccination against veterinary diseases and, most notably, for rabies in humans (Pasteur, 1885). Since then, the portfolio of vaccine formats expanded rapidly and now includes live attenuated pathogens, killed whole organisms, purified proteins or polysaccharides, or more recently, genetically engineered vaccines (Plotkin, 2014). The first genetically engineered vaccine was developed to prevent Hepatitis B virus (HBV) infection, based on a Hepatitis B surface antigen virus-like particle (VLP) that is expressed in yeast cells (McAleer et al., 1984). A list of licensed VLP vaccines is shown in Table 1.1. Since then, the number of exploratory and approved genetically engineered vaccines has been on the rise, with the VLP-based

human papillomavirus (HPV) (Bryan, Buckland, Hammond, & Jansen, 2016) and meningococcus group B vaccine (Giuliani et al., 2006) as prominent examples. With the numbers of available vaccines increasing, the technological portfolio is expanding as well. From messenger ribonucleic acid (mRNA) vaccines, to deoxyribonucleic acid (DNA) vaccines, to various types of VLP vaccines, recent research is exploring many ways to address previously unmet medical needs. This increased diversity of the vaccine portfolio comes with an urgent demand for novel production technologies. This includes synthesis of the product (upstream processing, USP), purification from product- or process-related contaminants (downstream processing, DSP), formulation to ensure immunogenicity and stability, and logistics (Kaufmann, Juliana McElrath, Lewis, & Del Giudice, 2014; Ladd Effio & Hubbuch, 2015; Plotkin, Robinson, Cunningham, Iqbal, & Larsen, 2017; S. Wang, Liu, Zhang, & Qian, 2015).

Table 1.1: Licensed VLP vaccine products (Huzair & Sturdy, 2017; Lua et al., 2014).

Targeted disease	**Trade name**	**Country and year of first approval**	**Producer at time of approval**
Hepatitis B	*Recombivax HB*	West Germany, 1986	Merck Sharp & Dohme
Hepatitis B	*Engerix-B*	Belgium, 1986	SmithKline Biologicals
Hepatitis B	*GenHevac-B*	France, 1989	Pasteur Vaccins
Cervical cancer and genital warts	*Gardasil*	US, 2006	Merck Sharp & Dohme
Cervical cancer and genital warts	*Cervarix*	EU, 2007	GlaxoSmithKline
Hepatitis E	*Hecolin*	China, 2011	Xiamen Innovax Biotech

As one of the newer technologies, VLPs are studied for various applications, such as cancer and malaria vaccines and as carrier for the delivery of nucleic acids or proteins. The application of VLPs as recombinant vaccine technology brings along many advantages. These include the generally high application safety and their potential to induce both cellular and humoral immune responses, including the breakage of immunological self-tolerance towards cancer antigens. In order to fulfill their potential to combat cancer or other previously unmet medical

needs, challenges associated with their production process demand solutions to facilitate the development process. The challenges for process development include, but are not limited to, low dynamic binding capacity in bind-and-elute chromatography (Ladd Effio & Hubbuch, 2015), the necessity of additional, specific process steps, such as dis- and reassembly or conjugation reactions (Peacey, Wilson, Baird, & Ward, 2007; Pomwised, Intamaso, Teintze, Young, & Pincus, 2016; Q. Zhao, Allen, et al., 2012; Q. Zhao, Modis, et al., 2012), or the lack of knowledge about the relation between the amino acid sequence of the VLP candidates and their immunogenicity, structure, process and phase behavior (Karpenko et al., 2000; Roseman et al., 2012).

In the last decades, regulatory authorities have strongly encouraged to build quality into processes by implementing process analytical technology (PAT) (FDA & Others, 2004; ICH, 2009). Monitoring of processes with PAT does not only enable a direct response to process variations, but also generates a great amount of data that directly contributes to process understanding. The availability of large amounts of data comes with opportunities, but also challenges. Biopharmaceutical process development has seen the advent of data science and machine learning in recent years, aiming to harness the great amount of data available to generate predictive models and to increase product and process understanding (Charaniya, Hu, & Karypis, 2008; Gangadharan et al., 2019; Tulsyan, Garvin, & Ündey, 2018). Applications range from identifying critical process parameters during fermentation (Buck, Subramanian, & Block, 2002), over artificial neural networks (ANNs) for mechanistic chromatography modeling (G. Wang, Briskot, Hahn, Baumann, & Hubbuch, 2017b), to clustering algorithms for evaluation of protein formulations (Klijn & Hubbuch, 2019). With the rapidly increasing computational resources available at ever lower costs, opportunities arise for molecular modeling to be included in the research and development process, for example to facilitate the prediction immunogenicity from three-dimensional (3-D) structures (Joshi, Cheluvaraja, Somogyi, Brown, & Ortoleva, 2011).

In the following subchapters, the structure and function of VLPs are described (Chapter 1.1). A typical production process of VLPs is illustrated and its challenges are discussed (Chapter 1.2). Finally, data-driven approaches, such as PAT (Chapter 1.3), molecular modeling

(Chapter 1.4), and machine learning methods (Chapter 1.5) are described in the context of the processes investigated in this thesis.

1.1 Virus-Like Particles

VLPs are macromolecular protein-based nanostructures that resemble the virus they are derived from. VLPs are highly immunogenic but not infectious as they do not contain viral nucleic acids (Chackerian, 2007). They can be subdivided into non-enveloped and enveloped VLPs. The latter are formed by secretion and envelopment in the host cell membrane (Kushnir, Streatfield, & Yusibov, 2012). Enveloped VLPs are for example expressed in a Baculovirus/insect cell system or mammalian cells, while non-enveloped VLPs are produced in *Escherichia coli* (*E. coli*) or yeast systems. Enveloped VLPs pose very different challenges with regard to their stability, expression systems, and purification compared to non-enveloped VLPs. The challenges associated with enveloped VLPs are reviewed elsewhere (Dai, Wang, & Deng, 2018). In the following pages, VLPs are discussed with a focus on non-enveloped VLPs, as these were investigated in this thesis.

The recently developed VLPs for human use are almost exclusively chimeric VLPs (cVLPs) (Mohsen, Speiser, Knuth, & Bachmann, 2020; Mohsen, Zha, Cabral-Miranda, & Bachmann, 2017; Ong, Tan, & Ho, 2017). CVLPs can be created by recombinant insertion of foreign peptide sequences into viral structure proteins, as schematically shown in Figure 1.1. Another option is chemical linkage of peptides to the surface of VLPs, for example by click chemistry (Brune et al., 2016). This bears the potential of a universal VLP platform but adds another process step.

Recently, a great number of preclinical and clinical trials have been under way for diverse applications of VLPs, including vaccines against cancer (Klamp et al., 2011; Mohsen et al., 2020), Alzheimer's disease (Maphis et al., 2019), Malaria (Chan et al., 2019), and Influenza (Buffin et al., 2019). Animal health is another domain for which VLP vaccines are increasingly considered as high potential strategies (Crisci, Bárcena, & Montoya, 2013). The wide spectrum of investigated applications for VLP vaccines illustrates that research and industry both acknowledge the potential that originates in the VLPs' unique structural properties and their functional versatility (Lua et al., 2014).

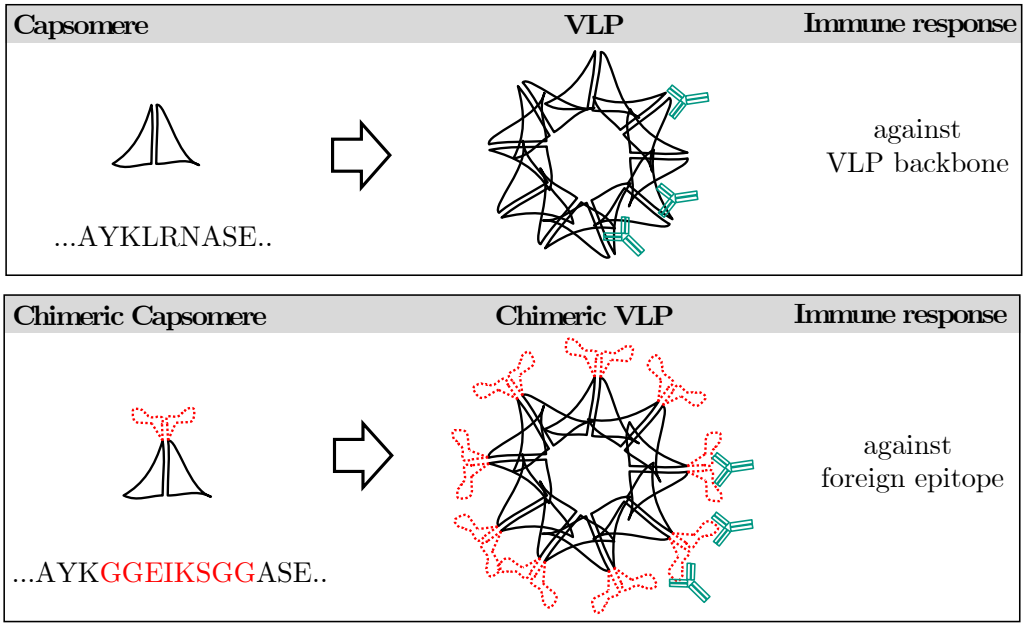


Figure 1.1: Schematic representation of a virus-like particle (VLP) and a chimeric VLP (cVLP) and their respective capsomere structures. The immune response towards a VLP is directed against the VLP-forming recombinant virus protein. The immune response towards a cVLP is ideally directed against the inserted foreign epitope (illustrated in red).

The biophysical prerequisite for the formation of VLPs is the natural property of some virus proteins to self-assemble to capsids after recombinant expression in various hosts, such as bacteria, yeast, or mammalian cell lines (Grgacic & Anderson, 2006). The assembly is initiated from morphological capsid subunits, termed capsomeres. These can be as simple as a virus protein homodimer, as for hepatitis B core antigen (HBcAg) capsomeres (J. Kim, 2016), or more complicated, such as heterohexamers in the *Picornaviridae* family that includes polio or enterovirus (Rustmeier, Strebl, & Stehle, 2019). The structure of VLPs is similar to the virus they were derived from, while they lack infectious nucleic acids (Chackerian, 2007; Kushnir et al., 2012). Their size ranges from 25 nm to 200 nm (Chung et al., 2010; Reiter et al., 2019).

In a study on Ag-coated nano-beads, 40 nm was shown to be the ideal size for uptake into dendritic cells (Fifis et al., 2004). Thus, their particulate nature probably is the reason that many VLPs can induce cytotoxic T cell responses through the major histocompatibility complex I pathway (Storni et al., 2002). Additionally, the highly repetitive surface structure of VLPs results in high B cell immunogenicity, triggering strong and long-lasting IgG responses (Fehr, Skrastina, Pumpens, &

Zinkernagel, 1998). Depending on the VLP design, they are applied as vaccines against infectious diseases, autoimmune disorders, or cancer. Especially the latter requires breaking the self-tolerance of the immune system against tumor-associated epitopes (Ong et al., 2017). For an efficient immune response against tumor cells, the combination of strong B cell responses and T cell responses induced by many VLPs is promising (Chackerian, 2007). HBcAg and Qβ VLPs only induce cytotoxic T cell responses, when adjuvanted to stimulate the immune system, for example with aluminium salts (Chackerian, 2007). This said, more recent results in animal models point at the potency of a strong B cell response inducing auto-antibodies against a tumor-associated cell lineage marker (Klamp et al., 2011). These results encourage the development of VLPs to apply them as vaccine against infectious diseases, autoimmune disorders, and cancer.

Since VLPs are protein-based structures, many of the platform technologies for the purification of other biopharmaceuticals, such as monoclonal antibodies, serve as a valuable toolbox. However, the VLPs' unique structure and large size pose challenges and opportunities that are elucidated in the following discussion of VLP production processes.

1.2 Production Process of Virus-Like Particles

VLP production processes are built on the same principle as many other biopharmaceutical processes. They begin with the USP, the actual biotechnological synthesis of the molecules. Starting from a cryo-culture, the inoculation train is scaled to production scale. After the harvest of the feed stock or cells, DSP of the product is initiated, consisting of product capture, purification, and polishing. Finally, the product is formulated and filled.

1.2.1 Expression and Lysis

Recombinant expression of VLPs can be performed in genetically modified bacteria, yeast, insect, plant, and mammalian cells (Kushnir et al., 2012; Vicente, Roldão, Peixoto, Carrondo, & Alves, 2011). A typical process is illustrated in Figure 1.2. VLPs can be expressed either intra- or extracellularly, depending on the expression system and the viral protein (Ladd Effio & Hubbuch, 2015). While *E. coli* is a highly efficient

host for the expression of non-enveloped VLPs, it cannot be applied to enveloped VLPs and produces VLPs intracellularly (J. Liu et al., 2016). Intracellular products are released from the cell by lysis. Common techniques are ultrasonic disruption (Ladd Effio, Baumann, et al., 2016; Wenger, DePhillips, & Bracewell, 2008) for lab scale processes or high-pressure homogenization for large-scale production (Cook et al., 1999; Z. Jiang, Tong, Cai, Xu, & Lou, 2011; Lünsdorf, Gurramkonda, Adnan, Khanna, & Rinas, 2011). After expression or lysis for extracellular or intracellular products, respectively, a solid-liquid separation step follows to remove the cells or cell debris, leaving the product in the liquid phase.

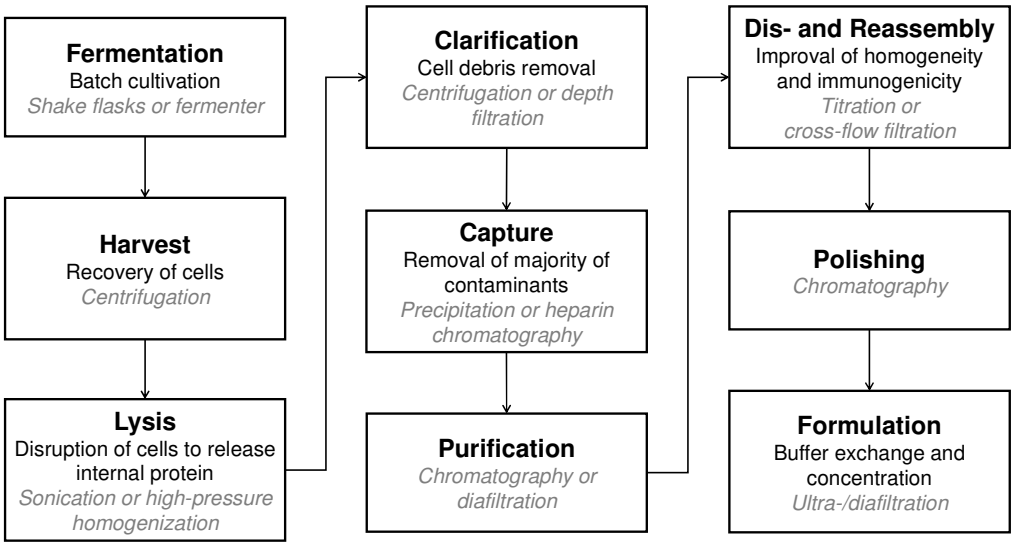


Figure 1.2: Typical production process of virus-like particles. Frequently used unit operations are indicated in gray italics.

1.2.2 Capture and Purification

The capture and purification train typically consists of unit operations such as chromatography, filtration, precipitation, and (ultra-) centrifugation (Ladd Effio & Hubbuch, 2015). Purification of virus-like particles has to deal with the challenge of increased particle size, which limits the diffusion into the pores of chromatography beads. Therefore, alternative chromatography technologies, such as monoliths (Burden, Jin, Podgornik, & Bracewell, 2012), membrane stacks (Ladd Effio, Hahn, et al., 2016; Vicente et al., 2008), or core bead resins (Lagoutte et al., 2016) are investigated. However, the VLPs' large size is not only a challenge but also an opportunity, which can be exploited in filtration processes (Negrete, Pai, & Shiloach, 2014; Vicente et al., 2014), precipitation processes (H. J. Kim et al., 2010), size-exclusion

chromatography (SEC) (Ladd Effio, Oelmeier, & Hubbuch, 2016), and ultracentrifugation (Ausar, Foubert, Hudson, Vedvick, & Middaugh, 2006).

1.2.2.1 Ultracentrifugation and Chromatography

While ultracentrifugation is a common technique for lab-scale purification of VLPs (Ausar et al., 2006; X. Jiang, Wang, Graham, & Estes, 1992; Mason et al., 1996), its application in large-scale processes is limited (Kleiner, Hooper, & Duerkop, 2015). Reasons for that include variability of the process and difficulty to scale up. In the study by Kleiner *et al.*, filtration was suggested as a viable alternative to ultracentrifugation (Kleiner et al., 2015). Another study circumvented ultracentrifugation by a combination of polyethylene glycole (PEG) precipitation combined with an anion exchange (AEX) chromatography step (Koho et al., 2012). While their process led to high purities, the concentration of recovered VLP was lower than with ultracentrifugation. Reasons for that could be that the interior of the applied Q Sepharose XL chromatography beads is inaccessible to the purified norovirus VLPs as they are four times larger than the bead pores (Yao & Lenhoff, 2004). This in turn leads to low capacity in the bind-and-elute chromatography step. The fact that VLPs are larger than the typical pore size is harnessed by the core bead technology, in which only the interior of the beads is functionalized (Weigel et al., 2014). Impurities enter the pores and are bound to the strong multimodal ligands, while VLPs flow through (Reiter et al., 2019). Affinity media based on heparin or metal ions have also shown promising results for the purification of HPV and norovirus VLPs, respectively (Koho et al., 2015; Minkner et al., 2018). Generally, various processes exist that apply chromatography to VLP processes, many of which report comparably low dynamic binding capacity (Ladd Effio & Hubbuch, 2015).

1.2.2.2 Precipitation and Re-Dissolution

Precipitation and re-dissolution has been used in several studies as highly selective and efficient VLP capture and purification step (Kazaks et al., 2017; H. J. Kim et al., 2010; Koho et al., 2012; Tsoka, Ciniawskyj, Thomas, Titchener-Hooker, & Hoare, 2000; Zahin et al., 2016). The selectivity derives from the size-dependency of precipitation methods along with the large size difference between VLPs and other solutes

(Rothstein, 1993). Precipitation agents used in above-mentioned studies are PEG and the kosmotropic salt ammonium sulfate. There are two theories explaining the mechanism of PEG precipitation. The excluded volume theory of Atha and Ingham assumes that a volume around PEG molecules is inaccessible for proteins (Atha & Ingham, 1981). By increasing the PEG concentration, the accessible volume decreases. Thus, the VLP concentration in the remaining volume increases, leading to precipitation. Another theory describing the micro-scale inhomogeneities in solutions is the preferential solvation theory (Ben-Naim, 1988). Applied to aqueous PEG-protein systems, Arakawa and Timasheff state that PEG interacts preferentially with water and therefore is excluded from the surface of the protein (Arakawa & Timasheff, 1985). Both theories imply that larger proteins, having a larger surface, are precipitating at lower concentration of the precipitant. Arakawa and Timasheff investigated this effect also for anorganic salts (Arakawa & Timasheff, 1982). The preferential solvation induced by ammonium sulfate results in similar conclusions as for precipitation with PEG – the proteins precipitate, where larger proteins are more prone to precipitation due to their larger surface. However, surface charge is thought to have a greater effect than size for precipitation by kosmotropic salts (Curtis, Montaser, Prausnitz, & Blanch, 1998). Therefore, precipitation of VLPs with ammonium sulfate is straightforward due to their large size, but the required ammonium sulfate concentration still depends on the VLP surface charge.

Re-dissolution of VLPs is most often realized by centrifugation and resuspension of the VLP-containing pellet in a precipitant-free buffer or solubilization buffer (Kazaks et al., 2017; Koho et al., 2012; Masuda et al., 2018; Tsoka et al., 2000; Zahin et al., 2016). After resuspension, additional purification can be realized by dialysis of re-dissolved product and centrifugation of undissolved contaminants, AEX chromatography, or SEC (Masuda et al., 2018; Zahin et al., 2016).

1.2.2.3 Filtration for Clarification, Capture, and Purification of Virus-Like Particles

Carvalho and colleagues point out that a platform process for VLP production should include techniques that exploit properties that are comparably constant for different candidates or products (Carvalho, Silva, Moleirinho, et al., 2019). Filtration serves this purpose as it is

based almost exclusively on the size of the solutes. Generally, filtration for bioprocesses can be divided into dead-end filtration and cross-flow filtration (CFF). High capacity dead-end filtration is realized by depth filtration, often applied to clarify solutions early in the process (Besnard et al., 2016). Size-selective dead-end membrane filtration is applied for sterile filtration (Carvalho, Silva, Moreira, et al., 2019). CFF retains and recirculates the product, which flows over a membrane in each recirculation, thus depleting smaller solutes and solvent (van Reis & Zydney, 2007). Therefore, CFF can be additionally used for concentration of the product. By implementing an input stream of diafiltration buffer to the recirculation loop, CFF serves to exchange the buffer.

Filtration has been successfully applied for clarification of VLP solutions (Carvalho, Silva, Moreira, et al., 2019; US Patent 6,602,697, 2003; Cook et al., 1999; Tretyakova et al., 2016), purification of VLPs (Carvalho, Silva, Moleirinho, et al., 2019; Kleiner et al., 2015), VLP reassembly (Liew, Chuan, & Middelberg, 2012), and formulation (Carvalho, Silva, Moleirinho, et al., 2019). Most of these applications are membrane-based filtration, such as micro-, ultra-, or diafiltration. The diverse application of membrane filtration technology to VLP processing in these studies is illustrated by Figure 1.3. Hereby, the selected pore size or molecular weight cut-off (MWCO) determines the applications of filtration. With 0.2 µm pore size, bacteria, spores, and dust are retained and VLPs and other solutes pass the filter (Huhti et al., 2010). In a publication on human immunodeficiency virus VLPs, capture and concentration was realized with a MWCO of 500 kDa, for example (Negrete et al., 2014). In a study on VLP reassembly by cross-flow filtration, a 30 kDa MWCO membrane was employed and compared to dialysis with 10 kDa MWCO (Liew et al., 2012).

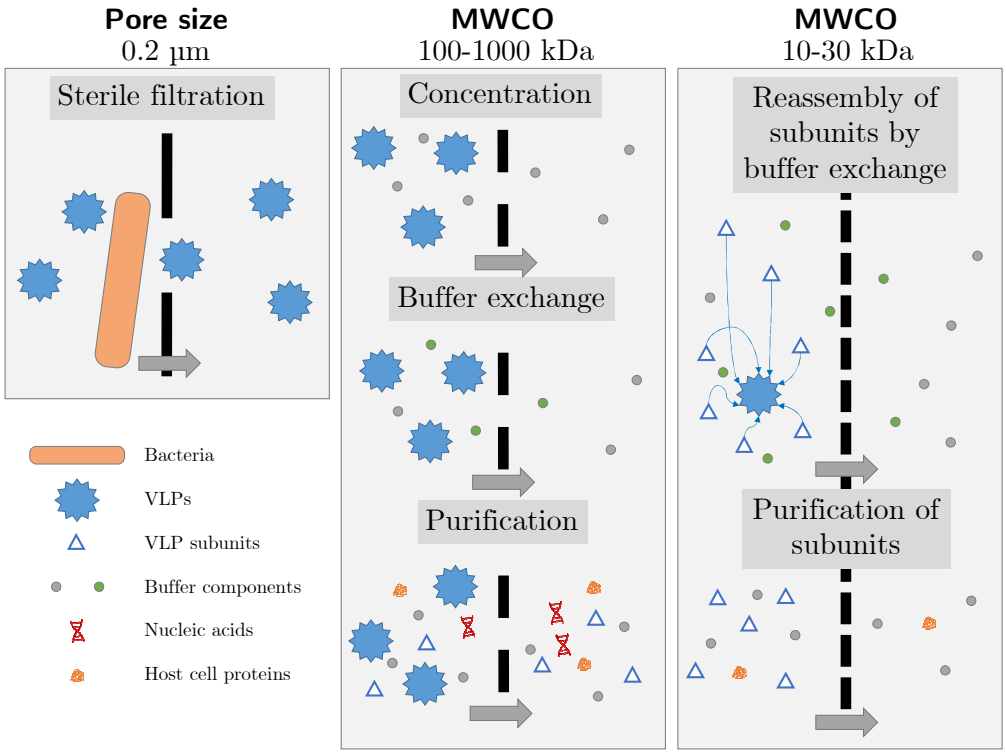


Figure 1.3: Illustration of the application of membrane filtration to virus-like particle (VLP) processes. With different pore sizes and molecular weight cut-offs (MWCOs), filtration serves as a tool for steps throughout the process ranging from clarification to sterile filtration of the formulated product. The illustrated steps can be found in various publications (Carvalho, Silva, Moleirinho, et al., 2019; Carvalho, Silva, Moreira, et al., 2019; Huhti et al., 2010; Liew et al., 2012; Negrete et al., 2014; Tretyakova et al., 2016).

In CFF, the flux across the membrane is determined by various parameters, including the membrane surface area, geometry, and pore size, the feed flow rate, and the viscosity and composition of the media (Van Reis et al., 1997; van Reis et al., 1997; van Reis & Zydney, 2007). Common problems in CFF processing are fouling, concentration polarization, or formation of a gel layer, all of which decrease the membrane flux. The occurrence of these events is dependent on a complex interplay of above-mentioned parameters (Bacchin, Si-Hassen, Starov, Clifton, & Aimar, 2002). Therefore, different approaches have been evolved to manage these challenges in process development. These include detailed studies of these parameters, modeling, and monitoring with process analytical technology (PAT) (Bacchin et al., 2002; Fernandez-Cerezo, Wismer, Han, & Pollard, 2019; Huter & Strube, 2019; Watson et al., 2016).

1.2.3 Dis- and Reassembly

Non-enveloped VLPs are regular oligomers, such as 240-mers (Wynne, Crowther, & Leslie, 1999) or 360-mers (Nilsson et al., 2005). As mentioned above, yeast and *E.coli* are popular expression systems for non-enveloped VLPs, resulting in high product yield of *in vivo* assembled particles (J. Liu et al., 2016). Often, these VLPs have structural defects and are inhomogeneous (Q. Zhao, Allen, et al., 2012). Therefore, VLPs are often dis- and reassembled, resulting in higher homogeneity and stability (Gallagher, Torian, McCraw, & Harris, 2017; Mach et al., 2006; McCarthy, White, Palmer-Hill, Koenig, & Suzich, 1998; Pattenden, Middelberg, Niebert, & Lipin, 2005; Q. Zhao, Allen, et al., 2012). Interestingly, dis- and reassembly also improved immunogenicity by increasing virion-like reactivity for HPV VLPs, as was done for the licensed HPV vaccine *Gardasil* (Q. Zhao, Modis, et al., 2012). Additionally, dis- and reassembly with an intermediate purification step allows for removal of contaminants contained in the void inside *in vivo* assembled VLPs (Link et al., 2012; Ren, Wong, & Lim, 2006).

The disassembled and assembled state is reached through a change in the quaternary structure of proteins. The mechanism behind this change is based on changes in disulfide linkages, weak electrostatic and hydrophobic interaction, temperature, and conformational changes (Ceres & Zlotnick, 2002; Hanslip, Zaccai, Middelberg, & Falconer, 2006; Kegel & Van Der Schoot, 2004; M Li et al., 1997; McCarthy et al., 1998; Sapp, Fligge, Petzak, Harris, & Streeck, 1998; Wingfield, Stahl, Williams, & Steven, 1995). While disulfide bridges stabilize the assembled capsids, they were found to not be required for assembly of HPV VLPs (Mukherjee, Thorsteinsson, Johnston, Dephillips, & Zlotnick, 2008). Mechanistic studies on HBcAg VLPs suggest that high ionic strengths induce a conformational change of capsomeres to an assembly-active state (Ceres & Zlotnick, 2002). A similar effect is thought to lead to an increase of assembly rate by bivalent cations (Choi, Gyoo Park, Yoo, & Jung, 2005; Stray, Ceres, & Zlotnick, 2004). Assembly from reduced HBcAg dimer structures was faster than from oxidized dimers, suggesting a geometrical effect induced by oxidation that is unfavorable for assembly (Selzer, Katen, & Zlotnick, 2014).

Disassembly of VLPs is achieved by lowering the ionic strength, adding of a reducing agent to break disulfide bridges, if present, the addition of

chaotropic agents, such as urea or guanidine hydrochloride, and increasing the pH (Mach et al., 2006; McCarthy et al., 1998; S. Singh & Zlotnick, 2003; Wingfield et al., 1995; A Zlotnick et al., 1996). This process step is typically realized by addition of NaOH, denaturant or chaotropic agent to the VLP sample, and by subsequent incubation.

The resulting capsomeres are then either directly reassembled or purified by filtration or chromatography, for example by SEC (Link et al., 2012; Mach et al., 2006; Ren et al., 2006; A Zlotnick et al., 1996). Reassembly is initiated by decreasing pH and increasing ionic strength, e.g. with NaCl or Na_2SO_4, typically by buffer exchange via dialysis or gel filtration (Mach et al., 2006; Wingfield et al., 1995; A Zlotnick et al., 1996). CFF has been employed as a scalable alternative to the lab-scale dialysis and gel filtration processes for the assembly of murine polyomavirus VLPs (Liew et al., 2012). Transfer to CFF increased yield and decreased buffer consumption.

1.2.4 Polishing, Formulation, and Logistics

Purity guidelines for vaccine products discriminate between product- and process-related contaminants (US Food and Drug Administartion & CBER, 1999). The majority of process-related contaminants, such as host cell proteins, DNA, cell debris and culture media are removed during capture and purification. The polishing step plays an important role in reducing product-related contaminants, such as aggregates, misfolded proteins or disassembled particles (Ladd Effio & Hubbuch, 2015). Typical unit operations for polishing are SEC (Carvalho et al., 2016; Lagoutte et al., 2016), CFF, and sterile filtration (Wagner et al., 2014). Polishing by CFF can easily be combined with a buffer exchange into the formulation buffer. Additives for increased storage stability include sugars, such as sucrose or sorbitol, surfactants, such as polysorbate 20 or 80, and amino acids, such as L-histidine (Mohr, Chuan, Wu, Lua, & Middelberg, 2013). To induce high and sustained immune responses, VLPs are often combined with adjuvants such as AS04 for *Cervarix* and Merck aluminium adjuvant for *Gardasil* (Garçon et al., 2011; Shi et al., 2005). Licensed HBV and hepatitis E virus vaccines are also formulated with aluminium adjuvant (Jain et al., 2015).

Considering the logistical challenge of distributing vaccines to remote parts of the earth, potentially with risks in breaking the cold chain, a

stable formulation is mandatory (Lloyd & Cheyne, 2017). Next to liquid formulations, freeze-drying is a frequent formulation strategy, leading to significantly enhanced product stability (Lang et al., 2009; Tumban et al., 2015). Another strategy to increase the stability of VLPs is to introduce disulfide bridges that stabilize the VLP structure (Ashcroft et al., 2005; Lu, Chan, Ko, Vanlang, & Swartz, 2016).

1.3 Process Analytical Technology

The 2004 published FDA *Guidance for Industry* on the implementation of PAT in pharmaceutical processes has been adopted more rapidly for small molecules than for biologics (FDA & Others, 2004; Rüdt, Briskot, & Hubbuch, 2017). Reasons for this include the biological molecules' complexity and the complexity of the associated processes. PAT is believed to increase process understanding and to enable automation, thus decreasing process costs (Rolinger, Rüdt, & Hubbuch, 2020). While at-line and on-line methods such as high-performance liquid chromatography (HPLC) have been implemented for biologics (Rathore, Yu, Yeboah, & Sharma, 2008; Tiwari, Kateja, Chanana, & Rathore, 2018), analysis of process streams with in-line sensors, such as spectroscopic tools, is desirable as it delivers real-time results, does not require sample preparation, and is non-destructive (Rolinger et al., 2020).

1.3.1 Implementation of Process Analytical Technology for Virus-Like Particle Processes

The field of PAT for VLP processing is still at an early stage. In a review on analytical technologies for Influenza VLPs, the lack of suitable methods for in-line analysis of the process stream has been pointed out (Thompson, Petiot, Lennaertz, Henry, & Kamen, 2013). Another review points out the potential of HPLC methods to serve as PAT for viral vaccines (Kramberger, Urbas, & Štrancar, 2015). As stated above, in-line methods have the potential to grant (near) real-time information about the process and should therefore be considered as PAT for VLP processing. Spectroscopic methods are regarded as the most promising toolbox for in-line sensors and include ultraviolet and visible (UV/Vis) absorbance spectroscopy, fluorescence spectroscopy, light-scattering spectroscopy, and infrared spectroscopy (Rolinger et al., 2020; Rüdt et

al., 2017). Various studies implemented spectroscopic techniques for the analysis of VLP stability or process steps, many of which refer to the potential to apply these techniques as a PAT tool.

1.3.1.1 Fluorescence Spectroscopy

Porcine circovirus type 2b VLP assembly and disassembly was analyzed by fluorescence spectroscopy analysis (M. Fang et al., 2016). The formation of disulfide bonds and the increase of define structures in assembled capsids was seen as one of the major drivers for the increased fluorescence. Intrinsic fluorescence has been monitored in studies investigating the pH and temperature stability of norovirus VLPs (Ausar et al., 2006), the HPV VLP dis- and reassembly process (Hanslip et al., 2006; Rajendar et al., 2013), disassembly of HBcAg VLPs (S. Singh & Zlotnick, 2003), and pH and temperature stability of the Marburg and Ebolavirus (Hu et al., 2011).

1.3.1.2 Light Scattering Spectroscopy

Light scattering methods, such as static or dynamic light scattering (SLS, DLS) are especially useful to detect the formation or disruption of the particulate assembled structures and have been applied to Marburg virus VLPs, Ebola virus VLPs, HBcAg VLPs, and HPV VLPs (Ausar et al., 2006; Hu et al., 2011; A. Zlotnick, Ceres, Singh, & Johnson, 2002; Adam Zlotnick, Johnson, Wingfield, Stahl, & Endres, 1999). In the product stream, the typical VLP concentration c is low compared to other bioprocesses. However, the intensity of scattered light I_R for Rayleigh scattering is strongly influenced by the particle diameter (Bohren & Huffman, 2004). It is proportional to the sum of the product of concentration c of species i with its diameter d_i to the power of six (Equation 1.1), making VLPs well detectable with light-scattering technologies.

$$I_R \propto \sum c_i d_i^6 \tag{1.1}$$

With lasers operating at >600 nm, most VLPs are typically in the size-range for Rayleigh scattering (Hosokawa, Nogi, Naito, & Yokoyama, 2012). Static light scattering models can therefore be applied to estimate molecular weight or root mean square radius (Lutomski et al., 2018; Pease et al., 2009; Somasundaram et al., 2016). This relation is less relevant for diffusion coefficient measurements by DLS, as particles with

a size equal to or larger than the wavelength of the laser can still be measured. However, in DLS, large particles overshadow smaller solutes due to their strong contribution to I_R. This in turn can facilitate the measurement of diffusion coefficients of large particles at low concentrations (Bohren & Huffman, 2004). A solute's diffusion coefficient is estimated by the method of cumulants in DLS (Koppel, 1972). The translation of the diffusion coefficient to a hydrodynamic diameter or radius is realized with the Stokes-Einstein-Equation (Equation 1.2).

$$D = \frac{k_B T}{3\pi\eta d}, \tag{1.2}$$

where k_B is Boltzmann's constant, T is the absolute temperature, η is the dynamic viscosity of the medium and d is the hydrodynamic diameter.

1.3.1.3 UV/Vis Absorbance Spectroscopy and Other Technologies

UV/Vis absorbance spectroscopy is routinely applied for concentration determination, based on the absorbance of aromatic side chains. Additionally, UV spectroscopy is sensitive to changes in disulfide bonding (Wetlaufer, 1963). This is particularly interesting, as disulfide bonds form in some VLPs upon assembly (Mukherjee et al., 2008). Another interesting approach is second derivative analysis of protein UV/Vis spectra, as has been pioneered by Ragone *et al.* and Mach and Middaugh (Mach & Middaugh, 1994; Ragone, Colonna, Balestrieri, Servillo, & Irace, 1984). While Mach & Middaugh point out the potential of wavelength shifts that occur when aromatic side chains move into different environments, Ragone *et al.* describe the ratio of peak-trough distances in the wavelength second derivative spectrum. The shift of the minimum of the second derivative at different wavelengths, for example at around 292 nm for tryptophan, was used in a study investigating norovirus VLP stability dependent on pH and temperature (Ausar et al., 2006).

Other technologies for the analysis of VLP dis- and reassembly are resistive-pulse sensing (Harms, Selzer, Zlotnick, & Jacobson, 2015), SEC (Ceres & Zlotnick, 2002; Ladd Effio, Oelmeier, et al., 2016), asymmetrical flow field-flow fractionation (Liew et al., 2012), or circular dichroism spectroscopy (Hu et al., 2011; S. Singh & Zlotnick, 2003). Fourier-transform infrared spectroscopy has been applied to detect poliovirus

infection in cell culture (Lee-Montiel, Reynolds, & Riley, 2011). It therefore seems to be a promising technology for VLP titer monitoring during USP.

1.3.2 Multivariate Data Analysis

Often used in combination with UV/Vis absorbance spectroscopy, multivariate data analysis is a powerful tool to analyze and steer processes (Rüdt et al., 2017). As such, principal component analysis (PCA) and partial least square (PLS) regression are the most popular techniques.

1.3.2.1 Principal Component Analysis

PCA is a common tool for classification, dimension reduction, and pattern identification (Wold, Esbensen, & Geladi, 1987). PCA approximates a given data table $\mathbf{X}$ by a matrix of scores $\mathbf{T}$ and loadings $\mathbf{P}$ resulting in error $\mathbf{E}$ as defined by Equation (1.3).

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \tag{1.3}$$

In most cases, the data are centered and scaled to unit variance for PCA. Often applied in the exploratory data analysis, PCA can be used to reveal patterns in a data set, as for example done for cell culture experiments (Bakker, Thomassen, & van der Pol, 2010; Mercier, Diepenbroek, Dalm, Wijffels, & Streefland, 2013; Suarez-Zuluaga, Borchert, Driessen, Bakker, & Thomassen, 2019). Furthermore, PCA is used to study the evolution of the seasonal influenza, potentially improving the forecast for the design of seasonal vaccines, and was used to help identify the origin of the 'swine flu' H1N1 virus (Konishi, 2019; Solovyov, Palacios, Briese, Lipkin, & Rabadan, 2009). It has been applied to study different vaccine formulations with respect to immune response data (Phanse et al., 2014). Fourier-transform infrared spectroscopy in combination with PCA was used to analyze freeze-dried vaccine formulations, for example to differentiate between formulations of different virus content (Hansen et al., 2015).

1.3.2.2 Partial Least Squares Regression

The estimation of response variables, such as a solute concentration, requires regression models. Compared to PCA, these models are applied to systems with both predictor variables X and one or more response

variables Y. PLS regression models are based on the transformation of the data to a latent variable space, similar to PCA (Wold, Sjöström, & Eriksson, 2001). In bioprocesses, PLS has been used to estimate component concentrations from spectral data (Andris, Rüdt, Rogalla, Wendeler, & Hubbuch, 2018; Brestrich, Rüdt, Büchler, & Hubbuch, 2018; Großhans et al., 2018). Similar to PCA, the data is often centered and scaled. However, for spectral data, scaling amplifies noise and should typically be omitted (Rüdt et al., 2017). While both X and Y data are reduced in their dimension similar to PCA, the explained variance is not maximized to create latent variables. For PLS, the goal of the algorithm is to maximize the covariance between the dimension-reduced X and Y data (Wold et al., 2001). The model results in an approximation of the X data by

$$\mathsf{X} = \mathsf{T}\mathsf{P}^{\mathsf{T}} + \mathsf{E}, \tag{1.4}$$

where T and P are scores and loadings, respectively, and E is the residual for the X data. The description of the Y data in reduced dimension is

$$\mathsf{Y} = \mathsf{U}\mathsf{Q}^{\mathsf{T}} + \mathsf{G}, \tag{1.5}$$

where U and Q are scores and loadings, respectively, and G is the residual for the Y data. The core of the PLS model is the maximization of the covariance between T and U, which is iteratively solved for the individual components of T and U. The regression model results from the latent X variables and the Y loadings (Equation (1.6).

$$\mathsf{Y} = \mathsf{T}\mathsf{Q}^{\mathsf{T}} + \mathsf{F}, \tag{1.6}$$

where F is the residual of the estimated Y data. With the transformation of X by the weights W to the latent variables $\mathsf{T} = \mathsf{X}\mathsf{W}$, the model can be rewritten to

$$\mathsf{Y} = \mathsf{X}\mathsf{W}\mathsf{Q}^{\mathsf{T}} + \mathsf{F}. \tag{1.7}$$

1.4 Molecular Design Challenges

The decision to approve and license a vaccine candidate is based on the demonstration of its efficacy and safety (Baylor, 2016). The safety of a VLP candidate depends largely on its purity and sterility. High purity and sterility are ensured with purification and polishing methods as

described above. The efficacy of a VLP is based on its capability to induce protecting and long-lasting specific immune responses against the target antigen, ideally inducing neutralizing antibodies (Kushnir et al., 2012). The immunogenicity of VLPs can depend on a variety of factors, one of which is the correct and high density display of the antigenic epitopes (Frietze, Peabody, & Chackerian, 2016; O'Rourke, Peabody, & Chackerian, 2015). Although the selection of epitopes to display on the VLP is often rational, a strong empirical element is involved in this process. This element can, for example, be simple screenings, but also vast library VLP display systems (O'Rourke et al., 2015). Another approach is the *in silico* analysis and design of vaccines. A great variety of models is described in literature, such as 3-D structure-based models or amino acid sequence-based models. These tackle different challenges, such as the prediction of immunogenicity, stability, or manufacturability. Therefore, the following section briefly touches upon 3-D structural analysis of capsid structures, and will then illustrate the most common problem in early VLP candidate process development, i.e. VLP solubility and assembly competence.

1.4.1 3-D Structural Analysis of VLPs

The scientific community has built databases containing 3-D structural data of tens of thousands of proteins and nucleic acids (www.rcsb.org) (Berman et al., 2000). 3-D structures contain valuable information on the geometry of the molecule(s), the interaction between side chains, and the nature and structure of the surface. While there is a great number of structures available, a vast number of protein structures, especially for newly discovered molecules, remains unresolved. The complexity of the structure determination workflow is high and the process is time-consuming (Steinbrener et al., 2010). Impressive time-lines have been realized in the structure determination of the main protease of SARS-CoV-2, due to the pressing need of a 3-D structure (Linlin Zhang et al., 2020). Computational technologies advance structure determination, such as homology modeling and molecular dynamics (MD) simulations. MD simulation aims to study the molecules' dynamic behavior or structure based on an atomistic or coarse grained simulation (Geng, Chen, Ye, & Jiang, 2019). For these simulations, 3-D structures are required, which can be retrieved from above-mentioned databases. If these structures are unavailable, homology modeling can serve with a

means to estimate the unknown 3-D structure by comparing sequence motifs with the proteins in the vast library of 3-D structures (Raval, Piana, Eastwood, Dror, & Shaw, 2012).

The application of 3-D structural analysis for vaccines includes description or prediction of immunogenicity, stability, and structure. Antibody-epitope complexes were studied to elucidate molecular interaction and identify potent VLP candidate structures (Roseman et al., 2012). Capsid structures have been investigated with MD simulations for analysis of stability and structure (Freddolino, Arkhipov, Larson, McPherson, & Schulten, 2006; Joshi et al., 2011; Roberts, Kuiper, Thorley, Smooker, & Hung, 2012; G. Zhao et al., 2013). Using small building blocks of murine polyomavirus VLPs, the suitability of linkers for the insertion of foreign epitopes has been evaluated (Lua, Fan, Chang, Connors, & Middelberg, 2015). 3-D structures of 1918 different Influenza VLPs have been investigated to elucidate possible routes for design of seasonal Influenza vaccines (McCraw et al., 2018).

3-D structure-based analysis can be envisaged to result in predictive models for various process- and product-relevant biophysical and physicochemical parameters, such as solubility, viscosity, or surface charge. However, with the high computational cost of MD simulations, it is hard to compete with advanced high-throughput laboratory techniques, when a large number of candidates have to be screened in a short time frame (Ladd Effio, Baumann, et al., 2016; Mohr et al., 2013). To simplify simulations, capsid subunits can be used, as mentioned above. However, this simplification often does not allow to describe the modeled properties, as was the case for an immunogenicity predictor (Joshi et al., 2011).

1.4.2 Virus-Like Particle Solubility and Assembly

Of the many properties that are of interest for the success of a product, protein solubility can lead to a molecule's failure very early in the process (Sormanni, Amery, Ekizoglou, Vendruscolo, & Popovic, 2017). For intra- and extracellular products, a well soluble product is found in the supernatant after cell lysis or cell removal, respectively. When VLPs or other proteins are overexpressed in hosts such as *E. coli*, low solubility means that the viral proteins are only detectable in inclusion bodies (Karpenko et al., 2000). Several approaches exist that extract VLPs from

inclusion bodies by solubilization (Bustos-Jaimes, Soto-Román, Gutiérrez-Landa, Valadez-García, & Segovia-Trinidad, 2017; Murthy, Ni, Meng, & Zhang, 2015; A. Singh, Upadhyay, Upadhyay, Singh, & Panda, 2015; Y. Zhang et al., 2020). While this is a possible DSP route, it comes with increased experimental effort, requirement of solution additives, and results in VLPs that are inherently more difficult to handle.

Besides solubilization of inclusion body proteins, there are VLP-specific strategies to circumvent the challenge of low solubility. In literature, the problem of low solubility often is linked to the lack of particle assembly, as insoluble proteins cannot assemble to VLPs. The problem of insoluble expression or lack of particle assembly is characteristic of recombinant cVLPs (Chackerian, 2007; Jegerlehner et al., 2002; Karpenko et al., 2000). This challenge can be overcome when working with chemical linkage of peptides to platform VLP structures, that themselves are soluble (Frietze et al., 2016). Techniques for coupling include click chemistry (Brune et al., 2016), chemical linkage to VLP surface groups (typically cysteine or lysine) by, for example, sulfo-SMCC chemistry (Peacey et al., 2007; Pomwised et al., 2016), or enzymatic linkage (Schoonen, Pille, Borrmann, Nolte, & van Hest, 2015). If a VLP scaffold and peptide linkage platform process is established, this can be a viable solution to deal with hydrophobic inserts. However, increased cost of production and decreased scalability via this process route explain that recombinant insertion is still the more popular strategy (Frietze et al., 2016).

Several studies on VLP assembly do not differentiate between soluble and insoluble expression but analyze total expression (including soluble and insoluble protein) and the presence of macromolecular structures, or analyze immunogenicity of purified lysates (Schödel et al., 1996; Ulrich et al., 1992). In other research on the solubility of *in vivo* assembled cVLPs, solubility is linked to the capability of the structure to self-assemble (Karpenko et al., 2000). In a study on woodchuck hepadnavirus VLPs, which is similar to HBcAg, low (soluble) expression levels correlated with low particle assembly (Billaud et al., 2005). This makes sense, as low concentration of soluble constructs point at aggregation and potentially presence of the product in inclusion bodies. This aggregated state of the proteins would make assembly of VLPs impossible. It is plausible that there are constructs, which are soluble and expressed at

high levels but that do not assemble to VLPs. However, researchers choose the insertion position rationally so that it theoretically does not interfere with assembly (Chackerian, 2007). It is therefore reasonable to believe that the case of strong soluble expression but low assembly is relatively rare. It can therefore be hypothesized, that constructs incapable of assembly to VLPs are probably found in the solid phase after cell lysis, i.e. they are aggregated and found in inclusion bodies. This hypothesis is supported by results of a study on RNA Phage MS2 VLPs, which found that the insertion of hydrophobic peptides can cause protein folding defects, leading to aggregation (Peabody et al., 2008). Earlier studies also acknowledge that inserted hydrophobic amino acids or amino acids with large residues, such as tryptophan, may interfere with assembly and lead to aggregation (Karpenko et al., 2000). Next to hydrophobicity, charge plays an important role for assembly competence (Billaud et al., 2005; Whitacre, Lee, & Milich, 2009).

Charge is an easily accessible property, dependent only on the pH and the presence of charged amino acids. Conversely, hydrophobicity in the context of proteins is a multifaceted property. Its relevance ranges from subnanoscale to nanoscale interactions in hydrophobic interaction chromatography (G. Wang, Hahn, & Hubbuch, 2016), protein folding, and aggregation (Lauer et al., 2012; Tanford, 1962; Valerio et al., 2005), to macroscale implications such as surface tension and viscosity (Galm, Amrhein, & Hubbuch, 2017). However, a hydrophobicity value cannot be derived as straightforward as the charge. For the derivation of the hydrophobic component of solvatization of proteins, no consistent and accurate approach exists (Harris & Pettitt, 2016). An early notion is the measurement of transfer free energies of amino acids from the protein interior, modeled by ethanol, for example, into water (Tanford, 1962). This idea of hydrophobicity was captured by various researchers in so-called hydrophobicity scales. A hydrophobicity scale assigns a particular hydrophobicity value to each amino acid. Meanwhile, a great number of hydrophobicity scales, derived from experimental or theoretical studies, have been described (Simm, Einloft, Mirus, & Schleiff, 2016). This said, the scales do not agree on the order of the amino acids' hydrophobicity values (Harris & Pettitt, 2016). The outcome they have in common is that large, hydrophobic amino acids are more likely to be buried in the interior of a protein. This agrees with the observation that large hydrophobic residues interfere with capsid assembly (Karpenko et al.,

2000), and that hydrophobic amino acids may disrupt folding of chimeric viral proteins (Peabody et al., 2008).

1.5 Machine Learning for Biopharmaceuticals

Machine learning tools begin to find their way into biopharmaceutical process development and manufacturing. The availability of today's powerful computational resources, the development and application of diverse machine learning tools, and the increasing amount of data available promote their implementation. A great diversity of methods are applied to various problems in manufacturing, development, and research, hinting towards the potential of machine learning in biopharmaceutical processes.

1.5.1 Machine Learning Applied to Biopharmaceutical Manufacturing, Development, and Research

In biopharmaceutical manufacturing, multivariate statistics have gained a foothold for batch process monitoring (Joeris, Frerichs, Konstantinov, & Scheper, 2002; Larsson, Liljas, & van der Spoel, 2012). Batch processes with little production history are a challenge, as limited amount of data is available for modeling (Tulsyan, Garvin, & Ündey, 2019). Gaussian processes have been used to generate *in silico* bioprocess data to address this problem (Tulsyan et al., 2018). Similar problems exist in chromatography modeling. The lack of data for calibration of the mechanistic model has been addressed by generating *in silico* data with an ANN model (G. Wang, Briskot, Hahn, Baumann, & Hubbuch, 2017a). The combination of mechanistic modeling and ANNs has also been used for root-cause investigation for chromatography process deviations (G. Wang et al., 2017b).

The development process for USP, DSP, and formulation can also benefit from machine learning by prediction of metabolite production and optimizing processes and formulations, for example. Generalized linear models have been applied to predict lactate production of Chinese hamster ovary cells by training on gene expression data at various conditions (Zampieri, Coggins, Valle, & Angione, 2017). With a deep Q-learning algorithm, liquid-liquid extraction processes were optimized based on model data (Hwangbo, Öner, & Sin, 2019). Optimized

formulations of pharmaceuticals have been predicted by deep learning methods trained on data of various existing formulations (Yilong Yang et al., 2019). This approach is also conceivable for formulation of biopharmaceutical drugs.

In addition, biopharmaceutical research has seen an advent of machine learning, for example in the prediction of protein solubility or the prediction of aggregating domains in proteins. Support-vector machines (SVM) and random forests (RF) have been successfully employed to predict protein solubility by various researchers (Agostini, Vendruscolo, & Tartaglia, 2012; Magnan, Randall, & Baldi, 2009; Samak, Gunter, & Wang, 2012). Another application was using SVM and RF as feature selection tool to identify features important to protein aggregation (Y. Fang, Gao, Tai, Middaugh, & Fang, 2013).

1.5.2 Fundamentals and Good Practice in Machine Learning

It stands to reason to apply machine learning to tasks, which are beyond human capabilities or which require highly repetitive action (Shalev-Shwartz & Ben-David, 2014). This includes very large and complex data sets, which can be handled with current computational power. Generally, machine learning algorithms can be subdivided into supervised and unsupervised learning methods. Unsupervised methods learn from patterns in the input data set, oblivious to response variables (Shalev-Shwartz & Ben-David, 2014). These are for example clustering algorithms or compression algorithms (Hinton, Sejnowski, Poggio, & others, 1999). Supervised methods learn from the input data, while being informed about the response data of a training set. Supervised machine learning methods include SVM, RF, ANN, or decision trees (Kubat, 2017). This said, most machine learning algorithms can be applied both in supervised and unsupervised form. In supervised learning, one can discriminate between classification and regression algorithms. Regression algorithms predict continuous variables, while classification algorithms predict dichotomous or multicategory data, i.e. discrete classes.

1.5.2.1 Model Evaluation

Several simple metrics exist to characterize the performance of regression learners, such as the root mean square error or the R^2 (Kubat, 2017).

Classification learners are characterized by metrics such as accuracy, precision, or recall, defined by the contingency matrix (Figure 1.4).

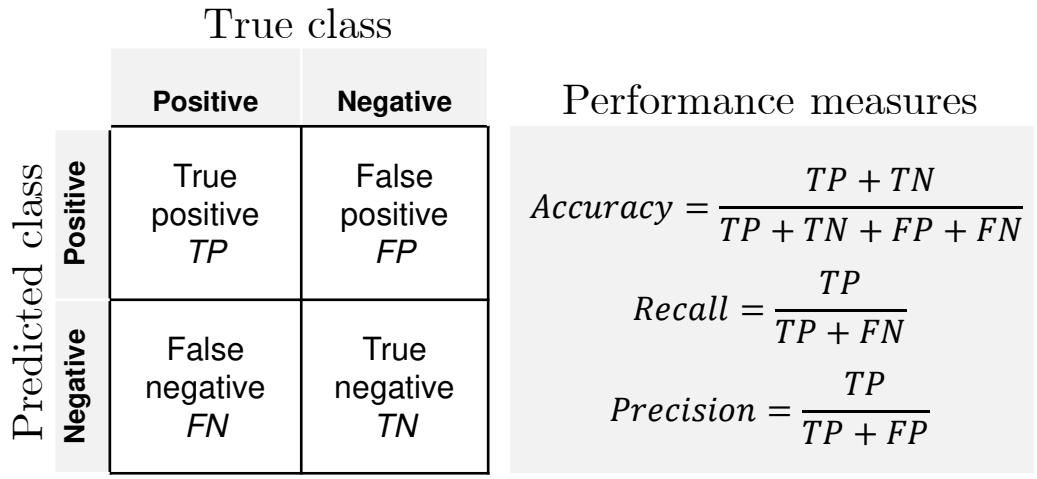


Figure 1.4: Contingency matrix for a binary classification problem. The derivation of accuracy, recall, and precision from the contingency matrix is shown.

Accuracy defines the percentage of correctly predicted classes. The recall defines the number of true positives correctly predicted of all positives and the precision describes the ratio of correctly predicted positives to all predicted positives. These measures are useful, but prone to bias for situations of class imbalance (Powers, 2011). Assume a situation, where 90% of training samples are negative and 10% are positive. The model could be the evaluation of a test for a disease, where it is very important to identify the few positives of the suspected ill. If trained on the accuracy, the model could favor solutions where all observations are predicted negative, since this would already lead to an accuracy of .9. This can be regarded as a failure of the algorithm, as positive samples are completely ignored. An alternative metric is Matthew's correlation coefficient

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}, \qquad (1.8)$$

where, similar to the accuracy, the entries of the contingency matrix are used for calculation. MCC ranges from -1 to +1, where +1 is perfect prediction, -1 is complete disagreement between prediction and reality, and 0 represents random prediction. It is considered the least biased singular metric for binary classification evaluation (Chicco & Jurman,

2020; Powers, 2011). The MCC for the example above is not defined since $TP + FP = 0$. If we assume a closely related case, in that one of the observations is predicted as TP, and the rest negative, this results in 90 observations TN, and 9 observations FN. An MCC of 0.31 and an accuracy of .91 would result. If we now consider a case, where nine out of the ten positive cases are identified, but 20 negative cases are predicted positive, the accuracy would decrease to .79, but the MCC would increase to .45. Compared to the previous example, the model deteriorates with regard to accuracy but improves, when the MCC is the evaluation metric. The prediction of 20 patients to have an illness they actually do not have could be mitigated by further analyses. Missing actually sick patients, however, could be fatal. Therefore, the MCC is the better metric for this example. Generally, because it is less prone to class imbalance, it is best practice to utilize MCC as a single performance evaluation metric instead of accuracy.

1.5.2.2 Bias-Variance Dilemma and Ensemble Learners

The design of a model always bears the potential of over- and underfitting. This is also referred to as the bias-variance dilemma (Geman, Bienenstock, & Doursat, 1992). A biased model ignores much of the training data, and is therefore underfitted. This model results in small differences when trained with different training sets. A model that includes a lot of information of the training set is an accurate predictor for the training set, but shows high variance with respect to other training data. It is overfitted. The optimal model lies in between but is often difficult to ascertain. There are different approaches to reduce variance or bias, one of which involves the combination of several learners into a single output model (Re & Valentini, 2012). These ensemble methods aggregate the prediction of individual models, which are often either biased or show high variance, so-called 'weak learners'. For example, this method was applied to genetic programming and improved the model with regard to its variance (Keijzer & Babovic, 2000). Classification and regression tree performance has been improved significantly with regard to variance, when ensemble methods were applied (Breiman, 1998; Schapire, 1999). In bioinformatics, ensemble learning was competitive with or superior to other machine learning methods for gene function prediction (M. Re & G. Valentini, 2010).

1.5.2.3 Data Pre-Treatment

Next to data evaluation, proper pre-treatment of the data is of utmost importance. There are sophisticated ways to deal with missing data, transform or scale data, select features, or reduce the dimension of the problem otherwise prior to building the actual model (Kubat, 2017). When designing a machine learning study, it is paramount to split the data into a training and a test set. The latter is not involved in model generation, but may only be used for evaluation of the model (Kubat, 2017). A larger training set benefits model construction but limits the evidence given by testing the model with the test set. To evaluate different models, cross-validation using the training data is useful. Validation methods include leave-one-out, k-fold, and Monte Carlo cross-validation (Shalev-Shwartz & Ben-David, 2014; Smyth, 1996).

While this section barely covers the vast field of machine learning, it gives an impression on how bioprocessing could benefit from these data-driven approaches, given the studies are well-designed, both from an experimental and a statistical point of view.

2

Thesis Outline

2.1 Research Proposal

Virus-like particles (VLPs) are an emerging class of vaccines, which is applied and investigated for the prevention or treatment of infectious diseases, immunological disorders, and cancer. VLPs are composed of viral proteins, resemble the virus they are derived from, but lack its infectious nucleic acids. Their repetitive surface and particulate structure are key to the strong immunogenic responses that VLP vaccines can induce. However, their complexity also poses challenges for VLP molecular design and process development. In the development of chimeric VLPs (cVLPs), which are VLPs presenting foreign antigenic epitopes, a significant fraction of candidate molecules is found to be insoluble or incapable of capsid assembly. The process to identify viable candidates is still largely empirical and therefore laborious. For purification of VLPs, technologies originally developed for much smaller biopharmaceutical molecules, such as antibodies, are utilized. However, the large size of VLPs can be challenging, for example with regard to their limited diffusion compared to smaller molecules or due to pore size limitations. Additionally, VLP processing adds process steps, which are not yet part of an already established platform process, calling for development and optimization of tailored methods. Important examples are the dis- and reassembly of VLPs, which are typically achieved by titration or buffer exchange by dialysis.

With limited experience available, tackling these challenges can significantly benefit from a data-driven approach, for example by implementing process analytical technology (PAT), data scientific methods, such as machine learning, or molecular modeling. The objective of this research is to implement data-driven approaches to advance the process development of VLPs, especially for VLP-specific processing challenges.

Regulatory authorities encourage the implementation of quality-by-design, which implies building quality into the process instead of testing it into the product. PAT is widely acknowledged as an important contributor to accomplish this goal. While many approaches for PAT in biopharmaceutical processing have been described, VLP processes generally lack PAT implementation. Additionally, VLP processing includes process steps that are unique, such as VLP dis- and reassembly. VLPs are disassembled by increasing pH and concentration of chaotropic solutes by titration, for example. The reassembly of VLPs is typically realized by dialysis exchanging the buffer for a neutral pH and high ionic strength buffer, thus removing the chaotrope. Higher yield and lower buffer consumption can be achieved by applying cross-flow filtration (CFF) to VLP assembly. However, no approach exists that implements PAT into the (re-)assembly process. In the first study of this thesis, CFF will be implemented to realize reassembly of different cVLPs based on hepatitis B core antigen (HBcAg). The goal of this study is, firstly, to identify the impact of process parameters, such as the transmembrane pressure (TMP), on the product yield and degradation. Secondly, the implementation of two different spectroscopic sensors into an on-line loop as PAT tool to monitor VLP formation will be realized. Lastly, factors affecting the required ionic strength for reassembly of the different species will be described. The questions that should be answered with this study are whether the VLP reassembly process by CFF can be monitored and potentially controlled with the integrated sensors and how mechanical process parameters and the different investigated cVLP constructs affect the assembly reaction.

Since the inserted amino acids in the different cVLPs are on the surface of the capsid, the physicochemical surface properties of cVLPs are probably relevant for processing, for example during capsid reassembly. It would be valuable to predict cVLP surface properties, such as surface

charge, to narrow down the design space for process development. The prediction of surface properties requires three dimensional (3-D) structures. For new molecules, these are not available and must either be generated experimentally, which is laborious, or predicted computationally based on the amino acid sequence. The number of investigative molecules and therefore structures is high in early vaccine development. It would therefore be beneficial to create a high-throughput and automated structure preparation workflow. Thus, the second study in this thesis aims to develop such a workflow for HBcAg dimers and to evaluate its potential with a case study on HBcAg surface charge. The predicted surface charge will be correlated with experimentally derived zeta potential data of entire VLP capsids. This study should evaluate whether an automated and data-dependent workflow can increase robustness of feature extraction with reasonable required computational resources, while allowing to extract meaningful surface data for correlation with experimental results.

The size of VLPs is much larger than other typical biopharmaceutical molecules, such as monoclonal antibodies. Their large size poses challenges for purification, as it for example limits the capacity in bind-and-elute chromatography, the work horse in traditional biopharmaceutical downstream processing. This said, VLP separation would benefit from their increased size when applying size-selective methods. These include filtration, size exclusion chromatography (SEC), and precipitation/re-dissolution. The third study will investigate the integration of these technologies for capture and purification of an HBcAg VLP. This will be realized by a sequence of precipitation, purification, and re-dissolution of VLPs on a CFF unit. Subsequently, the re-dissolved product will be purified by a multimodal SEC (mmSEC) in flow-through mode. These methods will be integrated into one unit operation. Consistent with the first study, process monitoring will be implemented, in order to make decisions on the fractionation of the purified product stream. The main hypothesis behind this study is that the combination of several size-selective methods in one unit operation is a powerful approach to VLP purification, which should be evaluated with the example of precipitation/re-dissolution, CFF and SEC.

While in this study, VLP solubility was decreased artificially using the precipitant ammonium sulfate, cVLP solubility during expression in

hosts, such as *Escherichia coli*, is a significant challenge. The process of identifying soluble and assembly-competent vaccine candidates is largely empirical and would benefit from a predictive toolbox. In the fourth study, a predictive solubility model based on a large dataset of several hundred VLP candidates' solubility and amino acid sequence data will be developed. Since solubility is strongly affected by the molecules' hydrophobicity, the implementation of hydrophobicity scales in a machine learning framework will be investigated as a cVLP solubility model. Additionally, statistical analysis of the models will be applied to reveal characteristics of the data to better understand the mechanisms behind VLP solubility. They hypotheses behind this study are I) that hydrophobicity scales are useful tools for solubility prediction, II) that feature selection is a potent tool to select hydrophobicity scales for a solubility model, and that III) a simple and interpretable machine learning model can help extract hidden information from the data.

In a final, fifth study, the potential to optimize the developed machine learning approach will be investigated. Characteristics of the resulting data of the preceding study will be used to fine tune the model. Additionally, an algorithm for synthesis of hydrophobicity scales specifically for the VLP solubility problem will be developed. This algorithm could serve to build better models and learn about the importance of certain amino acids with regard to their contribution to cVLP solubility. Precipitation of VLPs, as investigated in the third study, occurs through hydrophobic interaction. The concentration of ammonium sulfate required to precipitate cVLPs is therefore probably related to their hydrophobicity. Thus, the model will be redesigned to serve as a regression tool to predict precipitating ammonium sulfate concentrations from the cVLP amino acid sequences in a case study. The questions this study should answer are I) whether the sEVC solubility model performance can be boosted with optimization strategies, II) whether well-performing amino acid scales can be generated using this worklow, and III) whether the sEVC framework can be redesigned to function as a regression model for other biophysical parameters.

2.2 Outline and Author Statement

In several of the following manuscripts, first authorship was shared (contributed equally) among colleagues and me. This was undertaken to elevate the quality of our common publication. A detailed listing of author contributions signed by the respective authors is added as a separate supplement to the examination copy.

Chapter 3: Process Monitoring of Virus-Like Particle Reassembly by Diafiltration with UV/Vis Spectroscopy and Light Scattering

Matthias Rüdt*, Philipp Vormittag*, Nils Hillebrandt, Jürgen Hubbuch

* contributed equally

Biotechnology and Bioengineering (2019), Volume 116, Pages 1366-1379

In Chapter 3, a set-up for process monitoring of VLP reassembly by CFF was developed. Three different cVLPs were reassembled at three different TMPs. The goal of this study was to implement two different spectroscopic sensors to monitor VLP assembly. A UV/Vis absorbance spectrometer was implemented to monitor concentration and the hydrophobic environment of tyrosine and tryptophan. A light scattering photometer provided with SLS and DLS data, informing about the quaternary structure of the VLPs. The combination of the sensors allowed to differentiate between HBcAg dimers, VLPs, and aggregates.

The implementation of this elaborated measurement loop was undertaken to learn about the VLP reassembly by CFF and to lay the groundwork for monitoring and control of this process. The VLPs are faced with a high-ionic strength environment during assembly, which led to aggregation and product degradation for long process times, as could be observed with all sensors and derived metrics. Additionally, the evaluation of three different cVLPs could grant an insight on the influence of the inserted epitope for the VLP reassembly process. For example, the strongest charged HBcAg construct required higher ionic strength to reach the maximum VLP concentration. The evaluation of the process data allowed to identify three different phases for the assembly of HBcAg VLPs in a CFF unit. After an initial lag-phase, the second phase describes VLP assembly until a maximum was reached, which was followed by a degradation phase. The identification of this third phase is paramount in process control, as it has to be avoided for maximum yield and product quality.

This study was a cooperation between my colleague Matthias Rüdt and me. While his focus was the implementation of light scattering and absorbance

spectroscopy sensors for measurement of protein quaternary structure on a CFF unit, my focus lay in developing a PAT method for monitoring and potential control of the VLP reassembly process and to learn about the influence of mechanical process parameters as well as the biophysical properties of the cVLPs on the reassembly process.

The experimental and theoretical work behind this study was extensive and benefitted greatly from the collaboration of my colleague Matthias Rüdt and me. Our joint contribution to this study includes a literature review, experimental realization of the processes together with our master student Nils Hillebrandt, optimization of the CFF set-up, the online measurement loop, selection of the sensors, analysis of the data, critical discussion of the data, graphical illustration, and drafting and revising the manuscript. While my focus in the literature review lay on the VLP assembly process, both from a theoretical and practical point of view, Matthias Rüdt's focus were general PAT methods and the implementation of these in the process environment. My colleague and I chose the specific PAT methods in joint discussion.

The development of the reassembly process required establishment of the expression of different cVLPs, their purification, and development of the actual CFF reassembly process, which was conducted by me. Matthias Rüdt's focus in realizing this research was the implementation of multimodal spectroscopic tools in a software framework to retrieve data on the CFF process. He contributed majorly to software programming, accessing and treatment of the data of the different sensors, and implementation of the PLS model. The construction and optimization of the CFF set-up was conducted both by Matthias Rüdt and me. During the master thesis of Nils Hillebrandt, who worked on this project, we optimized the on-line measurement loop, so that pressure pulsations, and air bubbles – both deteriorating measurement quality – could efficiently be reduced. Measures included the implementation of a glass fiber depth filter to trap bubbles and large particles and the inclusion of a flow restrictor. These pre-experiments required lengthy preparations by all three parties to establish the knowledge required to conduct the final nine processes for the paper.

The joint realization of this project allowed Matthias Rüdt to learn about the implementation of multimodal spectroscopy for the assessment of tertiary and quaternary protein structure changes. It allowed me to learn

about the three technical phases of assembly, the inhibition of assembly by aggregation, the dependency of assembly on the VLP zeta potential, and the possibility to use the established PAT set-up as a control tool for VLP reassembly processing.

Chapter 4: High-Throughput Computational Pipeline for 3-D Structure Preparation and In Silico Protein Surface Property Screening: A Case Study on HBcAg Dimer Structures

Marieke Klijn*, Philipp Vormittag*, Nicolai Bluthardt, Jürgen Hubbuch

* contributed equally

International Journal of Pharmaceutics (2019), Volume 563, Pages 337-346

In Chapter 4, an automated, high-throughput computational pipeline was developed, which creates refined 3-D structures from an amino acid sequence input and which was evaluated with a case study on HBcAg surface charge. The knowledge on surface properties of VLP candidates can be indicative of developability and can narrow down the design space for processes. Surface properties can be assessed by the analysis of 3-D structures, which are not known in early development. Therefore, a computational method including template structure retrieval, homology modeling, and molecular dynamics (MD) simulation was developed.

The main outcomes of the study were I) that the automated workflow has the potential to significantly speed up 3-D structure generation, II) that the workflow could be run on desktop computers and did not require computer clusters, III) that the derived surface charge of the HBcAg dimer 3-D structures correlated with experimentally measured VLP zeta potential, and IV) that the evaluation of 1000 simulation snapshots resulted in more robust feature data than evaluation of a single simulation end-point.

This study was a cooperation between my colleague Marieke Klijn and me. While her focus was the development of a high-throughput (HT) computational pipeline for structure curation and preparation for consistent evaluation of 3-D molecular features, my focus lay in extending this workflow by a data-dependent multi-step MD simulation to refine homology modelled 3-D structures.

Selection of the VLPs for this study was done by me. The design of the curation and preparation workflow was mainly executed by Marieke Klijn. The actual application to the HBcAg molecules was done by

Nicolai Bluthardt. I designed the 3-step MD simulation workflow, which evolved from pre-experiments. In various literature studies a fixed simulation time is employed, e.g. 30-100 ns (simulated time). During pre-experiments, I observed that the structural fluctuations, indicated by the root mean square deviation (RMSD) of atom coordinates, were very diverse at fixed time points for different HBcAg constructs. This was the case, even though only few amino acids were different between the molecules. Therefore, I programmed a MATLAB script, which terminated the MD simulation after a stability criterion was reached. Additionally, I implemented a 3-step simulation approach, in which a continuously larger part of the molecule was simulated, while the positions of the other atoms were constrained. This allowed to focus computational power on the structurally most uncertain regions. These regions were the inserted epitopes and adjacent amino acids, whereas the conserved region of the HBcAg template required less structural refinement.

Marike Klijn and I selected the zeta potential as a case study, since it is a VLP experimental property of the whole capsid, while the surface charge was retrieved from the structural subunit of a dimer. A correlation of these data would allow VLP researchers to create the much simpler HBcAg dimer 3-D structure instead of the entire capsid structure to estimate VLP zeta potential, thus saving computational resources. I was responsible for the required production and measurement of pure VLPs, which was needed to retrieve the experimental data. Marieke Klijn and I analyzed the data, drafted the manuscript, created graphical illustrations, and revised the manuscript critically. All authors read and approved the final manuscript.

In summary, this project allowed Marieke Klijn to realize a HT computational 3-D structure generation pipeline, which is characterized by standardized, automated, and parallelizable workflows, allowing for consistent and robust 3-D structure generation for feature extraction. This project resulted in a computationally inexpensive workflow that allowed me to create HBcAg capsomer 3-D structures and to demonstrate that a correlation between *in silico*-derived surface charge and experimental VLP zeta potential exists, building the foundation for extraction of different features and the application to larger datasets.

Chapter 5: Integrated Process for Capture and Purification of Virus-Like Particles: Enhancing Process Performance by Cross-Flow Filtration

Nils Hillebrandt*, Philipp Vormittag*, Nicolai Bluthardt, Annabelle Dietrich, Jürgen Hubbuch

* contributed equally

Frontiers in Bioengineering and Biotechnology (2020), accepted article

The integration of three size-selective separation techniques for VLP capture and purification was investigated in Chapter 5. The large size of VLPs poses challenges for many traditional biopharmaceutical DSP unit operations, such as in bind-and-elute chromatography. Other techniques, such as precipitation/re-dissolution, filtration, and SEC can benefit from the large size difference between VLPs and impurities. Therefore, a process was developed that started by precipitating product, washing and re-dissolving the product in diafiltration mode on a CFF unit, and optionally leading the product-containing permeate stream through an mmSEC column, binding residual impurities. The inclusion of this mmSEC column led to the best purities combined with high yield and productivity. The permeate stream was monitored with an ultraviolet (UV) absorbance sensor of a chromatography system, which also integrated a fractionator, allowing for UV-based pooling decisions of the collected permeate fractions.

This study showed that I) the integration of three size-selective techniques results in high purities compared to literature processes and a centrifugation-based reference process, II) productivity and yield was higher than the reference process, and that III) data-dependent process control can be realized with the UV-monitored permeate stream, optimizing the output stream of this unit operation with regard to concentration and purity. Since VLPs share the attribute of large size, this process could be the foundation for a platform process for non-enveloped VLPs.

To enhance the quality of this research, this project was conducted together with my colleague Nils Hillebrandt. The study was designed

mutually throughout our joint work on VLP CFF processes. Nils Hillebrandt hereby worked on the implementation of a flow sensor for monitoring and control of the permeate flow rate, while I lay my concentrated on the precipitation behavior of cVLPs with regard to the CFF system. My main focus in this study was to evaluate the hypothesis that VLP purification can benefit from the combination of size-selective methods in one unit operation, while Nils Hillebrandt's focus rather lay on extending the typical mode of application of CFF.

The processes required substantial preparations, which were only realizable by the contributions of both Nils Hillebrandt and me, while we were supported in parts by our student Annabelle Dietrich. She prepared buffers, drew and partly analyzed samples, and helped with the system set-up. Prior to the published processes, several pre-experiments were carried out to optimize the set-up, which was done both by Nils Hillebrandt and me. Analytics were performed by Nils Hillebrandt and me, except for high-throughput capillary gel electrophoresis (HT-CGE), which was performed by Nicolai Bluthardt. Evaluation of the analytical results was done both by Nils Hillebrandt and me. While Nils Hillebrandt lay his focus on the automated time-alignment of on-line and off-line measurement data, I focused on the individual evaluation of the SEC, multi angle light scattering, and HT-CGE measurement data. The largest part of the evaluation was the consolidation and interpretation of all measurement results in the context of the aligned on-line data. This was done in collaborative work between Nils Hillebrandt and me. We interpreted the data in the context of the developed set-up and its utility for VLP DSP processes. Nils Hillebrandt and I drafted the manuscript, created the graphical illustrations, and revised the manuscript critically. For the benefit of the readers, we created a detailed supplementary material section to enable reproduction of our set-up. All authors read and approved the final manuscript.

Overall, the realization of this project allowed me to evaluate an integrated CFF-based precipitation and re-dissolution process for capture and purification of VLPs. This process utilizes the – in bind-and-elute chromatography problematic – large size of VLPs for efficient separation. Since the separation was mainly by size, its transferability to other VLPs seems straightforward and may therefore advance VLP process development into the direction of a platform process.

Chapter 6: Ensembles of Hydrophobicity Scales as Potent Classifiers for Chimeric Virus-Like Particle Solubility - an Amino Acid Sequence-Based Machine Learning Approach

Philipp Vormittag, Thorsten Klamp, Jürgen Hubbuch

Frontiers in Bioengineering and Biotechnology (2020), Volume 8, Article 395, Pages 1-15

Chapter 6 describes the establishment of a machine learning tool to predict VLP candidate solubility. The machine learning tool is a soft ensemble vote classifier (sEVC), which is based on individual one-level decision trees. The decision trees are trained on features from training data to predict test data. The entire dataset comprised 568 different HBcAg cVLP amino acid sequence and binary solubility data. The derived features were hydrophobicities calculated from the amino acid sequences and 91 different hydrophobicity scales. The models performed comparably or better than reported literature performance data of other solubility models. The simple architecture allowed to preserve the physicochemical information contained in the hydrophobicity scales largely. Thus, interpretation of the model led to the proposition of an arginine-mediated recruitment of HBcAg subunits during VLP assembly. While the experimental data was provided by Thorsten Klamp, the establishment of the model framework, the theoretical work, the statistical evaluation, drafting of the manuscript, and creating the graphical illustration was done by me. All authors read, critically revised, and approved the final manuscript.

In detail, the establishment of the classification model required a literature review of machine learning in general, machine learning applied to biopharmaceutical processes, existing solubility models, hydrophobicity scales, hydrophobicity in general, and VLP structural behavior. The implementation of the model was realized with a custom-written MATLAB code. In order to understand and characterize the model, a thorough study of model parameters, such as the training set size, and the number of included decision trees was investigated in a great number of randomized experiments. The design of the framework would

allow to apply it to other molecules or biophysical parameters. With minor adaptations, it could also be used for regression.

Chapter 7: Optimization of a Soft Ensemble Vote Classifier for the Prediction of Chimeric Virus-Like Particle Solubility and Other Biophysical Properties

Philipp Vormittag, Thorsten Klamp, Jürgen Hubbuch

Frontiers in Bioengineering and Biotechnology (2020), Volume 8, Article 881, Pages 1-17

In chapter 7, the established solubility prediction framework of chapter 6 is optimized with two different strategies and modified to serve as a regression tool. The dataset of 568 cVLPs, investigated both in chapter 6 and 7, is created by all possible combinations of 71 different inserts and 8 different insertion strategies. An insertion strategy defines, where in the HBcAg molecule the foreign epitope is inserted and which amino acids are deleted. Systematic misclassification based on these insertion strategies was observed and served as the basis for a first optimization strategy. This optimization algorithm identified systematic misclassification and adjusted the prediction of the model accordingly in an iterative process. A second optimization algorithm modified or synthesized amino acid hydrophobicity scales to better model the solubility of the training data set, resulting in better model performance on the external test set. Lastly, synthesized hydrophobicity scales were used in a modified model for regression of ammonium sulfate precipitant concentration data required for precipitation of ten cVLPs.

For this study, the same experimental data set as in chapter 6 was used, which was provided by Thorsten Klamp. The establishment of the optimization algorithms within the model framework, the theoretical work, the statistical evaluation, drafting of the manuscript, and creating the graphical illustration was done by me. All authors read, critically revised, and approved the final manuscript.

In detail, the establishment of the optimization algorithms required a literature review of machine learning in general, machine learning applied to biopharmaceutical processes, existing solubility models, hydrophobicity scales, hydrophobicity in general, VLP structural behavior, and optimization algorithms in general. A number of pre-experiments was required to determine the optimal optimization parameters to avoid early optimization termination, e.g. in a local

minimum. Repurposing of the model to work as a regression tool was achieved by utilizing aggregated decision tree child node probabilities as a continuous prediction value. The implementation of the optimization algorithms and the regression model was realized with a custom-written MATLAB code.

3

Process Monitoring of Virus-Like Particle Reassembly by Diafiltration with UV/Vis Spectroscopy and Light Scattering

Matthias Rüdt[a,*], Philipp Vormittag[a,*], Nils Hillebrandt[a], Jürgen Hubbuch[a,**]

[a] Institute of Process Engineering in Life Sciences, Section IV: Biomolecular Separation Engineering, Karlsruhe Institute of Technology, Fritz-Haber-Weg 2, 76131 Karlsruhe, Germany

* Contributed equally

** Corresponding author

Abstract

Virus-like particles (VLPs) have shown great potential as biopharmaceuticals on the market and in clinics. Non-enveloped, *in vivo*-assembled VLPs are typically dis- and reassembled *in vitro* to improve particle stability, homogeneity, and immunogenicity. At industrial scale, cross flow filtration (CFF) is the method of choice for performing reassembly by diafiltration. Here, we developed an experimental CFF setup with on-line measurement loop for the implementation of process analytical technology (PAT). The measurement loop included an

ultraviolet and visible (UV/Vis) spectrometer as well as a light-scattering photometer. These sensors allowed for monitoring protein concentration, protein tertiary structure, and protein quaternary structure. The experimental setup was tested with three hepatitis B core antigen (HBcAg) variants. With each variant, three reassembly processes were performed at different transmembrane pressures (TMPs). While light scattering provided information on the assembly progress, UV/Vis allowed for monitoring the protein concentration and the rate of VLP assembly based on the microenvironment of Tyrosine-132. VLP formation was verified by off-line dynamic light scattering (DLS) and transmission electron microscopy (TEM). Furthermore, the experimental results provided evidence of aggregate-related assembly inhibition and showed that off-line size exclusion chromatography (SEC) does not provide a complete picture of the particle content. Finally, a partial least squares (PLS) model was calibrated to predict VLP concentrations in the process solution. Q^2 values of 0.947 to 0.984 were reached for the three HBcAg variants. In summary, the proposed experimental setup provides a powerful platform for developing and monitoring VLP reassembly steps by CFF.

3.1 Introduction

Virus-like particles (VLPs) are biopharmaceuticals with potential applications against various diseases such as viral and bacterial infections, cancer, Alzheimer's disease, and autoimmune disorders (Bachmann & Whitehead, 2013; Klamp et al., 2011; Kushnir et al., 2012; Lua et al., 2014; Middelberg et al., 2011). They are generally designed to trigger an immune response by presenting antigens on their surface. These antigens are either part of the native viral capsid or introduced artificially. Chimeric VLPs were, for example, constructed based on hepatitis B core antigen (HBcAg) (Arora, Tyagi, Swaminathan, & Khanna, 2012; Klamp et al., 2011; Whitacre et al., 2009), hepatitis B surface antigen (Kaslow & Biernaux, 2015), GH1-Qβ(Low et al., 2014), and murine polyomavirus VP1 (MuPyVP1) (Middelberg et al., 2011). VLPs are resilient to most environmental stresses, have great potential to be produced inexpensively, and efficiently elicit potent immune responses due to their repetitive and particulate structure (Chuan, Wibowo, Lua, & Middelberg, 2014; Kumru et al., 2014).

