Chromatographic host cell protein removal in biopharmaceutical purification

Disela, R.C.

DOI
10.4233/uuid:5d0ce064-f5de-458f-92e7-7cd314ea9ae2

Publication date
2025

Document Version
Final published version

Citation (APA)
Disela, R. C. (2025). *Chromatographic host cell protein removal in biopharmaceutical purification*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:5d0ce064-f5de-458f-92e7-7cd314ea9ae2

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

CHROMATOGRAPHIC HOST CELL PROTEIN REMOVAL IN BIOPHARMACEUTICAL PURIFICATION

PhD thesis

Roxana C. Disela

Chromatographic host cell protein removal in biopharmaceutical purification

Chromatographic host cell protein removal in biopharmaceutical purification

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus,
Prof. dr. ir. T.H.J.J. van der Hagen,
chair of the Board for Doctorates
to be defended publicly on
Friday 24th, January 2025 at 10 o'clock

by

Roxana Clarissa DISELA,

Master of Science in Bioengineering,

Karlsruhe Institute of Technology, Germany

born in Basel, Switzerland

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus	Chairperson
Prof.dr.ir. M. Ottens	Delf University of Technology, promotor
Dr. M. Pabst	Delf University of Technology, promotor

Independent members:

Prof.dr. D. Bracewell	University College London, United Kingdom
Prof.dr.ing. M.H.M. Eppink	Delft University of Technology
Prof.dr. F. Hollmann	Delft University of Technology
Prof.dr. J.J. Hubbuch	Karlsruhe Institute of Technology, Germany
Prof.dr.ir. L.A.M. van der Wielen	Delft University of Technology

This research was funded by GlaxoSmithKline Biologicals S.A. under cooperative research and development agreement between GlaxoSmithKline Biologicals S.A. (Belgium) and the Technical University of Delft (The Netherlands).

Printed by Proefschriftspecialist
Cover illustration by Vanessa Disela

ISBN: 978-94-6384-665-3
An electronic version of this dissertation is available at
https://repository.tudelft.nl

Table of Contents

Summary

The COVID-19 pandemic stressed the need for accelerating the development of novel vaccines. Over the past decades, the bottleneck in the biopharmaceutical process development shifted from optimizing fermentation processes to developing suitable purification strategies. Thereby, improving process understanding can significantly accelerate the development of purification processes. Unlike other biopharmaceutical products, vaccines are often more complex products containing molecules from different origins. The manufacturing process is therefore also more demanding. Consequently, no platform process is available for protein subunit vaccine purification. In the case of expressing a novel antigen in a host cell system, knowledge of the possible impurities – in this case host cell proteins (HCPs) – allows for rational and systematic process development. Therefore, this thesis focuses on developing characterization strategies of recurrent HCP impurities (from *E. coli* host cells) and on the integration of this information into modeling tools that advance removal strategies of these proteins, with a focus on protein-based antigen vaccines.

Firstly, the complete host cell proteome from antigen expressing *E. coli* host cells (BLR(DE3) and HMS174(DE3)) were characterized in **chapter 2**. Around 2000 HCPs were identified from the *E. coli* harvest sample using mass spectrometry based proteomics. Furthermore, an extensive HCP database including their expression levels, and physicochemical properties was constructed. Additionally, the profiles of an antigen expressing and null plasmid strain were compared. From a downstream processing perspective, the differences may be minor and the findings from the BLR(DE3) null strain can be applied to determine a purification strategy for the BLR(DE3) antigen-producing strain and HMS174(DE3) strain. The dataset of identified proteins was connected to databases describing the physicochemical properties of HCPs. Finally, protein property maps that help to identify a suitable downstream processing (DSP) strategy in comparison with the physicochemical properties of the target antigen, were generated.

Preparative chromatography based on differences in physicochemical properties is one of the main techniques for purification of vaccines. As follow-up to chapter 2, an experimental retention map of the host cell

proteome during a salt gradient on hydrophobic interaction chromatography (HIC) and ion exchange chromatography (IEX) was constructed and reported and described in **chapter 3**. Furthermore, this study identified patterns in the retention behavior of HCPs based on their protein-protein interactions, molecular function, and cell location. To be able to predict the retention behavior of yet uncharacterized proteins, a quantitative structure-property relationship (QSPR) model was constructed using IEX retention data. Subsets of proteins, identified according to retention patterns, were used to build additional QSPR models, with monomer subsets yielding the most accurate predictions.

To achieve a higher level of process understanding, mechanistic models (MM) of chromatography columns are used in process development. These models primarily describe behavior of the target protein and selected process- or product-related impurities. However, it is beneficial to also include recurring HCP impurities in MMs. Hereby, critical HCPs causing issues when remaining in the product, are not necessarily abundant in the cell lysate and are often not individually described. A method for determining binding parameters of the entire host cell proteome including low abundant proteins to selected chromatography resins is still lacking.

Chapter 4 introduces a method to determine the above mentioned isotherm parameters of individual HCPs in a comprehensive manner. Fractions obtained from linear gradient elution experiments with different gradient lengths are analyzed by shotgun proteomics in order to extract the retention times of the individual HCPs. From the extracted retention volumes per gradient, isotherm parameters for all individual HCPs detected in the harvest were regressed. This method was exemplified using the BLR *E. coli* harvest, validated, and subsequently employed to optimize a capture step in silico.

Finally, **chapter 5** gives an overview of the additionally investigated high-throughput sample preparation and analysis methods. This involved packing filter plates with resin for batch adsorption, which was explored to determine isotherm parameters instead of low gradient elution (LGE) experiments. Additionally, ion exchange high-performance liquid chromatography (IEX-HPLC) was investigated as an analytical technique instead of mass spectrometry (MS).

In summary, this thesis presents a comprehensive, large-scale characterization of HCPs from widely employed *E. coli* host cell strains for

the production of protein vaccines. Moreover, a validated approach to determine isotherm parameters of all detectable HCPs in the harvest sample is presented.

Samenvatting

De COVID-19 pandemie heeft duidelijk gemaakt dat de ontwikkeling van nieuwe vaccins moet worden versneld. In de afgelopen decennia is het knelpunt in de ontwikkeling van biofarmaceutische processen verschoven van de optimalisatie van fermentatieprocessen naar de ontwikkeling van geschikte zuiveringsstrategieën. Een beter begrip van het proces kan de ontwikkeling van zuiveringsstrategieën aanzienlijk versnellen. In tegenstelling tot andere biofarmaceutische producten zijn eiwitvaccins vaak complexer om te produceren omdat ze moleculen van verschillende oorsprong bevatten. Als gevolg daarvan is er geen standaardproces voor de zuivering van eiwitsubunits voor vaccins. Wanneer een nieuw antigeen tot expressie wordt gebracht in de gastheercellen, maakt kennis van de mogelijke verontreinigingen - in dit geval bijvoorbeeld gastheerceleiwitten (*host cell proteins* - HCPs) - een rationele en systematische procesontwikkeling mogelijk. Daarom richt dit werk zich op de ontwikkeling van strategieën om terugkerende HCP-verontreinigingen (van E. coli-gastheercellen) te karakteriseren. Daarnaast wordt deze informatie gebruikt in een chromatografiemodel om strategieën te ontwikkelen om deze eiwitten *in silico* te verwijderen. De focus van dit proefschrift ligt op eiwitgebaseerde antigeenvaccins.

Eerst werd het volledige proteoom van de gastheercel van *E. coli* (BLR(DE3) en HMS174(DE3)) gekarakteriseerd in **hoofdstuk 2**. Ongeveer 2000 HCPs werden geïdentificeerd uit het *E. coli*-monster met behulp van massaspectrometrie. Daarnaast werd een uitgebreide HCP-database gemaakt met hun expressieniveaus en fysisch-chemische eigenschappen. Bovendien werden de profielen van een antigeen-producerende en een nulplasmide stam vergeleken. Vanuit een zuiveringsperspectief zijn de verschillen klein en de bevindingen van de BLR(DE3) null plasmid stam kunnen gebruikt worden om een zuiveringsstrategie te definiëren voor de BLR(DE3) antigeenproducerende stam en de HMS174(DE3) stam. De dataset van geïdentificeerde eiwitten werd gekoppeld aan databases die de fysisch-chemische eigenschappen van HCPs beschrijven. Tot slot werden eiwiteigenschapskaarten gegenereerd om te helpen bij het bepalen van een geschikte zuiveringsstrategie in vergelijking met de fysisch-chemische eigenschappen van het doelantigeen.

Preparatieve chromatografie, gebaseerd op verschillen in fysisch-chemische eigenschappen tussen moleculen, is een van de belangrijkste technieken voor de

zuivering van vaccins. Als vervolg op hoofdstuk 2 werd een experimentele retentiekaart van het gastheercelproteoom tijdens een zoutgradiënt op een hydrofobe interactiechromatografiekolom (*hydrophobic interaction chromatography* - HIC) en een ionenuitwisselingschromatografiekolom (*ion exchange chromatography* - IEX) geconstrueerd en in **hoofdstuk 3** beschreven. Daarnaast identificeerde deze studie patronen in het retentiegedrag van HCPs op basis van hun eiwit-eiwit interacties, hun moleculaire functie en hun cellulaire componenten. Om het retentiegedrag van ongekarakteriseerde eiwitten te voorspellen, werd een kwantitatief structuur-eigenschap relatie (*quantitative structure-property relationship* - QSPR) model gemaakt met behulp van de IEX retentiegegevens. Subgroepen van eiwitten geïdentificeerd uit retentiepatronen werden gebruikt om verdere QSPR-modellen te genereren, waarbij subgroepen van monomeren de meest nauwkeurige voorspellingen leverden.

Om het proces beter te begrijpen, worden bij de procesontwikkeling mechanistische modellen (MM) van chromatografiekolommen gebruikt. Deze modellen beschrijven voornamelijk de vloeistofdynamica en het adsorptiegedrag van het doeleiwit en geselecteerde proces- of productgerelateerde onzuiverheden. Het is echter voordelig om ook terugkerende HCP-verontreinigingen in MMs op te nemen. HCPs die problemen veroorzaken als ze in het product achterblijven, zijn niet noodzakelijk in grote aantallen aanwezig in het cellysaat en worden vaak niet afzonderlijk beschreven. Bovendien is er nog geen methode om de parameters van de adsorptie-isotherm te bepalen voor individuele eiwitten in het volledige proteoom van de gastheercel.

Hoofdstuk 4 presenteert een methode voor de uitgebreide bepaling van de bovengenoemde parameters van de adsorptie-isotherm van individuele HCPs. Fracties van lineaire gradiënt elutie-experimenten met verschillende gradiëntlengtes werden geanalyseerd met behulp van shotgun proteomics om de retentietijden van de individuele HCPs te extraheren. Uit de geëxtraheerde retentievolumes per gradiënt werden de parameters van de adsorptie-isotherm geregistreerd voor alle individuele HCPs die in het lysaatmonster werden gedetecteerd. Deze nieuwe methode werd geïllustreerd en gevalideerd met het BLR *E. coli* monster en vervolgens toegepast om een adsorptiestap *in silico* te optimaliseren.

Tot slot geeft **hoofdstuk 5** een overzicht van de aanvullende onderzochte methoden voor monstervoorbereiding en -analyse. Deze omvatten het pakken van filterplaten met chromatografiehars voor batch adsorptie, wat werd onderzocht om parameters van de adsorptie-isotherm te bepalen in plaats van lineaire gradiënt elutie (LGE)

experimenten. Daarnaast werd ionenuitwisselingsvloeistofchromatografie (IEX-HPLC) onderzocht als analysetechniek in plaats van massaspectrometrie (MS).

Samengevat presenteert dit werk een uitgebreide, grootschalige karakterisering van HCPs van veelgebruikte *E. coli*-gastheercelstammen voor de productie van eiwitvaccins. Daarnaast wordt een gevalideerde aanpak gepresenteerd voor de bepaling van adsorptie-isothermeparameters van alle detecteerbare HCPs in het lysaatmonster.

Zusammenfassung

Die COVID-19-Pandemie hat deutlich gemacht, dass die Entwicklung neuer Impfstoffe beschleunigt werden muss. In den letzten Jahrzehnten hat sich der Engpass in der biopharmazeutischen Prozessentwicklung von der Optimierung der Fermentationsprozesse zur Entwicklung geeigneter Aufreinigungsstrategien verlagert. Dabei kann ein besseres Prozessverständnis die Entwicklung von Aufreinigungsstrategien erheblich beschleunigen. Im Gegensatz zu anderen biopharmazeutischen Produkten sind Proteinimpfstoffe häufig komplexer in der Herstellung, da sie Moleküle abweichender Herkunft enthalten. Infolgedessen gibt es für die Aufreinigung von Proteinuntereinheiten für Impfstoffe kein Standardverfahren. Wird ein neuartiges Antigen in den Wirtszellen exprimiert, ermöglicht die Kenntnis der möglichen Verunreinigungen - in diesem Fall zum Beispiel Wirtszellproteine (*host cell proteins* - HCPs) - eine rationale und systematische Prozessentwicklung. Daher konzentriert sich diese Arbeit auf die Entwicklung von Strategien zur Charakterisierung wiederkehrender HCP-Verunreinigungen (aus *E. coli*-Wirtszellen). Zusätzlich werden diese Informationen in einem Chromatographiemodell benutzt, um eine Strategien zur Entfernung dieser Proteine *in silico* zu entwickeln. Dabei liegt der Schwerpunkt dieser Thesis auf proteinbasierten Antigen-Impfstoffen.

Zunächst wurde in **Kapitel 2** das komplette Wirtszellproteom von *E. coli*-Wirtszellen (BLR(DE3) und HMS174(DE3)) charakterisiert. Etwa 2000 HCPs wurden aus der *E. coli*-Probe mittels Massenspektrometrie identifiziert. Darüber hinaus wurde eine umfangreiche HCP-Datenbank mit ihren Expressionsniveaus und physikochemischen Eigenschaften erstellt. Außerdem wurden die Profile eines Antigen-exprimierenden und eines Null-Plasmid-Stammes verglichen. Aus Aufreinigungssicht sind die Unterschiede gering, und die Erkenntnisse aus dem BLR(DE3)-Null-Plasmid-Stamm können zur Festlegung einer Aufreinigungsstrategie für den BLR(DE3)-Antigen-produzierenden Stamm und den HMS174(DE3)-Stamm herangezogen werden. Der Datensatz der identifizierten Proteine wurde mit Datenbanken verknüpft, die die physikochemischen Eigenschaften von HCPs beschrieben. Schließlich wurden Proteineigenschaftskarten erstellt, die bei der Ermittlung einer geeigneten Aufreiningungsstrategie helfen, wenn diese

verglichen werden mit den physikochemischen Eigenschaften des Zielantigens.

Die präparative Chromatographie, die auf Unterschieden in den physikochemischen Eigenschaften zwischen Molekülen basiert, ist eine der wichtigsten Techniken für die Aufreinigung von Impfstoffen. Als Fortsetzung zu Kapitel 2 wurde eine experimentelle Retentionskarte des Wirtszellproteoms während eines Salzgradienten auf einer hydrophober Interaktionschromatographiesäule (*hydrophobic interaction chromatography* - HIC) und einer Ionenaustauschchromatographiesäule (*ion exchange chromatography* - IEX) erstellt und in **Kapitel 3** beschrieben. Darüber hinaus wurden in dieser Studie Muster im Retentionsverhalten von HCPs auf der Grundlage ihrer Protein-Protein-Wechselwirkungen, ihrer molekularen Funktion und ihrer Zellbestandteile identifiziert. Um das Retentionsverhalten von noch nicht charakterisierten Proteinen vorhersagen zu können, wurde anhand der IEX-Retentionsdaten ein quantitatives Struktur-Eigenschafts-Beziehungsmodell (*quantitative structure-property relationship* - QSPR) erstellt. Untergruppen von Proteinen, die anhand von Retentionsmustern identifiziert wurden, dienten zur Erstellung weiterer QSPR-Modelle, wobei Untergruppen von Monomeren die genauesten Vorhersagen lieferten.

Um ein besseres Prozessverständnis zu erreichen, werden in der Prozessentwicklung mechanistische Modelle (MM) von Chromatographiesäulen verwendet. Diese Modelle beschreiben in erster Linie Fluiddynamik und Adsorptionsverhalten des Zielproteins und ausgewählter prozess- oder produktbezogener Verunreinigungen. Es ist jedoch von Vorteil, auch wiederkehrende HCP-Verunreinigungen in MMs einzubeziehen. Dabei sind HCPs, die Probleme verursachen, wenn sie im Produkt verbleiben, nicht notwendigerweise in groβer Zahl im Zelllysat vorhanden und werden oft nicht einzeln beschrieben. Zusätzlich gibt es bisher noch keine Methode zur Bestimmung der Parameter der Adsorptionsisotherme für individuelle Proteine im gesamten Wirtszellproteom.

In **Kapitel 4** wird eine Methode zur umfassenden Bestimmung der oben genannten Parameter der Adsorptionsisotherme einzelner HCPs vorgestellt. Fraktionen aus linearen Gradientenelutionsversuchen mit unterschiedlichen Gradientenlängen wurden mittels Shotgun-Proteomics

analysiert, um die Retentionszeiten der einzelnen HCPs zu extrahieren. Aus den extrahierten Retentionsvolumina pro Gradient wurden die Parameter der Adsorptionsisotherme für alle einzelnen HCPs, die in der Lysateprobe nachgewiesen wurden, regressiert. Diese neue Methode wurde anhand der BLR-*E. coli* Probe exemplifiziert, validiert und anschließend zur Optimierung eines Capture-Schritts *in silico* angewandt.

Schließlich gibt **Kapitel 5** einen Überblick über die zusätzlich untersuchten Hochdurchsatz-Probenvorbereitungs- und Analysemethoden. Dazu gehörte das Packen von Filterplatten mit Chromatographieharz für die Batch-Adsorption, die zur Bestimmung von Parameter der Adsorptionsisotherme anstelle von Linear Gradient Elution (LGE)-Experimenten erforscht wurde. Außerdem wurde die Ionen-Austausch-Hochleistungsflüssigkeitschromatographie (IEX-HPLC) als Analysetechnik anstelle der Massenspektrometrie (MS) untersucht.

Zusammenfassend stellt diese Arbeit eine umfassende, groß angelegte Charakterisierung von HCPs aus weit verbreiteten *E. coli*-Wirtszellstämmen für die Produktion von Proteinimpfstoffen vor. Darüber hinaus wird ein validierter Ansatz zur Bestimmung von Parameter der Adsorptionsisotherme aller nachweisbaren HCPs in der Lysateprobe vorgestellt.

„Was du im Kopf hast, kann dir keiner nehmen."

(translates to: „What you have in your head can not be taken from you)"

Familie Disela/von Kimakowitz

Chapter 1 Introduction

1.1 Biopharmaceuticals

Biopharmaceuticals, also known as biologics or biologicals, are distinct from traditional pharmaceutical products, or small molecules, as they are produced through biological processes rather than chemical synthesis. Biopharmaceuticals provide highly specific and effective treatments for a range of diseases, including cancer, autoimmune disorders, and infectious diseases. These biologically derived medicines include monoclonal antibodies (mAbs), recombinant proteins, and vaccines.

1.1.1 Vaccines

Vaccines have revolutionized public health by significantly reducing the prevalence of once-common infectious diseases such as smallpox, polio, measles, and pertussis. Emerging infectious diseases, such as COVID-19, underscore the urgent need for continued vaccine innovation to protect global populations from novel pathogens.

A driving factor is the economic benefit of vaccines. Preventing diseases through vaccination reduces healthcare costs associated with treatment and hospitalization and minimizes the socioeconomic impact of disease outbreaks, such as workforce absenteeism and loss of productivity. Investing in vaccine research is cost-effective, with long-term savings far outweighing the initial development and implementation expenses.

Furthermore, the advancement of vaccine technology offers the potential to combat antibiotic resistance. As bacterial pathogens evolve resistance to existing antibiotics, vaccines present a viable alternative for preventing bacterial infections without contributing to the resistance problem. This is particularly important for pathogens, which have shown increasing resistance to antibiotic treatments.

1.1.2 Protein subunit vaccines

Recombinant protein vaccines, also known as protein subunit vaccines or protein antigens, are the focus of this research. Unlike whole-pathogen vaccines (live attenuated or inactivated), protein vaccines consist only of a subunit of the pathogen [1]. In this case, this is a purified protein antigen derived from the target pathogen (Figure 1.1). These antigens are selected based on their ability to elicit a protective immune response.

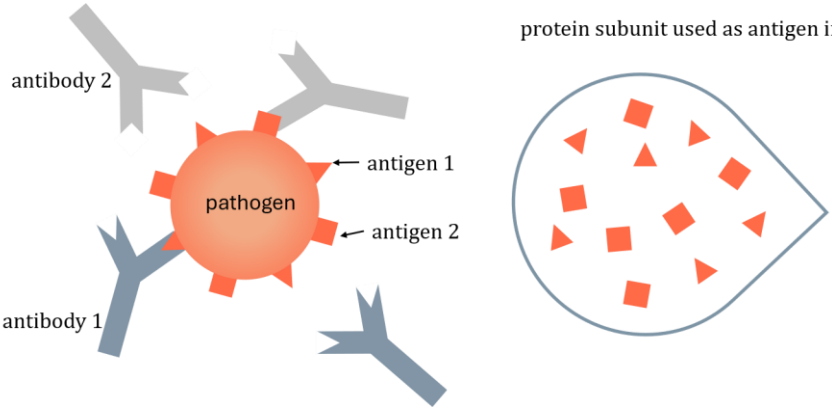


Figure 1.1: Scheme of immune response to protein subunit vaccines. The body forms antibodies as part of the immune response to pathogens. The protein subunit vaccines contain the antigen proteins from the surface of the pathogen responsible for the immune response. Hence an immune response can be induced without the disease.