Similar to viruses, VLPs are assemblies of one or several types of capsid proteins forming a higher-order structure (Lua et al., 2014). VLPs are expressed in genetically modified host organisms (Kushnir et al., 2012; Lua et al., 2014; Vicente, Roldão, et al., 2011). Subsequent production-scale purification most frequently consists of precipitation, chromatography, and ultrafiltration/diafiltration (UF/DF) (Ladd Effio & Hubbuch, 2015). In vivo self-assembled, non-enveloped VLPs are often disassembled and subsequently reassembled to remove impurities from within the capsid (Link et al., 2012; Ren et al., 2006). Disassembling and reassembling also leads to increased structural homogeneity, improved overall stability, and enhanced antigenicity (Mach et al., 2006; Q. Zhao, Allen, et al., 2012; Q. Zhao, Modis, et al., 2012). An overview of a typical VLP production process is given in Figure 3.1.

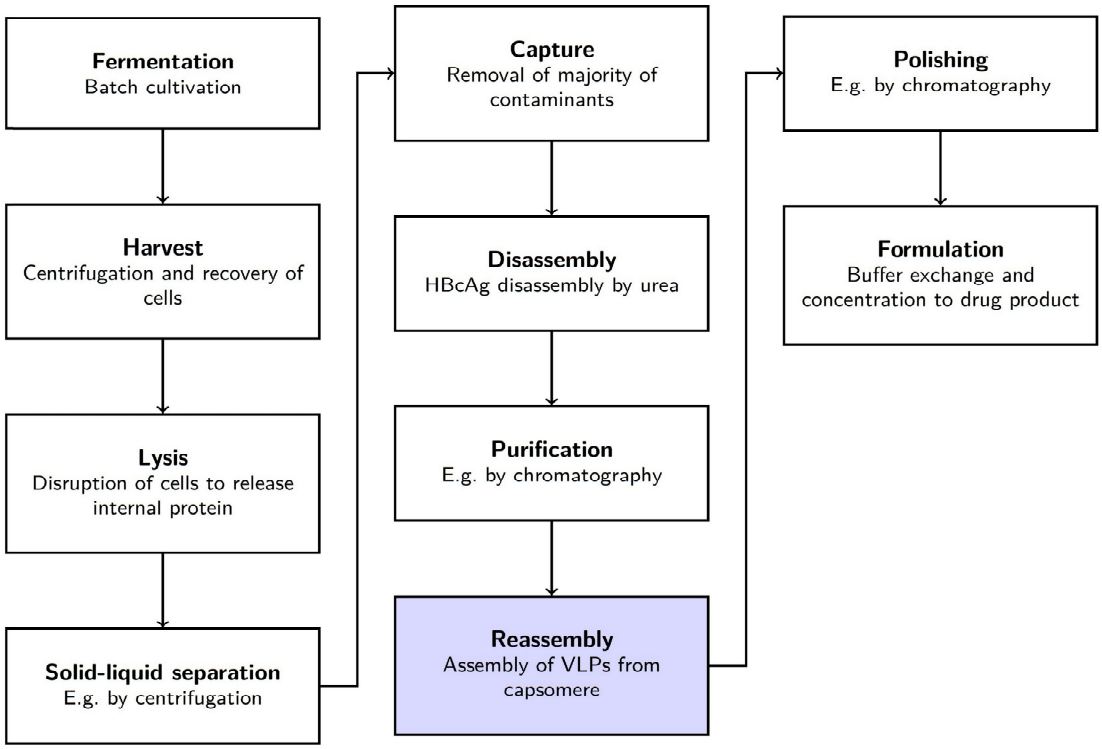


Figure 3.1: Illustration of a typical virus-like particle (VLP) production process. The downstream processing train may consist of eight or more unit operations. The unit operation investigated here – the VLP reassembly – is marked in blue.

Generally, a change in the quaternary structure of virus-like particles (VLPs) is induced by altering their physicochemical environment, i.e. the ionic strength of the protein solution, the pH, or the concentration of a reducing agent (Q. Zhao, Allen, et al., 2012). At lab scale, dialysis is the most common method for buffer exchanges (Mach et al., 2006). Dialysis has, however, some drawbacks such as long processing times and significant buffer consumption (Kurnik et al., 1995). In preparative downstream processes, cross-flow filtration (CFF) is more popular because of its simple scalability, reduced buffer consumption, and reduced processing time (Jornitz & Meltzer, 2008; Kurnik et al., 1995). CFF has been successfully applied to VLPs for capture, buffer exchange, and concentration (Russell et al., 2007; Vicente et al., 2014; Vicente, Roldão, et al., 2011). Compared to dialysis and batch diafiltration, assembly of VLPs by constant volume diafiltration was shown to increase VLP yield (Liew et al., 2012). Despite the many advantages, CFF may also cause problems due to protein-membrane interaction (Hanemaaijer, Robbertsen, van den Boomgaard, & Gunnink, 1989; Ko, Pellegrino,

Nassimbene, & Marko, 1993) which was observed to impact process performance (Peixoto, Sousa, Silva, Carrondo, & Alves, 2007). To reduce these problems, CFF process time has to be minimized while maximizing the process efficiency.

Process analytical technology (PAT) (Bakeev, 2010; Roch & Mandenius, 2016; Rüdt et al., 2017) is thus of interest to monitoring the assembly progress. Protein concentration measurements allow to detect protein adsorption to the membrane. Particle size measurements provide information on the assembly progress of the capsid proteins into VLPs. Previous publications have also reported effects of the VLP tertiary structure on ultraviolet and visible (UV/Vis) and fluorescence absorption spectra (Ausar et al., 2006; M. Fang et al., 2016; Hanslip et al., 2006; Hu et al., 2011; Rajendar et al., 2013). Following a systematic approach to process monitoring, a combination of PAT sensors should be chosen which allows to monitor protein concentration, protein tertiary structure, and protein size.

In this study, we developed a CFF setup consisting of a commercial lab-scale CFF device with a custom-made on-line measurement loop for process analytical instrumentation. The online measurement loop included a light-scattering photometer (dynamic light scattering (DLS) and static light scattering (SLS)) and a UV/Vis absorption spectrometer. DLS allowed for monitoring the mean hydrodynamic diameter of particles. SLS outputs an aggregated scattered-light intensity influenced by the particle concentrations and the diameters. Finally, UV/Vis spectroscopy provided information on the protein concentration and on changes in the tertiary structure by second derivative spectroscopy (W. Jiskoot & Crommelin, 2005). The usefulness of the custom-made setup was tested for monitoring the reassembly of three different chimeric HBcAg variants at three different transmembrane pressures (TMPs).

3.2 Materials and Methods

3.2.1 Experimental Setup

A custom-made setup was developed for the CFF experiments. Figure 3.2 shows the setup as a piping and instrumentation diagram. A KrosFlo KRIIi CFF unit with a modified polyethersulfone (mPES) hollow fiber

membrane module (10 kDa cutoff, 13 cm^2 membrane area) and a 50 mL conical tube retentate reservoir (all Spectrum Labs, Rancho Dominguez, USA) made up the core of the CFF unit. A Topolino magnetic stirrer (IKA Werke, Staufen im Breisgau, DE) ensured homogeneous mixing of the retentate reservoir. A T-piece with injection plug (Fresenius Kabi, Bad Homburg, DE) was inserted into the retentate line as sample port to draw liquid for off-line analytics. The retentate reservoir was modified with two additional polyether ether ketone capillaries to supply the on-line measurement loop with liquid from the process.

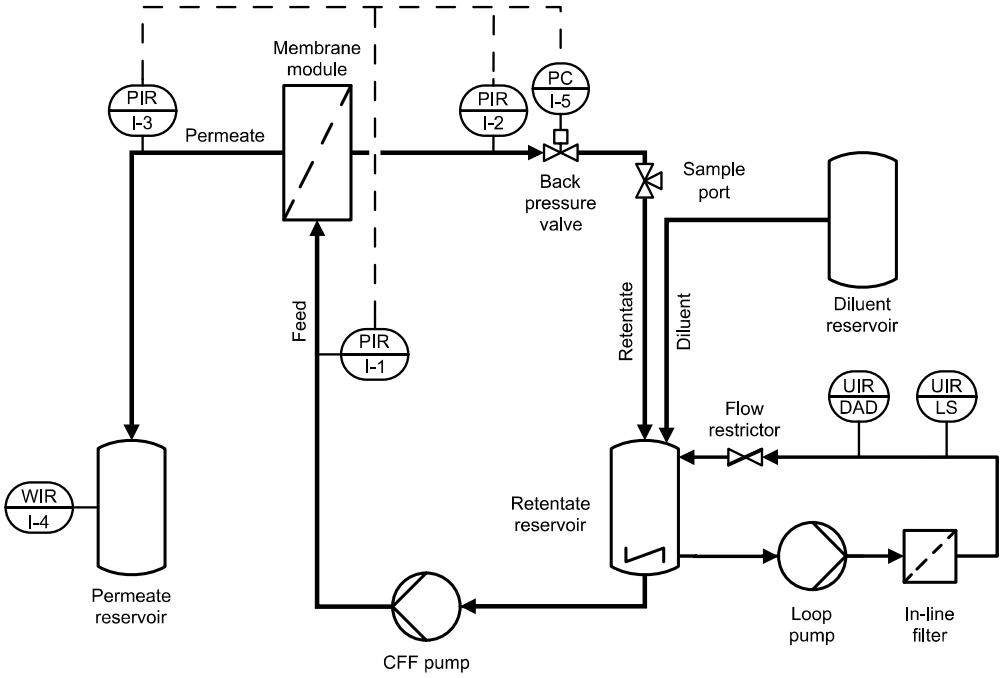


Figure 3.2: Piping and instrumentation diagram of the experimental setup. At the bottom right, the online measurement loop is shown. The remaining piping is required for the cross-flow filtration (CFF) unit. All sensors are connected to a computer for capturing the data centrally. Electronic communication lines are indicated by dashed lines. I-5 is a pinch valve actuated by a closed-loop controller for the transmembrane pressure. The letters indicate: C control, I indicate, P pressure, R record, U multivariable, W weight, DAD diode array detector, LS light scattering.

In the direction of flow, the on-line measurement loop consisted of a Gilson Minipuls 3 peristaltic pump, a 0.7 µm particle retention Minisart glass fiber syringe filter (Sartorius Stedim Biotech, Göttingen, DE), a Zetasizer Nano ZSP photometer (Malvern Instruments, Malvern, GB) with a 10 mm pathlength flowcell (Hellma Analytics, Müllheim, DE), an Ultimate DAD-3000 diode array detector (DAD) (Dionex Corporation,

Sunnyvale, US) with a 0.4 mm pathlength flowcell, and a FR-902 flow restrictor (GE Healthcare, Chalfont St Giles, GB). The pump of the on-line measurement loop was controlled via a NI USB-6008 data acquisition device (National Instruments, Austin, USA).

3.2.2 Proteins, Chemicals, and Buffers

Three chimeric HBcAg constructs , i.e. VLP A, B, and C provided by BioNTech Protein Therapeutics GmbH (Mainz, DE), were used in this study. The HBcAg variants were recombinantly modified in the major immunodominant region (MIR) to display three different peptides on their surfaces (see also Figure 3.3). All variants were present as homodimer stock solutions in disassembly buffer (3.5 M urea, 50 mM Tris(hydroxymethyl)-aminomethane, pH 9.0) as obtained after purification (see also Figure 3.1). Protein concentration calculations were based on extinction coefficients derived from the primary structure as provided by the ProtParam tool (Gasteiger et al., 2005) of the Swiss Institute of Bioinformatics. The purity of the stock solutions was characterized by reversed-phase chromatography based on the absorbance at 280 nm as described in the Appendix A, Supplementary Material S3.2. Immediately before each experiment, stock solutions were diluted with disassembly buffer to a protein concentration of 1 g/L (by ultraviolet (UV) absorbance at 280 nm) and filtered through a 0.2 µm polyethersulfone (PES) filter (VWR International, Radnor, US). The reassembly buffer was a high-salt buffer at pH 7.0.

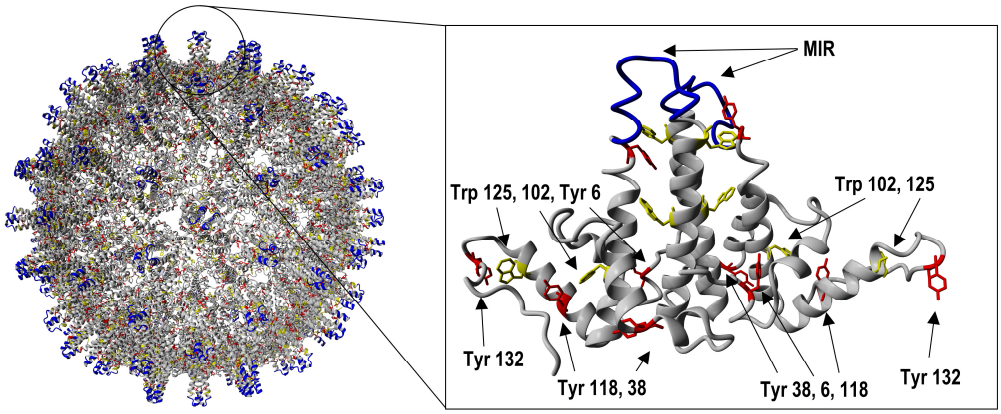


Figure 3.3: An assembled hepatitis B core antigen virus-like particle is shown on the left side (PDB ID 1QGT, *(Wynne et al., 1999)*). The right side shows a cartoon of a single homodimer (adapted from PDB ID 4BMG, (Alexander et al., 2013)). The tryptophan (Trp) and tyrosine (Tyr) side chains are depicted as

sticks and colored in yellow and red, respectively. Tyrosine and tryptophan side chains located in the base of the molecule are numbered. These residues undergo a change of hydrophobicity in their environment during assembly. The major immunodominant region loop, whereto the foreign epitope is inserted, is shown in blue.

For size-exclusion chromatography (SEC), 50 mM potassium phosphate at pH 7.0 was used as running buffer. If not mentioned otherwise, chemicals were purchased from Merck KGaA (Darmstadt, DE). All buffers and solutions were prepared with ultrapure water (arium pro UV, Sartorius, Göttingen, DE) and filtered through a 0.2 µm pore size Supor filter (Pall, Port Washington, US) immediately before each experiment.

3.2.3 VLP Reassembly Monitoring

The CFF unit and the measurement loop were filled with ultrapure water for pre-experimental preparation. The lamps of the DAD were turned on at least 1 h before starting the experiments.

At the end of the equilibration phase, the absorbance signal was zeroed in ultrapure water. Subsequently, the CFF unit and measurement loop were first flushed with disassembly buffer and then changed into 25 mL of protein solution. The CFF pump was set to 70 mL/min corresponding to a shear rate of approximately 6000 s^{-1} in the hollow fibers. The measurement loop pump 1 mL/min and data acquisition were started.

After 5 min, constant TMP diafiltration was initiated by applying a TMP of 0.25 bar, 0.5 bar, or 1 bar with reassembly buffer as diluent. 0.4 mL samples were taken every 0.5 diafiltration volumes (DVs) via the sample port. Experiments were stopped after 3 DV except for VLP C for which the runs had to be terminated early due to membrane clogging. After each run, the CFF membrane was cleaned with ultrapure water, a 0.1 M sodium hydroxide solution, and a 15 vol% ethanol solution.

3.2.4 Off-Line Sample Analysis

For SEC analysis, samples were centrifuged (Centrifuge 5810R, Eppendorf, Hamburg, DE) at 3220 rcf for 5 min to settle large particles. The supernatant was analyzed with a Sepax SRT SEC-1000 column (Sepax Technologies, Newark, US) on an Ultimate 3000 RS ultra high performance liquid chromatography (UHPLC) system consisting of a Pump HPG-3400RS, an Autosampler WPS-3000TFC, a Column

Compartment TCC-3000RS, and a DAD-3000 controlled by Chromeleon version 6.8 SR15 (all Thermo Fisher Scientific, Waltham, US). The run duration was 7 min with a flow rate of 0.8 mL/min and SEC buffer as a mobile phase. 20 µL were injected for each analysis. Samples were analyzed in triplicates.

Off-line DLS analysis was performed using a sample volume of 45 µL in a 3x3 mm quartz cuvette (Hellma Analytics, Müllheim, DE) and the same DLS photometer as mentioned above. Unfiltered samples were measured three times, each measurement consisting of 12 to 14 10 s runs at 25 °C, and 173 ° backscatter. Lower and upper limits for data processing were 1 nm and 6000 nm, respectively. The measurements were compared based on the VLP peak diameter in the regularization fit.

The photometer was also used for electrophoretic mobility measurements of pooled and formulated samples of each construct. The samples of different TMPs were pooled and dialyzed into a pH 7.2 buffer of 50 mM Tris and 100 mM sodium chloride. Samples were filtered with a 0.2 µm PES filter (VWR International, Radnor, USA) and concentration was adjusted with Vivaspin 20 filters with a 30 kDa pore rating (Sartorius, Göttingen, DE). 50 µL of sample was inserted into buffer-filled folded disposable capillary cells (DTS1070, Malvern Instruments Ltd., Malvern, UK) using a diffusion barrier technique (Patent WO2012083272A1). Samples were measured in pentaplicates in automatic mode. Each measurement comprised a 120 seconds equilibration and five runs with up to 15 sub runs. The measurements were performed at 60 mV and 25 °C. Zeta potential was calculated by Zetasizer Software version 7.12 (Malvern Instruments Ltd., Malvern, UK) assuming a material refractive index of 1.45, absorption of 0.001, a viscosity of 0.8872 mPas, a dielectric constant of 78.54, and a Smoluchowski approximation of 1.5 (Smoluchowski, 1921).

The VLPs were furthermore imaged by transmission electron microscopy (TEM) on a Titan³ 80-300 microscope (FEI Company, Hillsboro, US) at 80 kV in bright field mode. For sample preparation, carbon-coated 400-mesh copper grids (Plano GmbH, Wetzlar, DE) were first hydrophilized with a 1% (w/v) alcian blue 8GX (Alfa Aesar, Ward Hill, US) for 2 min and washed 5 times with ultrapure water. Subsequently, the grids were incubated for 2 min with the 0.2 µm filtered 0.3 g/L to 0.5 g/L VLP solutions. The samples were negatively stained with a 1% (w/v)

ammonium molybdate(VI) solution (Acros Organics, Geel, BE) at pH 7.2 for 45 s, washed, and air-dried. VLP diameters were measured with ImageJ 1.52a (NIH, USA). TEM images were processed by adjusting contrast and lightness to improve visibility of the VLP particles using RawTherapee version 5.5 (Gábor Horváth) image processing software.

3.2.5 Data Acquisition and Analysis

During experiments, all integrated sensors communicated with a custom application developed in MATLAB (version R2016b, The Mathworks, Natick, US). Next to starting and stopping measurements, the application gathered the sensor signals (3 pressure signals, the permeate weight, zaverage, and UV/Vis absorbance spectra). Communication and control were performed through software libraries provided by the different instrument softwares. The signals were displayed on the graphical user interface (GUI) and stored on the hard drive with a time stamp. For calculating the permeate volume, the density of the permeate was assumed to be 1 g/cm^3. Data acquisition and analysis of light scattering and UV/Vis measurements were performed as described below.

3.2.5.1 Light-Scattering Measurements

The Zetasizer Nano ZSP was utilized for DLS and SLS measurements using the chromatography flow standard operating procedure of the Zetasizer software (version 7.12, Malvern Instruments). The Zetasizer acquires data in a back-scattering geometry at 173 °. Each measurement duration was 10 s. While DLS measurements were exported on-line, SLS data was extracted off-line. From the DLS measurement, the z-average was obtained as calculated by the Zetasizer software by the method of cumulants (Koppel, 1972). Viscosity (0.8872 mPas), refractive indices (protein 1.45; water 1.33) (as provided by the Zetasizer software), temperature (25 °C), and flow rate (1 mL/min) were assumed to be constant for the calculation of the z-average. The z-average data was subsequently filtered by a moving median over 60 s to remove outliers. The SLS signal was not filtered. The transition from process phase I to process phase II was detected from the scattered-light intensity by the CUSUM algorithm (Grigg, Farewell, & Spiegelhalter, 2003; Page, 1954). The transition from process phase II to process phase III was set at the global maximum of the scattered-light intensity.

3.2.5.2 UV/Vis Absorption Measurements and Processing

During VLP assembly, UV/Vis spectra were continuously acquired at 1 Hz in the spectral range from 240 nm to 340 nm with a resolution of 1 nm. To gain information on the local environment of aromatic amino acids, the spectral data was filtered by a moving average over 30 s and the second derivatives were computed with a Savitzky-Golay filter (Savitzky & Golay, 1964) of order 5 with a 9-point window (Ausar et al., 2006; W. Jiskoot & Crommelin, 2005). An example spectrum with the subsequent data evaluation is shown in Figure S3.1 in the Appendix A. The resulting second-derivative spectra were interpolated with a cubic spline to a final resolution of 0.01 nm. From the interpolated data, the location of the minimum near 292.5 nm was used as a measure of tryptophan solvent exposure (W. Jiskoot & Crommelin, 2005; Mach & Middaugh, 1994). The exposure of tyrosine was assessed based on the a/b-ratio as defined by Ragone et al. (Ragone et al., 1984). Briefly, the vertical distance between trough and peak near 285 nm a was normalized by the trough-peak distance near 294 nm b. The inflection point of the a/b-ratio over time was computed by taking the first derivative with a second-order Savitzky-Golay filter (window width 501 points corresponding to 8.35 min) and finding the minimum.

3.2.5.3 Partial Least Squares Model Calibration

Partial least squares (PLS) model calibration was performed in MATLAB (version 2016a). For each VLP, a PLS model was calibrated based on the UV/Vis spectroscopic data in combination with the off-line SEC VLP concentration. Data of all three TMPs were included into one model. PLS model calibration was performed similarly as described previously (Großhans et al., 2018). The data were first preprocessed and subsequently fitted with a PLS-1 model by the SIMPLS algorithm (de Jong, 1993). For preprocessing, a Savitzky-Golay filter with a second-order polynomial was applied on the spectra and, optionally, the first or second derivative was taken. Cross-validation was performed by iteratively excluding one sample of each CFF run (1/7, resp. 1/6 of the data), calibrating a PLS model on the remaining samples (6/7, resp. 5/6 of the data), and calculating a residual sum of squares on the excluded run. This procedure was repeated until all runs had been excluded once. All residual sums of squares for the different submodels were subsequently accumulated yielding the predictive residual error sum of

squares (PRESS). The PRESS was scaled according to Wold et al. by the number of samples and latent variables used in the PLS model (Wold et al., 2001). Based on the scaled PRESS, an optimization was performed using the built-in genetic algorithm of MATLAB for integers (Deep, Singh, Kansal, & Mohan, 2009). The genetic algorithm optimized the window width of the Savitzky-Golay filter $5 \leq w \leq 21$, the order of derivative $0 \leq n \leq 2$, as well as the number of latent variables for the PLS-1 model $4 \leq N \leq 14$. The root mean square error of cross-validation (RMSECV) was calculated from the PRESS by dividing by the total number of samples. The Q^2 and R^2 values were calculated by dividing the PRESS, respectively the residual sum of squares, by the summed squares of the response corrected to the mean (Wold et al., 2001).

3.3 Results

In this study, a new UF/DF setup with on-line measurement loop was developed to monitor VLP reassembly steps. In the measurement loop, a UV/Vis spectrometer and a light-scattering photometer were integrated. Furthermore, an application was implemented in MATLAB providing a GUI, communication capabilities to the different sensors, as well as a common time base for all performed measurements. This allowed for acquiring and synchronizing measurements in a controlled manner. Within the application, UV/Vis spectra, DLS measurements, pressure, and weight readings were immediately available for processing and display. To demonstrate the advantages of this experimental setup, nine UF/DF runs with three different HBcAg constructs at three different TMPs were performed.

3.3.1 Monitoring of Standard Processes Parameters

During the UF/DF processes, the initial buffer was replaced by reassembly buffer to form HBcAg VLPs from homodimers. In Table 3.1, process data of all runs are summarized (original data presented in Figure S3.2 in the Appendix A). The table also shows that the feed stock purity of VLP A was higher than VLP C and VLP B. At 0.25 bar TMP, VLP A, B, and C showed nearly constant increases in permeate mass over time implicating constant fluxes. The average flux for these three runs was 25.8 L/m²h to 29.1 L/m²h. At 0.5 bar and 1 bar TMP, the

average flux was higher for all three VLPs (from 36.3 L/m²h to 48.7 L/m²h). CFF processes at 0.5 bar and 1 bar TMP showed a decreasing flux over time after an initial constant phase (except for VLP B at 0.5 bar). A decrease in flux at constant TMP indicates the formation of a fouling layer on the membrane (Huisman, Prádanos, & Hernández, 2000; van den Berg & Smolders, 1990).

Table 3.1: Process data is summarized for all performed runs.

	VLP A			**VLP B**			**VLP C**		
TMP / bar	**0.25**	**0.5**	**1**	**0.25**	**0.5**	**1**	**0.25**	**0.5**	**1**
Feed stock purity[a] / %		73.5			22.6			44.1	
Zeta potential[b] / mV		– 7.9(7)			– 11.8(6)			– 9.5(8)	
Total run time / min	118	78	75	133	75	79	108	71	70
Mean flux / ($Lm^{-2}h^{-1}$)	30.5	46.9	48.4	26.8	48.7	45.9	27.6	36.3	40.0
Max. VLP conc. / (g/L)	0.248	0.275	0.250	0.126	0.133	0.116	0.134	0.103	0.126
Inflection a/b-ratio / DV	1.5	0.8	0.7	1.5	1.4	0.7	1.6	0.9	0.6
VLP peak diameter[c] / nm	40(6)	46(11)	42(7)	35(5)	40(11)	46(10)	41(12)	48(5)	36(11)

Note. TMP: transmembrane pressure; VLP: Virus-like particle.
[a] assessed by reversed-phase chromatography as described in the Appendix A, Supplementary Material S3.2.
[b] denotes median and median absolute deviation in parenthesis.
[c] denotes mean and standard deviation of all DLS acquisitions (n=36–42) in parenthesis.

3.3.2 Process Monitoring with On-Line PAT Sensors

In Figures 3.4, 3.5, and 3.6, the on-line PAT sensor measurements as well as SEC off-line analytics are shown for VLP A, B, and C, respectively. All data were plotted over DV indicating the progress of buffer exchange. Each figure shows the absorbance at 280 nm, off-line VLP concentration measurements by SEC, second-derivative spectral analysis, and light-scattering data. It is important to note that an insufficient scattered-light intensity was recorded for VLP C at 1 bar TMP due to an incorrectly set laser attenuation. The corresponding light-scattering results were excluded. The run could not be repeated because of material constraints.

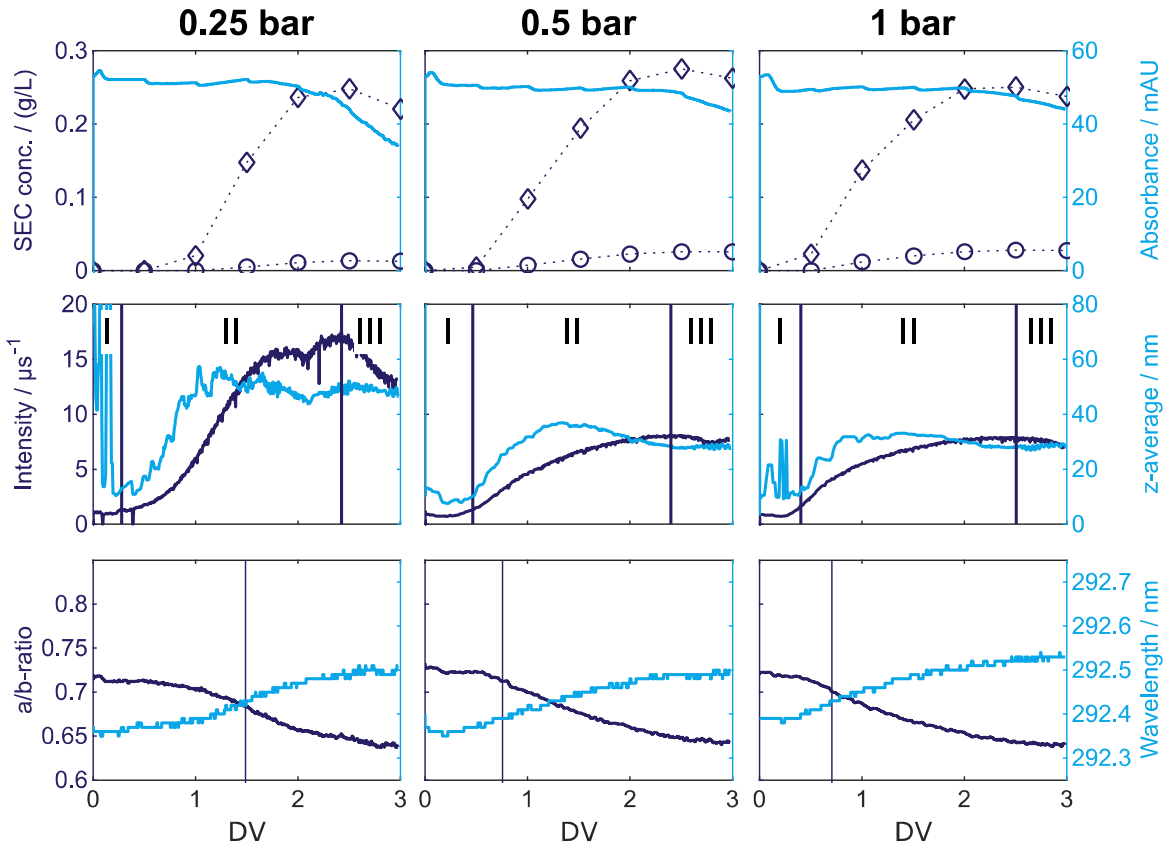


Figure 3.4: The figure displays the on-line sensor measurements as well as off-line analytics against the DV of VLP A. The rows display measurements of different sensors. Top row: Off-line VLP ◊ and aggregate ○ concentration measurements by SEC, UV absorbance at 280 nm –. Middle row: DLS and SLS measurements. Roman numbers indicate the different process phases. Bottom row: Second derivative spectral analysis for tyrosine (a/b-ratio) and tryptophan (location of the minimum around 292.5 nm). The inflection point of the a/b-ratio is marked by a vertical bar. The columns correspond to different TMPs. Left column: 0.25 bar, middle column: 0.5 bar, right column: 1 bar. At 0.25 bar TMP the z-average is corrupted with noise early in the process. DV, diafiltration volume; VLP, virus-like particle; SEC, size-exclusion chromatography; UV, ultraviolet; DLS, dynamic light scattering; SLS, static light scattering; TMP, transmembrane pressure.

Off-line SEC was performed in triplicates resulting in standard deviations smaller than 0.011 g/L. In all runs, the off-line VLP concentration first remained at zero followed by an increase to the maximum VLP concentration. Thereafter, the concentration was approximately constant or decreased slightly. Depending on the TMP, off-line VLP concentration started to increase at 0.5 DV to 1.5 DV. The onset occurred at a DV that was lower the higher the TMP. The maximum observed VLP concentration was between 0.248 g/L and 0.275 g/L for VLP A, between

0.116 g/L and 0.133 g/L for VLP B, and between 0.103 g/L and 0.134 g/L for VLP C. The SEC aggregate content was between 5% and 15% of the VLP concentration.

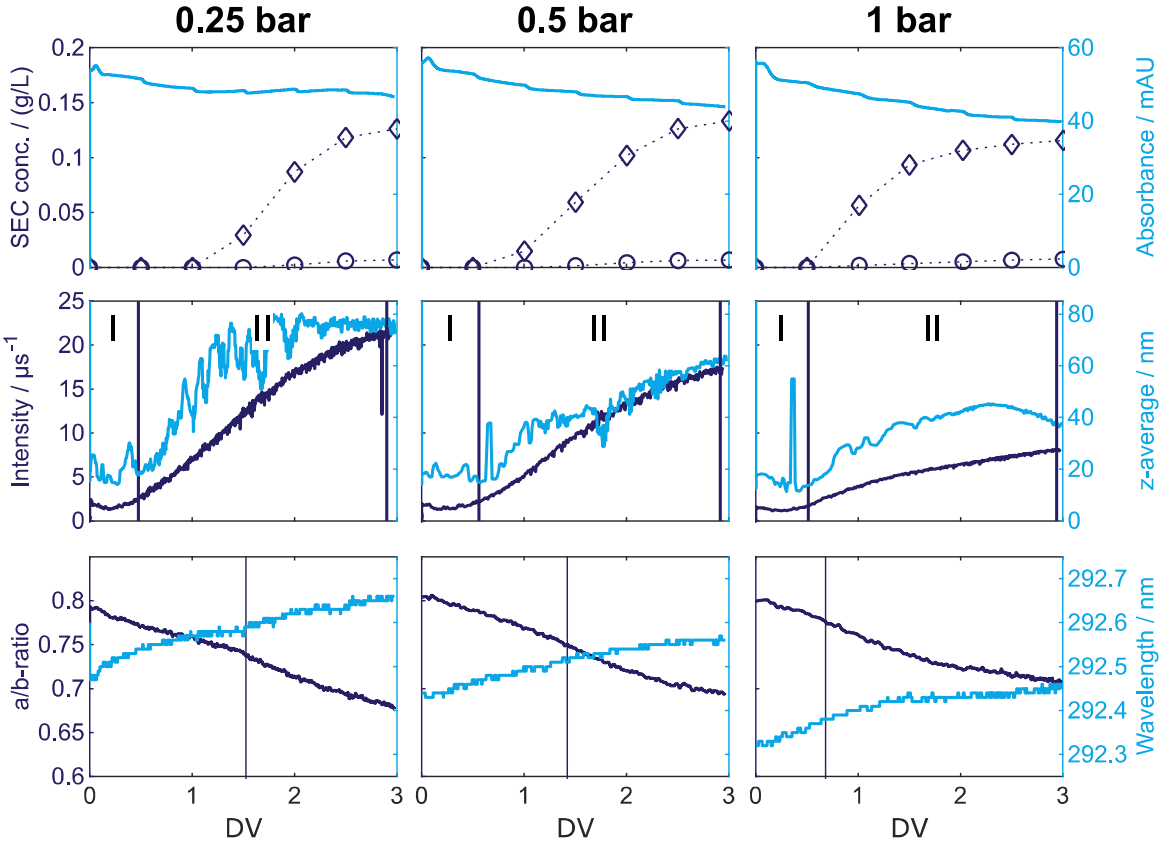


Figure 3.5: The figure displays the on-line sensor measurements as well as off-line analytics against the DV of VLP B. The rows display measurements of different sensors. Top row: Off-line VLP ◊ and aggregate ○ concentration measurements by SEC, UV absorbance at 280 nm –. Middle row: DLS and SLS measurements. Roman numbers indicate the different process phases. Bottom row: Second derivative spectral analysis for tyrosine (a/b-ratio) and tryptophan (location of the minimum around 292.5 nm). The inflection point of the a/b-ratio is marked by a vertical bar. The columns correspond to different TMPs. Left column: 0.25 bar, middle column: 0.5 bar, right column: 1 bar. DV, diafiltration volume; VLP, virus-like particle; SEC, size-exclusion chromatography; UV, ultraviolet; DLS, dynamic light scattering; SLS, static light scattering; TMP, transmembrane pressure.

UV absorbance at 280 nm decreased in all runs over time. Small step-like decreases were due to sampling for off-line analytics. The drawn sample volume was replaced by reassembly buffer resulting in dilution of the process liquid. For VLP A, B, and C, a rapid decrease in the absorbance at 0.25 bar TMP occurred towards the end of the runs, suggesting a loss of protein.

Solvatization of aromatic amino acids and particle formation were observed during CFF by on-line UV/Vis and light-scattering measurements. UV/Vis spectral data were examined by second derivative analysis. From the derived spectra, characteristics were calculated for the solvatization of tryptophan (location of the minimum around 292.5 nm) and tyrosine (a/b-ratio) (W. Jiskoot & Crommelin, 2005). For all runs, a shift towards longer wavelengths of the tryptophan minimum was observed, while the a/b-ratio decreased. Both trends indicated an increase in the mean hydrophobicity around tryptophans and tyrosines. Especially for higher TMPs, the characteristics followed a sigmoidal curve shape. The inflection points of the a/b-ratio in all runs were marked by a vertical line and were located either around 0.8 DV or 1.5 DV (see Table 3.1).

DLS measurements were interpreted based on the z-average. In all experiments, an initial phase of relatively constant z-average values below 20 nm was observed. The second phase was characterized by a rapid increase in z-average to around 40 nm for TMPs of 0.5 bar and 1 bar. At a TMP of 0.25 bar, the second phase showed a larger increase of the z-average to 50 nm to 80 nm. The third phase resulted in relatively constant z-averages over time.

SLS measurements are influenced by the particle diameter and concentration. Similar to the z-average, scattered-light intensities started to increase after an initial constant phase. The increase continued even after the z-average reached a plateau and eventually flattened. For VLP A and C at 0.25 bar TMP, scattered-light intensities rapidly decreased towards the end of the runs.

At 0.5 bar and 1 bar, z-averages, scattering intensities, and SEC VLP concentrations of each run started to increase simultaneously within off-line time resolution. Interestingly, for processes at 0.25 bar, the z-averages and scattering intensities increased earlier than VLP and aggregate concentration by SEC. The initial increase in phase II at 0.25 bar ended at high z-averages > 45 nm, not observed in the other processes. In all runs, the inflection point of the a/b-ratio occurred around the steepest increase in the VLP concentration by SEC.

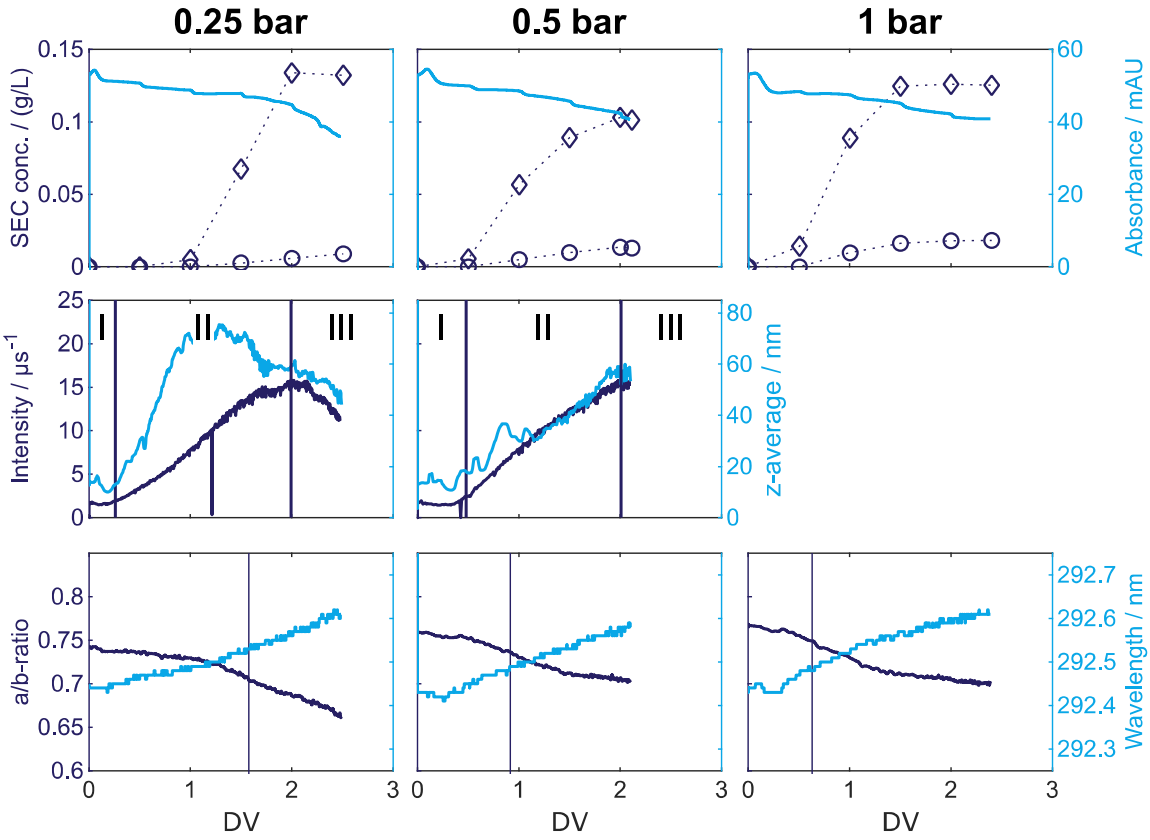


Figure 3.6: The figure displays the on-line sensor measurements as well as off-line analytics against the DV of VLP C. The rows display measurements of different sensors. Top row: Off-line VLP ◊ and aggregate ○ concentration measurements by SEC, UV absorbance at 280 nm –. Middle row: DLS and SLS measurements. Roman numbers indicate the different process phases. Bottom row: Second derivative spectral analysis for tyrosine (a/b-ratio) and tryptophan (location of the minimum around 292.5 nm). The inflection point of the a/b-ratio is marked by a vertical bar. The columns correspond to different TMPs. Left column: 0.25 bar, middle column: 0.5 bar, right column: 1 bar. DLS and SLS measurements at 1 bar were excluded because of an erratically set laser attenuator. DV, diafiltration volume; VLP, virus-like particle; SEC, size-exclusion chromatography; UV, ultraviolet; DLS, dynamic light scattering; SLS, static light scattering; TMP, transmembrane pressure.

3.3.3 Selective Prediction of VLP Concentration by PLS Modeling

The PLS model calibration results are shown in Figure 3.7 and Table 3.2. Figure 3.8 shows the PLS regression coefficients. All PLS models were fitted to the second derivative of the UV/Vis spectral data with 6 to 9 latent variables. The achieved Q^2 values were 0.984, 0.984, and 0.947 for VLP A, B, and C, respectively.

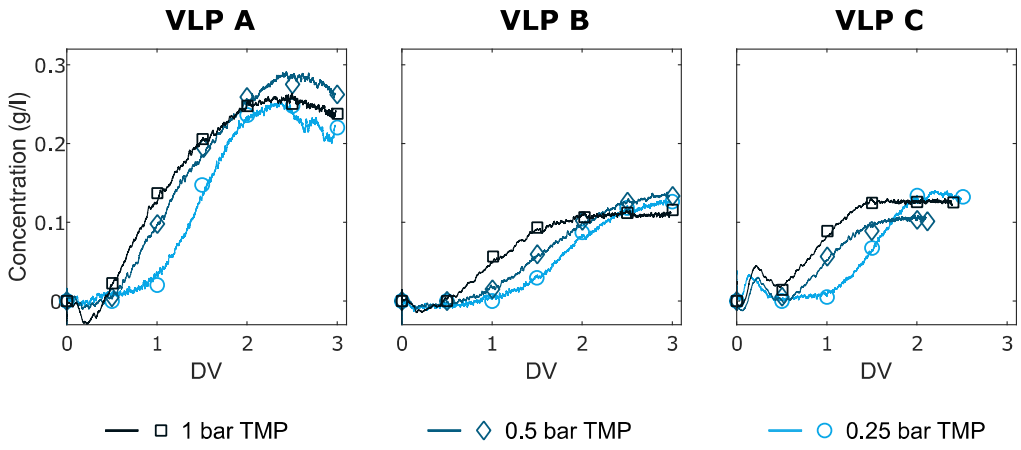


Figure 3.7: A PLS model was fitted to the UV/Vis spectral data for each construct to predict the concentration of assembled VLPs. The concentration estimated by the calibrated PLS model is compared to off-line analytics in the current plot. Each TMP is reflected by a color. The markers show the concentration measured by off-line analytics while the lines correspond to the concentrations estimated by the PLS model. PLS, partial least squares; UV/Vis, ultraviolet and visible, VLP, virus-like particle; TMP, transmembrane pressure.

Table 3.2: Spectral preprocessing parameters, parameters for the PLS model, and the prediction quality of the chemometric models are summarized. PLS, partial least squares, VLP, virus-like particle, RMSECV, root mean square error of cross-validation.

	VLP A	**VLP B**	**VLP C**
No. of samples	21	21	18
No. of cross-validation groups	7	7	6
No. of latent variables	6	9	7
Window Savitzky-Golay filter	7	9	9
Derivative	2	2	2
R^2	0.995	0.997	0.994
Q^2	0.984	0.984	0.947
RMSECV / (g/L)	0.01	0.01	0.01

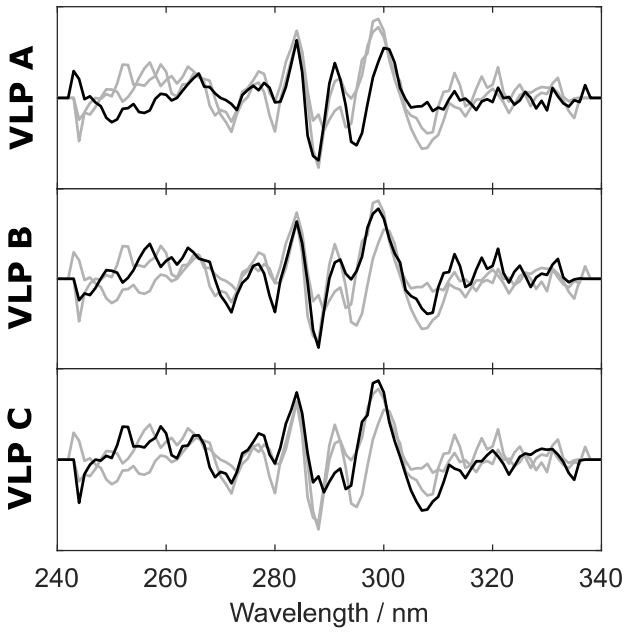


Figure 3.8: Regression coefficients of the three PLS models. Each row corresponds to the regression coefficients of one VLP in black while the other regression coefficients are supplemented in gray. PLS, partial least squares; VLP, virus-like particle.

3.3.4 Analysis of Post-Assembly Samples

Off-line DLS data was measured at the end of all processes. The VLP peak diameter data is shown in Table 3.1. The mean diameter across all runs was 41 nm with a standard deviation of 11 nm. VLP B had the most negative zeta potential with -11.8(6) mV, followed by VLP C with -9,5(8) mV, and VLP A with -7,9(7) mV.

TEM images (Figure 3.9) showed hollow spherical particles with a mean diameter of 33(3) nm, 32(2) nm, and 31(2) nm for the formulated and filtered solution of VLPs A, B, and C, respectively. This result is well in agreement with the DLS measurements and literature data (Crowther et al., 1994).

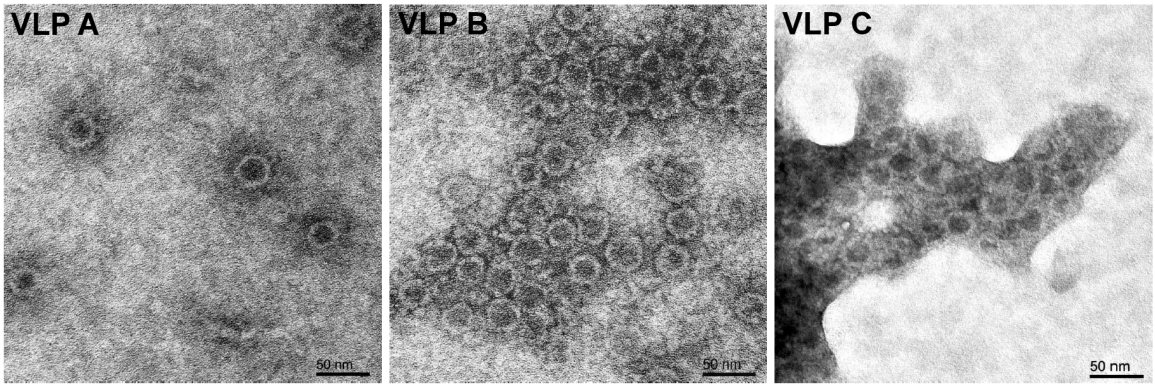


Figure 3.9: Transmission electron microscopy micrographs of the formulated virus-like particles (VLPs) A, B, and C after the end of the assembly process by cross-flow filtration.

3.4 Discussion

3.4.1 On-line Measurement Setup

As shown in Figure 3.2, the experimental setup included a flow restrictor and a filter next to the sensors in the on-line measurement loop. The flow restrictor and filter were added to improve the measurement quality. The flow restrictor set a minimal back pressure in the measurement loop reducing pressure fluctuations and air bubbles. The filter (cut-off 0.7 µm) retained bubbles and large particles adversely affecting light-scattering measurements. The light-scattering measurements depend strongly on the particle diameter d (Bohren and Huffman, 2004). Thus, large particles, such as air bubbles or large aggregates, can completely overshadow the light scattering of smaller particles in SLS and DLS measurements.

3.4.2 Interpretation of SLS and DLS Measurements

During VLP reassembly, anticipated particles in the process solution were homodimers, VLPs, VLP aggregates, and process-related impurities, all of which contributed to light scattering. Thus, the scattered-light intensity is a sum signal generated by all scattering species. By neglecting any interaction between the particles and assuming Rayleigh scattering, the scattered-light intensity I_R can be described as

$$I_R \infty \sum_i c_i d_i^6 \tag{3.1}$$

where i iterates over all species, c_i is the molar concentration of species i, and d_i is the diameter of species i (Bohren & Huffman, 2004). Based on this formula, it can be verified that particle agglomeration and concentration leads to increased scattered-light intensities.

The z-average is the intensity-weighted harmonic mean hydrodynamic diameter (J. C. Thomas, 1987). Therefore, the z-average is not proportional to the concentration but reflects an apparent mean particle diameter. A small fraction of large particles can still significantly increase the z-average. During reassembly, an increase of scattered-light intensity and z-average was expected because of the formation of VLPs and aggregates.

3.4.3 DLS Measurements in Flow

DLS measures the time correlation of scattered-light intensity. In contrast to the typical DLS measurement setup, the time correlation in the on-line measurement loop was not only influenced by diffusion but also by convective flow (Berne & Pecora, 2000). It has been previously demonstrated that the convective flow results in increased estimated diffusion coefficients and thus in reduced particle diameters (Leung, Suh, & Ansari, 2006). The effect was shown to be more pronounced for larger particles. Consequently, underestimation of particle sizes was expected to be more pronounced for aggregates than VLPs than homodimers. No effect on SLS was expected from convective flow.

3.4.4 General Considerations on the VLP Assembly Processes

During the diafiltration process, the disassembly buffer was gradually exchanged by an assembly buffer. The chemical environment of the HBcAg dimers increasingly favored assembly. This is different to the conventional approach in VLP kinetic studies where the composition of the assembly reaction liquid is usually adjusted by rapid dilution (Mukherjee et al., 2008; Adam Zlotnick et al., 1999). In said studies, assembly equilibrium phases were reached in a few minutes. Given the comparably large time frame of diafiltration experiments (75 min to 135 min), we assume that the VLP concentration was almost exclusively dependent on the buffer composition.

Figure 3.10 illustrates the formation of particles out of HBcAg dimers during a diafiltration process and expected sensor responses. The diafiltration process was split into phases I to III based on different reactions occurring during each phase.

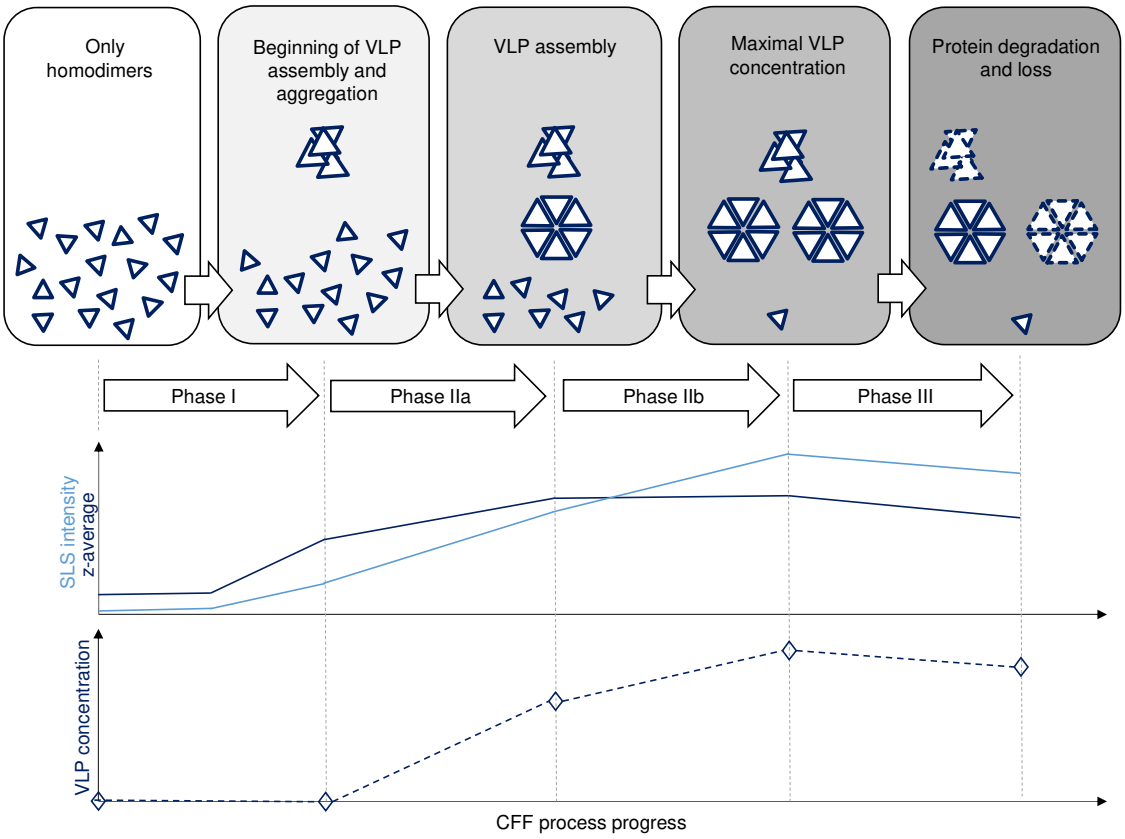


Figure 3.10: Theoretical consideration of particle formation during the assembly process by cross-flow filtration (CFF). Homodimers, aggregates, and virus-like particles (VLPs) are shown as schematics. The expected development of static light scattering (SLS), z-average, and VLP concentration signals is shown over the CFF process progress subdivided into four phases. In the process, the buffer of a homodimer solution is gradually exchanged by assembly buffer to initiate VLP assembly. In phase I, few aggregates are formed and no assembly takes place. The formation of aggregates increases the light-scattering signals while the VLP concentration remains at zero. As a consequence of exceeding a critical buffer composition, VLPs start to form in phase IIa, visualized by an increase in VLP concentration. The light-scattering signals continue to increase as a response to particle formation. In phase IIb, assembly continues, indicated by a further increase in VLP concentration and static light scattering. The z-average remains comparably constant as its value is already close to the actual VLP diameter and is thus only marginally influenced by further assembly. In phase III, the assembly reaction is no longer proceeding. Particles are depleted resulting in a decrease in the light-scattering signals.

In phase I, buffer exchange starts but no assembly occurs, i.e. the VLP concentration remains zero. However, aggregates may form resulting in an increase in scattered-light intensity and z-average, as seen in Figures 3.4, 3.5, and 3.6.

In phase II, homodimers assemble into VLPs. Native HBcAg VLPs are 30 nm to 34 nm in diameter (Crowther et al., 1994). VLP concentration increases to its maximum, while the scattered-light intensity and z-average continue to rise. To explain the sensor response more comprehensively, phase II was subdivided into two subphases, IIa and IIb. In subphase IIa, z-average and scattered-light intensity both increase. In subphase IIb, scattered-light intensity further increases while z-average remains constant. The increase in scattered-light intensity is caused by the ongoing formation of VLPs and aggregates. Conversely, the z-average stagnates, as it is an intensity-weighted harmonic mean. Native HBcAg VLPs are 30 nm to 34 nm in diameter (Crowther et al., 1994). When the z-average is close to the size of VLPs, further assembly has only a small effect on the z-average, while the scattered-light intensity still increases due to the formation of particles.

In phase III, the VLP concentration no longer increases. Thus, the end of the assembly process is reached. A loss of aggregates is reflected by a decrease in z-average and scattered-light intensity. A decrease in scattered-light intensity and UV absorbance with constant z-average reflects a decrease in overall protein concentration with constant particle size distribution.

Towards the end of some processes (most pronounced for VLP A and C at 0.25 bar), both light-scattering signals decreased combined with a decrease in the UV signal at 280 nm. Thus, the protein concentration decreased due to adsorption to the CFF membrane or retention on the measurement loop filter. The elevated salt concentration of the process liquid at this stage of the process may have promoted adsorption of protein to the hollow fiber membrane (Hanemaaijer et al., 1989). In both runs, the z-average started to decrease already earlier than the UV signal at 280 nm around the location of the inflection point of the a/b-ratio while the UV absorbance was still approximately constant. This could indicate a partial disintegration of aggregates. Phase III was generally short, as either its onset was close to the final DV or the process was stopped early due to membrane fouling.

The assembly of HBcAg VLPs also induces changes in mean hydrophobicity around aromatic amino acids as capsid assembly relies on hydrophobic interaction forces (Venkatakrishnan & Zlotnick, 2016; Wynne et al., 1999). Tyrosine-132 is especially important for the assembly (C. R. Bourne, Katen, Fulz, Packianathan, & Zlotnick, 2009). In homodimers, tyrosine-132 is highly solvent-exposed, as shown in Figure 3.3. After VLP assembly, tyrosine-132 is buried in a hydrophobic pocket of the neighboring homodimer. During diafiltration, the solvatization of tyrosine changes because of aggregation as well as VLP assembly. If the mean effect on hydrophobicity by aggregation is small compared to the mean effect caused by assembly, the change over time of the a/b-ratio correlates to the rate of assembly. As a result, the a/b-ratio's inflection point marks the point of the highest rate of assembly. Similarly, the increase in the wavelength of the tryptophan absorption minimum marks an increase in hydrophobicity around tryptophans. Since the change in the solvent exposure of tryptophans during VLP assembly is less pronounced, the effect is weaker and more biased by aggregation.

3.4.5 Cross-Flow Filtration for VLP Assembly

VLP A was assembled from the purest dimer stock solution of the three investigated VLPs. The process was thus expected to perform comparably well. This agreed with the experimental results at 0.5 bar and 1 bar TMP. The observed z-averages of 28 nm to 29 nm in phase III showed that there was a significant fraction of VLPs. Few large particles were generated while other factors such as the flow reduced the z-average compared to off-line DLS analytics (see Table 3.1). The higher final z-average and an elevated scattered-light intensity at 0.25 bar TMP provided evidence of the formation of large aggregates. The observations made for VLP A were in general also applicable to VLP B and C. Both VLPs were adversely affected at lower TMPs by aggregation reflected by increased z-averages and light-scattering intensities.

A further interesting result of this study was the clustering of the inflection points of the a/b-ratio either around 1.5 DV or around 0.8 DV. An early inflection point is consistent with early VLP formation. Conversely, a late inflection point correlated to an early increase in aggregates. By keeping in mind that the DV is indicative of the progress

of buffer exchange, the conclusion may be drawn that VLP assembly is inhibited by aggregates. Indeed, a similar conclusion was previously proposed for MuPyVP1 VLPs (Y. Ding, Chuan, He, & Middelberg, 2010). Ding and coworkers described a competition of capsomere association into aggregates and precursors of MuPyVP1 VLPs.

The results of the diafiltration experiments for all VLPs showed that a low TMP of 0.25 bar lead to an increased aggregation propensity and an increased process time compared to the other conditions. At 0.5 bar and 1 bar TMP, the process time, VLP concentration, and aggregate content depended on the VLP construct and stock purity but were not solely dependent on the TMP. For increased yield and decreased aggregate content, it could be helpful to introduce a further purification step for VLP B and C. In all runs, aggregate concentration by SEC did not reflect the data obtained by light scattering. The reason for this seemed to be that large aggregates were depleted during sample preparation or in the SEC column. As a consequence, light scattering provided a more complete picture of the aggregate content.

Process phase III is characterized by product loss. The process should therefore be terminated at the end of phase II. It is worth noting that the end of phase II is influenced by the VLP construct but seems to be independent of the applied TMP. No plateau or decrease in assembly was observed for VLP B. VLP B was charged strongest, requiring higher ionic strengths to overcome the electrostatic charges of the homodimers during assembly (see Table 3.1). Zeta potentials of VLP A and C are similar. For both, a transition into phase III was observed.

To compare the assembled VLPs with standard characterization methods, we performed DLS and TEM measurements on the assembled VLPs. Off-line DLS VLP peak diameters with a mean of 41 nm and a standard deviation of 11 nm are comparable to that of wild type HBcAg VLPs (typically 30 nm to 34 nm (Crowther et al., 1994)). No significant influence of the TMP or construct on the final VLP peak diameter could be observed. TEM measurements confirmed the existence of assembled VLPs for all three constructs in the expected size range.

In summary, the analytical measurements of the VLP size and structure confirm the information obtained from the PAT tools.

3.4.6 Benefits of Using PAT for Process Development and Production

PAT is currently a frequently investigated approach to increasing the acquired information about unit operations in biopharmaceutical process development and production by timely measurements. Generating information on the process in (near) real time potentially results in a better understanding, faster optimization, and reduced off-line analytical samples (Bakeev, 2010).

Here, the UV absorbance at 280 nm provided insight into changes in the concentration of protein and other absorbing species in real time. This can be of advantage for assessing the membrane performance (e.g. membrane fouling, pore rating out-of-specification, or membrane damage). A mechanistic understanding is, however, often not possible solely based on a single wavelength. A more in-depth view on the on-going processes during UF/DF could be realized based on the acquired UV/Vis spectra. For HBcAg, tyrosine-132 is especially important for the VLP assembly. The a/b-ratio provides a mechanistic insight into the assembly reaction based on the mean tyrosine solvatization. Next to means for quantification, the UV/Vis spectrometer implemented in the presented setup thus provides mechanistic process understanding. Furthermore, other UV/Vis chromophores are phenylalanine, tryptophan, and disulfide bridges (W. Jiskoot & Crommelin, 2005). These may be affected during the assembly of other VLPs. For example, during the assembly of human papilloma virus-like particles, disulfide bridges are the key to the formation of higher-order structures (Maolin Li, Beard, Estes, Lyon, & Garcea, 1998). An assembly process with these VLPs could therefore be monitored with a UV/Vis spectrometer.

Another changing protein attribute which can be monitored is the particle size. The significant increase in size has a large impact on the scattering characteristics of the process fluid. The light-scattering photometer thus allowed for the detection of the start of the assembly reaction and maximum VLP concentration. Light-scattering photometers are universal detectors that are not dependent on the protein primary structure. As a consequence, any VLP assembly reaction can be monitored with this technique. In development and production, light-scattering detectors provide the means for detecting the ideal point to

stop CFF or to initiate the next process step. This can improve the product quality (as process phase III is omitted) and allow for process intensification.

Generally, the on-line sensors provide data with high temporal resolution which typically is difficult to achieve with off-line analytics. In consequence, smaller changes in process characteristics (e.g. assembly onset, end of phase II) can be detected. This may be helpful for the further assessment of different processes in development or for detecting deviations or hidden trends in production.

For process monitoring in production, it may be beneficial to retrieve VLP concentrations in real time. A PLS model was thus developed to demonstrate the possibility to monitor VLP concentration on-line by UV/Vis spectroscopy. The model was optimized by a constrained heuristic search algorithm. The minimal number of four latent variables was set to reflect the minimal amount of independent UV-active species (VLP concentration, deoxyribonucleic acid (DNA) concentration, urea concentration, and aggregates). Reliable VLP concentration estimations were possible for all three constructs. In production, UV/Vis measurements in conjunction with a PLS model could thus be used for the real-time assessment of the assembly progress and ultimately for process control. Based on the regression coefficients of the PLS model (Figure 3.8), it is clearly visible that the fine structure of the tyrosine and tryptophan absorption is of major importance for the regression. Therefore, the PLS model accesses information similar to that provided by the a/b-ratio and the tryptophan minimum. The differences between the regression coefficients for VLP A, B, and C were attributed to the changing purity of the stock solutions. Provided that no additional chromophores are introduced into the MIR, a universally applicable PLS model for different HBcAg constructs is conceivable. This may be evaluated further in future studies.

3.5 Conclusion and Outlook

In this study, we investigated HBcAg assembly by diafiltration of three different constructs at three different TMPs. We developed an on-line measurement setup consisting of a UV/Vis and a light-scattering sensor (DLS and SLS) with a unified software platform. This setup allowed for

monitoring mean particle sizes, hydrophobicity around tyrosine and tryptophan as well as the protein concentration. VLP particle formation was verified by off-line DLS measurements and TEM imaging. Based on the acquired UV/Vis spectra, we calibrated three PLS models for estimating VLP concentrations in real-time. Regarding process performance, we observed that processes with hollow fiber modules at 0.25 bar TMP resulted in increased aggregation. In all processes, the maximum rate of assembly occurred around two characteristic DV. This behavior was interpreted as a result of aggregation-related inhibition of VLP assembly, which makes it especially important to prevent aggregation in a VLP assembly process. The maximum VLP concentration coincided with the maximum light-scattering intensity. Thus, the light scattering peak or the calibrated PLS model could potentially be used as PAT decision tools for VLP assembly process control leading to improved product quality and intensified processes. In summary, the established setup has shown great potential for improving process monitoring, development, and understanding during VLP assembly by diafiltration.

In the future, strategies may have to be developed for process control during VLP reassembly. The proposed setup allowed for monitoring central quality attributes during the process with and without calibrated chemometric models. It is therefore a good starting point for any further research in this direction. From a process development point of view, the current results have not yet shown a reduced process efficiency at the highest TMP. A further increase in TMP may thus be attractive. Alternative membrane options, such as membrane cassettes, could strongly affect the process and may be interesting to evaluate with the setup.

Acknowledgment

This project received funding from Deutsche Forschungsgemeinschaft (DFG) in the frame of SPP 1934, Project number 273937032. We would kindly like to thank BioNTech Protein Therapeutics GmbH for providing the VLP constructs and the production process scheme, especially Anja Wilming, Thomas Hiller, and Thorsten Klamp for their collaboration. We express our gratitude to Reinhard Schneider for technical and scientific support in preparing and performing TEM imaging. We are

thankful for the thorough review of the manuscript by Dr.-Ing. Josefine Morgenstern, Laura Rolinger, and Heidemarie Knieriem.

Appendix A: Supplementary Material

The Supplementary Material associated with this chapter contain the following information:

- S3.1: Calculation of Local Hydrophobicity around Aromatic Amino Acids
- S3.2: Reversed-Phase Chromatography
- S3.3: Cross-Flow Filtration (CFF) Process Progress

4

High-Throughput Computational Pipeline for 3-D Structure Preparation and In Silico Protein Surface Property Screening: A Case Study on HBcAg Dimer Structures

Marieke E. Klijn[a,*], Philipp Vormittag[a,*], Nicolai Bluthardt[a], Jürgen Hubbuch[a,**]

[a] Institute of Process Engineering in Life Sciences, Section IV: Biomolecular Separation Engineering, Karlsruhe Institute of Technology, Fritz-Haber-Weg 2, 76131 Karlsruhe, Germany

* Contributed equally

** Corresponding author

Abstract

Knowledge-based experimental design can aid biopharmaceutical high-throughput screening (HTS) experiments needed to identify critical manufacturability parameters. Prior knowledge can be obtained via computational methods such as protein property extraction from three dimensional (3-D) protein structures. This study presents a high-

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

4.1 Introduction

Virus-like particles (VLPs) are macromolecular assemblages, which in their simplest form consist of multiple copies of one viral structural protein (Kushnir et al., 2012). Their particulate and highly repetitive structure invokes an immune response similar to that of native viruses, but VLPs are incapable of reproduction as viral nucleic acids are lacking (Chackerian, 2007; Kushnir et al., 2012). VLPs can therefore provide immunization against the virus they were derived from, as was done for hepatitis B virus (HBV; Engerix B, Recombivax) (McAleer et al., 1984) and human papilloma virus (HPV, Cervarix; Gardasil) (Bryan et al., 2016). Immunization unrelated to the native virus can be achieved with chimeric VLPs (cVLPs), which are VLPs containing a foreign antigenic epitope. These antigenic epitopes can be inserted into a capsid forming protein at either the N-terminus, C-terminus, or major immunodominant region (MIR) (Pumpens & Grens, 2001). This insertion aims to trigger an immune response, adjuvanted by the particulate and repetitive VLP structure (Kratz, Böttcher, & Nassal, 1999). CVLPs are increasingly used in preclinical and clinical studies (Mohsen et al., 2017). An example of a cVLP that received positive opinion of the European Medical Agency is a malaria vaccine based on a HBV surface antigen VLP with an inserted segment of the *Plasmodium falciparum* circumsporozoite protein (Nielsen et al., 2018). Another platform for chimeric antigen display is the HBV core antigen (HBcAg) protein. Chimeric HBcAg VLPs with foreign and self-epitopes have been shown to induce strong B cell responses, a characteristic that can be used to develop VLPs for treatment of cancer (Fehr et al., 1998; Klamp et al., 2011; Milich, Sallberg, & Maruyama, 1995).

CVLP development involves screening large numbers of candidate epitope insertions (Pumpens et al., 2008). During screenings, cVLPs are evaluated based on immunogenicity, structure stability, and assembly-competence (Chackerian, 2007; X. Ding, Liu, Booth, Gao, & Lu, 2018). For example, fewer than 50% of inserted peptides in the HBcAg platform resulted in a properly assembled and soluble VLP (Jegerlehner et al., 2002). Structural stability and solubility are not only desired in the final formulation to ensure product efficacy, quality, and safety, but also throughout downstream processing to ensure manufacturability

(Buckland, 2005; X. Ding et al., 2018; Vicente, Roldão, et al., 2011). During manufacturing, VLPs are exposed to different environmental conditions such as changes in pH, ionic strength, and temperature. These conditions influence physicochemical properties of VLPs, which in turn determine critical evaluation parameters such as the structural stability and assembly-competence (Priddy & Middaugh, 2014). High-throughput screening (HTS) experiments allow for workload reduction in virus and VLP studies to determine optimal processing (Hämmerling, Ladd Effio, Andris, Kittelmann, & Hubbuch, 2017; Ladd Effio & Hubbuch, 2015) and formulation (Hämmerling, Lorenz-Cristea, Baumann, & Hubbuch, 2017) parameters. HTS design for VLP studies can be further optimized by search space minimization and manufacturability assessment using prior knowledge of physicochemical properties obtained computationally from three-dimensional (3-D) protein structures (X. Ding et al., 2018; Lua et al., 2014; Vicente, Mota, Peixoto, Alves, & Carrondo, 2011). Physicochemical properties that are most important for virus particles include electrostatic surface charge (Ghanem et al., 2016; Mellado et al., 2009; Schijven & Hassanizadeh, 2010). Research on bacteriophage MS2 showed correlation between experimentally determined virus surface charge using zeta potential measurements, and computationally calculated protein charge (Penrod, Olson, & Grant, 1996). Moreover, experimentally determined protein zeta potential showed stronger correlation with calculated protein charge using only capsid surface atoms compared to protein charge calculated using all MS2 capsid atoms. Other research showed that calculated protein charge using the surface of a single MS2 capsid protein was in agreement with experimentally determined protein zeta potential of the entire MS2 capsid (Lošdorfer Božič & Podgornik, 2017). Ionizable groups of a protein determine protein properties such as surface charge, structure, and stability (Johnston, Søndergaard, & Nielsen, 2011). Therefore, both 3-D structure preparation and *in silico* determination of surface charge require an estimation of the pKa of titratable groups. Fast and fairly accurate pKa estimation methods have been developed, such as methods to monitor pKa shifts during a molecular dynamics (MD) simulation (Krieger, Nielsen, Spronk, & Vriend, 2006) or to process a large number of structures parallelized in a short time (Anandakrishnan, Aguilar, & Onufriev, 2012).

Candidate cVLP 3-D structures have to be available for computational physicochemical property extraction. As it would be impractical to produce all candidates and experimentally determine their 3-D structures, an *in silico* 3-D structure preparation approach is needed. This approach would require an automated and high-throughput framework to support screening a large number of cVLPs to minimize manual effort. These requirements can be met with homology modeling. Homology modeling can be performed using several approaches (Forster, 2002; Venselaar et al., 2010), but all resulting 3-D structures remain only an estimation of reality. Further model refinement is needed to meet structure quality requirements and should therefore include an MD simulation step (Fan & Mark, 2004). Structure refinement requires the selection of a force field. The choice depends on the application and it can be notoriously difficult to identify the best-performing force field for a particular application. Novel self-parameterizing knowledge-based force fields, such as YASARA2, have been developed to improve the calculation of torsional angles and have shown to be useful and accurate for the physical correction of proteins by energy minimization (Krieger et al., 2009). Several authors have analyzed the performance of different open-source force fields by comparing *in silico* structural data to NMR experimental data (Beauchamp, Lin, Das, & Pande, 2012; Best, Buchete, & Hummer, 2008; Lange, Van Der Spoel, & De Groot, 2010). In general, modern force fields perform reasonably accurate and reproducible for MD simulation of proteins (Martín-García, Papaleo, Gomez-Puertas, Boomsma, & Lindorff-Larsen, 2015).

For VLPs, *in silico* experiments have most frequently been applied to study capsid stability using complete VLP capsid 3-D structures. All-atom MD simulations of complete capsids are as challenging as they are computationally expensive and can only be done using relatively short *in silico* timescales. Reported simulations reach <10 ns per day on super-computers (Freddolino et al., 2006; Roberts et al., 2012; G. Zhao et al., 2013) or 30 ns/day when using constrained bond-lengths (Larsson et al., 2012). However, modeling VLP structural transitions (e.g., self-assembly, capsid disintegration) requires a much larger timescale (μs or ms) (Mansour, Sereda, Yang, & Ortoleva, 2015). Compared to all-atom MD simulations, computational expense has been reduced to reach these relatively large timescales using coarse-grained (Arkhipov, Freddolino, & Schulten, 2006; Reddy & Sansom, 2016; Reddy et al., 2015) or multi-

scale (Ayton & Voth, 2010; Cheluvaraja & Ortoleva, 2010; Joshi et al., 2011; Machado, González, & Pantano, 2017; Miao, Johnson, & Ortoleva, 2011) models in various capsid studies. Supercomputers, such as the Blue Waters supercomputer with 128000 cores, were used and a simulation duration of several days for a single VLP was reported (G. Zhao et al., 2013). In silico candidate screening would require an equal amount of simulations as available cVLP candidates. Depending on the application, this could involve screening of hundreds of cVLP candidates. In this case, simulation time would go up to the order of magnitude of a year, even with the use of a supercomputer. Time requirement, super computer availability, and respective expertise hamper the implementation of these methods in computational high-throughput candidate screenings. Simulation simplification, by using only a single capsid protein or capsid building blocks models (Lua et al., 2015; Lin Zhang et al., 2013), aids in resolving these limitations. Monomers and pentamers were compared to an entire VLP 3-D capsid model to evaluate the applicability to immunogenicity prediction (Joshi et al., 2011). Joshi and coworkers showed that the immunogenicity predictor (epitope flexibility) was dependent on the complete capsid construct and thus a complete VLP capsid 3-D model was required to capture this effect. This requirement is not expected for the evaluation of surface charge as it has been shown that MS2 capsid protein surface charge descriptors have a high correlation to experimental zeta potential data of entire structure (Lošdorfer Božič & Podgornik, 2017; Penrod et al., 1996). In addition, this case study used chimeric HBcAg structures that differ only in the epitope located on the outer VLP surface. Therefore, the influence of dimer contact area on possible zeta potential changes observed for entire chimeric HBcAg VLP structures was considered to be minimal. Thus, surface charge after 3-D structure preparation of HBcAg dimers was evaluated based on its correlation to experimental zeta potential obtained for entire HBcAg VLP structures. Monomers were not considered as model simplification, since only dimers or larger assemblies (i.e., capsids) are present under physiological conditions (Adam Zlotnick, Tan, & Selzer, 2013).

This study presents a computationally inexpensive, high-throughput, and entry-level pipeline to obtain 3-D structures. Time and computational effort were minimized by automated homology modeling including novel, data-dependent, and stepwise MD simulation for homology model

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

4.2 Materials and Methods

4.2.1 Sample Preparation

Recombinant chimeric HBcAg constructs used in this study (referred to as VLP A, VLP B, and VLP C irrespective of being an HBcAg dimer or VLP) were modified in the MIR to display foreign epitopes on the VLP surface. Constructs were expressed and purified according to the production protocol generously provided by BioNTech Protein Therapeutics GmbH (Mainz, DE). Purified and assembled VLPs were stored at -20 °C and dialyzed into a 50 mM Tris (Merck KGaA, Darmstadt, DE) buffer at pH 7.2 containing 100 mM NaCl (Merck KGaA, Darmstadt, DE) for analysis. Buffer was prepared with ultrapure water (PURELAB Ultra, ELGA LabWater, Lane End, UK) and filtered through a 0.20 µm pore size Supor filter (Pall, Port Washington, NY, USA). Samples were brought to room temperature and filtered through

a 0.20 µm polyethersulfone (PES) filter (VWR International, Radnor, PA, USA) before measurements. Required VLP sample concentrations were obtained using Vivaspin 20 filters with a 30 kDa pore rating (Sartorius, Goettingen, DE). VLP concentration was determined with a NanoDrop2000c UV-Vis spectrophotometer (Thermo Fischer Scientific, Waltham, MA, USA). The E1% (280 nm) extinction coefficient was calculated by the online Swiss Institute of Bioinformatics ProtParam tool (https://web.expasy.org/protparam.html) based on the primary structure of the HBcAg monomer (Gasteiger et al., 2005).

4.2.2 Zeta Potential

Electrophoretic mobility measurements were performed with the Zetasizer Nano ZSP (Malvern Instruments Ltd., Malvern, UK). Folded disposable capillary cells (DTS1070, Malvern Instruments Ltd., Malvern, UK) were filled with the appropriate buffer and 50 µL of a 1 g/L VLP sample. VLP samples were inserted by employing the diffusion barrier technique (US Patent 2017/0269030 A1, 2017) using a 200 µL round, 0.5 mm thick Corning Costar gel-loading tip (Corning Inc., Corning NY, USA). Six replicates were measured at 25 °C in automatic mode, where each measurement consisted of 120 seconds equilibrium time and five runs with a maximum of 15 sub runs. The applied voltage was set to 60 mV and the dispersant was set to water. A material refractive index of 1.45 and absorption of 0.001 AU was used. The average zeta potential was calculated by Zetasizer Software (version 7.12, Malvern Instruments Ltd., Malvern, UK) with the measured average electrophoretic mobility, a viscosity of 0.8872 mPas, a dielectric constant of 78.54, and Smoluchowski's approximation of 1.5 (Smoluchowski, 1921). For each VLP sample, outlier detection was performed with MATLAB (version 2017b, MathWorks, Natick, MA, USA), using the inter quartile range rule with a whisker length of 0.75 (Moore, McCabe, & Craig, 2009), followed by median zeta potential calculation.

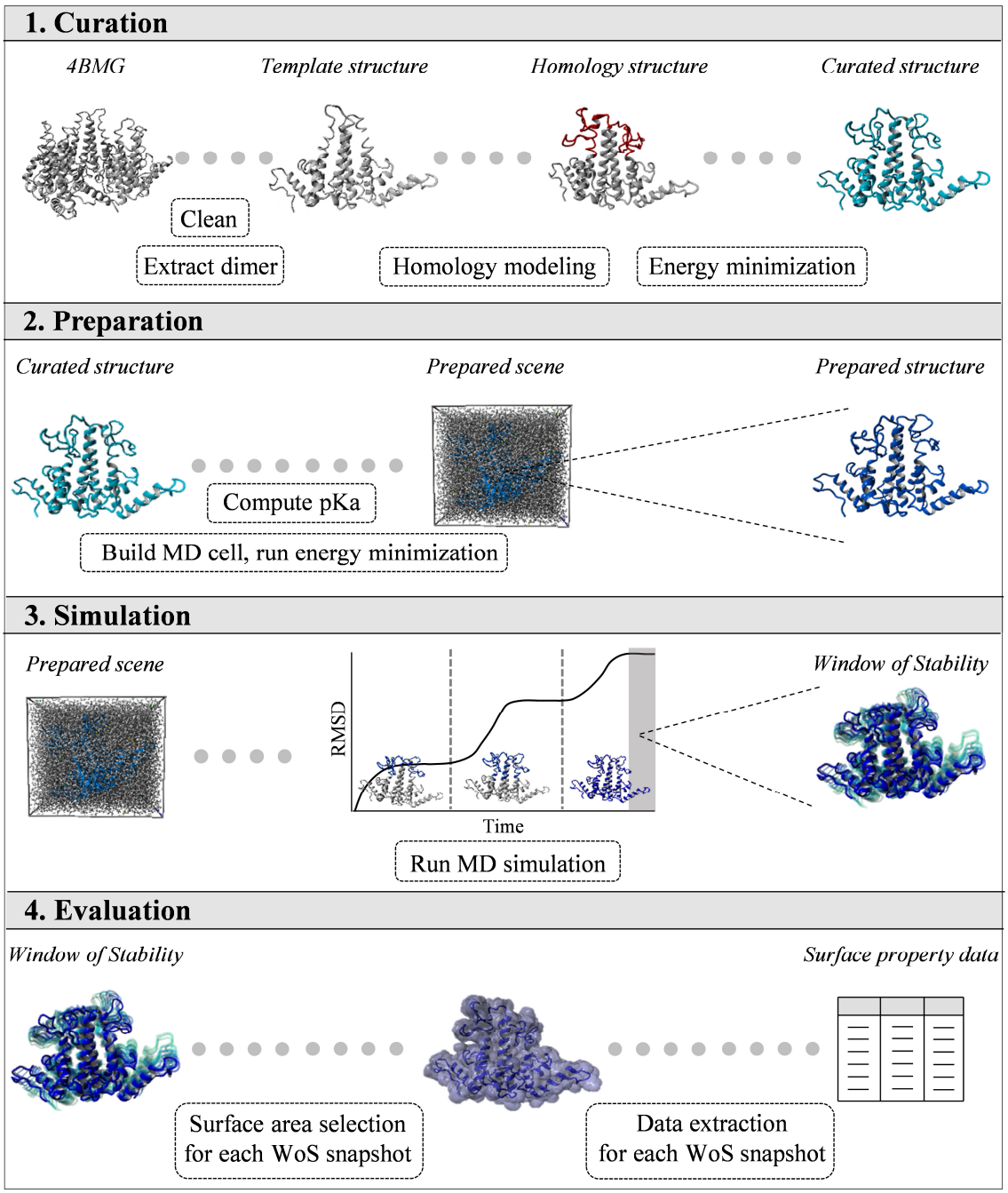


Figure 4.1: Computational pipeline for high-throughput homology model surface property data extraction. Four stages are depicted: (1) Curation: epitope insertion using homology modeling (Modeller), followed by an energy minimization run (YASARA); (2) Preparation: computed pKa values (H++) are assigned, followed by an energy minimization in a simulation cell (YASARA); (3) Simulation: 3-step data-dependent molecular dynamic (MD) simulation (YASARA) terminated by identification of a 2 ns Window of Stability (WoS); (4) Evaluation: surface area selection and extraction of surface property data for each snapshot in the WoS (YASARA).

4.2.3 Computational Methods

Figure 4.1 depicts the computational pipeline used to compute surface property information from dimer chimeric HBcAg structures. Required input is a template 3-D structure, the target sequences, and experimental conditions (i.e., oligostate, pH, and salt concentration). 3-D structure curation and MD scene preparation, described in section "Structure and Scene Preparation", were performed fully automated by employing an in-house developed MATLAB script (version 2017b, MathWorks, Natick, MA, USA). All depicted steps in section Curation and Preparation in Figure 4.1 were an automated operation of either MATLAB, YASARA (version 16.9.23, YASARA Biosciences GmbH, Vienna, AT), Modeller (version 9.18, University of California, San Francisco, CA, USA) (Fiser & Šali, 2003), H++ (Virginia Tech, Blacksburg, VA, USA, biophysics.cs.vt.edu) or Python (version 2.7.13, Python Software Foundation, Wilmington, DE, USA) sub scripts. These steps resulted in prepared scenes for MD simulation of each VLP construct. MD simulation of the prepared scene is described in section "Molecular Dynamics" and extraction of VLP surface properties is described in section "Data Processing". The 3-D structure quality was monitored throughout the workflow with the quality Z-score. This is the mean value of the WHAT IF parameters Packing1, PhiPsi and Backbone (Krieger et al., 2009; Vriend, 1990). Quality parameters were calculated using the YASARA2 force field in a TIP3P water (Jorgensen, Chandrasekhar, Madura, Impey, & Klein, 1983) filled cubic cell, with walls extended 10 Å from the 3-D structure.

4.2.3.1 Structure and Scene Preparation

The three HBcAg structures used in this study were based on C-terminally truncated and histidine(His)-tagged HBcAg, which were modified in the MIR. All experimental structures have an identical C-terminus. Therefore, it was assumed that the His-tag would not have a significant impact on the relative assessment of 3-D structural biophysical parameters. To avoid homology modeling of the His-tag, the C-termini of the input target sequences matched the template structure C-terminus. The 3-D crystal structure of C-terminally truncated (1-149) hexameric HBc Y132A was obtained from the online research collaborator for structural bioinformatics protein data bank (RCSB PDB, www.rcsb.org), under PDB ID 4BMG with a resolution of 3 Å (Alexander

et al., 2013; Berman et al., 2000). All non-protein molecules were removed and the hydrogen bonding network was optimized with YASARA (Krieger, Dunbrack, Hooft, & Krieger, 2012). The multimeric state was corrected to obtain a dimeric 3-D structure, which resulted in the template structure shown in Figure 4.1. Subsequently, homology modeling was performed to adjust the template structure to the target sequence using Modeller. The *automodel* function constructed five homology models, where gap initiation and extension penalties for sequence alignment were set to -600 and -400, respectively. Obtained homology models were superposed in YASARA and their atom coordinates averaged (referred to as homology structure). The hydrogen network was optimized and an energy minimization was run with the averaged structure at experimental pH and using the AMBER99 force field (J. Wang, Cieplak, & Kollman, 2000). After steepest descent minimization, the procedure continued by simulating annealing using 2 fs time steps. Atom velocities scaled down by 0.9 every 10^{th} step until the energy improved by less than 0.05 kJ/mol per atom during 200 steps. The resulting structure is referred to as the curated structure in Figure 4.1. The curated structure was uploaded to the H++ webserver using a Python web scraping algorithm (*selenium* library) to compute pKa values (Anandakrishnan et al., 2012). The external and internal dielectric constant were set to 80 and 10, respectively, and salinity and pH were set equal to experimental conditions (i.e., 0.1 molar salinity and pH 7.2). Obtained pKa values and the resulting 3-D structure were automatically downloaded and used to build an MD simulation cell. Additionally, to investigate the effect of H++ computed pKa values, pKa values computed by YASARA were used instead of H++ (Krieger et al., 2006). The simulation cell contained the prepared 3-D structure, which included computed pKa values as well as (de)protonated termini based on the experimental pH and computed pKa values. Cell walls were built at a distance of 10 Å from the refined 3-D structure. After simulation cell construction, a neutralization run was performed. TIP3P water molecules (Jorgensen et al., 1983) were added to the simulation cell (water density was set to 0.997) as well as salt ions (set to experimental conditions).The final step of MD scene preparation was an energy minimization using identical settings as described before. This resulted in the prepared MD scene depicted in Figure 4.1.

4.2.3.2 Molecular Dynamics

Prepared MD scenes with H++ pKa values were simulated using the YASARA2 or the AMBER03 force field (Duan et al., 2003), and with YASARA pKa values using YASARA2, (Krieger et al., 2009; Krieger & Vriend, 2015) with a cutoff of 7.86 Å (Krieger, Darden, Nabuurs, Finkelstein, & Vriend, 2004) and long range Coulomb interactions using the particle mesh Ewald method (Essmann et al., 1995). Temperature was controlled by rescaling velocities using a modified Berendsen Thermostat (Berendsen, Postma, Gunsteren, Dinola, & Haak, 1984; Krieger et al., 2004). Hardware consisted of two Windows 10 computers with an Intel i7-6700 CPU and a GeForce GTX 1080 GPU. Results of the second computer are shown in Appendix B, Figure S4.2, S4.3, S4.4, and S5. Intramolecular forces were calculated every 2 fs (1 fs for AMBER03) and intermolecular, non-bonded Van der Waals, and electrostatic forces every 4 fs (2 fs for AMBER03) to improve performance and subsequently scaled by 2 (Grubmüller & Tavan, 1998). MD scene snapshots were saved every 2 ps and superposed on the prepared structure to calculate a root-mean-square deviation (RMSD) of atom coordinates. The simulation was automatically performed in three RMSD-controlled steps. In step 1, only the epitope and five adjacent amino acids were simulated. All other amino acid atom positions were constrained. In step 2, 18 additional amino acids towards the N-terminus and ten amino acids towards the C-terminus (i.e., the dimer spike consisting of two alpha-helical hairpins) were simulated without position constraints. Other amino acids were simulated with free side chain atoms but fixed backbone atom positions. In step 3, all atom positions were unconstrained. All H-bonds were constrained during step 1 and step 2 using the linear constraint solver (LINCS) algorithm (Hess, Bekker, Berendsen, & Fraaije, 1997). In step 3, all H-bond constraints were removed after 0.2 ns and the time steps for intermolecular forces and intramolecular forces were reduced to 2 fs and 1 fs, respectively. The simulation advanced to the next step when the moving average (window: 0.15 ns, sampling rate: 10 ps) RMSD change was below a set threshold of 0.75 Å/ns for 0.1 ns. A penalty of 0.02 ns was used if the rate of RMSD change was above the threshold. Step 3 was terminated based on the RMSD coefficient of variance (CV) in a window of the last 2 ns of simulation. MD simulation was terminated when the window CV fell below 2.5%, using a sampling rate of 2 ps. The snapshots of the obtained

WoS were used for the calculation of quality and descriptors. Simulations that did not reached a WoS within 30 ns were manually stopped.

4.2.3.3 Data Processing

The homology structure and all MD snapshots of the WoS obtained with H++ or YASARA pKa values and YASARA2 or AMBER03 force field were analyzed based on their solvent accessible surface area (SASA). Structure SASA was calculated by finding all points a 1.4 Å water probe's oxygen nucleus can reach while rolling over the protein surface approximated by YASARA's numeric algorithm. Contribution of the intra-dimer surface was excluded. Molecular parameters were automatically extracted using similar settings as in the MD simulation. Surface charge was *calculated* for all atoms contributing to the SASA and the resulting surface charge was divided by the total SASA. This was done to exclude size effects that can occur between different epitope insertions. In silico zeta potential values were obtained via linear transformation of surface charge data. Linear transformation included normalization of *in silico* data between 0 and 1 and transformation using the minimum and maximum of the experimental data, as shown by Equation (4.1).

$$\tilde{y}_{Transform} = [\tilde{y}_{norm} \cdot (y_{max} - y_{min})] + y_{min} \qquad (4.1)$$

Normalized *in silico* data is indicated as $\tilde{y}_{norm}$, experimental minimum and maximum data are represented by y_{max} and y_{min}, respectively. Descriptors derived from each snapshot in the WoS are reported as medians and corresponding median absolute deviation (MAD). Correlation between linear transformed *in silico* data and experimental data was evaluated based on the Pearson correlation coefficient (PCC). PCC was calculated with the *corrcoef* function available in MATLAB. The error between *in silico* and experimental data was evaluated with the mean squared error (MSE), obtained with Equation (4.2).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 \qquad (4.2)$$

where n is the sample size, y_i experimental data, and $\tilde{y}_i$ *in silico* generated data.

4.3 Results and Discussion

4.3.1 Quality

Figure 4.2 shows an overview of structural quality Z-scores during curation, preparation, and simulation of each chimeric HBcAg dimer. The structural quality Z-score is an average of three parameters: (1) 3-D direction-dependent packing normality, (2) position normality of residues and secondary structural motifs in the Ramachandran plot, and (3) backbone conformation normality (Krieger et al., 2009). A value below -2 is considered to represent a poor structure and Z-scores close to or above zero indicate more reliable structures. Separate parameter values can be found in Appendix B, Figure S4.1. Quality Z-score differences were observed throughout the structure preparation workflow and between different identified windows of stability. The template structure quality Z-score (-1.18) increased after homology modeling with 0.12 and 0.16 for VLP B and VLP C, respectively. VLP A showed a 0.03 quality Z-score decrease compared to the template structure. Underlying parameters showed that VLP A's backbone conformation quality decreased roughly 1.5 times more than the other constructs. This is attributed to the amount of additional atoms included in the homology model. VLP A contains 17% additional atoms compared to the template structure, while VLP B and VLP C contain 13% and 11% additional atoms, respectively. The other two underlying quality parameters (packing normality and Ramachandran plot position normality) show a similar trend between VLP constructs when comparing the template and homology structure (data shown in Appendix B S.4.1, Figure S4.1). The observed quality improvement of homology structures VLP B and VLP C, which is dominated by Ramachandran position normality parameter improvement, might be an effect of the restraint-based homology modeling and knowledge-based loop modeling used by Modeller (Forster, 2002; Krieger, Nabuurs, & Vriend, 2003). Quality Z-scores of curated and prepared structures were between -1.44 and -1.62, which is between 22% and 37% lower compared to the template structure. Both structures are evaluated after energy minimization at experimental pH, where prepared structures included H++ computed pKa values and the curated structures did not. Energy minimization is used to remove global errors in 3-D structures, such as steric clashes. However, optimization of global and local structural quality with an energy minimization run is

not trivial. Energy minimization may result in lower quality structures because global errors are removed but local errors accumulate (Krieger et al., 2009; Xu & Zhang, 2011). This may explain quality decrease of curated and prepared structures, when compared to the template and homology structures. A similar decrease in quality Z-score after energy minimization with an AMBER99 force field has been reported before (Krieger et al., 2009). Structural issues present in curated and prepared structures were resolved by running an MD simulation with the YASARA2 force field, independent of the used pKa computation method. Mean quality Z-scores of all VLP constructs for MD simulation WoS without H++ (-1.17) and MD simulation WoS with H++ (-1.13) were comparable to the template. This shows there is no quality loss after completing the proposed structure preparation pipeline with the YASARA2 force field. Additionally, the coefficients of quality Z-score variance of 2.0% and 1.6% for the MD simulation with and without H++ pKa values, respectively, reflected that there is no quality influence of the inserted epitope length. However, a decrease in quality is seen for the WoS obtained with the MD simulation using the AMBER03 force field (WoS A03), represented by a mean quality Z-score of -1.87 considering all VLP constructs. This corresponds to observations previously reported about diverse structure quality values obtained with different force fields (Spronk, Linge, Hilbers, & Vuister, 2002). Quality Z-scores for intermediate structures and final MD simulation WoS showed that chimeric HBcAg dimer structure quality in this dataset was mostly influenced by the force field and an MD simulation, independent of the used pKa computation method.

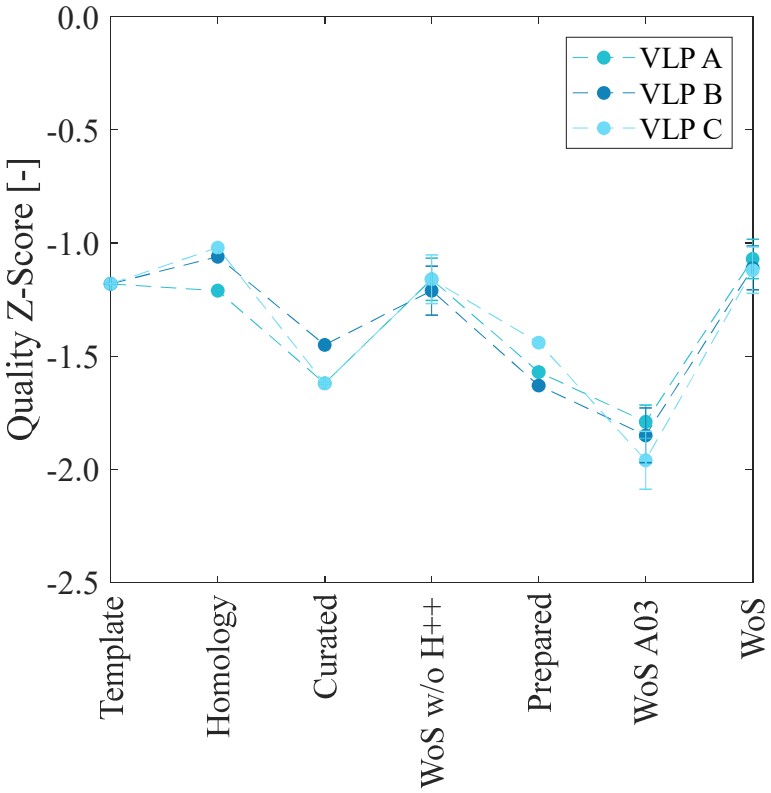


Figure 4.2: Overview of quality Z-scores for the template, homology structure, curated structure, Window of Stability (WoS) without H++ and the YASARA2 force field ("WoS w/o H++"), the prepared structure, WoS obtained with H++ and the AMBER03 force field ("WoS A03"), and WoS obtained with H++ and the YASARA2 force field ("WoS"). The quality Z-score is an average value of the WHAT IF quality factors 3-D packing (QUACHK), Ramachandran Z-score (RAMCHK) and backbone conformation (BBCCHK) *(Krieger et al., 2009)*. A median value and median absolute deviation as error bar is shown for the WoS quality Z-scores. A dashed line is used to guide the eye between the different quality Z-scores. VLP: virus-like particle.

4.3.2 MD Simulations

All chimeric HBcAg homology models were refined with MD simulations. This was done because MD simulations correct structural errors present in homology models (Fan & Mark, 2004). An MD simulation results in a change of atom coordinates, which is measured by the RMSD of those atom coordinates. Structure refinement is achieved upon stabilization of atom positions, referred to as the equilibrium state. This state is identified by a plateau of the RMSD value over simulation time. Plateau identification is frequently done subjectively based on visual inspection of RMSD plots. This approach is not recommended as it was shown to

be biased in a survey among researchers in the field (Knapp, Frantal, Cibena, Schreiner, & Bauer, 2011). To avoid subjective plateau identification, this study employed automated equilibrium state determination based on the average RMSD slope. Automated determination was used within a 3-step MD simulation. In each step a growing part of the chimeric HBcAg dimer structure was refined until an equilibrium was identified. Separate refinement of structure parts was used to reduce simulation time in addition to automated identification of the equilibrium state. The simulation was terminated when equilibrium was reached for the full chimeric HBcAg dimer structure. This state is referred to as the Window of Stability (WoS), which was defined as a 2 ns simulation window where the coefficient of variance of the RMSD in step 3 was below 2.5%. The 3-step MD simulation was specifically implemented for the HBcAg dimer structure, as sequences differ only in the MIR. For other applications (i.e., formulation condition screening of a single protein or a diverse protein dataset) a 3-step MD simulation may not be necessary, and a WoS could be determined in one simulation step.

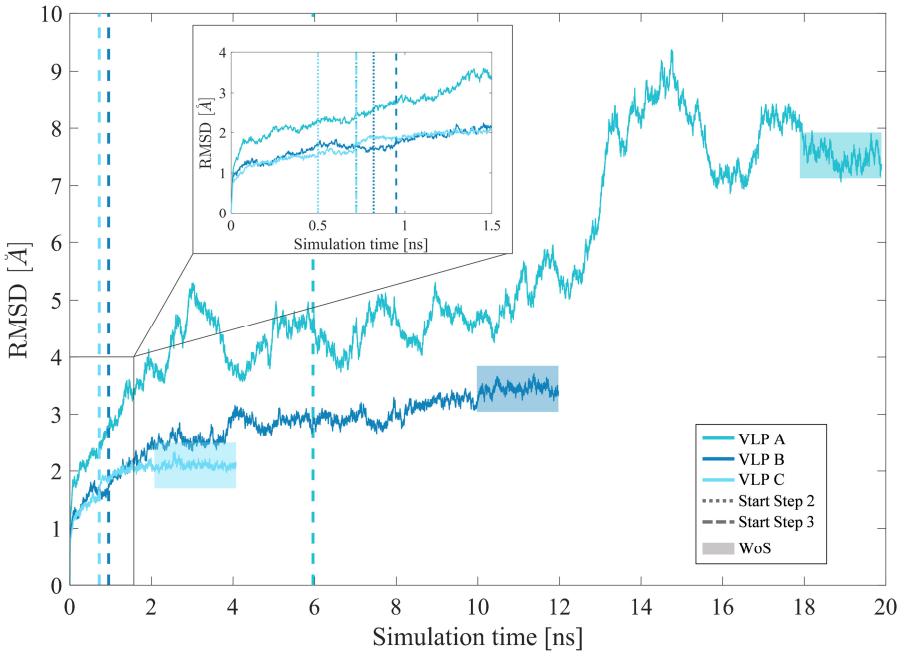


Figure 4.3: Progress of molecular dynamics simulations for virus-like particles (VLP) A, B, and C presented by root-mean-square deviation (RMSD) of atom coordinates (Å) over simulation time (ns). Three different simulation steps are separated by vertical lines, where vertical lines indicate simulation transition points. From 0 ns to dotted line: simulation of epitope and five adjacent amino acids; from dotted to dashed line: simulation of hepatitis B core antigen (HBcAg) dimer spike; from dashed line to the end of simulation: full

dimer simulation. The highlighted area is defined as the 2 ns Window of Stability (WoS).

Figure 4.3 shows the progress of 3-step MD simulations with the YASARA2 force field and H++ computed pKa values for all three VLPs. Every 0.002 ns the atom coordinate RMSD was calculated by superposing a simulation snapshot on the prepared structure. Overall simulation time ranged from 4.0 ns to 19.9 ns and the absolute RMSD increased to 2.10±0.04 Å to 7.52±0.15 Å during MD simulation. The *in silico* time span difference between structures to reach the WoS is in agreement with other work, where structure stability was achieved earlier, later, or not at all, depending on the protein (Fan & Mark, 2004). VLP C showed the lowest RMSD increase (2.1 Å±0.04 in the WoS) and shortest simulation time (6.6 h; *in silico*: 4.01 ns). VLP A resulted in the largest RMSD increase (7.52±0.15 Å in the WoS) and longest simulation time (37.5 h; *in silico*: 19.89 ns). Simulation time increased from VLP C to VLP B to VLP A, which corresponds to the number of inserted atoms of 11%, 13% and 17%, respectively. Step 1, which simulates the inserted epitope and five adjacent amino acids, showed 32.1% to 69.2% of the total RMSD change. This is a relatively large percentage considering step 1 accounted for 3.6% to 12.5% of the total simulation time. The epitope was not part of the template 4BMG crystal structure and therefore it was inserted with homology modeling. Homology models typically have errors in the secondary structure and atomic packing which should be resolved during MD simulation (Fan & Mark, 2004). This is presumably one factor contributing to the relatively large RMSD change observed in step 1, which only refined the inserted epitope and five adjacent amino acids. Another factor that can influence the observed RMSD profile of the epitope is its flexible design. It was stated that epitope flexibility allows for efficient presentation to the immune system (Schumacher et al., 2018), but increased structure flexibility can also result in larger RMSD change during MD simulation. Other parts of the HBcAg dimer are less flexible. Therefore, only small deviations in atom coordinates of the less flexible and conserved region of chimeric HBcAg (i.e., the molecule base and lower part of the spike) were observed when comparing MD simulation steps. This is also illustrated by Figure 4.4, where the RMSD per residue number is shown.

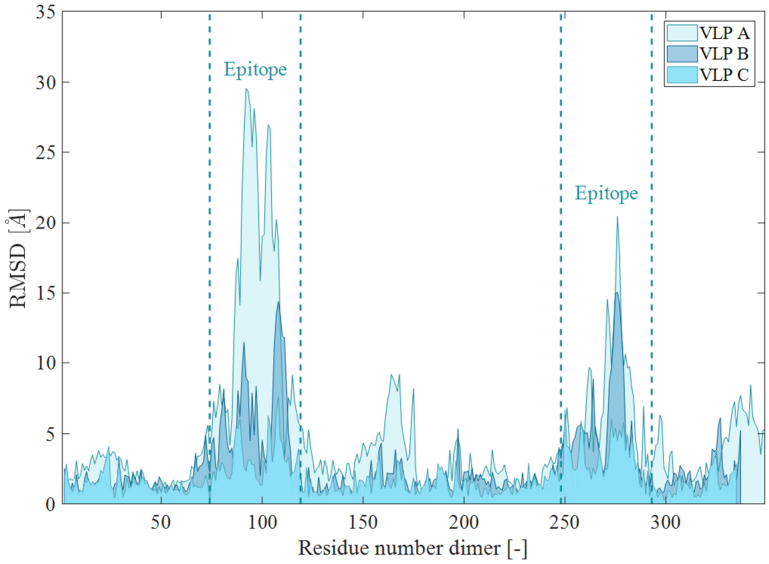


Figure 4.4: Local structural changes during molecular dynamics (MD) simulation represented by root-mean-square deviation (RMSD) of atom coordinates (Å) over residue number (-). Initial structures were compared with last MD simulation snapshots of virus-like particles (VLP) A, B, and C, respectively, with the YASARA2 force field and H++ computed pKa values. Vertical lines mark the inserted epitope exemplarily for VLP A.

Figure 4.4 shows that regions around the epitope show higher RMSD values than other regions. Simulation speed improved due to bond and regional atom constraints and due to an increased time step for force calculation in the first two steps of the simulation. On average, step 1 was 72% (21.26 ns/day) and step 2 was 69% (20.82 ns/day) faster compared to step 3 without constraints and with a smaller time step (12.32 ns/day). This supports the expected simulation speed improvement by employing a data-dependent 3-step method. This corresponds to the previous statement that simulation design should be adjusted to the application and starting structure to obtain optimal speed and stability output. With the used simulation approach, the 2 ns WoS of three chimeric HBcAg dimers were created on a Windows 10 computer with an Intel i7-6700 CPU and a GeForce GTX 1080 GPU in 66.0 h of computational time using the YASARA2 force field and H++ computed pKas. Simulations with H++ pKa values and YASARA2 as force field were also run on another computer containing similar hardware to evaluate reproducibility. No significant difference in simulation outcome were found, including calculation of quality and surface charge. More detailed information on reproducibility can be found in Appendix B, Figure S4.2 to S4.5.

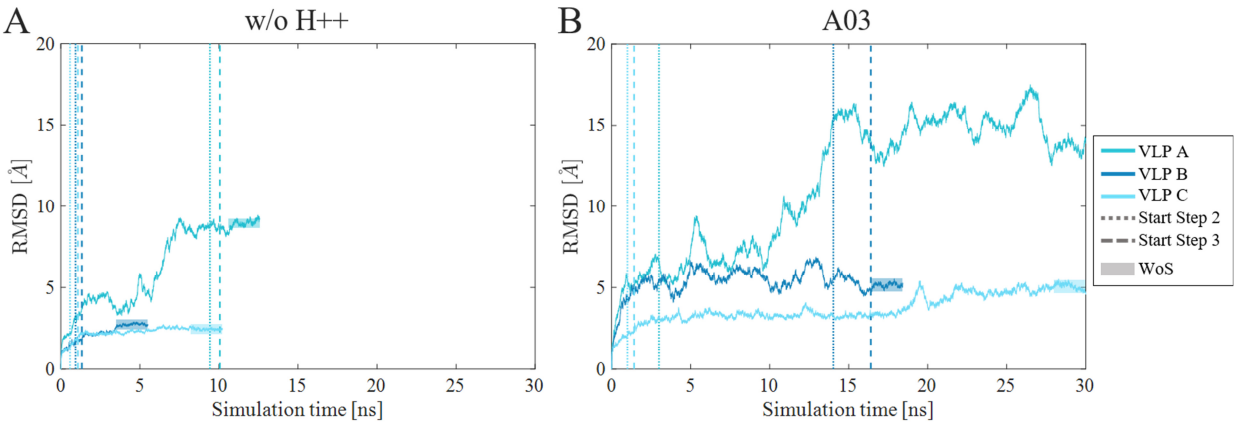


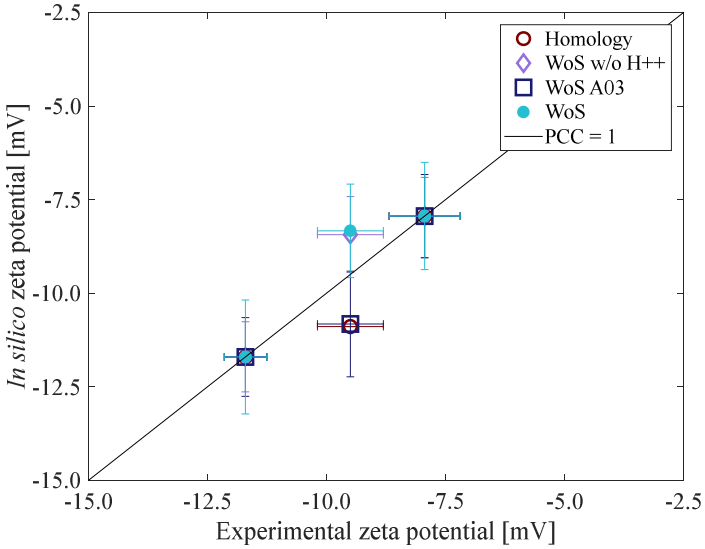
Figure 4.5: Progress of MD simulation for virus-like particles (VLP) A, B, and C presented by root-mean-square deviation (RMSD) of atom coordinates (Å) over simulation time (ns) for **(A)** MD simulation without H++ with YASARA2 as force field ("w/o H++") and **(B)** MD simulation with H++ and AMBER03 as force field ("A03"). Three different simulation steps are separated by vertical lines, where vertical lines indicate simulation transition points. From 0 ns to dotted line: simulation of epitope and five adjacent amino acids; from dotted to dashed line: simulation of Hepatitis B core antigen (HBcAg) dimer spike; from dashed line to the end of simulation: full dimer simulation. The highlighted area is defined as the 2 ns Window of Stability (WoS).

Two additional simulations were performed, the first to evaluate the effect of different pKa value computation methods and the second to compare MD simulation with YASARA2 to a standard force field for protein simulations, AMBER03. Figure 4.5A shows the progress of MD simulations using the YASARA2 force field and with YASARA computed pKas (w/o H++) and Figure 4.5B shows MD simulations with the AMBER03 force field with H++ computed pKas (A03). During MD simulations w/o H++, RMSD increased by 2.46±0.05 Å to 8.95±0.17 Å in 5.5 ns to 12.6 ns corresponding to 11.0 h to 30.5 h of computational time. The total computational time of 59.6 h for MD simulations without H++ computed pKa values was comparable to 66.0 h for MD simulations with H++ pKa values. This shows that the pKa calculation method did not have a significant influence on MD simulation performance. MD simulations with AMBER03 resulted in RMSD values of 5.10±0.16 Å to 13.66±0.25 Å. MD simulation took 18.42 ns to 30.0 ns which corresponds to a total computational time of 156 h. For A03, the MD time step had to be reduced to 1 fs for intramolecular and to 2 fs for intermolecular forces to avoid simulation failure. Structure instability also prevented the transition to MD simulation step 3 for VLP A, which is elucidated by a

fluctuating RMSD curve in Figure 4.5. Furthermore, VLP C did not reach a WoS within 30 ns. Both results indicated that using AMBER03 resulted in less stable simulations compared to simulations with YASARA2. Simulations with H++ or YASARA computed pKa values using the YASARA2 force field have shown the best performance based on simulation time, simulation stability and overall completion of the 3-step MD simulation method. This indicates that MD simulations evaluated in this study benefitted from the empirical data that is embodied in a force field containing knowledge-based potentials (Krieger et al., 2009; Sippl, 1990). Evaluation of this method based on other (refined) force fields and other software platforms would give more detailed insight into simulation performance.

4.3.3 Zeta Potential

Zeta potential was experimentally determined for all three HBcAg VLP constructs and compared to *in silico* determined total surface charge based on the HBcAg dimer structures. This was done to determine the applicability of the prepared structures for computational surface property extraction. Surface charge was extracted as the HBcAg VLP structures only differ in the surface exposed MIR. Therefore, it was assumed that the observed zeta potential differences occur due to the changes on the outer surface of the entire VLP structure (Lošdorfer Božič, Siber, & Podgornik, 2012). The obtained *in silico* surface charge extracted from the homology model and three different WoS, for each of the three chimeric HBcAg dimer structures, are shown in Figure 4.6. Linear transformation of *in silico* data was applied to obtain comparable scales and different MD simulation refinement settings were used to determine the effects on *in silico* generated data and the respective correlation to experimentally determined zeta potential. Linear transformation resulted in ranking three VLPs according to their zeta potential. Figure 4.6 shows that zeta potentials of complete cVLPs were ranked correctly by all dimer structures, which causes overlaying symbols at [-11.70, -11.70] and [-7.94, -7.94]. The main difference is seen for VLP C, which has an experimental zeta potential of-9.50±0.69 mV. This data point was used to evaluate the influence of pKa value computation method and force field selection on *in silico* surface charge calculations.



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

Figure 4.6: In silico computed zeta potential (mV) plotted against experimentally determined zeta potential (mV). Symbols represent *in silico* data based on the homology structure ("Homology", red open circle), Window of Stability (WoS) obtained without H++ and with YASARA2 ("WoS w/o H++", purple diamond), WoS obtained with H++ and AMBER03 ("WoS A03", purple square), and WoS obtained with H++ and YASARA2 ("WoS", blue filled circle). The diagonal line represents theoretical data with a Pearson correlation coefficient of 1 (PCC = 1). X-axis error bars represent the median absolute deviation (MAD) of experimental data and y-axis error bars represent MAD for *in silico* data points. For each *in silico* data series the PCC and mean squared error (MSE) are calculated (n = 3) and listed.

The evaluation parameters, PCC and MSE, are listed for each data series in Figure 4.6. A PCC value above 0.900 indicates a strong linear dependency with experimental data (Rodgers & Nicewander, 1988). This was seen for all evaluated data series because of the limited dataset size, but small differences were observed for VLP C's surface charge. WoS simulated without H++ pKa values and WoS with H++ pKa values showed the highest PCC, with values of 0.954 and 0.946, respectively. Transformed VLP C surface charges for WoS w/o H++ (-8.44±1.18) and WoS with H++ (-8.33±1.43) were also comparable, which resulted in a 0.07 MSE difference in favor of WoS w/o H++. The WoS transformed surface charge distribution, represented by the MAD, shows an overlap between these two values. This indicates there is no significant influence of the used pKa value computation methods in correlation to experimental data. This result was reproducible (data shown in Appendix B, Figure S4.2 to S4.5). Transformed surface charges based on the homology structure (-10.89) and WoS A03 (-10.82±1.11) showed a

weaker correlation than the WoS previously discussed. This is shown by MSE values of 0.64 and 0.58, respectively. Linear dependency is also weaker compared to the other two WoS, where the homology structure showed a PCC of 0.915 and WoS obtained with AMBER03 showed a PCC of 0.922. As mentioned during the discussion of the MD simulations, VLP A did not complete step 2 and VLP C did not reach a WoS when the AMBER03 force field was used during MD simulation. Presumably this also caused the decreased correlation to experimental data. This leads to the conclusion that for this case study the largest positive effect was obtained with the YASARA2 force field, regardless of the used pKa values, when evaluating the correlation between *in silico* HBcAg dimer surface charge and complete cVLP zeta potential. The observed force field effect should be confirmed with a larger dataset. Nevertheless, results indicate that surface properties extracted from structures obtained with the presented pipeline can represent experimental behavior. It should be noted that the applicability of chimeric dimer 3-D structure surface charge to quantitatively predict complete cVLP zeta potential lies outside the scope of this case study, and should be investigated using a more diverse sample space.

All evaluated WoS show a relatively large coefficient of variation (10%–16%) regarding the *in silico* zeta potential, which means there is a significant variation in protein surface property value within the WoS. For example, VLP A simulated with H++ pKa values and the YASARA2 force field resulted a maximum *in silico* zeta potential of -5.74 mV and minimum of -11.07 mV within its 2 ns WoS. This emphasizes cautiousness regarding the use of a single MD simulation snapshot because a snapshot can theoretically take any random value within the WoS. The use of a single snapshot can decrease correlation accuracy and thereby reduce the reliability of computational protein structure-based models. Therefore, a robust central tendency describing statistic which is less sensitive for outliers, such as the median (Andersen, 2008), is considered appropriate for the extraction of protein surface property information within a WoS. The presented computational pipeline did not only show the potential of a high-throughput approach for 3-D structure preparation, but also how a WoS can provide an objective MD simulation termination to reduce computational effort and a robust descriptor extraction platform. The approach could be used for other proteins, such as antibodies, and other prediction targets, such as

assembly competence, solubility, or surface hydrophobicity. A variety of proteins and other prediction targets should be investigated to determine the full potential of the proposed computational 3-D structure preparation pipeline.

4.4 Conclusion

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

a larger dataset. The simple, standardized, and automated framework allows for the implementation of the computational pipeline in manufacturability and formulability screening studies for early candidate assessment.

Acknowledgements

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

Appendix B: Supplementary Material

The Supplementary Material associated with this chapter contains the following information:

- S4.1 Quality Parameters
- S4.2 Reproducibility of Simulation

5

Integrated Process for Capture and Purification of Virus-Like Particles: Enhancing Process Performance by Cross-Flow Filtration

Nils Hillebrandt[a,*], Philipp Vormittag[a,*], Nicolai Bluthardt[a], Annabelle Dietrich[a], Jürgen Hubbuch[a,**]

[a] Institute of Process Engineering in Life Sciences, Section IV: Biomolecular Separation Engineering, Karlsruhe Institute of Technology, Fritz-Haber-Weg 2, 76131 Karlsruhe, Germany

* Contributed equally

** Corresponding author

Abstract

Virus-like particles (VLPs) are emerging nanoscale protein assemblies applied as prophylactic vaccines and in development as therapeutic vaccines or cargo delivery systems. Downstream processing (DSP) of VLPs comes both with challenges and opportunities, depending on the complexity and size of the structures. Filtration, precipitation/re-dissolution and size-exclusion chromatography (SEC) are potent

technologies exploiting the size difference between product and impurities. In this study, we therefore investigated the integration of these technologies within a single unit operation, resulting in three different processes, one of which integrates all three technologies. VLPs, contained in clarified lysate from *Escherichia coli*, were precipitated by ammonium sulfate, washed, and re-dissolved in a commercial cross-flow filtration (CFF) unit. Processes were analyzed for yield, purity, as well as productivity and were found to be largely superior to a reference centrifugation process. Productivity was increased 2.6-fold by transfer of the wash and re-dissolution process to the CFF unit. Installation of a multimodal SEC column in the permeate line increased purity to 96% while maintaining a high productivity and high yield of 86%. In addition to these advantages, CFF-based capture and purification allows for scalable and disposable DSP. In summary, the developed set-up resulted in high yields and purities, bearing the potential to be applied as an integrated process step for capture and purification of *in vivo*-assembled VLPs and other protein nanoparticles.

5.1 Introduction

Vaccination has reduced morbidity and mortality world-wide, especially since the introduction of the World Health Organization's Expanded Programme on Immunization (Greenwood, 2014). Expansion of the vaccine portfolio by virus-like particles (VLP) has opened up new opportunities, such as the prevention or treatment of cancer (Bolli et al., 2018; Bryan et al., 2016; F.-X. Ding et al., 2009; Goldinger et al., 2012; Klamp et al., 2011; Lizotte et al., 2016; Mohsen, Heath, et al., 2019; Mohsen, Vogel, et al., 2019; Palladini et al., 2018). However, especially VLP downstream processing (DSP) faces major challenges, such as low yields and the lack of platform processes or rapid analytical techniques. This is due to the complexity of the product and the associated processes, resulting in high development and production costs (Ladd Effio & Hubbuch, 2015). The structural properties of VLPs are similar or identical to the corresponding virus structure they are derived from (Zeltins, 2013). Composed of at least one type of viral structural protein, they are in a size range of approximately 25 nm to 200 nm (Chung et al., 2010; Reiter et al., 2019). Incorporation of foreign epitopes into VLP-forming viral structural proteins results in so-called chimeric VLPs (Pumpens & Grens, 2001). In a previous study, we observed that upon insertion of smaller peptides, the size of chimeric hepatitis B core antigen (HBcAg) VLPs remained comparable to native HBcAg VLPs with a diameter of 31±2 to 33±3 nm (Rüdt, Vormittag, Hillebrandt, & Hubbuch, 2019; Selzer & Zlotnick, 2017). During production, the size difference between VLPs and host cell proteins (HCPs) as well as other smaller contaminants can be exploited for DSP of VLPs (Ladd Effio & Hubbuch, 2015).

A typical VLP production process is shown in Figure 5.1 including unit operations such as centrifugation, filtration, and chromatography. Bind and elute chromatography, the work horse in biopharmaceutical manufacturing for capture, purification, and polishing, suffers from low dynamic binding capacities (Ladd Effio & Hubbuch, 2015), diffusion limitations (Kramberger et al., 2015), and often too small pore sizes (Kattur Venkatachalam, Szyporta, Kiener, Balraj, & Kwang, 2014) for the purification of VLPs. Size differences between VLPs and the bulk of host cell contaminants can be exploited by size-sensitive techniques such

as size-exclusion chromatography (SEC) – especially for analytical purposes (Ladd Effio, Hahn, et al., 2016) –, precipitation, filtration, and ultracentrifugation (Ladd Effio & Hubbuch, 2015). While ultracentrifugation is applied to lab-scale processes (Ausar et al., 2006; X. Jiang et al., 1992; Mason et al., 1996), scalability and variability issues, among others, hamper its application to industrial-scale processes (Kleiner et al., 2015; Koho et al., 2012).

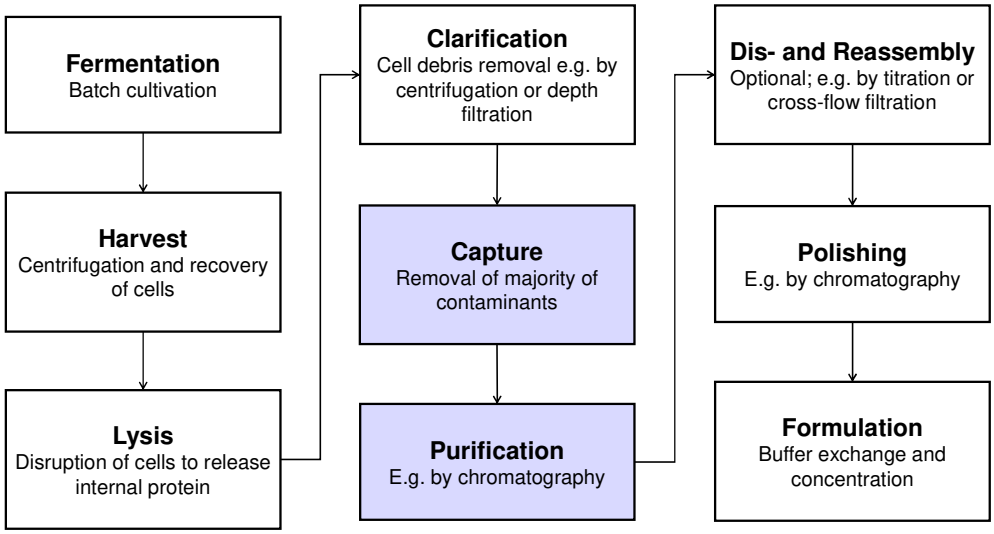


Figure 5.1: Typical production process for intracellularly produced, *in vivo*-assembled virus-like particles (VLPs). Virus structural proteins can be expressed in a variety of host systems, such as *Escherichia coli*, yeast or plant cells (Ladd Effio & Hubbuch, 2015). After harvest and lysis, cell debris are removed by solid-liquid separation and the VLPs remain in solution. VLPs are then captured and purified, followed by an optional dis- and reassembly step, which has shown to increase VLP stability, homogeneity and immunogenicity (Klamp et al., 2011; Mach et al., 2006; Q. Zhao, Allen, et al., 2012). Finally, the product is polished and formulated. The process steps that were investigated as integrated unit operations in this study are highlighted in blue.

Originally developed for the fractionation of blood by Edward Cohn and coworkers in the 1940s (Cohn, 1941; Cohn et al., 1946), precipitation of contaminants or native precipitation of the product are promising alternatives for protein separation and purification (Martinez, Spitali, Norrant, & Bracewell, 2019). In this context, native precipitation has been reported as highly selective for VLPs (H. J. Kim et al., 2010; Koho et al., 2012; Tsoka et al., 2000; Zahin et al., 2016), since larger proteins or protein assemblies are more susceptible to precipitation (Rothstein, 1993). The steric exclusion effect associated with the frequently applied

precipitant polyethylene glycol (PEG) generally leads to steeper slopes in the precipitation curves for larger proteins (Iverius & Laurent, 1967; Sim et al., 2012). For precipitation with kosmotropic salts, surface charge is however thought to have a greater effect than size (Curtis et al., 1998). Separation of product-containing precipitate and supernatant can be achieved by centrifugation or filtration. While PEG has been successfully applied to VLP precipitation (Koho et al., 2012; Tsoka et al., 2000), its application is limited when filtration is used as solid-liquid separation technique, as filtration performance is impaired by a PEG-induced viscosity increase (Z. Li & Zydney, 2017; Plisko, Bildyukevich, Ososky, & Volkov, 2016). Next to PEG of various molecular weights, the kosmotropic salt ammonium sulfate ($(NH_4)_2SO_4$) is a commonly applied precipitant (Kazaks et al., 2017; H. J. Kim et al., 2010; Zahin et al., 2016). In a study on adenovirus (Schagen et al., 2000), dead-end filtration has been applied to retain $(NH_4)_2SO_4$-precipitated virus but exhibited only 46-61% recovery from the filter. As an alternative to dead-end filtration, cross-flow filtration (CFF) in diafiltration (DF) mode has been applied to recover precipitated monoclonal antibodies (mAbs) (Hammerschmidt, Hobiger, & Jungbauer, 2016; Kuczewski, Schirmer, Lain, & Zarbis-Papastoitsis, 2011; Venkiteshwaran, Heider, Teysseyre, & Belfort, 2008). Precipitate was retained by a microfilter, allowing for a wash in DF mode. In CFF, turbulent flow along the membrane surface ensures better recovery from the filter (Davies & Smith, 2010), also reducing concentration polarization and fouling (van Reis & Zydney, 2007). A main advantage of precipitate recovery by CFF over centrifugation lies in avoiding the compaction of precipitate that occurs during centrifugation, which allows for shorter precipitate re-dissolution times using CFF (Hammerschmidt et al., 2016). Additionally, in the above-mentioned studies, precipitation and wash were conducted as integrated CFF-based process steps that showed a higher wash efficiency as compared to centrifugation (Hammerschmidt et al., 2016; Kuczewski et al., 2011). In these studies, the precipitate was re-dissolved by dilution.

This said, it seems promising to dissolve precipitated product by DF into a re-dissolution buffer. Product could subsequently be recovered in the permeate stream as it passes the microfilter. Implementing this approach, the permeate can be separated into fractions allowing for purity increase and concentration adjustment by strategic pooling while undissolved contaminants are retained by the microfilter.

In our experience with DSP of *Escherichia coli* (*E. coli*)-derived VLPs, HCP reduction poses a minor challenge as compared to nucleic acid depletion, demanding for a purification method to reduce the nucleic acid burden. One commonly applied strategy is the supplementation of lysate with Benzonase, a nucleic acid digestion enzyme (U.S. Patent No. 5,173,418, 1992). In recent years, a novel multimodal SEC (mmSEC) medium Capto Core 400/700 has been developed that found successful application in the purification of VLPs, decreasing impurity levels significantly (Lagoutte et al., 2016; Somasundaram et al., 2016; D. Zhao et al., 2015). Integration of a precipitation, wash, and re-dissolution step on a CFF system together with this novel mmSEC medium seems therefore promising.

In the light of the above, the objective of our study was to develop an integrated membrane-aided precipitation, wash, and re-dissolution process for capture and purification of VLPs. The set-up was realized on a commercial CFF unit coupled to a basic preparative chromatography system for monitoring of ultraviolet (UV) absorbance at 280 nm and fractionation. Three process variants were developed, the simplest of which comprised precipitation, wash, and re-dissolution within an integrated CFF-based set-up (Figure 5.2, Process *Basic*). To improve product purity, this method was further either extended by installation of a Capto Core 400 column in the CFF permeate line (Process *mmSEC*) or by pretreatment of the lysate with Benzonase prior to the precipitation step (Process *Nuclease*). As a model VLP, a C-terminally truncated chimeric HBcAg VLP was investigated. The three process variants were compared to a centrifugation-based precipitation, wash and re-dissolution process (Process *Reference*).

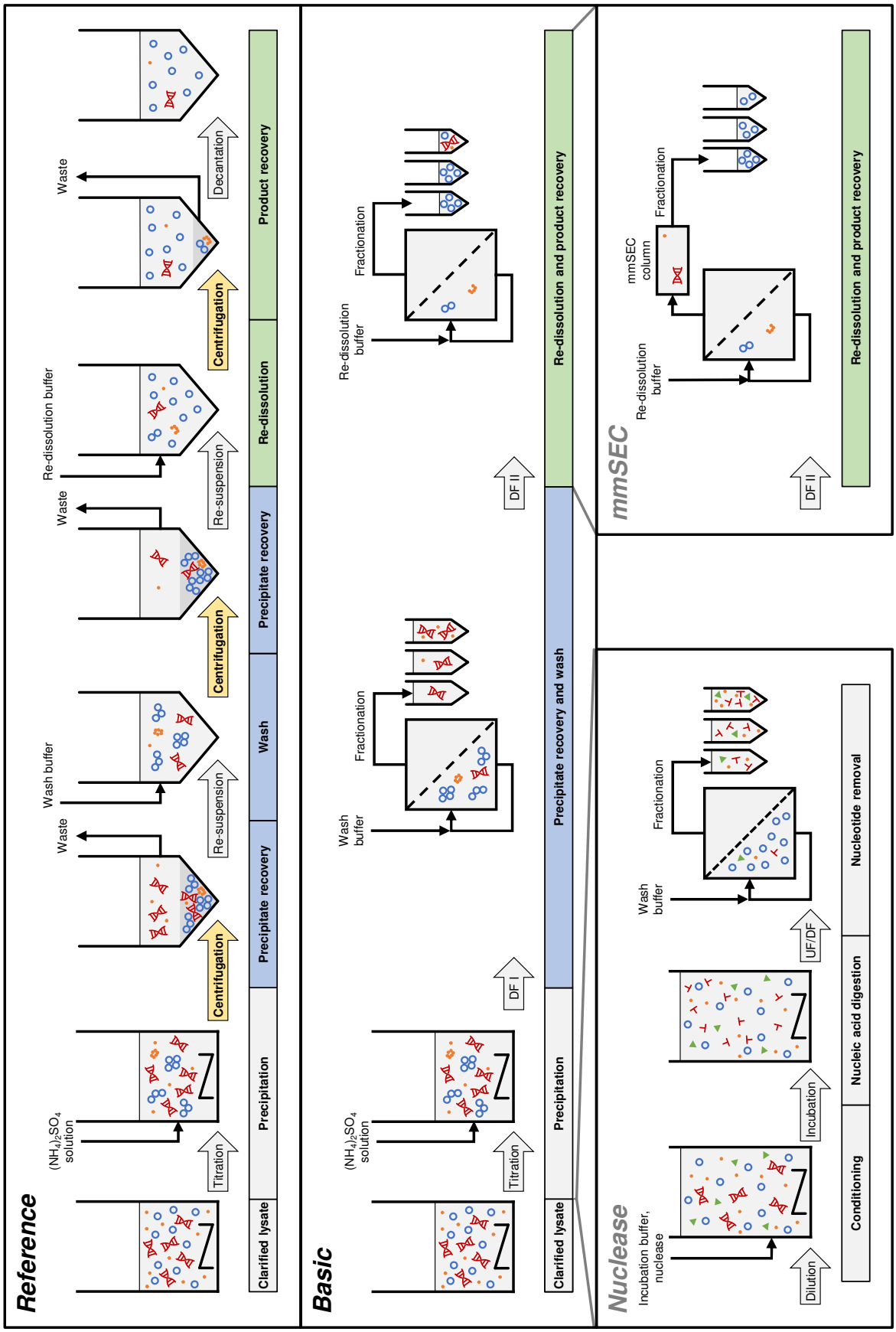


Figure 5.2: Schematic overview of the processes investigated in this study. The *Reference* process is shown at the top, consisting of

centrifugation-based precipitation, wash, and re-dissolution. Process transfer to a cross-flow filtration (CFF) unit resulted in the *Basic* process. Transferred process steps are wash and re-dissolution, highlighted in blue and green, respectively. Wash and re-dissolution are multiple process steps consisting of repeated centrifugation (highlighted in yellow) in the *Reference* process. In the *Basic* process, these are reduced to two consecutive diafiltration (DF) steps by simply switching between diafiltration buffers (Figure 5.3). Alternative CFF process variants, either *Nuclease* or *mmSEC*, are modifications from the *Basic* process. The *Nuclease* process adds a nucleic acid digestion and a 300 kDa wash step preceding precipitation and continues like the *Basic* process. The *mmSEC* process sequence is identical to the *Basic* process sequence but has a modified re-dissolution step (DF II) including a multimodal size-exclusion chromatography (mmSEC) column in the permeate line. $(NH_4)_2SO_4$: ammonium sulfate; HCP: host cell protein. UF/DF: ultrafiltration/diafiltration, VLP: virus-like particle.

5.2 Materials and Methods

5.2.1 Materials, Buffers, and VLPs

All chemicals were purchased from Merck Millipore (Darmstadt, DE), unless otherwise stated. Solutions and buffers were prepared with ultrapure water (PURELAB Ultra, ELGA LabWater, Lane End, UK). A buffer consisting of 50 mM Tris, 100 mM NaCl, 1 mM EDTA (AppliChem GmbH, Darmstadt, DE), pH 8 was used as lysis buffer. The wash buffer was created from lysis buffer that was adjusted to 0.25% (v/v) polysorbate 20 (AppliChem GmbH, Darmstadt, DE) with a 10% (v/v) polysorbate 20 stock solution and to 150 mM $(NH_4)_2SO_4$ (AppliChem GmbH, Darmstadt, DE) with a 1 M $(NH_4)_2SO_4$ stock solution. In the *Nuclease* process and respective experiments, the digestion and nuclease wash buffers were both 50 mM Tris at pH 8, containing 20 mM NaCl, 0.2 mM EDTA, and 2 mM $MgCl_2$. The re-dissolution buffer was 50 mM Tris at pH 8 for all experiments. All buffers were pH-adjusted with 32% HCl. BioNTech Protein Therapeutics generously provided the chimeric HBcAg VLP plasmid. HBcAg was expressed in *E. coli* and liberated by lysis as described in Appendix C, Supplementary Material S5.1. Its extinction coefficient at 280 nm of 1.558 L g^{-1} cm^{-1} was derived from the web-tool ProtParam (Gasteiger et al., 2005) and used for all methods. *E. coli* lysate was diluted to ensure a consistent HBcAg content, resulting in HBcAg concentrations between 2.60 g/L and 2.66 g/L, used as lysate for all processes and experiments.

5.2.2 Precipitation and Re-Dissolution Screening

For processes *Reference*, *Basic*, *mmSEC*, and *Nuclease*, optimal parameters for the precipitation were determined in screening experiments. Screening experiments for precipitant concentration were performed at a small scale in reaction tubes. Lysate was used either untreated or pretreated. Pretreatment comprised overnight dialysis with Slide-A-Lyzer G2 cassettes (10 kDa, 3 mL, Thermo Scientific, Rockford, US-IL) into the digestion buffer with or without addition of >114 U/mL of Benzonase (Sigma Aldrich, Saint Louis, US-MO) to the lysate. In 1.5 mL reaction tubes, 170 µL or 200 µL of these solutions, adjusted to 0.25% (v/v) polysorbate 20, were mixed with different volumes of $(NH_4)_2SO_4$ stock solution and incubated for 30 min at room temperature (RT), which was between 22 °C and 23 °C for all experiments. The solution was spun down at 17000 rcf for 2 min in a tabletop centrifuge and the supernatant was recovered. For screening of the incubation time during precipitation, untreated lysate was precipitated in a 20 mL batch, sampled at 10 min intervals, and treated as described above.

Small-scale re-dissolution experiments were conducted to test the influence of solution components on re-dissolution efficiency. Pooled fractions F3-F11 of the *mmSEC* process were concentrated to 7.74 g/L using 50 mL VivaSpins with 100 kDa molecular weight cut-off (MWCO) (Sartorius Stedim Biotech GmbH, Göttingen, DE). In 1.5 mL tubes, 0.5 mL of concentrated HBcAg solution was mixed with 0.5 mL of five different solutions. Solutions were a) 200 mM NaCl, 50 mM Tris, 2 mM EDTA, pH 8.0, b) 40 mM NaCl, 50 mM Tris, 2 mM EDTA, pH 8.0, c) 200 mM NaCl, 50 mM Tris, 0.4 mM EDTA, 4 mM $MgCl_2$, pH 8.0, d) supernatant of the precipitation step during the *Reference* (Section 2.5) process, and e) supernatant of the wash step during the *Reference* process. Solutions were adjusted to 0.25% (v/v) polysorbate 20 and then to 150 mM $(NH_4)_2SO_4$ for precipitation. Samples were incubated for 30 min at 300 rpm and 23 °C in a thermo-shaker Thermomixer comfort (Eppendorf, Hamburg, DE) and subsequently centrifuged at 15294 rcf in an Eppendorf 5810R centrifuge for 20 min at 20 °C. Supernatant was removed by pipetting. A volume of 1 mL re-dissolution buffer was added and the pellet was resuspended. The reaction tubes were incubated at 10 rpm at RT in an overhead shaker LD-79 (Labinco, Breda, NL) for

60 min, centrifuged with identical settings, and the supernatant was recovered.

5.2.3 Cross-Flow Filtration Instrumentation and Set-Up

The CFF precipitation, wash, and re-dissolution set-up (Figure 5.3) was based on a KrosFlo Research KRIIi CFF system with automatic backpressure valve (Spectrum Labs, Rancho-Dominguez, US-CA) with a stirred cell (Sartorius Stedim Biotech GmbH, Göttingen, DE) as reservoir, and 0.2 µm 200 cm^2 Hydrosart or 300 kDa MWCO 200 cm^2 polyether sulfone (PES) membranes (both Sartocon Slice 200) with corresponding membrane holders (all Sartorius Stedim Biotech GmbH, Göttingen, DE). The three stirred cell inlet ports were connected to retentate, wash buffer, and re-dissolution buffer lines. A Sensirion Liquid Flow Meter SLS-1500 (Sensirion AG, Stäfa, CH) was installed at the permeate outlet of the membrane holder and connected with a 1/16" polyether ether ketone capillary with 0.75 mm inner diameter to the wash valve of an ÄKTA Start (GE Healthcare, Uppsala, Sweden). On-line ÄKTA Start UV sensor data were converted to on-line concentration data applying Beer's law using the HBcAg extinction coefficient. The permeate was fractionated in either 15 mL (wash) or 5 mL (re-dissolution) fractions in 15 mL tubes (Corning, Reynosa, MX-TAM). In all presented filtration processes, a constant permeate flow rate of 2 mL/min was set and maintained using the automatic backpressure valve either by manual valve control (Process *Basic*) or automatic control (Processes *mmSEC* and *Nuclease*). Therefore, the backpressure valve controller was fed with flow rate data of the flow meter (at >1 Hz) instead of transmembrane pressure data as in normal operation mode using a custom-written communication MATLAB 2018b script (The Mathworks, Natick, US-MA). Flow rate, path, and control were optimized in pre-experiments, and data were temporally aligned considering delay volumes (for more detail see Appendix C, Supplementary Material S5.2).

5.2.4 Precipitation, Wash, and Re-Dissolution Process by Cross-Flow Filtration

Diluted lysate, adjusted to 0.25% (v/v) polysorbate 20, was filled into the aforementioned stirred cell with three inlets and two outlets. One

outlet was capped with an injection plug (Fresenius Kabi, Bad Homburg, DE) for sampling, the other outlet either closed or connected to the suction port of the CFF feed pump. A Minipuls 3 peristaltic pump (Gilson, Villiers le Bel, FR) was used to pump 1 M $(NH_4)_2SO_4$ solution at 1 mL/min through one of the inlet ports of the cell up to a final concentration of 150 mM $(NH_4)_2SO_4$ (Figure 5.3). The flow rate was monitored using a Sensirion Liquid Flow Meter SLS-1500. The stirred cell was set to minimal stirring speed. The solution was incubated for 30 min at RT. During incubation, 250 µL samples were taken every 10 min.

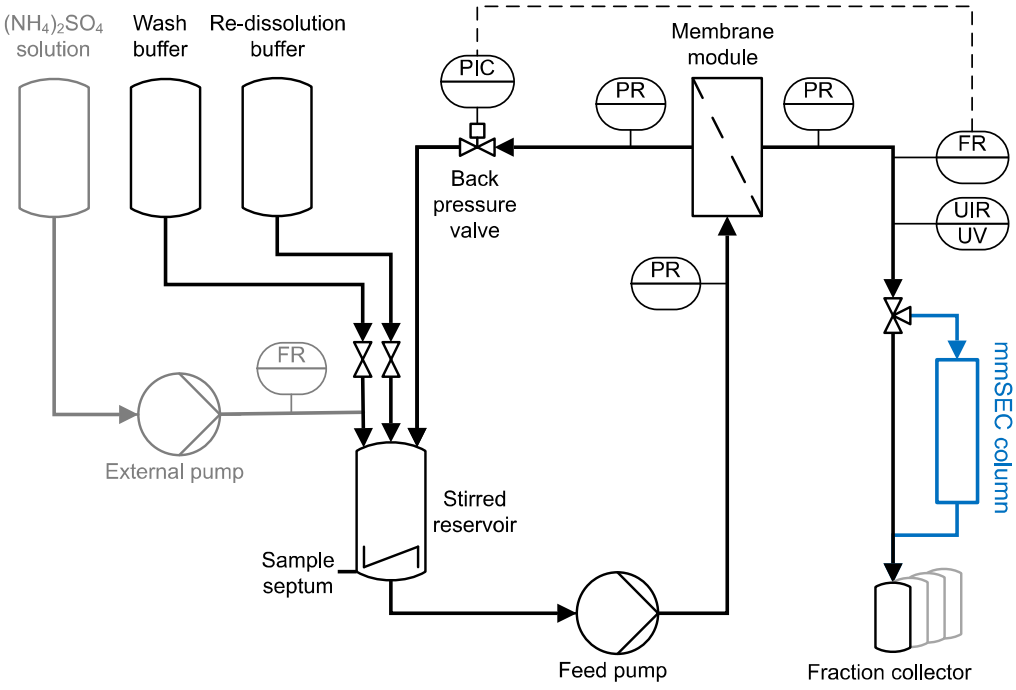


Figure 5.3: Piping and instrumentation diagram of the precipitation and cross-flow filtration (CFF) setup. The set-up used for wash and re-dissolution of the CFF processes *Basic* and *Nuclease* is shown. For process *Nuclease*, the depicted set-up was used with different membranes (300 kDa and 0.2 µm) for the respective wash steps. The *mmSEC* process included an additional multimodal size-exclusion chromatography column (mmSEC) in the permeate stream, highlighted in blue. The precipitation set-up consists of the components highlighted in gray on the left and the stirred reservoir. Precipitant was ammonium sulfate ($(NH_4)_2SO_4$). Gray highlighted components were removed after completion of precipitation. C: control; F: flow rate; I: indicate; P: pressure; R: record; U: multivariable; UV: ultraviolet.

Three wash and re-dissolution process variants were examined, referred to as *Basic*, *mmSEC,* and *Nuclease* (Figure 5.2). The *Basic* process consisted of wash and re-dissolution of precipitate suspension by constant

volume DF against wash and re-dissolution buffer, respectively, and fractionation of the permeate. CFF feed flow rate in all filtration steps was 30 mL/min. Compared to the *Basic* process, the *mmSEC* process included a Capto Core 400 HiScreen column (GE Healthcare, Uppsala, SE) with a nominal column volume of 4.7 mL in the permeate line downstream of the fractionation valve of the ÄKTA Start (Figure 5.3). The *Nuclease* process was conducted like the *Basic* process with additional pretreatment of the lysate prior to precipitation. The lysate was diluted 1:5 with a buffer containing 50 mM Tris and 2.5 mM $MgCl_2$ at pH 8 to optimize the conditions for the digestion of nucleic acids by Benzonase, resulting in the composition of the digestion buffer. Benzonase was added to a concentration of ≥114 Units/mL and incubated overnight for 16 h at 80 rpm and 23 °C in a 225 mL tube in a MaxQ 6000 Shaker (Thermo Scientific, Marietta, US-OH). The solution was concentrated fivefold by ultrafiltration (UF) in the CFF unit with the 300 kDa membrane. The solution was diafiltered for five diafiltration volumes (DV) using nuclease wash buffer. The permeate of UF and DF was fractionated into 15 mL fractions. The retentate was processed analogous to the lysate in the other processes.

5.2.5 Centrifugation-Based Wash and Re-Dissolution

In a centrifugation-based process (Figure 5.2, process *Reference*), precipitation was performed identically to the experimental procedure for the CFF runs, whereas wash and re-dissolution were performed as a centrifugation protocol. The suspension of 20 mL was centrifuged at 17387 rcf at 20 °C for 20 min. Supernatant was removed and the pellet was resuspended. The procedure including centrifugation and resuspension was repeated with re-dissolution buffer. The suspension was transferred into a stirred cell and stirred at minimal speed. After 1 h, 2 h, and 3 h, a sample was taken, spun down at 17000 rcf for 2 min in a table top centrifuge Heraeus Pico 17 (Thermo Electron LED GmbH, Osterode am Harz, DE), and the supernatant was recovered.

5.2.6 Analytical Characterization

Size-exclusion chromatography was coupled with a diode array detector (DAD), multi-angle light scattering (MALS), and quasi-elastic light scattering (QELS) to quantify and specify differently sized species. An

Agilent BioSEC-5 4.6 × 300 mm, 5 µm, 1000 Å column (Agilent, Santa Clara, US-CA) was used at a Dionex Ultimate 3000 RS UHPLC system controlled by Chromeleon version 6.8 SR15 (Thermo Fisher Scientific, Waltham, US-MA). The method was isocratic for 14 min at a flow rate of 0.4 mL/min with 50 mM potassium phosphate buffer at pH 7.4. The injection volume was 20 µL. The outlet of the DAD was connected to a Dawn Heleos 8 MALS/QELS system (Wyatt Technology Corporation, Santa Barbara, US-CA). MALS and QELS data were analyzed with the ASTRA V software (Version 5.3.4.15, Wyatt Technology Corporation, Santa Barbara, US-CA) and resulted in root mean square radius (rms) and molecular weight (both assessed by MALS) and hydrodynamic radius (assessed by QELS). For protein separation and quantitation, a Caliper LabChip GX II (PerkinElmer, Waltham, US-MA) high-throughput capillary gel electrophoresis (HT-CGE) device was employed. An HT Protein Express LabChip and the corresponding HT Protein Express Reagent Kit were used and results analyzed with LabChip GX software (Version 4.2.1745.0, PerkinElmer, Waltham, US-MA). Analyses were performed using the HT Protein Express 200 assay in reduced mode using dithiothreitol (DTT, Amresco, Solon, US-OH) according to the assay standard operation procedure provided by the manufacturer. For data analysis, all peaks of 21.5 ± 1 kDa were regarded as HBcAg monomers, which is the form in which HBcAg is present after sample preparation. The range derived from experiments with pure HBcAg. For SDS PAGE, LDS sample buffer, MES running buffer, and NuPage 4-12% BisTris Protein Gels were used and run on a PowerEase 500 Power Supply (all Invitrogen, Carlsbad, US-CA) in reduced mode with 50 mM DTT in the sample solution according to the manufacturer's manual with minor adaptations. The gel was stained with a Coomassie blue solution. CFF re-dissolution samples of fractions with maximum concentration were analyzed by transmission electron microscopy (TEM) on a Fecnei Titan^3 80 – 300 microscope (FEI company, Hillsboro, US-OR). Samples were adjusted to 0.5-1 g/L with ultrapure water and filtered with a 0.2 µm syringe filter. Sample preparation and image analysis were conducted similarly to previous studies with chimeric HBcAg VLPs (Rüdt et al., 2019). Hydrophilization and staining solutions were 1% (w/v) alcian blue 8GX (Alfa Aesar, Ward Hill, US-MA) in 1% acetic acid solution and 2% ammonium molybdate(VI) (Acros Organics, Geel, BE) solution (pH 6.25, adjusted with NaOH), respectively.

5.2.7 Calculation of Yield, Purity, and Productivity Measures

The yield Y of a process was calculated by

$$Y = \frac{\sum_{i=start}^{end} m_{F_i}}{m_{lysate}}, \tag{5.1}$$

where m_{lysate} is the mass of HBcAg, calculated from the processed lysate volume and HBcAg concentration as determined by HT-CGE, and m_{F_i} is the mass of HBcAg in re-dissolution fraction F as determined by SEC, where fractions were considered from fraction F_{start} to F_{end}. HT-CGE purity was determined by the ratio of HBcAg concentration to total protein concentration in HT-CGE samples. SEC purity was calculated by the ratio of HBcAg peaks to total peak area at 280 nm (for details on peak identification, the reader is referred to Appendix C, Supplementary Material S5.3). A260/A280 was calculated by dividing the cumulated peak areas at 260 nm by the cumulated peak areas at 280 nm. Absolute spatial productivity P was calculated by

$$P = \frac{m_{HBcAg,recovered}}{t_{process}}, \tag{5.2}$$

where $m_{HBcAg,recovered}$ is the accumulated mass of pooled fractions and $t_{process}$ the time to complete the process starting with precipitated material through to recovery of the product. Relative spatial productivity was derived by the ratio of absolute productivities to the absolute productivity of the *Reference* process.

5.3 Results

5.3.1 Precipitation

In pre-experiments, 150 mM $(NH_4)_2SO_4$ was determined as optimal concentration for all process variants, where most of the product is found in the precipitate. Figure 5.4 shows HT-CGE and SDS PAGE data of the clarified supernatant of small-scale precipitation experiments from I) lysate, II) lysate with added Benzonase dialyzed against digestion buffer overnight, and III) lysate dialyzed against digestion buffer over night without addition of Benzonase. The total protein concentration in the supernatant (Figure 5.4A) was higher for almost all $(NH_4)_2SO_4$ concentrations for precipitation from untreated lysate than for dialyzed

samples, as had been expected due to depletion of molecules during dialysis. HBcAg concentrations in all three experiments (Figure 5.4B) were comparable, except for the region between 100 mM and 150 mM $(NH_4)_2SO_4$, where supernatant HBcAg concentrations during precipitation from non-dialyzed lysate dropped significantly at 100 mM $(NH_4)_2SO_4$, while the dialyzed samples remained at comparably constant HBcAg concentrations from 0 mM to 100 mM $(NH_4)_2SO_4$. SDS PAGE analysis (Figure 5.4C) showed similar results based on band intensities.

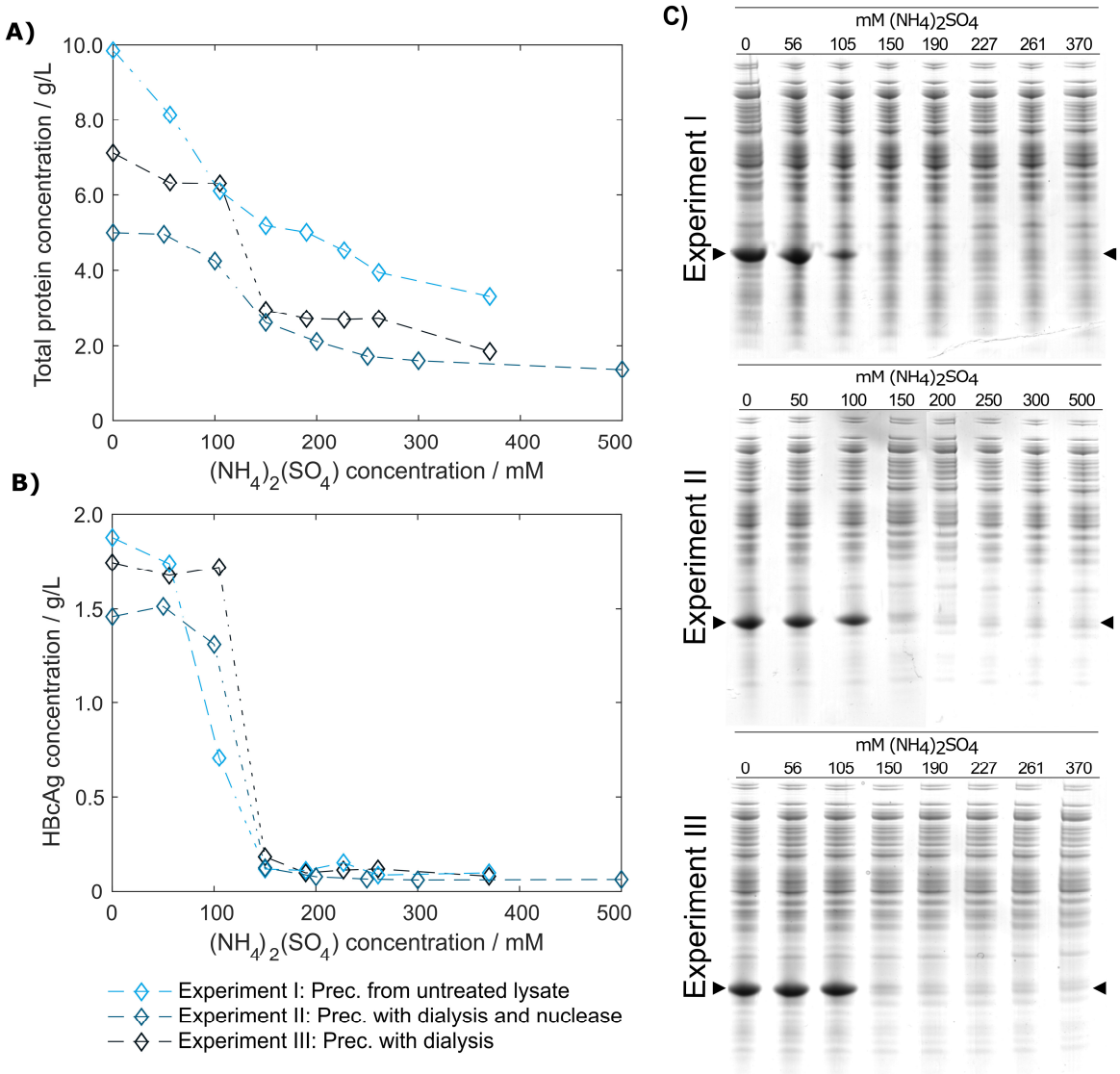


Figure 5.4: Total protein and hepatitis B virus core antigen (HBcAg) concentration in the supernatant after precipitation depending on ammonium sulfate ($(NH_4)_2SO_4$) concentration. Total protein concentration by reducing high-throughput capillary gel electrophoresis (HT-CGE) is shown in **(A)**, HBcAg concentration by HT-CGE in **(B)**. Experiments I-III represent precipitation (Prec.) from I) lysate (-◇-), II) lysate with added Benzonase dialyzed

against digestion buffer overnight (-◊-), and III) lysate dialyzed against digestion buffer overnight without addition of Benzonase (-◊-). Experiments I-III are also shown as reducing SDS PAGE scans **(C)**, where lanes 1-8 show $(NH_4)_2SO_4$ concentrations. The HBcAg band is indicated by arrows.

To validate that precipitation incubation time is sufficient at larger scale, HBcAg concentration in the supernatant was investigated in 10 min intervals at the previously determined 150 mM $(NH_4)_2SO_4$. Precipitation of HBcAg was already completed directly after addition of $(NH_4)_2SO_4$, judging visually based on SDS PAGE scans (Figure 5.5). It has to be noted that to the first sampling time 2-3 min have to be added, accounting for drawing of samples, transferring the samples into reaction tubes, and centrifugation of the samples. Interestingly, during titration of the untreated lysate with $(NH_4)_2SO_4$, we observed a rapid increase in turbidity when a concentration of 100 mM $(NH_4)_2SO_4$ was exceeded. Nevertheless, 150 mM $(NH_4)_2SO_4$ and a precipitation duration of 30 min were chosen to include a safety margin, which was successful in all processes.

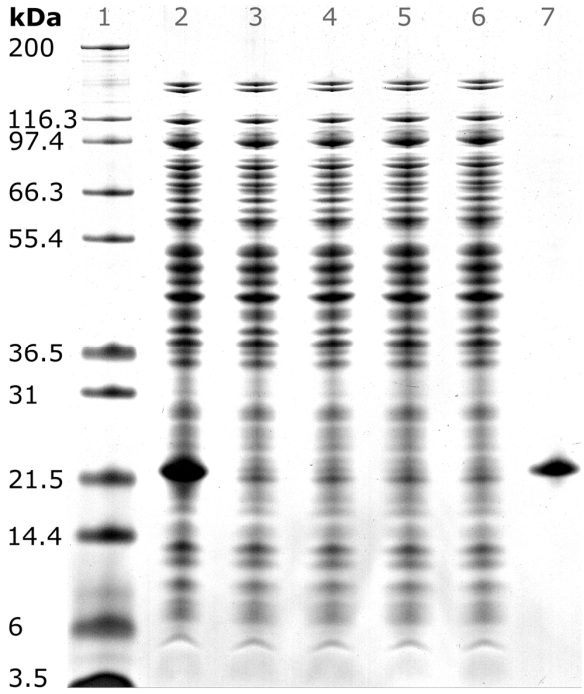


Figure 5.5: SDS PAGE scan of 1) Invitrogen Mark 12 Unstained Standard, 2) hepatitis B virus core antigen (HBcAg)-containing *E. coli* lysate, 3-6) supernatant of precipitation experiments with 150 mM ammonium sulfate directly, 10, 20, and 30 min after ammonium sulfate addition, and 7) pure chimeric HBcAg sample. Molecular weights of the proteins contained in the standard are shown on the left.

5.3.2 Centrifugation-Based *Reference* Process

After precipitation, solid-liquid separation aims at separating the contaminant solutes and precipitation buffer from the precipitated product. A wash step increases the efficiency of contaminant removal. The *Reference* process was based on centrifugal solid-liquid-separation for precipitate recovery, wash, and re-dissolution. HBcAg concentration of re-dissolution supernatant increased over the first 3 h and was 1.67 g/L, 1.80 g/L, and 1.85 g/L, respectively (Figure 5.6A). Table 5.1 shows the re-dissolution concentration and purity measures after 3 h, where SEC purity was 76%, HT-CGE purity was 83%, and A260/280 was 0.87. After precipitation, which was conducted identically for all CFF processes and the *Reference* process, the *Reference* process was completed in 4.5 h. Time-specific productivities of all processes were calculated based on mg HBcAg per hour relative to the *Reference* process productivity. Therefore, the relative productivity of the *Reference* process is 100%, as shown in Table 5.1. Assuming a similar area foot print of the unit operations, a spatial component of the productivity was neglected.

Table 5.1. Summary of re-dissolution process data for centrifugation (*Reference*) and cross-flow filtration (*Basic*, *mmSEC*, *Nuclease*) processes. Process data above the thin horizontal border are calculated based on a pool of all fractions. Results below this border are based on a fraction pool that aimed for a product concentration of at least 1 g/L and a maximum yield. This was not possible for the *Nuclease* process. Values are calculated using total hepatitis B virus core antigen concentrations except A260/A280, which is based on all species in the size-exclusion chromatography (SEC) chromatogram (Appendix C, Supplementary Material Figure S5.3.1). Best results of each table column are underlined.

	Mass†	**Yield**‡	**Conc.**†	**SEC purity**†	**A260/ A280**†	**HT-CGE purity**	**Relative productivity**†
	mg	%	gL^{-1}	% Area	-	%	%
Reference	30.73	72	<u>1.85</u>	76	0.87	83	100
Basic CFF,§	36.26	82	0.38	73	1.02	96	264
mmSEC CFF,§	<u>37.82</u>	<u>86</u>	0.34	96	0.73	96	239
Nuclease CFF,§	9.72	22	0.18	86	0.82	<u>98</u>	8
Basic CFF,¶	25.19	57	1.01	78	0.96	95	248
mmSEC CFF,‖	30.01	68	1.00	<u>98</u>	<u>0.70</u>	96	<u>269</u>

[CFF] cross-flow filtration process, [†] assessed by SEC, [‡] for definition see Material and Methods Equation (4.1), [§] Pool of all fractions, [¶] Pool of fractions F3-F7, [||] Pool of fractions F3-F8. Process data for pools were calculated by accumulating fraction process data. A260/A280: absorbance ratio of the sample at 260 nm to 280 nm; Conc.: concentration; HT-CGE: high-throughput capillary gel electrophoresis; SEC: size-exclusion chromatography

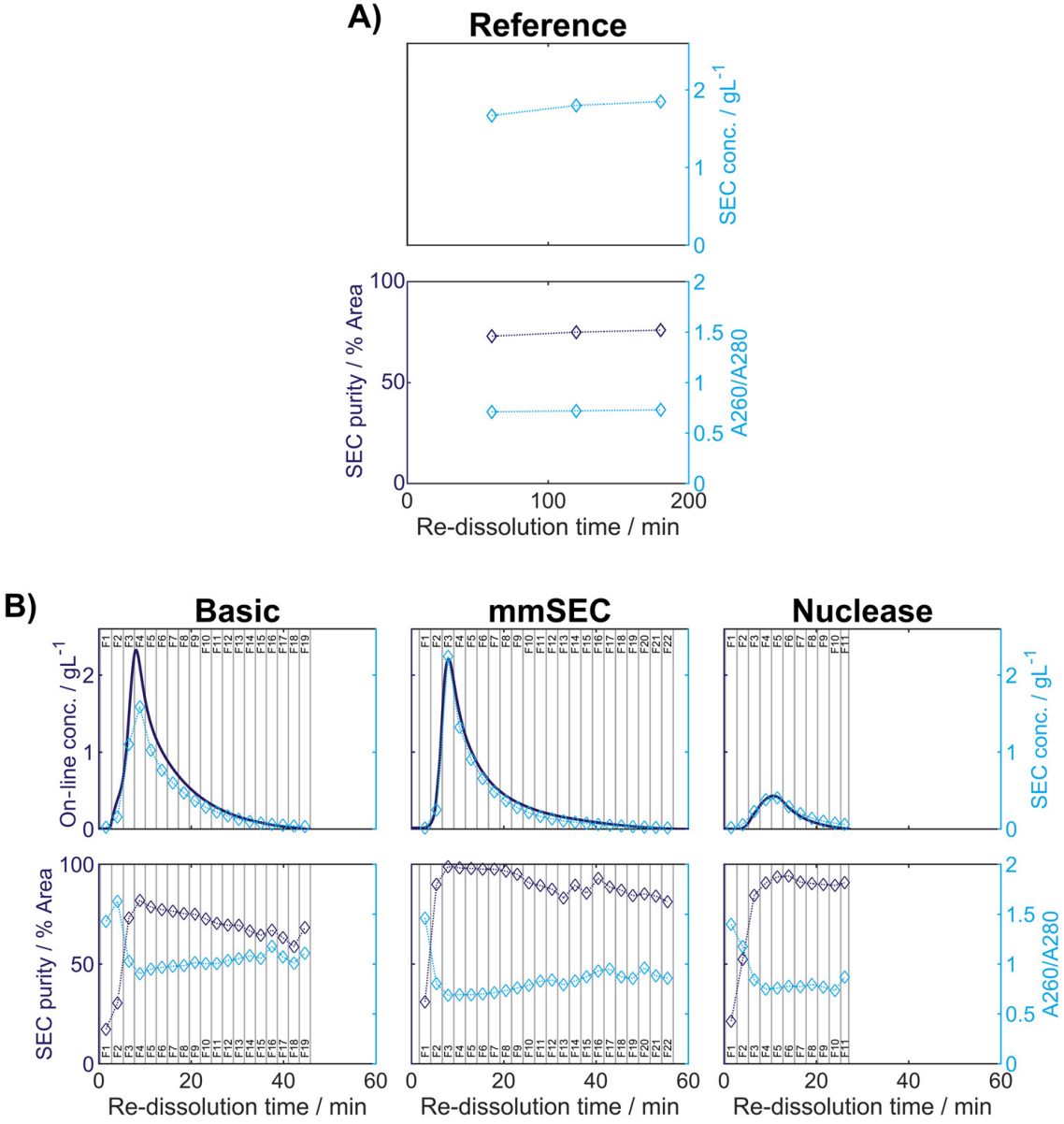


Figure 5.6: Re-dissolution protein concentration (conc.) and purity. Each figure column represents a re-dissolution process variant: **(A)** *Reference* and **(B)** *Basic*, *mmSEC* and *Nuclease*. In subfigure **(A)**, the *Reference* process concentration and purity data is shown based on off-line analysis of the supernatant after centrifugation. Top row: Off-line concentrations (◊) were derived from size-exclusion chromatography (SEC) peak areas of hepatitis B virus core antigen (HBcAg) species (Appendix C, Supplementary Material S5.3). Bottom row: SEC purity (◊) is defined as percentage of HBcAg peak

area at 280 nm with respect to the area of all SEC peaks at 280 nm. A260/A280 (◊) is defined as quotient of the cumulated SEC peak areas at 260 and 280 nm, respectively. Dotted lines are added to guide the eye. In subfigure **(B)**, on-line monitoring of the permeate concentration and off-line analysis of the corresponding permeate fractions (F, indicated by vertical lines) during cross-flow filtration (CFF) are shown. The metrics of subfigure **(A)** are shown in subfigure **(B)** using the same symbols. Additional to these metrics, protein concentrations (–) are shown. Protein concentrations are based on absorbance at 280 nm assuming the chimeric HBcAg extinction coefficient.

5.3.3 Cross-Flow Filtration-based Wash and Re-Dissolution Processes – On-Line Monitoring and Off-Line Analysis

While in the centrifugation-based *Reference* process, wash, re-dissolution, and product recovery steps have to be performed individually (Figure 5.2, *Reference*), the CFF set-up allows for process step integration. Diafiltration with a wash buffer retains the product while depleting solutes continuously. Diafiltration into a re-dissolution buffer replaces the wash/precipitation buffer and re-dissolves the product, which is then able to pass the 0.2 µm membrane. This additionally ensures that larger particles, such as insoluble precipitate, are removed by retention. The developed set-up facilitates fractionation of the permeate stream enabling individual analysis of the fractions (Figure 5.3).