One advantage of protein vaccines is their high degree of purity and specificity, which minimizes the risk of adverse reactions compared to whole-pathogen vaccines. However, their effectiveness may be limited by the need for adjuvants to enhance immunogenicity and multiple doses to achieve optimal protection. Furthermore, recombinant protein vaccines may require complex purification processes and can be costly to produce compared to other vaccine types.

In vaccine production, challenges include identifying suitable antigens, ensuring manufacturing scalability, selecting effective adjuvants, navigating regulatory requirements, and addressing cold chain logistics. Overcoming these challenges is crucial for the successful development, production, and deployment of vaccines to combat infectious diseases.

1.2 Process development of protein subunit vaccines

Vaccine production processes vary according to the type of vaccine [2], [3]. Protein subunit vaccines are produced recombinantly in a genetically modified host organism, in this case the gram-negative bacterium *Escherichia coli (E. coli)*. The process is comparable to other protein based biopharmaceuticals (see Figure 1.2), e.g. mAbs. During the fermentation, the host population is growing and eventually the expression of the target protein is induced when sufficient host cells are present. After fermentation of the host organism, the cells are disrupted to obtain the intracellular product. The main challenge of the following purification is to remove impurities such as host cell proteins (HCPs), DNA, RNA and endotoxins from the crude mixture. The purified target protein, called antigen should achieve a purity around 95%. It is further processed to the final drug product, in this case the vaccine, in the formulation step. In most cases the antigen is combined with other antigens originating from the same host to the final vaccine. Unlike other biopharmaceutical products, for example mAbs, the production process of the antigens is more various since vaccines are complex molecules from diverse origin. Hence no general platform process can be employed using affinity chromatography such as employing protein A to capture mAbs.

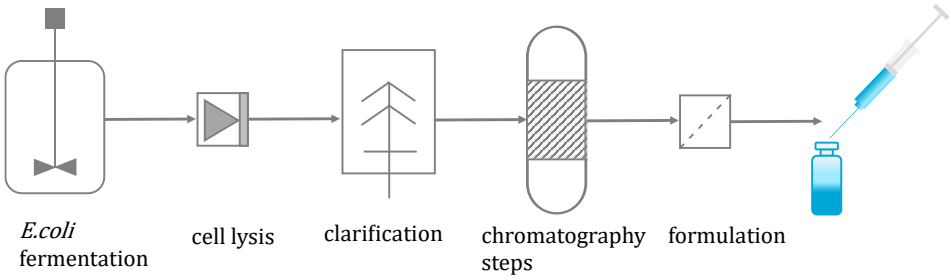


Figure 1.2: Scheme of a protein subunit vaccine production process.

While regulatory authorities demand a broader process and product understanding, from an economic and societal point of view the demand exists to shorten the development time of previously 10-20 years without compromising the product quality. The development of the purification and especially the chromatography steps pose the biggest bottleneck in the development of the process. Hence these are in the focus of this thesis (Figure 1.3).

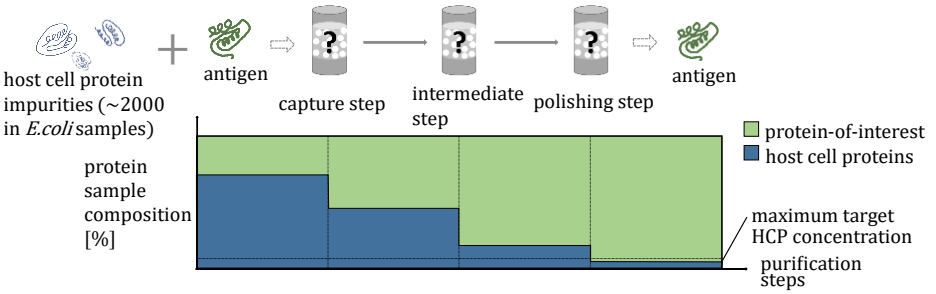


Figure 1.3: Scheme of the purification process development. Here the focus is on the host cell protein impurities that are present after the clarification together with the antigen.

Three main process steps are commonly used in the purification of proteins [4]. The first chromatography step is the "capture step", which serves as a coarse purification step, removing the bulk of impurities and concentrating the protein product. Subsequent intermediate purification steps use various chromatography resins to further reduce impurities. Finally, the polishing step is used to remove the low abundance and minor impurities [4]. Commonly used chromatographic separation techniques include ion exchange, hydrophobic interaction, mixed mode, size exclusion, or affinity-based chromatography, with packed bed resins being the current state of the art [5]. Suitable chromatography resins and buffer conditions must be identified when developing a new purification process.

The removal of HCPs is very challenging and therefore the focus of intensive process development for biopharmaceuticals. In the case of vaccines, the acceptable level of HCP is defined by the regulatory authorities on a case by case basis [6]. For a malaria vaccine candidate expressed in *E. coli* with a planned administration of 80 µg protein antigen per dose, Zhu et al. [7] specified that 1 µg/dose of total HCP impurities could be considered the limit, with a target of 100 ng/dose well below. In this case, the total HCP concentration was set at 90 ng or < 1100 ppm per dose [8]. Tolerated HCP levels for vaccines are generally higher than those for chronic disease medications (< 100 ppm) [6].

1.3 Purification process development tools

The biopharmaceutical industry requires novel methods to support process development at pandemic pace, while maintaining the highest product quality and robust material supply for clinical trials [9]. In the past, the

development of new processes still required expert knowledge and trial-and-error based screening approaches to identify suitable conditions for the development of effective purification steps [4]. A more rational and systematic approach is needed [10]–[12]. This rational and systematic approach build on models aims to enable shorter process development for future processes. In the following sub-chapters, the used process development tools are explained in detail.

1.3.1 High throughput experimentation/screening

High throughput experimentation (HTE), also referred to as high throughput screening (HTS), is a powerful method used extensively in scientific discovery, notably in drug development and across fields like biology, materials science, and chemistry. This technique allows the performance of a large number of experiments under variable conditions [13]. The miniaturization and parallelization of these HTS techniques result in decreased experimentation time and sample volumes. Additionally, the often used set-up in 96-well plates also allows for automatization of the experiments employing robotic Liquid Handling systems (by Tecan, Hamilton etc.). However, HTE also presents challenges such as data management complexities. Despite this limitation, HTS remains a crucial tool for advancing scientific understanding and driving innovation in various disciplines.

Increasing regulatory demands for a better understanding of purification processes, a critical aspect of the Quality by Design (QbD) initiative of the Food and Drug Administration (FDA), coupled with rising costs in biopharmaceutical research and development, have driven process development teams to seek more efficient tools and workflows. This quest has given rise to a new field known as high-throughput process development (HTPD). HTPD has shown promising applications in several areas of biomanufacturing process development, such as clone selection, upstream process optimization, and downstream purification process development [14].

A popular application of the HTPD concept is in the development of chromatographic separations. Three HTPD-compatible formats are suitable for the initial screening of chromatography resins and/or process conditions, categorized into packed-bed-based and slurry-based formats.

Packed-bed formats encompass minicolumns and pipette tips filled with chromatography resins, with minicolumns commercially available in sizes ranging from 50 to 200 µL (also known as Robocolumns). The other packed-bed format utilizes pipette tips filled with chromatography resins, with resin volumes ranging from 5 to 320 µL, offering versatility in experimentation. The third format used in HTPD studies of chromatographic separations combines the well-known principle of batch adsorption with the high-throughput advantages of microtiter plate formats [14]. The resin is dispensed into the microtiter plate [15], [16] or available in preloaded formats, then washed with buffer before being incubated with protein under specific mobile phase conditions.

The HTPD-compatible formats designed for chromatography resin screening can also be used as fast methods for mechanistic model parameter determination. This is further described in the next sub-chapter.

1.3.2 Mechanistic modeling

Mechanistic models describe the physicochemical mechanisms that occur during chromatography *in silico*. The big advantage of such models is that they allow *in silico* extrapolation of separation processes for conditions outside the experimentally tested parameter space, such as e.g. other resin or buffer volumes [17], [18]. This allows for more process understanding, and the potential of *in silico* optimization of the chromatography step provided that all components and their parameters are known.

System, column and chromatographic data are used as calibration input (Figure 1.4). The mechanistic model itself is based on differential equations describing solute transport (inter- and intraparticle) [18], [19]. The transport and adsorption elements of a chromatography model are combined [18]. Extensive work has been done to develop binding isotherms that describe the interaction between the molecule of interest and the ligand on the resin bead for various chromatographic modalities [20], [21].

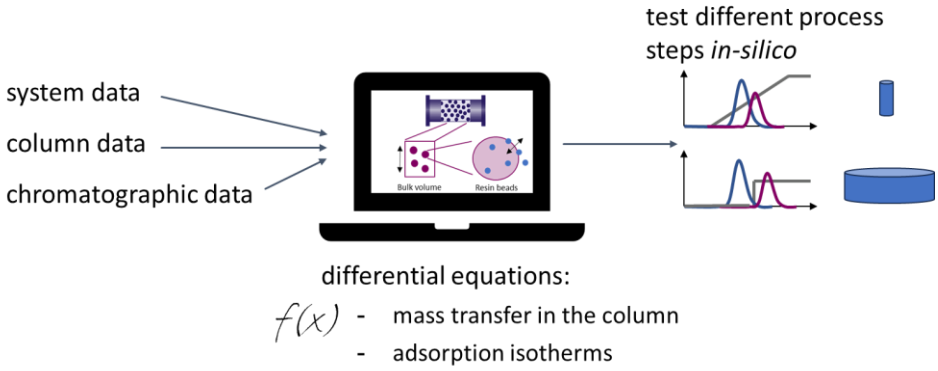


Figure 1.4.: Scheme to explain mechanistic models for chromatography.

The parameters used in the binding isotherms, so called isotherm parameters have to be determined experimentally since these are specific depending on the protein, resin, binding and elution conditions. Isotherm parameter determination experiments can be performed using static or dynamic methods [22].

In static methods (see HTE methods batch adsorption), protein dilution series are exposed to resin in a 96-well plate format for sufficient time to reach equilibrium [23], [24]. The differences in protein concentrations before and after the experiment are needed to determine the equilibrium constants [25], [26]. However, these methods often lack precision due to difficulties in reproducibly aliquoting resins and accounting for residual liquid [27].

Dynamic methods involve continuous flow through a packed bed column and measurement of protein concentration at the column outlet over time. Typically a UV signal is measured during linear gradient elution (LGE). This can be used to determine the elution volume of proteins. When varying the elution gradient length, the isotherm parameters can be regressed [28]–[30]. Another common approach is the model-based inverse method, in which isotherm parameters are fitted to the experimental data [31].

In typical purification process development, the primary focus is on the target protein and the conditions under which it binds or elutes. It is generally assumed that all impurities exhibit different binding behaviors and can be subsequently removed. Consequently, most efforts are concentrated on developing isotherms and models for the target molecule, with minimal

attention given to HCP impurities. Finding experimental techniques capable of determining the necessary isotherm parameters for impurities rather than pure standard proteins is a key challenge in applying these approaches to real-world purification problems. Targeted removal of impurities such as HCPs from the protein of interest is only possible if they are described in the mechanistic model.

1.3.3 Quantitative structure-property relationship modeling

Quantitative Structure-Property Relationship (QSPR) modeling is a computational approach used to correlate the structure of molecules with their physicochemical properties. In the realm of biopharmaceuticals, QSPR modeling plays a crucial role in predicting various properties of proteins, including their retention behavior in chromatographic processes [22], [32]. Over the last 20 years, successful models have been trained for a variety of globular proteins or antibodies, highlighting how structural knowledge of proteins can be used to describe chromatographic behavior [9], [33]–[37].

Protein retention in chromatography is influenced by several factors, including size, charge, hydrophobicity, and surface characteristics of the protein molecule, as well as the physicochemical properties of the chromatographic stationary phase. QSPR models leverage mathematical algorithms and statistical techniques to analyze the relationship between these molecular descriptors and protein retention, thereby facilitating the prediction of chromatographic behavior. Nonetheless, modeling HCPs is particularly challenging due to the mixture of proteins, complex interactions, and the fact that tertiary structures can deviate from theoretical predictions. Additionally, not all protein structures have been experimentally measured, though prediction tools like AlphaFold offer the potential to predict these structures with high accuracy.

Once developed and validated, QSPR models can be employed to predict the retention behavior of new undetected proteins under the same chromatographic conditions, such as mobile phase composition, pH, temperature, and column type. The major advantage of QSPR models lies in this ability to predict the behavior of new proteins that have never been measured, expanding beyond the experimental space. However, the inherent complexity of biological samples, especially those containing multiple

proteins and protein-protein interactions, bear a big challenge for QSPR applications in biopharmaceutical purifications.

1.4 Host cell proteins

HCPs are process-related impurities that can cause problems for the safety and efficacy of biopharmaceuticals, here vaccines. Reducing HCPs early in clinical development is important, as is using robust and sensitive methods for product purity testing and process development [38].

Intensive analytical development has focused on monitoring the purification process and measuring residual HCPs. Currently, anti-HCP enzyme-linked immunosorbent assays (ELISAs) are the gold standard for determining total HCP content at detection levels as low as 1 ng/mL [8], [10]. However, ELISAs only detect the proteins for which they were designed, and total protein ELISAs do not provide information about individual proteins present in the drug substance or drug product. Knowledge of HCP identities is limited. Thus, it is recommended that orthogonal methods are used to support process development and validation [39].

In addition, over the last few decades, significant advances in high-resolution mass spectrometry have pushed the limits of large-scale proteomics in the direction of higher accuracy, sensitivity, and throughput. Mass-spectrometry-based proteomics has emerged as a powerful alternative to identify and quantify HCPs down to detection limits of known and unknown components of up to 5 ppm [40]. This allows knowledge-based risk mitigation of critical HCPs [10]. As a result, host cell proteomics is increasingly used to monitor purification progress and confirm the absence of specific HCPs in the final drug substance or product [6], [7], [39]–[41].

Describing the HCP content of different expression hosts (*pichia pistoris*, Chinese hamster ovary (CHO),…) has been of interest over the past two decades [8], [42], [43]. At present, most of the literature is describing HCP from CHO cells, more specifically the HCP content after the protein A capture step in antibody production [38], [44]–[46]. From these, high-risk HCPs for CHO have been identified that have potential immunogenic reactions or compromise product quality due to degradation [44]. Studies have shown that HCP aggregates with mAbs can promote the persistence of HCPs during the protein A capture step [47]–[49]. A recent correlation analysis of HCPs

identified co-elution of HCPs in groups associated with protein-protein interactions (PPIs) [46].

Interactions between the product and production cell enzymes during cell disruption or enzyme release from dying cells are a potential source of significant damage to the intended native configuration [50]. This may result in irreversible aggregation of the product, significantly reducing the yield and raising concerns such as immunogenicity, as demonstrated by recent evidence of the involvement of HCPs in product aggregation [47]–[49]. Likewise, product stability can be affected by small amounts of HCPs such as host cell lipases capable of degrading the excipient polysorbate 20 or polysorbate 80 [51]. These examples highlight the importance of monitoring seemingly unimportant, low abundant proteins because they could lead to issues later on in the process.

Fewer studies, however, have been conducted on *E. coli* HCPs. Bartlow et al. analyzed a range of elution buffer concentrations by SDS/PAGE combined with MALDI-TOF MS and found 26 proteins that co-eluted during green fluorescent protein purification [52]. Recently, Lingg et al. investigated the effect of metal and chelator type on HCPs found in a similar process eluate [53]. Swanson et al. studied the elution of *E. coli* HCPs in a 5-step isocratic elution [54], [55] for cation and anion exchange chromatography. Using the experimentally determined molecular weight, isoelectric point (pI), and aqueous two-phase partition coefficients of the HCPs, random forest regressor models were trained to predict the retention of the proteins.

For products such as vaccine antigens produced in *E. coli*, however, no common persistent proteins are known. In particular, proteolytic digestion is a challenge when working with *E. coli* as a host [56]. Awareness of the importance of early removal, particularly of production cellular enzymes such as proteases, has proven to be beneficial in maintaining product integrity [50]. Another critical group to remove are chaperones, proteins involved in correct folding and associated with human diseases due to immunogenicity [57]. Although the elimination of these protein groups is a high priority, they are not always abundant in cell lysates and are often not described individually.

There are several approaches for the determination of isotherm parameters of the main HCP impurities during the production of mAbs or a therapeutic enzyme [58]–[60]. The identities of the HCPs are described according to

their experimentally determined physicochemical properties. Fractionation was used to construct multidimensional property maps, and isotherm parameters for these fractions of CHO HCP impurities were determined using orthogonal chromatographic methods. Similarly, characterization of process-related impurities (including HCPs) in *pichia pastoris* was performed on a library of chromatographic resins to describe their affinities [42], [61]. Wierling et al. [62] approached the determination of HCP impurities from CHO cells during the purification of a mAb by the combination of HTS with mass spectrometry detection.

1.5 Project setting and aim of the thesis

This thesis is part of a collaboration between GSK (Belgium) and Delft University of Technology (The Netherlands) relating to the development and establishment of a model-based high throughput development platform for downstream processing. The collaboration aims to ultimately reduce development time while increasing process understanding and is focused on protein subunit vaccines produced in the host *E. coli*. Two other PhD projects are part of the collaboration. Each PhD project has their main focus on one of the process development tools described earlier. One PhD project focusing on mechanistic modeling that aims to computationally describe and optimize the entire downstream process [63]. The other project is focused on molecular modeling of chromatographic separation by means of e.g. retention prediction using QSPR.

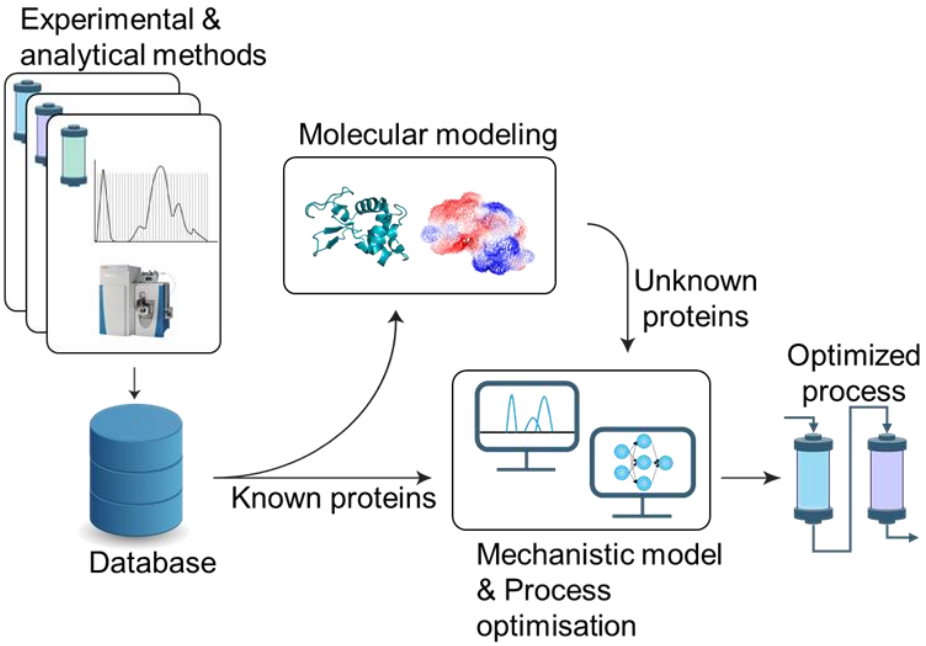


Figure 1.5: Schematic overview of project context.

The scope of this thesis is to characterize the host cell proteome. By building comprehensive databases and advanced methods, we can establish a general approach that applies to any new antigen produced in the characterized host. The strategy is shown in Figure 1.5. It involves identifying proteins originating from *E. coli* using mass spectrometry and describing the chromatographic binding behavior of the HCPs. Gradient elution experiments are conducted with the HCP mixture, and analysis of the fractions via mass spectrometry determines the retention times of individual proteins. These retention times allow for the extraction of model parameters through correlation, which can then be used in a mechanistic model to optimize the process *in silico*. Additionally, based on the database, a QSPR model can be developed to predict the retention times of unknown proteins. Unknown proteins could be new antigens, HCPs under the detection limit or HCPs from a new host. In the future, this approach can enable a new *in silico*-driven process development.

1.6 Outline of thesis

Chapter 1 gives an introduction to the field of vaccine process development and the work performed in this thesis.

Chapter 2 focuses on characterizing the host cell proteome of two commercially used *E. coli* strains (BLR(DE3), HMS174(DE3)) utilizing mass spectrometry based proteomics. A particular emphasis is set on comparing protein profiles between the two different strains and BLR(DE3) antigen-expressing with BLR(DE3) null plasmid strain. The identified proteins are connected to their theoretical physicochemical properties and protein property maps are generated. These protein property maps are used to give an indication for suitable process development strategies.

In **chapter 3**, the retention behavior of HCPs on chromatography resins during LGE experiments is investigated experimentally. The LGE experiments are conducted on ion exchange (IEX) and hydrophobic interaction chromatography (HIC) resins. This leads to an experimental protein retention map. Furthermore, the IEX retention data is used to build a descriptive QSPR model. Co-elution patterns based on cellular location, molecular function, or protein-protein interactions can be observed in the retention data. These observed patterns are then utilized to select subsets of proteins that are suited to predict retention times with more accuracy in the QSPR model.

Chapter 4 introduces a method to determine isotherm parameters of individual HCPs in the holistically measured host cell proteome. LGE experiments with varying gradient lengths are conducted and the extracted retention volumes of individual HCPs are used to regress the isotherm parameters. A selection of proteins covering varying concentration ranges is used to validate the mechanistic model and optimize a chromatography capture step.

Transitioning to **chapter 5**, the thesis explores the implementation of HTE to enhance the method of experimental isotherm determination.

Finally, **chapter 6** concludes the thesis, offering insights and reflections on the research conducted, along with potential avenues for future exploration.

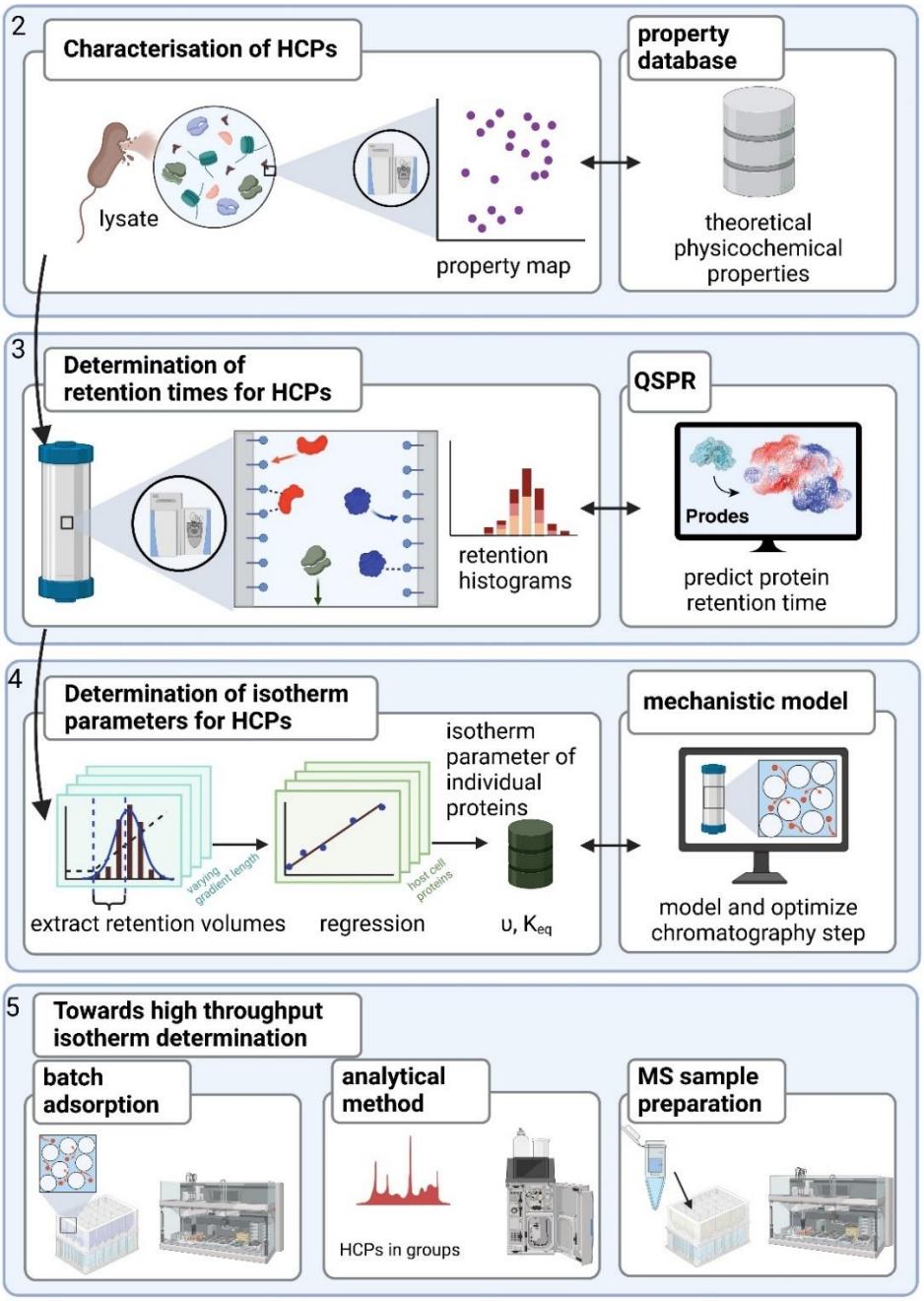


Figure 1.6: Schematic overview of the thesis outline (Illustration created using BioRender.com.).

1.7 References

[1] M. Hansson, P.-Å. Nygren, and S. Ståhl, "Design and production of recombinant subunit vaccines," *Biotechnol. Appl. Biochem.*, vol. 32, no. 2, p. 95, 2000.

[2] J. O. Josefsberg and B. Buckland, "Vaccine process technology," *Biotechnol. Bioeng.*, vol. 109, no. 6, pp. 1443–1460, 2012.

[3] B. Buckland *et al.*, "Vaccine process technology — A decade of progress," no. March, pp. 1–32, 2024.

[4] E. Wen, R. Ellis, and N. S. Pujar, Eds., *Vaccine Development and Manufacturing*, First. Wiley, 2015.

[5] U. Gottschalk, K. Brorson, and A. A. Shukla, "The need for innovation in biomanufacturing," *Nat. Biotechnol.*, vol. 30, no. 6, pp. 489–492, 2012.

[6] K. Reiter, M. Suzuki, L. R. Olano, and D. L. Narum, "Host cell protein quantification of an optimized purification method by mass spectrometry," *J. Pharm. Biomed. Anal.*, vol. 174, pp. 650–654, 2019.

[7] D. Zhu, A. J. Saul, and A. P. Miles, "A quantitative slot blot assay for host cell protein impurities in recombinant proteins expressed in E. coli," *J. Immunol. Methods*, vol. 306, no. 1–2, pp. 40–50, Nov. 2005.

[8] A. L. Tscheliessnig, J. Konrath, R. Bates, and A. Jungbauer, "Host cell protein analysis in therapeutic protein bioprocessing - methods and applications," *Biotechnol. J.*, vol. 8, no. 6, pp. 655–670, 2013.

[9] D. Saleh *et al.*, "A multiscale modeling method for therapeutic antibodies in ion exchange chromatography," *Biotechnol. Bioeng.*, vol. 120, no. 1, pp. 125–138, 2023.

[10] C. E. M. Hogwood, D. G. Bracewell, and C. M. Smales, "Measurement and control of host cell proteins (HCPs) in CHO cell bioprocesses," *Curr. Opin. Biotechnol.*, vol. 30, no. July, pp. 153–160, 2014.

[11] D. Keulen, G. Geldhof, O. Le Bussy, M. Pabst, and M. Ottens, "Recent advances to accelerate purification process development: A review with a focus on vaccines," *J. Chromatogr. A*, vol. 1676, p. 463195, Aug. 2022.

[12] A. T. Hanke and M. Ottens, “Purifying biopharmaceuticals: Knowledge-based chromatographic process development,” *Trends Biotechnol.*, vol. 32, no. 4, pp. 210–220, 2014.

[13] A. S. Rathore, “Quality by Design (QbD)-Based Process Development for Purification of a Biotherapeutic,” *Trends Biotechnol.*, vol. 34, no. 5, pp. 358–370, 2016.

[14] K. M. Łacki, “High-throughput process development of chromatography steps: Advantages and limitations of different formats used,” *Biotechnology Journal*, vol. 7, no. 10. pp. 1192–1202, 2012.

[15] T. Herrmann, M. Schröder, and J. Hubbuch, “Generation of equally sized particle plaques using solid-liquid suspensions,” *Biotechnol. Prog.*, vol. 22, no. 3, pp. 914–918, 2006.

[16] K. M. Lacki and E. Brekkan, “High throughput screening techniques in protein purification.,” *Methods of biochemical analysis*, vol. 54. pp. 489–506, 2011.

[17] S. W. Benner, J. P. Welsh, M. A. Rauscher, and J. M. Pollard, “Prediction of lab and manufacturing scale chromatography performance using mini-columns and mechanistic modeling,” *J. Chromatogr. A*, vol. 1593, pp. 54–62, 2019.

[18] V. Kumar and A. M. Lenhoff, “Mechanistic Modeling of Preparative Column Chromatography for Biotherapeutics,” *Annu. Rev. Chem. Biomol. Eng.*, vol. 11, pp. 235–255, 2020.

[19] F. Rischawy, D. Saleh, T. Hahn, S. Oelmeier, J. Spitz, and S. Kluters, “Good modeling practice for industrial chromatography: Mechanistic modeling of ion exchange chromatography of a bispecific antibody,” *Comput. Chem. Eng.*, vol. 130, p. 106532, 2019.

[20] C. a Brooks and S. M. Cramer, “Steric mass-action ion exchange: Displacement profiles and induced salt gradients,” *AIChE J.*, vol. 38, no. 12, pp. 1969–1978, 1992.

[21] B. K. Nfor, M. Noverraz, S. Chilamkurthi, P. D. E. M. Verhaert, L. A. M. van der Wielen, and M. Ottens, “High-throughput isotherm determination and thermodynamic modeling of protein adsorption on mixed mode adsorbents,” *J. Chromatogr. A*, vol. 1217, no. 44, pp. 6829–6850, 2010.

[22] C. R. Bernau, M. Knödler, J. Emonts, R. C. Jäpel, and J. F. Buyel, "The use of predictive models to develop chromatography-based purification processes," *Front. Bioeng. Biotechnol.*, vol. 10, no. October, pp. 1–24, 2022.

[23] S. Ghose, B. Hubbard, and S. M. Cramer, "Binding capacity differences for antibodies and Fc-fusion proteins on protein A chromatographic materials," *Biotechnol. Bioeng.*, vol. 96, no. 4, pp. 768–779, Mar. 2007.

[24] M. Moreno-González, P. Chuekitkumchorn, M. Silva, R. Groenewoud, and M. Ottens, "High throughput process development for the purification of rapeseed proteins napin and cruciferin by ion exchange chromatography," *Food Bioprod. Process.*, vol. 125, pp. 228–241, 2021.

[25] G. Guiochon, "Preparative liquid chromatography," *J. Chromatogr. A*, vol. 965, no. 1–2, pp. 129–161, Aug. 2002.

[26] A. Seidel-Morgenstern, "Experimental determination of single solute and competitive adsorption isotherms," *J. Chromatogr. A*, vol. 1037, no. 1–2, pp. 255–272, 2004.

[27] J. L. Coffman, J. F. Kramarczyk, and B. D. Kelley, "High-throughput screening of chromatographic separations: I. method development and column modeling," *Biotechnol. Bioeng.*, vol. 100, no. 4, pp. 605–618, 2008.

[28] E. S. Parente and D. B. Wetlaufer, "Relationship between isocratic and gradient retention times in the high-performance ion-exchange chromatography of proteins. Theory and experiment," *J. Chromatogr. A*, vol. 355, no. C, pp. 29–40, 1986.

[29] S. Yamamoto, K. Nakanishi, R. Matsuno, and T. Kamijubo, "Ion exchange chromatography of proteins—predictions of elution curves and operating conditions. II. Experimental verification," *Biotechnol. Bioeng.*, vol. 25, no. 5, pp. 1373–1391, May 1983.

[30] S. Yamamoto, K. Nakanishi, R. Matsuno, and T. Kamikubo, "Ion exchange chromatography of proteins—prediction of elution curves and operating conditions. I. Theoretical considerations," *Biotechnol. Bioeng.*, vol. 25, no. 6, pp. 1465–1483, Jun. 1983.

[31] T. Hahn, P. Baumann, T. Huuk, V. Heuveline, and J. Hubbuch, "UV absorption-based inverse modeling of protein chromatography," *Eng. Life Sci.*, vol. 16, no. 2, pp. 99–106, 2016.

[32] J. Emonts and J. F. Buyel, "An overview of descriptors to capture protein properties – Tools and perspectives in the context of QSAR modeling," *Comput. Struct. Biotechnol. J.*, vol. 21, pp. 3234–3247, 2023.

[33] A. T. Hanke *et al.*, "Prediction of protein retention times in hydrophobic interaction chromatography by robust statistical characterization of their atomic-level surface properties," *Biotechnol. Prog.*, vol. 32, no. 2, pp. 372–381, 2016.

[34] R. Hess *et al.*, "Predicting multimodal chromatography of therapeutic antibodies using multiscale modeling," *J. Chromatogr. A*, vol. 1718, no. February, p. 464706, 2024.

[35] J. Kittelmann, K. M. H. Lang, M. Ottens, and J. Hubbuch, "Orientation of monoclonal antibodies in ion-exchange chromatography: A predictive quantitative structure–activity relationship modeling approach," *J. Chromatogr. A*, vol. 1510, pp. 33–39, 2017.

[36] C. B. Mazza, N. Sukumar, C. M. Breneman, and S. M. Cramer, "Prediction of protein retention in ion-exchange systems using molecular descriptors obtained from crystal structure," *Anal. Chem.*, vol. 73, no. 22, pp. 5457–5461, 2001.

[37] T. Yang, M. C. Sundling, A. S. Freed, C. M. Breneman, and S. M. Cramer, "Prediction of pH-dependent chromatographic behavior in ion-exchange systems," *Anal. Chem.*, vol. 79, no. 23, pp. 8927–8939, 2007.

[38] M. Vanderlaan, J. Zhu-Shimoni, S. Lin, F. Gunawan, T. Waerner, and K. E. Van Cott, "Experience with host cell protein impurities in biopharmaceuticals," *Biotechnol. Prog.*, vol. 34, no. 4, pp. 828–837, Jul. 2018.

[39] H. Falkenberg *et al.*, "Mass spectrometric evaluation of upstream and downstream process influences on host cell protein patterns in biopharmaceutical products," *Biotechnol. Prog.*, vol. 35, no. 3, p. e2788, May 2019.

[40] S. Eliuk and A. Makarov, "Evolution of Orbitrap Mass Spectrometry Instrumentation," *Annu. Rev. Anal. Chem.*, vol. 8, pp. 61–80, 2015.

[41] V. Reisinger, H. Toll, R. Ernst, J. Visser, and F. Wolschin, "A mass spectrometry-based approach to host cell protein identification and its

application in a comparability exercise," *Anal. Biochem.*, vol. 463, pp. 1–6, 2014.

[42] S. M. Timmick *et al.*, "An impurity characterization based approach for the rapid development of integrated downstream purification processes," *Biotechnol. Bioeng.*, vol. 115, no. 8, pp. 2048–2060, 2018.

[43] X. Wang, A. K. Hunter, and N. M. Mozier, "Host cell proteins in biologics development: Identification, quantitation and risk assessment," *Biotechnol. Bioeng.*, vol. 103, no. 3, pp. 446–458, 2009.

[44] M. Jones *et al.*, "'High-risk' host cell proteins (HCPs): A multi-company collaborative view," *Biotechnol. Bioeng.*, vol. 118, no. 8, pp. 2870–2885, Aug. 2021.

[45] D. Migani, C. M. Smales, and D. G. Bracewell, "Effects of lysosomal biotherapeutic recombinant protein expression on cell stress and protease and general host cell protein release in Chinese hamster ovary cells," *Biotechnol. Prog.*, vol. 33, no. 3, pp. 666–676, 2017.

[46] S. Panikulam *et al.*, "Host cell protein networks as a novel co-elution mechanism during protein A chromatography," *Biotechnol. Bioeng.*, Mar. 2024.

[47] C. E. Herman *et al.*, "Behavior of host-cell-protein-rich aggregates in antibody capture and polishing chromatography," *J. Chromatogr. A*, vol. 1702, p. 464081, 2023.

[48] C. E. Herman *et al.*, "Analytical characterization of host-cell-protein-rich aggregates in monoclonal antibody solutions," *Biotechnol. Prog.*, vol. 39, no. 4, pp. 1–16, 2023.

[49] Y. H. Oh *et al.*, "Characterization and implications of host-cell protein aggregates in biopharmaceutical processing," *Biotechnol. Bioeng.*, vol. 120, no. 4, pp. 1068–1080, Apr. 2023.

[50] G. Jagschies and K. M. Łącki, "Process Capability Requirements," in *Biopharmaceutical Processing*, Elsevier, 2018, pp. 73–94.

[51] X. Li, F. Wang, H. Li, D. D. Richardson, and D. J. Roush, "The measurement and control of high-risk host cell proteins for polysorbate

degradation in biologics formulation," *Antib. Ther.*, vol. 5, no. 1, pp. 42–54, 2022.

[52] P. Bartlow *et al.*, "Identification of native Escherichia coli BL21 (DE3) proteins that bind to immobilized metal affinity chromatography under high imidazole conditions and use of 2D-DIGE to evaluate contamination pools with respect to recombinant protein expression level," *Protein Expr. Purif.*, vol. 78, no. 2, pp. 216–224, 2011.

[53] N. Lingg *et al.*, "Proteomics analysis of host cell proteins after immobilized metal affinity chromatography: Influence of ligand and metal ions," *J. Chromatogr. A*, vol. 1633, p. 461649, Dec. 2020.

[54] R. K. Swanson, R. Xu, D. S. Nettleton, and C. E. Glatz, "Accounting for host cell protein behavior in anion-exchange chromatography," *Biotechnol. Prog.*, vol. 32, no. 6, pp. 1453–1463, Nov. 2016.

[55] R. K. Swanson, R. Xu, D. Nettleton, and C. E. Glatz, "Proteomics-based, multivariate random forest method for prediction of protein separation behavior during cation-exchange chromatography," *J. Chromatogr. A*, vol. 1249, pp. 103–114, 2012.

[56] E. K. Lindskog, S. Fischer, T. Wenger, and P. Schulz, "Host Cells," in *Biopharmaceutical Processing*, Elsevier, 2018, pp. 111–130.

[57] J. C. Ranford, A. R. M. Coates, and B. Henderson, "Chaperonins are cell-signalling proteins: The unfolding biology of molecular chaperones," *Expert Rev. Mol. Med.*, vol. 2, no. 8, pp. 1–17, 2000.

[58] B. K. Nfor *et al.*, "Multi-dimensional fractionation and characterization of crude protein mixtures: Toward establishment of a database of protein purification process development parameters," *Biotechnol. Bioeng.*, vol. 109, no. 12, pp. 3070–3083, Dec. 2012.

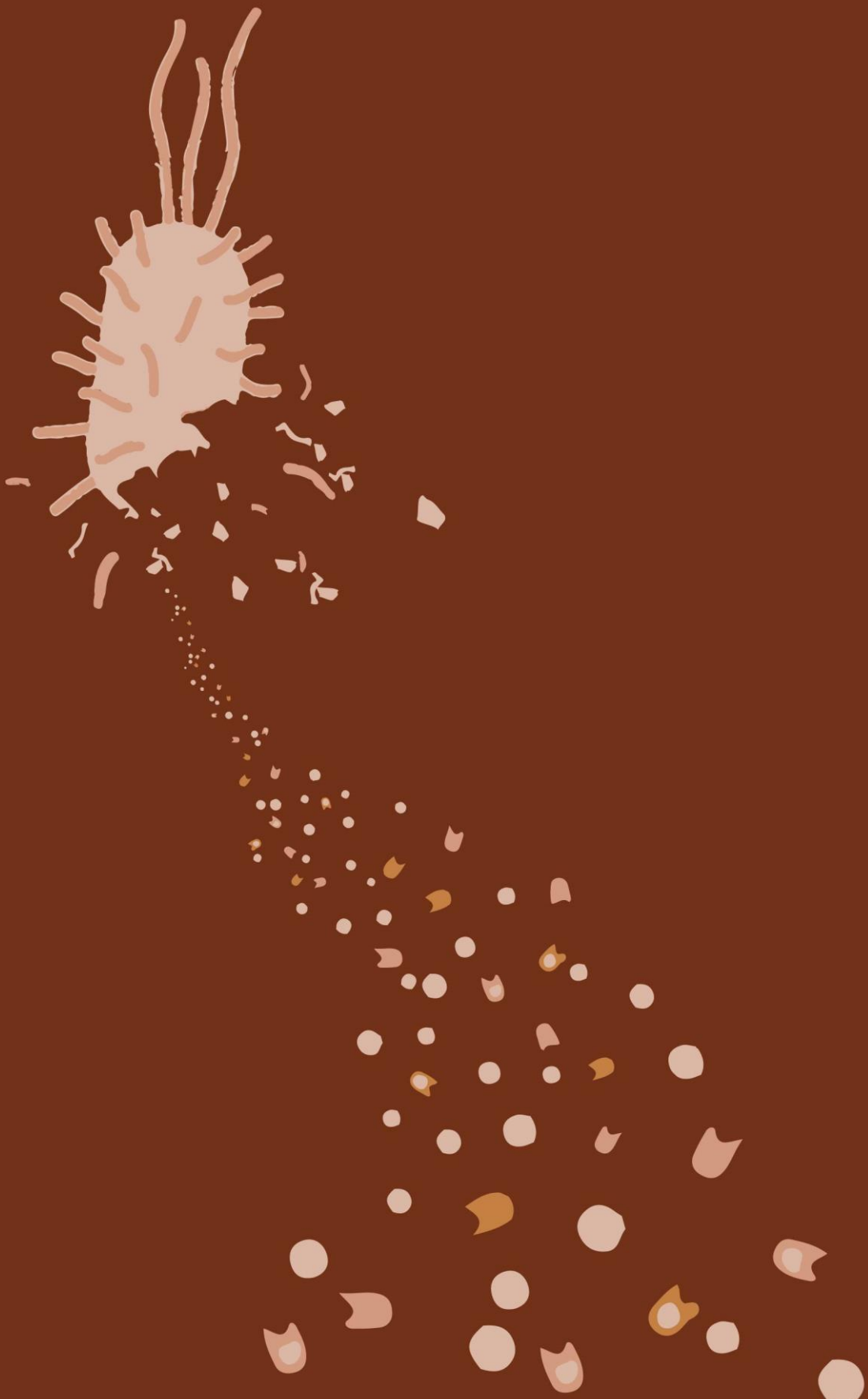
[59] A. T. Hanke *et al.*, "3D-liquid chromatography as a complex mixture characterization tool for knowledge-based downstream process development," *Biotechnol. Prog.*, vol. 32, no. 5, pp. 1283–1291, 2016.