In the presented CFF processes, the wash step was stopped when the initially saturated on-line UV absorbance in the permeate fell below 4 mAU (for visualization of this process see Appendix C, Supplementary Material S5.4). Product loss during the wash step was determined by HT-CGE. HBcAg concentrations in wash fractions were 0.02-0.03 g/L. The additional wash step prior to precipitation of the *Nuclease* process resulted in less than 0.1 mg HBcAg loss (analyzed by SEC). After precipitation and wash, re-dissolution of the product was initiated by switching DF buffer lines from wash buffer to re-dissolution buffer. Figure 5.6B depicts on-line and off-line process data over time for the re-dissolution step in the three CFF process variants. Upon DF into re-dissolution buffer, on-line permeate concentrations for all process variants increased to a maximum after a lag phase of nearly 2 min and subsequently decreased exponentially. The process was stopped as soon as the on-line absorbance dropped below 4 mAU (on-line concentration

of 0.01 g/L). The final retentate was analyzed for unrecovered product by HT-CGE. It showed a negligible HBcAg mass of <0.5 mg for processes *Basic* and *mmSEC*, as opposed to 22.4 mg in the *Nuclease* process. The maximum on-line concentrations were 2.3 g/L, 2.2 g/L, and 0.4 g/L for processes *Basic*, *mmSEC* and *Nuclease*, respectively. The curve shapes of the off-line HBcAg concentration are in good agreement with the on-line data. In all three CFF processes, SEC purities were the lowest in fraction F1 and constantly increased to the purity maximum which coincided with the concentration maximum. Maximum purities were 82%, 99%, and 94% for processes *Basic*, *mmSEC,* and *Nuclease*, respectively. The SEC A260/A280 coefficient showed a nearly inverse progression compared to SEC purity data.

5.3.4 Comparison of Process Data

As seen from summarized process data (Table 5.1), processes *Basic* and *mmSEC* showed higher HT-CGE purities and VLP yields compared to the *Reference* process. SEC purity was comparable between the *Reference* and the *Basic* process, while it was highest for the *mmSEC* process. The *mmSEC* process also showed lowest A260/A280 with 0.73. The relative productivities of processes *Basic* and *mmSEC* were higher than the *Reference* and the *Nuclease* process with >239%. While processes *Basic* and *mmSEC* were superior with regard to aforementioned process data, their concentrations were lower with 0.34-0.38 g/L as compared to 1.85 g/L for the *Reference* process. To increase pool concentrations, higher concentrated fractions can be selected for pooling. Strategic pooling increased concentrations for processes *Basic* and *mmSEC* to 1 g/L while maintaining purity and productivity. However, the yield decreased to 57-68%. Overall, the *mmSEC* process showed highest recovered mass, yield, SEC purity, and lowest A260/A280, along with high productivity and HT-CGE purity, both for strategic pooling and pooling of all fractions.

The *Nuclease* process showed great product loss during re-dissolution, as mentioned above. It exhibited the lowest yield and relative productivity of 22% and 8%, respectively. Due to low concentrations, purity is not comparable to the other processes. For completeness, these values are plotted in Figure 5.6B and shown in Table 5.1. Compared to the other processes, the precipitation process following nuclease treatment started

with altered solution conditions regarding NaCl, $MgCl_2$, EDTA, and impurity concentrations. Five screening experiments were designed to investigate the influence of solution conditions during precipitation on re-dissolution efficiency. The recovery of HBcAg in the re-dissolution experiments was 82±1%, indicating no significant difference in HBcAg recovery between the investigated experimental conditions.

5.3.5 VLP Size Analysis

SEC, coupled to DAD, MALS, and QELS, detected three peaks attributed to HBcAg (compare Appendix C, Supplementary Material S5.3 for peak identification). A main peak was identified with 15.3-15.5 nm rms radius and 16.4-17.7 nm hydrodynamic radius, corresponding to 79-84% of the HBcAg peak area in the CFF processes. In the *Reference* process, it was 65%. The two earlier-eluting peaks showed 24.4-25.2 nm and 30.4-32.0 nm radius, respectively. The molecular weights were 3.8-4.1 MDa, 7.5-7.8 MDa, and 12.2-12.7 MDa for the three peaks in ascending order by radius. Figure 5.7 shows TEM micrographs of the processes *Basic*, *mmSEC*, *Nuclease*, and the *Reference* process. Graphical analysis resulted in average radii of 13.4±1.2 nm, 14.6±1.5 nm, 13.6±1.2 nm, and 15.3±1.8 nm, respectively, not showing distinct species as observed in SEC. While samples from processes *mmSEC* and *Reference* showed a spatially equal distribution of VLPs, *Basic* and *Nuclease* samples appeared clustered.

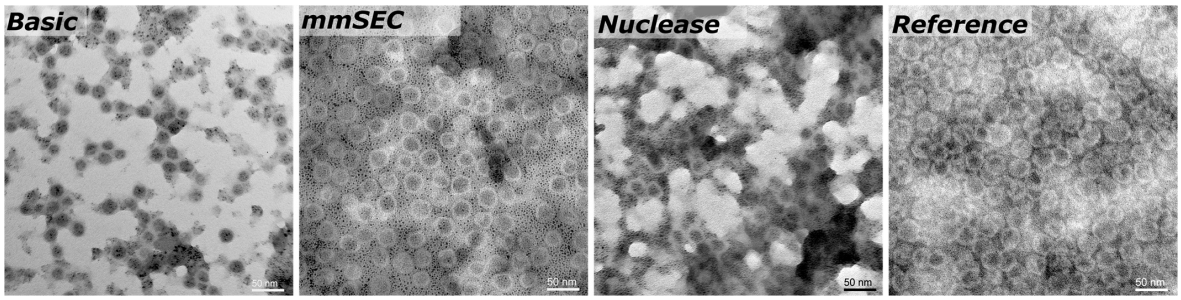


Figure 5.7: Transmission electron microscopy micrographs of re-dissolution peak samples of four processes: *Basic*, *mmSEC*, *Nuclease*, and the *Reference* centrifugation process. The magnification was 27,000-fold.

5.4 Discussion

5.4.1 Interpretation of Analytical Methods

In this study, SEC and HT-CGE have been applied to determine concentrations and to identify the quantified species. It is therefore important to discuss the meaning of the analytical data as determined for the presented processes. HT-CGE has been employed as, compared to SDS PAGE, a high-throughput compatible and quantitative size-dependent concentration analytical technique. HT-CGE purity informs about the relative HBcAg fraction of the total protein content, i.e. HBcAg protein purity. SEC is applied to assess particle size and molecular weight, HBcAg and contaminant concentrations, and additionally provides spectral data of the sample.

The ratio of the absorbance at 260 nm to the absorbance at 280 nm (A260/A280) is characteristic for the ratio of nucleic acid to protein concentration, whereby higher A260/A280 values indicate a larger fraction of nucleic acids (Wilfinger, Mackey, & Chomczynski, 1997). SEC purity describes purity based on all species absorbing at 280 nm, such as proteins and nucleic acids.

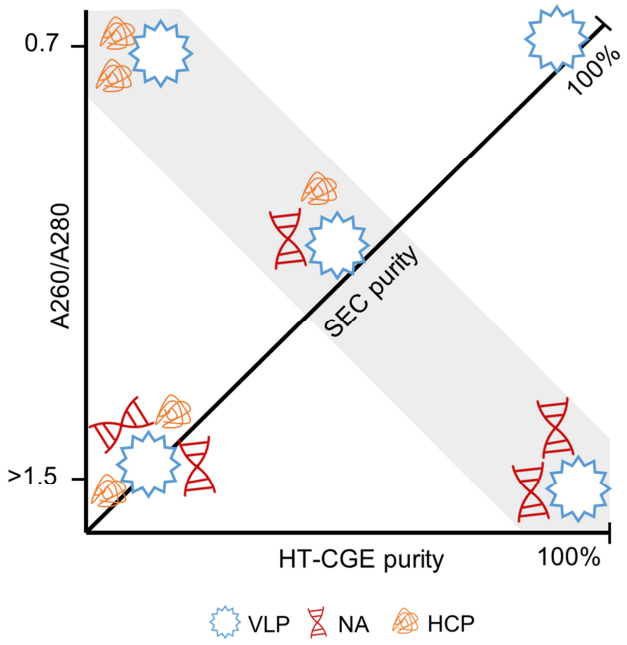


Figure 5.8: Illustration of the interdependence of derived purity measures. Virus-like particles (VLPs) with different degree of contamination by host cell proteins (HCPs) and nucleic acids (NAs) are shown. Size-exclusion chromatography (SEC) provides the A260/A280 (ordinate) and SEC purity (diagonal axis). A high-

throughput capillary gel electrophoresis (HT-CGE) protein assay provides the HT-CGE purity (abscissa). The gray highlighted area is characterized by identical SEC purity, while HT-CGE purity and/or A260/A280 describe the composition of the contaminants. A pure hepatitis B virus core antigen VLP sample is characterized by 100% SEC purity, 100% HT-CGE purity and an A260/A280 of ~0.7

The combination of these two purity measures together with the A260/A280 are thus seen to be powerful to describe a sample. Figure 5.8 illustrates the connection between these measures. For example, samples with high HT-CGE purity but lower SEC purity therefore probably also show increased A260/A280 values, indicating nucleic acid contamination. It is important to note that SEC measurements are more accurate than HT-CGE measurements for concentration determination. This being said, SEC could only be applied to rather clean, non-turbid samples (see also Appendix C, Supplementary Material S5.1). Therefore, SEC rather was applied to assess concentrations during re-dissolution while lysate and precipitation/wash samples were assessed by HT-CGE. Yields were calculated based on lysate HBcAg concentrations and re-dissolution sample concentrations and are therefore based on both HT-CGE and SEC measurements. Discussion on comparability of yields can be found in Appendix C, Supplementary Material S5.2.

Off-line SEC and HT-CGE analysis indicated that mainly HBcAg species pass through the membrane upon re-dissolution. It was therefore reasonable to convert the on-line UV absorbance into an on-line HBcAg concentration value, applying the HBcAg coefficient. The good agreement between on-line and off-line concentration profiles underlines the usefulness of this approach. However, the *mmSEC* process set-up included an additional purification step between the UV flow cell and the fraction collector, making off-line samples purer than the on-line measured permeate stream.

The MALS detector coupled to the SEC system provides an estimate of molecular weight. HBcAg capsids naturally occur as 180-mer with icosahedral symmetry T=3 and as 240-mer with symmetry T=4 (Wynne et al., 1999). As SEC is incapable of separating different capsid symmetries, the molecular weight measured is the average weight of T=3 and T=4 capsid species. The theoretical molecular weight for a chimeric T=4 capsid is 4.8 MDa and a T=3 capsid is 3.6 MDa. The SEC-MALS-derived molecular weights of the latest-eluting HBcAg peak were between

3.8 and 4.1 MDa, representing 18%/82% and 43%/57% mixture of T=3/T=4 capsids, respectively. *In vitro*, HBcAg VLPs are predominantly T=4, but can shift towards higher percentage of T=3 symmetry capsids upon VLP modification (Böttcher, Wynne, & Crowther, 1997; Rybka et al., 2019; A Zlotnick et al., 1996). As an orthogonal method, TEM imaging confirmed the presence of approximately 30 nm sized nearly spherical particles. TEM image-based size measurements did not result in significant differences between the VLP sizes in samples of the different processes. Due to graphical sizing inaccuracies, TEM was unable to resolve different HBcAg species as observed with SEC. These three differently sized HBcAg species, of which the smallest corresponds to the typical size of an HBcAg VLP, were observed in all CFF processes and the *Reference* process. Interestingly, the VLP fraction of these three peaks was similar in all the CFF processes but higher than in the reference process. It would be interesting to analyze these species separately in the following process steps, such as disassembly, which is, however, out of the scope of this study.

5.4.2 Precipitation of Chimeric HBcAg VLPs

Precipitation of complex mixtures involves interactions that are only partly understood (Przybycien, 1998). This has also recently been pointed out in a study on PEG-induced precipitation of mAbs (Großhans, Suhm, & Hubbuch, 2019). Although differences were small in our study, variations of HBcAg concentrations were observed especially at 100 mM $(NH_4)_2SO_4$, where supernatant concentrations after precipitation from untreated lysate were lowest. This is in accordance with previously reported results on mAb precipitation from complex mixtures in the study mentioned above, where precipitation from a complex mixture led to higher precipitation propensity of product molecules (Großhans et al., 2019). This rapid decrease in HBcAg solubility at 100 mM concurs with the observed rapid turbidity increase at 100 mM $(NH_4)_2SO_4$ at a larger scale during the CFF and centrifugation processes. Experiments on precipitation incubation time revealed that the investigated HBcAg VLPs precipitate almost immediately, which is fast compared to incubation times of 15 min – 4 h for different VLPs and precipitants stated in literature (Koho et al., 2012; Schagen et al., 2000; Tsoka et al., 2000).

5.4.3 Product Loss in the *Nuclease* Process

The *Nuclease* process showed significantly lower concentrations of recovered HBcAg, making it difficult to compare this process variant to the other processes. Due to its low relative productivity and comparably complicated process route, it is not competitive with the *Reference* process and the other CFF processes *Basic* and *mmSEC*. The low yield observed in this process is mainly due to incomplete re-dissolution, with 22.4 mg of HBcAg in the final retentate. In order to reveal the effect of different solution conditions during the precipitation step, this was investigated in small-scale re-dissolution experiments. However, no significant differences could be identified when investigating the influence of NaCl, EDTA, $MgCl_2$, and contaminants with regard to this problem. Further reasons could be the additional wash step by DF on a membrane of different material or overnight incubation at RT, resulting in irreversible precipitation. Apart from low yields, its low relative productivity derives from the 16 h Benzonase incubation, yet only increases to 42% if an incubation time of 1 h at optimized digestion conditions would be considered. From a scientific standpoint, it would be interesting to identify which factors contributed to the low re-dissolution yields, whereas from a technical standpoint this process route cannot be justified.

5.4.4 Benefits of Process Transfer to a Cross-Flow Filtration Unit

The main advantage in implementing CFF for precipitation/re-dissolution lies in the combination of product recovery by membrane retention with the capability of exchanging the product-containing buffer in a single process step. During CFF wash steps, impurities smaller than 0.2 µm are expected to be washed out with the permeate. Impurity depletion was observed in all processes indicated by the decrease of on-line UV absorbance. HBcAg VLPs are expected to be retained by the membrane due to the size of their precipitate, as was seen for mAb precipitate in previous studies (Hammerschmidt et al., 2016; Kuczewski et al., 2011). Although HT-CGE results point at minor product loss during wash, it is important to note, that all proteins of 19.5-21.5 kDa were assigned to HBcAg in our analysis due to sizing inaccuracies. Therefore, product loss is expected to be lower than reported. The wash process step was comparable for processes *Basic* and *mmSEC*. Higher protein purities in the CFF processes are probably due to a more efficient

wash as compared to the centrifugation-based *Reference* process, whereby interstitial pellet liquid cannot be removed. However, in the *Basic* process, SEC purity was slightly lower and A260/A280 higher than in the *Reference* process. This indicates that the main impurity in the *Basic* process are nucleic acids. This is in accordance with previous unpublished results of CFF-based processes from our group. It may be suggested that DNA interacts with the VLPs in the kosmotropic environment during precipitation and wash which hampers its depletion during the wash step.

As opposed to re-dissolution of the compact pellet in the *Reference* process, re-dissolution from a turbid solution in CFF-based processes was expected to improve process performance. This was for example observed by the increased yields of processes *Basic* and *mmSEC* compared to the *Reference* process. Product loss in the *Reference* process can be attributed to unrecoverable interstitial pellet liquid and high precipitate compaction (Hammerschmidt et al., 2016), which leads to slower and incomplete re-dissolution. This is in agreement with comparably slow re-dissolution in the *Reference* process. As a result, CFF processes *Basic* and *mmSEC* showed strongly enhanced relative productivities. Additionally, CFF process durations are reduced by minimizing manual handling compared to the *Reference* process. The *mmSEC* process showed superior SEC purity compared to all other processes. As discussed above, the main contaminant in the *Basic* process are nucleic acids. These were efficiently depleted in the *mmSEC* process, leading to excellent purity, while maintaining the increased yield of the *Basic* compared to the *Reference* process, underpinning the usefulness of the mmSEC column in the permeate line (Figure 5.3).

In summary, process transfer to the CFF set-up led to improved yields, accelerated re-dissolution kinetics, and process intensification by integrating multiple process steps into one unit operation. Compared to literature VLP processes showing a 31-76% recovery (Carvalho, Silva, Moleirinho, et al., 2019; D. Zhao et al., 2015), up to 95% protein purity (Wetzel et al., 2018), and a 78% nucleic acid reduction (Carvalho, Silva, Moleirinho, et al., 2019), the process data of the *mmSEC* process are comparable or superior while applying only a single unit operation after lysate clarification. The main drawback of the CFF-based processes were lower product concentrations as compared to the *Reference* process. The

exponential permeate concentration decrease observed for all re-dissolution processes, as expected for non-retained species in DF (Kurnik et al., 1995), results in decreased concentrations when aiming for a maximized process step yield. Although the re-dissolution concentration profile cannot be improved from a technical point of view, this effect can be ameliorated by strategic pooling. This was exemplified by creating 1 g/L pools, which resulted in improved purity and 18-25% yield decrease. Alternatively, collection of all fractions followed by a concentration process via UF could maximize both yield and concentration. Another interesting option would be loading the permeate onto an anion exchange column or membrane as a polishing step to bind VLPs, deplete $(NH_4)_2SO_4$, and achieve further purification from other contaminants while obtaining concentrated VLPs in the elution step. While it seems reasonable to dissolve the precipitated product by dilution to avoid DF-associated concentration decrease, DF shows several advantages. Considering 0% retention, 40% of $(NH_4)_2SO_4$ is theoretically found in fractions 1-2, which could be discarded due to low VLP concentrations. On the contrary, all $(NH_4)_2SO_4$ remains in the product solution for re-dissolution by dilution as used in several concepts for mAb capture processes (Hammerschmidt et al., 2016; Kuczewski et al., 2011; Z. Li, Gu, Coffman, Przybycien, & Zydney, 2019). This drawback may be circumvented by employing dead-end filtration to drain precipitate before re-dissolution (Chen et al., 2016; W. Liu et al., 2019; Lohmann & Strube, 2020). This approach was not considered in this study to avoid unknown effects of draining, precipitate compaction on the membrane, and uncontrolled concentration increase on product stability and yield. DF allows for highly efficient $(NH_4)_2SO_4$ removal in the retentate enabling maximum re-dissolution and therefore yield. Conversely, comparable levels of $(NH_4)_2SO_4$ can only be reached by dilution to very large volumes. Especially if a UF step is established after re-dissolution, a simple DF step after concentration can remove residual $(NH_4)_2SO_4$ efficiently.

To the best of our knowledge, this is the first study to present a fully integrated CFF system-based precipitation, wash and re-dissolution set-up for VLP capture and purification that includes DF-based re-dissolution. The presented approach showed exceptionally good performance with regard to yield, purity, and productivity while being based on a simple lab-scale set-up with basic commercial devices. As a

filtration-based process, it exhibits good scalability and the possibility of disposable manufacturing (van Reis & Zydney, 2007). For vaccines, especially cancer vaccines, which are envisaged to be produced as personalized medicine (Buonaguro, Aurisicchio, Buonaguro, & Ciliberto, 2013; Castiblanco & Anaya, 2015; Rammensee & Singh-Jasuja, 2013), this highly efficient, easy-to-control, and scalable process could enable distributed manufacturing of personalized protein nanoparticle-based therapeutics.

5.4.5 Considerations for Method Transfer

From a technical point of view, CFF process control of the presented method can be achieved by maintaining a constant transmembrane pressure (TMP) or permeate flow rate. In case of TMP-based control, low TMP values are required to obtain the target permeate flow rate due to the large membrane pore size of 0.2 µm. During wash and re-dissolution in processes *Basic* and *mmSEC*, the TMP was in the range of 0.01 bar to 0.02 bar. Therefore, a careful adjustment of the TMP is recommended to avoid exceeding the maximum flow rate of the mmSEC column. Nevertheless, a constant flow rate is advantageous for fractionation and mmSEC separation.

The prerequisites for the successful application of this process to the purification of other VLPs are the ability I) to precipitate the target product, II) to retain the majority of impurities in solution, III) to re-dissolve the product, and IV) to avoid electrostatic or hydrophobic interaction between product and impurities or matrices, such as the membrane material. These prerequisites are probably fulfilled – to varying degrees – for most non-enveloped VLPs.

Precipitation of the target product might require adaption of the precipitant concentration or agent for different VLPs. From unpublished results of our group, we learned that the precipitation of other chimeric HBcAg VLPs required ammonium sulfate concentrations of 0.1 M to 1 M. Their large size compared to the typical contaminants facilitates the precipitation of VLPs while retaining most impurities in solution. The application of this process to smaller product molecules (such as capsomers) could also be feasible, if a suitable precipitation method is developed, which retains impurities in solution. Product re-dissolution and hydrophobic or electrostatic interactions are influenced by the

solution conditions, which might need to be optimized, presumably with a focus on the optimum solution pH.

Compared to the here investigated non-enveloped VLPs, enveloped VLPs might pose a challenge due to their lower stability (Dai et al., 2018). VLPs derived from other hosts such as yeast or plants require changes in the lysis procedure and bring along a different impurity profile than *E. coli*. This said, the separation in the presented process is largely based on the size difference between product and impurities, which should be comparable for other hosts. Extracellularly produced VLPs could benefit from the higher purity of the starting material and therefore potentially result in yet higher purities using this process. Conclusively, the transfer of this method to the purification of other VLPs probably requires few adaptations, mainly regarding the development of optimal solution conditions for VLP precipitation and re-dissolution in small scale.

5.5 Conclusion and Outlook

In this study, we have developed a set-up for integrated capture and purification of VLPs within a CFF unit. Clarified lysate was precipitated, washed, and re-dissolved. Three CFF process variants were investigated and characterized for yield, purity, and relative productivity and were compared to a centrifugation-based *Reference* process. Process transfer of the *Reference* process to the CFF unit led to increased purities, probably attributed to a more efficient wash step. The *mmSEC* process, integrating an additional purification step by an mmSEC column in the permeate line, was superior to all tested variants and the *Reference* process resulting in the highest purity and productivity. As one single unit operation, it compares favorably to entire DSP processes found in the literature and shows great potential for disposable and scalable manufacturing. Another key advantage of CFF processes is the possibility to fractionate the VLP-containing permeate, allowing for efficient pooling with regard to the desired target process data and product analytical profile. In the future, this mainly size-based DSP step could be applied to other VLPs or similarly sized therapeutics with only minor adaptations, laying the foundation for a platform process for protein nanoparticles.

Acknowledgements

The authors would like to thank Matthias Rüdt and Thorsten Klamp for proofreading as well as Steffen Großhans and Sebastian Andris for inspiring discussions. The authors express their gratitude to Reinhard Schneider for technical and scientific support in performing TEM imaging. The authors would also like to thank BioNTech Protein Therapeutics, especially Thorsten Klamp and Anja Wilming, for the provision with VLP sequence data and production plasmids, without which this work would not have been possible.

Appendix C: Supplementary Material

The Supplementary Material associated with this article contains the following information:

- S5.1: Chimeric HBcAg Expression and Cell Lysis
- S5.2: CFF Set-up and Temporal Alignment
- S5:3: SEC Analysis
- S5.4: CFF Wash and Re-dissolution Process Data
- S5.5: Analytical Considerations

6

Ensembles of Hydrophobicity Scales as Potent Classifiers for Chimeric Virus-Like Particle Solubility – an Amino Acid Sequence-based Machine Learning Approach

Philipp Vormittag[a], Thorsten Klamp[b], Jürgen Hubbuch[a] *

[a] Institute of Process Engineering in Life Sciences, Section IV: Biomolecular Separation Engineering, Karlsruhe Institute of Technology, Fritz-Haber-Weg 2, 76131 Karlsruhe, Germany

[b] BioNTech SE, An der Goldgrube 12, 55131 Mainz, Germany

* Corresponding author

Abstract

Virus-like particles (VLPs) are protein-based nanoscale structures that show high potential as immunotherapeutics or cargo delivery vehicles. Chimeric VLPs are decorated with foreign peptides resulting in structures that confer immune responses against the displayed epitope. However, insertion of foreign sequences often results in insoluble proteins, calling for methods capable of assessing a VLP candidate's solubility *in silico*.

The prediction of VLP solubility requires a model that can identify critical hydrophobicity-related parameters, distinguishing between VLP-forming aggregation and aggregation leading to insoluble virus protein clusters. Therefore, we developed and implemented a soft ensemble vote classifier (sEVC) framework based on chimeric hepatitis B core antigen (HBcAg) amino acid sequences and 91 publicly available hydrophobicity scales. Based on each hydrophobicity scale, an individual decision tree was induced as classifier in the sEVC. An embedded feature selection algorithm and stratified sampling proved beneficial for model construction. With a learning experiment, model performance in the space of model training set size and number of included classifiers in the sEVC was explored. Additionally, seven models were created from training data of 24-384 chimeric HBcAg constructs, which were validated by 100-fold Monte Carlo cross-validation. The models predicted external test sets of 184-544 chimeric HBcAg constructs. Best models showed a Matthew's correlation coefficient of >0.6 on the validation and the external test set.

Feature selection was evaluated for classifiers with best and worst performance in the chimeric HBcAg VLP solubility scenario. Analysis of the associated hydrophobicity scales allowed for retrieval of biological information related to the mechanistic backgrounds of VLP solubility, suggesting a special role of arginine for VLP assembly and solubility. In the future, the developed sEVC could further be applied to hydrophobicity-related problems in other domains, such as monoclonal antibodies.

6.1 Introduction

New formats of targeted therapies are emerging, such as virus-like particles (VLPs) (Ong et al., 2017). VLPs are highly immunogenic macromolecular assemblages based on viral proteins, resembling the structure of the virus they were derived from (Kushnir et al., 2012). Since they lack viral nucleic acids, the particles are non-infectious (Chackerian, 2007; Kushnir et al., 2012). VLPs are on the market as vaccines against the virus they were derived from, e.g. human papillomavirus-VLPs against human papillomavirus infection to prevent cervical cancer, or hepatitis B surface antigen-VLPs against hepatitis B virus infection (Bryan et al., 2016; McAleer et al., 1984). An approach increasingly investigated is the display of foreign epitopes on a VLP scaffold resulting in chimeric VLPs (cVLPs) (Ong et al., 2017). They benefit from the inherent immunogenicity of a viral structure coupled with the structure of a foreign target antigenic epitope (Pumpens & Grens, 1999). Hepatitis B core antigen (HBcAg) has been widely applied as a VLP platform for chimeric antigen display due to its excellent stability, successful production in a high diversity of expression systems, and induction of strong B- and T-cell responses (Jegerlehner et al., 2002; Klamp et al., 2011; Pumpens & Grens, 1999). The foreign peptide is typically introduced genetically into the VLP at the N-terminus, C-terminus, or preferably in the major immunodominant region (MIR) (Karpenko et al., 2000; Pumpens & Grens, 2001). However, insertion of a foreign epitope often results in insoluble, misassembled or aggregated capsids, lacking the desired immunogenicity (Billaud et al., 2005; Gillam & Zhang, 2018; Karpenko et al., 2000). The process of identifying soluble cVLP constructs is highly empirical and time-consuming (Chackerian, 2007). While few reports studied cVLP solubility based on sequence data, the number of observations included in these studies is limited (Billaud et al., 2005; Janssens et al., 2010; Karpenko et al., 2000). Early development of cVLPs would therefore greatly benefit from a model to predict solubility which is probed using a large data set.

A variety of general approaches to predict protein solubility exists that are based on information from three-dimensional (3-D) structures and simulations and/or amino acid sequence information. For a detailed overview of aggregation and solubility prediction tools, we refer to a

recent review (Trainor, Broom, & Meiering, 2017). 3-D structure-based methods include the prediction of soluble expression by molecular dynamics (MD)-simulated unfolding combined with a support vector machine (SVM) architecture (Schaller, Connors, Oelmeier, Hubbuch, & Middelberg, 2015), dynamic exposure of hydrophobic patches in MD simulations (Chennamsetty, Voynov, Kayser, Helk, & Trout, 2009; Jamroz, Kolinski, & Kmiecik, 2014), and projection of sequence-based methods onto 3-D structures (Sormanni, Aprile, & Vendruscolo, 2015; Zambrano et al., 2015). Although high-throughput 3-D structure generation of VLP building blocks has been described previously (Klijn, Vormittag, Bluthardt, & Hubbuch, 2019), the computational cost of creating 3-D structures is still high, limiting the applicability of this approach in candidate selection for several hundred molecules. Amino acid sequence-based methods can be distinguished into amino acid composition-based algorithms such as machine learning approaches using SVM or random forest classifiers (Agostini et al., 2012; Magnan et al., 2009; Samak et al., 2012; Xiaohui, Feng, Xuehai, Jingbo, & Nana, 2014; Yang Yang, Niroula, Shen, & Vihinen, 2016) and sliding-window-based algorithms, such as AGGRESCAN, Zyggregator, and CamSol (Conchillo-Sole et al., 2007; Sormanni et al., 2015; Tartaglia et al., 2008). Interestingly, sequence-based methods have been reported to be superior to solvent-accessible surface-based methods in the prediction of monoclonal antibody aggregation (Hebditch, Roche, Curtis, & Warwicker, 2019).

The above-mentioned methods have in common that their goal is to identify proteins or patterns in proteins that are prone to aggregation and therefore have a higher chance to be insoluble upon expression. In the following section, we will discuss why hydrophobic interactions play a special role for VLP solubility and why the application of current models is difficult for the cVLP solubility problem.

HBcAg has been extensively studied in a C-terminally truncated form with amino acids 150-183 removed, termed $Cp_{1\text{-}149}$. It assembles to VLPs while being easier to handle in experiments and processes than the full-length HBcAg. This is attributed to the removal of the strongly positively charged C-terminal amino acids that bind nucleic acids (Alexander et al., 2013; Gallina et al., 1989; Wizemann & von Brunn, 1999; A Zlotnick et al., 1996). The smallest HBcAg species observed in physiological solutions

are dimers, stabilized by an intermonomer disulfide bridge and a hydrophobic core (Wynne et al., 1999). The dimeric $Cp_{1\text{-}149}$ aggregates aggressively and readily forms capsids at low concentrations, neutral pH, and low salt (Ceres & Zlotnick, 2002). Capsid formation is an entropy-driven process relying on hydrophobic interaction and is therefore similar to protein aggregation (Ceres & Zlotnick, 2002; Gorbenko & Trusova, 2011). Since capsids can exist in much higher concentrations in physiological buffers than dimers, the solubility limitation introduced by insertion of a foreign epitope is most probably related to an interference with the assembly reaction (Billaud et al., 2005; Chackerian, 2007). Investigation of chimeric HBcAg VLP assembly by diafiltration showed a dependence of the assembly reaction on the inserted epitope sequence (Rüdt et al., 2019). When the assembly reaction is hampered by the insertion of a foreign epitope, the strong entropic drive for protein-protein interaction probably leads to insoluble aggregates as opposed to soluble capsids. Ordered aggregation of dimers to capsids can therefore be assumed as a prerequisite for high-level soluble expression, whereby hydrophobicity plays a major role. Therefore, an appropriate measure of hydrophobicity is paramount to describing the cVLP solubility problem.

The hydrophobic effect is described by the free energy change of water surrounding a solute. For amino acids in specific, it has been investigated based on organic solvent-water partition coefficients, for example (Nozaki & Tanford, 1971). This partition coefficient might be suitable for the description of solute distribution in such systems. Protein folding, however, is a much more complicated matter influenced by more and different properties of the amino acids than their tendency to accumulate in a certain phase and is still not fully understood (Garde & Patel, 2011; Harris & Pettitt, 2016). To overcome the limitation of this definition of hydrophobicity for biological systems, so-called hydrophobicity scales have been developed. These are, for example, based on the analysis of the distribution of amino acids in the core or surface of the protein (Naderi-Manesh, Sadeghi, Arab, & Moosavi Movahedi, 2001), thermodynamic calculations elevating different aspects of solvation (Chothia, 1976; von Heijne & Blomberg, 1979), or peptide retention times in reversed-phase chromatography (Wilce, Aguilar, & Hearn, 1995). These scales have in common that they try to describe the hydrophobic effect in the interplay with other factors related to geometries and electrostatic contributions. It is therefore important to note that

throughout this manuscript, hydrophobicity is referred to as a value describing the tendency of proteins, amino acids, or functional groups to influence a biological process towards an outcome that is thought to be connected with hydrophobicity, such as aggregation, rather than its strict thermodynamic definition for smaller solutes (Harris & Pettitt, 2016). Hydrophobicity scales assign each proteinogenic amino acid a particular hydrophobicity value. These hydrophobicity values can be used to calculate overall protein hydrophobicity or regions within the molecule. Simm and colleagues (Simm et al., 2016) identified 98 protein hydrophobicity scales in the literature. These scales have been derived using experimental and theoretical techniques based on a great variety of training data, ranging from small to large sets of proteins, peptides, single amino acids, or 3-D structures. Application of a hydrophobicity scale to a new problem requires that an appropriate scale is chosen. This can be based on comparison of the investigated experimental conditions to the framework in which the hydrophobicity scales were derived or the choice of frequently applied hydrophobicity scales. None of these two approaches is recommended as they both introduce bias into the model. Feature selection algorithms can help overcoming this bias and selecting the appropriate scales using a set of training data. In a study on aggregation-prone regions of 354 peptides, feature selection has been successfully employed to derive critical features for peptide or protein aggregation (Y. Fang et al., 2013). The most important 16 critical features were incorporated in SVM and random forest architectures. In another study, an SVM architecture using 40 features was applied (Tian, Wu, Guo, & Fan, 2009). These methods project the problem onto a space of a dimension of the number of feature variables. This could also be applied to hydrophobicity values calculated by several hydrophobicity scales. Another approach is to regard each hydrophobicity scale individually to be included in a classifier in a one-dimensional input data space. Reflecting upon hydrophobicity scales, this is reasonable since each of the scales were derived to be individual measures of hydrophobicity. Considering them individually, the physicochemical meaning behind the scales remains largely unchanged. The strength of this method comes with the combination of several classifiers. This results in potent ensembles that incorporate the classifiers' strengths, while ideally overshadowing their weaknesses in classification (Re & Valentini, 2012). In an article on hydrophobicity scale optimization by a genetic algorithm,

the authors pointed out, that statistical methods may have strong prediction performance, but may not be applicable to new or even to similar problems and small data sets (Zviling, Leonov, & Arkin, 2005). Therefore, preserving physicochemical information in hydrophobicity scales was one important goal in this research.

In summary, ensemble methods based on hydrophobicity scales promise to be a potent tool to describe classification or regression problems related to hydrophobicity. The cVLP solubility problem calls for a method that ascertains critical features of the molecules in an aggregating environment, capable of distinguishing between structures that probably aggregate to soluble VLPs and those that aggregate to insoluble structures. The objective of this study was to create an interpretable protein solubility model framework and to uncover information about the VLP solubility problem that will aid in engineering soluble cVLP candidates. Therefore, a soft ensemble vote classifier (sEVC) was developed and implemented, which consists of individual decision trees, each based on a hydrophobicity scale including an embedded feature selection algorithm. Physicochemical information contained in the hydrophobicity scales was largely conserved by I) using each scale as an individual classifier within an ensemble and II) by implementing a simple one-level decision tree as classifier. Feature selection was implemented to boost model performance and identify the most relevant hydrophobicity scales for chimeric HBcAg VLP solubility. The applicability of the model was evaluated with 568 chimeric C-terminally truncated HBcAg VLP constructs using 91 hydrophobicity scales.

6.2 Materials and Methods

6.2.1 VLP Solubility Data

Chimeric HBcAg VLP constructs were based on His-tagged C-terminally truncated HBcAg (Schumacher et al., 2018). The molecules were created using 82 different peptide inserts and eight different insertion strategies, a total of 691 chimeric VLP constructs, which were experimentally tested for solubility. The peptides are inserted into the HBcAg molecule in the MIR. An insertion strategy defines where exactly in the MIR the peptide is inserted and which amino acids are deleted. Inserts that have not been tested with all eight insertion/deletion strategies were excluded from this

study. The final data set comprised 568 chimeric HBcAg VLPs with all possible combinations of strategies A-H and inserts 1-71. Solubility was evaluated by SDS-PAGE after lysis of the expression host *Escherichia coli* (*E. coli*). Solubility was treated as a binary class system with class labels 'soluble' or 'positive' or '1' and 'insoluble' or 'negative' or '0'. Throughout this paper, an 'observation' is referred to one of the 568 chimeric HBcAg VLP constructs. Class 'soluble' was attributed to 283 of 568 observations, while 285 of 568 were class 'insoluble'. With 49.8%/50.2% class division, the data set can be considered as a balanced classification problem.

6.2.2 Data Set Division

Model training, model evaluation, and data processing were performed with MATLAB R2018a (The Mathworks, Natick, US-MA). Models were always generated by and calculated on randomly selected validation subsets (or with stratified random sampling). In this article, randomization is achieved by using the *randn* command of MATLAB, which generates pseudorandom values. Seven data subsets containing n_{train} observations were created prior to model construction, where $n_{\text{train}} = \{24, 24\ ,24, 48, 96, 192, 384\}$. These data sets were constructed once and the remainder of available data n_{test} was used as an external test set. Observations were drawn from the data set by stratified sampling aiming at a balanced representation of strategies and inserts, i.e. the respective strata. Stratified sampling was achieved by limiting the occurrence of strategies and inserts in the data set. The maximum allowed number of inserts in the sampled data set was $n_{\text{insert,max}} = roundup\left(\frac{n_{\text{train}}}{n_{\text{inserts}}}\right)$, where $n_{\text{inserts}} = 71$. The maximum allowed number of strategies was accordingly $\frac{n_{\text{train}}}{n_{\text{strategies}}}$, where $n_{\text{strategies}} = 8$. When the maximum number of a certain insert or strategy in the training set was reached, all identical inserts or strategies, respectively, were made unavailable to random selection in order to sample the strata evenly.

6.2.3 Hydrophobicity Scales

98 hydrophobicity scales were retrieved from a recent article on peptide classification (Simm et al., 2016), originally derived from AAindex (Kawashima et al., 2007), the SPLIT 4.0 server (Juretić, Trinajstic, & Lucić, 1993), and ProtScale (Gasteiger et al., 2005). Each scale was

centered and scaled to unit variance. Reversed scales were excluded if there was a complementary, non-reversed scale available, resulting in 91 scales (see Appendix D, Supplementary Material Table S6.1).

6.2.4 Hydrophobicity Scale-Based Soft Decision Tree Ensemble Vote Classifier

The model generation comprised a feature selection, an sEVC informed by classifiers based on hydrophobicity scales, and a Monte Carlo cross-validation (MC-CV) procedure. Figure 6.1 illustrates the construction of the sEVC. Feature values were computed from amino acid sequences and hydrophobicity scales. A hydrophobicity scale assigns each amino acid a hydrophobicity value. The sum over the amino acids results in the feature value. 73 amino acids in N- and 71 amino acids in C-terminal direction were omitted in the calculation, as they were identical for all constructs. Each classifier in the sEVC was constructed from feature values calculated for each observation in the training set using one hydrophobicity scale. The individual classifiers based on this feature value were decision trees with one split (also called decision stumps) which were trained based on Gini's diversity index (Gini, 1912; Windeatt & Ardeshir, 2004). This one-level tree design ensures that a simple hydrophobicity threshold decides about the predicted class. Decision trees were constructed using the *fitctree* function of MATLAB's *statistics and machine learning toolbox*. The resulting n_{trees} decision trees assigned each observation a class decision and a class probability. For a hard ensemble vote classifier, the probabilities are equal to 1. In the here applied sEVC, the class probability is the probability estimate derived from the associated child node in the decision tree. The class decision and the associated class probability becomes the decision tree's *vote*. The *votes* of all decision trees for a particular observation are summed up in the sEVC. The class that has a higher sum of probability values is the elected class.

Figure 6.2 shows the procedure for model construction from stratified training set selection, over model selection by MC-CV through to model construction and prediction. Model performance was evaluated by 100-fold MC-CV. During validation, 50% of the data was used for training and the remaining data was predicted. MC-CV samples $n_{\mathsf{vali,train}}$ randomly without replacement. Compared to k-fold cross-validation, the

number of cross-validation groups in MC-CV is not governed by the choice of their sizes, and observations can be sampled in different cross-validation sets. The information on the model performance can then be used to inform about optimal classifier numbers for construction of the model. For the final model, the entire training data set is used for model training and feature selection. The embedded feature selection sorts the features with decreasing feature importance. In 91 models, the best 1-91 classifiers are included. The resulting classifiers are used to predict the external test set.

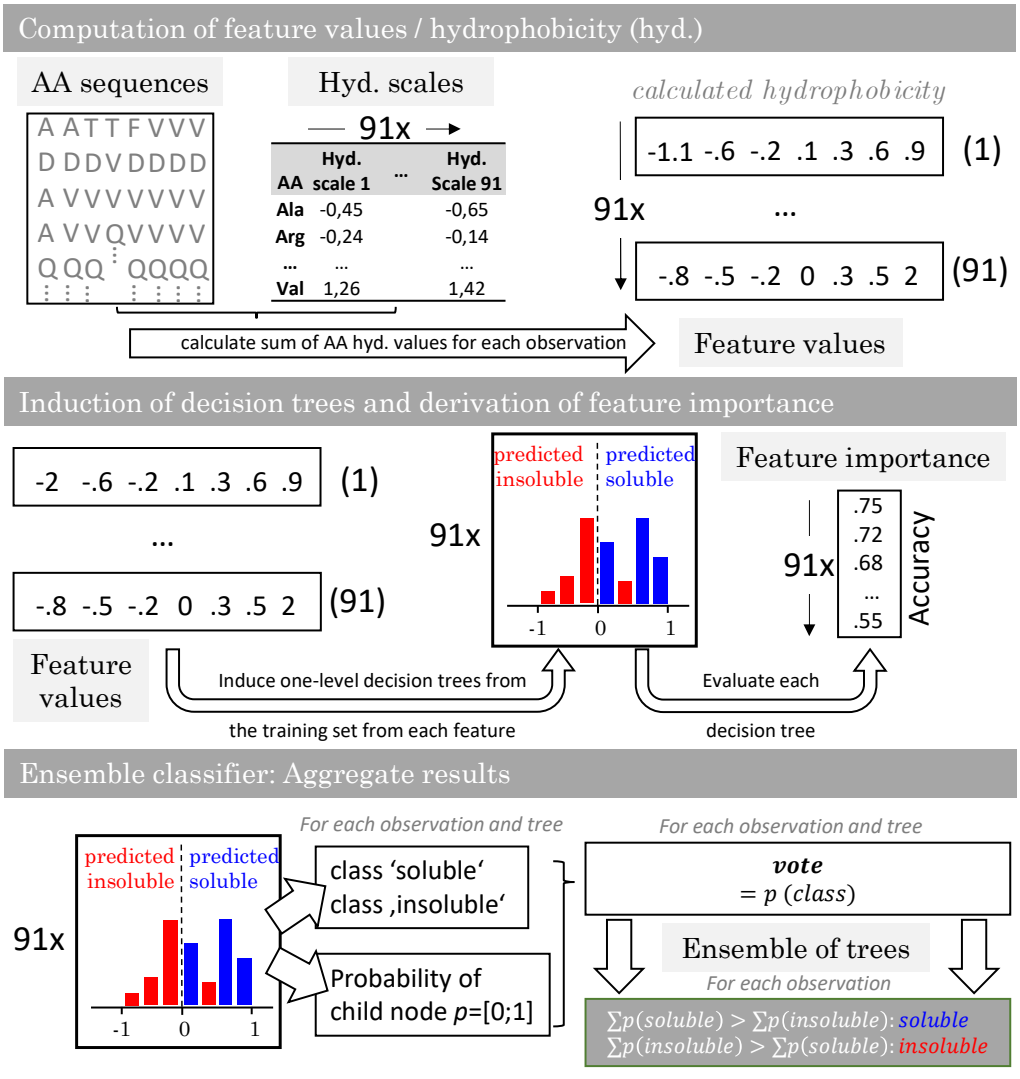


Figure 6.1: Workflow of the ensemble vote classifier. The ensemble vote classifier is constructed by computation of feature values from virus-like particle sequence data and 91 hydrophobicity scales. With the training set features, one-level decision trees are induced. The individual decision trees' accuracy in predicting the training set is defined as the feature importance. In the ensemble model, each decision tree contributes a solubility decision with associated

probability. The results are aggregated and the most probable class is chosen by the ensemble.

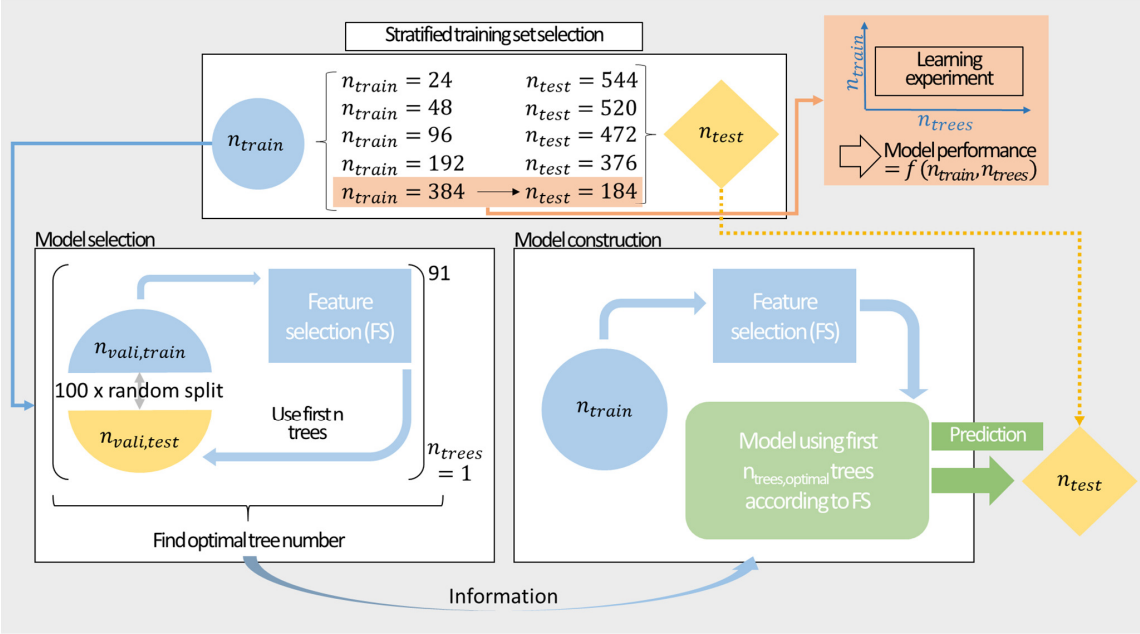


Figure 6.2: Modeling workflow comprising stratified sampling, a learning experiment, model selection, and construction. Stratified sampling results in training sets of $n_{\text{train}} = \{24, 48, 96, 192, 384\}$ data points. These training sets are split in 100-fold Monte Carlo cross-validation to inform about the optimal number of classifiers. The training set is then used to construct a model with preceding feature selection to predict the external test set. The largest training set is additionally utilized for a learning experiment exploring the performance of the model in the space of training set size and number of included decision trees.

6.2.5 Model Performance Evaluation

The performance of the sEVC was evaluated based on Matthew's correlation coefficient

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}, \tag{6.1}$$

where TP, TN, FP, and FN stand for true positive, true negative, false positive, and false negative classification of the model subsets, respectively (train, validation, and test contingency matrix). The MCC is considered to be the least biased singular metric to describe the performance of binary classifiers, especially for cases of class imbalance (Chicco & Jurman, 2020; Powers, 2011). Another metric that was used is the accuracy A as defined in Equation (6.2).

$$A = \frac{TP + TN}{TP + TN + FP + FN} \tag{6.2}$$

6.2.6 Feature Selection

The model generation was preceded by an embedded feature selection. The decision trees were evaluated individually to assess feature importance (Figure 6.1). The feature importance was defined as the accuracy of the individual decision trees for the prediction of the training set. While the MCC is a less biased metric (Powers, 2011), it is not defined for cases where terms in the denominator are zero, which was the case for the smallest training sets. For comparability, accuracy was subsequently used as the feature importance metric throughout this study. Feature importance was computed for every model and for each validation run. The features (and thus the respective decision trees) were then sorted in descending order according to their importance, so that most important features were chosen first during model generation.

6.2.7 Learning Experiment

To explore the model design space with further scrutiny, the model's performance was characterized in a learning experiment. To investigate the effect of training set size, the number of training observations was varied in steps of 5% of the $n_{\text{train}} = 384$ data set from 5% to 95%, resulting in 19 different training set sizes (see also Figure 6.2). The external test set was composed of the remaining 184 observations. The training sets were drawn randomly without stratified sampling from the stratified $n_{\text{train}} = 384$ data set. The remainder of the 384 observations was not used or evaluated. The number of included decision trees (1-91) was screened in addition to the training set size. Thus, a matrix of 19×91 individual model settings was created. Each model setting was repeated ten times resulting in 19×10×91 models. Each training set was sampled individually for all 19×10×91 models, resulting in 17290 training sets. Feature selection was performed 17290 times, i.e. individually for each of the models. Model performance was evaluated based on the external test set and the training set. The median and median absolute deviation (MAD) of the ten model repetitions were computed. They were the basis for the discussion of model performance at respective training set sizes and included number of decision trees.

6.2.8 Systematic Misclassification

To evaluate systematic misclassification, the frequency of true and false predictions were evaluated for each insertion strategy. The relative frequency of strategies found within the classification groups TP, TN, FP, and FN was calculated by summing up their occurrence in the respective groups in the 17290 models of the learning experiment and normalizing it by the overall occurrence of the strategies in all classification groups and all models.

6.2.9 Model Generation

The sEVC workflow comprises stratified training set selection, model validation by MC-CV and prediction of an external test set (Figure 6.2). The number of included decision trees was a hyperparameter that was screened for the model generation on the $n_{\text{train}} = \{24, 24, 24, 48, 96, 192, 384\}$ data sets. The optimal number of included decision trees in the MC-CV validation procedure should inform about the best model for the prediction of the external test set. This relationship was investigated for all seven training sets.

6.3 Results and Discussion

6.3.1 Data Set Construction

The data set consists of observations that can be assigned to 71 unique peptide inserts and 8 unique insertion strategies. Stratified sampling was used to build a representative training set from the full data set. Figure 6.3 shows seven training sets comprising $n_{\text{train}} = \{24, 24, 24, 48, 96, 192, 384\}$ observations sampled by 2-D stratified sampling in a grid of inserts over insertion strategies. Soluble constructs are marked in blue and insoluble constructs are marked in red. The fraction of soluble constructs in the training set f_{sol} is between .46 and .54, resulting in a maximum deviation of .04 from the expected value $f_{\text{sol,total}} = .498$. In the seven models, deviation of f_{sol} from the theoretically expected value of $f_{\text{sol,total}} = .498$ derives from random sampling but is limited due to stratified sampling. From Figure 6.3G, it can be seen that the choice of the insert is strongly influencing solubility, while the insertion strategy only has an effect on solubility for a small

number of constructs. This pattern is confirmed when considering the entire solubility matrix (Solubility Data Table in the Appendix D, Supplementary Material S6.6), underpinning the usefulness of stratified sampling especially for smaller data sets. With 24 training examples, only a third of the 71 inserts are represented by the training set. To investigate the influence of this potential lack of information during model training, three different training sets with 24 samples have been created.

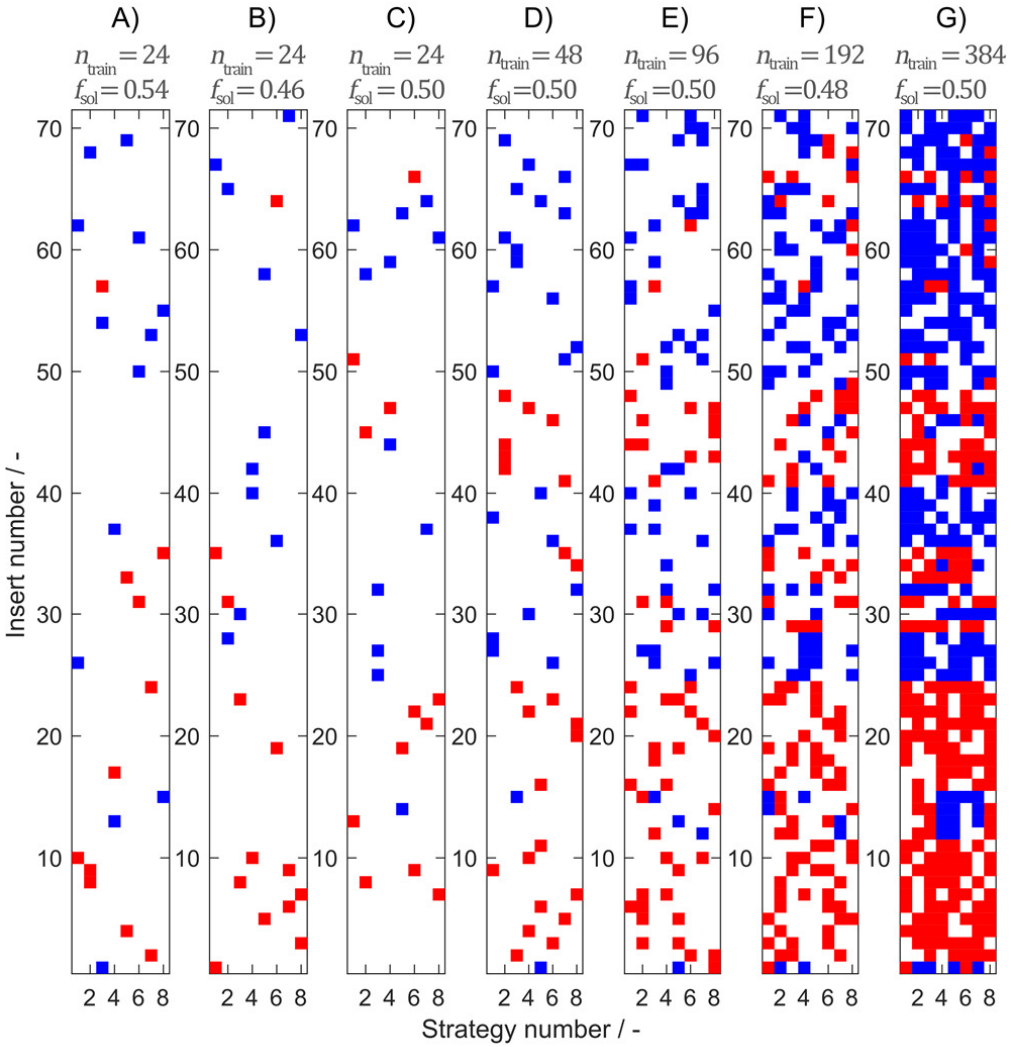


Figure 6.3: Model training sets 1-7 (**(A)-(G)**) created by stratified sampling of $n_{train} = \{24, 24, 48, 96, 192, 384\}$ data points. Stratified sampling was informed by the construction of the entire data set, where eight insertion strategies were used for 71 different inserts, amounting to a total of 568 observations. While for the sampling procedure the solubility data of the observations were unknown, their solubility class is illustrated for interpretation purposes. Blue represents soluble and red represents insoluble observations. The fraction of soluble observations is indicated above the plots by f_{sol}.

6.3.2 Influence of Training Set Size and Number of Decision Trees in the Ensemble Vote Classifier

The sEVC's performance characteristics were evaluated in a learning experiment exploring the space of training set size and number of included classifiers, i.e. decision trees. In this experiment, 190×91 models were created using the best 1 - 91 decision trees as determined individually for each model by the feature selection algorithm. The highest median training MCC can be observed at low training set sizes of ≤ 57 (Figure 6.4A). Note that a brighter color corresponds to better model performance (higher MCC) and lower model variability (lower MAD of MCC). Most models with an MCC $> .80$ are found at the smallest training set size of 19. This concurs with the area where the MAD is greatest with $\geq .06$ (Figure 6.4B). Larger MAD values indicate greater variation between the model repetitions. This suggests a high dependency of model performance on the individual random sampling of the training set. Increasing training set size results in lower training MCC and MAD of MCC. MAD is smaller since more information is available during model training. Additionally, large training sets have a higher probability to contain a significant fraction of identical training observations in the ten different random samplings.

Decision trees with lowest feature importance are included in the models with the largest number of included decision trees due to feature selection. Model performance aggravation due to inclusion of these decision trees was the case for larger training sets, where median training MCC decreases with the number of included decision trees.

The external test set observations are identical for all models, while the training set and therefore the resulting model is individually different. Median test set MCC is $< .48$ for low training set sizes $n_{\text{train}} \leq 38$ (Figure 6.4C). Most models of this size produce a test set median MCC of $\leq .54$, compared to training MCC of $\geq .8$ for most models, suggesting an overfitted model for small training data sets. The largest MCC of the external test set predictions are found at training set sizes ≥ 249 and at > 23 and < 65 number of included decision trees, which is also overlapping largely with the region of lowest MCC MAD with many models showing a test set MAD $\leq .01$ (Figure 6.4D). Therefore, in this area, the best models are found having high MCC (most $\geq .6$) on the test set and low MAD of MCC. The difference between training and

external test set MCC in this area is ≤.1. Thus, the training set is a good indicator of model performance in said area.

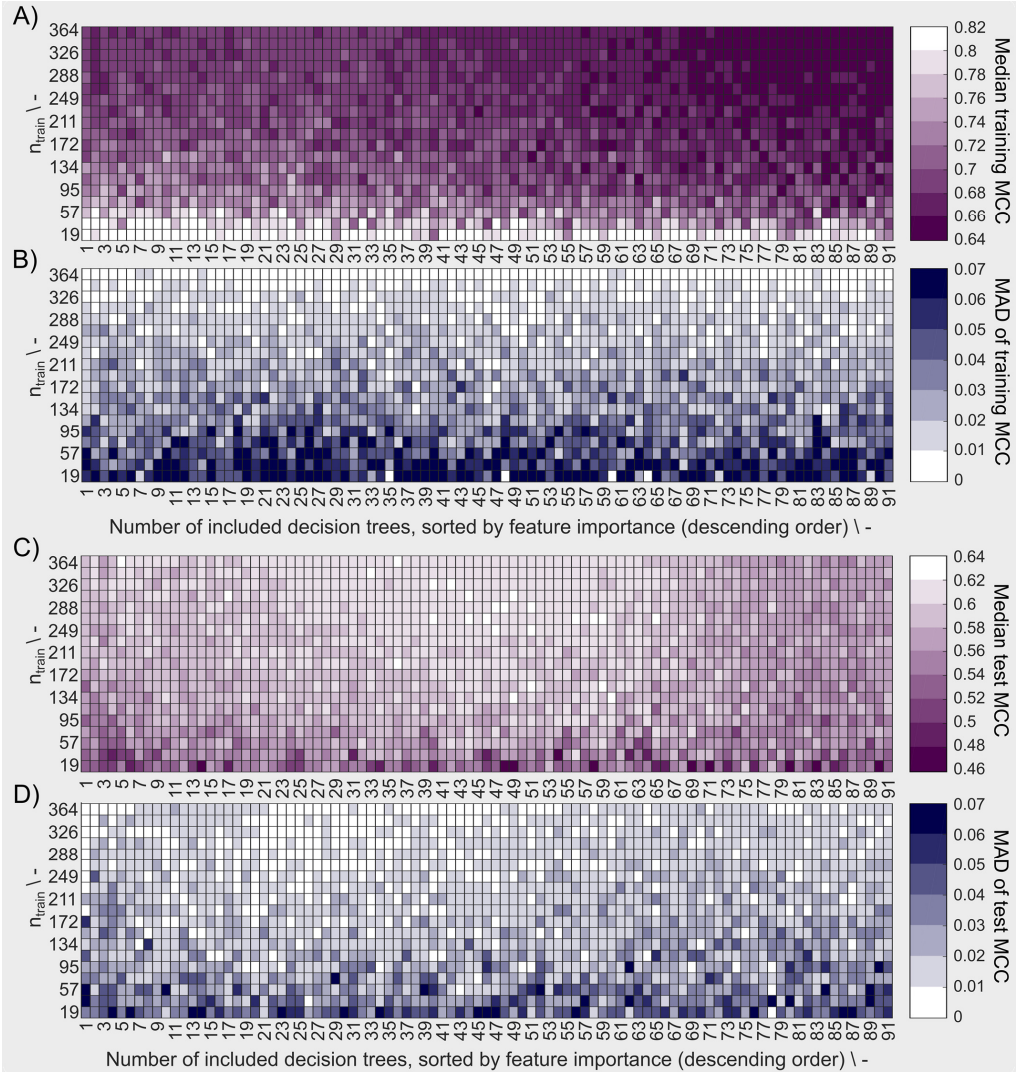


Figure 6.4: Model performance based on training and external test data described by **(A)** the median Matthew's correlation coefficient (MCC) and **(B)** the median absolute deviation (MAD) of the MCC for training data and **(C)** median MCC and **(D)** MAD of MCC for external test data. Each rectangle represents a decaplicate of a model with the number of training examples shown on the y-axis and the number of included decision trees in the ensemble classifier on the x-axis. The training observations were randomly sampled from the stratified dataset with $n_{\text{train}} = 384$. The decision trees are sorted in descending order by feature importance. This means, that at point n on the x-axis, the results of the models including the n best decision trees are shown. White/bright color denotes high median MCC values and low MAD of the MCC, dark (violet or blue) color denotes low median MCC values and high MAD of the MCC, relative to all MCC data in the learning experiment. A well-predicting and reproducible model has high MCC and low MAD, respectively (both bright).

It has to be noted that the model performance was evaluated on randomly chosen subsets of the stratified $n_{\text{train}} = 384$ data set, constraining the benefits from stratified sampling. Stratified sampling can be expected to decrease overfitting and variation seen for low training set sizes, as the probability of drawing a non-representative sample set is drastically reduced, as discussed both above and below.

6.3.3 Selection of Models Based on Stratified Training Sets

In the seven models created by stratified samplings, the shaded area, representing the model's MAD of validation MCC, is decreasing with increasing training set size (Figure 6.5). This was expected due to the increasing amount of information available during model training. The sets with $n_{\text{train}} = 24$ were constructed three times to investigate the robustness of small training set sizes. Of these, the first has a smaller MAD area than the other two, which can possibly be attributed to a 'lucky' stratified sampling. With only 24 samples, the MC-CV comprises 12 validation training and 12 validation test observations, resulting in potentially larger artifacts of random selection. Also, stratified sampling based on the insert does not have a strong effect, since only a maximum of 24 different inserts of the total 71 inserts are chosen, leaving 47 inserts unrepresented. Validation MCCs are comparably stable over the number of included decision trees. Some of the models show a slight MCC increase over the first number of included decision trees (models 1-3 and 5) and some show a gradual but shallow decline at larger numbers of included decision trees (> 50; models 4, 6, and 7). This underlines the effect seen in the learning graph while being less pronounced, which probably can be attributed to the more balanced training data set. Most of the models result in validation MCCs of around .6, whereas model 3 of training set size $n_{\text{train}} = 24$ has a significantly lower validation MCC of around .2. When considering the MCC of the external test sets, each of the models shows adequate performance when a minimum number of decision trees was included. At around 30 decision trees, the models have an external test set MCC that is either above or close to their overall median MCC. The test set MCCs at 30 decision trees are $> .6$ for all models, except model 2 with an MCC of .56. The mean MCC of all seven models' external test data over decision tree numbers was computed to inform about the average model performance dependent on the number of

included decision trees (data not shown). It was optimal at 29 and 30 decision trees, both with a mean MCC of .61.

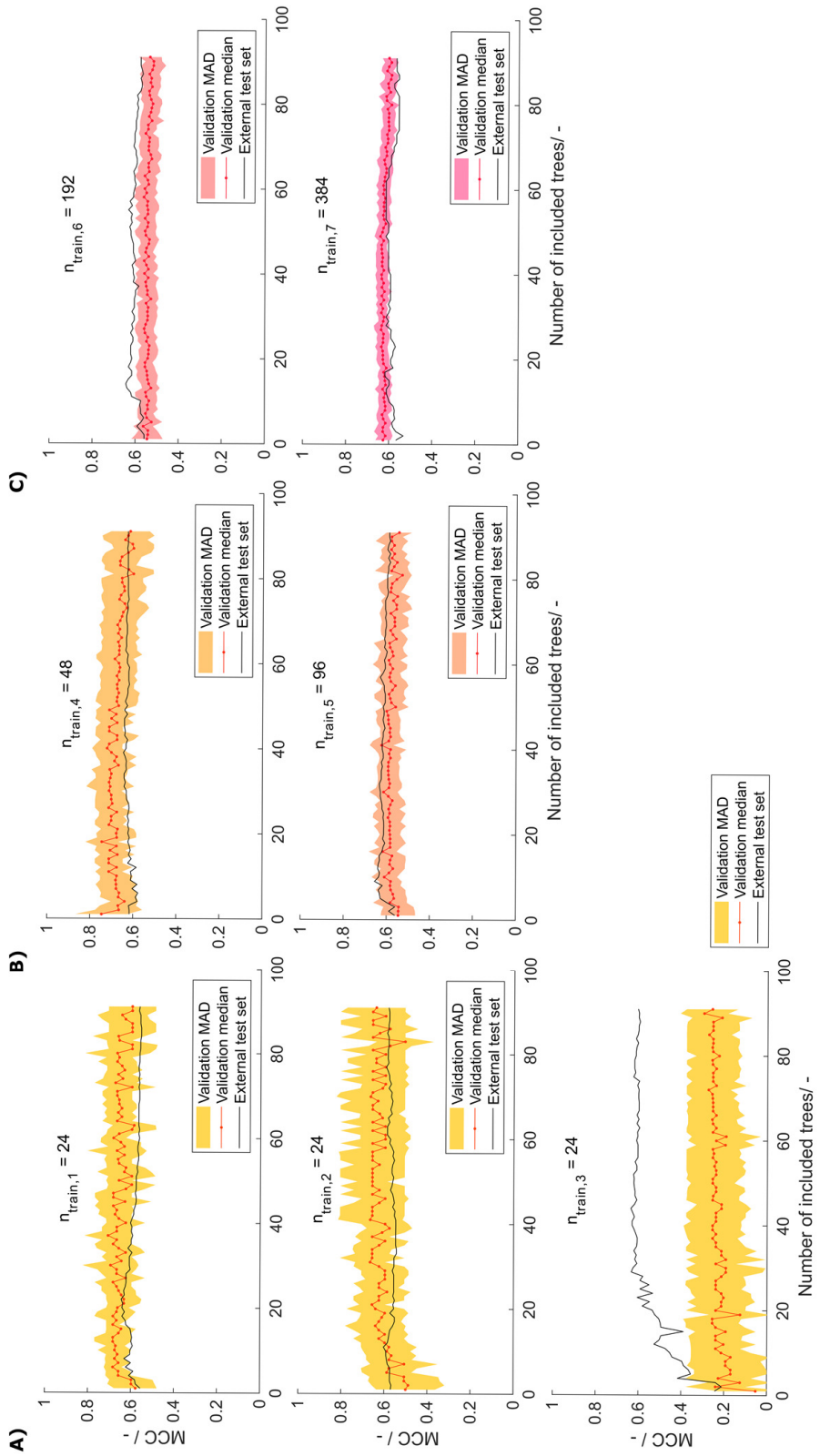


Figure 6.5: Validation and external test set Matthew's correlation coefficient (MCC) on stratified training data with training set size

of **(A)** 24, **(B)** 48 and 96, and **(C)** 192 and 384 observations, depending on the number of included decision trees. Validation median and median absolute deviation (MAD) are calculated from 100-fold Monte Carlo cross-validation. The median is shown as red dots connected by a line to guide the eye. The shaded area represents the MAD around the median and are colored from yellow over orange to pink with increasing training set size. The black line represents the performance of the model on the external test set of size $n_{\text{test}} = 568 - n_{\text{train}}$.

At larger training set sizes, the trend in validation data is more translatable to the trend in external test data, while at low training set sizes of 24 observations, the model should only partially rely on validation data and may include a minimum number of decision trees to avoid overfitting. Another reason for this is feature selection, which is performed on a potentially unrepresentative data set. It can therefore result in prioritizing decision trees that fit unrepresentative data but not the entire data set well. Comparison of the model's performance to other published models can be difficult since many report their results as accuracies - typically in the range of .62 - .83 (Hebditch, Carballo-Amador, Charonis, Curtis, & Warwicker, 2017; Idicula-Thomas, Kulkarni, Kulkarni, Jayaraman, & Balaji, 2006; Magnan et al., 2009; Smialowski et al., 2006). In the ideal balanced case, the MCC of these models would be .24 - .66 (for the explanation on the relation of MCC and accuracy see Appendix D, Supplementary Material S.6.2). However, many of those models are not based on a balanced data set, which would then lead to a lower MCC. The model presented in this paper shows MCC values close to the best MCC estimates of previously published models. In a review on HBcAg cVLPs, insert charge was described to be the most important parameter for solubility of cVLP candidates (Whitacre et al., 2009). Construction of a decision tree on a scale that rates aspartic acid and glutamic acid with -1, arginine and lysine with +1, and all other amino acids with 0, resulted in an MCC of 0.38 on the external test set using the $n_{\text{train}} = 384$ training and corresponding test set (data not shown). The correlation of insert charge to cVLP solubility, as described in the above-mentioned review, was therefore observed with the data set investigated in this article. It was not as strong as the predictions of the sEVC based on hydrophobicity scales.

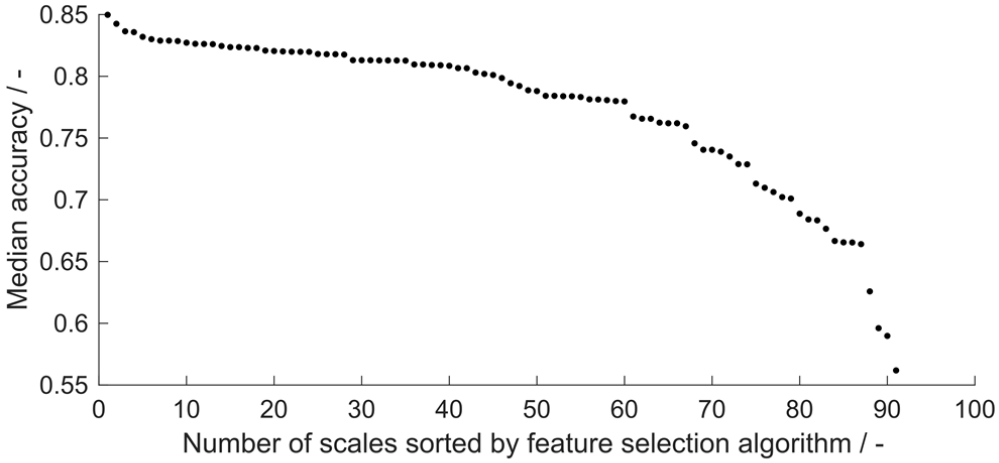
Some trends with regard to the number of included decision trees have been uncovered and discussed. However, it has to be noted that the effects are quite small over a wide range of included decision trees,

especially with stratified training sets. On the one hand, this indicates that the model performs well over a large space of a chosen number of classifiers and training set sizes. On the other hand, it highlights the potential of the ensemble classifier to include more orthogonal scales that could describe more aspects in the data and therefore result in even better models. In a very simple way, the orthogonality of the scales can be analyzed by principal component analysis (PCA). PCA on the 91 normalized hydrophobicity scales revealed that the first principal component already explains 68.8% of the variance (Appendix D, Supplementary Material Figure S6.1). This may be expected, when considering how the scales were derived. Many of the hydrophobicity scales originate in some way from other scales being only slightly modified. It would therefore be highly interesting to investigate the sEVC framework constructed with a set of scales that complement each other to explain more of the variance found in the data and result in even better models.

6.3.4 The Potential of Feature Selection to Retrieve Biological Information

Feature selection is an important tool to boost model performance. It can also serve to retrieve biological information with respect to the modeled problem. Accuracy was chosen as metric for feature importance to avoid cases where the MCC is not defined. Decision trees that individually classify more observations of the training set correctly therefore have higher feature importance. In the learning experiment, 19 different training set sizes ($n_{\text{train}} = 19, 38, \ldots, 364$) were evaluated 910 times, giving a statistically strong insight into feature importance in the range of tested training set sizes.

Median feature importance ranges from .54 to .85 (Figure 6.6) and shows an MAD of .02 to .04, while MAD increases towards lower accuracies (data of accuracies and MAD in Appendix D, Supplementary Material Table S6.2). This median feature importance value is valid for the entire training data set of 384 observations. It describes, in the framework of the presented chimeric HBcAg VLP solubility problem, which decision tree, and therefore hydrophobicity scale, is most suitable for the distinction of soluble and insoluble constructs independent of the training set size.



1	SET3	20	NNEIG	39	NADH010106	58	PONP800106	77	WILM950101
2	VHEG790101	21	PONP800101	40	PONG3	59	MIYS850101	78	Wimley
3	EISEN	22	JACWH	41	CowanWhittacker	60	SweetEisenberg	79	Bishop
4	ESID840101	23	PONNU	42	CIDAB	61	CIDH920105	80	BULDG
5	MANP780101	24	MDK1	43	NADH010104	62	Welling	81	CHOTA
6	ROSM880101	25	ROSM880102	44	CIDA	63	PONP800104	82	Wimleyreverse
7	CHOC760103	26	NADH010102	45	SET2	64	PARJ860101	83	ROSEAreverse
8	CHOTH	27	NADH010101	46	NOZY710101Tanford	65	FASG890101	84	ARGP820101
9	PONG1	28	KYTJ820101	47	NADH010105	66	GIBRA	85	JONES
10	PONP930101	29	KUHL950101reverse	48	BLAS910101	67	PONG2	86	JOND750101
11	WOLR810101reverse	30	ROSG850102	49	COWR900101	68	CIDH920101	87	COHEN
12	RaoArgos	31	ROSEF	50	GUYH850101	69	Wilson	88	ROSEB
13	JANJ790102	32	PRAM900101reverse	51	LEVM760101reverse	70	KIDA850101reverse	89	NADH010107
14	JANIN	33	ENGD860101reverse	52	LEVIT	71	KIDER	90	ZIMJ680101
15	JURD980101	34	ENGEL	53	FAUJ830101	72	SWER830101	91	WILM950103reverse
16	MDK0	35	PONP800102	54	FAUCH	73	SWEET		
17	ROSEM	36	NADH010103	55	ROSM880103reverse	74	WILM950102		
18	WOLR790101	37	CASSI	56	GUYFE	75	CIDBB		
19	SET1	38	PONP800103	57	WILM950104	76	PONP800105		

Figure 6.6: Median feature importance measured by the median training set accuracy of the individual scales in all 190×91 models in the learning experiment. Median accuracy of scales was sorted in descending order, so that scale 1 has highest accuracy and therefore highest feature importance, while scale 91 has lowest feature importance. The IDs of the scales (i.e. feature names) are noted below in respective order.

Feature importance can be used to obtain a biological interpretation of the model based on the characteristics of the hydrophobicity scales with highest feature importance (best) and lowest feature importance (worst). Feature importance describes their quality to predict within the cVLP solubility problem. With the first three scales, feature importance declines more than with the following 57 scales. The last four scales decrease markedly in feature importance. The best scale SET3 originates from a study on the prediction of transmembrane helical regions (Zviling et al., 2005). In this study, a genetic algorithm optimization approach amplified hydrophobicity and hydrophilicity of hydrophobic and hydrophilic amino acids, respectively, compared to the input scales. Insoluble expression results from protein-protein interaction leading to the formation of aggregates, potentially leading to inclusion bodies (Carrio & Villaverde, 2005). Transmembrane regions in proteins are naturally hydrophobic (Silverman, 2003) and in the absence of

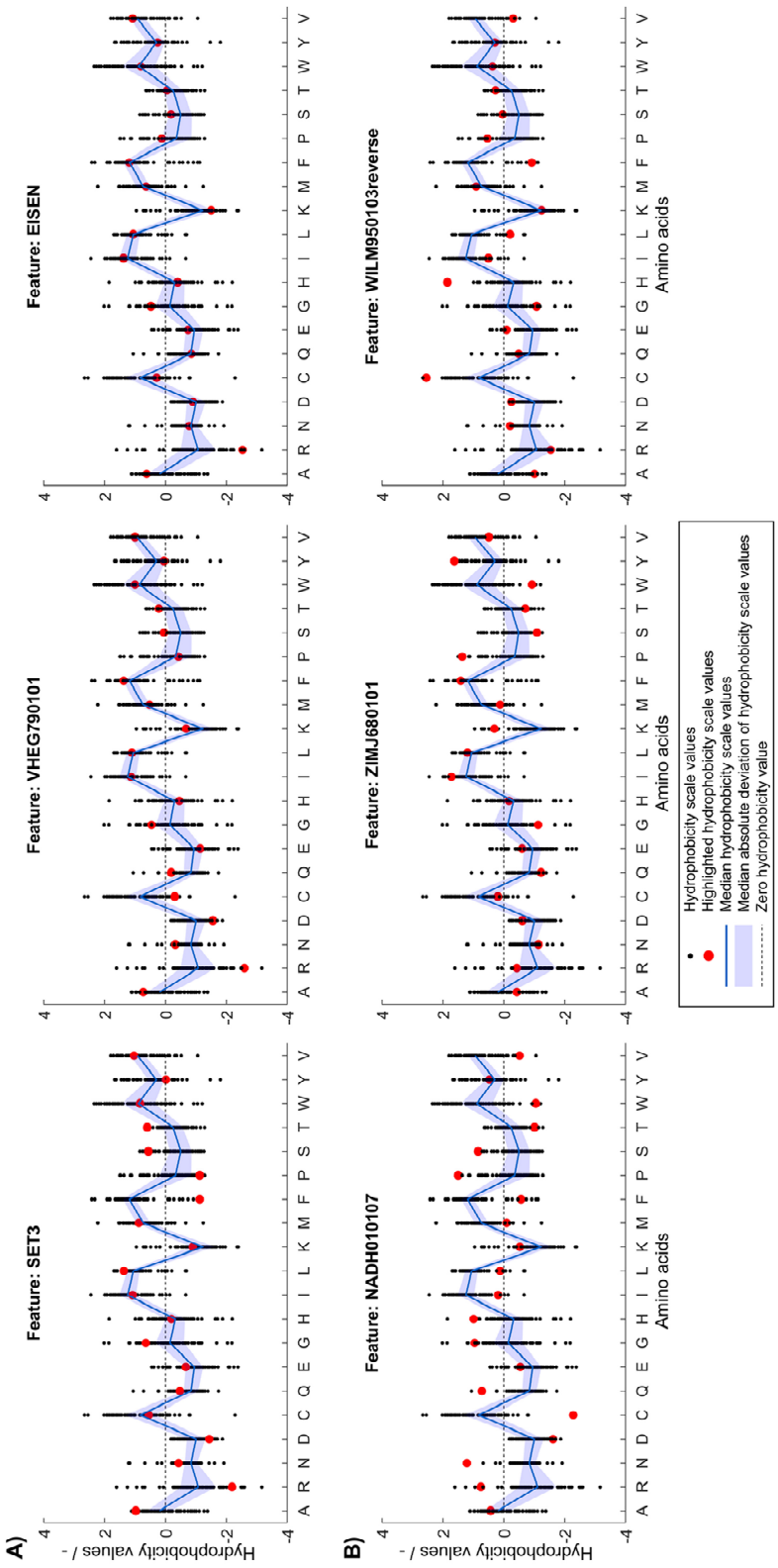
membranes therefore prone to protein-protein aggregation. The good performance of this scale suggests that findings on hydropathy based on the propensity to form transmembrane helices is comparable to the solubility investigated in this study. The scale VHEG790101 has second-highest feature importance and was derived using surface accessibility data from Chothia's study of 1976 (Chothia, 1976; von Heijne & Blomberg, 1979). Free transfer energies of residues from polar to non-polar solvent were calculated adding protonation energies for charged residues. This highlights the benefit of amplifying hydrophilicity of charged residues for application in the VLP solubility scenario. In a machine learning study on protein aggregation, VHEG790101 was also rated as an important feature to predict aggregation propensity (Y. Fang et al., 2013). The EISEN scale has similar but slightly lower feature importance, which can be explained by the fact that its hydrophobicity values are simply the average of five scales' normalized hydrophobicity values among which is the scale of Chothia and Von Heijne (Eisenberg, Weiss, Terwilliger, & Wilcox, 1982).

WILM950103reverse, the scale that has lowest feature importance, is based on C4 reversed phase chromatography retention times of peptides (Wilce et al., 1995). Retention on chromatography columns is often based on small fractions of the molecular surface and cannot directly be translated to properties related to the entire molecule (Hebditch et al., 2019). ZIMJ680101 was created by statistical analysis based on only 40 proteins. The space of applicability of this scale is probably limited and it therefore performs badly in the VLP solubility scenario. In the above-mentioned study by Fang and colleagues, ZIMJ680101 also was rated as an important feature (Y. Fang et al., 2013), highlighting that a direct comparison between the presented model and Fang's model cannot easily be drawn. The third-worst scale is one of seven hydrophobicity scales derived from a study on solvent-accessibility of amino acids (Naderi-Manesh et al., 2001). Scales NADH010101-7 are created by information theory and represent the self-information derived from different thresholds of solvent-accessibility. These aim to describe the amino acids' surface accessibility within a protein. From scales 1 to 7, the threshold for surface accessibility was increased from 5% to 50% of its maximum accessibility. Scales 1 and 2 with 5% and 9% threshold ranked at 27 and 26 in feature importance analysis. Increased thresholds of 16%, 20%, 25%, and 36% resulted in feature importance ranks of 36, 43, 47, and 39. Scale

7 with 50% accessibility threshold is significantly worse with position 89 of 91 scales in feature selection. This comparison suggests that there is a dependence of the threshold set in the study by Naderi-Manesh and colleagues on the performance of these scales in the solubility model. The scales with low threshold (5% and 9%) only count residues as inaccessible if they are almost completely buried, thus boosting the hydrophobicity of very hydrophobic amino acids relative to the other amino acids. On the contrary, with the cut-off of 50% accessibility, amino acids that have a significant share of solvent-accessible surface but still below 50% will be regarded as hydrophobic. The lower threshold is probably more applicable to the cVLP solubility problem, since the aim of the epitope design is to expose rather than bury it. This in turn means that the insertion of typically strongly buried hydrophobic residues can corrupt protein folding by their orientation to a protein core, potentially leading to misfolded proteins that aggregate to insoluble clusters. The feature importance ranking of these scales might indicate that, for solubility, strongly hydrophobic residues have a significant influence on solubility, while residues that are somewhat hydrophobic, but still have some solvent-accessible area are not as critical. This argument also supports that chromatography-based scales may not be the best choice to describe macro-properties such as solubility, as discussed above.