[60] S. M. Pirrung *et al.*, "Chromatographic parameter determination for complex biological feedstocks," *Biotechnol. Prog.*, vol. 34, no. 4, pp. 1006–1018, 2018.

[61] N. Vecchiarello *et al.*, "A combined screening and in silico strategy for the rapid design of integrated downstream processes for process and product-related impurity removal," *Biotechnol. Bioeng.*, vol. 116, no. 9, pp. 2178–2190, 2019.

[62] P. S. Wierling, R. Bogumil, E. Knieps-Grünhagen, and J. Hubbuch, "High-throughput screening of packed-bed chromatography coupled with SELDI-TOF MS analysis: monoclonal antibodies versus host cell protein," *Biotechnol. Bioeng.*, vol. 98, no. 2, pp. 440–450, Oct. 2007.

[63] D. Keulen, "Computational modeling and optimization of biopharmaceutical downstream processes," 2024, PhD Thesis Delft University of Technology.



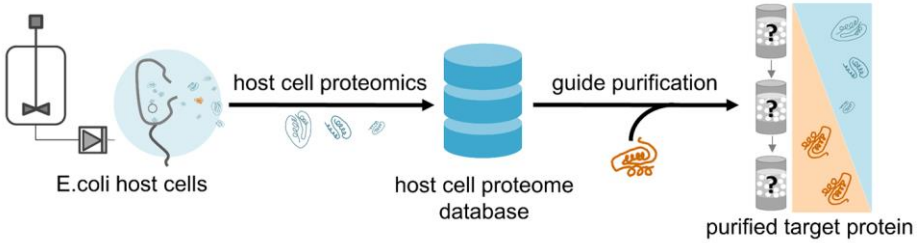
Chapter 2 Characterisation of the *E. coli* HMS174 and BLR host cell proteome to guide purification process development

This chapter has been published as:

Disela, R., Le Bussy, O., Geldhof, G., Pabst, M. Ottens, M., Characterisation of the E. coli HMS174 and BLR host cell proteome to guide purification process development, *Biotechnology Journal*, 18 (9) (2023), Article e2300068, https://doi.org/10.1002/biot.202300068

Abstract

Mass-spectrometry-based proteomics is increasingly employed to monitor purification processes or to detect critical host cell proteins in the final drug substance. This approach is inherently unbiased and can be used to identify individual host cell proteins without prior knowledge. In process development for the purification of new biopharmaceuticals, such as protein subunit vaccines, a broader knowledge of the host cell proteome could promote a more rational process design. Proteomics can establish qualitative and quantitative information on the complete host cell proteome before purification (i.e., protein abundances and physicochemical properties). Such information allows for a more rational design of the purification strategy and accelerates purification process development. In this study, we present an extensive proteomic characterisation of two *E. coli* host cell strains widely employed in academia and industry to produce therapeutic proteins, BLR and HMS174. The established database contains the observed abundance of each identified protein, information relating to their hydrophobicity, the isoelectric point, molecular weight, and toxicity. These physicochemical properties were plotted on proteome property maps to showcase the selection of suitable purification strategies. Furthermore, sequence alignment allowed integration of subunit information and occurrences of post-translational modifications from the well-studied *E. coli* K12 strain.



2.1 Introduction

Throughout the history of biopharmaceutical production, the effective removal and detection of host cell protein (HCP) impurities from the final drug product have been the subject of intensive research and development [1]–[3]. The presence of such impurities can have adverse effects on patient safety or product stability when present in the final drug product. For example, significant amounts of HCP impurities could be linked to strong side effects of the recent ChAdOx1 nCoV-19 vaccine [4]. To minimise impacts on patients and improve product quality, the effective removal of such impurities is of utmost importance [3]. At the same time, the pressure to accelerate the process development of biopharmaceuticals , especially the downstream processing[5]–[7], is high. Vaccines in particular require accelerated development to ensure timely responses to emerging pandemics, which has only recently become evident with the COVID-19 outbreak and pandemic.

Impurities can originate from the process or the product itself (e.g. the degraded or aggregated form of the product). Process-related impurities originate from the host cell expression system used to produce the protein therapeutic. When host cells are disrupted to obtain the intracellular or periplasmic products, impurities from the host such as HCPs, DNA, RNA, and endotoxins are released. Therefore, extensive purification must be performed, where, the HCP content is reduced in every purification step until the target quality is reached (Figure 2.1). The structural and physicochemical properties of HCPs may closely resemble those of the protein therapeutic produced such that the elimination of such HCPs poses a significant challenge and is, therefore, the subject of extensive analytical development.

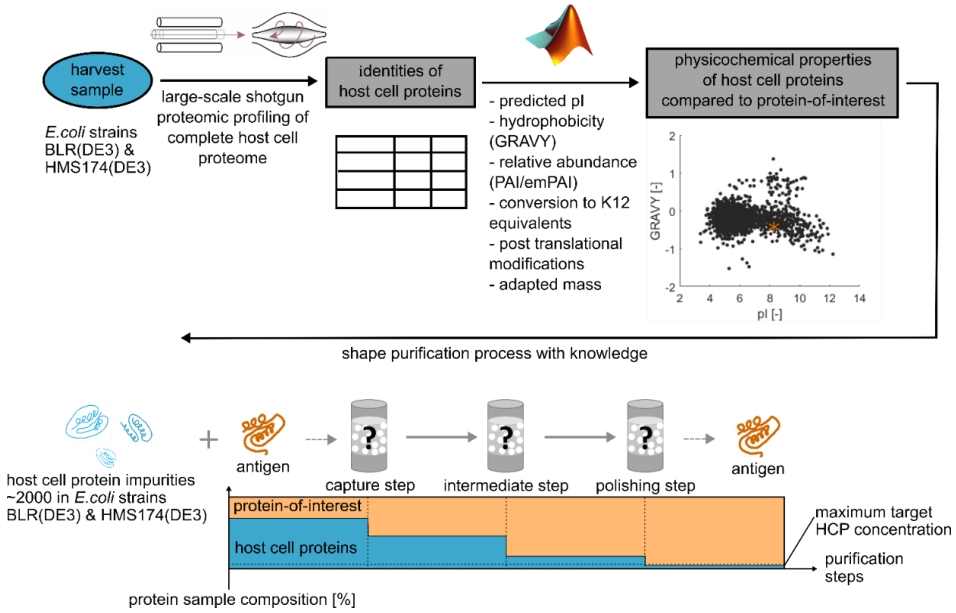


Figure 2.1: Schematic overview of the process development approach guided by large-scale host cell proteomics described in this study. The clarified harvest sample from the fermentation process was analysed using mass-spectrometry-based proteomics to identify all detectable HCPs. Further, a range of physicochemical properties was calculated for every possible gene product. One can guide the selection of the most suitable purification process by comparing the properties of protein therapeutics with those in the established database resource.

The acceptable levels of HCPs in vaccines are defined on a case-by-case basis by regulatory authorities [8]. For example, Zhu et al [9] investigated a malaria vaccine candidate expressed in *E. coli.* The total HCP concentration was specified to be 90 ng or $<$ 1100 ppm per dose in this case [10]. Tolerated HCP levels for vaccines are generally higher compared to those of drugs for chronic diseases ($<$ 100 ppm) [8].

Jones et al. identified high-risk, immunogenic, biologically active, or enzymatically active HCPs, which showed the potential to degrade either the product molecules or the excipients in the formulation [2]. Using this knowledge, Chiu et al. furthermore knocked out genes from CHO cells to prevent the expression of high-risk and difficult-to-remove HCPs [11]. The types of persistent HCP(s), however, not only depend on the employed host cell expression system, but also on the produced protein therapeutics. These may have very different physicochemical properties and therefore different critical HCPs than previously purified products.

Monitoring the purification process and measuring residual HCPs are the focus of intensive analytical development. Anti-HCP enzyme-linked immunosorbent assays (ELISAs) are the gold standard for determining overall HCP content to detection levels as low as 1 ng/mL [10], [12]. However, the ELISA technique can only detect proteins against which it is developed, and total protein ELISAs do not provide information on individual proteins present in the drug substance or product. Therefore, the use of orthogonal methods to support process development and validation is recommended [13].

Significant advancements in high-resolution mass spectrometry in recent decades have enabled large-scale proteomics with greater accuracy, sensitivity and throughput. Mass-spectrometry-based proteomics have emerged as a powerful alternative to identify and quantify HCPs to detection limits of up to 5 ppm for known and unknown components [14]. Consequently, host cell proteomics have been increasingly employed to monitor purification progress and to confirm the absence of specific HCPs in the final drug substance or product [8], [9], [13]–[16].

When a new purification process is designed, suitable chromatography resins and buffer conditions have to be identified. Three main process steps are commonly used in protein purification [15]. The first is the “capture step”, which serves as the gross purification step. The bulk of the impurities is removed, thereby concentrating the protein product. The subsequent intermediate purification steps use various chromatographic resins to further reduce impurities. Finally, the polishing step removes low-abundance and minor impurities [15]. Frequently applied chromatographic separation techniques are ion exchange, hydrophobic interaction, mixed mode, size-exclusion or affinity-based chromatography, where packed bed resins are currently state-of-the-art [6].

Identifying the most effective technique for the removal of HCPs is difficult without extensive experimental and predictive data. In particular, anticipating the presence of critical HCPs that are difficult to remove or that are retained by the product during processing remains challenging [12]. Currently, the development of new processes still requires expert knowledge and high-throughput screening approaches to identify suitable conditions for the development of effective purification steps [15], [17].Advanced process development tools are needed that use a more rational and

systematic approach [12], [18]. In previous work, mechanistic models have been used to describe the binding behaviour of HCP on several chromatographic columns[19]–[21]. Isotherm parameters of HCP were determined from the chromatographic separations. Alternatively, the affinity of process-related impurities (including HCPs) to a library of resins was described[22], [23].

Notably, extensive data are available on model organisms commonly employed in clinical and medical studies, such as *E. coli* K12 ,CHO cells or *Pichia pastoris*. Conversely, limited studies have been conducted on the proteomes of strains developed and optimised for biotechnological applications, including the widely employed host strains of *E. coli* BLR and HMS174. The advantages of comprehensively analyzing the proteome present in the harvest before the capture step is often overlooked. Knowledge of protein impurities, including their abundance and characteristics relative to the expressed protein therapeutic, can facilitate the development of an effective purification strategy.

In this study, we characterise the complete host cell proteome of two widely employed *E. coli* strains BLR and HMS174, using state-of-the-art Orbitrap mass spectrometry. The established proteomic data were further used to construct a database resource containing information regarding observed expression levels, hydrophobicity, isoelectric points (pI), molecular weights (MW), subunit information, possible post-translational modifications (PTMs), and toxicity for every possible gene product. The properties of the expressed protein therapeutics can then be evaluated in the context of the complete host cell proteome. This extensive resource generated by mass spectrometry analysis of the host cell proteome, therefore, leads to a more rational and accelerated purification process development. Furthermore, we exemplify the use of the database resource for purification process development of the capture step for two model antigens used in a protein subunit vaccine produced with the *E. coli* strains BLR and HMS174.

2.2 Material and methods

2.2.1 E. coli fermentation and harvest sample

The cultivation was performed as a standard fed-batch process using semi-synthetic media. Working seed for the pre-culture was first amplified in a

shake flask until it reached an OD_{650} of about 2.0. Then ca. 20 mL of pre-culture is added in a 20 L fermenter filled with 9 L of culture medium. In the first part of the fermentation bacterial biomass was produced in fed-batch mode taking approximately 18 hours to reach a volume of 12 L. Afterwards in the second phase of the fermentation, Isopropyl ß-D-1-thiogalactopyranoside (IPTG) was added to induce the production of the model antigen (same procedure in null plasmid strains). After 24 hours the fermentation harvest was obtained and clarified.

The harvest samples were derived from the *E. coli* strains BLR(DE3) and HMS174(DE3) (further called BLR and HMS174). For both strains a fermentation was conducted using an empty plasmid cassette which did not encode the gene of the antigen. These 2 samples from null plasmid cell lines were frozen at -80 °C before the clarification step. The third sample was obtained from the *E. coli* strain BLR producing the model antigen recombinantly. In the clarification, the *E. coli* cells in all samples were disrupted by homogenisation with a French pressure cell (*Sim Aminco* Spectronic Instruments) to obtain the intracellular soluble products. In the further clarification, the samples were centrifuged for 45 minutes at 15,000 *g* and filtered with a 0.2 µm PES filter. All harvest material for the analysis of the host cell proteome was provided by GSK (Rixensart, Belgium).

2.2.2 Sample preparation for host cell proteomic analysis

The E. coli host cell proteome samples from BLR and HMS174 were prepared in accordance to recently published protocols by den Ridder et al. [24]. An extended description is provided in the supplementary information.

2.2.3 Shotgun host cell proteomics

In the supplementary information the protocol using a nano-liquid-chromatography separation system consisting of an EASY-nLC 1200, equipped with an Acclaim PepMap RSLC RP C18 separation column (50 µm x 150 mm, 2 µm and 100 Å), and a QE plus Orbitrap mass spectrometer (Thermo Scientific, Germany) is described in detail. The Orbitrap was operated in data-dependent acquisition (DDA) mode. The mass spectrometry proteomics raw data for the null plasmid cell lines of *E. coli* strains BLR and HMS174, have been deposited in the ProteomeXchange consortium database with the dataset identifier PXD035590.

2.2.4 Processing of mass spectrometric raw data

Mass spectrometric raw data were analysed using PEAKS Studio X (Bioinformatics Solutions Inc., Canada) described in detail in the supplementary information. The mass spectrometric raw data were further analysed using strain specific proteome sequence databases obtained from NCBI (*E. coli* BLR: BioProject PRJNA379778 and *E. coli* HMS174 BioProject PRJEB6353) and the GPM crap contaminant proteins sequences (https://www.thegpm.org/crap/). Relative protein abundances (or content) was estimated using the protein abundance index (PAI) and the exponentially modified PAI (emPAI) according to Ishihama et al.[25]. Label free quantification of protein abundance changes between the null plasmid *E. coli* strain and the corresponding antigen producing strain was performed using the PEAKSQ module [26].

2.2.5 Construction of host cell proteome property databases for E.coli BLR and HMS174

The two databases accessible in the supplementary data were based on the mass spectrometry measurement of the clarified harvest samples originating from null plasmid cell lines from the *E. coli* strains BLR and HMS174. Each protein has a protein group, protein ID and accession assigned. The average mass, area and coverage is determined via the MS measurement and proteins are ranked according to their spectral count. PAI is defined as the number of sequenced peptides (fragmentation spectra assigned with significant score and as the top match to an individual identified protein) divided by the number of its calculated, observable peptides [27]. This value was used as the abundance measure in the comparison between proteomes. Furthermore, the PAI was converted to the emPAI, equal to 10^{PAI} minus one as described in reference [25]. With help of the emPAI, the protein content was calculated in molar percent and weight percent as described by Ishihama et al. [25]. Each individual protein was assigned their calculated physicochemical properties. Calculated pI, calculated charge and grand average of hydropathy (GRAVY) as a measure of hydrophobicity were chosen as properties that define the most useful separation mechanism. For this purpose an in-house Matlab program was written that sorted the proteins according to their accession and assigned the physicochemical parameter predicted based on the amino acid sequence. The isoelectric point was predicted using the Matlab function “isoelectric”

and the "Isoelectric Point Calculator 2.0" software [28], that predicts the pI based on 21 different models. The average pI of the different calculation methods was used in the plotted graphs thereafter. The charge of the proteins was calculated in Matlab with the function "isoelectric" based on the amino acid sequence of the protein. The hydrophobicity was extracted in form of the GRAVY based on the amino acid sequence of the HCP (http://www.gravy-calculator.de/). A GRAVY value below 0 describes a hydrophilic protein, while scores above 0 are describing hydrophobic proteins. The sum of GRAVY values of the amino acids in the protein sequence divided by the number of amino acids is used as the GRAVY value of the protein. The toxicity is predicted using the ToxinPred2 tool [29]. Selected machine learning technique was hybrid (RF + BLAST + MERCI) with a threshold value of 0.6. Protein subunit information and knowledge about possible occurrence of PTMs for *E. coli* BLR and HMS174 were inferred from the *E. coli* K12 strain, which proteome sequence was obtained from Uniprot reference proteome sequence database (UP000000625_83333). The alignment of sequences was performed for this purpose using the Diamond sequence aligner [30] where the quality of the match was assessed by considering % sequence identify and e-values.

2.2.6 Codes and functions used for visualization of host cell proteome properties

In-house Matlab scripts were used to plot the physicochemical properties of the identified proteins into property maps using scatter plots. The database including all proteins identified in the sample was used as input and the abundance was plotted over the mass, pI and GRAVY of the identified proteins. In the next step, the pI was plotted over the GRAVY and the charge at pH 7.0 over the GRAVY values. This analysis was conducted for all identified proteins, the 20 top abundant HCPs and the antigen properties.

2.3 Results and discussion

2.3.1 A comprehensive host cell proteome database for E. coli BLR and HMS174

Characterising the host cell proteome (i.e. protein abundances and predicted properties) is expected to streamline the development of purification processes significantly. Hence, we performed a proteomic characterisation

of the widely employed *E. coli* BLR and HMS174 strains and predicted the physicochemical properties for all possible gene products. For example, differences in isoelectric point (pI) and hydrophobicity (GRAVY) affect the selection of the most common chromatographic methods, which are ion-exchange chromatography (IEX) and hydrophobic interaction chromatography (HIC). The proteome database was further expanded with parameters such as protein coverage, area, and protein content indices (protein abundance index PAI and the exponentially modified protein abundance index emPAI). The most abundant proteins in the database for the BLR and HMS174 strains are presented in Table 1 and Table 2, respectively. The complete database for both strains is available in the supplementary material. From the 4,295 proteins of the complete proteome of *E. coli* BLR, 1,993 HCPs were detected in the null plasmid strain, and 2,006 were identified when additionally expressing the model antigen. In *E. coli* HMS174, 4,216 proteins are found in the theoretical proteome, of which 1,886 were detected in the null plasmid strain. Most of the abundant proteins have functions in biosynthesis or are ribosomal proteins. The most abundant protein in both strains, appeared to be the ATP synthase F1 subunit epsilon. This protein generates ATP from ADP in the presence of a proton gradient across the membranes. However, this protein is relatively small and has only one theoretically observable peptide (in the considered mass range 800–2,400 Da) according to the original definition of PAI [31]. Therefore, the observed peptides divided by the number of theoretically observable peptides provides disproportionally high PAI values. Furthermore, we linked all protein sequences to homologue counterparts of the well investigated model organism *E. coli* K12 using sequence alignment. This enabled inferring information about possible complex formation and occurrence of PTMs. The latter could alter the protein size and net charge. For *E. coli* BLR, 224 PTMs are listed in the database, while 221 PTMs are listed for *E. coli* HMS174.

Table 2.1: The top 20 HCPs (according to the PAI values) observed in the null plasmid fermentation using the E. coli strain BLR and their physicochemical properties.

Protein accession	Protein name	Avg. Mass [Da]	Area BLR [-]	PAI [-]	emPAI [-]	protein content [mol %]	protein content [weight %]	GRAVY [-]	Net charge pH 7.0 Matlab [-]	Average pI [-]	Accession closest *E coli* K12 analogue
ARH99394.1_3613	F1 sector of membrane-bound ATP synthase epsilon subunit	15,068	3.41E+08	5.00	99,999	73.03	63.65	-0.095	-4.6	5.4	P0A6E6
ARH98063.1_2282	phosphohistidinoprotein-hexose phosphotransferasecomponent of PTS system (Hpr)	9,119	1.04E+09	4.00	9,999	7.30	3.85	-0.166	-1.5	5.6	P0AA04
ARH98931.1_3150	50S ribosomal subunit protein L22	12,226	6.06E+09	3.25	1,777	1.30	0.92	-0.349	10.9	10.4	P61175
ARH99273.1_3492	50S ribosomal subunit protein L33	6,372	2.29E+09	3.00	999	0.73	0.27	-0.804	10.5	10.5	P0A7N9
ARH97584.1_1803	CopC family protein	13,410	1.34E+08	3.00	999	0.73	0.57	-0.262	5.4	9.5	P0AA57
ARH97040.1_1259	putative uncharacterized protein YciS	11,351	16,900,000	3.00	999	0.73	0.48	0.680	4.2	9.6	P0ACV4
ARH99646.1_3865	50S ribosomal subunit protein L7/L12	12,295	6.16E+09	2.75	561	0.41	0.29	0.295	-8.0	4.5	P0A7K2
ARH98925.1_3144	50S ribosomal subunit protein L24	11,316	8.61E+09	2.75	561	0.41	0.27	-0.381	11.2	10.5	P60624
ARH96394.1_613	glutamate/aspartate ABC transporter substrate-binding protein	33,375	7.26E+09	2.73	532	0.39	0.75	-0.445	3.4	8.4	P37902
ARH98288.1_2507	glycine betaine ABC transporter substrate-binding protein	36,046	1.28E+09	2.57	372	0.27	0.57	-0.320	-2.9	6.0	P0AFM2

Table 2.1: The top 20 HCPs (according to the PAI values) observed in the null plasmid fermentation using the E. coli strain BLR and their physicochemical properties.

Protein accession	Protein name	Avg. Mass [Da]	Area BLR [-]	PAI [-]	emPAI [-]	protein content [mol %]	protein content [weight %]	GRAVY [-]	Net charge pH 7.0 Matlab [-]	Average pI [-]	Accession closest *E coli* K12 analogue
ARH98851.1_3070	50S ribosomal subunit protein L13	16,019	8.02E+09	2.50	315	0.23	0.21	-0.540	11.7	10.1	P0AA10
ARH98514.1_2733	fructose-bisphosphate aldolase class II	39,147	4.59E+09	2.50	315	0.23	0.52	-0.224	-9.0	5.5	P0AB71
ARH96740.1_959	methylglyoxal synthase	16,919	63000000	2.50	315	0.23	0.23	0.038	-1.4	6.2	P0A731
ARH99864.1_4083	30S ribosomal subunit protein S6	15,173	6.41E+09	2.40	250	0.18	0.16	-0.745	-6.6	5.1	P02358
ARH96687.1_906	30S ribosomal protein S1	61,158	1.64E+10	2.33	214	0.16	0.55	-0.300	-27.1	4.7	P0AG67
ARH98856.1_3075	malate dehydrogenase NAD(P)-binding	32,337	7.68E+09	2.33	214	0.16	0.29	0.194	-2.6	5.5	P61889
ARH98791.1_3010	30S ribosomal subunit protein S15	10,269	5.8E+09	2.33	214	0.16	0.09	-0.673	8.2	10.6	P0ADZ4
ARH97768.1_1987	galactitol-specific enzyme IIB component of PTS	10,270	1.7E+09	2.33	214	0.16	0.09	0.291	-1.4	5.8	P37188
ARH99054.1_3273	aspartate-semialdehyde dehydrogenase NAD(P)-binding	40,034	9.87E+08	2.33	214	0.16	0.36	-0.040	-5.9	5.2	P0A9Q9
ARH98920.1_3139	50S ribosomal subunit protein L18	12,770	3.64E+09	2.25	177	0.13	0.10	-0.395	10.7	10.6	P0C018

Table 2.2: The top 20 most abundant HCPs (according to the PAI values) identified in the null plasmid fermentation of the E. coli strain HMS174 and their physicochemical properties. he top 20 HCPs (according to the PAI values) observed in the null plasmid fermentation using the E. coli strain BLR and their physicochemical properties.

Accession	Protein name	Avg. Mass [Da]	Area HMS [-]	PAI [-]	emPAI [-]	protein content [mol %]	protein content [weight %]	GRAVY [-]	Net charge pH 7.0 Matlab [-]	Average pI [-]	Accession closest *E coli* K12 analogue
CDY62850.1	ATP synthase F1 complex-epsilon subunit subunit of ATP synthase F1 complex	15,068	3.1E+08	6.00	999,999	95.69	93.53	-0.095	-4.6	5.4	P0A6E6
CDY60193.1	phosphohistidinoprotein-hexose phosphotransferasecomponent of PTS system (Hpr)	9,119	4.81E+09	4.00	9,999	0.96	0.57	-0.166	-1.5	5.6	P0AA04
CDY64403.1	major type 1 subunit fimbrin (pilin) subunit of fimbrial complex	18,111	7.18E+07	4.00	9,999	0.96	1.12	0.310	-2.6	4.9	P04128
CDY56811.1	methylglyoxal synthase	16,919	3.63E+08	3.50	3161	0.30	0.33	0.038	-1.4	6.2	P0A731
CDY58853.1	conserved protein	13,410	4.22 E+08	3.00	999	0.10	0.08	-0.262	5.4	9.5	P0AA57
CDY56069.1	conserved protein involved in translation	17,526	1.06 E+08	3.00	999	0.10	0.11	-0.311	-22.6	4.0	P0A898
CDY63707.1	50S ribosomal subunit protein L22 subunit of ribosome	12,226	1.39E+09	2.75	561	0.05	0.04	-0.349	10.9	10.4	P61175
CDY56061.1	glutamate / aspartate ABC transporter-periplasmic binding protein subunit of GltIJKL glutamate ABC transporter	33,420	3.56E+09	2.73	533	0.05	0.11	-0.472	3.4	8.4	P37902
CDY61370.1	fructose bisphosphate aldolase class II	39,147	5.41E+09	2.50	315	0.03	0.08	-0.224	-9.0	5.5	P0AB71
CDY63356.1	universal stress global stress response regulator	16,066	4.35E+09	2.50	315	0.03	0.03	-0.056	-7.4	4.9	P0AED0

Table 2.2: The top 20 most abundant HCPs (according to the PAI values) identified in the null plasmid fermentation of the E. coli strain HMS174 and their physicochemical properties. he top 20 HCPs (according to the PAI values) observed in the null plasmid fermentation using the E. coli strain BLR and their physicochemical properties.

Accession	Protein name	Avg. Mass [Da]	Area HMS [-]	PAI [-]	emPAI [-]	protein content [mol %]	protein content [weight %]	GRAVY [-]	Net charge pH 7.0 Matlab [-]	Average pI [-]	Accession closest *E coli* K12 analogue
CDY57666.1	lipid hydroperoxide peroxidase	17,835	6.28E+09	2.40	250	0.02	0.03	0.256	-4.9	4.6	P0A862
CDY56210.1	SucB subunit of dihydrolipoyltranssuccinylase and 2-oxoglutarate dehydrogenase complex	44,011	5.43E+09	2.36	230	0.02	0.06	-0.217	-6.3	5.5	P0AFG6
CDY62241.1	malate dehydrogenase	32,337	1.47E+10	2.33	214	0.02	0.04	0.194	-2.6	5.5	P61889
CDY62031.1	30S ribosomal subunit protein S15 subunit of ribosome	10,269	1.78E+09	2.33	214	0.02	0.01	-0.673	8.2	10.6	P0ADZ4
CDY62519.1	superoxide dismutase (Mn)	23,097	2.17E+09	2.29	192	0.02	0.03	-0.429	-0.1	6.5	P00448
CDY57747.1	nucleotide binding filament protein	16,017	5.19E+08	2.20	157	0.02	0.02	0.022	-4.4	5.6	P37903
CDY60191.1	CysK subunit of cysteine synthase A and bifunctional CysEK cysteine biosynthesis complex	34,490	1.55E+10	2.17	146	0.01	0.03	-0.078	-2.3	5.8	P0ABK5
CDY62230.1	50S ribosomal subunit protein L13 subunit of ribosome	16,019	2.02E+09	2.17	146	0.01	0.01	-0.540	11.7	10.1	P0AA10
CDY60673.1	phage lambda replication; host DNA synthesis; heat shock protein; protein repair subunit of DnaJ/DnaK/GrpE	21,798	3.17E+08	2.17	146	0.01	0.02	-0.372	-14.3	4.5	P09372
CDY57250.1	isocitrate dehydrogenase	45,785	2.3E+10	2.16	143	0.01	0.04	-0.154	-11.0	5.0	P08200

2.3.2 HCP differences between E. coli strains, null plasmid and antigen expressing strains

Furthermore, we compared the proteome and expression pattern between the BLR and HMS174 (null plasmid) strains. Out of all the identified proteins, approximately 80 % (1,590 proteins) were detected in both strains. A correlation graph using the abundance values (expressed by the PAI metric) provided for a linear regression an R^2 of 0.69 (Figure 2.2a). This overlap shows that the bulk amount of HCPs are comparable even between different *E. coli* strains. Furthermore, we compared the identified proteins and abundances between the BLR null plasmid and the corresponding antigen-expressing strain. Here, approximately 90 % (1,779) of the identified proteins were identical in both samples. After plotting the abundances of the observed proteins, an R^2 value of 0.81 was obtained (Figure 2.2b). Differences in the abundances, however, may also be partly due to slight differences in the sample preparation procedure (e.g. the antigen-containing harvest was exposed to one freeze/thaw cycle before the clarification step). Nevertheless, the expression of the antigen is expected to have some impact on the observed host cell proteome. The differences may be minor and the findings from the null strain can be applied to determine a purification strategy for the antigen-producing strain.

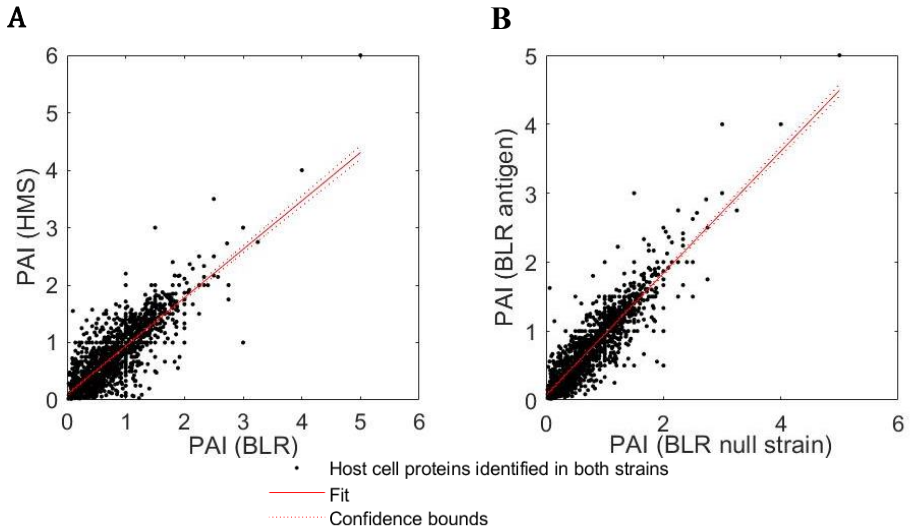


Figure 2.2: Scatterplots of the HCPs identified in the investigated E. coli strains. (A) presents a comparison of null plasmid E. coli strains BLR and HMS174. The correlation of 1,590 proteins that were common to both strains resulted in an R^2 value of 0.6899. In (B), the null plasmid BLR was compared to the corresponding antigen-expressing strain; 1,779 proteins were common to both samples, resulting in an R^2 value of 0.8078.

2.3.3 Visualizing the host cell proteome using global property maps

The properties of the host cell proteomes were further visualised using proteome property maps. The use of global property maps can be an effective tool for designing an optimal purification strategy [19]. For example, differences in the properties between the most abundant HCPs (or critical HCPs) and the antigen allow identification of the most promising resins for the first purification step. In the following subsequent sections various property maps (abundance versus pI/GRAVY; pI versus GRAVY; and net charge versus GRAVY) are discussed. The data of two model antigens are shown and possible purification strategies for the capture step are discussed based on differences between the antigens and the most abundant proteins.

2.3.4 Abundances versus molecular weight (MW), isoelectric point (pI) and hydrophobicity (GRAVY)

The (null plasmid) BLR and HMS174 strains were compared based on properties such as MW (mass), pI, and GRAVY. Utilizing this approach enabled the search for conditions in which the majority of the HCPs differ from the expressed protein therapeutics (in this case, antigens). The properties of “antigen 1” expressed in BLR and “antigen 2” expressed in HMS174 are shown in the graph in relation to the properties of the HCPs (Figure 2.3) to define a purification strategy. Both strains show similar distributions of abundances compared to their protein properties, which is unsurprising, as a large number of proteins are identified in both strains with relatively similar abundances.

The MWs of the HCPs vary between 2 and 250 kDa, with the majority of proteins having a MW $<$ 50 kDa (Figure 2.3a and 2.3d). The high-abundance proteins are in the lower MW range. Antigen 1 has a MW of 59 kDa, while antigen 2 (28 kDa) is comparatively small. Separating the antigens from the host cell proteins with a separation mechanism based on the size of the molecules, for example size exclusion chromatography (SEC), seem to be suitable for later purification steps [15]. The discrepancy between mass of abundant HCP to the antigens seems to be a poor separation property for the capture step.

The pI spectrum of the identified HCPs ranges from pH 3.4–12.2, where the majority of the proteins are acidic (Figure 2.3b and 2.3e). A trough with fewer proteins is visible between a pI of 7 and 8. This trough can be explained

by the intracellular pH for *E. coli* (approx. pH = 7.5) that would decrease the stability of proteins with a similar pI. Antigen 1 is located at the lower end of the pI spectrum with a pI of 4.4, while antigen 2 is close to the trough with fewer identified proteins with a pI of 8.4. Both antigens have pIs that are significantly different from the HCPs. One could consider a separation based on charge, such as IEX, as a promising capture step.

The estimated GRAVY values of the proteins range from -1.526 to +1.369. Most of the identified proteins have a slightly negative GRAVY value and are, hence, slightly hydrophilic (Figure 2.3c and 2.3f). Antigen 1 has a GRAVY of -0.749, which is relatively different to the values obtained for most HCPs. For antigen 1, a separation based on hydrophobicity (e.g., using HIC) therefore appears highly promising as a capture step.

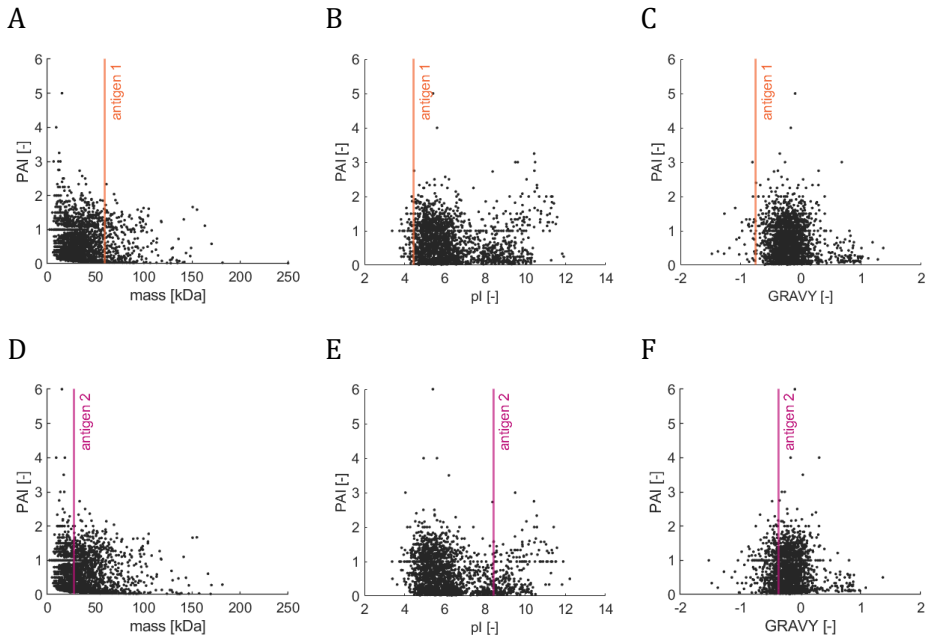


Figure 2.3: Abundances of the detected HCPs from null plasmid fermentations of the E. coli strains BLR (A–C) and HMS174 (D–F) are compared: (A and D) the mass of the proteins according to mass spectrometric measurements, (B and E) the average predicted isoelectric points (pI), and (C and F) hydrophobicity (GRAVY). Positive and negative GRAVY values describe hydrophobic and hydrophilic proteins, respectively. The model antigens 1 and 2 are indicated in red and purple. The abundances are expressed by the PAI parameter.

2.3.5 Isoelectric point (pI) versus hydrophobicity (GRAVY)

We furthermore plotted the predicted pI against the hydrophobicity (GRAVY) of the identified host cell proteome, as shown in Figures 2.4a and 2.4c. Additionally, we generated a plot for the 20 most abundant HCPs and model antigens (Figures 2.4b and 2.4d, also listed in Table 1 and Table 2). In this example case, it was chosen to focus on the most abundant HCPs in the sample to design a capture step targeting the removal of the main HCP impurities. Antigen 1 has alow pI and GRAVY value compared to the most abundant HCPs. IEX together with HIC, or their combination in mixed mode chromatography, appear highly suitable for purifying this antigen.

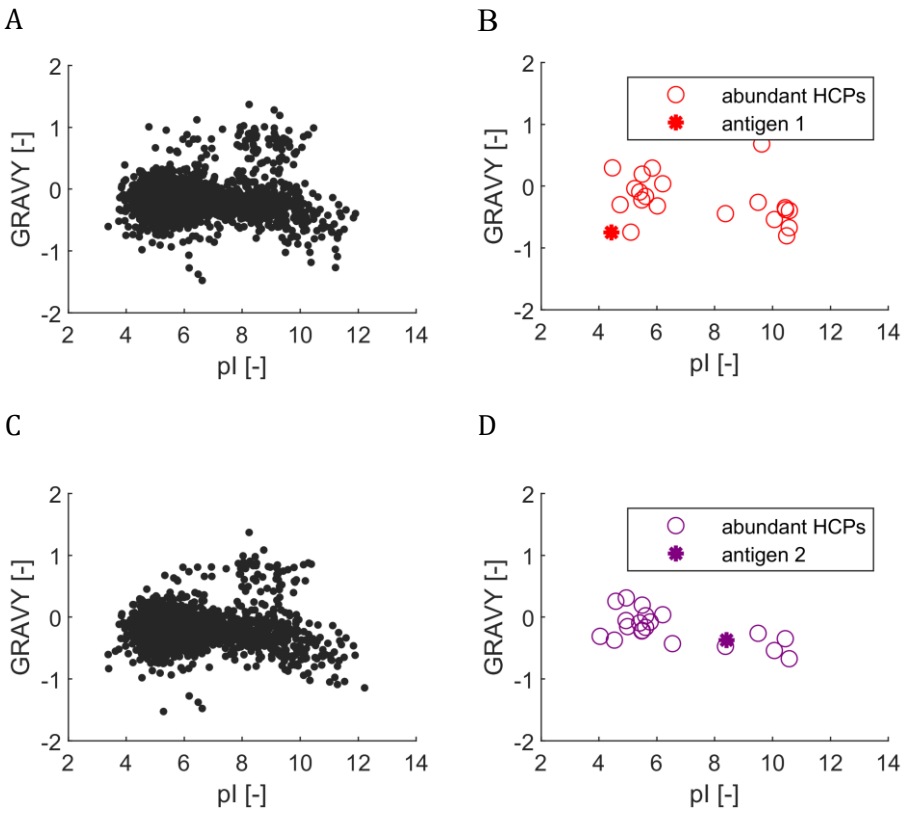


Figure 2.4.: Comparison of HCPs from the null plasmid fermentations of the E. coli strains BLR (A and B) and HMS174 (C and D). The predicted hydrophobicity (GRAVY) is plotted against the predicted isoelectric point (pI). Displayed are (A) the properties of the complete, identified host cell proteome of BLR; (B) the properties of the most abundant HCPs and antigen 1 in BLR; (C) the properties of the complete, identified host cell proteome of HMS174; and (D) the properties of the most abundant HCPs and antigen 2 in HMS174.

Antigen 2 is located in close proximity to the centre of the pI spectrum of the HCPs. However, apart from the glutamate/aspartate ABC transporter periplasmic binding protein, the most abundant HCPs have a significantly different pI. IEX appears to be a suitable purification method. The GRAVY value of antigen 2, on the other hand, is not significantly different to the values of the most abundant HCPs.

2.3.6 Net charge versus hydrophobicity (GRAVY)

The net charge of a protein depends on the pI and the pH value of the environment (solvent or buffer). Therefore, knowing the net charge of the HCPs at different pH values helps in selecting the most suitable conditions when using IEX. Plots at a pH of 7.0 were generated so that typically no buffer exchange (or pH adjustment) is required before the capture step, thus reducing time and costs e.g. for titration. We calculated the net charge of the HCPs at pH 7.0 and we plotted them against the predicted GRAVY values, which is shown in Figure 2.5. Net charges for a range of different pH conditions are furthermore included in the database resource of the supplementary information material.

In the case of BLR, 11 of the 20 most abundant proteins have a negative net charge at pH 7.0. Antigen 1 has a predicted net charge of -46.78, which is low compared to that of the other HCPs. Considering a bind-and-elute mode, anion-exchange chromatography at pH 7.0 seems highly suitable for the capture step. The other abundant HCPs with a positive net charge would be repelled by the ligands and would not bind to the resin under the identified conditions. The 11 negatively charged HCPs would bind to the resin at pH 7.0 but could be eluted earlier using (low) salt-washing steps. A flow-through mode, on the other hand, seems suboptimal for this antigen at the specified pH. However, this approach might be suitable at pH values lower than the antigen pI. The majority of the abundant proteins in HMS174 – 15 out of 20 proteins – are negatively charged at pH 7.0. Antigen 2, on the other hand, has a slightly positive charge. In the case of a bind-and-elute mode, a cation-exchange step, combined with a salt elution step (at low ionic strength) to elute the antigen, could be suitable. Another option would be the use of an anion exchange resin in flow-through mode.