To retrieve information related to the hydrophobicity of individual amino acids, it is valuable to analyze hydrophobicity values of the normalized best and worst scales compared to the median of all hydrophobicity scales. If a particular scale performs better in feature selection than average, this can probably be attributed to the fact that the hydrophobicity values of certain amino acids are different from the median value. Figure 6.7 shows normalized hydrophobicity scale values for all amino acids for A) the three best and B) the three worst scales in the scope of this study. In the following, the hydrophobicity values of the highlighted scales are discussed in reference to the individual amino acid median hydrophobicity value and its MAD (of all 91 hydrophobicity scales). The three best scales are SET3, VHEG790101, and EISEN, in descending order. They have in common that arginine (single-letter code R, as indicated in Figure 6.7) has a significantly lower hydrophobicity value than in most other scales. This value falls well below the MAD range of the hydrophobicity value of all scales for arginine. Alanine (A) and glycine (G) are attributed a slightly larger hydrophobicity value than

within the MAD of their distribution. SET3 and VHEG790101 also rate hydrophobicity of aspartic acid (D) lower and asparagine (N) higher compared to the MAD range. Compared to the median and the other two scales, SET3 has a markedly low hydrophobicity value for phenylalanine (F).



|←

Figure 6.7: Normalized hydrophobicity scale values for the 20 proteinogenic amino acids. Amino acids are shown on the x-axis indicated by single-letter code. In each graph, 91 scales' amino acid hydrophobicity values are represented by black dots, their median is shown as a blue line with a shaded blue area representing the median absolute deviation. A dashed, horizontal line through zero is shown to guide the eye. The scale indicated by a subfigure title is highlighted in red for **(A)** the three scales with highest median feature importance and **(B)** the three scales with lowest median feature importance as determined in the learning experiment. The normalized scales' sign was changed so that hydrophobicity of aspartic acid is always negative.

The three worst scales are NADH010107, ZIMJ680101, and WILM950103reverse, in descending order. For the worst scales, it is more difficult to identify patterns in the deviation from the median hydrophobicity values. However, a general trend in the worst three scales is that hydrophobicity values of hydrophobic amino acids are particularly low while charged amino acids are about average or above. Overall, the consensus of the best scales is that arginine should be attributed a lower hydrophobicity value than the population of scales would suggest, while alanine should be more hydrophobic than in most of the scales. Other charged amino acids are partly rated slightly less hydrophobic, such as aspartic acid and lysine. The worst scales' accuracies are probably lower since a number of hydrophobic amino acids' hydrophobicity values are comparably low, while, compared to the population of hydrophobicity scales, a number of charged or polar amino acids' hydrophobicity values are relatively high. Amino acids such as cysteine or phenylalanine show conflicting trends in the worst and best scales and thus no conclusion thereof can be drawn.

The above analysis suggests that a unique property of arginine might contribute to its special scale position having both absolute lowest and relatively low hydrophobicity in the three best scales. Arginine's role as an agent to reduce protein-protein interactions and increase solubility is only partially understood (Arakawa et al., 2007). As free amino acid in high concentrations (1 M), it has been shown to interact favorably with almost all amino acid side chains and peptide bonds. This means, that arginine can reduce both hydrophobic and electrostatic interactions (Arakawa et al., 2007). As an additive, arginine favorably interacts with tryptophan and therefore can suppress hydrophobic interactions leading

to aggregation (Tsumoto et al., 2004). Exactly this effect is thought to bear the potential of introducing protein-protein interactions when arginine is present in abundance in the amino acid sequence of a protein (Warwicker, Charonis, & Curtis, 2014). This conclusion was drawn by investigating the ratio of lysine to arginine (K/R) to highlight the specific effects of arginine on protein solubility as compared to lysine, since both bear one positive charge. It has also been shown that decreasing arginine content could increase solubility of a single-chain variable fragment (Austerberry et al., 2019). A negative arginine-related solubility effect was also seen in a study on a large data set of *E. coli* expressed proteins (Price et al., 2011). In this study on cVLPs, higher arginine content leads to decreased hydrophobicity values, which in turn leads to higher probability for soluble classification. This effect was observed although the K/R ratio ($mean(K/R) = .32$) was strongly unfavorable considering the results of Warwicker and colleagues (Warwicker et al., 2014). Another study showed that mutations from surface lysines to arginine in GFP could enhance its chemical stability (Sokalingam, Raghunathan, Soundrarajan, & Lee, 2012). However, protein folding was found to be aggravated. Protein solubility is a very complex topic and depends on a variety of factors of different dimensions, which is illustrated by the cVLP solubility problem. HBcAg dimers have low solubility in physiological pH and ionic strength, since hydrophobic and other interactions strongly favor VLPs (Ceres & Zlotnick, 2002). The assembly is an entropy-driven mechanism and is therefore similar to protein aggregation (Ceres & Zlotnick, 2002; Gorbenko & Trusova, 2011). Association relies on weak protein-protein interactions of HBcAg, such as hydrophobic interaction in a tyrosine pocket (C. R. Bourne et al., 2009) in the base of the molecule. This ordered aggregation is probably mandatory for the soluble state of HBcAg at the high expression levels in *E. coli*'s cytosol, since HBcAg dimers were found to be aggressively aggregating and forming capsids already at low concentrations (Ceres & Zlotnick, 2002). Protein insolubility during expression can therefore exhibit an entirely different origin than for other proteins – the association of the HBcAg proteins to structures that are not VLPs but unordered aggregates, which themselves are not soluble. Truncated wild-type HBcAg ($Cp_{1\text{-}149}$, based on UniprotID: P03147 (The UniProt Consortium, 2018)) contains eight arginine residues with a K/R of .25, even though the arginine-rich C-terminus of the full-length HBcAg is not

considered. To investigate this relationship with cVLPs, a one-level decision tree solely based on K/R ratio of all 568 observations was constructed, resulting in an inverse relationship as observed by Warwicker and colleagues (Warwicker et al., 2014). With the VLP solubility data, K/R values below the cut point are predicted as soluble with an Accuracy of .65 and MCC of .3, while in Warwicker's study lower K/R lead to higher chances of insolubility. Arginine-based interactions could therefore be hypothesized to be of great importance in the recruitment of other HBcAg molecules to form VLPs eventually. Therefore, arginine's property to increase protein-protein interactions when present in the amino acid sequences can be assumed to enhance VLP assembly, which is mandatory for significant levels of soluble HBcAg. Following this reasoning, substitution of arginine with lysine would maintain overall protein charge but probably promote the existence of either soluble HBcAg dimers incapable of assembly or insoluble HBcAg aggregates.

The role of arginine can also be discussed with respect to other amino acids. Tryptophan is not present in abundance in truncated wild-type HBcAg. Four tryptophan residues build the core of the HBcAg helices and are paired with either tyrosine, phenylalanine, or arginine residues (see also Appendix D, Supplementary Material Figure S6.2). Since arginine-tryptophan interactions were found to be extraordinarily strong (Arakawa et al., 2007), additional tryptophans in the epitope may result in misfolding during protein expression, since abundant arginines and other residues interact favorably with the tryptophans in the epitope region. This would give reason to the low feature importance observed in the three worst scales, which underrate tryptophan hydrophobicity. In the data set, 0, 1, or 2 tryptophans are introduced compared to the wild-type $Cp_{1\text{-}149}$. Interestingly, observations containing two additional tryptophans are all insoluble, while observations containing zero or one additional tryptophan are found both in the soluble and in the insoluble group. As discussed above, valine probably also plays a vital role that is shown by its low hydrophobicity values in the worst scales. Arginine-valine interaction was found to be the only unfavorable interaction in single amino acid experiments (Arakawa et al., 2007). With the above reasoning, this could hamper assembly of HBcAg VLPs and therefore decrease VLP solubility.

6.3.5 Systematic Classification Errors Based on Insertion Strategies

The average model performance with respect to the eight insertion strategies is related to properties of the utilized molecules that are strongly associated with the presented problem. If a strategy has a significantly higher relative frequency of *FP* classifications than *FN* classifications, the sEVC model systematically overestimates the solubility of observations created with this strategy. This indicates a particularly bad performance of this strategy. From Figure 6.8, it can be seen that this is the case for strategy H, both in training (A) and external test sets (B). During model construction, it would therefore be interesting to tweak strategy H's solubility prediction so that the numbers of strategy H's $FP = FN$. This can of course only be done for constructs where there is already a significant influence visible in the training set and when the training set is large enough. If a strategy is more numerous in the *FN* than in the *FP* group, the opposite case is true, where the model underestimates its solubility. These strategies are systematically good for solubility with respect to the model. This can, for example, be observed for strategy E. Its solubility prediction could be tweaked to higher solubility during model training. From the insertion sites and deletions that are different in the eight strategies we could, however, not find a relation that would explain the above-mentioned behavior. This relationship could be related to 3-D properties that cannot be explained by the 2-D amino acid sequence information only. The same approach has been tested on the 71 inserts, where no significant effects could be observed on the comparably large amount of different inserts and therefore their rarer occurrence (data not shown). The presented model performance analysis, which is based on the particular training set structure, can be used as a potent tool to learn about the characteristics of the data set and to boost model performance.

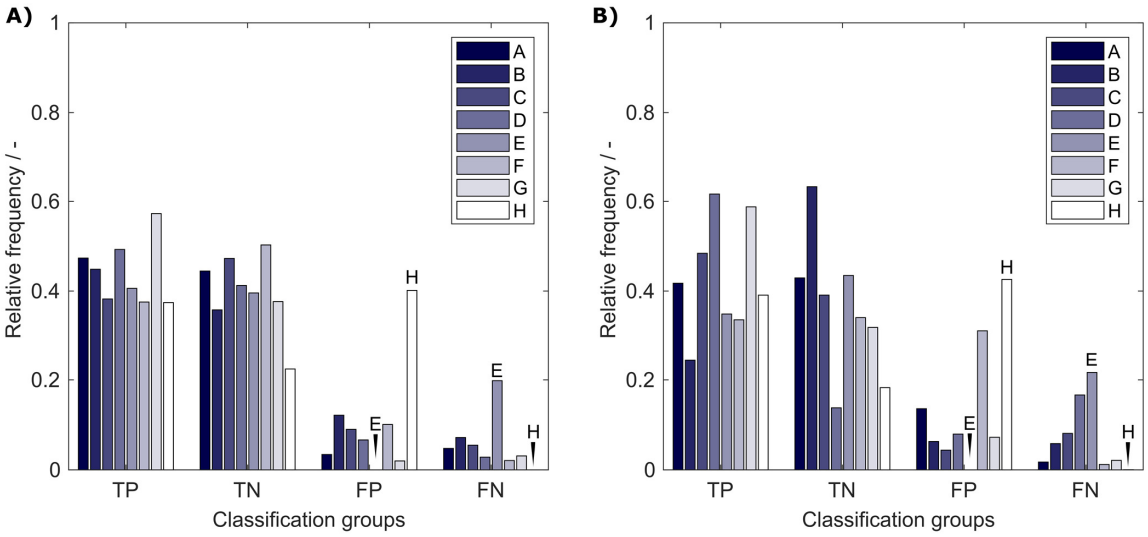


Figure 6.8: Relative frequency of classification groups based on insertion strategies A-H in **(A)** the training set and **(B)** the external test set of the 17920 models in the learning experiment. Strategy E and H are marked additionally to guide the eye. TP: true positive; TN: true negative; FP: false positive; FN: false negative.

6.4 Conclusion and Outlook

In this article, we presented a novel solubility prediction framework based on experimental and theoretical hydrophobicity scales that was applied to the prediction of chimeric HBcAg VLP solubility. In summary, little information was fed into our model, i.e. publicly available sequences, hydrophobicity scales, and solubility data.

The best models predicted with an MCC of $> .6$ on the external test set. Stratified training set sampling based on information on the inserted peptide sequence and the insertion strategy proved beneficial especially for small training set sizes. Evaluation of the contingency matrix revealed that certain epitope insertion strategies were overrepresented in the *FP* or the *FN* group of both training and test set and were therefore particularly limiting or promoting, respectively, for cVLP solubility. Detailed assessment of the best and worst features, i.e. hydrophobicity scales, suggested a special role for arginine for soluble cVLP expression. Contrary to reports on the solubility of other proteins, a large arginine content did not disrupt but rather improved cVLP solubility. We hypothesized that arginine's positive interaction with almost all amino acids plays a crucial role in recruiting HBcAg dimers or larger building

blocks to form a capsid, which in turn is required for meaningful levels of HBcAg concentrations in physiological buffers.

The presented framework proved to be applicable to small and larger training set sizes and could, with minor adaptations, be transferred to the prediction of monoclonal antibody solubility or even other biophysical properties. In the future, an informed design of scales that are orthogonal could greatly benefit the presented approach, as it would diversify the classifiers' performances and therefore benefit the ensemble classifier. Additionally, it would be interesting to evaluate the model as a regression tool, avoiding the discretization that is performed during the sEVC procedure. Our results also suggest that building a global solubility model for all proteins is highly challenging and may only be feasible if a balanced data set of equally represented protein classes at very high observation numbers is available.

Acknowledgements

The authors would like to thank Sebastian Andris, Marieke Klijn, Heidemarie Knieriem, and Adrian Sanden for proofreading and inspiring discussions.

Appendix D: Supplementary Material

The Supplementary Material associated with this article contains the following information:

- S6.1 Normalized Hydrophobicity Scales
- S6.2 Comparison of MCC and Accuracy
- S6.3 Feature Importance
- S6.4 Principal Component Analysis
- S6.5 Location and Interaction of Tryptophan within the HBcAg Dimer
- S6.6 Solubility Data Table

7

Optimization of a Soft Ensemble Vote Classifier for the Prediction of Chimeric Virus-Like Particle Solubility and Other Biophysical Properties

Philipp Vormittag[a], Thorsten Klamp[b], Jürgen Hubbuch[a] *

[a] Institute of Process Engineering in Life Sciences, Section IV: Biomolecular Separation Engineering, Karlsruhe Institute of Technology, Fritz-Haber-Weg 2, 76131 Karlsruhe, Germany

[b] BioNTech SE, An der Goldgrube 12, 55131 Mainz, Germany

* Corresponding author

Chimeric virus-like particles (cVLPs) are protein-based nanostructures applied as investigational vaccines against infectious diseases, cancer, and immunological disorders. Low solubility of cVLP vaccine candidates is a challenge that can prevent development of these very substances. Solubility of cVLPs is typically assessed empirically, leading to high time and material requirements. Prediction of cVLP solubility *in silico* can aid in reducing this effort. Protein aggregation by hydrophobic interaction is an important factor driving protein insolubility. In this article, a recently developed soft ensemble vote classifier (sEVC) for the prediction of cVLP solubility was used based on 91 literature amino acid

hydrophobicity scales. Optimization algorithms were developed to boost model performance, and the model was redesigned as a regression tool for ammonium sulfate concentration required for cVLP precipitation. The present dataset consists of 568 cVLPs, created by insertion of 71 different peptide sequences using eight different insertion strategies.

Two optimization algorithms were developed that (I) modified the sEVC with regard to systematic misclassification based on the different insertion strategies, and (II) modified the amino acid hydrophobicity scale tables to improve classification. The second algorithm was additionally used to synthesize scales from random vectors. Compared to the unmodified model, Matthew's Correlation Coefficient (MCC) and accuracy of the test set predictions could be elevated from .63 and .81 to .77 and .88, respectively, for the best models. This improved performance compared to literature scales was suggested to be due to a decreased correlation between synthesized scales. In these, tryptophan was identified as the most hydrophobic amino acid, i.e. the amino acid most problematic for cVLP solubility, supported by previous literature findings. As a case study, the sEVC was redesigned as a regression tool and applied to determine ammonium sulfate concentrations for the precipitation of cVLPs. This was evaluated with a small dataset of ten cVLPs resulting in an R^2 of .69.

In summary, we propose optimization algorithms that improve sEVC model performance for the prediction of cVLP solubility, allow for the synthesis of amino acid scale tables, and further evaluate the sEVC as regression tool to predict cVLP-precipitating ammonium sulfate concentrations.

7.1 Introduction

Protein solubility is a generally recognized problem in biopharmaceutical drug development. The fact that poor solubility can hamper a molecule's development is a well-known challenge in chimeric virus-like particle (VLP) process development. VLPs are multimeric structures based on viral proteins, which are employed as vaccines or delivery vehicles for proteins or nucleic acids (Bryan et al. 2016; Kaczmarczyk et al. 2011; McAleer et al. 1984; Muratori, Bona, and Federico 2010; Strods et al. 2015). For example, VLPs are applied as vaccines against hepatitis B virus or human papillomavirus (McAleer et al. 1984; Bryan et al. 2016). Chimeric VLPs (cVLPs) are VLPs decorated with foreign epitopes altering the function of the unmodified VLPs by, for example, directing the patient's immune response towards the inserted epitope (Klamp et al. 2011; Yoshikawa et al. 1993). While this flexibility of antigenic display is one of the major advantages of VLPs (Pumpens et al. 2008), recombinant insertion of epitopes often results in expression of insoluble structures (Billaud et al. 2005; Karpenko et al. 2000). Factors affecting cVLP solubility have been described as, for example, insert charge (Whitacre, Lee, and Milich 2009), amino acid side chain volume (Karpenko et al. 2000), or the content of specific amino acids, such as tryptophan or arginine (Vormittag, Klamp, and Hubbuch 2020). None of these individual attributes describe the cVLP solubility landscape comprehensively. This is underlined by findings, in which combining different attributes improved the solubility model's performance (Vormittag, Klamp, and Hubbuch 2020). Each amino acid makes a unique contribution to protein solubility, e.g. based on its charge, volume or specific interactions. In recent years, a great number of so-called hydrophobicity scales have been derived that aim to serve in an (almost) calibration-free model to describe hydrophobicity-related problems based on amino acid-specific hydrophobicity values.

Already in 1962, Tanford pointed out that hydrophobic interaction is a key factor influencing the stability of globular protein conformation (Tanford 1962). Nozaki and Tanford measured transfer free energies of amino acid side chains into ethanol and dioxane, deriving an early hydrophobicity scale (Nozaki and Tanford 1971). They describe the hydrophobicity scale value of an amino acid as, for example, its tendency

to be located in the interior of a protein. This idea of a scale to describe an amino acid's tendency to partition into exterior or interior regions of a protein is an assumption that does not take into account 3-D-specific effects. If 3-D-specific effects were negligible, a linear or non-linear function should exist that perfectly describes a protein's solubility based on its amino acid composition. The fact that this is probably not the case has been extensively shown, directly or indirectly, for example, by several studies on protein solubility prediction yielding only about 60-80% accuracy (Hebditch et al. 2017; Idicula-Thomas et al. 2006; Magnan, Randall, and Baldi 2009; Smialowski et al. 2006), or detailed mechanistic studies on protein structure and assembly. The latter is illustrated by the complex behavior of VLPs. Tyrosine can be regarded as a hydrophobic (aromatic ring) or polar (hydroxyl group) amino acid. Interestingly, it is required for Hepatitis B core antigen (HBcAg) to form capsids, buried in a hydrophobic pocket (Wynne, Crowther, and Leslie 1999). A mutational form, replacing tyrosine 132 by alanine, prohibits particle assembly (Bourne et al. 2009). The predominant quaternary structure of this HBcAg mutant is therefore a dimer instead of the 180- or 240-meric capsid. This comes with great changes in physicochemical and biophysical behavior as the mass of a solvatized entity differs by 90- to 120-fold. Obviously, this behavior cannot be explained by one universal hydrophobicity scale, as this is an effect with a strong 3-D spatial component.

In a recent article by our group, we applied a soft ensemble vote classifier (sEVC) with embedded feature selection to predict cVLP solubility, based on 91 hydrophobicity scales (Vormittag, Klamp, and Hubbuch 2020) to harness the information contained in multiple scales. This can help overcome the limitations of a sequence-based approach by expanding the dimensionality of the sequence-based descriptions by using different scales in one model. In said study, a feature selection algorithm selected the best features to be included in the model based on a training set. Individual hydrophobicity scale performance for classification ranged from 54 to 85%, which underpins that hydrophobicity scales cannot be universal. The choice of hydrophobicity scales by the algorithm and the analysis of the best- and worst-performing scales revealed dominant roles for arginine and tryptophan in cVLP solubility. In another study on the prediction of peptide aggregation propensity, feature selection has been successfully employed to select the best of 560 features, showing some

overlap with regard to best features with our previous study (Fang et al. 2013). Both these publications combine theoretical physicochemical data with statistical methods to predict a biophysical property by selecting appropriate physicochemical measures. Compared to pure statistical regression, these models therefore contain physicochemical information, which is advantageous for calibration on smaller datasets and for interpretation of the data.

Zviling and colleagues came to similar conclusions in their work on the prediction of transmembrane helical regions (Zviling, Leonov, and Arkin 2005). Based on two existing amino acid scales, provided by Kyte and Doolittle and Goldman, Engelmann and Steitz (Engelman, Steitz, and Goldman 1986; Kyte and Doolittle 1982), they generated a set of new hydrophobicity scales by optimization using a genetic algorithm on a cross-validation set. Both Zviling's and our approach combine real experimental physicochemical data, contained in hydrophobicity scales, with a statistical adjustment to the problem to be modeled. This ensures that prediction is based on actual physicochemical groundwork. The degree of statistical adjustment, however, is larger, when a 20-dimensional function is optimized, such as by optimization of scale tables in Zviling's work, than with calibration of decision trees that only shift classification borders in the one-dimensional target function space.

The present article describes approaches to optimize and tweak our recently developed model to improve prediction accuracy, learn more about the data, and to extend the model to other biophysical parameters. An optimization procedure for the synthesis of amino acid scale tables is one approach to improve model performance. To ensure that overfitting is avoided, this approach would benefit from a large balanced dataset, as was used in our recent study. These synthesized scales would be tailored to the problem they are optimized on and therefore have the potential to improve model performance and reveal dominant roles of amino acids for classification of the dataset.

In our previous study, we demonstrated the potential of optimizing the model's prediction based on the contingency matrices of the individual insertion strategies (Vormittag, Klamp, and Hubbuch 2020). The dataset used consisted of 568 chimeric HBcAg constructs, created by a grid of 71 different inserts and eight insertion strategies. The eight different insertion strategies in this study define where in the major

immunodominant region of the HBcAg molecule the epitope is inserted and which amino acids are deleted. The different strategies are meant to optimize the integration of the foreign epitope into the VLP sequence and would ideally result in an integration that produces a soluble cVLP. Analyzing the strategies showed that the model systematically overestimated or underestimated certain insertion strategies with respect to the predicted solubility (Vormittag, Klamp, and Hubbuch 2020). To recapitulate briefly, a strategy that is overrepresented in the training false-positive (FP) group has overestimated solubility in relation to the other strategies. This means that this strategy is particularly bad for solubility in the perspective of the training dataset. The model is, at this stage, incapable of describing this different behavior. As previously suggested, this could be related to 3-D phenomena that cannot be described by a sequence-based approach (Vormittag, Klamp, and Hubbuch 2020). Knowledge of the above-described systematic misclassification helps a) to conclude that this insertion strategy may be disadvantageous with respect to solubility, and b) to adjust the model, so that the model's blind spot is compensated. The latter can be achieved by modifying the model predictions specifically for those insertion strategies, for which systematic misclassification can be observed in the training set.

The introduction of a foreign epitope to be displayed on the VLP surface has implications on many facets of the product and the process. The main question addressed by our work – the solubility of cVLP candidates after cell lysis – is typically a decision point where candidates drop out of the candidate pool. In this large dataset, this leaves half of the candidates to choose from (Vormittag, Klamp, and Hubbuch 2020). This number will be cut down to very few candidates throughout the development process. Besides solubility, several other biophysical or physicochemical parameters are determinants in the development process of a cVLP candidate. The most important property is the candidates' ability to induce an immune response against the target structure, the basis for its efficacy (Roseman et al. 2012; Frietze, Peabody, and Chackerian 2016; Klamp et al. 2011). Therefore, the introduced foreign epitope has to be properly displayed and accessible on the molecular surface, which is something that can very probably not be described by amino acid scale-based models and requires detailed 3-D structural studies (Roseman et al. 2012). Another process-related property that can vary among the

candidates is their structural and phase behavior as a function of the solution environment. VLPs are complex nanostructures which are held together by intra- and intermolecular forces, such as electrostatic and hydrophobic interactions and disulfide bonds. Their complex structural behavior is dependent on the introduced foreign epitope. In a previous work by our group, we investigated the re-assembly of disassembled HBcAg cVLPs (in the form of HBcAg dimers) by increasing ionic strength by diafiltration (Rüdt et al. 2019). We observed that the diafiltration volumes – an indicator of progress in buffer exchange and therefore ionic strength – that were required to complete the VLP assembly reaction varied between the three constructs. Based on zeta potential measurements, this behavior could be related to surface charge. In another study, a high-throughput 3-D structure generation workflow was developed that we applied on exactly these three constructs in their disassembled form to calculate a surface charge that correlated well with the zeta potential measurements (Klijn et al. 2019). This is a good example of in silico representations of physicochemical properties, which pave the way for model-assisted rather than empirical process development. This said, it seemed promising to test the sEVC model to predict other process-related properties. One such property is the required concentration of ammonium sulfate to precipitate cVLPs, a typical process step in cVLP downstream processing (Hillebrandt et al. 2020). Precipitation of cVLPs can typically be achieved with an ammonium sulfate concentration that leaves most of the contaminants in solution (Kazaks et al. 2017). Once the supernatant containing these contaminants is discarded, the cVLPs can be resolubilized, resulting in high yields with the potential of increasing product concentration. The ammonium sulfate concentration required for cVLP precipitation is typically determined in screening experiments (Hillebrandt et al. 2020). To reduce required time and resources, regression for the estimation of the ammonium sulfate concentration for different cVLPs would therefore be highly interesting.

We have recently shown that ensembles of individual classifiers based on hydrophobicity scales and amino acid sequences are potent classifiers for cVLP solubility. The objective of this study is to evaluate the potential of different optimization strategies to improve our recently developed sEVC framework and to apply the sEVC to another biophysical parameter. We therefore combined the sEVC with optimization

algorithms to improve generated models and to learn more about the data obtained. Optimization algorithms employed in this study aimed to (I) reduce systematic misclassification based on insertion strategies, (II) optimize and generate amino acid scale tables, and (III) combine both optimization strategies to maximize model performance. Finally, we show some perspective on how to apply the model to another biophysical parameter, i.e. ammonium sulfate concentration for cVLP precipitation, by transforming the model to a regression tool.

7.2 Materials and Methods

7.2.1 Dataset

For an overview of the methodology applied to this work, we recommend reading our previous article on the sEVC for chimeric VLP solubility prediction (Vormittag et al., 2020). The dataset is equivalent to that used in said previous study, comprising amino acid sequence and binary solubility data of chimeric HBcAg constructs. Chimeric HBcAg was based on C-terminally truncated, His-tagged Hepatitis B virus core protein, modified with 71 different inserts and eight unique insertion strategies. An insertion strategy describes where in the major immunodominant region of HBcAg the foreign epitope is inserted and how many amino acids of the native protein are deleted. All possible combinations of the 71 inserts and eight strategies result in 568 constructs/observations. The literature hydrophobicity scales used in this study can be found in our recent work and the Supplementary Material Table S7.1, originally derived from AAindex (Kawashima et al., 2007), the SPLIT 4.0 server (Juretić et al., 1993), and ProtScale (Gasteiger et al., 2005) and put together by Simm and colleagues (Simm et al., 2016). Reversed scales were treated as duplicates, and therefore removed if a non-reversed scale was available, resulting in 91 hydrophobicity scales. For all models, a training set of 384 observations was used, which was created once by stratified sampling based on the identity of the inserts and insertion strategies (Vormittag et al., 2020). The remaining 184 observations were used as an external test set. For Monte Carlo cross-validation (MC-CV), a 1:1 random split of the training set was applied to each validation run.

7.2.2 Soft Ensemble Vote Classifier

The sEVC applied in this study is described in detail in the above-mentioned recent study by our group and was only slightly modified. Briefly, the sEVC aggregates the solubility predictions of individual classifiers, which classify based on hydrophobicity features calculated by hydrophobicity scales and sequence data. The features F_p are derived by accumulating hydrophobicity values of amino acids $Hyd(aa)$, as prescribed by a hydrophobicity scale, for the entire amino acid sequence $[aa_1, aa_2, \dots, aa_n]$ of each observation p (Equation [7.1]).

$$F_p = \sum_{aa_{1,p}}^{aa_{n,p}} Hyd(aa_{i,p}) \tag{7.1}$$

Classifiers are one-level decision trees induced from these hydrophobicity features, trained using Gini's diversity index as impurity measure (Gini 1912; Windeatt and Ardeshir 2004). The classifiers return a class ('soluble'/+1 or 'insoluble'/-1) with a probability associated with the respective child node in the decision tree. The classifier's vote v is the probability with the sign associated with the respective class and therefore falls between -1 and +1. Aggregation of all votes v_i results in the continuous prediction value p of the sEVC, which is normalized by the number of included scales n, again falling between -1 and +1, as explained by Equation (7.2).

$$p = \frac{\sum_{i=1}^{n} v_i}{n} \tag{7.2}$$

This continuous prediction value is subsequently discretized, where for $p > 0$, the prediction is 'soluble' or +1 and for $p \leq 0$, it is 'insoluble' or -1. In the sEVC, an embedded feature selection algorithm informs about the potency of the individual classifiers to predict solubility and sorts them according to their feature importance, namely their Matthew's correlation coefficient (MCC) on prediction of the training data as defined in Equation (7.3).

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{7.3}$$

TP, TN, FP, and FN are true-positive, true-negative, false-positive, and false-negative classifications (contingency matrix of training, validation, and test set). In our previous study, feature selection was based on

accuracy, which is, however, biased when unbalanced datasets are considered (Powers, 2011). For model validation, an MC-CV procedure with embedded feature selection is run to inform about the optimal number of included classifiers. The sEVC could theoretically be composed of any combination of available classifiers, where each combination is an individual model. In this study, the n best classifiers, according to feature selection, are included, where n ranges from 1 to the maximum number of available classifiers. For the dataset of literature scales, n is 91. In the scale generation procedure described below, n ranges from 1 to 16.

The training/validation set is selected randomly from the full training set in each MC-CV run. The sEVC including 1 to n classifiers, sorted by descending feature importance, is probed on these MC-CV datasets. In the original study on the sEVC, model validation datasets were newly constructed for each of the 91 models, while in this study, the same MC-CV dataset within one MC-CV run is used for all n models. This is reasonable as each of the models is evaluated on the same dataset within one MC-CV run, thus increasing comparability while reducing computational resources. The validation procedure informs about the model performance dependent on the number of included classifiers. This information can be used for generation of the final model based on all training data to predict the external test set.

7.2.3 Optimization Based on Insertion Strategies

An optimization algorithm focusing on the insertion strategies was developed, based on a loop of model generation, evaluation, and modification (Figure 7.1A). The dataset used for this optimization procedure is the training set of 384 observations and all 91 literature hydrophobicity scales (Supplementary Material Table S1). In the first iteration, a 25-fold MC-CV is computed and an accumulated validation set contingency matrix is calculated. This matrix contains information on the accumulated number of validation TP, TN, FP, and FN classifications dependent on the insertion strategy for all 25 MC-CV runs. The largest absolute value of $FN - FP$ classifications defines which strategy's prediction is modified in this iteration and what the sign of this modification is. If $FN > FP$, it is positive, if $FN < FP$, it is negative. A strategy that has more FN than FP should be classified more positively by the classifier, in order to push FN observations into the TP group.

This is realized by modifying the accumulated continuous prediction values (compare also Equation [7.2]) of the sEVC. A modification vector m contains the information on how this aggregated prediction value is modified individually for each strategy, imaginable as shifting the classification boundary (Figure 7.1A) (while, in fact, the predictions are shifted instead of the classification boundary). In each iteration, the vector is changed by an absolute 0.01 for the strategy and sign identified as described above. The first iteration is calculated on an unmodified model, providing the modification vector for the second iteration.

In each iteration, the sEVC modified by the previous modification vector is evaluated in the 25-fold MC-CV, resulting in a new median validation MCC value, i.e. target function to be maximized. This value is the median of all models' median validation MCC. Therefore, the target function takes into account the entire model space of 1 to 91 included classifiers. If this is in an acceptable range (equal to or at maximum 0.05 worse than the best previous median MCC), then the accumulated contingency matrix of this iteration is used to calculate a new modification vector for the next iteration. The acceptable range was determined in pre-experiments, so that early termination of the algorithm was avoided while limiting model deterioration. If the MCC is better than the best MCC so far, the modification vector is stored as best modification vector, and future iterations are compared to this MCC. If the MCC is worse and outside the acceptable range, the modification vector is reset to the current best modification vector. If for 50 times, no improvement on the MCC has been made, the optimization is stopped. For the evaluation of optimized model performance, the model with the best MCC during optimization is chosen and the respective modification vector is applied within the sEVC in order to predict in 1000-fold MC-CV and to predict the external test set. For each of the 91 models, the validation procedure results in a median validation MCC and accuracy. For evaluation, the median of these values ('overall median') is compared for the MCC and accuracy, respectively. The metric to compare initial and optimized model performance is the change of these values in percent, where the difference between optimized and initial performance values is divided by their maximum range, i.e. 1 and 2 for accuracy and MCC, respectively.

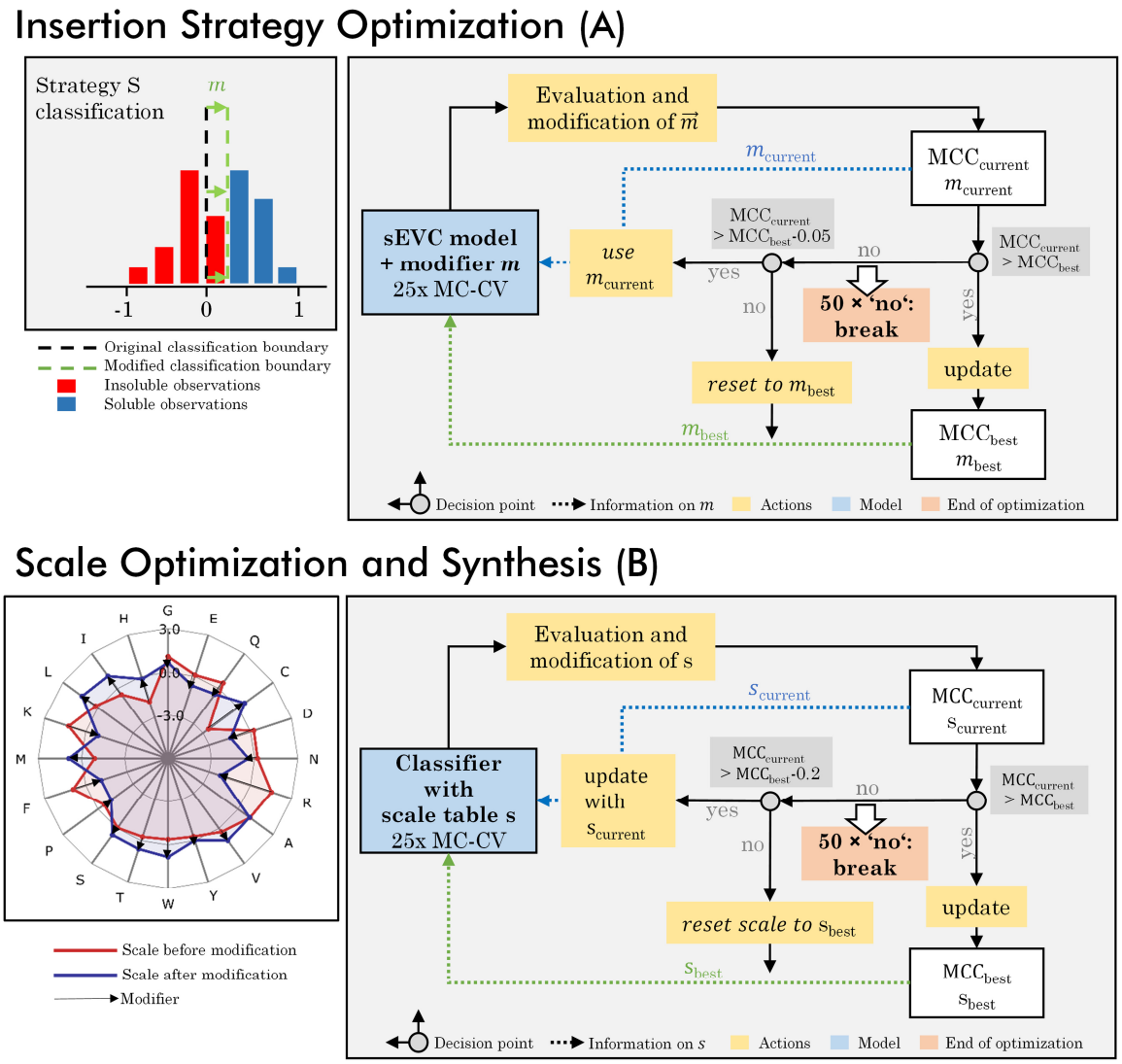


Figure 7.1: Workflow of optimization procedures. **(A)** The insertion strategy optimization is based on modifying the classifier for a specific strategy to increase model performance. In a 25-fold Monte Carlo cross-validation (MC-CV), the previous modifier is evaluated and a new modifier is derived, based on systematic misclassification in the false-positive and negative group, specific for certain insertion strategies. This results in the current modifier $m_{current}$ and current Matthew's correlation coefficient $MCC_{current}$. If $MCC_{current}$ is better than previous best MCC, the best MCC and modifier are updated and used for the next iteration. If the $MCC_{current}$ is lower than the best MCC within a defined acceptance margin of 0.05, the current modifier is used for the next iteration. If it is below this acceptance margin, the MCC and modifier are reset to the previous best MCC and modifier. If for 50 times, no improvement on the MCC has been made, the iteration is stopped. **(B)** Scale optimization and synthesis are based on an optimization of the individual amino acid's hydrophobicity values in the hydrophobicity scale. In each iteration the previous scale is modified and probed in 25-fold MC-CV, resulting in a current MCC and scale s. The iteration rules are comparable to the insertion strategy optimization, with the

difference that the acceptance margin is higher with 0.2 and the modified scales, as opposed to the modifiers in the insertion strategy optimization, are stored and updated.

7.2.4 Synthesis of Amino Acid Scales

A second algorithm was created to modify amino acid scales to I) synthesize new amino acid scales and II) optimize existing scales specifically for the presented VLP solubility problem (Figure 7.1B). The two algorithms are almost identical and are, in the following, explained by the example of scale synthesis. Each scale is optimized from an initial scale that contains normally distributed pseudorandom numbers for each of the 20 encoded amino acids. In each iteration, the scale s of the preceding run is adjusted with a modifier m (Equation [7.8] and [7.9]). The modifier is designed to move the average FP and FP feature value in the direction of the classification boundary, which is the cut point of the one-level decision tree, thus aiming to decrease false classification. This is done independent of insertion strategies. The modifier's direction is determined by average feature values of the classification groups FN and FP and the difference in frequency of individual amino acids. The average feature values F_{FN} and F_{FP} and mean amino acid frequency vectors a_{FN} and a_{FP} are derived from a 25-fold MC-CV run with the scale s of the previous iteration. Herein, the amino acid frequency vectors describe the frequency of the individual amino acids within the groups of FN and FP classification, respectively. The average feature values f_{FN}, f_{FP}, and their differences are

$$f_{FN} = a'_{FN} s, \tag{7.4}$$

$$f_{FP} = a'_{FP} s, \tag{7.5}$$

and

$$\Delta f = f_{FN} - f_{FP}. \tag{7.6}$$

The vector of the difference in amino acid frequency is given by

$$\Delta a_{FN,FP} = (a_{FN} - a_{FP}) \tag{7.7}$$

The modifier used in this optimization loop is vector

$$m = \Delta f \Delta a_{FN,FP}, \tag{7.8}$$

which is used in a centered and unit variance-scaled form $\bar{m}$. In each iteration, a scale is modified as prescribed in Equation (7.9).

$$s_{i+1} = s_i + r\bar{m}, \tag{7.9}$$

where s_i and s_{i+1} are the previous and the modified scale, respectively, and r is the modification rate, which was 1% for scale synthesis. After modification, the new scale is also centered and scaled to unit variance. Therefore, the extent of modification is comparable in all iterations, as it corresponds to an average of 1% of unit variance.

The modified scale is probed in a 100-fold MC-CV run resulting in a median MCC value. This current MCC value is compared to the best previous MCC value. If it is better, it is stored as the new best MCC value with associated new best scale. If it is worse, the new scale is still accepted, as long as the MCC does not fall below an acceptance margin, which is 0.2, where the scale and MCC are reset to the previous best iteration. The acceptance margin is larger than in the insertion strategy-based optimization, as model performance fluctuated more with this second optimization strategy. If no new best scale is created for 50 times consecutively, the optimization is stopped and the best scale and MCC are returned. For the generation of scales, either the full training set or subsets thereof were used. When the full training set is used, one scale is generated by the algorithm. When two (equally sized) subsets are used, two scales are generated by the algorithm. The algorithm was run with up to 16 subsets, which in turn resulted in 16 different scales. Subsets were either created by random split into evenly sized subsets or split by insertion strategies.

Additionally, this algorithm was used to optimize literature scales. Based on the full training set with 384 observations, the 91 literature scales were used as scales in an initial iteration, where the optimization was performed for each scale individually at a rate of 5%. Other parameters were identical to the scale generation procedure.

7.2.5 Analysis of Performance Data for Optimizations

Evaluation of optimized and non-optimized models was always based on 1000-fold MC-CV, returning median MCC and median absolute deviation (MAD) of MCC. The external test set consisted of 184 observations,

remaining after stratified sampling of 384 training observations from the full dataset. For all models, the same external test set was used.

7.2.6 Redesigning the Model for Regression of Precipitation Data

Ten cVLP constructs of strategy H were experimentally evaluated for cVLP-precipitating ammonium sulfate concentration. The ammonium sulfate concentration screening procedure was performed as described in a recent article on precipitation of HBcAg VLPs (Hillebrandt et al., 2020). Briefly, clarified *E. coli* lysate, containing HBcAg VLPs, was adjusted to 0.25% polysorbate 20 and then precipitated with 4 M ammonium sulfate stock solution to different target concentrations. The ammonium sulfate concentration required to precipitate most of the cVLPs was determined visually based on SDS PAGE scans.

Scales generated by the above-described algorithm, that is those derived from randomly splitting the training set into eight equal parts, were used to train a model based on all observations with insertion strategy H. Evaluation of the model was performed on the prediction of the 1000-fold MC-CV set for eight models composed of 1-8 classifiers. As opposed to the classification for solubility, the continuous prediction value of the models (compare also Equation [7.2]) was not discretized. The mean resulting prediction value of the MC-CV runs was subsequently used to be correlated with the experimental data in linear regression. The order of the scales was derived from feature selection. The data were fit using MATLAB's *fitlm* function and evaluated by the ordinary R^2.

7.3 Results and Discussion

7.3.1 Optimization Based on Insertion Strategies

An optimization procedure was developed, which, based on 25-fold MC-CV, adjusts the model's predictions for the insertion strategies individually based on a modification vector. This modification vector is applied before discretizing the continuous scale of the aggregated sEVC votes by increasing (higher solubility) or decreasing (lower solubility) the continuous prediction value.

The optimized model, obtained after 130 iterations, showed an increase in median validation MCC values from .63 to .69 (Table 7.1). Most

notable modifications are made on predictions of insertion strategies E and H, resulting in a strategy-specific accuracy increase of 12% for both strategies, while the MCC increased by 8% for strategy E and decreased by 1% for strategy H. This is also illustrated by the number decrease of these strategies in the respective false classification groups (Figure 7.2). Overall, there is a similar true-positive (TP) and false-negative (FN) number, indicated by the mean over insertion strategies (red line), while true negative (TN) is increased and false positive (FP) decreased. The constant numbers in FN are explained by the increase in the number of strategy H in this group, balancing out the decrease of strategy E in FN. Making strategy H more negative pushed FP-classified observations to TN, but also TP to FN. Strategy E performed better in this regard, as we only see a minor increase in the number of E in FP.

|→

Table 7.1: Modification vector and summarized model Monte Carlo cross-validation (MC-CV) performance data for the insertion strategy optimization. The final modification vector (m) with elements for each strategy is shown, affecting the continuous prediction values after aggregation of all votes (Figure 7.1A). Accuracy (A) and Matthew's Correlation Coefficient (MCC) before (start) and after optimization (opt) are shown for each strategy. The percent change values in accuracy and MCC are calculated by the absolute change in accuracy and MCC, respectively, divided by the range of the statistical value (1 for accuracy, 2 for MCC). Overall median is based on the median performance data of the 1000-fold MC-CV. The MC-CV results in 91 values, which are the median of the 1000 MC-CV runs for each number of included classifiers individually. These values are illustrated for the optimized model in Figure 7.3A, right. The overall median describes the median of these 91 values, which is the optimization target function. Overall median MCC and accuracy are also shown for strategy-based optimization of scales optimized and synthesized with the scale table optimization algorithm. *Change of MCC and accuracy of optimized/generated scales are calculated relative to the performance of literature scales without optimization (A_{start} and MCC_{start}).

Best modification vector for insertion strategy-based optimization only

Strategy	A	B	C	D	E	F	G	H
m	0.01	0	-0.01	0	0.59	-0.1	0	-0.5

Insertion strategy-based optimization only

Strategy	A	B	C	D	E	F	H	I	Overall median
$\mathbf{A_{start}}$	0.87	0.78	0.83	0.87	0.78	0.83	0.92	0.58	0.81
$\mathbf{A_{opt}}$	0.88	0.78	0.83	0.87	0.90	0.83	0.92	0.70	0.84
$\mathbf{A_{change}}$	1%	0%	0%	0%	12%	0%	0%	12%	4%
$\mathbf{MCC_{start}}$	0.75	0.56	0.66	0.74	0.64	0.67	0.83	0.39	0.63
$\mathbf{MCC_{opt}}$	0.76	0.56	0.65	0.74	0.79	0.66	0.83	0.37	0.69
$\mathbf{MCC_{change}}$	1%	0%	0%	0%	8%	0%	0%	-1%	3%

Insertion strategy-based optimization combined with scale generation and optimization

Optimization of literature scales:	A_{opt}	0.86
	$A_{change*}$	**5%**
	MCC_{opt}	0.72
	$MCC_{change*}$	**5%**
Generation of scales: Example subset $S_{8,1}$:	A_{opt}	0.86
	$A_{change*}$	**5%**
	MCC_{opt}	0.73
	$MCC_{change*}$	**5%**

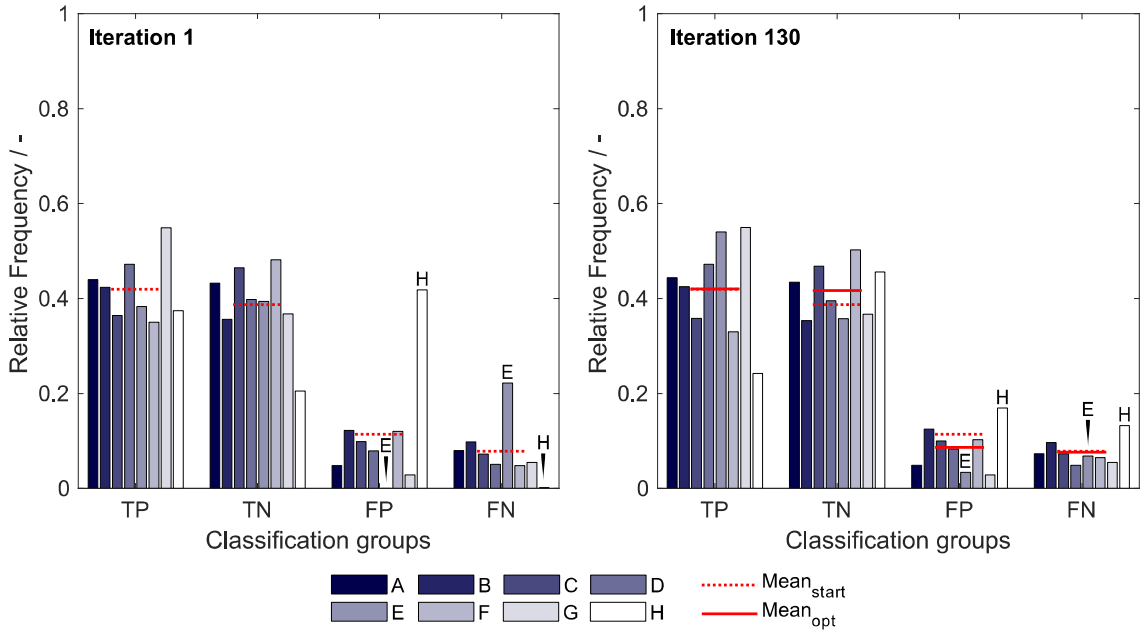


Figure 7.2: Relative frequency of classification groups based on insertion strategies A-H in the first iteration (left) and the best optimization iteration (right) during insertion strategy-based

optimization with the 91 literature scales. The mean of the relative frequencies within a classification group is shown for the first iteration ($Mean_{start}$) and for the best optimization ($Mean_{opt}$), indicating that through optimization, the FP group decreases in mean relative frequency while the TN group increases in mean relative frequency. Strategies E and H are marked additionally to guide the eye. TP: true positive; TN: true negative; FP: false positive; FN: false negative.

During the optimization, both median validation and training MCC as well as external test set MCC of 91 models are increasing (Figure 7.3A, left). Their maxima approximately coincide, underpinning the usefulness of the validation MCC-based optimization procedure. While this median MCC of all models (including 1-91 classifiers) describes the general tendency of model improvement, it is also valuable to have a closer look on the improvement of the individual 91 models.

During the optimization, both training and test MCC increase for most models, when 1 to about 80 classifiers are included (Figure 7.3B). However, models deteriorate at roughly >80 included classifiers. This said, the most important area is where the MCC is maximal (30-40 for the test set). Here, the optimization algorithm continuously improves the models with regard to training and test set MCC, where the last iteration shows highest MCC values for the individual models. To select the appropriate number of included classifiers, validation data is useful. The validation data of the optimized model generally follows the course of the external test data (Figure 7.3A, right). Their maxima do not coincide. However, choice of the best model with regard to validation MCC also produces a reasonable model for the prediction of test data with a test MCC of .65 at 48 included scales. Interestingly, the optimum number of included classifiers with regard to test MCC is 34 with an MCC of .70 (Table 7.2), similar to an optimum of 29-30 included classifiers as described in our previous study with the basic sEVC (Vormittag et al., 2020).

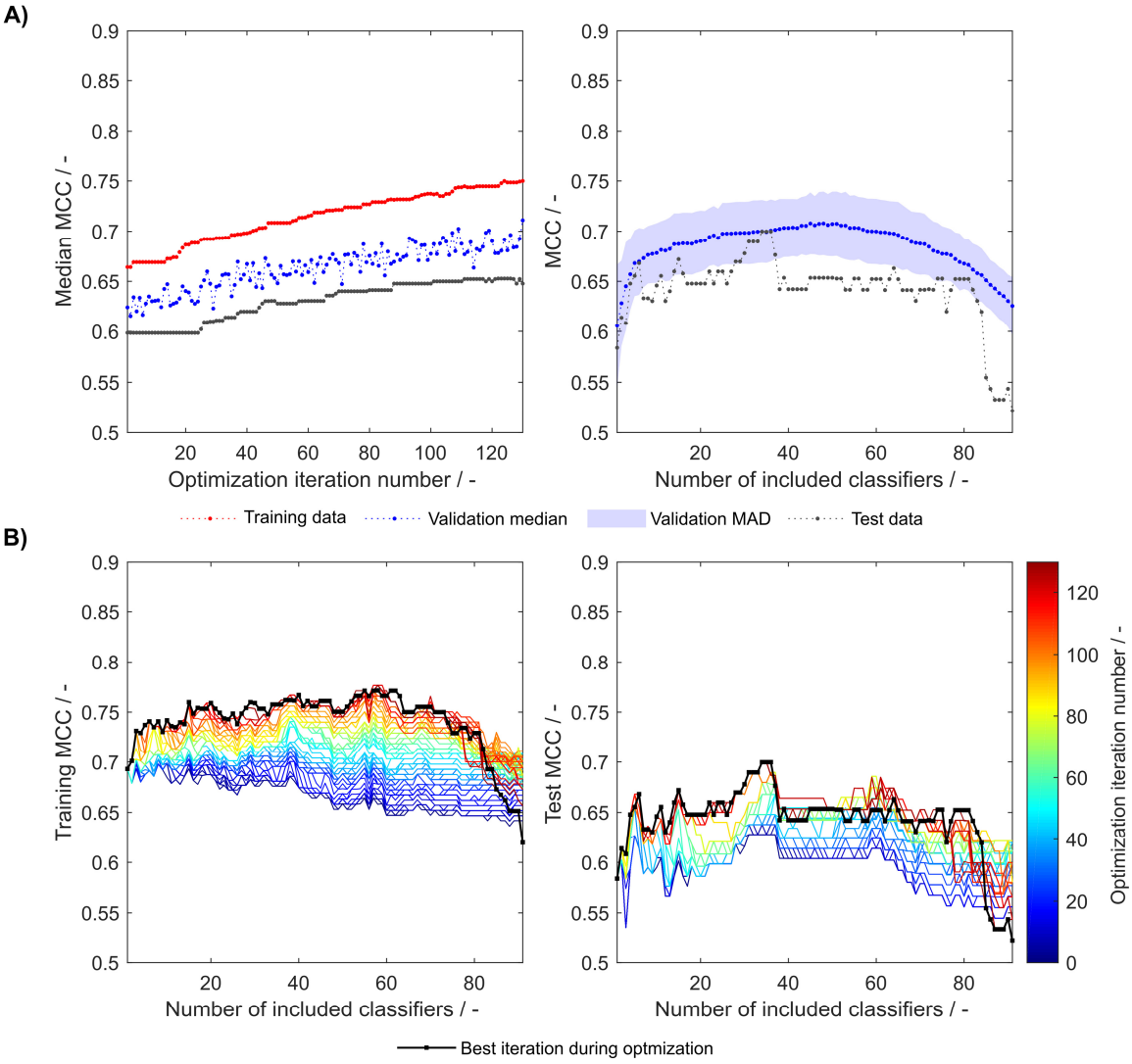


Figure 7.3: Matthew's correlation coefficient (MCC) during insertion strategy-based model optimization. Scales used were the unmodified 91 literature scales. Median MCC are shown for training, validation, and test data over optimization iterations **(A, left)**. Validation and test MCC are shown over number of included classifiers in the soft ensemble vote classifier (sEVC) **(A, right)** for the best model in the optimization procedure. The median absolute deviation (MAD) of the validation MCC above and below the median validation MCC is visualized with a shaded area. Training and test MCC over number of included classifiers are shown for the optimization iterations until the best iteration, where median validation MCC was highest **(B)**. Optimization iterations are illustrated by a colormap, where dark blue represents the first iteration and dark red the best iteration, highlighted by the black dots.

7.3.2 Synthesis and Optimization of Amino Acid Scale Tables

Another option to optimize the model relates to the amino acid scale tables. The target for such optimizations was seen in the feature values and amino acid composition in the *FP* and *FN* groups. Adaptation of the scale tables was performed, in a way that amino acids predominant in the respective groups were altered in their scale table values to push observations that have been predicted falsely over the classification boundary, i.e. decision tree cut point. This is illustrated by the following example. Let us assume that *FP* has a lower mean feature value than *FN*, and, for example, that valine has a higher frequency in *FP* than *FN*. Observations in *FP* are classified positive but their data label is negative or insoluble. If we wanted observations of *FP* to be classified rather insoluble, their feature value would have to be increased for false observations to cross the classification boundary. This said, the classification boundary is not static, but changes with alterations in the amino acid scale table. Therefore, small increments are made and scale improvement is monitored. Note that the aim is to increase *FP* hydrophobicity, but decrease *FN* hydrophobicity. Considering that valine is more frequently observed in the *FP* group, increasing valine's hydrophobicity value in the scale would be beneficial, as it would increase the average *FP* hydrophobicity value more than the average *FN* hydrophobicity value. If this is executed for all amino acids, the *FP* feature value would ideally be increased and the *FN* feature value decreased, increasing overall correlation.

This optimization procedure has been performed on the entire training dataset (384 observations) and equally sized subsets, where the number of subsets, and therefore synthesized scales, was 1-16, resulting in subsets with 384 to 24 training examples. Each of the optimization procedures was performed 20 times, resulting in 320 scale tables $S_{[\text{number of scales}],[\text{number of repetition}]}$. For evaluation of the synthesized scales, the MCC of the external test set prediction at optimal number of classifiers as determined by validation (maximum validation MCC) is compared (Figure 7.4). This validation-based model selection is a useful strategy to select the optimal number of included classifiers and thus the model. Additionally, the maximum MCC of the external test set prediction is evaluated. Both metrics increase to a maximum from one to five generated scales, where the median of maximum test MCC is .71 and the median of test MCC

at maximum validation MCC is .70. From this maximum towards a higher number of training subsets and likewise number of synthesized scales, there is a tendency of decreased model performance. Also, best test MCC and test MCC at best validation diverge more, probably since training subsets become smaller and the probability of more unrepresentative scales being synthesized rises, thus potentially decreasing the power of validation for model selection. From this data and with the present dataset, it would be recommended to synthesize scales from five subsets, although most other scales also perform reasonably well.

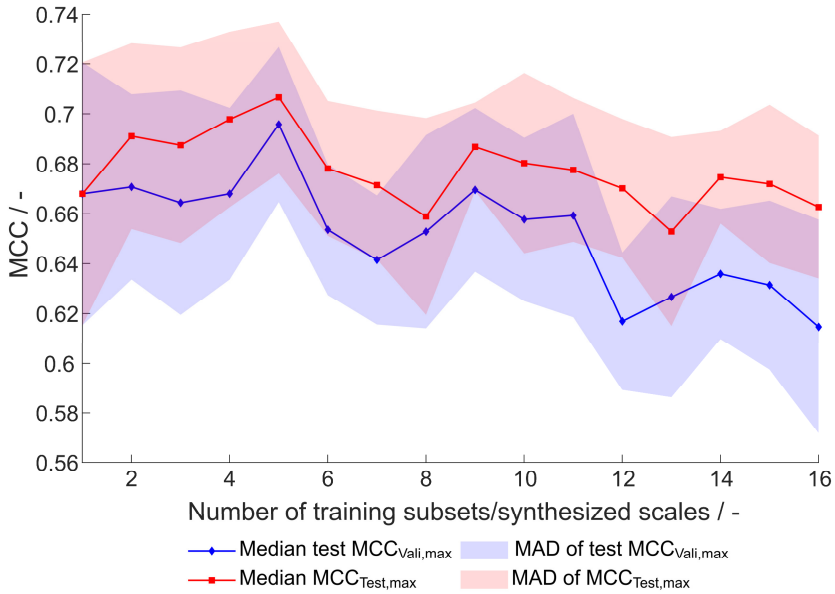


Figure 7.4: Test set Matthew's correlation coefficient (MCC) of synthesized scales. For each number of training subsets/synthesized scales, 20 repetitions of scale generation were performed. Median and median absolute deviation (MAD) of best test set MCC and test set MCC of best model by validation MCC are shown.

The overall best scale table with regard to best test MCC at maximum validation MCC is $S_{1,5}$ with test $MCC_{vali,max}$=.77 (Table 7.3), which is significantly better than with literature scale tables with test $MCC_{vali,max}$=.63. With respect to the 20 repetitions of scale synthesis, median test MCC at maximum validation MCC for one scale is worse than for five scales (Figure 7.4). The best scale table with five scales, $S_{5,17}$, shows $MCC_{vali,max}$=.75. For the best model by validation MCC, these scale tables show a test set accuracy of .86 and .88, respectively, corresponding to 155 and 158 correctly classified constructs in the test set of 184 observations.

The generation of subsets for scale synthesis was additionally investigated with subsets containing one insertion strategy each, amounting to eight different subsets. The median of maximum test set MCC and test MCC at maximum validation MCC were .67±.03 and .65±.03, respectively (data not shown). They were comparable to randomized training subset generation with eight subsets, showing a maximum test set MCC and a test MCC at maximum validation MCC of 0.66±0.04 and 0.65±0.04, respectively (Figure 7.4). Therefore, strategy-based generation of subsets for scale synthesis was not advantageous to random subset generation.

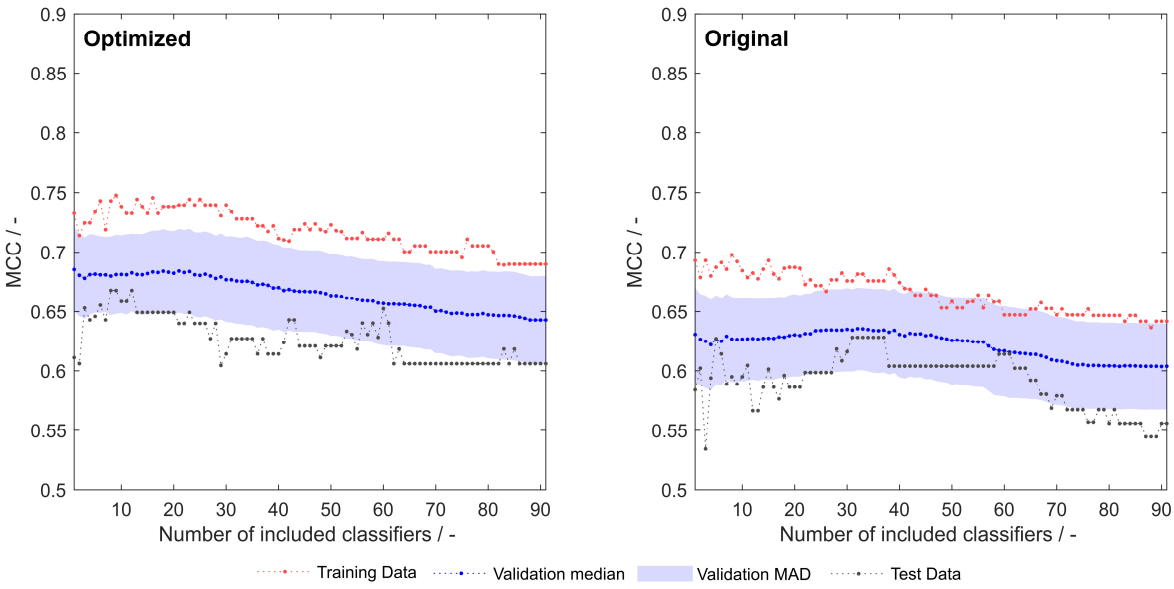


Figure 7.5: Training, validation and test sets Matthew's correlation coefficient (MCC) of optimized (left) and original (right) 91 literature hydrophobicity scales. The shaded area represents the median absolute deviation (MAD) during 1000-fold Monte Carlo cross-validation.

Optimization of the 91 literature scales with the same algorithm that synthesized scales as described above resulted in an improvement of training, validation and test set MCC over the whole model space (Figure 7.5). A greater rate during optimization was chosen (5%), as the lower rate employed for scale synthesis resulted in early optimization termination with no significant improvements in model performance (data not shown).

7.3.3 Combination of the Optimization Procedures

As both above-described optimization procedures tackle different challenges, it seems promising to combine these by adding an insertion

strategy-based optimization procedure after synthesizing or optimizing hydrophobicity scales. Strategy-based optimization of optimized literature scale tables results in similar trends but increased model performance compared to models with unmodified literature scale tables (compare Figure 7.3 and Figure 7.6). With optimized scale tables, the resulting MCC values are higher for the training, validation and test sets (Figure 7.6A). The maximum test set MCC and the test set MCC at maximum validation MCC are increased to 0.72 and 0.71, respectively, as compared to 0.70 and 0.65 before scale table optimization (test set performance data summarized in Table 7.2). Additionally, the performance of the model at very low and high numbers of included classifiers is benefitted, never falling below an MCC of 0.6 for the test set (Figure 7.6B). The number of insertion strategies in the false classification groups show similar trends as without scale table optimization, underlining that the insertion strategy-based optimization procedure is effective (Appendix E, Supplementary Material Figure S7.1).

Table 7.2: Performance data of selected models on the external test set. Best test set Matthew's Correlation Coefficient (MCC) and test MCC of model with best validation MCC in 1000-fold MC-CV are shown. *Accuracy is shown for the model with best validation MCC. Best models' performance data are written in bold. LS: Literature scales; SO: Strategy optimization; LS_{opt}: Scales optimized with scale table optimization workflow; $S_{x,y}$: Generated scale table with x scales, optimization procedure y of 20.

	91 LS	91 LS,SO	91 LS_{opt}, SO	$S_{1,5}$	$S_{5,17}$	$S_{8,1}$	$S_{8,1}$,SO
$\mathbf{MCC_{max}}$	0.63	0.70	0.72	**0.77**	**0.77**	0.72	0.76
$\mathbf{MCC_{vali,max}}$	0.63	0.65	0.71	**0.77**	0.75	0.72	0.71
$\mathbf{A_{vali,max}}$*	0.81	0.83	0.85	0.86	**0.88**	0.86	0.85

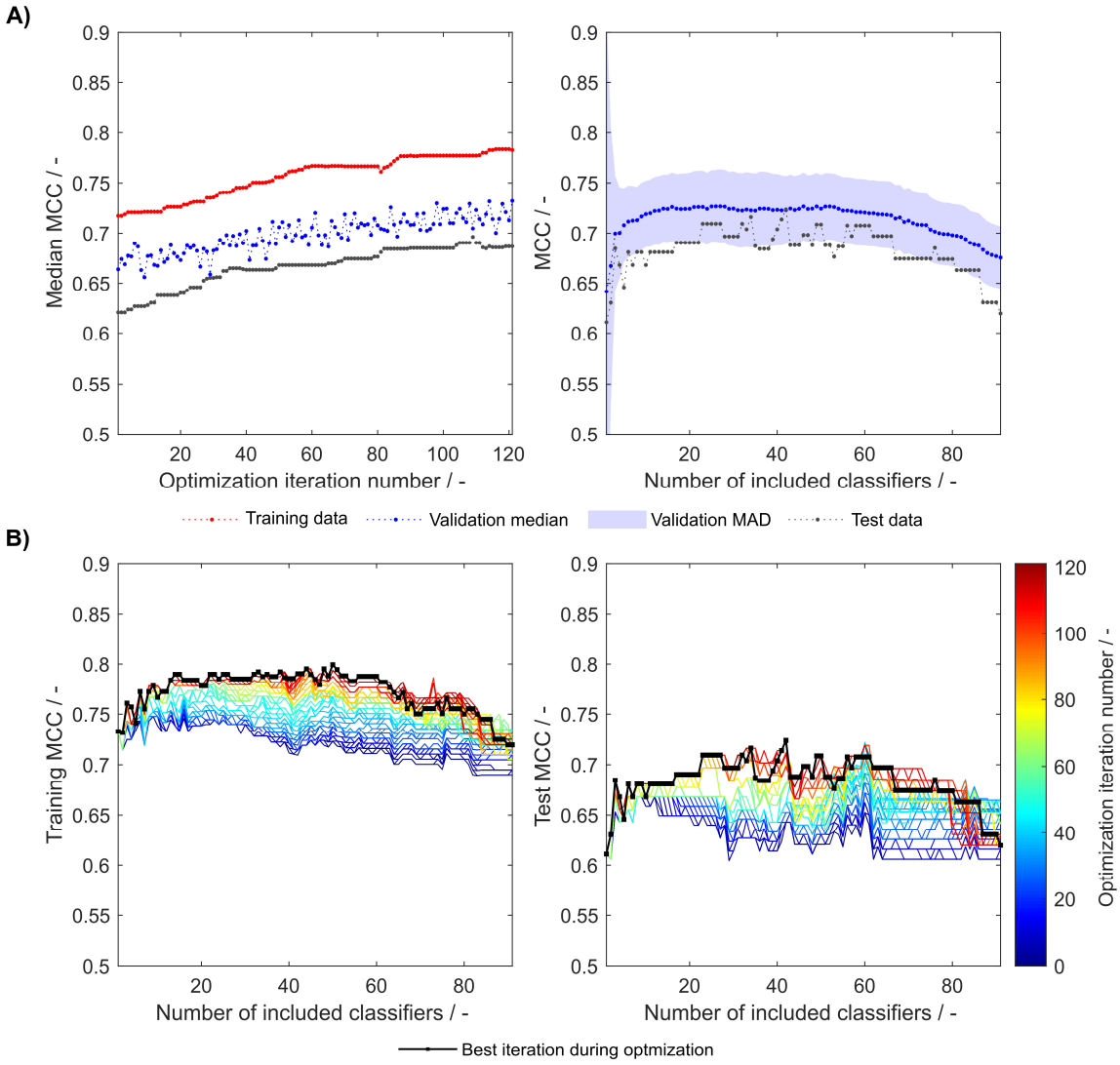


Figure 7.6: Matthew's correlation coefficient (MCC) during insertion strategy-based model optimization. Scales used were the optimized 91 literature scales (see also Figure 7.4). Median MCC are shown for training, validation, and test data over optimization iterations **(A, left)**. Validation and test MCC are shown over number of included classifiers in the soft ensemble vote classifier (sEVC) **(A, right)** for the best model in the optimization procedure. The median absolute deviation (MAD) of the validation MCC above and below the median validation MCC is visualized by a shaded area. Training and test MCC over number of included classifiers are shown for the optimization iterations until the best iteration, where median validation MCC was highest **(B)**. Optimization iterations are illustrated by a colormap, where dark blue represents the first iteration and dark red the best iteration, highlighted by the black dots.

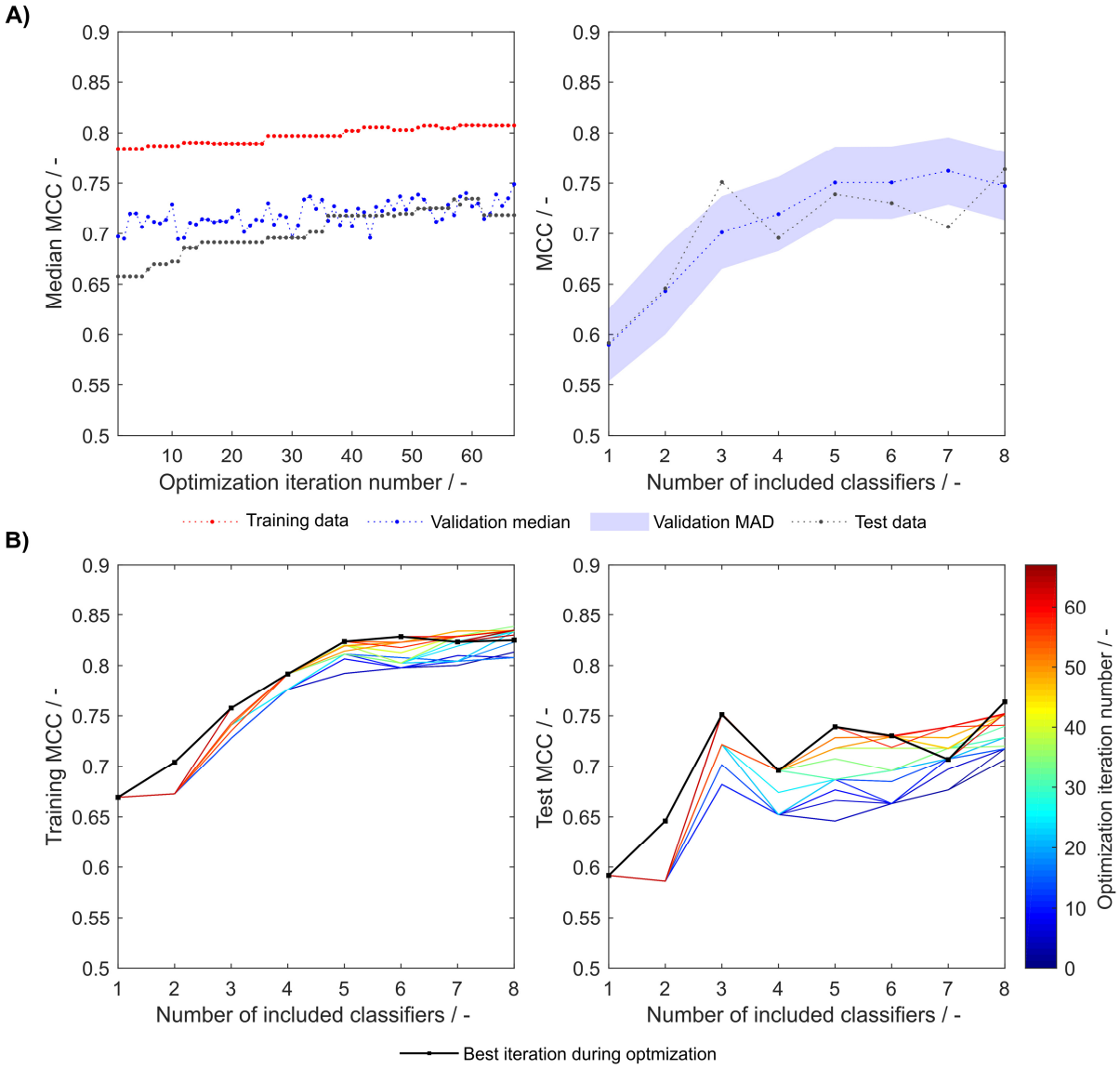


Figure 7.7: Matthew's correlation coefficient (MCC) during insertion strategy-based model optimization. Scales used were eight generated scales from scale table set $S_{8,1}$. Median MCC are shown for training, validation, and test data over optimization iterations **(A, left)**. Validation and test MCC are shown over number of included classifiers in the soft ensemble vote classifier (sEVC) **(A, right)** for the best model in the optimization procedure. The median absolute deviation (MAD) of the validation MCC above and below the median validation MCC is visualized by a shaded area. Training and test MCC over number of included classifiers are shown for the optimization iterations until the best iteration, where median validation MCC was highest **(B)**. Optimization iterations are illustrated by a colormap, where dark blue represents the first iteration and dark red the best iteration, highlighted by the black dots.

Table 7.3: Best five scale tables, measured by highest test set Matthew's Correlation Coefficient (MCC) at maximum validation MCC. Scale tables are centered and scaled to unit variance.

	1 scale		**3 scales**						**5 scales**				
	$\mathbf{S_{1,5}}$	$\mathbf{S_{1,20}}$	$\mathbf{S_{3,3}}$			$\mathbf{S_{3,16}}$			$\mathbf{S_{5,17}}$				
$MCC_{vali,max}$/ Amino acid	**0.77**	**0.76**	**0.76**			**0.76**			**0.75**				
A	-0.320	0.092	1.087	-0.088	-0.281	1.371	-0.161	-0.056	0.847	-0.143	0.493	0.784	0.418
R	0.168	0.312	0.454	0.867	0.201	-0.122	-0.460	0.194	-1.902	-0.845	-1.492	0.044	-0.306
N	-0.668	-0.017	0.046	1.390	0.217	-0.403	-0.519	-1.132	0.050	0.377	-1.687	1.458	-1.545
D	-0.751	-1.758	-0.771	-2.164	-0.555	-0.817	-0.375	-1.697	0.159	0.165	-0.368	-0.836	0.724
C	0.182	1.395	0.801	-0.660	2.344	2.143	-1.505	-1.437	-0.719	0.893	0.549	0.205	0.302
Q	0.410	0.256	-0.263	0.240	0.009	-0.898	0.285	0.181	0.015	0.386	0.913	0.047	-0.444
E	-0.345	-0.656	-0.283	-0.942	-0.318	-0.396	-0.198	-0.298	-0.439	-0.444	-0.984	-0.744	-0.262
G	0.422	0.284	-0.289	-0.563	-0.190	-1.103	-1.659	-1.955	-0.073	1.105	1.775	2.545	0.446
H	-0.552	-0.659	-2.698	0.134	-0.891	-2.272	0.337	0.884	1.179	-1.168	-1.083	-0.052	-1.557
I	-0.321	0.795	0.271	0.059	-0.316	0.446	-0.310	-0.338	0.982	0.223	0.128	-1.003	-0.436
L	0.090	-0.118	-0.008	0.298	-0.478	-0.010	-0.276	0.371	-0.789	-1.022	0.432	-0.893	-0.393
K	0.383	0.269	-0.286	-0.014	0.018	-0.266	0.216	0.034	-1.569	1.379	-0.576	-0.050	0.505
M	-0.730	-0.445	-0.601	-0.904	-2.212	-0.072	-1.271	-0.301	1.773	-2.126	-0.550	-1.816	1.239
F	3.038	-0.027	0.944	0.552	-0.294	1.117	0.822	1.835	-0.206	-0.410	0.055	-0.426	2.380
P	0.747	1.745	-0.576	1.395	0.980	0.227	0.395	1.206	0.751	0.865	-0.661	0.760	1.058
S	0.645	0.077	-1.055	0.426	0.385	-0.687	0.475	0.367	-0.062	-0.027	1.154	-0.702	-0.276
T	-1.853	-2.579	-0.003	-1.159	-0.334	-0.107	1.314	1.178	1.089	-1.397	0.735	-0.759	-0.249
W	0.604	1.002	2.331	1.474	2.216	1.450	2.898	0.673	0.224	1.988	0.980	0.854	-0.010
Y	-1.469	-0.766	0.743	-1.394	-0.766	-0.022	-0.204	-0.498	-1.803	0.030	-1.130	-0.281	-1.918
V	0.320	0.797	0.157	1.054	0.267	0.421	0.195	0.788	0.491	0.171	1.317	0.863	0.326

As another example, the first set of scales generated from eight training subsets ($S_{8,1}$) was tested with the strategy-based optimization algorithm. Figures 7.7A and B show that synthesized scales still can benefit from this optimization procedure resulting in higher MCC values for training, validation and test sets for most models. Compared to the 91 literature scales, these models perform 5% better with regard to validation accuracy and MCC (Table 7.1). Model test set MCC is increased for all models except for the sEVC including seven scales, which remains at a comparable test set MCC as before strategy-based optimization (Figure 7.7B), resulting in slightly decreased test MCC and accuracy at best validation MCC (Table 7.2). This shows that an improvement in median performance of models does not necessarily result in an improved prediction outcome. Additionally, well-performing scale tables such as $S_{1,5}$ and $S_{5,17}$ were tested with the strategy-optimization workflow. However, strategy-based optimization failed to improve model performance using these scales, suggesting that with these scales, systematic misclassification based on insertion strategies is not an issue. This in turn shows, that this systematic misclassification can be reduced by the use of other scale tables, and not only by the strategy-based optimization. This contradicts the assumption that insertion strategy-based misclassification is a 3-D-specific effect that cannot be captured by

an amino acid sequence-based approach (Vormittag et al., 2020). The strategy-based optimization could theoretically be performed for all 20×16 generated scale sets, but would go beyond the scope of this research.

7.3.4 Correlation of Scales within Scale Tables

As pointed out earlier, the explained variance of the first principal component (PC) from a principal component analysis (PCA) on the 91 literature scales revealed that already 69% of the variance is explained with one single PC (Vormittag et al., 2020). This indicates that a significant degree of correlation between the 91 literature scales is present. After the optimization procedure, this explained variance remained at a comparable level of 66% (data not shown). The explained variance of the synthesized scale tables' first PC after PCA varied from 100% to 20% (Figure 7.8). An explained variance of 100% is predefined for the situation where only one scale was generated from the training set, as the first PC equals this scale. From 2 to 16 scales, the explained variance is below the above- mentioned value for the literature scales. It can therefore be deduced that the correlation between synthesized scales is reduced as compared to literature scales. This can be interpreted as increased orthogonality, which was expected to increase model performance of the sEVC. Decreased correlation between scales could explain the improved performance of synthesized amino acid scales in the ensemble of classifiers, as described above. PCA of the group of scales synthesized from dataset division by the eight insertion strategies reveals that with 28.3%±3.3% of explained variance, this approach is comparable to random division into eight insertion strategies with 31.6%±3.0% (data not shown). This suggests that the correlation between generated scales can probably not be reduced by splitting the dataset by the insertion strategies. As discussed above, model performance did not improve with this subset generation strategy either.

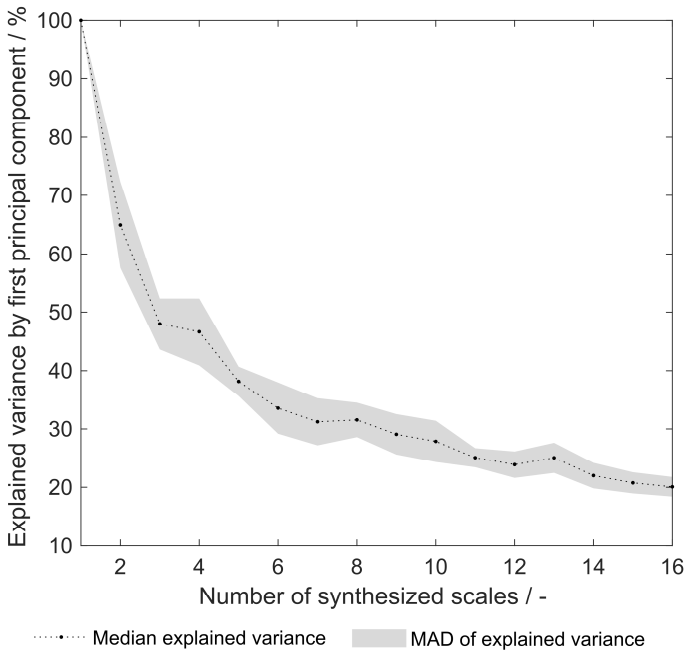


Figure 7.8: Explained variance by first principal component of synthesized hydrophobicity scale tables. For each number of synthesized scales, corresponding to the number of training subsets that were generated for scale synthesis, the median of 20 repetitions is shown. The shaded area represents the median absolute deviation (MAD).

7.3.5 Amino Acids with Characteristic Hydrophobicities

The three best literature scales by feature selection in 1000-fold MC-CV show very similar hydrophobicity values (Figure 7.9). This is partly the case because they are either related to each other or because they were generated from similar original scales (Eisenberg et al., 1982; von Heijne & Blomberg, 1979; Zviling et al., 2005). All synthesized scales taken together show a rather broad distribution around zero, with few more prominent exceptions, such as valine (V) or tryptophan (W). Interestingly, the three best synthesized scales also seem to agree quite well on most of the amino acids' hydrophobicities, which are, however, for many examples different from the best literature scales (Best 25 individual scales shown in Appendix E, Supplementary Material Table S7.2). The largest difference can be seen in the arginine (R) hydrophobicity values. The literature scales' low arginine value, indicating lowest hydrophobicity, make them exceptional with respect to worse-performing literature scales, suggesting an important role of arginine for VLP assembly and solubility (Vormittag et al., 2020). This is not confirmed with the synthesized amino acid scales. This being said,

it is also not contradicted. To be able to interpret what the synthesized amino acid scale tables mean, one has to consider the mechanism behind the optimization algorithm. In the algorithm, the scales are analyzed for misclassification, and the resulting feature values of misclassified observations. On the basis of the amino acid frequency distribution within the classification groups, the hydrophobicity scale is optimized, thus fitting the scale to the training data of 384 observations through the MC-CV-based procedure. Therefore, the synthesized scales can be regarded as hydrophobicity scales that describe the cVLP solubility problem well. Their application to other molecules or biophysical data would yet have to be probed. (A small case study regarding other biophysical data is shown below.) The discrepancy between the hydrophobicity values, for example for arginine, is probably due to the dominance of other amino acids with respect to their influence during the optimization procedure. This underpins the usefulness of approaching the solubility problem both from a physicochemical and statistical perspective. Tryptophan plays a very important role being one of the most hydrophobic amino acids in the synthesized scales, while its hydrophobicity is less pronounced for literature scales. Its high hydrophobicity value contributes to insoluble classification. In accordance with this finding, amino acids with large side chains, such as tryptophan, have been described to be problematic for HBcAg cVLP assembly (Karpenko et al., 2000).

Methionine (M) and histidine (H) show low hydrophobicity values in the best three synthesized scales, but have a median hydrophobicity close to zero considering all scales. A one-level decision tree based on histidine content was constructed on the entire dataset and showed a low MCC of 0.17 (data not shown), indicating that its low hydrophobicity might be an artifact of the random scale initiation along with its irrelevance to classify the observations. A decision tree on methionine resulted in an MCC of 0.41. However, observations with large methionine content would be rather classified insoluble with this decision tree. This speaks

for a high hydrophobicity, as opposed to what can be seen for the three best synthesized scales.

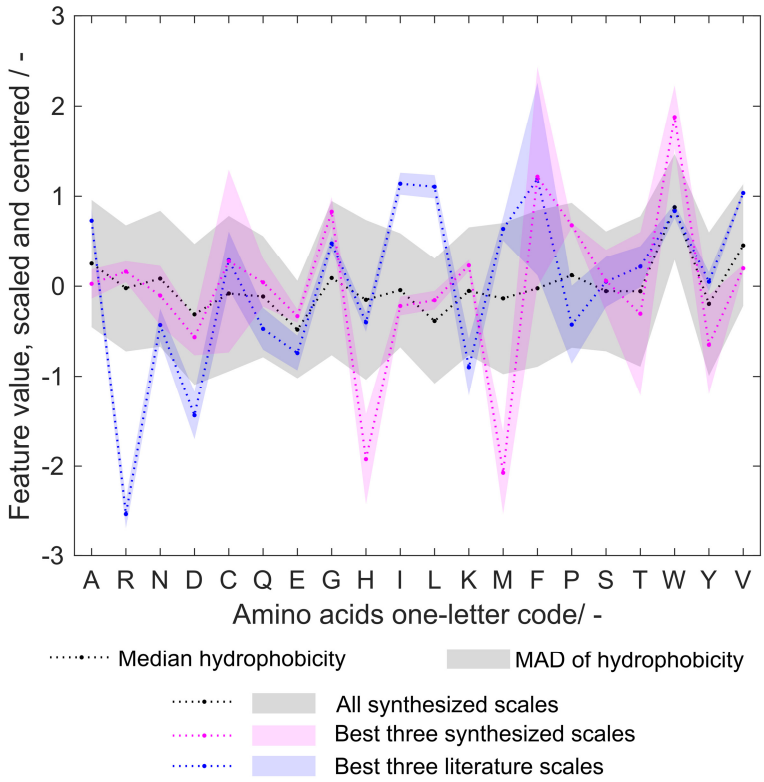


Figure 7.9: Median and median absolute deviation (MAD) of amino acid hydrophobicity for all synthesized, the three best synthesized, and the three best literature scales. The hydrophobicity scales are centered and scaled to unit variance. For comparison purposes, the sign of hydrophobicity scales was changed so that tryptophan (W) hydrophobicity was always positive. Amino acids are represented with one-letter code. The MAD is visualized by a shaded area.

In summary, model performance was significantly enhanced by the synthesis of scales. The above- described cases yet underline the potential to further optimize the procedure for scale synthesis. However, when scales are increasingly optimized, it is important to bear in mind the danger of overfitting.

7.3.6 Redesigning the Soft Ensemble Vote Classifier for Estimation of Ammonium Sulfate Concentrations for VLP Precipitation

Apart from cVLP solubility, there is a variety of other biophysical properties that are interesting with regard to cVLP processing. In a previous study, we investigated precipitation and redissolution of a cVLP candidate (Hillebrandt et al., 2020). In this work, ammonium sulfate

concentration to precipitate the cVLP is determined in a screening experiment before running the process. The screening method to determine optimal ammonium sulfate concentrations for precipitation of the cVLPs was applied to ten cVLPs, all constructed with insertion strategy H, contained in the present dataset. As an example model, synthesized scales from eight training subsets were fitted to solubility data of all observations with insertion strategy H. Synthesized scales were used instead of literature scales, as these were generated based on the model space of interest. Eight models were created including 1-8 of the scales sorted by feature importance. Instead of discretizing the prediction of the models, their continuous value was retrieved. Thus, the individual classifiers become regression models. However, we will still call them 'classifiers' in this section for consistency. In principle, this continuous prediction value should be positive for all constructs as they had to be soluble to be investigated experimentally for precipitation behavior. The rationale behind using the continuous value is that constructs for which the classifier is uncertain have biophysical properties that are actually close to insolubility and therefore probably easier to precipitate.

The ammonium sulfate concentration required to precipitate the investigated ten constructs was mostly between 0.5 M and 0.7 M, except one concentration with 0.1 M and another concentration with 1 M ammonium sulfate (SDS PAGE scans not shown). Linear regression with an sEVC based on scales from set $S_{8,1}$ including all eight synthesized scales resulted in an ordinary R^2 of 0.69. This indicates a linear correlation between the continuous solubility prediction and the ammonium sulfate concentration required for precipitation (Figure 7.10). Confidence bounds are wider at the edge data points of 0.1 M and 1 M ammonium sulfate. This is due to a higher data density in the middle region. The linear fit almost crosses the y-axis at 0 M ammonium sulfate concentration, which, as discussed above, reflects a behavior of this model that would be expected. The construct with lowest continuous solubility prediction value precipitates at low ammonium sulfate concentrations of only 0.1 M. Interestingly, it would be classified as insoluble by the algorithm, while in fact being a soluble construct. Its closeness to the solubility classification border is probably the reason for the low associated precipitating ammonium sulfate concentration.

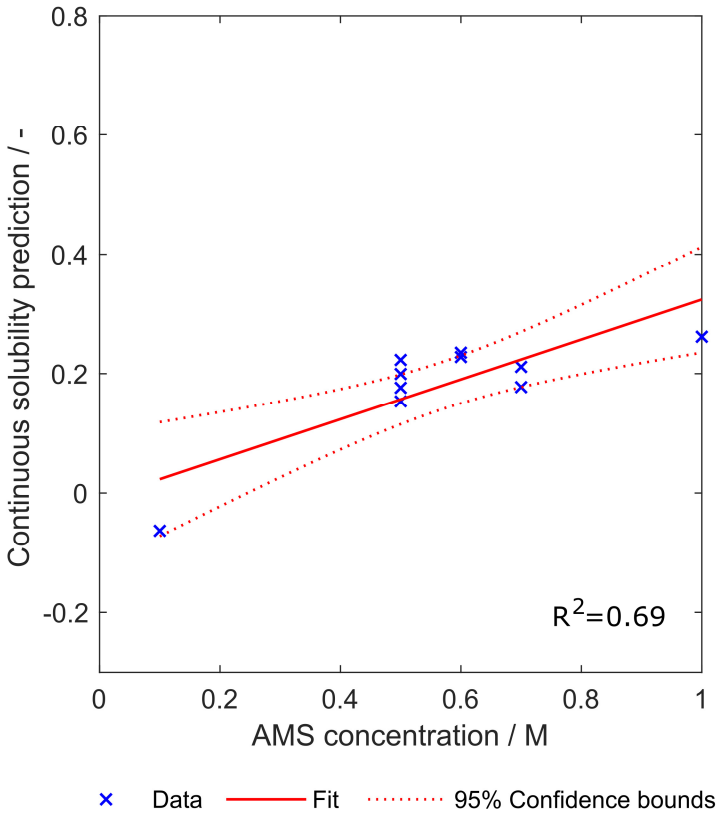


Figure 7.10: Relationship between the continuous solubility prediction value and optimal ammonium sulfate concentration for precipitation of ten constructs. Eight scales were used, which were generated with a scale table optimization procedure (Set $S_{8,1}$). Goodness of fit is indicated by 95% confidence bounds and R^2.

It is important to note that the dataset of ammonium sulfate concentrations is comparably small. This regression study therefore serves as a proof-of-concept, demanding a larger dataset for confirmation of the results and for refinement of the method. With the limited amount of data available, it cannot be deduced which number of included classifiers is optimal. While for this set of scales $S_{8,1}$, it seems that increasing classifier numbers boost regression performance (see also Appendix E, Supplementary Material Figure S7.3), the use of other scales shows inverse trends, where using the first (and according to feature selection best) classifier results in the best R^2, e.g. set $S_{9,1}$ (data not shown). This indicates that regression for the estimation of required ammonium sulfate concentration for precipitation of cVLPs would benefit from a validation procedure, realizable with larger datasets. Additionally, the relationship between the continuous prediction value and ammonium sulfate concentration was assumed to be linear, due to the limited data available. However, this might also be inappropriate,

which again could be answered with a larger dataset. Not all 16x20 scale tables have been tested, since it was deemed inappropriate given the small dataset. Finding the right set of scales by testing all 320 scale table sets for 10 experimental data points can quickly lead to overfitting. The first scale table of the set with eight scales has been chosen, as it represents an average number of generated scales. From some additional tests with other scale tables, it might be assumed that a small number of generated scales perform worse than a greater number (data not shown), which would have to be confirmed with a larger dataset of ammonium sulfate concentration data.

7.4 Conclusion and Outlook

In this study, we have developed and evaluated two different optimization algorithms to improve the performance of an sEVC for the prediction of cVLP solubility based on amino acid sequences and hydrophobicity scale tables. The dataset in this study consisted of 568 chimeric HBcAg constructs, created by insertion of 71 different foreign peptide sequences using 8 different insertion strategies. The sEVC algorithm was originally developed to classify based on 91 literature hydrophobicity scales but showed systematic misclassification for some of the insertion strategies. This was tackled by optimizing the prediction specific for these insertion strategies, resulting in a strategy-specific increase in validation accuracy and MCC of up to 12% and 8%, respectively. The second optimization algorithm modified amino acid scale tables and was also used to synthesize 320 different hydrophobicity scale table sets showing an MCC and accuracy of up to .77 and .88, respectively, on the external test set of 184 HBcAg constructs. The presented models are therefore better than other protein solubility models, typically reporting accuracies of about .60 to .80. A combination of both procedures could elevate the prediction performance data of worse-performing synthesized scales to similar levels. Finally, extension of the model to regression of the required ammonium sulfate concentration for precipitation of ten cVLPs was evaluated, and the linear correlation showed a promising R^2 of .69. The results of this study encourage to further explore the model for other biophysical parameters and molecules.

Acknowledgements

The authors would like to thank Heidemarie Knieriem for proofreading and Sebastian Andris for inspiring discussions.

Appendix E: Supplementary Material

The Supplementary Material associated with this article contain the following information:

- Insertion strategy-based optimization of optimized literature hydrophobicity scales
- Insertion strategy-based optimization of synthesized hydrophobicity scales
- Ninety-one literature hydrophobicity scales
- Twenty-five best individual synthesized hydrophobicity scales
- Relationship between continuous solubility prediction value and optimal ammonium sulfate concentration for precipitation of ten constructs

8

General Discussion and Conclusion

The goal of this thesis was the development of data-driven approaches to advance process development for biopharmaceutical production of virus-like particles (VLPs). Process analytical technology (PAT) was implemented in process steps for downstream processing (DSP) of VLPs (Chapter 3). A tailored DSP unit operation was developed that integrates three different size-dependent separation techniques (Chapter 5). Prediction of process and product parameters was realized with different models, ranging from amino acid sequence data-based models to three-dimensional (3-D) structural analysis (Chapters 4, 6, and 7). The developed data-driven concepts for process development resulted in efficient and well-controlled processes and well-performing models.

VLP dis- and reassembly are unique VLP-specific process steps. In this thesis, VLP reassembly was realized by increasing the ionic strength and lowering the pH from the disassembly solution by buffer exchange applying diafiltration (DF) on a cross-flow filtration (CFF) unit. Chapter 3 describes the implementation of an on-line measurement loop into a CFF unit to monitor VLP reassembly. The on-line measurement loop comprised an ultraviolet and visible (UV/Vis) absorbance spectrometer and a light scattering photometer. The changes in quaternary structure from hepatitis B core antigen (HBcAg) homodimers to VLPs consisting of 180-240 HBcAg molecules was detected with static and dynamic light scattering. The maximum of the static light scattering signal coincided with the maximum VLP concentration. Therefore, light scattering was a good indicator of the end of the process. This is important, as a degradation phase was observed after reaching the

maximum VLP concentration. For the first time, it was shown that the VLP reassembly process by CFF can be monitored and potentially be controlled by PAT based on spectroscopic methods. Orthogonal information on the assembly process was obtained from second derivative analysis of acquired UV/Vis spectra. Information from the second derivative spectra was used to track the changes in the solvatization of the aromatic side chains of tryptophan and tyrosine. During the reassembly processes, measures derived from second derivative spectroscopy were in accordance with the assembly reaction and light scattering data. Especially the a/b-ratio, the metric describing tyrosine solvatization, resulted in a trend timely correlating with the off-line VLP concentrations. Tyrosine 132 is known to be buried in a hydrophobic pocket upon assembly, explaining the measurement results. To optimize the VLP reassembly by CFF, the transmembrane pressure (TMP) was varied in the processes. It influenced the process time, where a lower TMP led to longer process times, which in turn led to increased aggregation. Another outcome of this study was that different VLPs had different assembly end points, where the VLP with strongest negative zeta potential showed the latest end point. The reasons for this were seen in the larger repulsion that had to be overcome by increasing the ionic strength.

Physicochemical properties, such as the zeta potential, influence process steps, as for example seen in Chapter 3. Experimental determination of these properties requires pure material, which is scarce in early development. The prediction of these physicochemical properties by computational methods is therefore useful. In Chapter 4, a case study is shown, which predicts the zeta potential of the VLPs that were investigated in Chapter 3. The prediction was based on *in silico* extraction of surface charge from chimeric HBcAg dimer 3-D structures. These dimer structures were not previously available and had to be generated. The focus of Chapter 4 is the automated and high-throughput generation of 3-D structures from unknown molecules based on similar template structures. In this case, this was the known structure of a C-terminally truncated HBcAg molecule. After automated structure cleaning and dimer extraction, homology models were created and then energy minimized. The resulting structures were simulated in a 3-step, data-dependent MD simulation that was terminated when a Window of Stability (WoS) of 2 ns was reached. The median of the surface charge

of structure snapshots in the WoS lead to more robust results than extraction of the single last simulation snapshot. The workflow was computationally inexpensive, so that it could be run on an ordinary desktop computer requiring a reasonable computational time of 6.6-37.5 h per chimeric HBcAg dimer. Dimer surface charge and experimental zeta potential showed strong linear dependence, but would benefit from a larger dataset to allow further conclusions about this correlation. Overall, the developed workflow was shown to be robust, computationally efficient, and automated, and therefore required minimal user interaction.

Apart from VLP-specific process steps, purification of VLPs often faces challenges when applying traditional biopharmaceutical DSP methods, which were established for other products, such as for monoclonal antibodies. One of these challenges is the reduced binding capacity in bind-and-elute chromatography compared to traditional biopharmaceuticals, such as monoclonal antibodies. This is mainly due to the VLPs' large size. Their size, however, not only poses challenges but also opportunities. Precipitation and re-dissolution, filtration, and size-exclusion chromatography (SEC) profit from the size difference to typical impurities, resulting in potentially better selectivity. In Chapter 5, the integration of these three technologies was investigated, resulting in high purities, yields and productivities compared to a centrifugation-based reference process. The reason for the increased purity was partly seen in an intense wash step of precipitated material in DF mode. Purities and yields were comparable or superior to literature VLP processes. The hypothesis that VLP processing can benefit from combining different size-selective methods in one unit operation was therefore proven with regard to purity, productivity and yield. Consistent with a data-driven approach to process development, ultraviolet (UV) absorbance of the product-containing permeate stream was measured during re-dissolution. Monitoring of the permeate line allowed to identify product-containing fractions. Subsequently, selective pooling could serve to adjust concentration and purity of the recovered material.

While in the above-mentioned study the solubility was reduced artificially, low solubility of recombinantly expressed proteins is a typical challenge in biopharmaceutical product development. This also applies to chimeric VLP (cVLP) expression in various hosts, such as *Escherichia*

coli or yeast. The determination of a cVLP candidate's solubility is typically empirical and is therefore time-consuming and laborious. In Chapter 6, a machine learning framework was developed to predict cVLP solubility based on amino acid sequences and 91 different hydrophobicity scales. The hydrophobicity scales were used to derive hydrophobicity features using the candidates' amino acid sequence. The model was a soft ensemble classifier (sEVC), which was an ensemble of one-level decision trees, each based on an individual hydrophobicity scale. The sEVC was trained on training sets of different sizes (24-384). Its accuracy and Matthew's correlation coefficient (MCC) were comparable or superior to reported literature solubility model performance data. Stratified sampling and feature selection were beneficial for model construction. Feature selection also proved useful for interpretation of the model, suggesting a special role of arginine for VLP assembly.

Chapter 7 is built on the foundation of Chapter 6, aiming to optimize the model performance, to derive new hydrophobicity scales, and to extend the model to function as a regression tool. The first implemented optimization strategy was based on the systematic misclassification observed in Chapter 6. Certain insertion strategies, defining where in the HBcAg molecule the foreign epitope is inserted and which amino acids are deleted, were systematically over- or underestimated with regard to their solubility. This information was used in an iterative process to modify the sEVC's prediction based on this observation within the training set, thus increasing overall model performance. A second optimization strategy was based on the amino acid frequency of falsely classified constructs and the model's classification boundary. This strategy was used to optimize existing scales and to synthesize novel hydrophobicity scales from random vectors. The best model was created with synthesized hydrophobicity scales, resulting in an MCC of .77 (accuracy of .88). Since hydrophobic interaction drives precipitation, the concept of the sEVC was redesigned to be applied as a regression tool for the prediction of ammonium sulfate concentrations required for VLP precipitation. Initial tests with ten experimental data points were promising, showing an R^2 of .69 for the correlation of experimental data with the prediction of the regression tool. The concept of the developed sEVC based on amino acid sequences and hydrophobicity scales therefore seems promising for the prediction of other hydrophobicity-related

biophysical properties and may be applied to other cVLPs or even different classes of molecules.

In conclusion, this thesis presents an array of data-driven approaches, including machine learning, 3-D structure generation, and PAT, which will aid in VLP molecular and process design and to efficiently control VLP processes. Additionally, it presents advanced methods for capture, purification, and reassembly of VLPs, which are competitive or superior to literature processes. This thesis is therefore a contribution to a potential platform process, implementing state-of-the-art data-driven methods.

9

Outlook

The potential of virus-like particles (VLPs) has been shown in various pre-clinical and clinical studies. Their application ranges from immunotherapy against Malaria, Alzheimer's disease, and cancer, to the utilization of VLPs as cargo delivery vehicles for proteins and nucleic acids. One of the major bottlenecks for VLP development is the provision of VLP material for these studies, which is often realized in inefficient lab-scale processes. The absence of VLPs in big pharma – apart from few products – is one reason contributing to this issue. The breakthrough of monoclonal antibodies (mAbs) on the pharmaceutical market has been reached through successful initial products, trust in the safety and efficacy of antibodies, and drastic reduction in development and production costs (Vertès & Dowden, 2015). The reduction in development and production costs is largely due to the establishment of a platform process (Kelley, 2009). This platform process enabled researchers to develop and produce mAbs in sufficient quantities for *in vitro* and animal testing reducing the effort in the pre-clinical development phase. The comparably less advanced field of VLPs will need this platform process to achieve a higher degree of success on the pharmaceutical market.

VLP processing often still relies on techniques that are either tailored to other molecules, such as mAbs, or not easily scalable, such as ultracentrifugation. The future of VLP processing will therefore see an increase in the application of methods that are tailored to VLP processing. Due to their large size, filtration, precipitation and re-dissolution, and size exclusion chromatography are promising methods

for the purification of VLPs. The reason behind this is that the large size is a property which is common for all VLPs and which discriminates them from most impurities, leading to good selectivities. Therefore, these methods will probably be increasingly used in VLP processes and may lay the foundation for a VLP platform process. Application of the methods presented in this thesis to other VLPs would help evaluate their potential as components of a platform process.

Consistent with the regulatory authorities' requirement for quality by design – i.e. building the quality into the processes instead of testing it into the product – the process of establishing a platform process should be accompanied by the implementation of state-of-the-art process analytical technology (PAT). The establishment of sophisticated PAT tools for VLP-specific process steps, such as presented in this thesis for VLP reassembly, is an important starting point in this direction. However, the application of simple PAT tools, such as monitoring of UV absorbance at a single wavelength, should not be disregarded, as they provide a simple means for process monitoring and control. When processes, such as the VLP reassembly, require prompt action (e.g. at the end of the assembly reaction), at-line PAT methods, such as high-performance liquid chromatography, are inadequate to fulfill the requirements to a data-based process control tool. Therefore, (near) real-time spectroscopic on-line or in-line techniques will be predominantly applied to processes, where a quick decision has to be made. Hereby, light scattering, second derivative analysis of ultraviolet and visible (UV/Vis) spectra, and partial least squares modeling have the potential to be applied to other VLPs than examined in this thesis and to different process steps, such as disassembly.

Protein engineering is seeing the advent of machine learning and other data scientific methods. When vast libraries of physicochemical data, structural data, and process data become available, the importance of these data scientific methods will grow. The application can be quite diverse and range from predictive models, to soft sensors, to data analytical applications. Soft sensors are based on the fusion of advanced data analytical techniques and sensor data and have the potential to efficiently steer processes in real-time. Predictive models can point out promising VLP candidates early on or reduce the design space of processes, such as the ionic strength in VLP reassembly. The machine

learning framework developed in this thesis could be applied to predict different process-relevant properties, such as solution viscosity or retention times in chromatography. Statistical analysis of process data within a machine learning framework or with other techniques, such as simple cluster algorithms will aid in understanding the system better and provide a basis to make decisions about molecular design and the design of better processes.

The application of VLPs as cargo delivery vehicles will be evaluated in detail with the onset of the gene therapy era. For a breakthrough, gene therapy would benefit from a reproducible formulation step. This includes packaging of the nucleic acid material, which is susceptible to degradation. VLPs – although some VLPs are too small for the packaging of large nucleic acids – seem promising vehicles, as they are better understood than lipid based nanoparticles, would act close to their actual function (of delivering viral nucleic acids), are stable upon dilution, and their surface can be modified for targeted delivery (Rohovie, Nagasawa, & Swartz, 2017; Sandra, Khaliq, Sunna, & Care, 2019). The realization of their potential will rely on the development of tailored downstream processing unit operations which include monitoring techniques to speed up development and increase process understanding.

In summary, the following years will see an increasing diversification of the application of VLPs due to their great versatility. Selection and combination of adequate computational and physical process tools will greatly advance the field of VLPs in the biopharmaceutical industry, especially by paving the way to a platform process that may form the basis for a swift, flexible, economic, and efficient VLP process development.

Bibliography

Agostini, F., Vendruscolo, M., & Tartaglia, G. G. (2012). Sequence-Based Prediction of Protein Solubility. *Journal of Molecular Biology*, *421*(2–3), 237–241. https://doi.org/10.1016/j.jmb.2011.12.005

Alexander, C. G., Jurgens, M. C., Shepherd, D. A., Freund, S. M. V, Ashcroft, A. E., & Ferguson, N. (2013). Thermodynamic origins of protein folding, allostery, and capsid formation in the human hepatitis B virus core protein. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(30), E2782–E2791. https://doi.org/10.1073/pnas.1308846110

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

Andersen, R. (2008). *Modern Methods for Robust Regression*. (V. Knight, Ed.), *Modern methods for robust regression* (152nd ed.). Thousand Oaks, US: SAGE Publications, Inc. https://doi.org/10.4135/9781412985109

Andris, S., Rüdt, M., Rogalla, J., Wendeler, M., & Hubbuch, J. (2018). Monitoring of antibody-drug conjugation reactions with UV/Vis spectroscopy. *Journal of Biotechnology*, *288*, 15–22. https://doi.org/10.1016/j.jbiotec.2018.10.003

Arakawa, T., Ejima, D., Tsumoto, K., Obeyama, N., Tanaka, Y., Kita, Y., & Timasheff, S. N. (2007). Suppression of protein interactions by arginine: A proposed mechanism of the arginine effects. *Biophysical Chemistry*. https://doi.org/10.1016/j.bpc.2006.12.007

Arakawa, T., & Timasheff, S. N. (1982). Preferential interactions of proteins with salts in concentrated solutions. *Biochemistry*, *21*(25), 6545–6552. https://doi.org/10.1021/bi00268a034

Arakawa, T., & Timasheff, S. N. (1985). Mechanism of polyethylene glycol interaction with proteins. *Biochemistry*, *24*(24), 6756–6762. https://doi.org/10.1021/bi00345a005

Arkhipov, A., Freddolino, P. L., & Schulten, K. (2006). Stability and Dynamics of Virus Capsids Described by Coarse-Grained Modeling. *Structure*, *14*(12), 1767–1777.

https://doi.org/10.1016/j.str.2006.10.003

Arora, U., Tyagi, P., Swaminathan, S., & Khanna, N. (2012). Chimeric Hepatitis B core antigen virus-like particles displaying the envelope domain III of dengue virus type 2. *Journal of Nanobiotechnology*, *10*(1), 30. https://doi.org/10.1186/1477-3155-10-30

Ashcroft, A. E., Lago, H., Macedo, J. M. B., Horn, W. T., Stonehouse, N. J., & Stockley, P. G. (2005). Engineering thermal stability in RNA phage capsids via disulphide bonds. *Journal of Nanoscience and Nanotechnology*, *5*(12), 2034–2041. https://doi.org/10.1166/jnn.2005.507

Atha, D. H., & Ingham, K. C. (1981). Mechanism of precipitation of proteins by polyethylene glycols. Analysis in terms of excluded volume. *Journal of Biological Chemistry*, *256*(23), 12108–12117.

Ausar, S. F., Foubert, T. R., Hudson, M. H., Vedvick, T. S., & Middaugh, C. R. (2006). Conformational Stability and Disassembly of Norwalk Virus-like Particles: Effect of pH and Temperature. *Journal of Biological Chemistry*, *281*(28), 19478–19488. https://doi.org/10.1074/jbc.M603313200

Austerberry, J. I., Thistlethwaite, A., Fisher, K., Golovanov, A. P., Pluen, A., Esfandiary, R., ... Curtis, R. (2019). Arginine to Lysine Mutations Increase the Aggregation Stability of a Single-Chain Variable Fragment through Unfolded-State Interactions. *Biochemistry*, *58*(32), 3413–3421. https://doi.org/10.1021/acs.biochem.9b00367

Ayton, G. S., & Voth, G. A. (2010). Multiscale computer simulation of the immature HIV-1 virion. *Biophysical Journal*, *99*(9), 2757–2765. https://doi.org/10.1016/j.bpj.2010.08.018

Bacchin, P., Si-Hassen, D., Starov, V., Clifton, M. ., & Aimar, P. (2002). A unifying model for concentration polarization, gel-layer formation and particle deposition in cross-flow membrane filtration of colloidal suspensions. *Chemical Engineering Science*, *57*(1), 77–91. https://doi.org/10.1016/S0009-2509(01)00316-5

Bachmann, M. F., & Whitehead, P. (2013). Active immunotherapy for chronic diseases. *Vaccine*, *31*(14), 1777–1784. https://doi.org/10.1016/j.vaccine.2013.02.001

Bakeev, K. A. (Ed.). (2010). *Process analytical technology: Spectroscopic tools and implementation strategies for the chemical and pharmaceutical industries* (2nd Editio). Chichester, GB-WSX: John

Wiley & Sons.

Bakker, W. A. M., Thomassen, Y. E., & van der Pol, L. A. (2010). Scale-Down Approach for Animal-Free Polio Vaccine Production. In T. Noll (Ed.), *Cells and Culture. ESACT Proceedings* (vol 4, pp. 541–550). Dordrecht: Springer.

Baylor, N. W. (2016). The Regulatory Evaluation of Vaccines for Human Use. In S. Thomas (Ed.), *Vaccine Design: Methods and Protocols, Volume 2: Vaccines for Veterinary Diseases* (pp. 773–787). New York, NY: Springer New York. https://doi.org/10.1007/978-1-4939-3389-1_51

Beauchamp, K. A., Lin, Y. S., Das, R., & Pande, V. S. (2012). Are protein force fields getting better? A systematic benchmark on 524 diverse NMR measurements. *Journal of Chemical Theory and Computation*, *8*(4), 1409–1414. https://doi.org/10.1021/ct2007814

Ben-Naim, A. (1988). Theory of preferential solvation of nonelectrolytes. *Cell Biophysics*, *12*(1), 255–269. https://doi.org/10.1007/BF02918361

Berendsen, H. J. C., Postma, J. P. M., Gunsteren, W. F. Van, Dinola, A., & Haak, J. R. (1984). Molecular dynamics with coupling to an external bath Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, *81*(8), 3684–3690. https://doi.org/10.1063/1.448118

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., ... Bourne, P. E. (2000). The Protein Data Bank, *28*(1), 235–242.

Berne, B. J., & Pecora, R. (2000). *Dynamic light scattering: with applications to chemistry, biology, and physics*. Courier Corporation.

Besnard, L., Fabre, V., Fettig, M., Gousseinov, E., Kawakami, Y., Laroudie, N., ... Pattnaik, P. (2016). Clarification of vaccines: An overview of filter based technology trends and best practices. *Biotechnology Advances*, *34*(1), 1–13. https://doi.org/10.1016/j.biotechadv.2015.11.005

Best, R. B., Buchete, N. V., & Hummer, G. (2008). Are current molecular dynamics force fields too helical? *Biophysical Journal*, *95*(1), 7–9. https://doi.org/10.1529/biophysj.108.132696

Billaud, J.-N., Peterson, D., Barr, M., Chen, A., Sallberg, M., Garduno, F., ... Milich, D. (2005). Combinatorial Approach to Hepadnavirus-Like Particle Vaccine Design. *Journal of Virology*, *79*(21), 13656–

13666. https://doi.org/10.1128/JVI.79.21.13656-13666.2005

Bohren, C. F., & Huffman, D. R. (2004). *Absorption and scattering of light by small particles*. Wiley-VCH.

Bolli, E., O'Rourke, J. P., Conti, L., Lanzardo, S., Rolih, V., Christen, J. M., ... Cavallo, F. (2018). A Virus-Like-Particle immunotherapy targeting Epitope-Specific anti-xCT expressed on cancer stem cell inhibits the progression of metastatic cancer in vivo. *OncoImmunology*, *7*(3), e1408746. https://doi.org/10.1080/2162402X.2017.1408746

Böttcher, B., Wynne, S. A., & Crowther, R. A. (1997). Determination of the fold of the core protein of hepatitis B virus by electron cryomicroscopy. *Nature*, *386*(6620), 88–91. https://doi.org/10.1038/386088a0

Bourne, C. R., Katen, S. P., Fulz, M. R., Packianathan, C., & Zlotnick, A. (2009). A Mutant Hepatitis B Virus Core Protein Mimics Inhibitors of Icosahedral Capsid Self-Assembly †. *Biochemistry*, *48*(8), 1736–1742. https://doi.org/10.1021/bi801814y

Breiman, L. (1998). Arcing Classifiers. *The Annals of Statistics*, *26*(3), 801–824. Retrieved from http://www.jstor.org/stable/120055

Brestrich, N., Rüdt, M., Büchler, D., & Hubbuch, J. (2018). Selective protein quantification for preparative chromatography using variable pathlength UV/Vis spectroscopy and partial least squares regression. *Chemical Engineering Science*, *176*, 157–164. https://doi.org/10.1016/j.ces.2017.10.030

Brune, K. D., Leneghan, D. B., Brian, I. J., Ishizuka, A. S., Bachmann, M. F., Draper, S. J., ... Howarth, M. (2016). Plug-and-Display: decoration of Virus-Like Particles via isopeptide bonds for modular immunization. *Scientific Reports*, *6*, 1–13. https://doi.org/10.1038/srep19234

Bryan, J. T., Buckland, B., Hammond, J., & Jansen, K. U. (2016). Prevention of cervical cancer: Journey to develop the first human papillomavirus virus-like particle vaccine and the next generation vaccine. *Current Opinion in Chemical Biology*, *32*, 34–47. https://doi.org/10.1016/j.cbpa.2016.03.001

Buck, K. K. S., Subramanian, V., & Block, D. E. (2002). Identification of Critical Batch Operating Parameters in Fed-Batch Recombinant E. coli Fermentations Using Decision Tree Analysis. *Biotechnology Progress*, *18*(6), 1366–1376. https://doi.org/10.1021/bp020112p

Buckland, B. C. (2005). The process development challenge for a new vaccine. *Nature Medicine*, *11*(4S), S16. https://doi.org/10.1038/nm1218

Buffin, S., Peubez, I., Barrière, F., Nicolaï, M.-C., Tapia, T., Dhir, V., ... Legastelois, I. (2019). Influenza A and B virus-like particles produced in mammalian cells are highly immunogenic and induce functional antibodies. *Vaccine*, *37*(46), 6857–6867. https://doi.org/10.1016/j.vaccine.2019.09.057

Buonaguro, L., Aurisicchio, L., Buonaguro, F. M., & Ciliberto, G. (2013). New developments in cancer vaccines. *Expert Review of Vaccines*, *12*(10), 1109–1110. https://doi.org/10.1586/17476348.2013.838013

Burden, C. S., Jin, J., Podgornik, A., & Bracewell, D. G. (2012). A monolith purification process for virus-like particles from yeast homogenate. *Journal of Chromatography B*, *880*, 82–89. https://doi.org/10.1016/j.jchromb.2011.10.044

Bustos-Jaimes, I., Soto-Román, R. A., Gutiérrez-Landa, I. A., Valadez-García, J., & Segovia-Trinidad, C. L. (2017). Construction of protein-functionalized virus-like particles of parvovirus B19. *Journal of Biotechnology*, *263*(May), 55–63. https://doi.org/10.1016/j.jbiotec.2017.09.014

Carrio, M. del M., & Villaverde, A. (2005). Localization of Chaperones DnaK and GroEL in Bacterial Inclusion Bodies. *Journal of Bacteriology*, *187*(10), 3599–3601. https://doi.org/10.1128/JB.187.10.3599-3601.2005

Carvalho, S. B., Freire, J. M., Moleirinho, M. G., Monteiro, F., Gaspar, D., Castanho, M. A. R. B., ... Peixoto, C. (2016). Bioorthogonal Strategy for Bioprocessing of Specific-Site-Functionalized Enveloped Influenza-Virus-Like Particles. *Bioconjugate Chemistry*, *27*(10), 2386–2399. https://doi.org/10.1021/acs.bioconjchem.6b00372

Carvalho, S. B., Silva, R. J. S., Moleirinho, M. G., Cunha, B., Moreira, A. S., Xenopoulos, A., ... Peixoto, C. (2019). Membrane-Based Approach for the Downstream Processing of Influenza Virus-Like Particles. *Biotechnology Journal*, *14*(8), 1–12. https://doi.org/10.1002/biot.201800570

Carvalho, S. B., Silva, R. J. S., Moreira, A. S., Cunha, B., Clemente, J. J., Alves, P. M., ... Peixoto, C. (2019). Efficient filtration strategies for the clarification of influenza virus-like particles derived from insect cells. *Separation and Purification Technology*, *218*, 81–88. https://doi.org/10.1016/j.seppur.2019.02.040

Castiblanco, J., & Anaya, J.-M. (2015). Genetics and vaccines in the era of personalized medicine. *Current Genomics*, *16*(1), 47–59. https://doi.org/10.2174/1389202916666141223220551

Centers for Disease Control and Prevention. (1999). Impact of vaccines universally recommended for children--United States, 1990-1998. *MMWR. Morbidity and Mortality Weekly Report*, *48*(12), 243–248.

Ceres, P., & Zlotnick, A. (2002). Weak Protein– Protein Interactions Are Sufficient To Drive Assembly of Hepatitis B Virus Capsids. *Biochemistry*, *41*(39), 11525–11531. https://doi.org/10.1021/bi0261645

Chackerian, B. (2007). Virus-like particles: flexible platforms for vaccine development. *Expert Review of Vaccines*, *6*(3), 381–390. https://doi.org/10.1586/14760584.6.3.381

Chan, J.-A., Wetzel, D., Reiling, L., Miura, K., Drew, D. R., Gilson, P. R., ... Beeson, J. G. (2019). Malaria vaccine candidates displayed on novel virus-like particles are immunogenic and induce transmission-blocking activity. *PloS One*, *14*(9), e0221733–e0221733. https://doi.org/10.1371/journal.pone.0221733

Charaniya, S., Hu, W.-S., & Karypis, G. (2008). Mining bioprocess data: opportunities and challenges. *Trends in Biotechnology*, *26*(12), 690–699. https://doi.org/10.1016/j.tibtech.2008.09.003

Cheluvaraja, S., & Ortoleva, P. (2010). Thermal nanostructure: An order parameter multiscale ensemble approach. *Journal of Chemical Physics*, *132*(7), 1–9. https://doi.org/10.1063/1.3316793

Chen, Q., Abdul Latiff, S. M., Toh, P., Peng, X., Hoi, A., Xian, M., ... Gagnon, P. (2016). A simple and efficient purification platform for monoclonal antibody production based on chromatin-directed cell culture clarification integrated with precipitation and void-exclusion anion exchange chromatography. *Journal of Biotechnology*, *236*, 128–140. https://doi.org/10.1016/j.jbiotec.2016.08.014

Chennamsetty, N., Voynov, V., Kayser, V., Helk, B., & Trout, B. L. (2009). Design of therapeutic proteins with enhanced stability. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(29), 11937–11942. https://doi.org/10.1073/pnas.0904191106

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), 1–13.

https://doi.org/10.1186/s12864-019-6413-7

Choi, Y., Gyoo Park, S., Yoo, J. H., & Jung, G. (2005). Calcium ions affect the hepatitis B virus core assembly. *Virology*, *332*(1), 454–463. https://doi.org/10.1016/j.virol.2004.11.019

Chothia, C. (1976). The nature of the accessible and buried surfaces in proteins. *Journal of Molecular Biology*, *105*(1), 1–12. https://doi.org/10.1016/0022-2836(76)90191-1

Chuan, Y. P., Wibowo, N., Lua, L. H. L., & Middelberg, A. P. J. (2014). The economics of virus-like particle and capsomere vaccines. *Biochemical Engineering Journal*, *90*, 255–263. https://doi.org/10.1016/j.bej.2014.06.005

Chung, C.-Y., Chen, C.-Y., Lin, S.-Y., Chung, Y.-C., Chiu, H.-Y., Chi, W.-K., ... Hu, Y.-C. (2010). Enterovirus 71 virus-like particle vaccine: Improved production conditions for enhanced yield. *Vaccine*, *28*(43), 6951–6957. https://doi.org/10.1016/j.vaccine.2010.08.052

Cohn, E. J. (1941). The Properties and Functions of the Plasma Proteins, with a Consideration of the Methods for their Separation and Purification. *Chemical Reviews*, *28*(2), 395–417. https://doi.org/10.1021/cr60090a007

Cohn, E. J., Strong, L. E., Hughes, W. L., Mulford, D. J., Ashworth, J. N., Melin, M., & Taylor, H. L. (1946). Preparation and Properties of Serum and Plasma Proteins. IV. A System for the Separation into Fractions of the Protein and Lipoprotein Components of Biological Tissues and Fluids 1a,b,c,d. *Journal of the American Chemical Society*, *68*(3), 459–475. https://doi.org/10.1021/ja01207a034

Conchillo-Sole, O., de Groot, N. S., Aviles, F. X., Vendrell, J., Daura, X., & Ventura, S. (2007). AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. *BMC Bioinformatics*, *8*, 65. https://doi.org/10.1186/1471-2105-8-65

Cook III, J. C. (2003). *US Patent 6,602,697*. US: Google Patents.

Cook, J. C., Joyce, J. G., George, H. a, Schultz, L. D., Hurni, W. M., Jansen, K. U., ... Lehman, E. D. (1999). Purification of virus-like particles of recombinant human papillomavirus type 11 major capsid protein L1 from Saccharomyces cerevisiae. *Protein Expression and Purification*, *17*(3), 477–484. https://doi.org/10.1006/prep.1999.1155

Corbett, J. C. W., Connah, M., & Mattison, K. (2017). *US Patent*

2017/0269030 A1.

Crisci, E., Bárcena, J., & Montoya, M. (2013). Virus-like particle-based vaccines for animal viral infections. *Inmunología*, *32*(3), 102–116. https://doi.org/10.1016/j.inmuno.2012.08.002

Crowther, R. A., Kiselev, N. A., Böttcher, B., Berriman, J. A., Borisova, G. P., Ose, V., & Pumpens, P. (1994). Three-dimensional structure of hepatitis B virus core particles determined by electron cryomicroscopy. *Cell*, *77*(6), 943–950. https://doi.org/10.1016/0092-8674(94)90142-2

Curtis, R. A., Montaser, A., Prausnitz, J. M., & Blanch, H. W. (1998). Protein-protein and protein-salt interactions in aqueous protein solutions containing concentrated electrolytes. *Biotechnology and Bioengineering*, *58*(4), 451–451. https://doi.org/10.1002/(sici)1097-0290(19980520)58:4<451::aid-bit13>3.3.co;2-6

Dai, S., Wang, H., & Deng, F. (2018). Advances and challenges in enveloped virus-like particle (VLP)-based vaccines. *Journal of Immunological Sciences*, *2*(2), 36–41. https://doi.org/10.29245/2578-3009/2018/2.1118

Davies, J. L., & Smith, M. P. (2010). Membrane Applications in Monoclonal Antibody Production. In Z. F. Cui & H. S. Muralidhara (Eds.), *Membrane Technology* (pp. 79–120). Oxford: Butterworth-Heinemann. https://doi.org/10.1016/B978-1-85617-632-3.00006-9

de Jong, S. (1993). SIMPLS: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laborary Systems*, *18*(3), 251–263. https://doi.org/10.1016/0169-7439(93)85002-X

Deep, K., Singh, K. P., Kansal, M. L., & Mohan, C. (2009). A real coded genetic algorithm for solving integer and mixed integer optimization problems. *Applied Mathematics and Computation*, *212*(2), 505–518. https://doi.org/10.1016/j.amc.2009.02.044

Ding, F.-X., Wang, F., Lu, Y.-M., Li, K., Wang, K.-H., He, X.-W., & Sun, S.-H. (2009). Multiepitope peptide-loaded virus-like particles as a vaccine against hepatitis B virus–related hepatocellular carcinoma. *Hepatology*, *49*(5), 1492–1502. https://doi.org/10.1002/hep.22816

Ding, X., Liu, D., Booth, G., Gao, W., & Lu, Y. (2018). Virus-Like Particle Engineering: From Rational Design to Versatile Applications. *Biotechnology Journal*, *13*(5), 1–7. https://doi.org/10.1002/biot.201700324

Ding, Y., Chuan, Y. P., He, L., & Middelberg, A. P. J. (2010). Modeling

the competition between aggregation and self-assembly during virus-like particle processing. *Biotechnology and Bioengineering*, *107*(3), 550–560. https://doi.org/10.1002/bit.22821

Duan, Y., Wu, C., Chowdhury, S., Lee, M. C., Xiong, G., Zhang, W., ... Kollman, P. (2003). A Point-Charge Force Field for Molecular Mechanics Simulations of Proteins Based on Condensed-Phase Quantum Mechanical Calculations. *Journal of Computational Chemistry*, *24*(16), 1999–2012. https://doi.org/10.1002/jcc.10349

Eisenberg, D., Weiss, R. M., Terwilliger, T. C., & Wilcox, W. (1982). Hydrophobic moments and protein structure. *Faraday Symposia of the Chemical Society*, *17*, 109. https://doi.org/10.1039/fs9821700109

Engelman, D. M., Steitz, T. A., & Goldman, A. (1986). Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annual Review of Biophysics and Biophysical Chemistry*, *15*(1), 321–353.

Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., Pedersen, L. G., ... Pedersen, L. G. (1995). A smooth particle mesh Ewald method A smooth particle mesh Ewald method. *The Journal of Chemical Physics*, *103*(19), 8577–8593. https://doi.org/10.1063/1.470117

Fan, H. A. O., & Mark, A. E. (2004). Refinement of homology-based protein structures by molecular dynamics simulation techniques. *Protein Science*, *13*, 211–220. https://doi.org/10.1110/ps.03381404.normally

Fang, M., Diao, W., Dong, B., Wei, H., Liu, J., Hua, L., ... Wan, M. (2016). Detection of the Assembly and Disassembly of PCV2b Virus-Like Particles Using Fluorescence Spectroscopy Analysis. *Intervirology*, *58*(5), 318–323. https://doi.org/10.1159/000442751

Fang, Y., Gao, S., Tai, D., Middaugh, C. R., & Fang, J. (2013). Identification of properties important to protein aggregation using feature selection. *BMC Bioinformatics*, *14*(1), 314. https://doi.org/10.1186/1471-2105-14-314

FDA, & Others. (2004). Guidance for Industry: PAT—a framework for innovative pharmaceutical development, manufacturing, and quality assurance. *Rockville, MD.* Retrieved from https://www.fda.gov/media/71012/download

Fehr, T., Skrastina, D., Pumpens, P., & Zinkernagel, R. M. (1998). T cell-independent type I antibody response against B cell epitopes

expressed repetitively on recombinant virus particles. *Proc. Natl. Acad. Sci. U. S. A.*, *95*(16), 9477–9481. https://doi.org/10.1073/pnas.95.16.9477

Fenner, F., Henderson, D. A., Arita, I., Jezek, Z., Ladnyi, I. D., & Organization, W. H. (1988). *Smallpox and its eradication*. Geneva: World Health Organization.

Fernandez-Cerezo, L., Wismer, M. K., Han, I., & Pollard, J. M. (2019). High throughput screening of ultrafiltration and diafiltration processing of monoclonal antibodies via the ambr(R) crossflow system. *Biotechnology Progress*, e2929. https://doi.org/10.1002/btpr.2929

Fifis, T., Gamvrellis, A., Crimeen-Irwin, B., Pietersz, G. A., Li, J., Mottram, P. L., ... Plebanski, M. (2004). Size-Dependent Immunogenicity: Therapeutic and Protective Properties of Nano-Vaccines against Tumors. *The Journal of Immunology*, *173*(5), 3148–3154. https://doi.org/10.4049/jimmunol.173.5.3148

Fiser, A., & Šali, A. (2003). Modeller: Generation and Refinement of Homology-Based Protein Structure Models. In *Method in Enzymology* (Vol. 374, pp. 461–491). https://doi.org/10.1016/S0076-6879(03)74020-8

Forster, M. J. (2002). Molecular modelling in structural biology. *Micron*, *33*(4), 365–384. https://doi.org/10.1016/S0968-4328(01)00035-X

Freddolino, P. L., Arkhipov, A. S., Larson, S. B., McPherson, A., & Schulten, K. (2006). Molecular dynamics simulations of the complete satellite tobacco mosaic virus. *Structure*, *14*(3), 437–449. https://doi.org/10.1016/j.str.2005.11.014

Frietze, K. M., Peabody, D. S., & Chackerian, B. (2016). Engineering virus-like particles as vaccine platforms. *Current Opinion in Virology*, *18*, 44–49. https://doi.org/10.1016/j.coviro.2016.03.001

Gallagher, J. R., Torian, U., McCraw, D. M., & Harris, A. K. (2017). Characterization of the disassembly and reassembly of the HBV glycoprotein surface antigen, a pliable nanoparticle vaccine platform. *Virology*, *502*, 176–187. https://doi.org/10.1016/j.virol.2016.12.025

Gallina, A., Bonelli, F., Zentilin, L., Rindi, G., Muttini, M., & Milanesi, G. (1989). A recombinant hepatitis B core antigen polypeptide with the protamine-like domain deleted self-assembles into capsid particles but fails to bind nucleic acids. *Journal of Virology*, *63*(11), 4645–4652. Retrieved from

https://jvi.asm.org/content/63/11/4645.long

Galm, L., Amrhein, S., & Hubbuch, J. (2017). Predictive approach for protein aggregation: Correlation of protein surface characteristics and conformational flexibility to protein aggregation propensity. *Biotechnology and Bioengineering*, *114*(6), 1170–1183. https://doi.org/10.1002/bit.25949

Gangadharan, N., Turner, R., Field, R., Oliver, S. G., Slater, N., & Dikicioglu, D. (2019). Metaheuristic approaches in biopharmaceutical process development data analysis. *Bioprocess and Biosystems Engineering*, *42*(9), 1399–1408. https://doi.org/10.1007/s00449-019-02147-0

Garçon, N., Morel, S., Didierlaurent, A., Descamps, D., Wettendorff, M., & Van Mechelen, M. (2011). Development of an AS04-Adjuvanted HPV Vaccine with the Adjuvant System Approach. *BioDrugs*, *25*(4), 217–226. https://doi.org/10.2165/11591760-000000000-00000

Garde, S., & Patel, A. J. (2011). Unraveling the hydrophobic effect, one molecule at a time. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(40), 16491–16492. https://doi.org/10.1073/pnas.1113256108

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D., & Bairoch, A. (2005). Protein Identification and Analysis Tools on the ExPASy Server. In J. M. Walker (Ed.), *The Proteomics Protocols Handbook* (pp. 571–607). Totowa, NJ: Humana Press. https://doi.org/10.1385/1-59259-890-0:571

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural Networks and the Bias/Variance Dilemma. *Neural Computation*, *4*(1), 1–58. https://doi.org/10.1162/neco.1992.4.1.1

Geng, H., Chen, F., Ye, J., & Jiang, F. (2019). Applications of Molecular Dynamics Simulation in Structure Prediction of Peptides and Proteins. *Computational and Structural Biotechnology Journal*, *17*, 1162–1170. https://doi.org/10.1016/j.csbj.2019.07.010

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

Gillam, F., & Zhang, C. (2018). Epitope selection and their placement for increased virus neutralization in a novel vaccination strategy for

porcine epidemic diarrhea virus utilizing the Hepatitis B virus core antigen. *Vaccine*, *36*(30), 4507–4516. https://doi.org/10.1016/j.vaccine.2018.06.015

Gini, C. (1912). Variabilità e mutabilità. *Reprinted in Memorie Di Metodologica Statistica (Ed. Pizetti E, Salvemini, T). Rome: Libreria Eredi Virgilio Veschi.*

Giuliani, M. M., Adu-Bobie, J., Comanducci, M., Aricò, B., Savino, S., Santini, L., ... Pizza, M. (2006). A universal vaccine for serogroup B meningococcus. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(29), 10834–10839. https://doi.org/10.1073/pnas.0603940103

Goldinger, S. M., Dummer, R., Baumgaertner, P., Mihic-Probst, D., Schwarz, K., Hammann-Haenni, A., ... Speiser, D. E. (2012). Nano-particle vaccination combined with TLR-7 and -9 ligands triggers memory and effector CD8+ T-cell responses in melanoma patients. *European Journal of Immunology*, *42*(11), 3049–3061. https://doi.org/10.1002/eji.201142361

Gorbenko, G., & Trusova, V. (2011). *Protein aggregation in a membrane environment. Advances in Protein Chemistry and Structural Biology* (1st ed., Vol. 84). Elsevier Inc. https://doi.org/10.1016/B978-0-12-386483-3.00002-1

Greenwood, B. (2014). The contribution of vaccination to global health: Past, present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1645). https://doi.org/10.1098/rstb.2013.0433

Grgacic, E. V. L., & Anderson, D. A. (2006). Virus-like particles: Passport to immune recognition. *Methods*, *40*(1), 60–65. https://doi.org/10.1016/j.ymeth.2006.07.018

Grigg, O. A., Farewell, V. T., & Spiegelhalter, D. J. (2003). Use of risk-adjusted CUSUM and RSPRTcharts for monitoring in medical contexts. *Statistical Methods in Medical Research*, *12*(2), 147–170. https://doi.org/10.1177/096228020301200205

Gross, C. P., & Sepkowitz, K. A. (1998). The myth of the medical breakthrough: Smallpox, vaccination, and Jenner reconsidered. *International Journal of Infectious Diseases*, *3*(1), 54–60. https://doi.org/10.1016/S1201-9712(98)90096-0

Großhans, S., Rüdt, M., Sanden, A., Brestrich, N., Morgenstern, J., Heissler, S., & Hubbuch, J. (2018). In-line Fourier-transform infrared

spectroscopy as a versatile process analytical technology for preparative protein chromatography. *Journal of Chromatography A*, *1547*, 37–44. https://doi.org/10.1016/j.chroma.2018.03.005

Großhans, S., Suhm, S., & Hubbuch, J. (2019). Precipitation of complex antibody solutions: influence of contaminant composition and cell culture medium on the precipitation behavior. *Bioprocess and Biosystems Engineering*, *42*(6), 1039–1051. https://doi.org/10.1007/s00449-019-02103-y

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

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

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

Hammerschmidt, N., Hobiger, S., & Jungbauer, A. (2016). Continuous polyethylene glycol precipitation of recombinant antibodies: Sequential precipitation and resolubilization. *Process Biochemistry*, *51*(2), 325–332. https://doi.org/10.1016/j.procbio.2015.11.032

Hanemaaijer, J. H., Robbertsen, T., van den Boomgaard, T., & Gunnink, J. W. (1989). Fouling of ultrafiltration membranes. The role of protein adsorption and salt precipitation. *Journal of Membrane Science*, *40*(2), 199–217. https://doi.org/10.1016/0376-7388(89)89005-2

Hansen, L., De Beer, T., Pierre, K., Pastoret, S., Bonnegarde-Bernard, A., Daoussi, R., ... Remon, J. P. (2015). FTIR spectroscopy for the detection and evaluation of live attenuated viruses in freeze dried vaccine formulations. *Biotechnology Progress*, *31*(4), 1107–1118. https://doi.org/10.1002/btpr.2100

Hanslip, S. J., Zaccai, N. R., Middelberg, A. P. J., & Falconer, R. J. (2006). Assembly of Human Papillomavirus Type-16 Virus-Like

Particles: Multifactorial Study of Assembly and Competing Aggregation. *Biotechnology Progress*, *22*(2), 554–560. https://doi.org/10.1021/bp0502781

Harms, Z. D., Selzer, L., Zlotnick, A., & Jacobson, S. C. (2015). Monitoring Assembly of Virus Capsids with Nanofluidic Devices. *ACS Nano*, *9*(9), 9087–9096. https://doi.org/10.1021/acsnano.5b03231

Harris, R. C., & Pettitt, B. M. (2016). Reconciling the understanding of 'hydrophobicity' with physics-based models of proteins. *Journal of Physics: Condensed Matter*, *28*(8), 083003. https://doi.org/10.1088/0953-8984/28/8/083003

Hebditch, M., Carballo-Amador, M. A., Charonis, S., Curtis, R., & Warwicker, J. (2017). Protein-Sol: A web tool for predicting protein solubility from sequence. *Bioinformatics*, *33*(19), 3098–3100. https://doi.org/10.1093/bioinformatics/btx345

Hebditch, M., Roche, A., Curtis, R. A., & Warwicker, J. (2019). Models for Antibody Behavior in Hydrophobic Interaction Chromatography and in Self-Association. *Journal of Pharmaceutical Sciences*, *108*(4), 1434–1441. https://doi.org/10.1016/j.xphs.2018.11.035

Hess, B., Bekker, H., Berendsen, H. J. C., & Fraaije, J. G. E. M. (1997). LINCS: A Linear Constraint Solver for molecular simulations. *Journal of Computational Chemistry*, *18*(12), 1463–1472. https://doi.org/10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H

Hillebrandt, N., Vormittag, P., Bluthardt, N., Dietrich, A., & Hubbuch, J. (2020). Integrated Process for Capture and Purification of Virus-Like Particles: Enhancing Process Performance by Cross-Flow Filtration. *Frontiers in Bioengineering and Biotechnology*, *8*, 489:1-16. https://doi.org/10.3389/fbioe.2020.00489

Hinton, G. E., Sejnowski, T. J., Poggio, T. A., & others. (1999). *Unsupervised learning: foundations of neural computation*. MIT press.

Hosokawa, M., Nogi, K., Naito, M., & Yokoyama, T. B. T.-N. T. H. (Second E. (Eds.). (2012). Evaluation methods for properties of nanostructured body. In *Nanoparticle Technology Handbook* (pp. 317–383). Amsterdam: Elsevier. https://doi.org/10.1016/B978-0-444-56336-1.50006-0

Hu, L., Trefethen, J. M., Zeng, Y., Yee, L., Ohtake, S., Lechuga-

Ballesteros, D., ... Middaugh, C. R. (2011). Biophysical Characterization and Conformational Stability of Ebola and Marburg Virus-Like Particles. *Journal of Pharmaceutical Sciences*, *100*(12), 5156–5173. https://doi.org/10.1002/jps.22724

Huhti, L., Blazevic, V., Nurminen, K., Koho, T., Hytönen, V. P., & Vesikari, T. (2010). A comparison of methods for purification and concentration of norovirus GII-4 capsid virus-like particles. *Archives of Virology*, *155*(11), 1855–1858. https://doi.org/10.1007/s00705-010-0768-z

Huisman, I. H., Prádanos, P., & Hernández, A. (2000). The effect of protein–protein and protein–membrane interactions on membrane fouling in ultrafiltration. *Journal of Membrane Science*, *179*(1), 79–90. https://doi.org/10.1016/S0376-7388(00)00501-9

Huter, M. J., & Strube, J. (2019). Model-Based Design and Process Optimization of Continuous Single Pass Tangential Flow Filtration Focusing on Continuous Bioprocessing. *Processes*, *7*(6). https://doi.org/10.3390/pr7060317

Huzair, F., & Sturdy, S. (2017). Biotechnology and the transformation of vaccine innovation: The case of the hepatitis B vaccines 1968–2000. *Studies in History and Philosophy of Science Part C :Studies in History and Philosophy of Biological and Biomedical Sciences*, *64*, 11–21. https://doi.org/10.1016/j.shpsc.2017.05.004

Hwangbo, S., Öner, M., & Sin, G. (2019). *Design of smart liquid-liquid extraction columns for downstream separations of biopharmaceuticals using deep Q-learning algorithm. Computer Aided Chemical Engineering* (Vol. 46). Elsevier Masson SAS. https://doi.org/10.1016/B978-0-12-818634-3.50046-1

ICH. (2009). ICH Harmonised Tripartite Guideline - Pharmaceutical Development Q8(R2). Retrieved from https://database.ich.org/sites/default/files/Q8_R2_Guideline.pdf

Idicula-Thomas, S., Kulkarni, A. J., Kulkarni, B. D., Jayaraman, V. K., & Balaji, P. V. (2006). A support vector machine-based method for predicting the propensity of a protein to be soluble or to form inclusion body on overexpression in Escherichia coli. *Bioinformatics*, *22*(3), 278–284. https://doi.org/10.1093/bioinformatics/bti810

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

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

Jamroz, M., Kolinski, A., & Kmiecik, S. (2014). CABS-flex predictions of protein flexibility compared with NMR ensembles. *Bioinformatics*, *30*(15), 2150–2154. https://doi.org/10.1093/bioinformatics/btu184

Janeway, C., Murphy, K. P., Travers, P., & Walport, M. (2008). Manipulation of the Immune Response. In *Janeway's Immunobiology*.

Janssens, M. E., Geysen, D., Broos, K., De Goeyse, I., Robbens, J., Van Petegem, F., ... Guisez, Y. (2010). Folding properties of the hepatitis B core as a carrier protein for vaccination research. *Amino Acids*, *38*(5), 1617–1626. https://doi.org/10.1007/s00726-009-0365-1

Jegerlehner, A., Tissot, A., Lechner, F., Sebbel, P., Erdmann, I., Kündig, T., ... Bachmann, M. F. (2002). A molecular assembly system that renders antigens of choice highly repetitive for induction of protective B cell responses. *Vaccine*, *20*(25–26), 3104–3112. https://doi.org/10.1016/S0264-410X(02)00266-9

Jiang, X., Wang, M., Graham, D. Y., & Estes, M. K. (1992). Expression, self-assembly, and antigenicity of the Norwalk virus capsid protein. *Journal of Virology*, *66*(11), 6527–6532. Retrieved from https://jvi.asm.org/content/jvi/66/11/6527.full.pdf

Jiang, Z., Tong, G., Cai, B., Xu, Y., & Lou, J. (2011). Purification and immunogenicity study of human papillomavirus 58 virus-like particles expressed in Pichia pastoris. *Protein Expression and Purification*, *80*(2), 203–210. https://doi.org/10.1016/j.pep.2011.07.009

Jiskoot, W., & Crommelin, D. (2005). *Methods for structural analysis of protein pharmaceuticals*. (Wim Jiskoot & D. Crommelin, Eds.). American Association of Pharmaceutical Scientists.

Joeris, K., Frerichs, J.-G., Konstantinov, K., & Scheper, T. (2002). In-situ microscopy: Online process monitoring of mammalian cell cultures. *Cytotechnology*, *38*(1–3), 129–134. https://doi.org/10.1023/A:1021170502775

Johnston, M. A., Søndergaard, C. R., & Nielsen, J. E. (2011). Integrated

prediction of the effect of mutations on multiple protein characteristics. *Proteins: Structure, Function and Bioinformatics*, *79*(1), 165–178. https://doi.org/10.1002/prot.22870

Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., & Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, *79*, 926–935. https://doi.org/10.1063/1.445869

Jornitz, M. W., & Meltzer, T. H. (Eds.). (2008). *Filtration and purification in the biopharmaceutical industry* (2nd ed.). CRC Press.

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

Juretić, D., Trinajstic, N., & Lucić, B. (1993). Protein secondary structure conformations and associated hydrophobicity scales. *Journal of Mathematical Chemistry*, *14*(1), 35–45.

Kaczmarczyk, S. J., Sitaraman, K., Young, H. a, Hughes, S. H., & Chatterjee, D. K. (2011). Protein delivery using engineered virus-like particles. *Proceedings of the National Academy of Sciences*, *108*(41), 16998–17003. https://doi.org/10.1073/pnas.1101874108/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1101874108

Karpenko, L. I., Ivanisenko, V. A., Pika, I. A., Chikaev, N. A., Eroshkin, A. M., Veremeiko, T. A., & Ilyichev, A. A. (2000). Insertion of foreign epitopes in HBcAg: how to make the chimeric particle assemble. *Amino Acids*, *18*(4), 329–337. https://doi.org/10.1007/s007260070072

Kaslow, D. C., & Biernaux, S. (2015). RTS,S: Toward a first landmark on the Malaria Vaccine Technology Roadmap. *Vaccine*, *33*(52), 7425–7432. https://doi.org/10.1016/j.vaccine.2015.09.061

Kattur Venkatachalam, A. R., Szyporta, M., Kiener, T. K., Balraj, P., & Kwang, J. (2014). Concentration and purification of enterovirus 71 using a weak anion-exchange monolithic column. *Virology Journal*, *11*(1), 99. https://doi.org/10.1186/1743-422X-11-99

Kaufmann, S. H. E., Juliana McElrath, M., Lewis, D. J. M., & Del Giudice, G. (2014). Challenges and responses in human vaccine development. *Current Opinion in Immunology*, *28*(1), 18–26. https://doi.org/10.1016/j.coi.2014.01.009

Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., & Kanehisa, M. (2007). AAindex: amino acid index database, progress report 2008. *Nucleic Acids Research*, *36*(Database), D202–D205. https://doi.org/10.1093/nar/gkm998

Kazaks, A., Lu, I.-N., Farinelle, S., Ramirez, A., Crescente, V., Blaha, B., ... Rosenberg, W. M. (2017). Production and purification of chimeric HBc virus-like particles carrying influenza virus LAH domain as vaccine candidates. *BMC Biotechnology*, *17*(1), 79. https://doi.org/10.1186/s12896-017-0396-8

Kegel, W. K., & Van Der Schoot, P. (2004). Competing hydrophobic and screened-coulomb interactions in hepatitis B virus capsid assembly. *Biophysical Journal*, *86*(6), 3905–3913. https://doi.org/10.1529/biophysj.104.040055

Keijzer, M., & Babovic, V. (2000). Genetic programming, ensemble methods and the bias/variance tradeoff – Introductory investigations. In *Genetic Programming* (Vol. 1802, pp. 76–90). https://doi.org/10.1007/978-3-540-46239-2_6

Kelley, B. (2009). Industrialization of mAb production technology: The bioprocessing industry at a crossroads. *MAbs*, *1*(5), 443–452. https://doi.org/10.4161/mabs.1.5.9448

Kim, H. J., Kim, S. Y., Lim, S. J., Kim, J. Y., Lee, S. J., & Kim, H.-J. (2010). One-step chromatographic purification of human papillomavirus type 16 L1 protein from Saccharomyces cerevisiae. *Protein Expression and Purification*, *70*(1), 68–74. https://doi.org/10.1016/j.pep.2009.08.005

Kim, J. (2016). Molecular Models for Hepatitis B Virus Capsid Formation, Maturation, and Envelopment: Theory and Simulation. In *Self-Assembling Systems: Theory and Simulation* (pp. 134–185). https://doi.org/10.1002/9781119113171.ch6

Klamp, T., Schumacher, J., Huber, G., Kühne, C., Meissner, U., Selmi, A., ... Sahin, U. (2011). Highly specific auto-antibodies against claudin-18 isoform 2 induced by a chimeric HBcAg virus-like particle vaccine kill tumor cells and inhibit the growth of lung metastases. *Cancer Research*, *71*(2), 516–527. https://doi.org/10.1158/0008-5472.CAN-10-2292

Kleiner, M., Hooper, L. V, & Duerkop, B. A. (2015). Evaluation of methods to purify virus-like particles for metagenomic sequencing of intestinal viromes. *BMC Genomics*, *16*(1), 7. https://doi.org/10.1186/s12864-014-1207-4

Klijn, M. E., & Hubbuch, J. (2019). Correlating multidimensional short-term empirical protein properties to long-term protein physical stability data via empirical phase diagrams. *International Journal of Pharmaceutics*, *560*, 166–174. https://doi.org/10.1016/j.ijpharm.2019.02.006

Klijn, M. E., Vormittag, P., Bluthardt, N., & Hubbuch, J. (2019). High-throughput computational pipeline for 3-D structure preparation and in silico protein surface property screening: A case study on HBcAg dimer structures. *International Journal of Pharmaceutics*, *563*, 337–346. https://doi.org/10.1016/j.ijpharm.2019.03.057

Knapp, B., Frantal, S., Cibena, M., Schreiner, W., & Bauer, P. (2011). Is an Intuitive Convergence Definition of Molecular Dynamics Simulations Solely Based on the Root Mean Square Deviation Possible? *Journal of Computational Biology*, *18*(8), 997–1005. https://doi.org/10.1089/cmb.2010.0237

Ko, M. K., Pellegrino, J. J., Nassimbene, R., & Marko, P. (1993). Characterization of the adsorption-fouling layer using globular proteins on ultrafiltration membranes. *Journal of Membrane Science*, *76*(2), 101–120. https://doi.org/10.1016/0376-7388(93)85210-N

Koho, T., Ihalainen, T. O., Stark, M., Uusi-Kerttula, H., Wieneke, R., Rahikainen, R., ... Hytönen, V. P. (2015). His-tagged norovirus-like particles: A versatile platform for cellular delivery and surface display. *European Journal of Pharmaceutics and Biopharmaceutics*, *96*, 22–31. https://doi.org/10.1016/j.ejpb.2015.07.002

Koho, T., Mäntylä, T., Laurinmäki, P., Huhti, L., Butcher, S. J., Vesikari, T., ... Hytönen, V. P. (2012). Purification of norovirus-like particles (VLPs) by ion exchange chromatography. *Journal of Virological Methods*, *181*(1), 6–11. https://doi.org/10.1016/j.jviromet.2012.01.003

Konishi, T. (2019). Re-evaluation of the evolution of influenza H1 viruses using direct PCA. *Scientific Reports*, *9*(1), 1–10. https://doi.org/10.1038/s41598-019-55254-z

Koppel, D. E. (1972). Analysis of Macromolecular Polydispersity in Intensity Correlation Spectroscopy: The Method of Cumulants. *The Journal of Chemical Physics*, *57*(11), 4814–4820. https://doi.org/10.1063/1.1678153

Kramberger, P., Urbas, L., & Štrancar, A. (2015). Downstream processing and chromatography based analytical methods for

production of vaccines, gene therapy vectors, and bacteriophages. *Human Vaccines & Immunotherapeutics*, *11*(4), 1010–1021. https://doi.org/10.1080/21645515.2015.1009817

Kratz, P. A., Böttcher, B., & Nassal, M. (1999). Native display of complete foreign protein domains on the surface of hepatitis B virus capsids. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(5), 1915–1920. https://doi.org/10.1073/pnas.96.5.1915

Krieger, E., Darden, T., Nabuurs, S. B., Finkelstein, A., & Vriend, G. (2004). Making optimal use of empirical energy functions: Force-field parameterization in crystal space. *Proteins: Structure, Function and Genetics*, *57*(4), 678–683. https://doi.org/10.1002/prot.20251

Krieger, E., Dunbrack, R., Hooft, R., & Krieger, B. (2012). Assignment of protonation states in protein and ligands: Combining pKa prediction with hydrogen bonding network optimization. In R. Baron (Ed.), *Computational Drug Discovery and Design* (Vol. 819, pp. 405–421). New York, NY, USA: Springer. https://doi.org/10.1007/978-1-61779-465-0

Krieger, E., Joo, K., Lee, J., Lee, J., Raman, S., Thompson, J., ... Karplus, K. (2009). Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: Four approaches that performed well in CASP8. *Proteins: Structure, Function and Bioinformatics*, *77*(SUPPL. 9), 114–122. https://doi.org/10.1002/prot.22570

Krieger, E., Nabuurs, S. B., & Vriend, G. (2003). Homology modeling. In P. E. Bourne & H. Weissig (Eds.), *Methods of biochemical analysis* (pp. 507–521). Hoboken, NJ, USA: Wiley-Liss, Inc.

Krieger, E., Nielsen, J. E., Spronk, C. A. E. M., & Vriend, G. (2006). Fast empirical pKa prediction by Ewald summation. *Journal of Molecular Graphics and Modelling*, *25*, 481–486. https://doi.org/10.1016/j.jmgm.2006.02.009

Krieger, E., & Vriend, G. (2015). New ways to boost molecular dynamics simulations. *Journal of Computational Chemistry*, *36*(13), 996–1007. https://doi.org/10.1002/jcc.23899

Kubat, M. (2017). *An introduction to machine learning* (Vol. 2). Springer.

Kuczewski, M., Schirmer, E., Lain, B., & Zarbis-Papastoitsis, G. (2011). A single-use purification process for the production of a monoclonal

antibody produced in a PER.C6 human cell line. *Biotechnology Journal*, *6*(1), 56–65. https://doi.org/10.1002/biot.201000292

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

Kurnik, R. T., Yu, A. W., Blank, G. S., Burton, A. R., Smith, D., Athalye, A. M., & van Reis, R. (1995). Buffer exchange using size exclusion chromatography, countercurrent dialysis, and tangential flow filtration: Models, development, and industrial application. *Biotechnology and Bioengineering*, *45*(2), 149–157. https://doi.org/10.1002/bit.260450209

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

Kyte, J., & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, *157*(1), 105–132.

Ladd Effio, C., Baumann, P., Weigel, C., Vormittag, P., Middelberg, A., & Hubbuch, J. (2016). High-throughput process development of an alternative platform for the production of virus-like particles in Escherichia coli. *Journal of Biotechnology*, *219*, 7–19. https://doi.org/10.1016/j.jbiotec.2015.12.018

Ladd Effio, C., Hahn, T., Seiler, J., Oelmeier, S. A., Asen, I., Silberer, C., ... Hubbuch, J. (2016). Modeling and simulation of anion-exchange membrane chromatography for purification of Sf9 insect cell-derived virus-like particles. *Journal of Chromatography A*, *1429*, 142–154. https://doi.org/10.1016/j.chroma.2015.12.006

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

Ladd Effio, C., Oelmeier, S. A., & Hubbuch, J. (2016). High-throughput characterization of virus-like particles by interlaced size-exclusion chromatography. *Vaccine*, *34*(10), 1259–1267. https://doi.org/10.1016/j.vaccine.2016.01.035

Lagoutte, P., Mignon, C., Donnat, S., Stadthagen, G., Mast, J., Sodoyer,

R., ... Werle, B. (2016). Scalable chromatography-based purification of virus-like particle carrier for epitope based influenza A vaccine produced in Escherichia coli. *Journal of Virological Methods*, *232*, 8–11. https://doi.org/10.1016/j.jviromet.2016.02.011

Lang, R., Winter, G., Vogt, L., Zürcher, A., Dorigo, B., & Schimmele, B. (2009). Rational Design of a Stable, Freeze-Dried Virus-Like Particle-Based Vaccine Formulation. *Drug Development and Industrial Pharmacy*, *35*(1), 83–97. https://doi.org/10.1080/03639040802192806

Lange, O. F., Van Der Spoel, D., & De Groot, B. L. (2010). Scrutinizing molecular mechanics force fields on the submicrosecond timescale with NMR Data. *Biophysical Journal*, *99*(2), 647–655. https://doi.org/10.1016/j.bpj.2010.04.062

Larsson, D. S. D., Liljas, L., & van der Spoel, D. (2012). Virus capsid dissolution studied by microsecond molecular dynamics simulations. *PLoS Computational Biology*, *8*(5), 1–8. https://doi.org/10.1371/journal.pcbi.1002502

Lauer, T. M., Agrawal, N. J., Chennamsetty, N., Egodage, K., Helk, B., & Trout, B. L. (2012). Developability Index: A Rapid In Silico Tool for the Screening of Antibody Aggregation Propensity. *Journal of Pharmaceutical Sciences*, *101*(1), 102–115. https://doi.org/10.1002/jps.22758

Lee-Montiel, F. T., Reynolds, K. A., & Riley, M. R. (2011). Detection and quantification of poliovirus infection using FTIR spectroscopy and cell culture. *Journal of Biological Engineering*, *5*, 16. https://doi.org/10.1186/1754-1611-5-16

Leszczyszyn, O. (2012). Hydrodynamic Radius Vs Radius of Gyration. Retrieved December 17, 2019, from http://www.materials-talks.com/blog/2012/11/15/size-matters-rh-versus-rg/

Leung, A. B., Suh, K. I., & Ansari, R. R. (2006). Particle-size and velocity measurements in flowing conditions using dynamic light scattering. *Applied Optics*, *45*(10), 2186–2190. https://doi.org/10.1364/AO.45.002186

Li, M, Cripe, T. P., Estes, P. A., Lyon, M. K., Rose, R. C., & Garcea, R. L. (1997). Expression of the human papillomavirus type 11 L1 capsid protein in Escherichia coli: characterization of protein domains involved in DNA binding and capsid assembly. *Journal of Virology*, *71*(4), 2988–2995.

Li, Maolin, Beard, P., Estes, P. A., Lyon, M. K., & Garcea, R. L. (1998). Intercapsomeric Disulfide Bonds in Papillomavirus Assembly and Disassembly. *Journal of Virology*, *72*(3), 2160–2167. Retrieved from https://jvi.asm.org/content/72/3/2160

Li, Z., Gu, Q., Coffman, J. L., Przybycien, T., & Zydney, A. L. (2019). Continuous precipitation for monoclonal antibody capture using countercurrent washing by microfiltration. *Biotechnology Progress*, (June), 1–8. https://doi.org/10.1002/btpr.2886

Li, Z., & Zydney, A. L. (2017). Effect of zinc chloride and PEG concentrations on the critical flux during tangential flow microfiltration of BSA precipitates. *Biotechnology Progress*, *33*(6), 1561–1567. https://doi.org/10.1002/btpr.2545

Liew, M. W. O. O., Chuan, Y. P., & Middelberg, A. P. J. J. (2012). Reactive diafiltration for assembly and formulation of virus-like particles. *Biochemical Engineering Journal*, *68*(2010), 120–128. https://doi.org/10.1016/j.bej.2012.07.009

Link, A., Zabel, F., Schnetzler, Y., Titz, A., Brombacher, F., & Bachmann, M. F. (2012). Innate Immunity Mediates Follicular Transport of Particulate but Not Soluble Protein Antigen. *The Journal of Immunology*, *188*(8), 3724–3733. https://doi.org/10.4049/jimmunol.1103312

Liu, J., Dai, S., Wang, M., Hu, Z., Wang, H., & Deng, F. (2016). Virus like particle-based vaccines against emerging infectious disease viruses. *Virologica Sinica*, *31*(4), 279–287. https://doi.org/10.1007/s12250-016-3756-y

Liu, W., Fan, X., Wang, X., Bao, Z., Sun, Y., Rai, K., ... Nian, R. (2019). Salt-enhanced permeabilization for monoclonal antibody precipitation and purification in a tubular reactor with a depth filtration membrane with advanced chromatin extraction. *Biochemical Engineering Journal*, *151*(May). https://doi.org/10.1016/j.bej.2019.107332

Lizotte, P. H., Wen, A. M., Sheen, M. R., Fields, J., Rojanasopondist, P., Steinmetz, N. F., & Fiering, S. (2016). In situ vaccination with cowpea mosaic virus nanoparticles suppresses metastatic cancer. *Nature Nanotechnology*, *11*(3), 295–303. https://doi.org/10.1038/nnano.2015.292

Lloyd, J., & Cheyne, J. (2017). The origins of the vaccine cold chain and a glimpse of the future. *Vaccine*, *35*(17), 2115–2120. https://doi.org/10.1016/j.vaccine.2016.11.097

Lohmann, L. J., & Strube, J. (2020). Accelerating biologics manufacturing by modeling: Process integration of precipitation in mAb downstream processing. *Processes*, *8*(1). https://doi.org/10.3390/pr8010058

Lošdorfer Božič, A., & Podgornik, R. (2017). pH Dependence of Charge Multipole Moments in Proteins. *Biophysical Journal*, *113*(7), 1454–1465. https://doi.org/10.1016/j.bpj.2017.08.017

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

Low, J. G. H., Lee, L. S., Ooi, E. E., Ethirajulu, K., Yeo, P., Matter, A., ... Novotny-Diermayr, V. (2014). Safety and immunogenicity of a virus-like particle pandemic influenza A (H1N1) 2009 vaccine: Results from a double-blinded, randomized Phase I clinical trial in healthy Asian volunteers. *Vaccine*, *32*(39), 5041–5048. https://doi.org/10.1016/j.vaccine.2014.07.011

Lu, Y., Chan, W., Ko, B. Y., Vanlang, C. C., & Swartz, J. R. (2016). Assessing sequence plasticity of a virus-like nanoparticle by evolution toward a versatile scaffold for vaccines and drug delivery. *Proc Natl Acad Sci U S A*, *112*(40), 12360–12365. https://doi.org/10.1073/pnas.1510533112

Lua, L. H. L., Connors, N. K., Sainsbury, F., Chuan, Y. P., Wibowo, N., & Middelberg, A. P. J. (2014). Bioengineering virus-like particles as vaccines. *Biotechnology and Bioengineering*, *111*(3), 425–440. https://doi.org/10.1002/bit.25159

Lua, L. H. L., Fan, Y., Chang, C., Connors, N. K., & Middelberg, A. P. J. (2015). Synthetic biology design to display an 18kDa rotavirus large antigen on a modular virus-like particle. *Vaccine*, *33*(44), 5937–5944. https://doi.org/10.1016/j.vaccine.2015.09.017

Lünsdorf, H., Gurramkonda, C., Adnan, A., Khanna, N., & Rinas, U. (2011). Virus-like particle production with yeast: ultrastructural and immunocytochemical insights into Pichia pastoris producing high levels of the hepatitis B surface antigen. *Microbial Cell Factories*, *10*, 48. https://doi.org/10.1186/1475-2859-10-48

Lutomski, C. A., Lyktey, N. A., Pierson, E. E., Zhao, Z., Zlotnick, A., & Jarrold, M. F. (2018). Multiple Pathways in Capsid Assembly. *Journal of the American Chemical Society*, *140*(17), 5784–5790. https://doi.org/10.1021/jacs.8b01804

M. Re, & G. Valentini. (2010). Simple ensemble methods are competitive with state-of-the-art data integration methods for gene function prediction. *Journal of Machine Learning Research, W&C Proceedings*, *8*, 98–11. Retrieved from http://jmlr.csail.mit.edu/proceedings/papers/v8/re10a/re10a.pdf

Mach, H., & Middaugh, C. R. (1994). Simultaneous Monitoring of the Environment of Tryptophan, Tyrosine, and Phenylalanine Residues in Proteins by Near-Ultraviolet Second-Derivative Spectroscopy. *Analytical Biochemistry*, *222*(2), 323–331. https://doi.org/10.1006/abio.1994.1499

Mach, H., Volkin, D. B., Troutman, R. D., Wang, B. e. i., Luo, Z., Jansen, K. U., & Shi, L. i. (2006). Disassembly and reassembly of yeast-derived recombinant human papillomavirus virus-like particles (HPV VLPs). *Journal of Pharmaceutical Sciences*, *95*(10), 2195–2206. https://doi.org/10.1002/jps.20696

Machado, M. R., González, H. C., & Pantano, S. (2017). MD simulations of viruslike particles with Supra CG solvation afforadble to desktop computers. *Journal of Chemical Theory and Computation*, *13*, 5106–5116. https://doi.org/10.1021/acs.jctc.7b00659

Magnan, C. N., Randall, A., & Baldi, P. (2009). SOLpro: Accurate sequence-based prediction of protein solubility. *Bioinformatics*, *25*(17), 2200–2207. https://doi.org/10.1093/bioinformatics/btp386

Mansour, A. A., Sereda, Y. V., Yang, J., & Ortoleva, P. J. (2015). Prospective on multiscale simulation of virus-like particles: Application to computer-aided vaccine design. *Vaccine*, *33*(44), 5890–5896. https://doi.org/10.1016/j.vaccine.2015.05.099

Maphis, N. M., Peabody, J., Crossey, E., Jiang, S., Jamaleddin Ahmad, F. A., Alvarez, M., ... Bhaskar, K. (2019). Qß Virus-like particle-based vaccine induces robust immunity and protects against tauopathy. *Npj Vaccines*, *4*(1), 26. https://doi.org/10.1038/s41541-019-0118-4

Martín-García, F., Papaleo, E., Gomez-Puertas, P., Boomsma, W., & Lindorff-Larsen, K. (2015). Comparing Molecular Dynamics Force Fields in the Essential Subspace. *PLOS ONE*, *10*(3), e0121114. Retrieved from https://doi.org/10.1371/journal.pone.0121114

Martinez, M., Spitali, M., Norrant, E. L., & Bracewell, D. G. (2019). Precipitation as an Enabling Technology for the Intensification of Biopharmaceutical Manufacture. *Trends in Biotechnology*, *37*(3), 237–241. https://doi.org/10.1016/j.tibtech.2018.09.001

Mason, H. S., Ball, J. M., Shi, J. J., Jiang, X., Estes, M. K., & Arntzen, C. J. (1996). Expression of Norwalk virus capsid protein in transgenic tobacco and potato and its oral immunogenicity in mice. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(11), 5335–5340. https://doi.org/10.1073/pnas.93.11.5335

Masuda, A., Lee, J. M., Miyata, T., Sato, T., Hayashi, S., Hino, M., ... Kusakabe, T. (2018). Purification and characterization of immunogenic recombinant virus-like particles of porcine circovirus type 2 expressed in silkworm pupae. *Journal of General Virology*, *99*(7), 917–926. https://doi.org/10.1099/jgv.0.001087

McAleer, W. J., Buynak, E. B., Maigetter, R. Z., Wampler, D. E., Miller, W. J., & Hilleman, M. R. (1984). Human hepatitis B vaccine from recombinant yeast. *Nature*, *307*(5947), 178–180. https://doi.org/10.1038/307178a0

McCarthy, M. P., White, W. I., Palmer-Hill, F., Koenig, S., & Suzich, J. A. (1998). Quantitative disassembly and reassembly of human papillomavirus type 11 viruslike particles in vitro. *Journal of Virology*, *72*(1), 32–41. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=109346&tool=pmcentrez&rendertype=abstract

McCraw, D. M., Gallagher, J. R., Torian, U., Myers, M. L., Conlon, M. T., Gulati, N. M., & Harris, A. K. (2018). Structural analysis of influenza vaccine virus-like particles reveals a multicomponent organization. *Scientific Reports*, *8*(1), 1–16. https://doi.org/10.1038/s41598-018-28700-7

McGovern, M. E., & Canning, D. (2015). Vaccination and all-cause child mortality from 1985 to 2011: global evidence from the Demographic and Health Surveys. *American Journal of Epidemiology*, *182*(9), 791–798. https://doi.org/10.1093/aje/kwv125

Mellado, M. C. M., Mena, J. A., Lopes, A., Ramírez, O. T., Carrondo, M. J. T., Palomares, L. A., & Alves, P. M. (2009). Impact of physicochemical parameters on in vitro assembly and disassembly kinetics of recombinant triple-layered rotavirus-like particles. *Biotechnology and Bioengineering*, *104*(4), 674–686. https://doi.org/10.1002/bit.22430

Mercier, S. M., Diepenbroek, B., Dalm, M. C. F., Wijffels, R. H., & Streefland, M. (2013). Multivariate data analysis as a PAT tool for early bioprocess development data. *Journal of Biotechnology*, *167*(3), 262–270. https://doi.org/10.1016/j.jbiotec.2013.07.006

Miao, Y., Johnson, J. E., & Ortoleva, P. J. (2011). All-atom multiscale simulation of cowpea chlorotic mottle virus capsid swelling. *Journal of Physical Chemistry*, *114*(34), 11181–11195. https://doi.org/10.1021/jp102314e.All-atom

Middelberg, A. P. J., Rivera-Hernandez, T., Wibowo, N., Lua, L. H. L., Fan, Y., Magor, G., ... Batzloff, M. R. (2011). A microbial platform for rapid and low-cost virus-like particle and capsomere vaccines. *Vaccine*, *29*(41), 7154–7162. https://doi.org/10.1016/j.vaccine.2011.05.075

Milich, D. R., Sallberg, M., & Maruyama, T. (1995). The humoral immune response in acute and chronic hepatitis B virus infection. *Springer Seminars in Immunopathology*, *17*(2–3), 149–166. https://doi.org/10.1007/BF00196163

Minkner, R., Baba, R., Kurosawa, Y., Suzuki, S., Kato, T., Kobayashi, S., & Park, E. Y. (2018). Purification of human papillomavirus-like particles expressed in silkworm using a Bombyx mori nucleopolyhedrovirus bacmid expression system. *Journal of Chromatography B*, *1096*, 39–47. https://doi.org/10.1016/j.jchromb.2018.08.007

Mohr, J., Chuan, Y. P., Wu, Y., Lua, L. H. L., & Middelberg, A. P. J. (2013). Virus-like particle formulation optimization by miniaturized high-throughput screening. *Methods*, *60*(3), 248–256. https://doi.org/10.1016/j.ymeth.2013.04.019

Mohsen, M. O., Heath, M. D., Cabral-Miranda, G., Lipp, C., Zeltins, A., Sande, M., ... Bachmann, M. F. (2019). Vaccination with nanoparticles combined with micro-adjuvants protects against cancer. *Journal for ImmunoTherapy of Cancer*, *7*(1), 114. https://doi.org/10.1186/s40425-019-0587-z

Mohsen, M. O., Speiser, D. E., Knuth, A., & Bachmann, M. F. (2020). Virus-like particles for vaccination against cancer. *Wiley Interdisciplinary Reviews: Nanomedicine and Nanobiotechnology*, *12*(1), 1–17. https://doi.org/10.1002/wnan.1579

Mohsen, M. O., Vogel, M., Riether, C., Muller, J., Salatino, S., Ternette, N., ... Bachmann, M. F. (2019). Targeting Mutated Plus Germline Epitopes Confers Pre-clinical Efficacy of an Instantly Formulated Cancer Nano-Vaccine. *Frontiers in Immunology*, *10*, 1015. https://doi.org/10.3389/fimmu.2019.01015

Mohsen, M. O., Zha, L., Cabral-Miranda, G., & Bachmann, M. F. (2017). Major findings and recent advances in virus–like particle (VLP)-

based vaccines. *Seminars in Immunology*, *34*(July), 123–132. https://doi.org/10.1016/j.smim.2017.08.014

Molin, S., Givskov, M., & Riise, E. (1992). *U.S. Patent No. 5,173,418*. U.S.: Washington DC: U.S. Patent and Trademark Office. Retrieved from https://patents.google.com/patent/US5173418A/en

Moore, D. S., McCabe, G. P., & Craig, B. A. (2009). Describing distributions with numbers. In S. Burke & A. Scanlan-Rohrer (Eds.), *Introduction to the Practice of statistics* (6th ed.). New York, NY, USA: W.H. Freemand and Company.