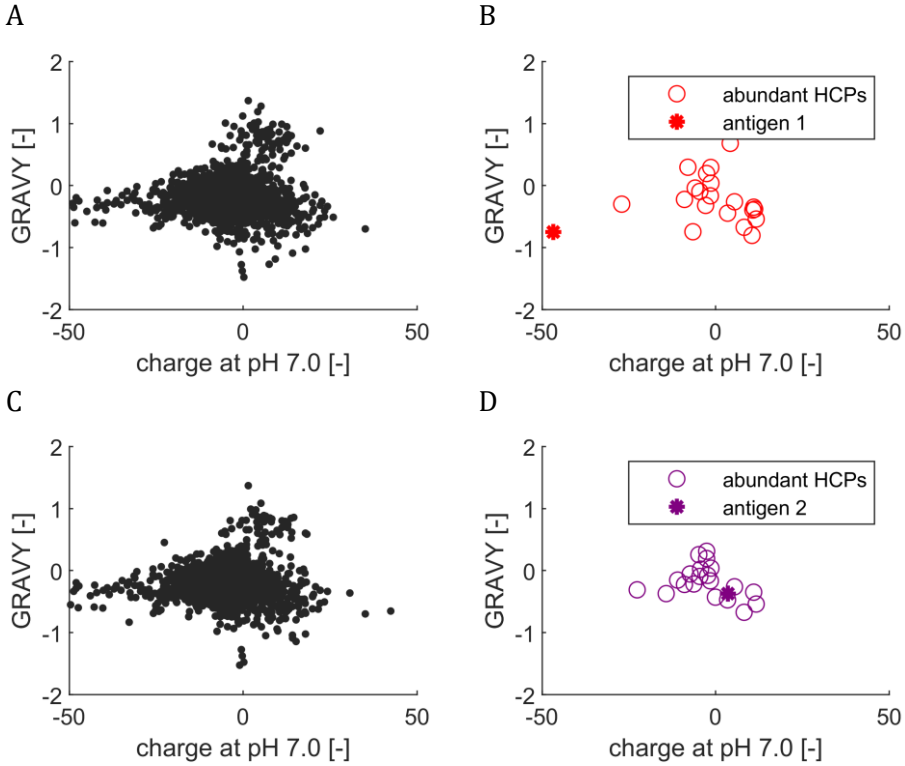


Figure 2.5: Comparison of HCPs from the null plasmid fermentations of the E. coli strains BLR (A and B) and HMS174 (C and D) . The predicted hydrophobicity (GRAVY) is plotted against the predicted net charge at pH 7.0 . Displayed are (A) the properties of the complete, identified host cell proteome of BLR; (B) the properties of the most abundant HCPs and antigen 1 in BLR; (C) the properties of the complete, identified host cell proteome of HMS174; and (D) the properties of the most abundant HCPs and antigen 2 in HMS174.

2.4 Conclusions

The avoidance and removal of HCP impurities when purifying protein targets is particularly challenging. Characterising protein abundances and physicochemical properties enables a more rational, systematic, and accelerated development of the purification process. In this study, we performed a comprehensive characterisation of the complete host cell proteome for the widely employed *E. coli* strains BLR and HMS174. Furthermore, we constructed an extensive proteome property resource by integrating physicochemical properties such as hydrophobicity (GRAVY), calculated pI, and the predicted net charge at different pH values. Additionally, we determined PAI and emPAI parameters to estimate protein abundances and relative protein content. We then linked proteins with homologues of the well-investigated *E. coli* K12 strain shedding light on

possible PTMs and complex formation. Furthermore, the protein abundances of null plasmid and antigen-expressing strains were compared, which demonstrated high similarity for the most abundant proteins.

We demonstrated the use of the established proteome resource database by creating global proteome property maps to support the design of new purification processes (or in particular to select the most promising capture step). This avoids extensive trial-and-error studies and sole expert-knowledge-dependent choices.

2.5 References

[1] D. G. Bracewell, R. Francis, and C. M. Smales, "The future of host cell protein (HCP) identification during process development and manufacturing linked to a risk-based management for their control," *Biotechnol. Bioeng.*, vol. 112, no. 9, pp. 1727–1737, 2015.

[2] M. Jones *et al.*, "'High-risk' host cell proteins (HCPs): A multi-company collaborative view," *Biotechnol. Bioeng.*, vol. 118, no. 8, pp. 2870–2885, Aug. 2021.

[3] M. Vanderlaan, J. Zhu-Shimoni, S. Lin, F. Gunawan, T. Waerner, and K. E. Van Cott, "Experience with host cell protein impurities in biopharmaceuticals," *Biotechnol. Prog.*, vol. 34, no. 4, pp. 828–837, Jul. 2018.

[4] L. Krutzke, R. Roesler, and S. Wiese, "Process-related impurities in the ChAdOx1 nCoV-19 vaccine," *Res. Sq.*, 2021.

[5] A. T. Hanke and M. Ottens, "Purifying biopharmaceuticals: Knowledge-based chromatographic process development," *Trends Biotechnol.*, vol. 32, no. 4, pp. 210–220, 2014.

[6] U. Gottschalk, K. Brorson, and A. A. Shukla, "The need for innovation in biomanufacturing," *Nat. Biotechnol.*, vol. 30, no. 6, pp. 489–492, 2012.

[7] H. Narayanan *et al.*, "Bioprocessing in the Digital Age: The Role of Process Models," *Biotechnol. J.*, vol. 15, no. 1, pp. 1–10, 2020.

[8] K. Reiter, M. Suzuki, L. R. Olano, and D. L. Narum, "Host cell protein quantification of an optimized purification method by mass spectrometry," *J. Pharm. Biomed. Anal.*, vol. 174, pp. 650–654, 2019.

[9] D. Zhu, A. J. Saul, and A. P. Miles, "A quantitative slot blot assay for host cell protein impurities in recombinant proteins expressed in E. coli," *J. Immunol. Methods*, vol. 306, no. 1–2, pp. 40–50, Nov. 2005.

[10] A. L. Tscheliessnig, J. Konrath, R. Bates, and A. Jungbauer, "Host cell protein analysis in therapeutic protein bioprocessing - methods and applications," *Biotechnol. J.*, vol. 8, no. 6, pp. 655–670, 2013.

[11] J. Chiu, K. N. Valente, N. E. Levy, L. Min, A. M. Lenhoff, and K. H. Lee, "Knockout of a difficult-to-remove CHO host cell protein, lipoprotein lipase, for improved polysorbate stability in monoclonal antibody formulations," *Biotechnol. Bioeng.*, vol. 114, no. 5, pp. 1006–1015, May 2017.

[12] C. E. M. Hogwood, D. G. Bracewell, and C. M. Smales, "Measurement and control of host cell proteins (HCPs) in CHO cell bioprocesses," *Curr. Opin. Biotechnol.*, vol. 30, no. July, pp. 153–160, 2014.

[13] H. Falkenberg *et al.*, "Mass spectrometric evaluation of upstream and downstream process influences on host cell protein patterns in biopharmaceutical products," *Biotechnol. Prog.*, vol. 35, no. 3, p. e2788, May 2019.

[14] S. Eliuk and A. Makarov, "Evolution of Orbitrap Mass Spectrometry Instrumentation," *Annu. Rev. Anal. Chem.*, vol. 8, pp. 61–80, 2015.

[15] E. Wen, R. Ellis, and N. S. Pujar, Eds., *Vaccine Development and Manufacturing*, First. Wiley, 2015.

[16] V. Reisinger, H. Toll, R. Ernst, J. Visser, and F. Wolschin, "A mass spectrometry-based approach to host cell protein identification and its application in a comparability exercise," *Anal. Biochem.*, vol. 463, pp. 1–6, 2014.

[17] P. S. Wierling, R. Bogumil, E. Knieps-Grünhagen, and J. Hubbuch, "High-throughput screening of packed-bed chromatography coupled with SELDI-TOF MS analysis: monoclonal antibodies versus host cell protein," *Biotechnol. Bioeng.*, vol. 98, no. 2, pp. 440–450, Oct. 2007.

[18] D. Keulen, G. Geldhof, O. Le Bussy, M. Pabst, and M. Ottens, "Recent advances to accelerate purification process development: A review with a focus on vaccines," *J. Chromatogr. A*, vol. 1676, p. 463195, Aug. 2022.

[19] B. K. Nfor *et al.*, "Multi-dimensional fractionation and characterization of crude protein mixtures: Toward establishment of a database of protein purification process development parameters," *Biotechnol. Bioeng.*, vol. 109, no. 12, pp. 3070–3083, Dec. 2012.

[20] A. T. Hanke *et al.*, "3D-liquid chromatography as a complex mixture characterization tool for knowledge-based downstream process development," *Biotechnol. Prog.*, vol. 32, no. 5, pp. 1283–1291, 2016.

[21] S. M. Pirrung *et al.*, "Chromatographic parameter determination for complex biological feedstocks," *Biotechnol. Prog.*, vol. 34, no. 4, pp. 1006–1018, 2018.

[22] S. M. Timmick *et al.*, "An impurity characterization based approach for the rapid development of integrated downstream purification processes," *Biotechnol. Bioeng.*, vol. 115, no. 8, pp. 2048–2060, 2018.

[23] N. Vecchiarello *et al.*, "A combined screening and in silico strategy for the rapid design of integrated downstream processes for process and product-related impurity removal," *Biotechnol. Bioeng.*, vol. 116, no. 9, pp. 2178–2190, Sep. 2019.

[24] M. den Ridder, E. Knibbe, W. van den Brandeler, P. Daran-Lapujade, and M. Pabst, "A systematic evaluation of yeast sample preparation protocols for spectral identifications, proteome coverage and post-isolation modifications," *J. Proteomics*, vol. 261, no. January, p. 104576, 2022.

[25] Y. Ishihama *et al.*, "Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein," *Mol. Cell. Proteomics*, vol. 4, no. 9, pp. 1265–1272, 2005.

[26] M. den Ridder, P. Daran-Lapujade, and M. Pabst, "Shot-gun proteomics: Why thousands of unidentified signals matter," *FEMS Yeast Res.*, vol. 20, no. 1, pp. 1–9, 2020.

[27] J. Rappsilber, U. Ryder, A. I. Lamond, and M. Mann, "Large-scale proteomic analysis of the human spliceosome," *Genome Res.*, vol. 12, no. 8, pp. 1231–1245, 2002.

[28] L. P. Kozlowski, "IPC 2.0: prediction of isoelectric point and p K a dissociation constants," *Nucleic Acids Res.*, vol. 49, no. W1, pp. W285–W292, Jul. 2021.

[29] N. Sharma, L. D. Naorem, S. Jain, and G. P. S. Raghava, "ToxinPred2: an improved method for predicting toxicity of proteins," *Brief. Bioinform.*, May 2022.

[30] B. Buchfink, C. Xie, and D. H. Huson, "Fast and sensitive protein alignment using DIAMOND," *Nat. Methods*, vol. 12, no. 1, pp. 59–60, 2014.

2.6 Appendix

2.6.1 Mass spectrometric methods

Sample preparation for host cell proteomic analysis

The E. coli host cell proteome samples from BLR and HMS174 were prepared in accordance to recently published protocols by den Ridder et al. [24]. An extended description is provided in the supplementary information. Briefly, 0.5 µg of bovine serum albumin (BSA) was spiked to every lysate and then trichloroacetic acid (TCA) was added at a 1:4 ratio to precipitate the proteins. Samples were incubated at 4 °C for 20 minutes and then centrifuged at 14,000 rcf for 15 minutes, and the supernatant was removed. The protein pellet was washed with acetone, centrifuged at 14,000 rcf for 15 minutes, and the supernatant removed. The pellet was re-dissolved in 100 µL of 6 M urea by vortexing. 30 µL of 10 mM dithiothreitol (DTT) was added to the samples and incubated for 60 minutes at 37 °C under gentle shaking (300 rpm). Further, alkylation was performed by adding 30 µL of 20 mM iodoacetamide (IAA), which was incubated in the dark for 30 minutes. The samples were then diluted with 200 mM ammonium bicarbonate (ABC) to reach an urea concentration of <1 M. Proteolytic digestion was performed using trypsin (Promega) at a ratio of 1:25 trypsin:protein at 37 °C incubated overnight. Obtained peptides were cleaned by solid phase extraction using an Oasis HLB 96-well Plate (30 µm particle size, Waters). The resin was conditioned with methanol and equilibrated with water. The samples were loaded onto the elution plate and washed with 5 % methanol in water. The peptides were eluted in 2 steps, using 200 µL of 2 % formic acid in 80 % methanol and second with 200 µL of 10 mM ABC in 80 % methanol. The

eluates were combined and placed into a speedvac concentrator until dryness (using heating at 50 °C for 1 hour).

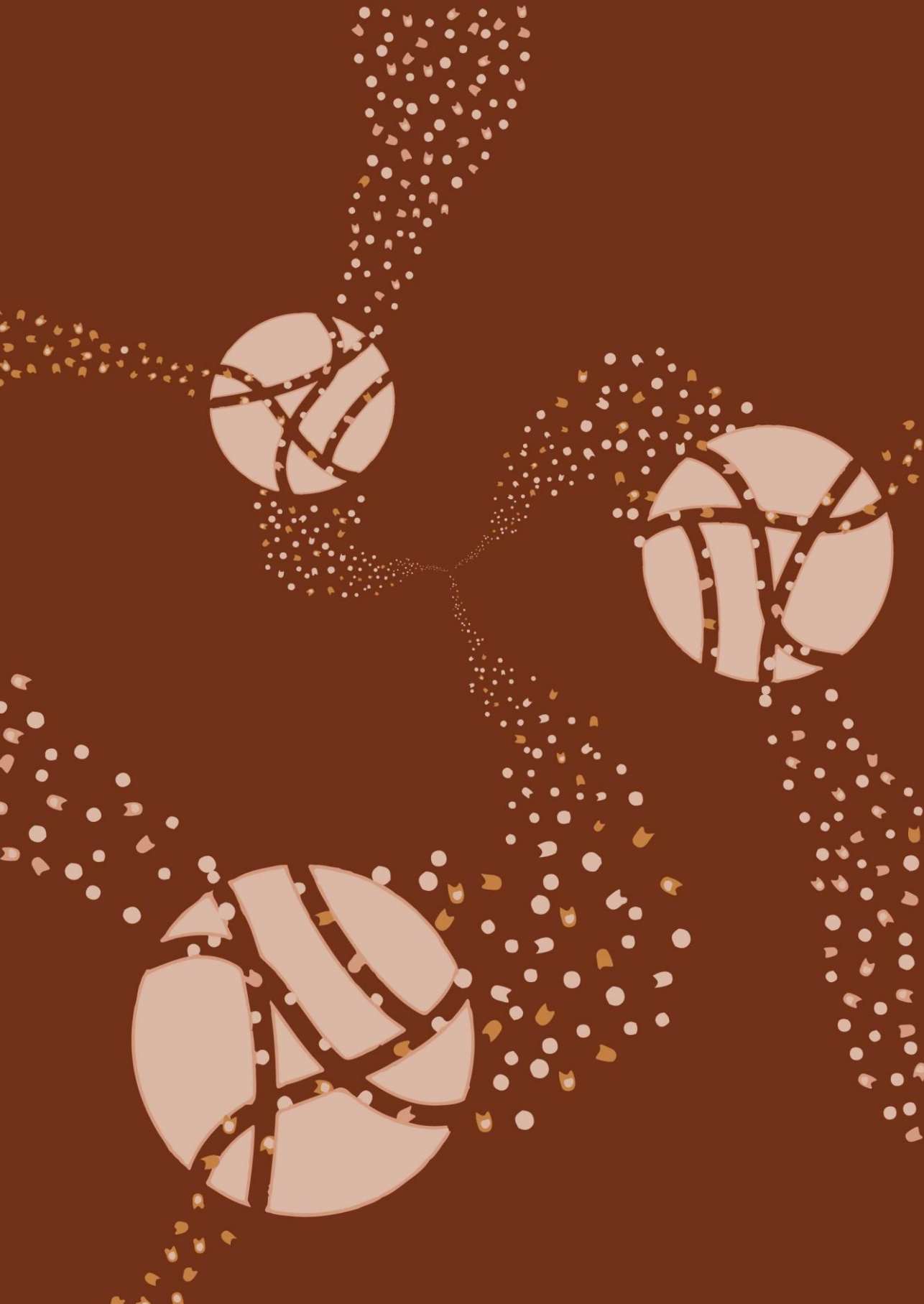
Shotgun host cell proteomics

The speedvac dried peptide fractions were resuspended in water containing 3 % acetonitrile and 0.01 % trifluoroacetic acid (TFA). An aliquot corresponding to approximately 500 ng digest were analysed using a nano-liquid-chromatography separation system consisting of an EASY-nLC 1200, equipped with an Acclaim PepMap RSLC RP C18 separation column (50 µm x 150 mm, 2 µm and 100 Å), and a QE plus Orbitrap mass spectrometer (Thermo Scientific, Germany). The flow rate was maintained at 350 nL/minutes with solvent A water containing 0.1 % formic acid, and solvent B consisted of 80 % acetonitrile in water and 0.1 % formic acid. A gradient consisting of a linear increase of solvent B from 5 to 25 % within 88 minutes, and finally to 55 % over 30 minutes. The Orbitrap was operated in data-dependent acquisition (DDA) mode acquiring spectra at 70 K resolution from 385–1,250 m/z, where the top 10 signals were isolated with a window 2.0 m/z and 0.1 m/z isolation offset, for fragmentation using a normalized collision energy (NCE) of 28. Fragmentation spectra were acquired at 17 K resolution, with an automatic gain control (AGC) target of 2e5, at a maximum injection time (IT) of 75 ms. Unassigned, singly charged, 6x and higher charge states were excluded from fragmentation. Dynamic exclusion was set to 60 seconds. The mass spectrometry proteomics raw data for the null plasmid cell lines of E. coli strains BLR and HMS174, have been deposited in the ProteomeXchange consortium database with the dataset identifier PXD035590.

Processing of mass spectrometric raw data

Mass spectrometric raw data were analysed using PEAKS Studio X (Bioinformatics Solutions Inc., Canada) allowing 20 ppm parent ion and 0.02 Da fragment ion mass error tolerance, considering 3 missed cleavages, carbamidomethylation as fixed and methionine oxidation and N/Q deamidation and N-terminal acetylation as variable modifications. The mass spectrometric raw data were further analysed using strain specific proteome sequence databases obtained from NCBI (E. coli BLR: BioProject PRJNA379778 and E. coli HMS174 BioProject PRJEB6353) and the GPM crap

contaminant proteins sequences (https://www.thegpm.org/crap/). Every sequence database contained additionally the sequence for BSA, which was spiked to every sample as process control. Additionally, decoy fusion was used for estimating false discovery rates (FDRs). Peptide spectrum matches were filtered against 1 % FDR and proteins with > 1 unique peptide sequences were considered significant.



Chapter 3

Experimental characterization and prediction of *Escherichia coli* host cell proteome retention during preparative chromatography

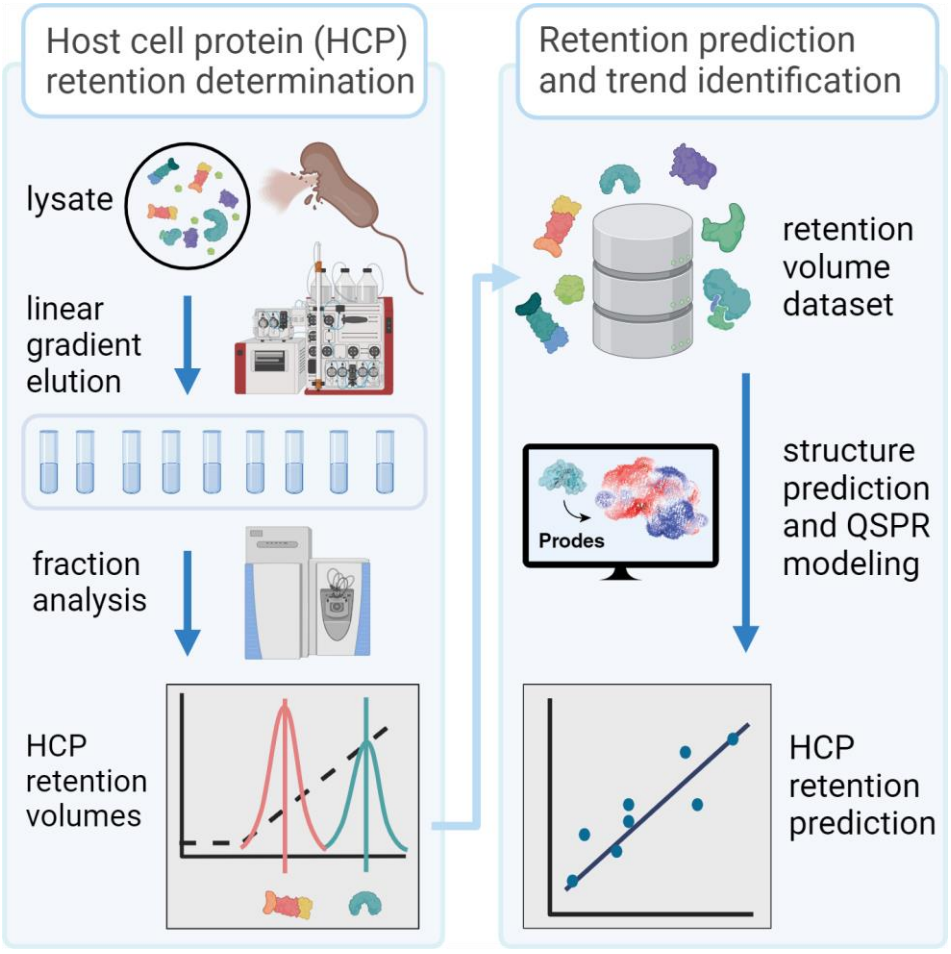
This chapter has been published as:

Disela, R., Neijenhuis, T., Le Bussy, O., Geldhof, G., Klijn, M., Pabst, M. Ottens, M., Experimental characterization and prediction of Escherichia coli host cell proteome retention during preparative chromatography, *Biotechnology and Bioengineering*, 121 (12) (2024), pp. 3848-3859,

https://doi.org/10.1002/bit.28840

Abstract

Purification of recombinantly produced biopharmaceuticals involves removal of host cell material, such as host cell proteins (HCPs). For lysates of the common expression host *Escherichia coli* (*E. coli*) over 1500 unique proteins can be identified. Currently, understanding the behavior of individual HCPs for purification operations, such as preparative chromatography, is limited. Therefore, we aim to elucidate the elution behavior of individual HCPs from *E. coli* strain BLR(DE3) during chromatography. Understanding this complex mixture and knowing the chromatographic behavior of each individual HCP improves the ability for rational purification process design. Specifically, linear gradient experiments were performed using ion exchange (IEX) and hydrophobic interaction chromatography, coupled with mass spectrometry-based proteomics to map the retention of individual HCPs. We combined knowledge on protein location, function and interaction available in literature to identify trends in elution behavior. Additionally, quantitative structure-property relationship models were trained relating the protein 3D structure to elution behavior during IEX. For the complete dataset a model with a cross validated R^2 of 0.55 was constructed, that could be improved to a R^2 of 0.70 by considering only monomeric proteins. Ultimately this study is a significant step towards greater process understanding.


Host cell protein (HCP) retention determination
lysate
linear gradient elution
fraction analysis
HCP retention volumes
Retention prediction and trend identification
retention volume dataset
Prodes
structure prediction and QSPR modeling
HCP retention prediction

3.1 Introduction

To ensure drug safety and efficacy, removal of impurities is essential. For protein-based pharmaceuticals (e.g., protein-based vaccines and monoclonal antibodies (mAbs)), removal of host cell proteins (HCPs) remains a major challenge [1]. Especially for recombinant biopharmaceuticals, produced intracellularly or in the periplasm, where harvest requires cell lysis, resulting in a complex mixture [1], [2].

For the purification of protein-based pharmaceuticals, packed bed chromatography has been the industry standard due to its high versatility and specificity [3]. Multiple orthogonal methods are often performed in sequence allowing to separate the target from the impurities based on different physicochemical properties. Selection of specific chromatographic methods and operation conditions currently remain to be primarily done by Trial-and-error, expert knowledge or Design of experiments [4], [5]. In recent years, tools like high throughput experimentation and *in silico* modeling have shown great potential to accelerate the design process [6]–[9]. These methods allow to not only consider the elution behavior of target molecules, but the behavior of HCP impurities. This leads to the development of the purification process in a rational and systematic manner.

Alternatively, for prediction of protein behavior at specific chromatographic conditions, quantitative structure-property relationship (QSPR) models aim to use specific features calculated from the protein structures [8], [10]. Over the last 20 years, successful models have been trained for a variety of globular proteins or antibodies [11]–[16]. Recently, Cai et al. trained predictive models using both resin and protein descriptors to predict the adsorption of globular proteins for different mixed mode resins [17]. These prediction methods become even more powerful in combination with mechanistic modeling, allowing full prediction of the elution profile [12], [15]. While these models highlight how structural knowledge of proteins can be used to describe chromatographic behavior, application for HCP removal process development remains challenging. Data available for these models is generally obtained for pure solutions containing only one protein. Therefore these models cannot take the full complexity of a lysate into account, where often countless of protein-protein interactions (PPIs) occur between HCPs [18], [19].

Describing the HCP content of various expression host has been of interest in the last two decades [2], [20], [21]. Mass spectrometry-based proteomics (MS) has gained popularity for analyzing HCPs, enabling the sensitive detection of individual HCPs during process development [1], [2], [22]–[24]. Advances in the field allow identification of specific proteins which are commonly remaining after the downstream processing [25]. Currently, most literature describe HCPs from Chinese hamster ovary (CHO) cells, more specifically the HCP content after the protein A capture step in antibody production [26]–[29]. From these, high-risk HCPs have been identified for CHO, that have potential immunogenic responses or compromise product quality due to degradation [27]. Studies showed that HCP aggregates with mAbs may promote the persistence of HCPs during the protein A capture step [30]–[33]. A recent correlation analysis of HCPs identified co-elution of HCPs in groups that are associated with PPIs [29].

However, less studies targeting *E. coli* HCPs have been conducted. To identify HCP co-elution in immobilized metal affinity chromatography, Bartlow et al., analysed a range of elution buffer concentrations using SDS-PAGE in combination with MALDI-TOF-MS finding 26 proteins co-eluting during a green fluorescent protein purification [34]. More recently, Lingg et al., investigated the effect of metal and chelator type on the HCPs found in the eluate of a similar process [35]. For cation- and anion-exchange chromatography, Swanson et al., studied *E. coli* HCP elution in a 5-step isocratic elution [36], [37]. Using the experimentally determined molecular weight, isoelectric point (pI) and aqueous two-phase partitioning coefficients of the HCPs, random forest regressor models were trained to predict the protein retention. In a more fundamental study, Disela et al., performed MS analysis on *E. coli* BLR(DE3) and HMS174(DE3) HCPs and plotted proteome property maps using the physicochemical properties of around 2000 HCPs to showcase the selection of suitable purification strategies [38].

Despite these efforts, knowledge on chromatographic retention behavior of *E. coli* lysates to aid process design is still lacking. This study aims to guide process development by elucidating the chromatographic behavior of specific HCPs of the *E. coli* BLR(DE3) strain for ion exchange (IEX) and hydrophobic interaction (HIC) chromatography (Figure 3.1). By analyzing fractions collected from linear gradient elution (LGE) experiments using MS, the identity and elution time of different HCPs were determined. For each

HCP the cellular location, function and potential interactions were identified to assess the effect on the elution. For the IEX retention data, predictive QSPR models were trained using protein descriptors calculated from predicted 3D structures. Finally, model accuracies using different HCP subsets were compared.

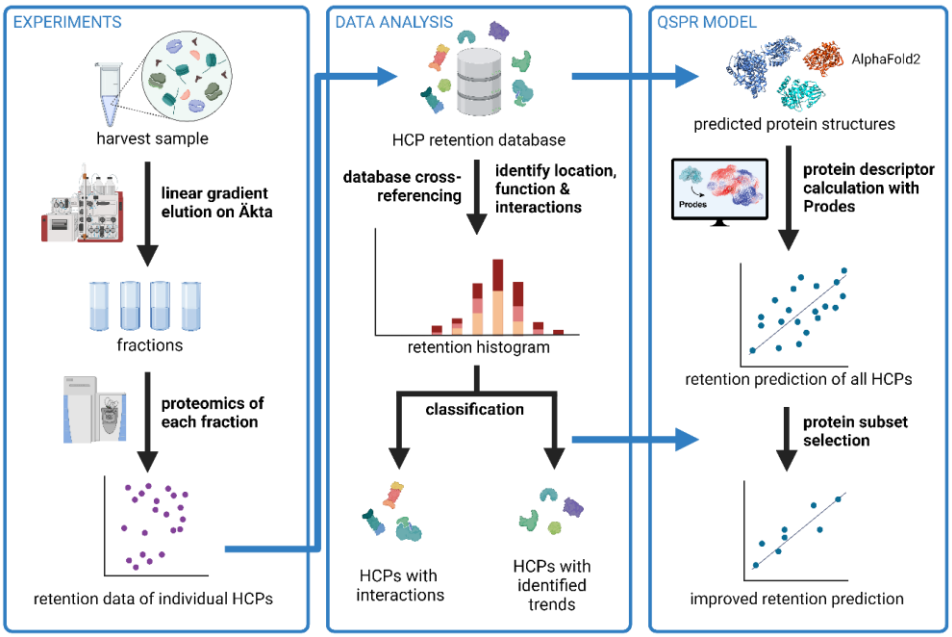


Figure 3.1: Schematic overview of this study. Chromatographic experiments are conducted using the lysate containing a mixture of host cell proteins (HCPs). The protein mixture is injected to the Äkta chromatography system and linear gradient elution experiments on IEX and HIC are conducted. From each of the gradient runs, fractions are taken and their proteome is analyzed via mass spectrometry. The obtained retention data of all HCPs is analyzed regarding elution trends occurring due to cellular location, molecular function and protein-protein interactions. The data is furthermore used to build a QSPR model and investigate several variations using filters based on the deviating retention trends (Illustration created using BioRender.com.).

3.2 Material and Methods

3.2.1 Chromatographic experiments and proteomeic analysis

E. coli harvest sample and equipment

The cells in the harvest sample originating from a null plasmid *E. coli* BLR(DE4) strain, used for the LGE experiments, were disrupted by use of a

French press. Proteins identified in this sample are extensively characterized and described elsewhere [38]. Chromatographic experiments were performed on an Äkta pure with a connected fraction collector F9-C from Cytiva (Uppsala, Sweden). Prepacked HiTrap Q XL (IEX, here: anion exchange chromatography) and Butyl FF (HIC) 5 ml columns from Cytiva (Uppsala, Sweden) were used for chromatographic experiments. The running buffer for the IEX experiment was 0.02 M Tris at pH 7.0 with 0.02 M NaCl added. The elution buffer during the IEX experiment consisted of the same buffer components with 1 M NaCl added. During the HIC experiment, the running buffer was 0.02 M sodium phosphate at pH 7.0 with 3 M NaCl added and as an elution buffer ultrapure water (MilliQ) was employed. Between experimental runs the chromatography columns were cleaned using 1 M NaOH solution. All buffers were filtered with 0.22 µm pore size and sonicated before use.

Linear gradient elution experiments

After injection of 1 ml of the dialyzed clarified harvest sample the column was washed with 5 column volume of running buffer. Then, the gradient elution was started by mixing the running buffer with the elution buffer over a gradient length of 10 column volume (50 ml). During the gradient elution runs conducted with a flow rate of 5 ml/min, fractions were continuously taken and afterwards analyzed using MS. During the IEX experiment, 1 ml fractions were taken and every other fraction was analyzed, as described in more detail in Disela et al., 2024. For the HIC experiment, 2.5 ml fractions were taken and every fraction was analyzed.

Proteomic analysis

Shotgun proteomics to identify individual E. coli proteins in each of the fractions taken during the LGE experiments was performed as described in Disela et al., 2024.

3.2.2 Data processing

The retention profiles (in peak area) of the proteins eluting during the gradient were fitted to a Gaussian function. If the shape could be fitted with a R^2 above 0.7, the maximum of the fitted Gaussian function was used as the retention volume $V_{R,i}$ of each protein i as exemplified in [39]. Since a constant

flow rate was used in the experiments, the dimensionless retention time (DRT) could be calculated as

$$DRT(i) = \frac{V_{R,i} - V_g}{V_G - V_g}, \tag{3.1}$$

where V_g is the volume in the beginning of the salt gradient and V_G in the end of the salt gradient. This measure has been used in literature to describe retention in a dimensionless manner [13].

Abundance measures (for the common scatter plot) and theoretical physicochemical properties were retrieved from a previous study of the harvest sample [38]. The cellular location and functions were retrieved from UniProt [40]. Hereby proteins that were exclusively located in the cytosol or cytoplasm, not in a membrane, were summarized as cytoplasm proteins. Comparable *E. coli* K-12 proteins were retrieved from Arifuzzaman et al., 2006 that show PPIs (Supplemental Table 1 in Arifuzzaman et al., 2006) and proteins without measured interactions (Supplemental Table 3.2 in Arifuzzaman et al., 2006).

3.2.3 QSPR

Protein model generation

Using the amino acid sequence, protein structures were predicted using AlphaFold2 to ensure full sequence coverage in the structure [41]. Of the predicted structures, only the Rank 0 structures were used throughout the study. For each protein, the *E. coli* K12 homolog was used to identify signal peptides which require removal. Protein descriptors were calculated using the open-source software package Prodes (https://github.com/tneijenhuis/prodes) in default settings [42]. Visualization of the protein structures was performed using UCSF Chimera [43].

QSPR model training

Multi Linear Regression (MLR) models were trained for the retention time prediction of the whole dataset and specific subsets of HCPs (SI Table 3.1). The selection of proteins for each subset was based on their presence in the cytoplasm, their multimeric state, described interactions and average per-

residue model confidence score (pLDDR). Initially, the datasets were randomly split into a train (67%) and a test set (33%). To reduce the number of features considered during the feature selection, a series of filter thresholds were screened by applying a range of feature-feature correlation filters (Pearson correlations of 0.8, 0.9, 0.99 and 1). Followed by feature-observation correlations filtering, maintaining a predefined percentage of features (10% to 100% in 10% increments). Features were selected using sequential forward selection for all filter thresholds, resulting in 40 models to be considered. Final models, and optimal filtering thresholds (Supplemental Figure S1), were selected based on the R^2 of a 10-fold cross-validation.

3.3 Results and discussion

3.3.1 Retention behavior of individual host cell proteins

Protein retention map

To identify retention behavior during HIC and IEX chromatography, clarified lysate of *E. coli* was injected, fractions were collected during LGE and subsequently analyzed using MS. For the orthogonal chromatographic methods, data was collected on specific DRT of 908 and 816 HCPs for IEX and HIC, respectively. Undetected HCPs elute either before or after the salt gradient experiments, or are below the detection limit.

Of the determined HCP DRTs, a total of 569 were found for both methods, which allows construction of a 2D retention map (Figure 3.2). As determination of protein abundance remains cumbersome using shotgun proteomics, relative abundance using peak area and the protein abundance index (PAI) were used (Figure 3.2a and Figure 3.2b, respectively). For the different abundance measures, a different order in abundance is caused by the strong dependence on the protein size in the definition of PAI. To estimate absolute protein contents in complex mixtures, the PAI is defined as the number of observed peptides divided by the number of observable peptides per protein [44]. The abundance of the most abundant protein according to the PAI value, ARH99394.1, was plotted over the volume during the IEX and HIC gradient (Figure 3.2c and Figure 3.2d, respectively.

During the IEX LGE, proteins eluted between 0.1 and 0.8 DRT whereas proteins eluted throughout the whole gradient for HIC. If the retention of the new target is known, the experimental HCP retention map can help forming an efficient HCP removal strategy using physicochemical property maps as discussed in Disela et al., 2023. While the physicochemical property maps provide a basis for process development, the experimental retention map provides an improved effective tool. The retention map reflects the actual retention behavior of the HCPs in the lysate including interactions with other proteins limited to the used system, resin and buffer conditions. In contrast to the target retention behavior, this map can be used to form a general approach to remove HCP impurities. This promotes a rational and systematic design of a purification process.

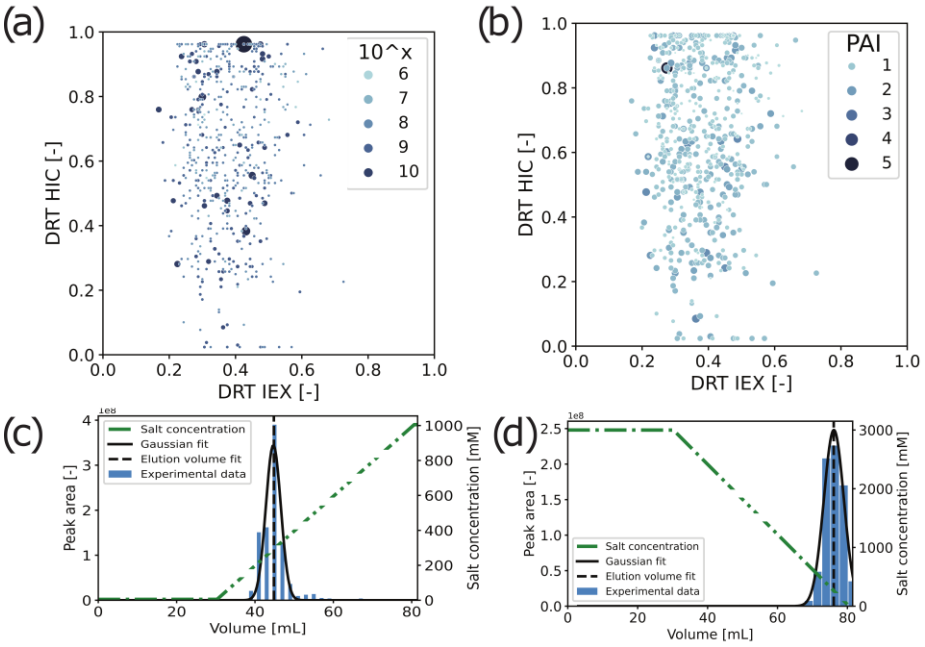


Figure 3.2: Host cell protein (HCP) retention map of individual HCPs in the E. coli lysate. Dimensionless retention times (DRTs) were obtained from MS analysis of fractions obtained from linear gradient experiments on Q Sepharose XL (IEX) and Butyl FF (HIC) HiTrap 5 ml columns at pH 7 using NaCl as salt in both cases. (a) abundance in peak area and (b) abundance as protein abundance index (PAI) obtained from (Disela et al., 2023). (c) elution of protein ARH99394.1 during salt gradient on IEX. (d) elution of protein ARH99394.1 during salt gradient on HIC.

Influence of cellular location

To better understand the behavior of specific HCPs, the extensive proteome dataset was explored regarding a variety of factors which may influence retention. Cellular location was first investigated, where proteins were divided according to their cellular localization (as obtained from UniProt) in the subgroups cytoplasm, plasma membrane, and outer membrane (Figure 3.3a&b).

For IEX, the histogram with all proteins shows the highest number of proteins in the fraction at 0.30 DRT (166 out of 908) and second highest number at 0.46 DRT (123 out of 908). The histogram of all proteins eluting on HIC shows an increase with increasing DRT over the whole gradient. This spread over the gradient leads to less protein per fraction in the HIC histograms compared to the IEX histograms.

During the IEX, the majority of the HCPs are cytoplasm proteins (total 572) and the elution follows the general trend of all proteins during IEX, with the exception of a lower number of proteins eluting at DRT 0.46. At this DRT, the histogram of plasma membrane proteins (total 79) shows the highest abundance (41 out of 79). The histogram of outer membrane proteins (total 27) shows a low general abundance throughout the gradient with a slightly higher abundance at 0.26 and 0.46 DRT. In IEX, retention is based on charge, meaning that a protein with a lower pI elutes later during the LGE. This trend holds true for the overall dataset, except for the plasma membrane HCPs (Supplemental Figure 3.S2a), suggesting interactions of these proteins leads to concurrent elution. This indicates that forces causing these interactions are stronger compared to electrostatic forces that are the main interaction as shown by the IEX trendline of the majority of the proteins. Plasma membrane proteins might interact with each other directly forming parts of known (sdhB, secY) or unknown complexes (hflC, arnC) [45]. We even observe the co-elution of yidC and secY, that are known to form a multi-protein complex for Sec-dependent membrane protein integration [46]. However, the joint elution of several plasma membrane proteins might indicate that they form liposomes or are parts of membrane vesicles [47]. Considering that HCPs are impurities, a concurrent elution could simplify the development of the chromatography step. However, for a retention prediction model, joint elution hampers the prediction for these proteins, when using calculated protein features.

During the HIC gradient, the histogram of cytoplasm proteins (total 532) shows a similar shape to the histogram of all proteins with a slightly lower number of proteins eluting toward the end of the gradient (Figure 3.3b). At the end of the HIC gradient, the plasma membrane proteins (total 66) show an increased occurrence. Outer membrane proteins (total 48) elute continuously throughout the gradient. In HIC, a correlation to hydrophobicity, such as the GRAVY value (grand average of hydropathy) is expected. However, none of the hydrophobicity measures, calculated from the predicted protein structure, showed a high correlation and hence it was not possible to identify protein groups that show deviating retention behavior (data not shown). This is thought due to the highly dynamic behavior of the proteins in the high salt conditions. Often complex phenomena such as nonspecific PPIs or partial unfolding upon binding occur, making the single, static, protein chain representation invalid. Additionally, preferred binding orientations might play an important role due to the short range interactions governing adsorption [13]. This complicates the retention prediction substantially, leaving room for future studies to develop new features to describe flexibility and local aggregation propensities, influencing protein elution in HIC.