Mukherjee, S., Thorsteinsson, M. V, Johnston, L. B., Dephillips, P. A., & Zlotnick, A. (2008). A Quantitative Description of In Vitro Assembly of Human Papillomavirus 16 Virus-Like Particles, *381*(1), 229–237. https://doi.org/10.1016/j.jmb.2008.05.079

Muratori, C., Bona, R., & Federico, M. (2010). Lentivirus-Based Virus-Like Particles as a New Protein Delivery Tool. In M. Federico (Ed.), *Lentivirus Gene Engineering Protocols: Second Edition* (pp. 111–124). Totowa, NJ: Humana Press. https://doi.org/10.1007/978-1-60761-533-0_7

Murthy, A. M. V., Ni, Y., Meng, X., & Zhang, C. (2015). Production and evaluation of virus-like particles displaying immunogenic epitopes of porcine reproductive and respiratory syndrome virus (PRRSV). *International Journal of Molecular Sciences*, *16*(4), 8382–8396. https://doi.org/10.3390/ijms16048382

Naderi-Manesh, H., Sadeghi, M., Arab, S., & Moosavi Movahedi, A. A. (2001). Prediction of protein surface accessibility with information theory. *Proteins: Structure, Function, and Bioinformatics*, *42*(4), 452–459. https://doi.org/10.1002/1097-0134(20010301)42:4<452::AID-PROT40>3.0.CO;2-Q

Negrete, A., Pai, A., & Shiloach, J. (2014). Use of hollow fiber tangential flow filtration for the recovery and concentration of HIV virus-like particles produced in insect cells. *Journal of Virological Methods*, *195*, 240–246. https://doi.org/10.1016/j.jviromet.2013.10.017

Nielsen, C. M., Vekemans, J., Lievens, M., Kester, K. E., Regules, J. A., & Ockenhouse, C. F. (2018). RTS,S malaria vaccine efficacy and immunogenicity during Plasmodium falciparum challenge is associated with HLA genotype. *Vaccine*, *36*(12), 1637–1642. https://doi.org/10.1016/j.vaccine.2018.01.069

Nilsson, J., Miyazaki, N., Xing, L., Wu, B., Hammar, L., Li, T. C., ...

Cheng, R. H. (2005). Structure and assembly of a T=1 virus-like particle in BK polyomavirus. *Journal of Virology*, *79*(9), 5337–5345. https://doi.org/10.1128/JVI.79.9.5337-5345.2005

Nozaki, Y., & Tanford, C. (1971). The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions. Establishment of a hydrophobicity scale. *Journal of Biological Chemistry*, *246*(7), 2211–2217.

O'Rourke, J. P., Peabody, D. S., & Chackerian, B. (2015). Affinity selection of epitope-based vaccines using a bacteriophage virus-like particle platform. *Current Opinion in Virology*, *11*, 76–82. https://doi.org/10.1016/j.coviro.2015.03.005

Ong, H. K., Tan, W. S., & Ho, K. L. (2017). Virus like particles as a platform for cancer vaccine development. *PeerJ*, *5*, e4053. https://doi.org/10.7717/peerj.4053

Page, E. S. (1954). Continuous Inspection Schemes. *Biometrika*, *41*(1), 100–115.

Palladini, A., Thrane, S., Janitzek, C. M., Pihl, J., Clemmensen, S. B., de Jongh, W. A., ... Sander, A. F. (2018). Virus-like particle display of HER2 induces potent anti-cancer responses. *OncoImmunology*, *7*(3), e1408749. https://doi.org/10.1080/2162402X.2017.1408749

Pasteur, L. (1885). Méthode pour prévenir la rage après morsure. *Comptes Rendus de l'Académie Des Sciences*, *101*, 765–772.

Pattenden, L. K., Middelberg, A. P. J., Niebert, M., & Lipin, D. I. (2005). Towards the preparative and large-scale precision manufacture of virus-like particles. *Trends in Biotechnology*, *23*(10), 523–529. https://doi.org/10.1016/j.tibtech.2005.07.011

Peabody, D. S., Manifold-Wheeler, B., Medford, A., Jordan, S. K., do Carmo Caldeira, J., & Chackerian, B. (2008). Immunogenic Display of Diverse Peptides on Virus-like Particles of RNA Phage MS2. *Journal of Molecular Biology*, *380*(1), 252–263. https://doi.org/10.1016/j.jmb.2008.04.049

Peacey, M., Wilson, S., Baird, M. A., & Ward, V. K. (2007). Versatile RHDV virus-like particles: Incorporation of antigens by genetic modification and chemical conjugation. *Biotechnology and Bioengineering*, *98*(5), 968–977. https://doi.org/10.1002/bit.21518

Pease, L. F., Lipin, D. I., Tsai, D. H., Zachariah, M. R., Lua, L. H. L., Tarlov, M. J., & Middelberg, A. P. J. (2009). Quantitative characterization of virus-like particles by asymmetrical flow field

flow fractionation, electrospray differential mobility analysis, and transmission electron microscopy. *Biotechnology and Bioengineering*, *102*(3), 845–855. https://doi.org/10.1002/bit.22085

Peixoto, C., Sousa, M. F. Q., Silva, A. C., Carrondo, M. J. T., & Alves, P. M. (2007). Downstream processing of triple layered rotavirus like particles. *Journal of Biotechnology*, *127*(3), 452–461. https://doi.org/10.1016/j.jbiotec.2006.08.002

Penrod, S. L., Olson, T. M., & Grant, S. B. (1996). Deposition Kinetics of Two Viruses in Packed Beds of Quartz Granular Media. *Langmuir*, *12*(23), 5576–5587. https://doi.org/10.1021/la950884d

Phanse, Y., Carrillo-Conde, B. R., Ramer-Tait, A. E., Broderick, S., Kong, C. S., Rajan, K., ... Wannemuehler, M. J. (2014). A systems approach to designing next generation vaccines: combining α-galactose modified antigens with nanoparticle platforms. *Scientific Reports*, *4*(1), 3775. https://doi.org/10.1038/srep03775

Plisko, T. V, Bildyukevich, A. V, Usosky, V. V, & Volkov, V. V. (2016). Influence of the concentration and molecular weight of polyethylene glycol on the structure and permeability of polysulfone hollow fiber membranes. *Petroleum Chemistry*, *56*(4), 321–329. https://doi.org/10.1134/S096554411604006X

Plotkin, S. (2014). History of vaccination. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(34), 12283–12287. https://doi.org/10.1073/pnas.1400472111

Plotkin, S., Robinson, J. M., Cunningham, G., Iqbal, R., & Larsen, S. (2017). The complexity and cost of vaccine manufacturing – An overview. *Vaccine*, *35*(33), 4064–4071. https://doi.org/10.1016/j.vaccine.2017.06.003

Pomwised, R., Intamaso, U., Teintze, M., Young, M., & Pincus, S. (2016). Coupling Peptide Antigens to Virus-Like Particles or to Protein Carriers Influences the Th1/Th2 Polarity of the Resulting Immune Response. *Vaccines*, *4*(2), 15. https://doi.org/10.3390/vaccines4020015

Powers, D. M. W. (2011). Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies*, *2*(1), 37–63. https://doi.org/10.9735/2229-3981

Price, W. N. 2nd, Handelman, S. K., Everett, J. K., Tong, S. N., Bracic, A., Luff, J. D., ... Hunt, J. F. (2011). Large-scale experimental

studies show unexpected amino acid effects on protein expression and solubility in vivo in E. coli. *Microbial Informatics and Experimentation*, *1*(1), 6. https://doi.org/10.1186/2042-5783-1-6

Priddy, T. S., & Middaugh, C. R. (2014). Stabilization and Formulation of Vaccines. *Vaccine Development and Manufacturing*, 237–261. https://doi.org/10.1002/9781118870914.ch8

Przybycien, T. M. (1998). Protein-protein interactions as a means of purification. *Current Opinion in Biotechnology*, *9*(2), 164–170. https://doi.org/10.1016/S0958-1669(98)80110-9

Pumpens, P., & Grens, E. (1999). Hepatitis B core particles as a universal display model: A structure-function basis for development. *FEBS Letters*, *442*(1), 1–6. https://doi.org/10.1016/S0014-5793(98)01599-3

Pumpens, P., & Grens, E. (2001). HBV Core Particles as a Carrier for B Cell/T Cell Epitopes. *Intervirology*, *44*(2–3), 98–114. https://doi.org/10.1159/000050037

Pumpens, P., Ulrich, R., Sasnauskas, K., Kazaks, A., Ose, V., & Grens, E. (2008). Construction of novel vaccines on the basis of the virus-like particles: Hepatitis B virus proteins as vaccine carriers. *Medicinal Protein Engineering. CRC Press Taylor & Francis Group, Boca Raton, FL*, (March 2016), 205–248.

Ragone, R., Colonna, G., Balestrieri, C., Servillo, L., & Irace, G. (1984). Determination of tyrosine exposure in proteins by second-derivative spectroscopy. *Biochemistry*, *23*(8), 1871–1875.

Rajendar, B., Sivakumar, V., Sriraman, R., Raheem, M., Lingala, R., & Matur, R. V. (2013). A simple and rapid method to monitor the disassembly and reassembly of virus-like particles. *Analytical Biochemistry*, *440*(1), 15–17. https://doi.org/10.1016/j.ab.2013.05.009

Rammensee, H.-G., & Singh-Jasuja, H. (2013). HLA ligandome tumor antigen discovery for personalized vaccine approach. *Expert Review of Vaccines*, *12*(10), 1211–1217. https://doi.org/10.1586/14760584.2013.836911

Rathore, A. S., Yu, M., Yeboah, S., & Sharma, A. (2008). Case study and application of process analytical technology (PAT) towards bioprocessing: Use of on-line high-performance liquid chromatography (HPLC) for making real-time pooling decisions for process chromatography. *Biotechnology and Bioengineering*, *100*(2),

306–316. https://doi.org/doi:10.1002/bit.21759

Raval, A., Piana, S., Eastwood, M. P., Dror, R. O., & Shaw, D. E. (2012). Refinement of protein structure homology models via long, all-atom molecular dynamics simulations. *Proteins: Structure, Function, and Bioinformatics*, *80*(8), 2071–2079. https://doi.org/doi:10.1002/prot.24098

Re, M., & Valentini, G. (2012). Ensemble methods: A review. In *Advances in Machine Learning and Data Mining for Astronomy* (pp. 563–594).

Reddy, T., & Sansom, M. S. P. (2016). The Role of the Membrane in the Structure and Biophysical Robustness of the Dengue Virion Envelope. *Structure*, *24*(3), 375–382. https://doi.org/10.1016/j.str.2015.12.011

Reddy, T., Shorthouse, D., Parton, D. L., Jefferys, E., Fowler, P. W., Chavent, M., ... Sansom, M. S. P. (2015). Nothing to Sneeze At: A Dynamic and Integrative Computational Model of an Influenza A Virion. *Structure*, *23*(3), 584–597. https://doi.org/10.1016/j.str.2014.12.019

Reiter, K., Aguilar, P. P., Wetter, V., Steppert, P., Tover, A., & Jungbauer, A. (2019). Separation of virus-like particles and extracellular vesicles by flow-through and heparin affinity chromatography. *Journal of Chromatography A*, *1588*, 77–84. https://doi.org/10.1016/j.chroma.2018.12.035

Ren, Y., Wong, S. M., & Lim, L. Y. (2006). In vitro-reassembled plant virus-like particles for loading of polyacids. *Journal of General Virology*, *87*(9), 2749–2754. https://doi.org/10.1099/vir.0.81944-0

Roberts, J. A., Kuiper, M. J., Thorley, B. R., Smooker, P. M., & Hung, A. (2012). Investigation of a predicted N-terminal amphipathic α-helix using atomistic molecular dynamics simulation of a complete prototype poliovirus virion. *Journal of Molecular Graphics and Modelling*, *38*, 165–173. https://doi.org/10.1016/j.jmgm.2012.06.009

Roch, P., & Mandenius, C.-F. (2016). On-line monitoring of downstream bioprocesses. *Current Opinion in Chemical Engineering*, *14*, 112–120. https://doi.org/10.1016/j.coche.2016.09.007

Rodgers, J. L., & Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician*, *42*(1), 59–66.

Rohovie, M. J., Nagasawa, M., & Swartz, J. R. (2017). Virus-like particles: Next-generation nanoparticles for targeted therapeutic

delivery. *Bioengineering & Translational Medicine*, *2*(1), 43–57. https://doi.org/10.1002/btm2.10049

Rolinger, L., Rüdt, M., & Hubbuch, J. (2020). A critical review of recent trends, and a future perspective of optical spectroscopy as PAT in biopharmaceutical downstream processing. *Analytical and Bioanalytical Chemistry*, *412*(9), 2047–2064. https://doi.org/10.1007/s00216-020-02407-z

Roseman, A. M., Borschukova, O., Berriman, J. A., Wynne, S. A., Pumpens, P., & Crowther, R. A. (2012). Structures of hepatitis b virus cores presenting a model epitope and their complexes with antibodies. *Journal of Molecular Biology*, *423*(1), 63–78. https://doi.org/10.1016/j.jmb.2012.06.032

Rothstein, F. (1993). Differential Precipitation of Proteins: Science and Technology. In R. Harrison (Ed.), *Protein Purification Process Engineering* (pp. 115–208). CRC Press.

Rüdt, M., Briskot, T., & Hubbuch, J. (2017). Advances in downstream processing of biologics – Spectroscopy: An emerging process analytical technology. *Journal of Chromatography A*, *1490*, 2–9. https://doi.org/10.1016/j.chroma.2016.11.010

Rüdt, M., Vormittag, P., Hillebrandt, N., & Hubbuch, J. (2019). Process monitoring of virus-like particle reassembly by diafiltration with UV/Vis spectroscopy and light scattering. *Biotechnology and Bioengineering*, *116*, 1366–1379. https://doi.org/10.1002/bit.26935

Russell, B. J., Velez, J. O., Laven, J. J., Johnson, A. J., Chang, G. J. J., & Johnson, B. W. (2007). A comparison of concentration methods applied to non-infectious flavivirus recombinant antigens for use in diagnostic serological assays. *Journal of Virological Methods*, *145*(1), 62–70. https://doi.org/10.1016/j.jviromet.2007.05.008

Rustmeier, N. H., Strebl, M., & Stehle, T. (2019). The Symmetry of Viral Sialic Acid Binding Sites-Implications for Antiviral Strategies. *Viruses*, *11*(10), 947. https://doi.org/10.3390/v11100947

Rybka, J. D., Mieloch, A. A., Plis, A., Pyrski, M., Pniewski, T., & Giersig, M. (2019). Assembly and characterization of HBc derived virus-like particles with magnetic core. *Nanomaterials*, *9*(2). https://doi.org/10.3390/nano9020155

Samak, T., Gunter, D., & Wang, Z. (2012). Prediction of protein solubility in E. coli. In *2012 IEEE 8th International Conference on E-Science* (pp. 1–8). IEEE.

https://doi.org/10.1109/eScience.2012.6404416

Sandra, F., Khaliq, N. U., Sunna, A., & Care, A. (2019). Developing protein-based nanoparticles as versatile delivery systems for cancer therapy and imaging. *Nanomaterials*, *9*(9). https://doi.org/10.3390/nano9091329

Sapp, M., Fligge, C., Petzak, I., Harris, J. R., & Streeck, R. E. (1998). Papillomavirus assembly requires trimerization of the major capsid protein by disulfides between two highly conserved cysteines. *Journal of Virology*, *72*(7), 6186–6189. Retrieved from https://pubmed.ncbi.nlm.nih.gov/9621087

Savitzky, A., & Golay. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, *36*(8), 1627–1639.

Schagen, F. H. E., Rademaker, H. J., Rabelink, M. J. W. E., van Ormondt, H., Fallaux, F. J., van der Eb, A. J., & Hoeben, R. C. (2000). Ammonium sulphate precipitation of recombinant adenovirus from culture medium: An easy method to increase the total virus yield. *Gene Therapy*, *7*(18), 1570–1574. https://doi.org/10.1038/sj.gt.3301285

Schaller, A., Connors, N. K., Oelmeier, S. A., Hubbuch, J., & Middelberg, A. P. J. (2015). Predicting recombinant protein expression experiments using molecular dynamics simulation. *Chemical Engineering Science*, *121*, 340–350. https://doi.org/10.1016/j.ces.2014.09.044

Schapire, R. E. (1999). A brief introduction to boosting. *IJCAI International Joint Conference on Artificial Intelligence*, *2*, 1401–1406.

Schijven, J. F., & Hassanizadeh, S. M. (2010). Removal of viruses by soil passage: Overview of modeling, processes, and parameters. *Environmental Science and Technology*, *30*(1), 49–127. https://doi.org/10.1080/10643380091184174

Schödel, F., Kelly, S., Tinge, S., Hopkins, S., Peterson, D., Milich, D., & Curtiss, R. (1996). Hybrid hepatitis B virus core antigen as a vaccine carrier moiety. In *Novel Strategies in the Design and Production of Vaccines* (pp. 15–21). Springer.

Schoonen, L., Pille, J., Borrmann, A., Nolte, R. J. M., & van Hest, J. C. M. (2015). Sortase A-Mediated N-Terminal Modification of Cowpea Chlorotic Mottle Virus for Highly Efficient Cargo Loading.

Bioconjugate Chemistry, *26*(12), 2429–2434. https://doi.org/10.1021/acs.bioconjchem.5b00485

Schumacher, J., Bacic, T., Staritzbichler, R., Daneschdar, M., Klamp, T., Arnold, P., ... Sahin, U. (2018). Enhanced stability of a chimeric hepatitis B core antigen virus-like-particle (HBcAg-VLP) by a C-terminal linker-hexahistidine-peptide. *Journal of Nanobiotechnology*, *16*(39), 1–21. https://doi.org/10.1186/s12951-018-0363-0

Selzer, L., Katen, S. P., & Zlotnick, A. (2014). The hepatitis B virus core protein intradimer interface modulates capsid assembly and stability. *Biochemistry*, *53*(34), 5496–5504. https://doi.org/10.1021/bi500732b

Selzer, L., & Zlotnick, A. (2017). Assembly and Release of Hepatitis B Virus, 1–18. https://doi.org/10.1101/cshperspect.a021394

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Shi, L., Sanyal, G., Ni, A., Luo, Z., Doshna, S., Wang, B., ... Volkin, D. B. (2005). Stabilization of human papillomavirus virus-like particles by non-ionic surfactants. *Journal of Pharmaceutical Sciences*, *94*(7), 1538–1551. https://doi.org/10.1002/jps.20377

Silverman, B. D. (2003). Hydrophobicity of transmembrane proteins: spatially profiling the distribution. *Protein Science : A Publication of the Protein Society*, *12*(3), 586–599. https://doi.org/10.1110/ps.0214903

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

Simm, S., Einloft, J., Mirus, O., & Schleiff, E. (2016). 50 years of amino acid hydrophobicity scales: revisiting the capacity for peptide classification. *Biological Research*, *49*(1), 31. https://doi.org/10.1186/s40659-016-0092-5

Singh, A., Upadhyay, V., Upadhyay, A. K., Singh, S. M., & Panda, A. K. (2015). Protein recovery from inclusion bodies of Escherichia coli using mild solubilization process. *Microbial Cell Factories*, *14*(1), 41. https://doi.org/10.1186/s12934-015-0222-8

Singh, S., & Zlotnick, A. (2003). Observed hysteresis of virus capsid disassembly is implicit in kinetic models of assembly. *Journal of*

Biological Chemistry, *278*(20), 18249–18255. https://doi.org/10.1074/jbc.M211408200

Sippl, M. (1990). Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge based prediction of local structures in globular proteins. *J Mol Biol*, *213*, 859–883. https://doi.org/10.1016/S0022-2836(05)80269-4

Smialowski, P., Martin-Galiano, A. J., Mikolajka, A., Girschick, T., Holak, T. A., & Frishman, D. (2006). Protein solubility: sequence based prediction and experimental verification. *Bioinformatics*, *23*(19), 2536–2542. https://doi.org/10.1093/bioinformatics/btl623

Smoluchowski, M. von. (1921). Handbuch der Elektrizität und des Magnetismus. *Band II, Barth-Verlag, Leipzig*.

Smyth, P. (1996). Clustering Using Monte Carlo Cross-Validation. *KDD-96 Proceedings*, 126–133. https://doi.org/10.5555/3001460.3001486

Sokalingam, S., Raghunathan, G., Soundrarajan, N., & Lee, S.-G. (2012). A Study on the Effect of Surface Lysine to Arginine Mutagenesis on Protein Stability and Structure Using Green Fluorescent Protein. *PLoS ONE*, *7*(7), e40410. https://doi.org/10.1371/journal.pone.0040410

Solovyov, A., Palacios, G., Briese, T., Lipkin, W. I., & Rabadan, R. (2009). Cluster analysis of the origins of the new influenza A(H1N1) virus. *Euro Surveillance : Bulletin Europeen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin*. Physics Department, Princeton University, Princeton, United States. https://doi.org/10.2807/ese.14.21.19224-en

Somasundaram, B., Chang, C., Fan, Y. Y., Lim, P.-Y., Cardosa, J., & Lua, L. (2016). Characterizing Enterovirus 71 and Coxsackievirus A16 virus-like particles production in insect cells. *Methods*, *95*, 38–45. https://doi.org/10.1016/j.ymeth.2015.09.023

Sormanni, P., Amery, L., Ekizoglou, S., Vendruscolo, M., & Popovic, B. (2017). Rapid and accurate in silico solubility screening of a monoclonal antibody library. *Scientific Reports*, *7*(1), 8200. https://doi.org/10.1038/s41598-017-07800-w

Sormanni, P., Aprile, F. A., & Vendruscolo, M. (2015). The CamSol method of rational design of protein mutants with enhanced solubility. *Journal of Molecular Biology*, *427*(2), 478–490. https://doi.org/10.1016/j.jmb.2014.09.026

Spronk, C. A. E. M., Linge, J. P., Hilbers, C. W., & Vuister, G. W.

(2002). Improving the quality of protein structures derived by NMR spectroscopy. *Journal of Biomolecular NMR*, *22*(3), 281–289. https://doi.org/10.1023/A:1014971029663

Steinbrener, J., Nelson, J., Huang, X., Marchesini, S., Shapiro, D., Turner, J. J., & Jacobsen, C. (2010). Data preparation and evaluation techniques for x-ray diffraction microscopy. *Optics Express*, *18*(18), 18598–18614. https://doi.org/10.1364/OE.18.018598

Storni, T., Lechner, F., Erdmann, I., Bächi, T., Jegerlehner, A., Dumrese, T., ... Bachmann, M. F. (2002). Critical Role for Activation of Antigen-Presenting Cells in Priming of Cytotoxic T Cell Responses After Vaccination with Virus-Like Particles. *The Journal of Immunology*, *168*(6), 2880–2886. https://doi.org/10.4049/jimmunol.168.6.2880

Stray, S. J., Ceres, P., & Zlotnick, A. (2004). Zinc Ions Trigger Conformational Change and Oligomerization of Hepatitis B Virus Capsid Protein. *Biochemistry*, *43*(31), 9989–9998. https://doi.org/10.1021/bi049571k

Strods, A., Ose, V., Bogans, J., Cielens, I., Kalnins, G., Radovica, I., ... Renhofa, R. (2015). Preparation by alkaline treatment and detailed characterisation of empty hepatitis B virus core particles for vaccine and gene therapy applications. *Scientific Reports*, *5*, 11639. https://doi.org/10.1038/srep11639

Suarez-Zuluaga, D. A., Borchert, D., Driessen, N. N., Bakker, W. A. M., & Thomassen, Y. E. (2019). Accelerating bioprocess development by analysis of all available data: A USP case study. *Vaccine*, *37*(47), 7081–7089. https://doi.org/10.1016/j.vaccine.2019.07.026

Tanford, C. (1962). Contribution of Hydrophobic Interactions to the Stability of the Globular Conformation of Proteins. *Journal of the American Chemical Society*, *84*(22), 4240–4247. https://doi.org/10.1021/ja00881a009

Tartaglia, G. G., Pawar, A. P., Campioni, S., Dobson, C. M., Chiti, F., & Vendruscolo, M. (2008). Prediction of Aggregation-Prone Regions in Structured Proteins. *Journal of Molecular Biology*, *380*(2), 425–436. https://doi.org/10.1016/j.jmb.2008.05.013

The UniProt Consortium. (2018). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, *47*(D1), D506–D515. https://doi.org/10.1093/nar/gky1049

Thomas, J. C. (1987). The determination of log normal particle size distributions by dynamic light scattering. *Journal of Colloid and Interface Science*, *117*(1), 187–192.

Thompson, C. M., Petiot, E., Lennaertz, A., Henry, O., & Kamen, A. A. (2013). Analytical technologies for influenza virus-like particle candidate vaccines: Challenges and emerging approaches. *Virology Journal*, *10*(November 2012). https://doi.org/10.1186/1743-422X-10-141

Tian, J., Wu, N., Guo, J., & Fan, Y. (2009). Prediction of amyloid fibril-forming segments based on a support vector machine. *BMC Bioinformatics*, *10*(Suppl 1), S45. https://doi.org/10.1186/1471-2105-10-S1-S45

Tiwari, A., Kateja, N., Chanana, S., & Rathore, A. S. (2018). Use of HPLC as an Enabler of Process Analytical Technology in Process Chromatography. *Analytical Chemistry*, *90*(13), 7824–7829. https://doi.org/10.1021/acs.analchem.8b00897

Trainor, K., Broom, A., & Meiering, E. M. (2017). Exploring the relationships between protein sequence, structure and solubility. *Current Opinion in Structural Biology*, *42*, 136–146. https://doi.org/10.1016/j.sbi.2017.01.004

Tretyakova, I., Hidajat, R., Hamilton, G., Horn, N., Nickols, B., Prather, R. O., ... Pushko, P. (2016). Preparation of quadri-subtype influenza virus-like particles using bovine immunodeficiency virus gag protein. *Virology*, *487*, 163–171. https://doi.org/10.1016/j.virol.2015.10.007

Tsoka, S., Ciniawskyj, O. C., Thomas, O. R. T., Titchener-Hooker, N. J., & Hoare, M. (2000). Selective Flocculation and Precipitation for the Improvement of Virus-Like Particle Recovery from Yeast Homogenate. *Biotechnology Progress*, *16*(4), 661–667. https://doi.org/10.1021/bp0000407

Tsumoto, K., Umetsu, M., Kumagai, I., Ejima, D., Philo, J. S., & Arakawa, T. (2004). Role of Arginine in Protein Refolding, Solubilization, and Purification. *Biotechnology Progress*, *20*(5), 1301–1308. https://doi.org/10.1021/bp0498793

Tulsyan, A., Garvin, C., & Ündey, C. (2018). Advances in industrial biopharmaceutical batch process monitoring: Machine-learning methods for small data problems. *Biotechnology and Bioengineering*, *115*(8), 1915–1924. https://doi.org/10.1002/bit.26605

Tulsyan, A., Garvin, C., & Ündey, C. (2019). Industrial batch process

monitoring with limited data. *Journal of Process Control*, *77*, 114–133. https://doi.org/10.1016/j.jprocont.2019.03.002

Tumban, E., Muttil, P., Escobar, C. A. A., Peabody, J., Wafula, D., Peabody, D. S., & Chackerian, B. (2015). Preclinical refinements of a broadly protective VLP-based HPV vaccine targeting the minor capsid protein, L2. *Vaccine*, *33*(29), 3346–3353. https://doi.org/10.1016/j.vaccine.2015.05.016

Ulrich, R., Borisova, G. P., Grens, E., Berzin, I., Pumpens, P., Eckert, R., ... Krüger, D. H. (1992). Immunogenicity of recombinant core particles of hepatitis B virus containing epitopes of human immunodeficiency virus 1 core antigen. *Archives of Virology*, *126*(1–4), 321–328. https://doi.org/10.1007/BF01309705

US Food and Drug Administartion, & CBER. (1999). Guidance for Industry: Content and Format of Chemistry, Manufacturing and Controls. Information and Establishment Description Information for a Vaccine or Related Product.

Valerio, M., Colosimo, A., Conti, F., Giuliani, A., Grottesi, A., Manetti, C., & Zbilut, J. P. (2005). Early events in protein aggregation: Molecular flexibility and hydrophobicity/charge interaction in amyloid peptides as studied by molecular dynamics simulations. *Proteins: Structure, Function, and Bioinformatics*, *58*(1), 110–118. https://doi.org/doi:10.1002/prot.20306

van den Berg, G. B., & Smolders, C. A. (1990). Flux decline in ultrafiltration processes, *77*, 101–133. https://doi.org/10.1016/0011-9164(90)85023-4

Van Reis, R., Gadam, S., Frautschy, L. N., Orlando, S., Goodrich, E. M., Saksena, S., ... Zydney, A. L. (1997). High performance tangential flow filtration. *Biotechnology and Bioengineering*, *56*(1), 71–82. https://doi.org/10.1002/(SICI)1097-0290(19971005)56:1<71::AID-BIT8>3.0.CO;2-S

van Reis, R., Goodrich, E. M., Yson, C. L., Frautschy, L. N., Dzengeleski, S., & Lutz, H. (1997). Linear scale ultrafiltration. *Biotechnology and Bioengineering*, *55*(5), 737–746. https://doi.org/doi:10.1002/(SICI)1097-0290(19970905)55:5<737::AID-BIT4>3.0.CO;2-C

van Reis, R., & Zydney, A. (2007). Bioprocess membrane technology. *Journal of Membrane Science*, *297*(1), 16–50. https://doi.org/10.1016/j.memsci.2007.02.045

Venkatakrishnan, B., & Zlotnick, A. (2016). The Structural Biology of Hepatitis B Virus: Form and Function. *Annual Review of Virology*, *3*(1), 429–451. https://doi.org/10.1146/annurev-virology-110615-042238

Venkiteshwaran, A., Heider, P., Teysseyre, L., & Belfort, G. (2008). Selective precipitation-assisted recovery of immunoglobulins from bovine serum using controlled-fouling crossflow membrane microfiltration. *Biotechnology and Bioengineering*, *101*(5), 957–966. https://doi.org/10.1002/bit.21964

Venselaar, H., Joosten, R. P., Vroling, B., Baakman, C. A. B., Hekkelman, M. L., Krieger, E., & Vriend, G. (2010). Homology modelling and spectroscopy, a never-ending love story. *European Biophysics Journal*, *39*(4), 551–563. https://doi.org/10.1007/s00249-009-0531-0

Vertès, A. A., & Dowden, N. J. (2015). History of Monoclonal Antibodies and Lessons for the Development of Stem Cell Therapeutics. *Stem Cells in Regenerative Medicine: Science, Regulation and Business Strategies*, 665–692. https://doi.org/10.1002/9781118846193.ch33

Vicente, T., Burri, S., Wellnitz, S., Walsh, K., Rothe, S., & Liderfelt, J. (2014). Fully aseptic single-use cross flow filtration system for clarification and concentration of cytomegalovirus-like particles. *Engineering in Life Sciences*, *14*(3), 318–326. https://doi.org/10.1002/elsc.201300093

Vicente, T., Mota, J. P. B., Peixoto, C., Alves, P. M., & Carrondo, M. J. T. (2011). Rational design and optimization of downstream processes of virus particles for biopharmaceutical applications: Current advances. *Biotechnology Advances*, *29*(6), 869–878. https://doi.org/10.1016/j.biotechadv.2011.07.004

Vicente, T., Roldão, A., Peixoto, C., Carrondo, M. J. T., & Alves, P. M. (2011). Large-scale production and purification of VLP-based vaccines. *Journal of Invertebrate Pathology*, *107*, S42–S48. https://doi.org/10.1016/j.jip.2011.05.004

Vicente, T., Sousa, M. F. Q., Peixoto, C., Mota, J. P. B., Alves, P. M., & Carrondo, M. J. T. (2008). Anion-exchange membrane chromatography for purification of rotavirus-like particles. *Journal of Membrane Science*, *311*(1), 270–283. https://doi.org/10.1016/j.memsci.2007.12.021

von Heijne, G., & Blomberg, C. (1979). Trans-membrane Translocation of Proteins: The Direct Transfer Model. *European Journal of*

Biochemistry, *97*(1), 175–181. https://doi.org/10.1111/j.1432-1033.1979.tb13100.x

Vormittag, P., Klamp, T., & Hubbuch, J. (2020). Ensembles of Hydrophobicity Scales as Potent Classifiers for Chimeric Virus-Like Particle Solubility – An Amino Acid Sequence-Based Machine Learning Approach. *Frontiers in Bioengineering and Biotechnology*, *8*, 395:1-15. https://doi.org/10.3389/fbioe.2020.00395

Vriend, G. (1990). WHAT IF: a molecular modeling and drug design program. *Journal of Molecular Graphics*, *8*(1), 29,52-56.

Wagner, J. M., Pajerowski, J. D., Daniels, C. L., McHugh, P. M., Flynn, J. A., Balliet, J. W., ... Subramanian, S. (2014). Enhanced Production of Chikungunya Virus-Like Particles Using a High-pH Adapted Spodoptera frugiperda Insect Cell Line. *PLOS ONE*, *9*(4), e94401. Retrieved from https://doi.org/10.1371/journal.pone.0094401

Wang, G., Briskot, T., Hahn, T., Baumann, P., & Hubbuch, J. (2017a). Estimation of adsorption isotherm and mass transfer parameters in protein chromatography using artificial neural networks. *Journal of Chromatography A*, *1487*, 211–217. https://doi.org/10.1016/j.chroma.2017.01.068

Wang, G., Briskot, T., Hahn, T., Baumann, P., & Hubbuch, J. (2017b). Root cause investigation of deviations in protein chromatography based on mechanistic models and artificial neural networks. *Journal of Chromatography A*, *1515*, 146–153. https://doi.org/10.1016/j.chroma.2017.07.089

Wang, G., Hahn, T., & Hubbuch, J. (2016). Water on hydrophobic surfaces: Mechanistic modeling of hydrophobic interaction chromatography. *Journal of Chromatography A*, *1465*, 71–78. https://doi.org/10.1016/j.chroma.2016.07.085

Wang, J., Cieplak, P., & Kollman, P. A. (2000). How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *Journal of Computational Chemistry*, *21*(12), 1049–1074.

Wang, S., Liu, H., Zhang, X., & Qian, F. (2015). Intranasal and oral vaccination with protein-based antigens: advantages, challenges and formulation strategies. *Protein & Cell*, *6*(7), 480–503. https://doi.org/10.1007/s13238-015-0164-2

Warwicker, J., Charonis, S., & Curtis, R. A. (2014). Lysine and arginine

content of proteins: Computational analysis suggests a new tool for solubility design. *Molecular Pharmaceutics*, *11*(1), 294–303. https://doi.org/10.1021/mp4004749

Watson, D. S., Kerchner, K. R., Gant, S. S., Pedersen, J. W., Hamburger, J. B., Ortigosa, A. D., & Potgieter, T. I. (2016). At-line process analytical technology (PAT) for more efficient scale up of biopharmaceutical microfiltration unit operations. *Biotechnology Progress*, *32*(1), 108–115. https://doi.org/10.1002/btpr.2193

Weigel, T., Solomaier, T., Peuker, A., Pathapati, T., Wolff, M. W., & Reichl, U. (2014). A flow-through chromatography process for influenza A and B virus purification. *Journal of Virological Methods*, *207*, 45–53. https://doi.org/10.1016/j.jviromet.2014.06.019

Wenger, M. D., DePhillips, P., & Bracewell, D. G. (2008). A Microscale Yeast Cell Disruption Technique for Integrated Process Development Strategies. *Biotechnology Progress*, *24*(3), 606–614. https://doi.org/doi:10.1021/bp070359s

Wetlaufer, D. B. (1963). Ultraviolet spectra Of Proteins and Amino Acids. *Advances in Protein Chemistry*, *17*(C), 303–390. https://doi.org/10.1016/S0065-3233(08)60056-X

Wetzel, D., Rolf, T., Suckow, M., Kranz, A., Barbian, A., Chan, J.-A., ... Piontek, M. (2018). Establishment of a yeast-based VLP platform for antigen presentation. *Microbial Cell Factories*, *17*(1), 17. https://doi.org/10.1186/s12934-018-0868-0

Whitacre, D. C., Lee, B. O., & Milich, D. R. (2009). Use of hepadnavirus core proteins as vaccine platforms. *Expert Review of Vaccines*, *8*(11), 1565–1573. https://doi.org/10.1586/erv.09.121

Wilce, M. C. J., Aguilar, M.-I., & Hearn, M. T. W. (1995). Physicochemical Basis of Amino Acid Hydrophobicity Scales: Evaluation of Four New Scales of Amino Acid Hydrophobicity Coefficients Derived from RP-HPLC of Peptides. *Analytical Chemistry*, *67*(7), 1210–1219. https://doi.org/10.1021/ac00103a012

Wilfinger, W. W., Mackey, K., & Chomczynski, P. (1997). Effect of pH and Ionic Strength on the Spectrophotometric Assessment of Nucleic Acid Purity. *BioTechniques*, *22*(3), 474–481. https://doi.org/10.2144/97223st01

Windeatt, T., & Ardeshir, G. (2004). Decision Tree Simplification For Classifier Ensembles. *International Journal of Pattern Recognition and Artificial Intelligence*, *18*(05), 749–776.

https://doi.org/10.1142/S021800140400340X

Wingfield, P. T., Stahl, S. J., Williams, R. W., & Steven, A. C. (1995). Hepatitis Core Antigen Produced in Escherichia Coli: Subunit Composition, Conformation Analysis, and in Vitro Capsid Assembly. *Biochemistry*, *34*(15), 4919–4932. https://doi.org/10.1021/bi00015a003

Wizemann, H., & von Brunn, A. (1999). Purification of E. coli-expressed HIS-tagged hepatitis B core antigen by Ni2+-chelate affinity chromatography. *Journal of Virological Methods*, *77*(2), 189–197. https://doi.org/10.1016/S0166-0934(98)00152-9

Wold, S., Esbensen, K. I. M., & Geladi, P. (1987). Principal Component Analysis, *2*, 37–52.

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, *58*(2), 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1

Wynne, S. ., Crowther, R. ., & Leslie, A. G. . (1999). The Crystal Structure of the Human Hepatitis B Virus Capsid. *Molecular Cell*, *3*(6), 771–780. https://doi.org/10.1016/S1097-2765(01)80009-5

Xiaohui, N., Feng, S., Xuehai, H., Jingbo, X., & Nana, L. (2014). Predicting the protein solubility by integrating chaos games representation and entropy in information theory. *Expert Systems with Applications*, *41*(4), 1672–1679. https://doi.org/10.1016/j.eswa.2013.08.064

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

Yang, Yang, Niroula, A., Shen, B., & Vihinen, M. (2016). PON-Sol: prediction of effects of amino acid substitutions on protein solubility. *Bioinformatics*, *32*(13), 2032–2034. https://doi.org/10.1093/bioinformatics/btw066

Yang, Yilong, Ye, Z., Su, Y., Zhao, Q., Li, X., & Ouyang, D. (2019). Deep learning for in vitro prediction of pharmaceutical formulations. *Acta Pharmaceutica Sinica B*, *9*(1), 177–185. https://doi.org/10.1016/j.apsb.2018.09.010

Yao, Y., & Lenhoff, A. M. (2004). Determination of pore size distributions of porous chromatographic adsorbents by inverse size-

exclusion chromatography. *Journal of Chromatography A*, *1037*(1–2), 273–282. https://doi.org/10.1016/j.chroma.2004.02.054

Yoshikawa, A., Tanaka, T., Hoshi, Y., Kato, N., Tachibana, K., Iizuka, H., ... Miyakawa, Y. (1993). Chimeric hepatitis B virus core particles with parts or copies of the hepatitis C virus core protein. *Journal of Virology*, *67*(10), 6064–6070. https://doi.org/10.1128/jvi.67.10.6064-6070.1993

Zahin, M., Joh, J., Khanal, S., Husk, A., Mason, H., Warzecha, H., ... Jenson, A. B. (2016). Scalable production of HPV16 L1 protein and VLPs from tobacco leaves. *PLoS ONE*, *11*(8), 1–16. https://doi.org/10.1371/journal.pone.0160995

Zambrano, R., Jamroz, M., Szczasiuk, A., Pujols, J., Kmiecik, S., & Ventura, S. (2015). AGGRESCAN3D (A3D): server for prediction of aggregation properties of protein structures. *Nucleic Acids Research*, *43*(W1), W306-13. https://doi.org/10.1093/nar/gkv359

Zampieri, G., Coggins, M., Valle, G., & Angione, C. (2017). A poly-omics machine-learning method to predict metabolite production in CHO cells. In *Proceedings of The 2nd International Electronic Conference on Metabolomics* (Vol. 2, p. 4993). Basel, Switzerland: MDPI. https://doi.org/10.3390/iecm-2-04993

Zeltins, A. (2013). Construction and characterization of virus-like particles: A review. *Molecular Biotechnology*, *53*(1), 92–107. https://doi.org/10.1007/s12033-012-9598-4

Zhang, Lin, Tang, R., Bai, S., Connors, N. K., Lua, L. H. L., Chuan, Y. P., ... Sun, Y. (2013). Molecular energetics in the capsomere of virus-like particle revealed by molecular dynamics simulations. *Journal of Physical Chemistry B*, *117*(18), 5411–5421. https://doi.org/10.1021/jp311170w

Zhang, Linlin, Lin, D., Sun, X., Curth, U., Drosten, C., Sauerhering, L., ... Hilgenfeld, R. (2020). Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. *Science*. https://doi.org/10.1126/science.abb3405

Zhang, Y., Yin, S., Zhang, B., Bi, J., Liu, Y., & Su, Z. (2020). HBc-based virus-like particle assembly from inclusion bodies using 2-methyl-2, 4-pentanediol. *Process Biochemistry*, *89*, 233–237. https://doi.org/10.1016/j.procbio.2019.10.031

Zhao, D., Sun, B., Jiang, H., Sun, S., Kong, F. T., Ma, Y., ... Jiang, C. (2015). Enterovirus71 virus-like particles produced from insect cells

and purified by multistep chromatography elicit strong humoral immune responses in mice. *Journal of Applied Microbiology*, *119*(4), 1196–1205. https://doi.org/10.1111/jam.12922

Zhao, G., Perilla, J. R., Yufenyuy, E. L., Meng, X., Chen, B., Ning, J., ... Zhang, P. (2013). Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics. *Nature*, *497*(7451), 643–646. https://doi.org/10.1038/nature12162

Zhao, Q., Allen, M. J., Wang, Y., Wang, B., Wang, N., Shi, L., & Sitrin, R. D. (2012). Disassembly and reassembly improves morphology and thermal stability of human papillomavirus type 16 virus-like particles. *Nanomedicine: Nanotechnology, Biology, and Medicine*, *8*(7), 1182–1189. https://doi.org/10.1016/j.nano.2012.01.007

Zhao, Q., Modis, Y., High, K., Towne, V., Meng, Y., Wang, Y., ... Sitrin, R. D. (2012). Disassembly and reassembly of human papillomavirus virus-like particles produces more virion-like antibody reactivity. *Virology Journal*, *9*(1), 52. https://doi.org/10.1186/1743-422X-9-52

Zlotnick, A., Ceres, P., Singh, S., & Johnson, J. M. (2002). A Small Molecule Inhibits and Misdirects Assembly of Hepatitis B Virus Capsids. *Journal of Virology*, *76*(10), 4848–4854. https://doi.org/10.1128/jvi.76.10.4848-4854.2002

Zlotnick, A, Cheng, N., Conway, J. F., Booy, F. P., Steven, A. C., Stahl, S. J., & Wingfield, P. T. (1996). Dimorphism of Hepatitis B Virus Capsids Is Strongly Influenced by the C-Terminus of the Capsid Protein. *Biochemistry*, *35*(23), 7412–7421. https://doi.org/10.1021/bi9604800

Zlotnick, Adam, Johnson, J. M., Wingfield, P. W., Stahl, S. J., & Endres, D. (1999). A Theoretical Model Successfully Identifies Features of Hepatitis B Virus Capsid Assembly †. *Biochemistry*, *38*(44), 14644–14652. https://doi.org/10.1021/bi991611a

Zlotnick, Adam, Tan, Z., & Selzer, L. (2013). One protein, at least three structures, and many functions. *Structure*, *21*(1), 6–8. https://doi.org/10.1016/j.str.2012.12.003

Zviling, M., Leonov, H., & Arkin, I. T. (2005). Genetic algorithm-based optimization of hydrophobicity tables. *Bioinformatics*, *21*(11), 2651–2656. https://doi.org/10.1093/bioinformatics/bti405

Abbreviations

(NH4)2SO4	ammonium sulfate
3-D	three-dimensional
AEX	anion exchange chromatography
ANN	artificial neural networks
CFF	cross-flow filtration
cVLP	chimeric virus-like particle
DAD	diode array detector
DF	diafiltration
DLS	dynamic light scattering
DNA	deoxyribonucleic acid
DSP	downstream processing
DTT	dithiothreitol
DV	diafiltration volumes
E. coli	*Escherichia coli*
FN	false negative
FP	false positive
GUI	graphical user interface
HBcAg	hepatitis B core antigen
HBV	hepatitis B virus
HCP	host cell protein
HPLC	high-performance liquid chromatography
HPV	human papillomavirus
HT	high-throughput
HT-CGE	high-throughput capillary gel electrophoresis
HTS	high-throughput screening
mAb	monoclonal antibody
MAD	median absolute deviation
MALS	multi-angle light scattering
MCC	Matthew's correlation coefficient
MC-CV	Monte Carlo cross-validation
MD	molecular dynamics
MIR	major immunodominant region
mmSEC	multimodal size-exclusion chromatography
mPES	modified polyethersulfone
mRNA	messenger ribonucleic acid
MSE	mean square error
MuPyVP1	murine polyomavirus protein 1

MWCO	molecular weight cut-off
P&ID	piping and instrumentation diagram
PAT	process analytical technology
PC	principal component
PCA	principal component analysis
PCC	Pearson's correlation coefficient
PEG	polyethylene glycole
PES	polyethersulfone
PLS	partial least squares
PRESS	predictive residual error sum of squares
QELS	quasi-elastic light scattering
RF	random forest
rms	root mean square
RMSD	root mean square deviation
RMSECV	root mean square error of cross-validation
RT	room temperature
SEC	size-exclusion chromatography
sEVC	soft ensemble vote classifier
SLS	static light scattering
SVM	support-vector machine
TEM	transmission electron microscopy
TMP	transmembrane pressure
TN	true negative
TP	true positive
UF	ultrafiltration
UF/DF	ultrafiltration/diafiltration
UHPLC	ultra high performance liquid chromatography
USP	upstream processing
UV	ultraviolet
UV/Vis	ultraviolet and visible
VLP	virus-like particle
WoS	Window of Stability

Amino Acid Codes

Amino Acid	Three-Letter Code	One-Letter Code
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic Acid	Asp	D
Cysteine	Cys	C
Glutamine	Gln	Q
Glutamic Acid	Glu	E
Glycine	Gly	G
Histidine	His	H
Isoleucine	Iso	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

Appendix A: Supplementary Material for Chapter 3

S3.1 Calculation of Local Hydrophobicity Around Aromatic Amino Acids

The local hydrophobicity around tryptophan and tyrosine was assessed by performing a second derivative on the ultraviolet and visible (UV/Vis) spectra and interpolating the resulting data. An interpolated derivative spectrum is shown in Supplementary Figure S3.1. The spectrum is annotated with the tryptophan minimum, the a-value, and the b-value. The a and b values are used for calculating the a/b-ratio by dividing the former through the latter

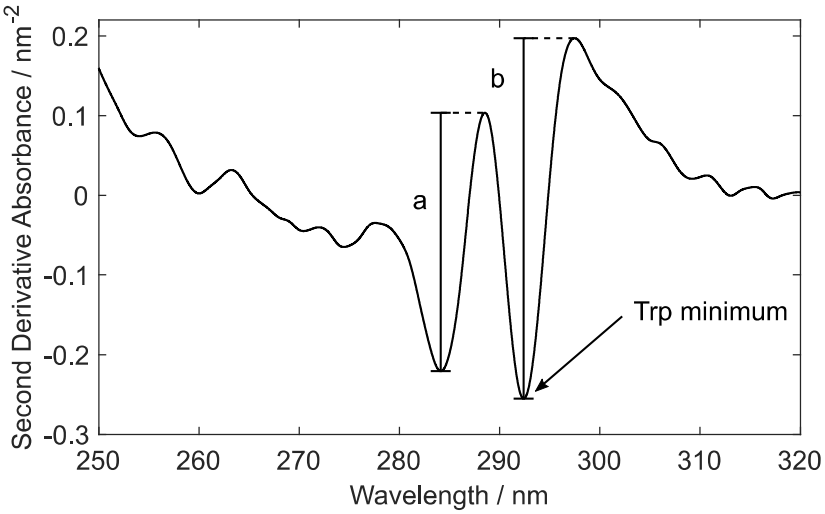


Figure S3.1: An interpolated second derivative spectrum of virus-like particle A is shown. The tryptophan (Trp) minimimum, the a-value, and the b-value are marked.

S3.2 Reversed-Phase Chromatography

The purity of the stock solutions was assessed by reverse-phase chromatography based on the absorbance of the eluting species at 280 nm. The stock solutions were analyzed with a Waters Acquity BEH300 C4 1.7 µm column (Waters Corporation, Milford, US) on an Ultimate 3000 RS ultra high-performance liquid chromatography (UHPLC) system consisting of a Pump HPG-3400RS, an Autosampler WPS-3000TFC, a Column Compartment TCC-3000RS, and a Diode Array Detector DAD-3000 controlled by Chromeleon version 6.8 SR15 (all Thermo Fisher Scientific, Waltham, US). The run duration was 6.8 min with

a flow rate of 0.45 mL/min at a temperature of 80 °C with solvent A as 0.1% trifluoroacetic acid (TFA) in water and solvent B as 0.1% TFA in acetonitrile. Equilibration was done at 5% B, and a gradient of 4.7 min was run from 23.5% to 63.5% B. The column was stripped with 95% B for 0.5 min and then reequilibrated at 5% B for 1.3 min. 2 µL were injected for each analysis. Samples were analyzed in triplicates. The purity of the stock solutions was calculated as the percentage of absorbance at 280 nm of the respective hepatitis B core antigen (HBcAg) construct of the total absorbance of all eluting species.

S3.3 Cross-Flow Filtration (CFF) Process Progress

For interested readers, the permeate mass over time of the different processes is shown in Supplementary Figure S3.2.

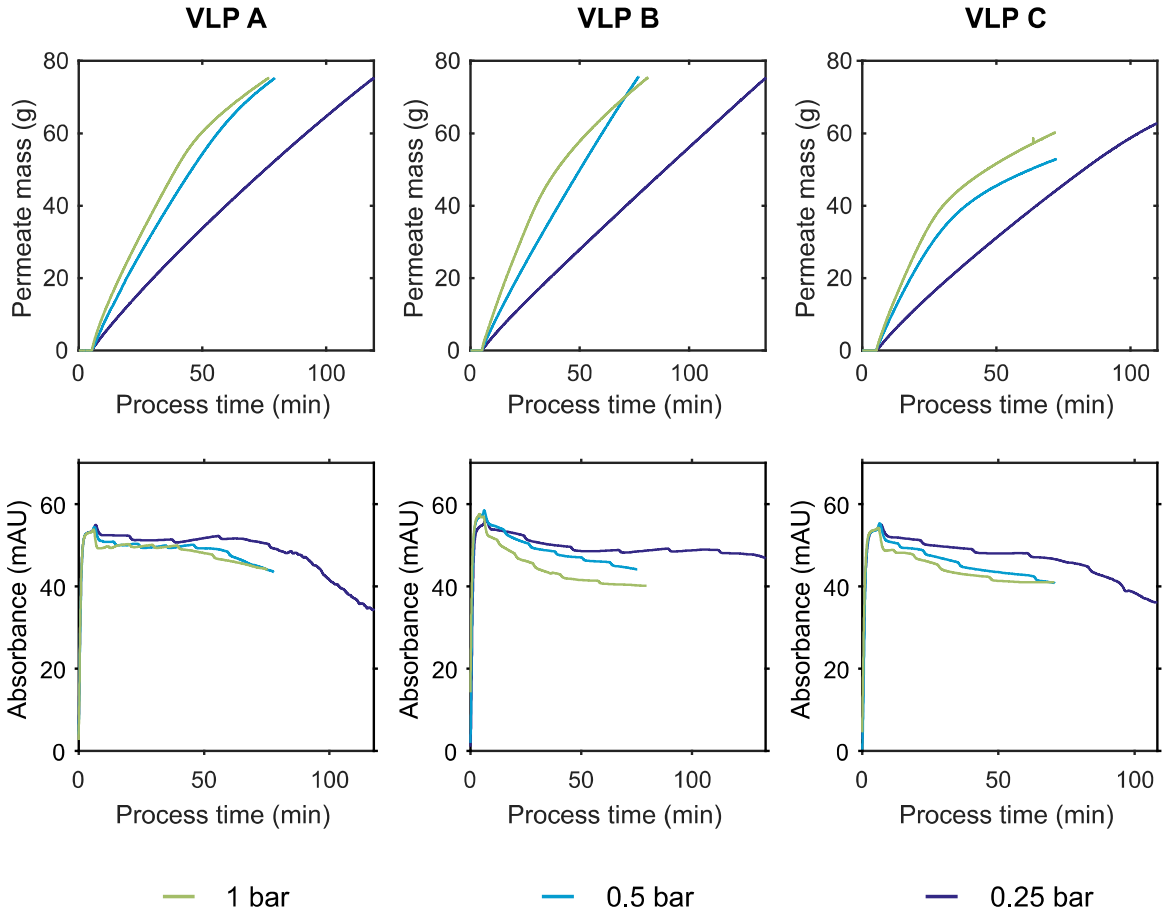


Figure S3.2: In the top row, the permeate mass over process time is shown. The bottom row shows the UV absorbance at 280 nm. VLP, virus-like particle, UV, ultraviolet.

Appendix B: Supplementary Material for Chapter 4

S4.1 Quality Parameters

Quality Z-score for each intermediate structure and WoS obtained in the proposed structure preparation pipeline is defined as the mean value of three separate WHAT IF parameters. The separate values are shown in Figure S4.1 for each intermediate structure and WoS for each VLP construct.

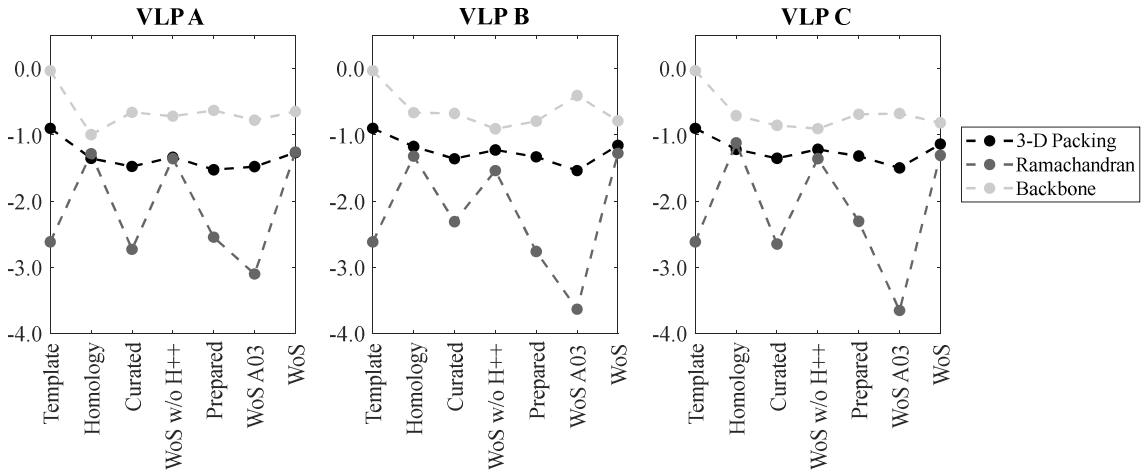


Figure S4.1: Overview of WHAT IF quality factors for the template, homology structure, curated structure, window of stability (WoS) without H++ and the YASARA2 force field ("WoS w/o H++"), the prepared structure, WoS obtained with H++ and the AMBER03 force field ("WoS A03"), and WoS obtained with H++ and the YASARA2 force field ("WoS"). WHAT IF quality factors 3-D packing (QUACHK, black), Ramachandran Z-score (RAMCHK, dark gray) and backbone conformation (BBCCHK, light grey) (Krieger et al., 2009). A dashed line is used to guide the eye between the different values. VLP, virus-like particle.

Figure S4.1 shows 3 quality parameters for each VLP construct for each intermediate structure and WoS obtained with the proposed 3-D structure preparation workflow. The backbone parameter shows a decrease from the template to the homology model for each VLP structure. The backbone quality parameter value remains stable for all VLP constructs and obtained structures, except for an increase seen at VLP B for WoS A03. The 3-D packing normality parameter also shows comparable trends for all VLP constructs. Here, a decrease is seen from the template to the homology model and an increase for WoS w/o H++ and WoS. The Ramachandran quality parameter shows fluctuation between intermediate structures and MD simulation WoS. The fluctuations are similar between VLP constructs. The lowest Ramachandran quality parameter is found for WoS A03,

followed by the curated, prepared and template structure. The homology structure, WoS w/o H++, and WoS show an increase of the Ramachandran quality parameter.

S4.2 Reproducibility of Simulation

To determine the reproducibility of the proposed protein 3-D structure preparation pipeline, all VLP constructs were simulated on two different computers using H++ computed pKa values and the YASARA2 force field. The hardware setup of the second computer was similar, using a Windows 10 computer with an Intel i7-6700 CPU and a GeForce GTX 1080 GPU. Reproducibility is evaluated based on obtained structure quality parameters, RMSD course during MD simulation, and correlation between the subsequent extracted surface charge descriptor and experimental zeta potential data. Figure S4.2 shows the quality Z-score plot for all intermediate structures and WoS obtained with the proposed structure preparation pipeline. All data is similar to the data presented in the main research article, except the WoS which was obtained using a different computer.

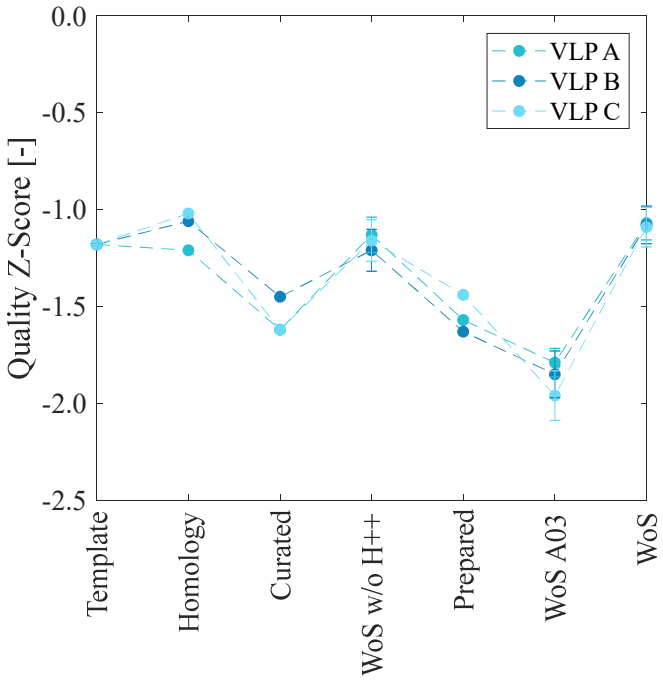


Figure S4.2: Overview of quality Z-scores for the template, homology structure, curated structure, window of stability (WoS) without H++ and the YASARA2 force field ("WoS w/o H++"), the prepared structure, WoS obtained with H++ and the AMBER03 force field ("WoS A03"), and WoS obtained with H++ and the YASARA2 force field ("WoS") on the second computer. The quality Z-score is an average value of the WHAT IF quality factors 3-D packing (QUACHK), Ramachandran Z-score (RAMCHK) and backbone conformation (BBCCHK)) *(Krieger et al., 2009)*. A median value and median absolute

deviation as error bar is shown for the WoS quality Z-scores. A dashed line is used to guide the eye between the different quality Z-scores. VLP, virus-like particle.

Figure S4.2 shows a quality Z-score of -1.07, -1.08, and -1.09 for VLP A, B, and C, respectively. These quality Z-scores have a mean difference of 0.06 compared to the quality Z-scored obtained with the first computer. An overview of the separate quality parameter obtained for the VLP constructs simulated with the second computer are shown in Figure S4.3.

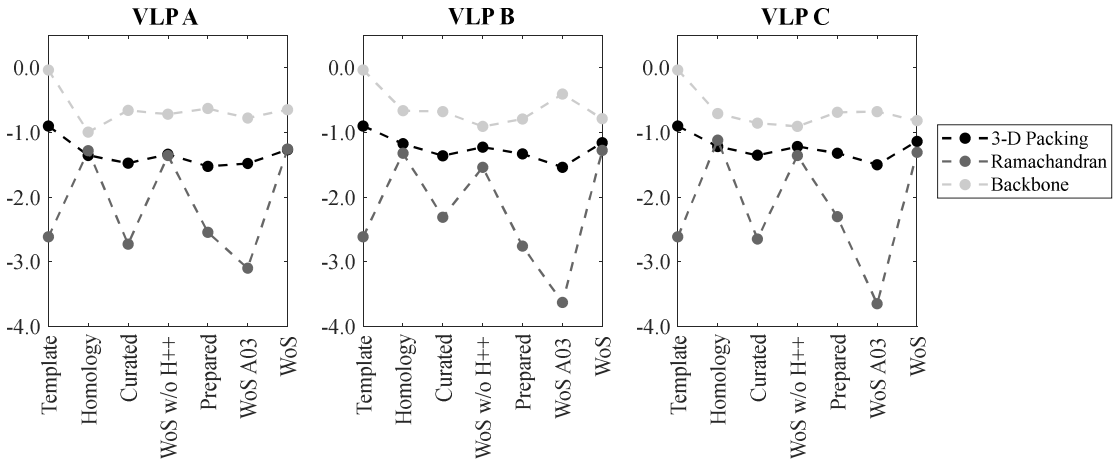


Figure S4.3: Overview of WHAT IF quality factors for the template, homology structure, curated structure, window of stability (WoS) without H++ and the YASARA2 force field ("WoS w/o H++"), the prepared structure, WoS obtained with H++ and the AMBER03 force field ("WoS A03"), and WoS obtained with H++ and the YASARA2 force field ("WoS") on the second computer. WHAT IF quality factors 3-D packing (QUACHK, black), Ramachandran Z-score (RAMCHK, dark gray) and backbone conformation (BBCCHK, light grey) *(Krieger et al., 2009)*. A dashed line is used to guide the eye between the different values.

Figure S4.3 shows 3 separate WHAT IF quality parameters. A mean difference of 0.05, 0.135, and 0.05 was calculated using all VLP constructs simulated on the second computer in values for 3-D packing normality, Ramachandran plot position normality, and the backbone conformation, respectively. This indicates that quality was not influenced by simulation of identical constructs on another computer. The course of the MD simulation, represented by the change of atom coordinates over time was monitored for the simulations with the second computer as well. The obtained data is shown in Figure S4.4.

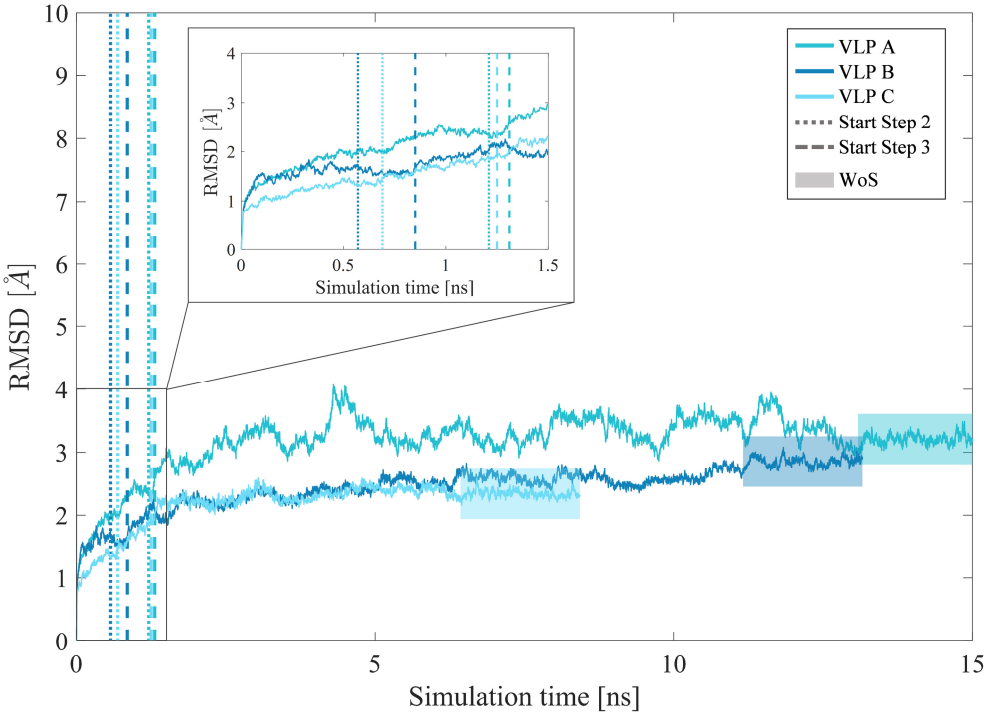


Figure S4.4: Reproducibility run of MD simulations for virus-like particles (VLP) A, B, and C presented by root-mean-square deviation (RMSD) of atom coordinates (Å) over simulation time (ns), using a second Windows 10 computer with an Intel i7-6700 CPU and a GeForce GTX 1080 GPU. Three different simulation steps are separated by vertical lines, where vertical lines indicate simulation transition points. From 0 ns to dotted line: simulation of epitope and five adjacent amino acids; from dotted to dashed line: simulation of Hepatitis B core antigen (HBcAg) dimer spike; from dashed line to the end of simulation: full dimer simulation. The highlighted area is defined as the 2 ns window of stability (WoS).

Figure S4.4 shows the MD simulation course of three VLP constructs when simulated with the second computer. VLP A reached the WoS after 15.08 ns instead of 19.89 ns seen in the main research article. VLP B and VLP C reached the WoS later compared to the first computer, with a difference of 1.19 ns and 4.36 ns, respectively. The simulation time is still in accordance with the length of epitope insertion, where VLP A contains the largest insert. The maximum RMSD reached for each VLP construct is different compared to the RMSD shown in the main research article. VLP A, B, and C have a median WoS RMSD of 3.21 ± 0.06 Å, 2.86 ± 0.06 Å, and 2.33 ± 0.05 Å, respectively, in the simulation on the second computer. This should be compared to the median WoS RMSD of 7.52 ± 0.15 Å, 3.45 ± 0.07 Å, and 2.09 ± 0.04 Å on the first computer.

Appendix C: Supplementary Material for Chapter 5

S5.1 Chimeric HBcAg Expression and Cell Lysis

The chimeric HBcAg construct was modified with a foreign epitope in the major immunodominant region and C-terminally truncated as previously described (Klamp 2011, Schumacher 2015). HBcAg protein was recombinantly overexpressed in *E. coli* BL21 DE3 (New England Biolabs, Ipswitch, US-MA). Expression was induced using a TB-based auto-induction medium developed by BioNTech Protein Therapeutics GmbH. Cells were cultured at 180 rpm and 37 °C for 7 h in a MaxQ 6000 Shaker (Thermo Scientific, Marietta, US-OH) with 250 mL medium in 1 L baffled glass shake flasks (Schott AG, Mainz, DE) up to an OD_{600} of 6. Cells were harvested by centrifugation at 4 °C at 3220 rcf for 30 min in an Eppendorf Centrifuge 5810 R (Eppendorf, Hamburg, DE), washing the pellet with phosphate-buffered saline at pH 7.4, and centrifugation at 4 °C at 17387 rcf for 20 min. Pellets were generated from 500 mL of culture volume and frozen at -30 °C for storage. For lysis, the pellet was thawed and resuspended in 20 mL of lysis buffer. Ultrasonic disruption was performed with a Digital Sonifier 450 (Branson Ultrasonic Corporation, Danbury, US-CT) at 80% amplitude for 2×40 s with a 3 min break. During this procedure, the sample was cooled in a stirred ice bath. Cell debris were separated from the supernatant by centrifugation at 4 °C and 17387 rcf for 20 min and filtration through a glass fiber and 0.45 µm cellulose acetate syringe filter (both Sartorius Stedim Biotech GmbH, Göttingen, DE). The lysate was stored at -30 °C. Prior to precipitation and re-dissolution experiments and processes, lysate was thawed and filtered again through a 0.45 µm syringe filter.

S5.2 CFF Set-Up and Temporal Alignment

In pre-experiments, flow rates for the CFF-DF steps were tested. Constraints were the linear range of permeate flowrate over TMP and the maximum tolerable flow rate of the mmSEC column. Resulting flowrates were 30 and 2 mL/min for feed and permeate flow rate, respectively. The pump of the ÄKTA Start chromatography system was bypassed and the flow generated and controlled by the CFF unit's backpressure valve. Setting the set-point as surrogate flow rate in the chromatography system settings was necessary to enable data collection and fractionation. Flow meter, chromatography fraction, and UV absorbance data were temporally aligned and processed volumes as well as fraction volumes were retrospectively corrected by integration of flow rate over time. Before integration,

flow rate data were smoothed using a moving mean with a window of 50 data points, corresponding to 3 s. Delay volumes of the chromatography system were automatically corrected. For the *mmSEC* process, the column was inserted after the fractionation valve to ensure UV absorbance monitoring during the wash procedure avoiding flow over the column. Contrary to the other processes, the wash step permeate had to be collected from the wash valve before the column. Fractions were collected manually based on the flow meter cumulative volume readings. During re-dissolution, a volume of 1.96 mL was needed for VLPs to pass the mmSEC column and was therefore manually added to the delay volume during alignment. Manual and automatic flow rate control in all processes resulted in maximum 3% deviation of the mean flow rate from the set-point and a coefficient of variation smaller than 9%. Flow rate data of the first three minutes showed transient oscillation and were omitted in the calculations.

S5.3 SEC Analysis

Samples were separated by size using analytical SEC. Three detectors were coupled to the UHPLC system, which were a DAD, MALS detector, and QELS detector. The DAD recorded spectra as well as single wavelengths, of which 260 nm and 280 nm were used for SEC purity and A260/A280 calculation. A typical UV chromatogram is shown in Figure S5.3.1.

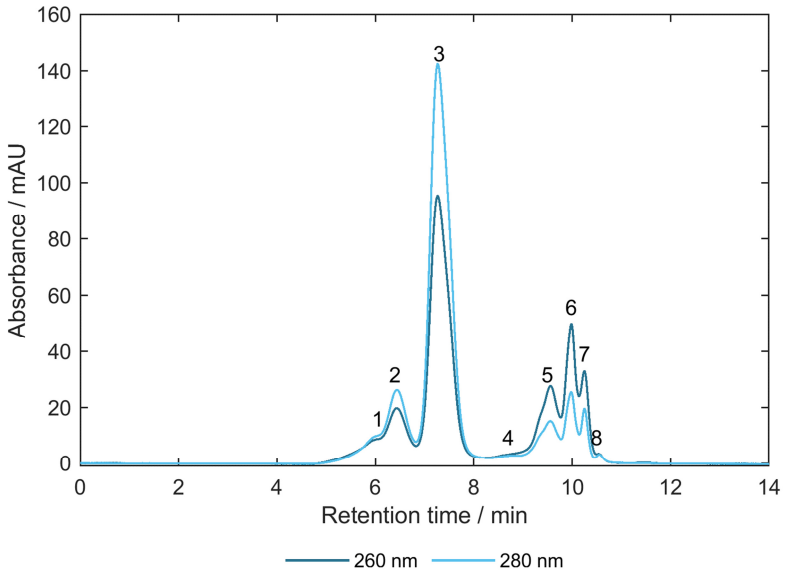


Figure S5.3.1: Size-exclusion chromatography chromatogram of the Basic process fraction F4 showing absorbance at 280 and 260 nm over retention time. Detected peaks are marked with numbers. Peaks 1-3 represent hepatitis B virus core antigen species; peaks 4-8 represent impurities.

Eight peaks were detected, whereby peaks 1-3, showing protein-typical A260/280 values of mostly <0.75, were attributed to HBcAg. This assumption was confirmed by HT-CGE analysis of samples that showed almost only peaks 1-3 (average 98% SEC purity, main text Table 5.1), such as samples of strategic pooling for process *mmSEC*. These samples exhibited one dominant peak in the HT-CGE electropherogram corresponding to monomeric HBcAg (average 96% HT-CGE purity, main text Table 5.1). During sample preparation for the protein HT-CGE assay, all proteins are denatured and reduced and therefore disassembled to monomers. It is therefore reasonable to assume that peaks 1-3 only differ in their quaternary structure while being based on HBcAg molecules. Peaks 4-8 showed higher absorbance at 260 nm and are therefore probably mainly nucleic acid species (Wilfinger et al., 1997). This scheme was observed for all CFF re-dissolution samples. For re-dissolution samples in the *Reference* process, peak 5 was dominated by protein contaminants, according to the UV spectral data (A260/A280 < 1.0, data not shown), not seen in the CFF processes. This is in accordance with lower protein purities seen for the *Reference* process samples (main text Table 5.1).

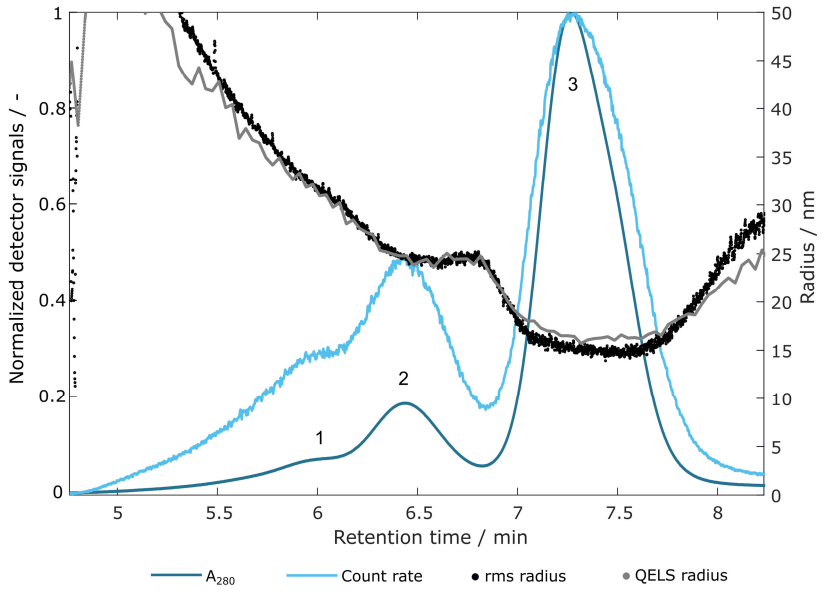


Figure S5.3.2: Absorbance at 280 nm and light scattering signals of a size-exclusion chromatography analysis of fraction F4 of the Basic process. Absorbance at 280 nm (A_{280}, —) and light scattering count rate (—) are normalized by their maximum value. Peaks 1, 2, and 3 are marked with numbers 1-3. Root mean square (rms, •) and quasi-elastic light scattering radius (QELS, •) are shown as absolute values.

Figure S5.3.2 shows an excerpt of the above shown SEC chromatogram with normalized signals of absorbance at 280 nm and count rate derived from the light scattering device focusing on peaks of HBcAg species. Root mean square (rms) radius and quasi-elastic light scattering (QELS) radius are shown which were calculated by a 1st degree Zimm model and by the manufacturers' QELS model, respectively. The size measurements were in good agreement and resulted in radii of 31-32 nm, 25 nm, and 15-16 nm for peaks 1, 2, and 3, respectively. Only peak 2 and 3 represent typical peak shapes and therefore likely represent a distinct species each, while peak 1 probably represents a broad range of aggregates of various sizes. The largest peak, peak 3, showed a radius typical for HBcAg VLPs (15-17 nm (Selzer & Zlotnick, 2017)). Other chromatograms were almost identical, but tended to diverge more at very low sample concentrations due to a disadvantageous signal-to-noise ratio (data not shown).

Figure S5.3.3 shows size and mass of species behind peak 1-3 as indicated in Figure S5.3.2 for processes *Basic*, *mmSEC*, and *Nuclease* (Figure S5.3.3A-C). The difference between the processes was small but most notably between *Nuclease* and the other two processes. This is probably due to lower sample concentrations and therefore lower signal-to-noise ratio. As discussed above, peaks represent HBcAg species, which was concluded due to a low A260/280 ratio (~0.7), and high HT-CGE purity (≥96%), and typical protein UV spectra (not shown). QELS and rms radii are in good agreement. Only for peak 3 representing VLPs, QELS radius was slightly larger than rms radius, which is expected for spherical particles (Leszczyszyn, 2012). In the following, peak radii are discussed indifferent of measurement type (rms or QELS) and processes. Peak 1 showed largest radius and weight with 30.4-32.0 nm and 11.2-12.7 MDa, respectively, and probably represents a broad size range of aggregates. The small range of the measured sizes for peak 1 is derived from the calculation method of peak data, which is based on a window of 0.15 min around the peak maximum as determined by SEC. Peak 2 was smaller with 24.4-25.2 nm and 7.5-8.0 MDa. Peak 3 was the smallest with 15.3-17.7 nm and 3.8-4.1 MDa. Its radius is consistent with HBcAg capsid size reported in literature.

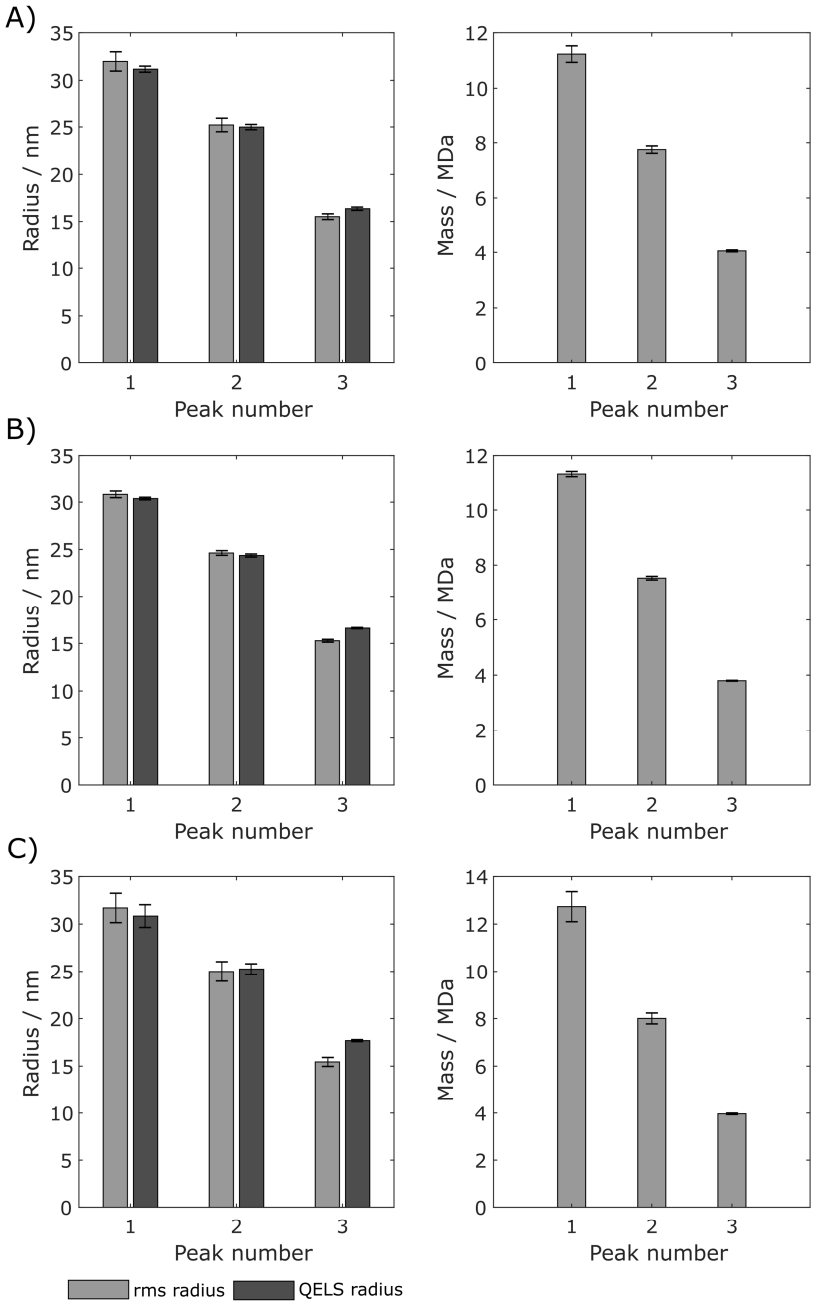


Figure S5.3.3: Size measurements of fractions with highest hepatitis B virus core antigen concentrations of the processes **(A)** Basic, **(B)** mmSEC, and **(C)** Nuclease. The left column shows root mean square radius (rms) and quasi-elastic light scattering (QELS) radius of peaks 1, 2, and 3, as indicated in Figure S5.3.2. The right column shows calculated mass of peak species 1-3. Error bars indicate standard deviations of cumulated measurement values within 0.15 min left and right of the SEC peak maximum from duplicate measurements.

Based on manual graphical size evaluation of TEM micrographs (main text Figure 5.7), it was not possible to identify distinct particle size species as seen with SEC (MALS/QELS), illustrating the limitation of this quantification method. The

difference between the even distribution of VLPs in the *Reference* and *mmSEC* process and the observed VLP clusters in *Basic* and *Nuclease* process can most probably be caused by TEM grid preparation, sample adsorption, negative staining, and washing steps, rather than by differences in the samples. Existence of such clusters were not reflected by the results of SEC. SEC, as opposed to TEM measurements, reflects solution conditions and is therefore the preferred size analytical method.

S5.4 CFF Wash and Re-Dissolution Process Data

Figure S5.4 depicts on-line concentrations for the processes *Basic*, *mmSEC*, and *Nuclease* showing both wash and re-dissolution process steps as a complement to Figure 5.6 in the main text. Initially, the signal was in saturation for processes *Basic* as well as *mmSEC* and decreased exponentially afterwards. The on-line concentration of the *Nuclease* process started below 1 g/L and also decreased exponentially. During wash, UV active impurities, such as proteins and nucleic acids, are depleted, leading to an elevated absorbance of the permeate at 280 nm which decreases over time. During the *Nuclease* process, the enzymatic digestion of nucleic acids and wash prior to precipitation leads to a lower initial contaminant level in the following wash step.

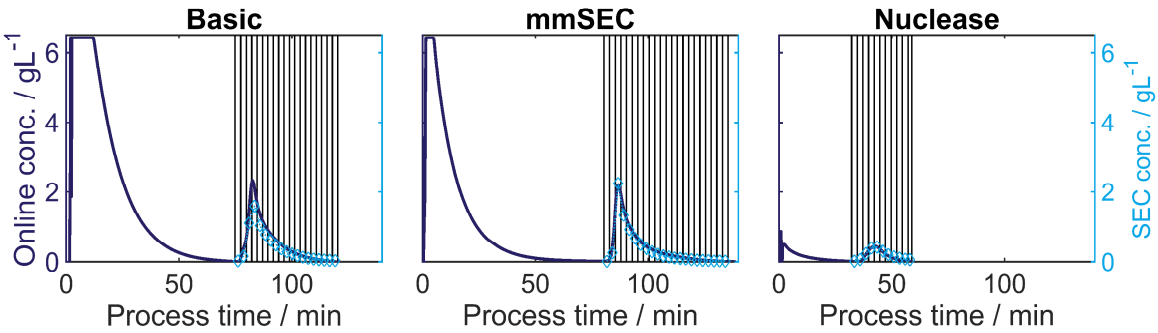


Figure S5.4: On-line monitoring of the permeate protein concentration (conc.) of wash and re-dissolution and off-line protein concentration of the re-dissolution fractions (indicated by vertical lines). Each column represents a process variant: Basic, mmSEC and Nuclease. Protein concentrations (—) are based on the absorbance at 280 nm, assuming the chimeric hepatitis B virus core antigen (HBcAg) extinction coefficient. Off-line concentrations (◊) were derived from size-exclusion chromatography (SEC) peak areas of HBcAg species (Figure S5.3.1).

S5.5: Analytical Considerations

S5.5.1 Analysis of Turbid Samples

Turbidity prohibits analysis of the samples with SEC due to presence of precipitate that would block the column. During the *mmSEC* process, turbidity was observed in fractions F2 and F3, probably due to an erroneous priming of the mmSEC column with wash buffer, containing 150 mM $(NH_4)_2SO_4$. The $(NH_4)_2SO_4$ permeates slower through the column than VLPs, as it can penetrate the pores. VLP solution therefore leaves the column in a buffer with higher $(NH_4)_2SO_4$ concentration than before entering the column, thus leading to precipitation. This effect can be circumvented by priming the column in a non-$(NH_4)_2SO_4$-containing buffer. Upon dilution, the samples became clear and could be measured by SEC. Wash samples were measured by HT-CGE as, in particular for early samples, heavy precipitation was observed.

S5.5.2 Comparability of Yields

Yields are calculated from re-dissolution and lysate HBcAg concentrations. Two separate methods have been employed to assess HBcAg concentration in the lysate and the re-dissolution samples, i.e. HT-CGE and SEC, respectively. SEC measurements exhibit much better reproducibility but could not be applied for lysate concentration measurements due to high impurity levels. Concentration determination by HT-CGE has a reproducibility of only 30% according to the manufacturer's manual. Reasons for that include low-volume liquid handling of sample and buffers, interfering particles, and baseline determination. Yields relative to each other are well comparable due to highly reproducible SEC HBcAg concentration measurements of the re-dissolution samples. Additionally, HT-CGE assessed lysate HBcAg concentrations were consistent between processes, which is owed to identical lysate preparation. However, absolute yield values are subject to variability related to HT-CGE reproducibility.

Appendix D: Supplementary Material for Chapter 6

S6.1 Normalized Hydrophobicity Scales

Table S6.1 shows normalized hydrophobicity scales that were used in this study. Scales were derived from a recent study on hydrophobicity scales for peptide classification (Simm et al., 2016). Reversed scales were excluded if there was a complementary, non-reversed scale available, and the remaining 91 scales were centered and scaled to unit variance. The scale IDs (feature names) were adapted from above-mentioned study removing spaces. For reference to the original publications of the hydrophobicity scales we refer to their article.

S6.2 Comparison of MCC and Accuracy

To compare the presented data with available models, we must understand the relation between the MCC and the typically reported accuracy. In the ideal balanced case, where $TP = TN$ and $FP = FN$ (implying $n_{\text{total,positive}} = n_{\text{total,negative}}$ which is a balanced data set), the relationship between accuracy A and the MCC is

$$A = 0.5 + 0.5 \times MCC. \tag{S6.1}$$

When $TP \neq TN$, it is

$$A < 0.5 + 0.5 \times MCC. \tag{S6.2}$$

Contrary to this, class imbalance in FP and FN increases MCC. It is however less pronounced in models that predict better than average random. Reported accuracy of other solubility models varies greatly and falls into the region of $.62 - .83$ (Hebditch et al., 2017; Idicula-Thomas et al., 2006; Magnan et al., 2009; Smialowski et al., 2006). Assuming the ideal balanced case, this translates to an MCC analogue of $0.24 - 0.66$. It is important to note, that this is the ideal case and therefore typically results in overestimated MCCs. Compared to these values, the best models in the learning experiment are close to the best reported accuracies of previous models. Class imbalance would be favorable to the presented model, as the assumed MCC of previous models is overestimated, while it is included in the actual MCC in this article.

S6.3 Feature Importance

In the learning experiment, 17290 models were created with varying training set size and number of included decision trees within the soft ensemble vote classifier. The median feature importance and the median absolute deviation (MAD) of the features were computed and are shown in Table S6.2.

|→

Table S6.1: Centered and unit-variance scaled hydrophobicity scales derived from (Simm et al., 2016). Reversed scales were excluded if there was a complementary, non-reversed scale available, resulting in 91 scales. Each amino acid, represented in single-letter code, is assigned a hydrophobicity value by each hydrophobicity scale.

SCALE-ID/AMINO ACID	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
CIDH920101	-0,611	-0,406	-0,367	-1,654	0,598	-1,137	-0,952	-1,147	0,871	0,569	1,086	-0,523	1,164	1,271	-0,289	-1,127	-0,854	1,173	1,280	1,056
CIDH920105	-0,117	-0,548	-0,891	-1,155	0,617	-1,214	-1,253	-0,920	0,118	1,636	0,980	-0,538	0,843	1,185	-0,225	-1,086	-0,891	1,538	0,950	0,970
ESID840101	0,623	-2,534	-0,775	-0,901	0,293	-0,854	-0,744	0,481	-0,398	1,376	1,062	-1,497	0,638	1,188	0,120	-0,178	-0,053	0,811	0,261	1,078
MANP780101	0,060	-0,733	-0,924	-1,285	1,114	-0,708	-0,625	-0,282	-0,454	1,774	1,285	-0,962	0,962	0,714	-0,955	-1,044	-0,752	0,670	0,346	1,800
PONP800101	-0,177	-0,713	-1,045	-1,066	1,620	-0,855	-0,917	-0,360	0,202	1,511	1,057	-1,181	1,213	0,603	-0,917	-0,869	-0,605	0,277	0,508	1,715
PONP800102	-0,172	-0,688	-1,096	-1,090	1,938	-0,777	-0,962	-0,369	-0,025	1,339	0,944	-1,383	1,237	0,701	-0,796	-0,612	-0,516	0,332	0,408	1,587
PONP800103	-0,158	-0,667	-1,176	-1,119	1,905	-0,667	-1,063	-0,384	-0,328	1,113	0,831	-1,600	1,396	0,944	-0,639	-0,243	-0,384	0,463	0,294	1,481
PONP800104	0,699	-1,129	-0,388	-1,360	1,347	-1,113	-0,149	2,018	-0,889	1,455	0,977	-0,604	0,506	1,031	-0,951	-1,144	0,198	-0,527	-0,080	0,105
PONP800105	0,244	-0,802	-1,917	-0,387	1,244	-1,740	-0,533	-0,079	0,821	-0,140	1,698	-0,771	1,498	-0,079	-0,140	-0,710	0,167	-0,294	0,367	1,552
PONP800106	-0,821	-0,553	-0,694	-1,145	1,631	-0,088	-0,088	-0,624	0,165	0,785	0,870	-1,343	2,230	1,011	-0,856	-0,462	-0,920	-0,300	-0,222	1,426
PRAM900101REVERSE	0,607	-2,235	-0,701	-1,600	0,690	-0,560	-1,395	0,485	-0,335	0,915	0,851	-1,517	0,973	1,037	0,241	0,402	0,524	0,666	0,138	0,812
SWER830101	-0,401	-0,591	-0,921	-1,311	0,170	-0,911	-1,221	-0,671	-0,641	1,250	1,220	-0,671	1,020	1,920	-0,491	-0,551	-0,281	0,500	1,670	0,910
NADH010101	0,716	-1,875	-0,900	-0,943	1,337	-1,393	-1,307	-0,022	-0,686	1,241	1,112	-0,162	0,930	1,080	-0,751	-0,269	0,020	0,727	-0,022	1,166
NADH010102	0,590	-1,354	-0,756	-0,696	1,448	-1,195	-1,065	-0,048	-0,467	1,139	1,109	-1,963	0,810	1,159	-0,706	-0,178	0,052	0,770	0,191	1,159
NADH010103	0,357	-1,240	-0,867	-0,580	1,720	-1,187	-1,038	-0,271	-0,452	1,028	1,017	-1,655	0,741	1,284	-0,942	-0,409	0,027	1,007	0,304	1,156
NADH010104	0,196	-1,190	-0,949	-0,469	1,834	-1,190	-0,960	-0,393	-0,426	1,004	0,982	-1,506	0,742	1,288	-1,048	-0,524	0,065	1,135	0,327	1,080
NADH010106	-0,232	-0,946	-1,176	0,228	2,288	-1,061	-0,382	-0,831	-0,865	0,665	0,654	-0,727	0,665	1,057	-1,475	-0,762	0,619	1,206	0,021	1,057
NADH010107	-0,434	-0,754	-1,213	1,617	2,281	-0,721	0,542	-0,959	-0,992	-0,188	-0,122	0,526	0,091	0,567	-1,500	-0,844	1,010	1,051	-0,475	0,518
WILM950101	-0,390	-0,910	-0,281	-0,538	-0,143	-0,246	-0,481	-0,304	-1,706	1,568	1,579	-1,351	-0,304	2,324	-0,017	-0,779	-0,052	0,887	0,658	0,486
WILM950102	0,475	0,069	-0,686	-1,155	-0,089	-0,811	-0,441	-0,650	-0,528	1,001	1,655	-1,137	-1,238	2,422	-0,343	-0,722	0,233	1,457	0,108	0,380
WILM950103REVERSE	1,008	1,541	0,206	0,248	-2,545	0,488	0,092	1,076	-1,853	-0,505	0,206	1,242	-0,908	0,917	-0,538	-0,041	-0,275	-0,372	-0,294	0,307
WILM950104	-1,374	0,196	0,678	-0,633	1,563	-0,378	0,077	-0,864	-1,637	2,452	-0,007	0,180	-0,194	0,583	-0,501	0,328	-0,366	0,989	-1,470	0,379
KUHL950101REVERSE	0,350	-2,003	-0,885	-1,326	1,026	-0,856	-1,621	0,644	-0,268	1,262	0,997	-0,591	0,703	1,262	0,615	-0,297	-0,444	0,585	-0,297	1,144
JURD980101	0,547	-1,503	-0,974	-1,007	1,010	-1,033	-0,874	-0,028	-0,874	1,671	1,439	-1,175	0,811	1,109	-0,445	0,018	-0,048	0,031	-0,246	1,572
KIDA850101REVERSE	0,263	-1,822	-0,789	-0,789	1,023	-1,072	-1,140	0,156	-0,273	0,750	1,072	-1,656	0,711	1,393	0,731	-0,409	-0,614	1,530	0,546	0,390
ENGD860101REVERSE	0,607	-2,234	-0,701	-1,600	0,689	-0,558	-1,396	0,484	-0,333	0,914	0,852	-1,519	0,975	1,036	0,239	0,403	0,525	0,668	0,137	0,811
KYTJ820101	0,767	-1,342	-1,008	-1,008	1,001	-1,008	-1,008	0,030	-0,907	1,671	1,436	-1,142	0,800	1,101	-0,372	-0,104	-0,070	-0,137	-0,271	1,570
ENGEL	0,607	-2,234	-0,701	-1,600	0,689	-0,558	-1,396	0,484	-0,333	0,914	0,852	-1,519	0,975	1,036	0,239	0,403	0,525	0,668	0,137	0,811
JONES	-0,468	-0,485	-1,131	-0,647	0,087	-1,209	-0,637	-1,124	-0,468	1,478	0,642	0,189	0,216	1,238	1,153	-1,148	-1,148	2,006	1,069	0,385
CIDBB	-0,358	-0,368	-0,963	-0,973	0,462	-0,671	-1,559	-1,100	0,140	1,077	0,930	-0,368	0,960	1,213	-0,290	-1,188	-0,856	1,916	1,213	0,784
CIDA	0,233	-0,623	-0,993	-1,178	0,564	-1,139	-0,925	-0,915	-0,273	1,994	1,031	-0,662	1,060	0,865	-0,176	-0,701	-1,285	1,157	0,904	1,060
CIDAB	0,090	-0,759	-0,954	-1,100	1,134	-1,247	-1,237	-0,632	-0,320	1,934	0,860	-0,681	0,509	1,182	-0,281	-0,886	-0,681	1,397	0,568	1,104
PONG1	0,204	-0,283	-0,792	-1,257	1,141	-0,747	-1,025	-0,433	-0,268	1,920	1,058	-1,324	0,631	0,833	-1,287	-0,822	-0,493	0,886	0,594	1,463
PONG2	0,376	-0,253	0,124	-1,498	1,695	-0,527	-2,068	0,243	0,791	0,799	-0,061	-1,476	-0,039	-0,135	-0,379	0,710	-0,646	2,117	0,036	0,191
PONG3	0,321	-0,300	-0,366	-1,534	1,580	-0,706	-1,724	-0,101	0,297	1,506	0,553	-1,559	0,330	0,388	-0,921	-0,060	-0,631	1,671	0,346	0,909
KIDER	0,263	-1,822	-0,789	-0,789	1,023	-1,072	-1,140	0,156	-0,273	0,750	1,072	-1,656	0,711	1,393	0,731	-0,409	-0,614	1,530	0,546	0,390
WOLR790101	1,119	-2,548	-0,830	-0,830	0,589	-0,780	-0,920	1,198	-0,930	1,159	1,178	-0,800	0,549	0,669	0,539	-0,050	-0,020	-0,190	-0,230	1,129
CHOTA	-1,264	1,253	-0,235	-0,463	-0,807	0,223	0,452	-2,180	0,566	0,109	-0,006	0,681	0,338	0,910	-0,578	-1,264	-0,692	1,940	1,367	-0,349
ROSEB	-1,107	0,683	-0,726	-0,845	-0,033	-0,344	-0,463	-1,680	0,539	0,587	0,730	-0,415	0,945	1,446	-0,964	-1,131	-0,654	2,186	1,064	0,181
ROSEAREVERSE	1,385	-1,608	0,344	0,496	0,777	-0,242	-0,090	2,035	-0,437	0,018	-0,242	-0,957	-0,459	-0,892	0,756	1,124	0,647	-1,825	-1,196	0,365
COHEN	-0,874	-0,665	-1,607	-1,189	-0,456	-0,874	-0,665	-1,189	0,382	0,696	0,801	0,487	1,534	1,220	1,115	-0,979	-0,246	1,744	0,487	0,278
JACWH	0,626	-3,164	-0,379	-1,099	0,687	-0,325	-0,922	0,565	-0,501	0,755	0,782	-1,214	0,803	0,844	0,368	0,232	0,273	0,497	0,450	0,721
CASSI	0,198	-0,671	-0,478	-1,347	1,839	-1,057	-1,250	-0,092	0,391	1,356	0,487	-1,540	0,487	0,970	-0,960	-0,671	-0,381	1,549	0,487	0,681
MDK0	0,673	-1,499	-0,973	-1,006	1,002	-1,032	-0,874	-0,097	-0,874	1,660	1,430	-1,173	0,804	1,101	-0,446	0,015	-0,051	0,028	-0,249	1,561

MDK1	0,746	-1,581	-0,949	-0,949	0,846	-1,248	-1,049	0,081	-0,750	1,710	1,378	-1,182	0,846	1,046	-0,317	-0,052	0,115	0,115	-0,218	1,411
BULDG	0,737	0,823	1,037	0,737	0,468	1,123	0,629	0,951	0,823	-1,475	-1,690	0,576	-0,627	-1,551	-0,101	0,533	0,393	-1,207	-1,454	-0,724
GUYH850101	0,175	1,832	0,523	0,798	-1,216	0,953	0,844	0,386	-0,374	-0,951	-0,996	1,365	-1,372	-1,857	0,752	0,560	0,148	-0,383	-0,108	-1,079
MIYS850101	-0,166	-0,704	-0,923	-0,979	1,050	-0,849	-0,951	-0,563	-0,274	1,470	1,302	-1,279	1,527	1,564	-0,849	-0,746	-0,559	0,923	0,096	0,909
WILSON	-0,631	-0,848	-0,604	-0,929	1,157	-0,604	-0,550	-0,225	-0,902	0,615	1,238	-1,525	0,127	1,482	0,046	-0,712	-1,146	1,590	1,373	1,048
CHOC760103	0,586	-1,490	-0,873	-0,704	1,260	-1,153	-0,536	0,474	-0,592	1,821	0,979	-1,378	0,699	1,260	-0,536	-0,311	-0,255	-0,031	-0,704	1,484
EISEN	0,620	-2,531	-0,780	-0,900	0,290	-0,850	-0,740	0,480	-0,400	1,380	1,060	-1,500	0,640	1,190	0,120	-0,180	-0,050	0,810	0,260	1,080
JANJ790102	0,589	-1,824	-0,547	-0,689	1,441	-0,831	-0,831	0,589	0,021	1,157	0,873	-2,392	0,731	0,873	-0,263	0,021	-0,121	0,589	-0,405	1,015
RAOARGOS	0,954	-1,390	-1,041	-1,467	0,780	-1,041	-1,196	0,431	-0,363	1,109	1,167	-1,506	1,070	1,361	-0,634	0,199	0,412	0,257	-0,073	0,973
NOZY710101TANFORD	0,491	-2,485	-0,832	-0,180	0,180	-0,898	-0,794	0,359	-0,472	1,209	1,351	-1,512	0,510	1,030	0,019	-0,265	-0,142	0,671	0,151	1,606
WELLING	0,946	0,585	-0,270	0,629	-0,542	0,148	-0,232	-0,947	2,193	-1,631	0,693	1,522	-2,219	-0,675	-0,118	0,053	-0,067	-0,504	0,300	0,135
PARJ860101	0,177	0,508	0,951	1,424	0,066	0,793	1,077	0,745	0,177	-1,418	-1,608	0,745	-0,818	-1,608	0,177	0,872	0,666	-1,734	-0,455	-0,739
ROSG850102	0,030	-0,837	-0,924	-1,010	1,505	-1,010	-1,010	-0,143	0,377	1,244	0,984	-1,877	0,984	1,244	-0,837	-0,663	-0,317	0,984	0,204	1,071
BISHOP	0,200	0,321	0,442	1,411	-0,406	0,684	2,381	-0,042	0,200	-0,406	-0,769	0,563	-0,285	-1,375	0,442	0,079	0,079	-2,344	-1,254	0,079
WIMLEYREVERSE	-0,372	-1,499	-0,673	-0,312	0,075	-0,604	-0,037	-0,931	-0,037	1,021	1,133	-2,350	0,634	1,529	-0,062	-0,338	-0,157	1,855	0,669	0,454
WIMLEY	-0,086	-0,832	-0,377	-1,322	0,393	-0,564	-2,244	0,101	-0,086	0,474	0,766	-1,042	0,381	1,431	-0,412	-0,039	-0,051	2,271	1,209	0,031
ARGP820101	-0,469	-0,481	-1,135	-0,651	0,088	-1,208	-0,639	-1,123	-0,469	1,480	0,645	0,185	0,221	1,238	1,153	-1,147	-1,147	2,001	1,068	0,390
FAUJ830101	-0,167	-1,449	-1,051	-1,216	1,027	-0,682	-1,089	-0,468	-0,342	1,280	1,183	-1,429	0,726	1,270	0,231	-0,507	-0,216	1,717	0,464	0,717
JOND750101	-0,467	-0,484	-1,132	-0,646	0,087	-1,209	-0,637	-1,123	-0,467	1,477	0,642	0,190	0,215	1,239	1,153	-1,149	-1,149	2,006	1,068	0,386
LEVM760101REVERSE	0,089	-1,795	-0,288	-1,526	0,358	-0,288	-1,526	-0,180	0,089	0,789	0,789	-1,795	0,519	1,165	0,573	-0,342	0,035	1,650	1,058	0,627
ZIMJ680101	-0,427	-0,427	-1,137	-0,609	0,196	-1,223	-0,600	-1,127	-0,168	1,720	1,193	0,311	0,119	1,414	1,366	-1,089	-0,705	-0,926	1,625	0,493
NADH010105	0,012	-1,206	-1,171	-0,272	2,022	-1,301	-0,946	-0,603	-0,639	0,934	0,946	-0,378	0,792	1,277	-1,277	-0,698	0,130	1,100	0,237	1,041
PONP930101	0,204	-0,283	-0,792	-1,257	1,141	-0,747	-1,025	-0,433	-0,268	1,920	1,058	-1,324	0,631	0,833	-1,287	-0,822	-0,493	0,886	0,594	1,463
COWR900101	0,238	-1,310	-0,896	-0,489	0,566	-0,841	-0,380	-0,091	-1,873	1,324	1,317	-1,678	0,832	1,270	0,582	-0,591	-0,294	1,051	0,308	0,957
BLAS910101	0,276	-1,563	-0,851	-1,474	0,454	-0,821	-1,444	-0,080	-1,059	1,225	1,225	-0,732	0,632	1,403	0,543	-0,495	-0,228	1,047	1,047	0,898
FASG890101	0,232	1,030	0,634	0,772	-1,772	0,827	1,095	0,304	-0,118	-1,349	-1,304	1,638	-0,954	-1,294	0,562	0,903	0,573	-0,837	-0,043	-0,899
FAUCH	-0,168	-1,449	-1,050	-1,216	1,028	-0,683	-1,089	-0,468	-0,341	1,279	1,184	-1,428	0,725	1,268	0,231	-0,507	-0,214	1,717	0,464	0,718
PONNU	-0,176	-0,712	-1,046	-1,066	1,621	-0,855	-0,916	-0,360	0,202	1,512	1,058	-1,182	1,212	0,601	-0,916	-0,868	-0,606	0,277	0,506	1,713
JANIN	0,245	-1,091	-0,839	-0,839	2,661	-0,924	-0,924	0,327	-0,505	1,410	0,828	-1,132	0,408	0,661	-0,672	-0,505	-0,590	0,160	0,078	1,243
GUYFE	-0,219	-1,057	-0,670	-1,015	1,308	-0,940	-0,983	-0,595	0,371	1,255	1,145	-1,423	1,209	1,653	-0,908	-0,691	-0,446	0,790	0,200	1,017
CHOTH	0,585	-1,489	-0,873	-0,705	1,259	-1,154	-0,534	0,475	-0,591	1,822	0,981	-1,379	0,699	1,259	-0,534	-0,309	-0,256	-0,032	-0,705	1,483
VHEG790101	0,731	-2,592	-0,323	-1,555	-0,303	-0,187	-1,139	0,460	-0,455	1,137	1,104	-0,676	0,526	1,375	-0,425	0,050	0,219	1,002	0,050	1,002
ROSEM	0,224	-1,909	-1,151	-0,824	0,603	-1,081	-0,922	-0,090	-0,599	1,323	1,323	-1,240	0,692	1,323	0,729	-0,136	0,107	1,248	-0,548	0,930
LEVIT	0,089	-1,795	-0,288	-1,526	0,358	-0,288	-1,526	-0,181	0,089	0,789	0,789	-1,795	0,520	1,165	0,575	-0,343	0,033	1,651	1,058	0,627
GIBRA	-0,232	-1,030	-0,634	-0,772	1,772	-0,827	-1,095	-0,304	0,118	1,349	1,304	-1,638	0,954	1,294	-0,562	-0,903	-0,573	0,837	0,043	0,899
ROSEF	0,029	-0,837	-0,923	-1,010	1,503	-1,010	-1,010	-0,143	0,378	1,244	0,985	-1,879	0,985	1,244	-0,837	-0,664	-0,316	0,985	0,205	1,071
SWEETEISENBERG	-0,183	-1,822	-0,528	-0,800	1,191	-0,952	-0,843	-0,392	0,358	0,929	1,234	-1,360	1,294	1,825	-0,753	-0,564	-0,156	0,368	0,096	1,055
NNEIG	0,162	-0,284	-0,891	-1,448	1,854	-0,828	-1,383	0,057	-0,143	1,549	0,825	-1,501	0,582	0,733	-0,979	-0,586	-0,156	0,762	0,378	1,297
SWEET	-0,506	-1,768	-1,067	-0,424	0,143	-0,678	-0,443	-0,226	-0,232	1,341	1,341	-1,787	0,475	1,392	0,596	-0,761	-0,487	1,404	0,812	0,876
WOLR810101REVERSE	1,099	-2,502	-0,816	-1,025	0,575	-0,766	-0,901	1,173	-0,913	1,133	1,155	-0,789	0,535	0,654	0,779	-0,055	-0,025	-0,190	-0,227	1,107
COWANWHITTACKER	0,240	-1,196	-0,800	-1,701	0,558	-0,754	-1,546	-0,032	-0,536	1,389	1,366	-1,227	0,822	1,280	0,620	-0,521	-0,241	1,016	0,271	0,993
ROSM880101	-0,600	2,139	1,094	1,414	-0,529	0,914	1,120	-0,456	0,363	-1,104	-1,104	0,858	-0,735	-1,151	-0,832	0,477	0,372	-1,070	-0,246	-0,924
ROSM880102	-0,203	1,895	1,150	0,828	-0,815	1,081	0,924	0,105	0,607	-1,285	-1,285	1,237	-0,663	-1,386	-0,700	0,151	-0,088	-1,211	0,556	-0,898
ROSM880103REVERSE	0,641	0,947	-0,885	-0,580	0,336	-0,275	-2,107	1,863	-1,191	0,641	0,031	0,641	0,947	-0,275	-0,885	0,641	0,641	0,031	-1,801	0,641
SET1	0,662	-2,482	-0,530	-1,870	1,751	-0,452	-0,836	0,880	0,145	0,677	1,267	-0,327	0,400	-0,460	-0,460	0,334	0,271	-0,135	0,249	0,917
SET2	0,850	-1,912	-1,430	-0,956	0,830	-0,511	-1,306	0,682	-0,093	1,075	1,260	-1,121	0,959	-0,740	-0,740	0,666	0,220	0,509	0,541	1,216
SET3	0,977	-2,185	-0,429	-1,437	0,551	-0,472	-0,666	0,648	-0,193	1,082	1,369	-0,898	0,877	-1,123	-1,123	0,563	0,601	0,838	-0,015	1,035

|→

Table S6.2: Median and median absolute deviation (MAD) of feature importance (scale accuracy on training data) in the learning experiment. Each data point represents 19 different training set sizes each evaluated 910 times. Scales are sorted according to feature importance in descending order. Scale no. 1 has highest feature importance, scale no. 91 has lowest feature importance.

NO.	MEDIAN ACCURACY	MAD OF ACCURACY	NO.	MEDIAN ACCURACY	MAD OF ACCURACY	NO.	MEDIAN ACCURACY	MAD OF ACCURACY	NO.	MEDIAN ACCURACY	MAD OF ACCURACY	NO.	MEDIAN ACCURACY	MAD OF ACCURACY
1	0.85	0.02	20	0.82	0.02	39	0.81	0.02	58	0.78	0.02	77	0.70	0.03
2	0.84	0.02	21	0.82	0.02	40	0.81	0.02	59	0.78	0.03	78	0.70	0.03
3	0.84	0.02	22	0.82	0.02	41	0.81	0.02	60	0.78	0.02	79	0.70	0.03
4	0.83	0.02	23	0.82	0.02	42	0.80	0.02	61	0.77	0.02	80	0.68	0.03
5	0.83	0.02	24	0.82	0.02	43	0.80	0.02	62	0.76	0.02	81	0.68	0.03
6	0.83	0.02	25	0.82	0.02	44	0.80	0.02	63	0.76	0.02	82	0.68	0.03
7	0.83	0.02	26	0.82	0.02	45	0.80	0.02	64	0.76	0.03	83	0.67	0.03
8	0.83	0.02	27	0.82	0.02	46	0.80	0.02	65	0.76	0.02	84	0.66	0.03
9	0.82	0.02	28	0.81	0.02	47	0.79	0.02	66	0.76	0.02	85	0.66	0.03
10	0.82	0.02	29	0.81	0.02	48	0.79	0.02	67	0.76	0.03	86	0.66	0.03
11	0.82	0.02	30	0.81	0.02	49	0.79	0.02	68	0.74	0.03	87	0.66	0.03
12	0.82	0.02	31	0.81	0.02	50	0.79	0.03	69	0.74	0.03	88	0.62	0.03
13	0.82	0.02	32	0.81	0.02	51	0.78	0.02	70	0.74	0.03	89	0.58	0.03
14	0.82	0.02	33	0.81	0.02	52	0.78	0.02	71	0.74	0.03	90	0.57	0.03
15	0.82	0.02	34	0.81	0.02	53	0.78	0.02	72	0.73	0.03	91	0.54	0.04
16	0.82	0.02	35	0.81	0.02	54	0.78	0.02	73	0.73	0.02			
17	0.82	0.02	36	0.81	0.02	55	0.78	0.02	74	0.73	0.03			
18	0.82	0.02	37	0.81	0.02	56	0.78	0.02	75	0.71	0.03			
19	0.82	0.02	38	0.81	0.02	57	0.78	0.02	76	0.71	0.02			

S6.4 Principal Component Analysis

The centered and unit-variance scaled hydrophobicity scales were analyzed by principal component analysis using MATLAB. Figure S6.1 shows the explained variance per principal component and as cumulative explained variance over principal components. 68.8% of the variance is already explained by the first component.

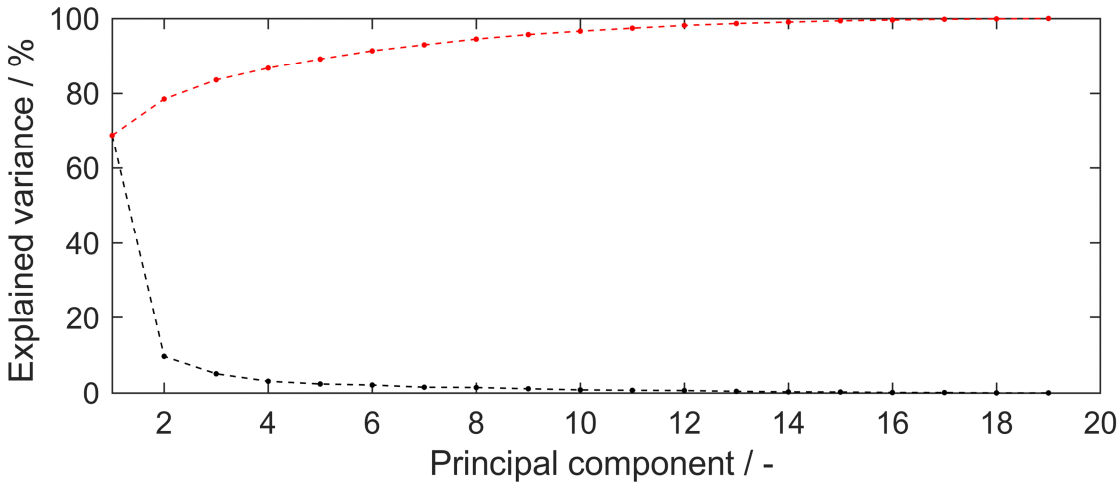


Figure S6.1: Explained variance (%) of principal components derived from 91 normalized hydrophobicity scales. The black dots represent the explained variance per principal component. The red rots represent the cumulative sum of the explained variance. Dashed lines are shown to guide the eye.

S6.5 Location and Interaction of Tryptophan within the HBcAg Dimer

Figure S6.2 shows a 3-D representation of HBcAg dimer backbone, with residues arginine, phenylalanine, tryptophan, and tyrosine shown and colored blue, yellow, red, and green, respectively. The four tryptophans are indicated by one-letter code and marked with their sequence position. All tryptophans are in the vicinity of either arginine, phenylalanine, and/or tyrosine.

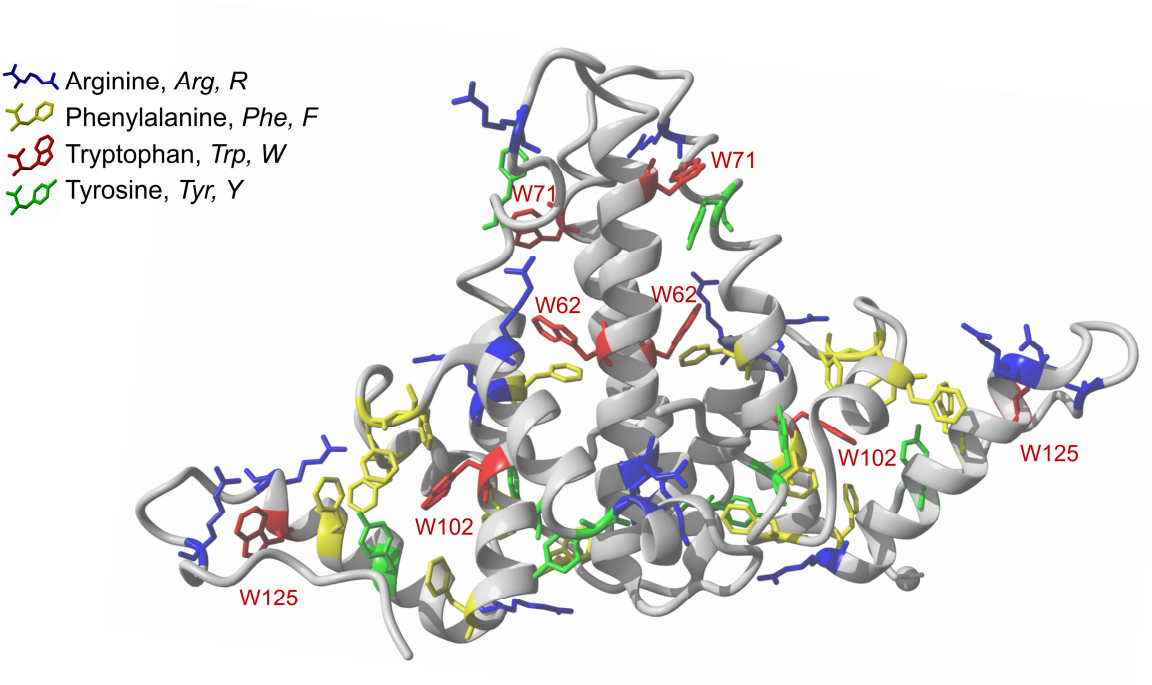


Figure S6.2: 3-D model of the truncated HBcAg crystal structure 4BMG (retrieved from www.rcsb.org, created with YASARA version 18.2.7). Four amino acids are highlighted: arginine, blue; phenylalanine, yellow; tryptophan, red; tyrosine, green. Tryptophan residues are labeled with one-letter code and their sequence positions. Each tryptophan side chain interacts either with the side chains of arginine, phenylalanine or tryptophan.

S6.6 Solubility Data Table

Table S6.3: Binary solubility Data Table: Binary solubility indicating soluble (1) and insoluble (0) observations, shown in a grid of insertion strategies (A-H) over inserts (1-71).

	INSERTION STRATEGY							
INSERT	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**
1	0	1	1	1	1	0	1	0
2	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0
12	0	0	0	1	1	0	1	0
13	0	0	0	1	1	0	1	0
14	1	0	0	1	1	0	1	0
15	1	0	1	1	1	1	1	1
16	0	0	0	0	0	0	0	0

17	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0
25	1	1	1	1	1	1	1	1
26	1	1	1	1	1	1	1	1
27	1	1	1	1	1	1	1	1
28	1	1	1	1	1	1	1	1
29	0	0	0	0	0	0	0	0
30	1	1	1	1	1	1	1	1
31	0	0	0	0	0	0	0	0
32	1	1	1	1	1	1	1	1
33	0	0	0	0	0	0	0	0
34	0	0	0	1	0	0	1	0
35	0	0	0	0	0	0	0	0
36	1	1	1	1	1	1	1	1
37	1	1	1	1	1	1	1	1
38	1	1	1	1	1	1	1	1
39	1	1	1	0	1	1	1	1
40	1	1	1	1	1	1	1	1
41	0	0	0	1	0	0	0	0
42	0	0	0	1	1	0	1	0
43	0	0	0	1	0	0	0	0
44	0	0	0	1	0	0	0	0
45	0	0	1	1	1	1	1	0
46	1	0	0	1	1	0	1	0
47	0	0	0	0	0	0	0	0
48	0	0	0	0	0	0	0	0
49	1	1	1	1	1	0	1	0
50	1	1	1	1	1	1	1	1
51	0	0	0	1	1	1	1	0
52	1	1	1	1	1	1	1	1
53	1	1	1	1	1	1	1	1
54	1	1	1	1	1	1	1	1
55	1	1	1	1	1	1	1	1
56	1	1	1	1	1	1	1	1
57	1	1	0	0	1	0	1	1
58	1	1	1	1	1	1	1	1
59	1	1	1	1	1	0	1	0
60	1	1	1	1	1	0	1	0
61	1	1	1	1	1	1	1	1
62	1	1	1	1	1	0	1	0
63	1	1	1	1	1	1	1	1
64	1	0	1	0	1	0	1	0
65	1	1	1	1	1	1	1	1
66	0	0	0	1	1	0	1	0
67	1	1	1	1	1	1	1	1
68	1	1	1	1	1	0	1	0
69	1	1	1	1	1	0	1	1
70	1	1	1	1	1	1	1	1
71	1	1	1	1	1	1	1	1

Appendix E: Supplementary Material for Chapter 7

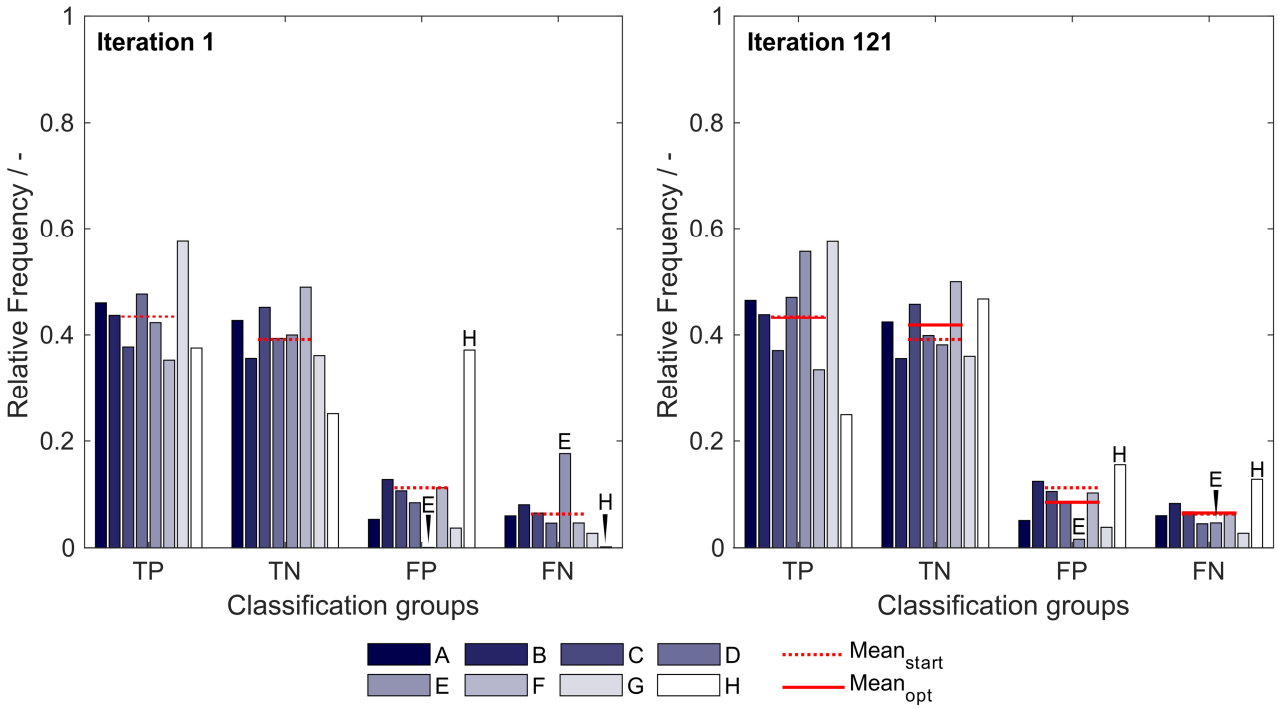


Figure S7.1: Relative frequency of classification groups based on insertion strategies A-H in the first iteration (left) and the best optimization iteration (right) during insertion strategy-based optimization with the 91 optimized literature scales. The mean of the relative frequencies within a classification group is shown for the first iteration ($Mean_{start}$) and for the best optimization ($Mean_{opt}$), indicating that through optimization the FP group decreases in mean relative frequency while the TN group increases in mean relative frequency. Strategy E and H are marked additionally to guide the eye. TP: true positive; TN: true negative; FP: false positive; FN: false negative.

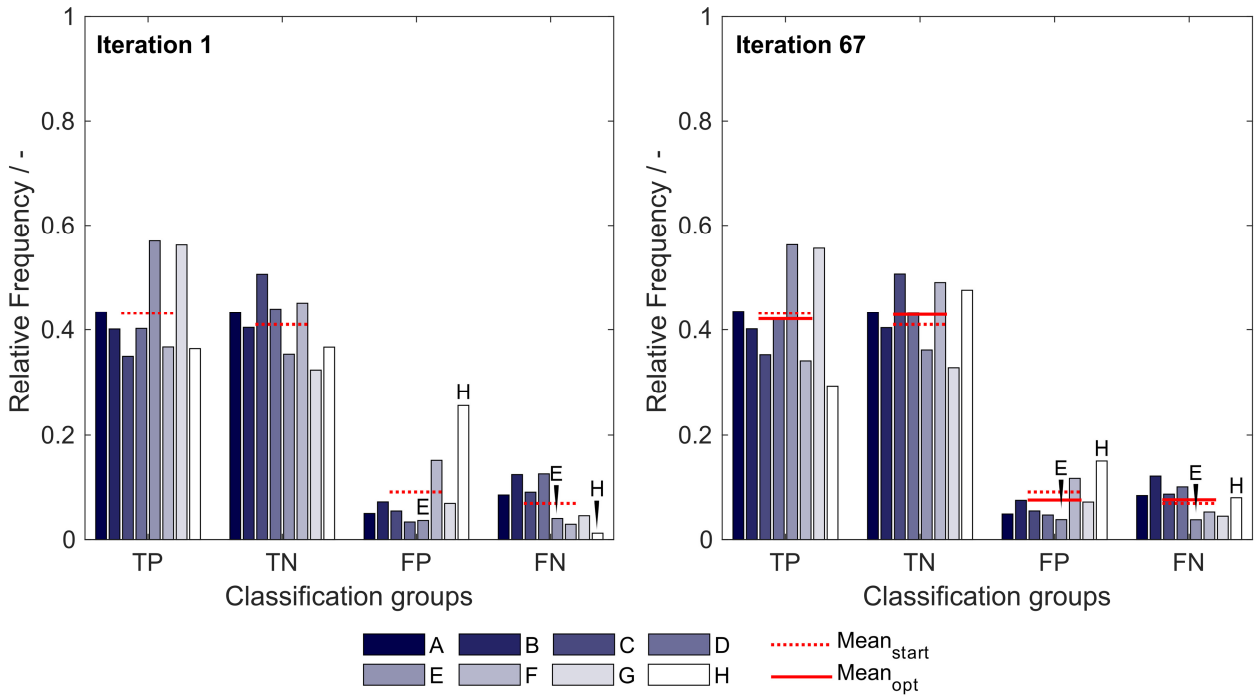


Figure S7.2: Relative frequency of classification groups based on insertion strategies A-H in the first iteration (left) and the best optimization iteration (right) during insertion strategy-based optimization with the synthesized scale set $S_{8,1}$. The mean of the relative frequencies within a classification group is shown for the first iteration ($Mean_{start}$) and for the best optimization ($Mean_{opt}$), indicating that through optimization the FP group decreases in mean relative frequency while the TN group increases in mean relative frequency. Strategy E and H are marked additionally to guide the eye. TP: true positive; TN: true negative; FP: false positive; FN: false negative.

| →

Table S7.1: Centered and unit-variance scaled literature hydrophobicity scales derived from (Simm et al., 2016). For original references of the scales, the reader is referred to the publication of Simm *et al.* (2016). Reversed scales were excluded if there was a complementary, non-reversed scale available, resulting in 91 scales. Each amino acid, represented in single-letter code, is assigned a hydrophobicity value by each hydrophobicity scale.

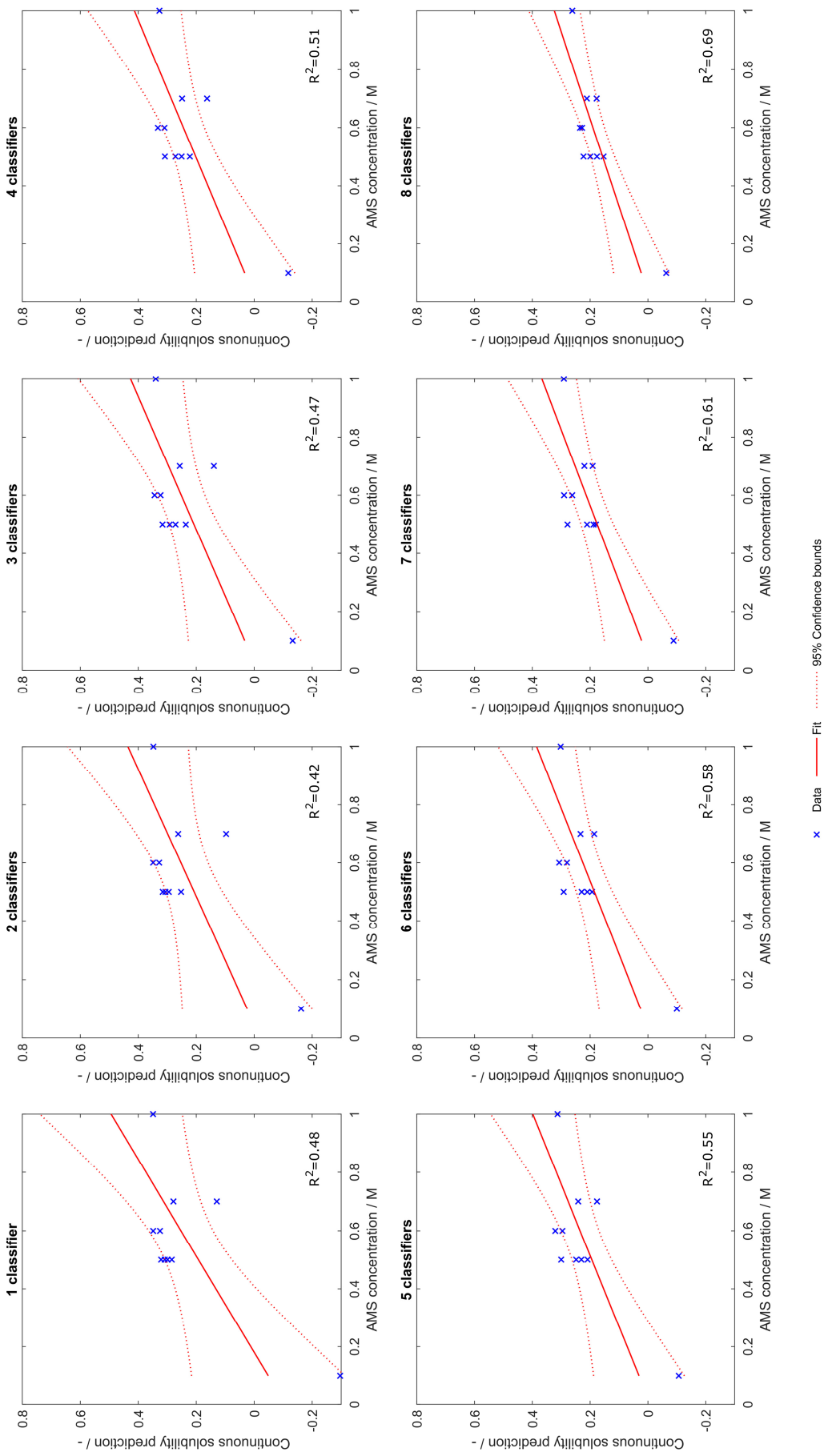
SCALE-ID/AMINO ACID	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
CIDH920101	-0,611	-0,406	-0,367	-1,654	0,598	-1,137	-0,952	-1,147	0,871	0,569	1,086	-0,523	1,164	1,271	-0,289	-1,127	-0,854	1,173	1,280	1,056
CIDH920105	-0,117	-0,548	-0,891	-1,155	0,617	-1,214	-1,253	-0,920	0,118	1,636	0,980	-0,538	0,843	1,185	-0,225	-1,086	-0,891	1,538	0,950	0,970
ESID840101	0,623	-2,534	-0,775	-0,901	0,293	-0,854	-0,744	0,481	-0,398	1,376	1,062	-1,497	0,638	1,188	0,120	-0,178	-0,053	0,811	0,261	1,078
MANP780101	0,060	-0,733	-0,924	-1,285	1,114	-0,708	-0,625	-0,282	-0,454	1,774	1,285	-0,962	0,962	0,714	-0,955	-1,044	-0,752	0,670	0,346	1,800
PONP800101	-0,177	-0,713	-1,045	-1,066	1,620	-0,855	-0,917	-0,360	0,202	1,511	1,057	-1,181	1,213	0,603	-0,917	-0,869	-0,605	0,277	0,508	1,715
PONP800102	-0,172	-0,688	-1,096	-1,090	1,938	-0,777	-0,962	-0,369	-0,025	1,339	0,944	-1,383	1,237	0,701	-0,796	-0,612	-0,516	0,332	0,408	1,587
PONP800103	-0,158	-0,667	-1,176	-1,119	1,905	-0,667	-1,063	-0,384	-0,328	1,113	0,831	-1,600	1,396	0,944	-0,639	-0,243	-0,384	0,463	0,294	1,481
PONP800104	0,699	-1,129	-0,388	-1,360	1,347	-1,113	-0,149	2,018	-0,889	1,455	0,977	-0,604	0,506	1,031	-0,951	-1,144	0,198	-0,527	-0,080	0,105
PONP800105	0,244	-0,802	-1,917	-0,387	1,244	-1,740	-0,533	-0,079	0,821	-0,140	1,698	-0,771	1,498	-0,079	-0,140	-0,710	0,167	-0,294	0,367	1,552
PONP800106	-0,821	-0,553	-0,694	-1,145	1,631	-0,088	-0,088	-0,624	0,165	0,785	0,870	-1,343	2,230	1,011	-0,856	-0,462	-0,920	-0,300	-0,222	1,426
PRAM900101REVERSE	0,607	-2,235	-0,701	-1,600	0,690	-0,560	-1,395	0,485	-0,335	0,915	0,851	-1,517	0,973	1,037	0,241	0,402	0,524	0,666	0,138	0,812
SWER830101	-0,401	-0,591	-0,921	-1,311	0,170	-0,911	-1,221	-0,671	-0,641	1,250	1,220	-0,671	1,020	1,920	-0,491	-0,551	-0,281	0,500	1,670	0,910
NADH010101	0,716	-1,875	-0,900	-0,943	1,337	-1,393	-1,307	-0,022	-0,686	1,241	1,112	-0,162	0,930	1,080	-0,751	-0,269	0,020	0,727	-0,022	1,166
NADH010102	0,590	-1,354	-0,756	-0,696	1,448	-1,195	-1,065	-0,048	-0,467	1,139	1,109	-1,963	0,810	1,159	-0,706	-0,178	0,052	0,770	0,191	1,159
NADH010103	0,357	-1,240	-0,867	-0,580	1,720	-1,187	-1,038	-0,271	-0,452	1,028	1,017	-1,655	0,741	1,284	-0,942	-0,409	0,027	1,007	0,304	1,156
NADH010104	0,196	-1,190	-0,949	-0,469	1,834	-1,190	-0,960	-0,393	-0,426	1,004	0,982	-1,506	0,742	1,288	-1,048	-0,524	0,065	1,135	0,327	1,080
NADH010106	-0,232	-0,946	-1,176	0,228	2,288	-1,061	-0,382	-0,831	-0,865	0,665	0,654	-0,727	0,665	1,057	-1,475	-0,762	0,619	1,206	0,021	1,057
NADH010107	-0,434	-0,754	-1,213	1,617	2,281	-0,721	0,542	-0,959	-0,992	-0,188	-0,122	0,526	0,091	0,567	-1,500	-0,844	1,010	1,051	-0,475	0,518
WILM950101	-0,390	-0,910	-0,281	-0,538	-0,143	-0,246	-0,481	-0,304	-1,706	1,568	1,579	-1,351	-0,304	2,324	-0,017	-0,779	-0,052	0,887	0,658	0,486
WILM950102	0,475	0,069	-0,686	-1,155	-0,089	-0,811	-0,441	-0,650	-0,528	1,001	1,655	-1,137	-1,238	2,422	-0,343	-0,722	0,233	1,457	0,108	0,380
WILM950103REVERSE	1,008	1,541	0,206	0,248	-2,545	0,488	0,092	1,076	-1,853	-0,505	0,206	1,242	-0,908	0,917	-0,538	-0,041	-0,275	-0,372	-0,294	0,307
WILM950104	-1,374	0,196	0,678	-0,633	1,563	-0,378	0,077	-0,864	-1,637	2,452	-0,007	0,180	-0,194	0,583	-0,501	0,328	-0,366	0,989	-1,470	0,379
KUHL950101REVERSE	0,350	-2,003	-0,885	-1,326	1,026	-0,856	-1,621	0,644	-0,268	1,262	0,997	-0,591	0,703	1,262	0,615	-0,297	-0,444	0,585	-0,297	1,144
JURD980101	0,547	-1,503	-0,974	-1,007	1,010	-1,033	-0,874	-0,028	-0,874	1,671	1,439	-1,175	0,811	1,109	-0,445	0,018	-0,048	0,031	-0,246	1,572
KIDA850101REVERSE	0,263	-1,822	-0,789	-0,789	1,023	-1,072	-1,140	0,156	-0,273	0,750	1,072	-1,656	0,711	1,393	0,731	-0,409	-0,614	1,530	0,546	0,390
ENGD860101REVERSE	0,607	-2,234	-0,701	-1,600	0,689	-0,558	-1,396	0,484	-0,333	0,914	0,852	-1,519	0,975	1,036	0,239	0,403	0,525	0,668	0,137	0,811
KYTJ820101	0,767	-1,342	-1,008	-1,008	1,001	-1,008	-1,008	0,030	-0,907	1,671	1,436	-1,142	0,800	1,101	-0,372	-0,104	-0,070	-0,137	-0,271	1,570
ENGEL	0,607	-2,234	-0,701	-1,600	0,689	-0,558	-1,396	0,484	-0,333	0,914	0,852	-1,519	0,975	1,036	0,239	0,403	0,525	0,668	0,137	0,811
JONES	-0,468	-0,485	-1,131	-0,647	0,087	-1,209	-0,637	-1,124	-0,468	1,478	0,642	0,189	0,216	1,238	1,153	-1,148	-1,148	2,006	1,069	0,385
CIDBB	-0,358	-0,368	-0,963	-0,973	0,462	-0,671	-1,559	-1,100	0,140	1,077	0,930	-0,368	0,960	1,213	-0,290	-1,188	-0,856	1,916	1,213	0,784
CIDA	0,233	-0,623	-0,993	-1,178	0,564	-1,139	-0,925	-0,915	-0,273	1,994	1,031	-0,662	1,060	0,865	-0,176	-0,701	-1,285	1,157	0,904	1,060
CIDAB	0,090	-0,759	-0,954	-1,100	1,134	-1,247	-1,237	-0,632	-0,320	1,934	0,860	-0,681	0,509	1,182	-0,281	-0,886	-0,681	1,397	0,568	1,104
PONG1	0,204	-0,283	-0,792	-1,257	1,141	-0,747	-1,025	-0,433	-0,268	1,920	1,058	-1,324	0,631	0,833	-1,287	-0,822	-0,493	0,886	0,594	1,463
PONG2	0,376	-0,253	0,124	-1,498	1,695	-0,527	-2,068	0,243	0,791	0,799	-0,061	-1,476	-0,039	-0,135	-0,379	0,710	-0,646	2,117	0,036	0,191
PONG3	0,321	-0,300	-0,366	-1,534	1,580	-0,706	-1,724	-0,101	0,297	1,506	0,553	-1,559	0,330	0,388	-0,921	-0,060	-0,631	1,671	0,346	0,909
KIDER	0,263	-1,822	-0,789	-0,789	1,023	-1,072	-1,140	0,156	-0,273	0,750	1,072	-1,656	0,711	1,393	0,731	-0,409	-0,614	1,530	0,546	0,390
WOLR790101	1,119	-2,548	-0,830	-0,830	0,589	-0,780	-0,920	1,198	-0,930	1,159	1,178	-0,800	0,549	0,669	0,539	-0,050	-0,020	-0,190	-0,230	1,129
CHOTA	-1,264	1,253	-0,235	-0,463	-0,807	0,223	0,452	-2,180	0,566	0,109	-0,006	0,681	0,338	0,910	-0,578	-1,264	-0,692	1,940	1,367	-0,349
ROSEB	-1,107	0,683	-0,726	-0,845	-0,033	-0,344	-0,463	-1,680	0,539	0,587	0,730	-0,415	0,945	1,446	-0,964	-1,131	-0,654	2,186	1,064	0,181
ROSEAREVERSE	1,385	-1,608	0,344	0,496	0,777	-0,242	-0,090	2,035	-0,437	0,018	-0,242	-0,957	-0,459	-0,892	0,756	1,124	0,647	-1,825	-1,196	0,365
COHEN	-0,874	-0,665	-1,607	-1,189	-0,456	-0,874	-0,665	-1,189	0,382	0,696	0,801	0,487	1,534	1,220	1,115	-0,979	-0,246	1,744	0,487	0,278
JACWH	0,626	-3,164	-0,379	-1,099	0,687	-0,325	-0,922	0,565	-0,501	0,755	0,782	-1,214	0,803	0,844	0,368	0,232	0,273	0,497	0,450	0,721
CASSI	0,198	-0,671	-0,478	-1,347	1,839	-1,057	-1,250	-0,092	0,391	1,356	0,487	-1,540	0,487	0,970	-0,960	-0,671	-0,381	1,549	0,487	0,681
MIDK0	0,673	-1,499	-0,973	-1,006	1,002	-1,032	-0,874	-0,097	-0,874	1,660	1,430	-1,173	0,804	1,101	-0,446	0,015	-0,051	0,028	-0,249	1,561

MDK1	0,746	-1,581	-0,949	-0,949	0,846	-1,248	-1,049	0,081	-0,750	1,710	1,378	-1,182	0,846	1,046	-0,317	-0,052	0,115	0,115	-0,218	1,411
BULDG	0,737	0,823	1,037	0,737	0,468	1,123	0,629	0,951	0,823	-1,475	-1,690	0,576	-0,627	-1,551	-0,101	0,533	0,393	-1,207	-1,454	-0,724
GUYH850101	0,175	1,832	0,523	0,798	-1,216	0,953	0,844	0,386	-0,374	-0,951	-0,996	1,365	-1,372	-1,857	0,752	0,560	0,148	-0,383	-0,108	-1,079
MIYS850101	-0,166	-0,704	-0,923	-0,979	1,050	-0,849	-0,951	-0,563	-0,274	1,470	1,302	-1,279	1,527	1,564	-0,849	-0,746	-0,559	0,923	0,096	0,909
WILSON	-0,631	-0,848	-0,604	-0,929	1,157	-0,604	-0,550	-0,225	-0,902	0,615	1,238	-1,525	0,127	1,482	0,046	-0,712	-1,146	1,590	1,373	1,048
CHOC760103	0,586	-1,490	-0,873	-0,704	1,260	-1,153	-0,536	0,474	-0,592	1,821	0,979	-1,378	0,699	1,260	-0,536	-0,311	-0,255	-0,031	-0,704	1,484
EISEN	0,620	-2,531	-0,780	-0,900	0,290	-0,850	-0,740	0,480	-0,400	1,380	1,060	-1,500	0,640	1,190	0,120	-0,180	-0,050	0,810	0,260	1,080
JANJ790102	0,589	-1,824	-0,547	-0,689	1,441	-0,831	-0,831	0,589	0,021	1,157	0,873	-2,392	0,731	0,873	-0,263	0,021	-0,121	0,589	-0,405	1,015
RAOARGOS	0,954	-1,390	-1,041	-1,467	0,780	-1,041	-1,196	0,431	-0,363	1,109	1,167	-1,506	1,070	1,361	-0,634	0,199	0,412	0,257	-0,073	0,973
NOZY710101TANFORD	0,491	-2,485	-0,832	-0,180	0,180	-0,898	-0,794	0,359	-0,472	1,209	1,351	-1,512	0,510	1,030	0,019	-0,265	-0,142	0,671	0,151	1,606
WELLING	0,946	0,585	-0,270	0,629	-0,542	0,148	-0,232	-0,947	2,193	-1,631	0,693	1,522	-2,219	-0,675	-0,118	0,053	-0,067	-0,504	0,300	0,135
PARJ860101	0,177	0,508	0,951	1,424	0,066	0,793	1,077	0,745	0,177	-1,418	-1,608	0,745	-0,818	-1,608	0,177	0,872	0,666	-1,734	-0,455	-0,739
ROSG850102	0,030	-0,837	-0,924	-1,010	1,505	-1,010	-1,010	-0,143	0,377	1,244	0,984	-1,877	0,984	1,244	-0,837	-0,663	-0,317	0,984	0,204	1,071
BISHOP	0,200	0,321	0,442	1,411	-0,406	0,684	2,381	-0,042	0,200	-0,406	-0,769	0,563	-0,285	-1,375	0,442	0,079	0,079	-2,344	-1,254	0,079
WIMLEYREVERSE	-0,372	-1,499	-0,673	-0,312	0,075	-0,604	-0,037	-0,931	-0,037	1,021	1,133	-2,350	0,634	1,529	-0,062	-0,338	-0,157	1,855	0,669	0,454
WIMLEY	-0,086	-0,832	-0,377	-1,322	0,393	-0,564	-2,244	0,101	-0,086	0,474	0,766	-1,042	0,381	1,431	-0,412	-0,039	-0,051	2,271	1,209	0,031
ARGP820101	-0,469	-0,481	-1,135	-0,651	0,088	-1,208	-0,639	-1,123	-0,469	1,480	0,645	0,185	0,221	1,238	1,153	-1,147	-1,147	2,001	1,068	0,390
FAUJ830101	-0,167	-1,449	-1,051	-1,216	1,027	-0,682	-1,089	-0,468	-0,342	1,280	1,183	-1,429	0,726	1,270	0,231	-0,507	-0,216	1,717	0,464	0,717
JOND750101	-0,467	-0,484	-1,132	-0,646	0,087	-1,209	-0,637	-1,123	-0,467	1,477	0,642	0,190	0,215	1,239	1,153	-1,149	-1,149	2,006	1,068	0,386
LEVM760101REVERSE	0,089	-1,795	-0,288	-1,526	0,358	-0,288	-1,526	-0,180	0,089	0,789	0,789	-1,795	0,519	1,165	0,573	-0,342	0,035	1,650	1,058	0,627
ZIMJ680101	-0,427	-0,427	-1,137	-0,609	0,196	-1,223	-0,600	-1,127	-0,168	1,720	1,193	0,311	0,119	1,414	1,366	-1,089	-0,705	-0,926	1,625	0,493
NADH010105	0,012	-1,206	-1,171	-0,272	2,022	-1,301	-0,946	-0,603	-0,639	0,934	0,946	-0,378	0,792	1,277	-1,277	-0,698	0,130	1,100	0,237	1,041
PONP930101	0,204	-0,283	-0,792	-1,257	1,141	-0,747	-1,025	-0,433	-0,268	1,920	1,058	-1,324	0,631	0,833	-1,287	-0,822	-0,493	0,886	0,594	1,463
COWR900101	0,238	-1,310	-0,896	-0,489	0,566	-0,841	-0,380	-0,091	-1,873	1,324	1,317	-1,678	0,832	1,270	0,582	-0,591	-0,294	1,051	0,308	0,957
BLAS910101	0,276	-1,563	-0,851	-1,474	0,454	-0,821	-1,444	-0,080	-1,059	1,225	1,225	-0,732	0,632	1,403	0,543	-0,495	-0,228	1,047	1,047	0,898
FASG890101	0,232	1,030	0,634	0,772	-1,772	0,827	1,095	0,304	-0,118	-1,349	-1,304	1,638	-0,954	-1,294	0,562	0,903	0,573	-0,837	-0,043	-0,899
FAUCH	-0,168	-1,449	-1,050	-1,216	1,028	-0,683	-1,089	-0,468	-0,341	1,279	1,184	-1,428	0,725	1,268	0,231	-0,507	-0,214	1,717	0,464	0,718
PONNU	-0,176	-0,712	-1,046	-1,066	1,621	-0,855	-0,916	-0,360	0,202	1,512	1,058	-1,182	1,212	0,601	-0,916	-0,868	-0,606	0,277	0,506	1,713
JANIN	0,245	-1,091	-0,839	-0,839	2,661	-0,924	-0,924	0,327	-0,505	1,410	0,828	-1,132	0,408	0,661	-0,672	-0,505	-0,590	0,160	0,078	1,243
GUYFE	-0,219	-1,057	-0,670	-1,015	1,308	-0,940	-0,983	-0,595	0,371	1,255	1,145	-1,423	1,209	1,653	-0,908	-0,691	-0,446	0,790	0,200	1,017
CHOTH	0,585	-1,489	-0,873	-0,705	1,259	-1,154	-0,534	0,475	-0,591	1,822	0,981	-1,379	0,699	1,259	-0,534	-0,309	-0,256	-0,032	-0,705	1,483
VHEG790101	0,731	-2,592	-0,323	-1,555	-0,303	-0,187	-1,139	0,460	-0,455	1,137	1,104	-0,676	0,526	1,375	-0,425	0,050	0,219	1,002	0,050	1,002
ROSEM	0,224	-1,909	-1,151	-0,824	0,603	-1,081	-0,922	-0,090	-0,599	1,323	1,323	-1,240	0,692	1,323	0,729	-0,136	0,107	1,248	-0,548	0,930
LEVIT	0,089	-1,795	-0,288	-1,526	0,358	-0,288	-1,526	-0,181	0,089	0,789	0,789	-1,795	0,520	1,165	0,575	-0,343	0,033	1,651	1,058	0,627
GIBRA	-0,232	-1,030	-0,634	-0,772	1,772	-0,827	-1,095	-0,304	0,118	1,349	1,304	-1,638	0,954	1,294	-0,562	-0,903	-0,573	0,837	0,043	0,899
ROSEF	0,029	-0,837	-0,923	-1,010	1,503	-1,010	-1,010	-0,143	0,378	1,244	0,985	-1,879	0,985	1,244	-0,837	-0,664	-0,316	0,985	0,205	1,071
SWEETEISENBERG	-0,183	-1,822	-0,528	-0,800	1,191	-0,952	-0,843	-0,392	0,358	0,929	1,234	-1,360	1,294	1,825	-0,753	-0,564	-0,156	0,368	0,096	1,055
NNEIG	0,162	-0,284	-0,891	-1,448	1,854	-0,828	-1,383	0,057	-0,143	1,549	0,825	-1,501	0,582	0,733	-0,979	-0,586	-0,156	0,762	0,378	1,297
SWEET	-0,506	-1,768	-1,067	-0,424	0,143	-0,678	-0,443	-0,226	-0,232	1,341	1,341	-1,787	0,475	1,392	0,596	-0,761	-0,487	1,404	0,812	0,876
WOLR810101REVERSE	1,099	-2,502	-0,816	-1,025	0,575	-0,766	-0,901	1,173	-0,913	1,133	1,155	-0,789	0,535	0,654	0,779	-0,055	-0,025	-0,190	-0,227	1,107
COWANWHITTACKER	0,240	-1,196	-0,800	-1,701	0,558	-0,754	-1,546	-0,032	-0,536	1,389	1,366	-1,227	0,822	1,280	0,620	-0,521	-0,241	1,016	0,271	0,993
ROSM880101	-0,600	2,139	1,094	1,414	-0,529	0,914	1,120	-0,456	0,363	-1,104	-1,104	0,858	-0,735	-1,151	-0,832	0,477	0,372	-1,070	-0,246	-0,924
ROSM880102	-0,203	1,895	1,150	0,828	-0,815	1,081	0,924	0,105	0,607	-1,285	-1,285	1,237	-0,663	-1,386	-0,700	0,151	-0,088	-1,211	0,556	-0,898
ROSM880103REVERSE	0,641	0,947	-0,885	-0,580	0,336	-0,275	-2,107	1,863	-1,191	0,641	0,031	0,641	0,947	-0,275	-0,885	0,641	0,641	0,031	-1,801	0,641
SET1	0,662	-2,482	-0,530	-1,870	1,751	-0,452	-0,836	0,880	0,145	0,677	1,267	-0,327	0,400	-0,460	-0,460	0,334	0,271	-0,135	0,249	0,917
SET2	0,850	-1,912	-1,430	-0,956	0,830	-0,511	-1,306	0,682	-0,093	1,075	1,260	-1,121	0,959	-0,740	-0,740	0,666	0,220	0,509	0,541	1,216
SET3	0,977	-2,185	-0,429	-1,437	0,551	-0,472	-0,666	0,648	-0,193	1,082	1,369	-0,898	0,877	-1,123	-1,123	0,563	0,601	0,838	-0,015	1,035

|→

Table S7.2: The amino acid hydrophobicity value of the 25 best individual synthesized scales, selected by feature selection in 1000-fold Monte-Carlo cross-validation. Scale notation $S_{x,y\text{-}z}$ shows from which number of subsets x the number x of scales was synthesized, where y is the repetition number (ranging from 1 to 20), defining a specific scale set. Within this scale set, z is the number of the classifier, ranging from 1 to x.

SCALE-ID/ AMINO ACID	MCC OF VALIDATION	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
$S_{2,1-2}$	**0,84**	0,148	0,400	-0,104	-0,657	2,115	-0,091	-0,470	0,830	-2,660	-0,219	-0,155	0,232	-1,232	1,214	0,646	0,059	-0,772	1,143	-0,648	0,220
$S_{1,7-1}$	**0,82**	0,027	0,163	0,301	-0,158	0,279	0,043	-0,277	0,923	-1,923	-0,206	-0,218	0,095	-2,435	-0,883	0,680	-0,564	1,507	1,998	0,535	0,114
$S_{1,8-1}$	**0,81**	-0,281	0,113	-0,647	-0,565	-0,603	0,581	-0,331	0,512	-1,171	-0,447	0,046	0,244	-2,074	2,504	0,722	0,383	-0,305	1,877	-0,757	0,198
$S_{8,12-6}$	**0,81**	0,657	-1,397	-0,051	0,624	0,095	-0,197	-0,750	2,613	-0,261	-0,173	-0,812	0,600	-0,747	0,681	0,557	-0,095	-1,883	1,101	-1,122	0,560
$S_{1,5-1}$	**0,81**	-0,320	0,168	-0,668	-0,751	0,182	0,410	-0,345	0,422	-0,552	-0,321	0,090	0,383	-0,730	3,038	0,747	0,645	-1,853	0,604	-1,469	0,320
$S_{1,14-1}$	**0,81**	-0,074	0,191	0,470	-0,274	2,543	-0,129	-0,294	0,670	-1,939	-0,165	-0,405	0,066	-2,105	-0,553	0,634	-0,118	-0,070	1,573	-0,177	0,158
$S_{2,15-1}$	**0,80**	0,570	-0,443	-0,272	0,587	-1,147	-0,262	-0,712	0,545	-1,100	-0,113	-0,922	0,528	0,210	1,112	0,703	-0,595	-0,628	2,784	-1,624	0,780
$S_{6,19-2}$	**0,80**	0,967	-0,347	1,165	-0,541	-0,280	-0,287	-0,838	1,725	-2,662	-0,572	-0,457	1,007	0,200	0,327	0,105	0,004	-1,150	1,440	-0,253	0,447
$S_{1,12-1}$	**0,80**	-0,302	0,369	-0,503	-0,698	1,481	0,282	-0,481	-1,267	-1,397	-0,180	-0,063	0,081	-1,487	1,779	1,217	0,222	0,402	1,498	-1,516	0,564
$S_{2,8-2}$	**0,80**	0,408	0,488	0,493	-0,548	0,141	-0,442	-0,580	3,006	-1,825	-0,082	-0,055	-0,070	-0,574	-0,601	1,222	-0,595	-1,191	0,877	-0,476	0,404
$S_{3,18-3}$	**0,79**	0,300	-0,627	0,533	0,410	0,161	-0,091	-0,559	2,035	-1,790	-0,263	-1,767	0,785	-0,843	-0,813	1,442	-0,817	-0,271	1,449	0,306	0,423
$S_{4,13-2}$	**0,79**	1,336	0,377	0,103	-0,003	0,060	-1,088	-0,682	1,376	-1,229	-0,457	-0,560	-0,095	0,254	1,494	0,113	-0,983	-1,474	1,875	-1,308	0,891
$S_{2,20-1}$	**0,79**	0,370	0,169	-0,040	-0,880	2,095	-0,042	-0,449	1,274	-2,804	0,395	0,086	0,414	0,489	0,103	0,525	-0,295	-1,674	-0,183	0,007	0,440
$S_{1,9-1}$	**0,79**	0,201	0,189	0,479	0,046	-0,679	-0,093	-0,250	3,421	-1,334	-0,077	-0,140	0,124	-1,054	-1,011	0,401	-0,692	-0,939	0,759	0,633	0,018
$S_{2,17-1}$	**0,79**	1,038	-0,054	-0,341	-1,319	0,792	-0,521	-0,698	-0,974	0,076	0,037	-1,118	0,627	-0,903	-0,769	1,175	-0,637	-0,113	2,891	0,100	0,712
$S_{1,18-1}$	**0,79**	0,108	0,354	-0,524	-0,468	-0,488	0,286	-0,222	-0,249	-2,629	-0,295	0,284	-0,104	0,672	3,077	-0,034	0,209	-0,137	0,654	-0,647	0,154
$S_{1,13-1}$	**0,79**	0,001	-0,165	-0,556	-0,970	0,570	0,112	-0,548	-1,051	0,395	-0,234	-0,901	0,438	-2,447	0,197	1,376	-0,189	0,745	2,412	0,176	0,640
$S_{2,5-2}$	**0,78**	-0,266	-0,015	0,315	-0,009	1,949	0,221	-0,185	-0,571	-2,571	-0,234	-0,229	0,316	-1,906	0,156	0,320	0,088	1,037	1,709	-0,189	0,063
$S_{2,13-2}$	**0,78**	-0,365	-0,036	0,093	-1,141	0,104	0,476	-0,303	-1,614	-0,574	-0,438	-0,180	0,108	-2,218	1,182	0,474	0,249	1,151	2,437	0,380	0,214
$S_{7,5-1}$	**0,78**	1,121	0,704	1,564	-1,824	-1,278	-0,079	-1,114	1,141	-0,782	-0,716	-0,989	-0,736	0,297	0,832	-0,114	-0,053	-0,897	1,260	0,634	1,028
$S_{3,3-3}$	**0,78**	-0,281	0,201	0,217	-0,555	2,344	0,009	-0,318	-0,190	-0,891	-0,316	-0,478	0,018	-2,212	-0,294	0,980	0,385	-0,334	2,216	-0,766	0,267
$S_{3,15-2}$	**0,78**	-0,029	0,459	0,349	-0,160	-1,829	0,308	-0,221	2,455	-1,405	-0,324	-0,090	-0,197	-1,302	0,704	0,524	-0,515	-0,948	1,812	0,154	0,256
$S_{8,5-3}$	**0,78**	0,970	0,053	-0,031	-0,607	-0,003	-1,245	-0,529	1,478	0,793	-0,399	-0,676	-0,996	-0,856	-2,010	0,938	-0,880	0,625	1,863	0,688	0,821
$S_{3,5-1}$	**0,78**	0,156	0,154	0,447	-0,863	-0,083	-0,167	-0,319	-1,004	-0,892	-0,375	-0,431	-0,316	-0,948	-0,226	1,095	0,158	0,172	3,605	-0,543	0,380
$S_{3,5-3}$	**0,78**	0,542	0,497	0,761	-0,518	1,459	-0,154	-0,865	1,831	-2,116	0,126	-0,539	0,098	0,724	-0,287	0,775	-0,152	-1,958	0,033	-0,985	0,730


1 classifier
2 classifiers
3 classifiers
4 classifiers
5 classifiers
6 classifiers
7 classifiers
8 classifiers
R2=0.48
R2=0.42
R2=0.47
R2=0.51
R2=0.55
R2=0.58
R2=0.61
R2=0.69
Continuous solubility prediction / -
AMS concentration / M
Data
Fit
95% Confidence bounds

|←

Figure S7.3: Relationship between continuous solubility prediction value and optimal ammonium sulfate concentration for precipitation of ten constructs. In eight models, 1-8 scales were used, which were generated with a scale table optimization procedure (Set $S_{8,1}$). 95% confidence bounds and R^2 indicate goodness of fit.