Influence of molecular function

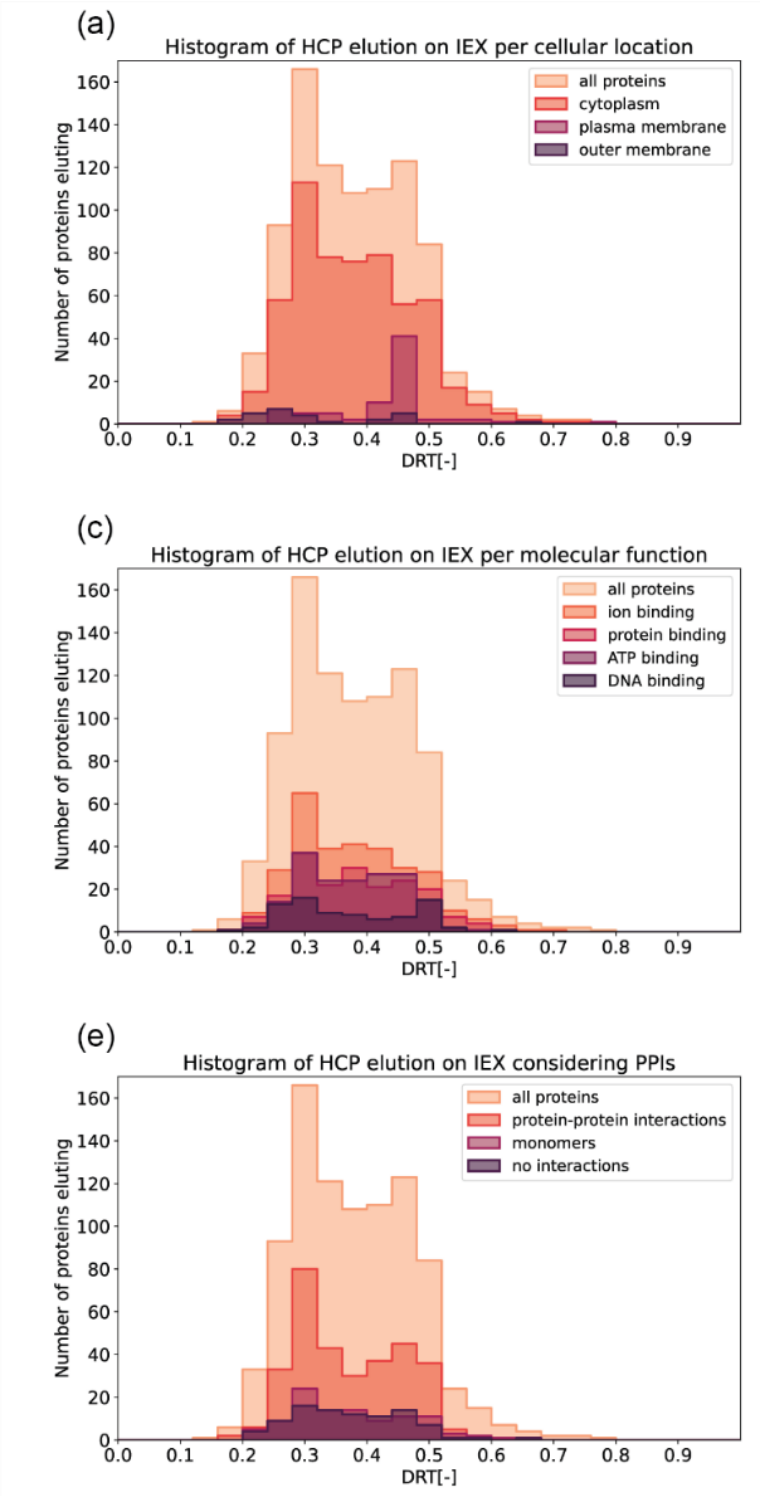
Molecular function as a discriminator for retention behavior was investigated and the results are shown in Figure 3c&d. Proteins that bind ions, other proteins, ATP, or DNA were identified using the UniProt entry. During the IEX gradient, the ion (302), protein (190) and ATP binding proteins (177) follow the trend seen for all proteins. Hence, the binding sites of ions, other proteins, and ATP seem to have little effect on retention behavior. In contrast, DNA binding proteins (80) show a second local maximum at 0.50 DRT. This second maximum is caused by polymerases and ribonucleases, while the first peak is caused by other translation proteins. In contrast to the plasma membrane proteins, the DNA binding proteins follow the trend given by the correlation to the pI (Supplemental Figure 3.S2b).

During the HIC gradient, the ion (272), protein (165), ATP (133), and DNA binding proteins (71) are distributed across all elution times with no clear elution points (Figure 3.3d).

Influence of protein-protein interactions

In the complex mixture of a host cell lysate proteins can interact, forming functional or non-functional complexes. The different PPIs at physiological conditions between *E. coli* proteins were identified by Arifuzzaman et al. [19]. Out of the interactions identified by Arifuzzaman et al., 1270 were found in the IEX dataset and 1225 in the HIC dataset. From these interactions, 349 protein pairs (27 %) in IEX and 178 protein pairs (14 %) in HIC showed close retention proximity (IEX $<$ 0.04 DRT; HIC $<$ 0.05 DRT). It is worth noting that close retention proximity depends on the chosen threshold, which was the fraction size. While conditions in the running buffer of IEX come close to the physiological conditions used in the study from Arifuzzaman et al., the HIC running buffer has a significant higher salt concentration that might dissociate complexes or induce additional PPIs [48]. Nevertheless, these interactions pose an interesting effect on the DRTs of involved HCPs as indicated in a recent study for CHO cells [29].

To identify the effect of PPIs, proteins described to interact from protein pairs in close proximity were selected (Figure 3.3e&f). Proteins described to have no interactions in Arifuzzaman et al. were also plotted as one group. Additionally, proteins known to be present as monomers were grouped. During the IEX gradient, the proteins with PPIs (319) show a high abundance at 0.30 and 0.46 DRT and the surrounding fractions. This shape impacts the histogram with all proteins significantly. Monomers (104) and non-interacting proteins (89), on the other hand, are eluting throughout the IEX gradient with a near Gaussian distribution. During the HIC gradient, less proteins with PPIs were detected (170). These proteins show an increased abundance at higher DRT, which might be related to the large size of the complexes which is reported to effect retention in HIC [49]. For the monomers (98) and non-interacting proteins (80) no such trend was observed as these elute throughout the gradient.


(a)
Histogram of HCP elution on IEX per cellular location
all proteins
cytoplasm
plasma membrane
outer membrane
Number of proteins eluting
DRT[-]
(c)
Histogram of HCP elution on IEX per molecular function
all proteins
ion binding
protein binding
ATP binding
DNA binding
Number of proteins eluting
DRT[-]
(e)
Histogram of HCP elution on IEX considering PPIs
all proteins
protein-protein interactions
monomers
no interactions
Number of proteins eluting
DRT[-]

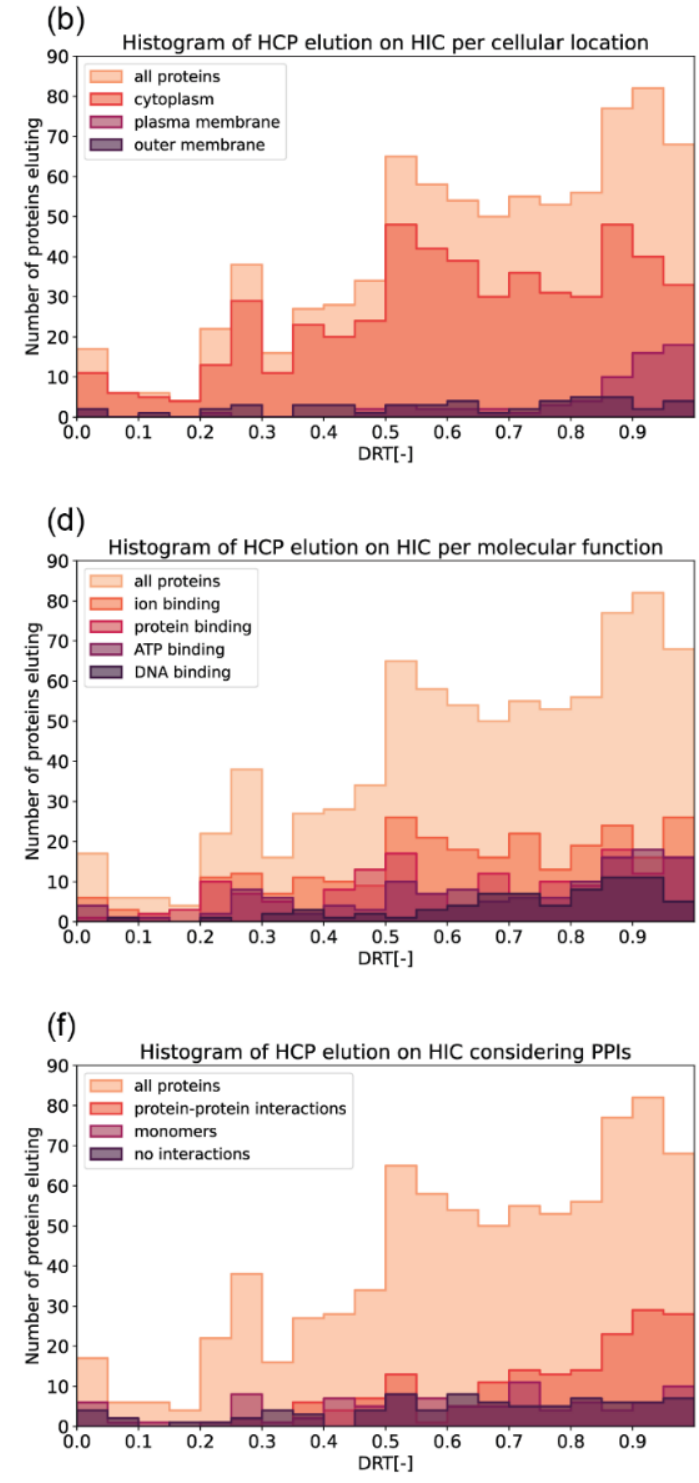


Figure 3.3: Histograms representing the elution of groups of host cell proteins (HCPs). The number of proteins with an elution maximum during a specific dimensionless retention time (DRT) is listed for ion exchange (IEX) and hydrophobic interaction chromatography (HIC). (a) histogram of cellular location groups during IEX. (b) histogram of cellular location groups during HIC. (c) histogram of molecular function groups during IEX. (d) histogram of molecular function groups during HIC. (e) histogram of protein-interaction groups during IEX. (f) histogram of protein-interaction groups during HIC.

In conclusion, the plasma membrane proteins, DNA binding, and proteins with PPIs were identified as protein groups that show a deviant elution behavior due to their location in the cell, molecular functions or PPIs. Not considering these characteristics during feature calculation might hinder accurate retention predictions. The proteins in the cytoplasm, without known interactions, and monomers seem to be more suited to build an improved model.

3.3.2 Prediction of retention time of individual HCPs in IEX

Descriptive QSPR model using the complete dataset

Using the DRTs obtained from IEX LGE of all single peak proteins, a predictive QSPR model was trained, correlating specific physicochemical features to protein retention. A final MLR model composed of 27 features was build achieving a 10-fold cross validated R^2 of 0.55 and a mean absolute error (MAE) of 0.049 (Figure3. 4 and Table 3.1 [ALL]). For the test set, data not involved during feature selection, a MAE of 0.048 was achieved. Due to the fractionation approach, the resolution of 25 fractions introduces an experimental error of 0.04 DRT, which requires consideration while assessing the final QSPR model. Therefore, the prediction can be considered successful, given the data resolution. As observed in the IEX histograms, a significant part of the proteins have a DRT around 0.3. For the QSPR model, this resulted in a general overprediction for proteins with a DRT < 0.3 and underprediction for protein with DRT > 0.3 (Figure 3.4). Despite this bias, the trend of the HCP elution behavior was still captured by the model.

The model captures the importance of charge in IEX since the majority of the selected features, 15 of the 27, directly describe the charge of the protein (Supplemental Figure S3). Additionally, the surface content of the four charged amino acids was found to be important. Due to the number of features and the inherent collinearity of the charge related features, specific feature importance cannot be identified. The remaining eight features describe the surface, hydrophobicity and the surface content of specific noncharged amino acids. Y-scrambling was performed before training as final validation (Supplemental Figure 3.S4). The resulting model was not able to predict scrambled protein retention (R^2 of -0.065) proving physical validity.

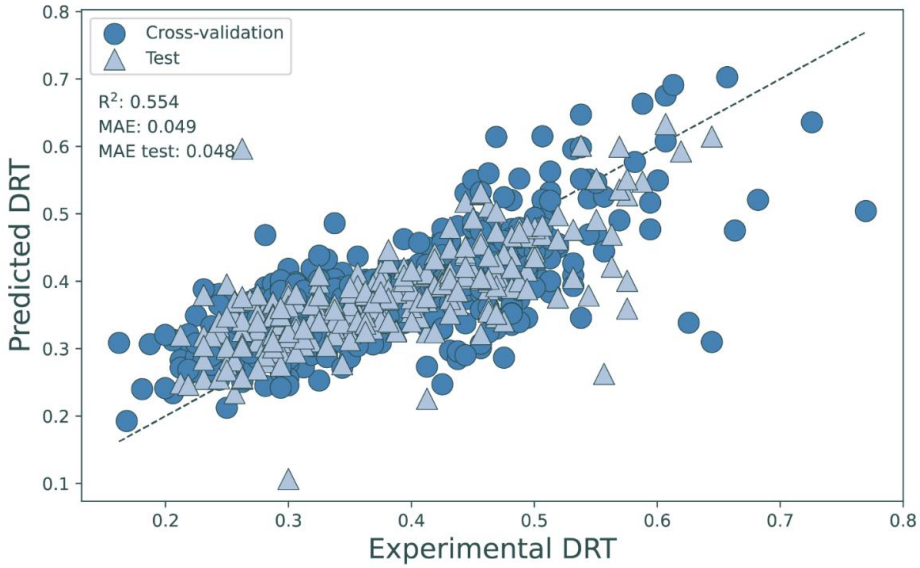


Figure 3.4: QSPR validation of the regression model trained to predict DRT, where the circles represent the 10-fold cross-validation and the triangles the test set.

A similar approach was performed to train elution prediction model for HIC albeit being less successful. No combination of features was found resulting in a model with a cross validated R^2 >0.2. It is thought this is due to the nonspecific protein interactions at high salt conditions and partial unfolding upon binding which often occur [48]. As was mentioned in 3.1.2, no correlation was found with HIC elution and any of the hydrophobicity features for the full dataset nor any subsets.

Influence of HCP subsets on model accuracy

One of the major challenges in accurately describing the HCPs is the countless interactions that can occur between proteins and other host cell components. As these interactions have not been taken into account for the first elution prediction model, the cross validated R^2 of 0.55 is thought to be a success. Nevertheless, the elution model would not be suitable for decision making as the residuals are not spread evenly. To increase the prediction accuracy, the dataset was simplified by selecting proteins which do not bind the cell membrane (cytoplasm proteins), or interact to form complexes (monomers, proteins without measured interactions) and combinations thereof (Table 3.1, Figure 3.5). All models resulting from the different

subsets provided a greater accuracy for the cross validated training set (MAE from 0.045 to 0.039). In contrast to the cross-validation, the accuracy of the test was not improved for most models (MAE of 0.058 to 0.043).

For the proteins in the cytoplasm, the overall trend in the model (Table 3.1 and Figure 3.5a) is similar to the trends observed in the model with all proteins. It was expected that removal of the membrane proteins would result in a better prediction as these proteins did not adhere to the correlation between pI and DRT (Supplemental figure 3.5.1a). In the contrary, the test set was predicted less accurately (MAE of 0.055) compared to the all HCP dataset (MAE of 0.048). This decrease in accuracy can be attributed to an increased bias towards a DRT close to 0.3 (Figure 3.2a).

The subset containing the proteins without PPIs were found to elute according to a normal distribution (Figure 3.3e), therefore the bias at 0.3 DRT observed for the other datasets should not pose a problem. However, the test set accuracy (MAE of 0.058) was found to be lower than the all HCP dataset (MAE 0.048) (Figure 3.5b, Table 3.1). Unlike the all HCP or cytoplasm datasets, no bias is observed for the prediction. While these proteins were described as noninteracting, they can still be prone to multimerization. Only nine proteins showed overlap between the noninteracting and monomer dataset (data not shown). The loss of accuracy is also thought to be due to the smaller training dataset, resulting in less general QSPR models. Therefore, complex behavior, such as oligomerization or complex formation, cannot be captured implicitly.

For the monomer subset a cross validated R^2 of 0.697 was achieved and the accuracy of the test set was improved to a MAE of 0.043, 7.5% off the experimental error (Table 3.1, Figure 3.5c). Additionally, the residuals of the model are spread more evenly compared to the initial elution model allowing prediction of parts of the dataset. The main reason for the improved accuracy is thought to be the structural representation used for the feature calculation, as the structures were predicted in a monomeric state. While PPIs were not filtered out, these do not seem to have a major influence. For this particular model, the average and sum of the negative electrostatic potential were found to be most important, as removing either features from the model results in a cross validated R^2 of 0.47 (Supplemental Figure 3.S7).

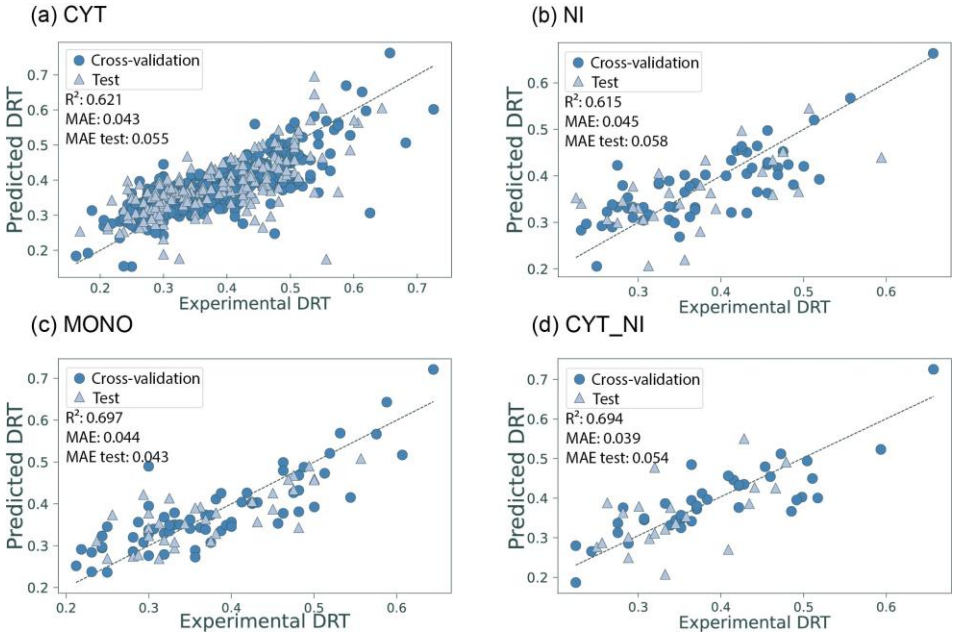


Figure 3.5: QSPR validation of the regression model trained to predict DRT of protein subsets, where the circles represent the 10-fold cross-validation and the triangles the test set. the presented subsets are the cytosolic proteins (a), the proteins without interactions (b), proteins reported to be present as monomers.

Table 3.1: Comparison of model performance for the different protein subsets. Protein subsets were generated based on all proteins (ALL), proteins present in the cytoplasm (CYT), proteins without PPIs (NI), proteins annotated as monomers (MONO) and proteins with an average pLDDR > 0.95 (HC) or combinations thereof.

	#Proteins for training	#Features selected	Cross-validation R^2	Cross-validation MAE	Test MAE	Difference Test MAE to experimental error (%)
ALL	560	27	0.554	0.049	0.048	20
CYT	373	10	0.621	0.043	0.055	37.5
NI	59	10	0.615	0.045	0.058	45
MONO	67	10	0.697	0.044	0.043	7.5
CYT_NI	40	8	0.694	0.039	0.054	35
HC	299	23	0.614	0.045	0.051	27.5
CYT_HC	189	10	0.587	0.048	0.049	22.5
NI_HC	31	6	0.829	0.029	0.069	72.5
CYT_NI_HC	24	4	0.852	0.029	0.080	100
MONO_HC	38	7	0.750	0.035	0.047	17.5

In contrast to recent literature, the retention data used in this work is obtained from a clarified lysate. The lower R^2 relative to those described in elsewhere cannot be compared to elution prediction of antibodies, due to an increase in sample heterogeneity, or other model proteins which are better understood at a fundamental level [12], [17], [42]. Nevertheless, This work is an initial step in better understanding elution behavior of HCPs which would ultimately be predicted with similar accuracies as antibodies or model proteins.

The increased accuracy of the monomer subset highlights the importance of accurate protein structure representation. Therefore, improvements in the model can be made by modeling the multimeric state of each protein for which it is known. As this information is not available for every protein, improving accurate PPI prediction is essential [50]. This would allow QSPR application to predict the behavior a full lysate rather than only protein subsets. Additionally, the structures obtained by AlphaFold are predicted and should therefore be used with caution. The per residue confidence score and the predicted aligned error provided by AlphaFold has the potential for template selection to increase model accuracy. However, setting confident

score thresholds for the predicted structures did not yield more accurate elution prediction models (Supplemental Figure 3.S9).

3.4 Conclusions and outlook

The observed host cell proteome after lysis of the *E. coli* BLR(DE3) host covers the retention times of around 900 unique proteins on IEX and HIC. By selecting protein subsets based on location, function, and interactions, trends in retention behavior were examined. For IEX, it was observed that proteins present in the plasma membrane would primarily co-elute, disregarding the general trend of the lower pI resulting in later retention. For HIC, an almost linear trend was observed for the number of proteins throughout the gradient. Only proteins located in the plasma membrane or that are known to engage in PPIs were found to deviate from this trend, primarily eluting at the end of the HIC gradient. Despite the complexity of the mixture, structure models predicted by AlphaFold2 were used to train a descriptive QSPR model (R^2 of 0.55) for IEX retention, approaching the experimental error. By selecting proteins annotated as monomer in UniProt, the accuracy of the QSPR model improved significantly (R^2 of 0.70). This work is the initial step towards understanding the HCP elution of the *E. coli* BLR(DE3) host cell proteome.

To further improve the understanding and implementation of QSPR in process development, future research should focus on the in-depth characterization of lysate compositions. Currently, lots of knowledge is available via databases such as UniProt, however many proteins remain underdetermined especially regarding PPIs. More experiments are needed to identify complex formation of proteins under different buffer conditions. Additionally, despite the improvements in structure prediction, automated protocols for assessing the plausibility of a structure to allow processing of large datasets. Ultimately, this research represents a significant step towards *in-silico* driven process development, increasing process understanding and reducing development times.

3.5 References

[1] D. G. Bracewell, R. Francis, and C. M. Smales, “The future of host cell protein (HCP) identification during process development and

manufacturing linked to a risk-based management for their control," *Biotechnol. Bioeng.*, vol. 112, no. 9, pp. 1727–1737, 2015.

[2] A. L. Tscheliessnig, J. Konrath, R. Bates, and A. Jungbauer, "Host cell protein analysis in therapeutic protein bioprocessing - methods and applications," *Biotechnol. J.*, vol. 8, no. 6, pp. 655–670, 2013.

[3] U. Gottschalk, K. Brorson, and A. A. Shukla, "The need for innovation in biomanufacturing," *Nat. Biotechnol.*, vol. 30, no. 6, pp. 489–492, 2012.

[4] D. Keulen, G. Geldhof, O. Le Bussy, M. Pabst, and M. Ottens, "Recent advances to accelerate purification process development: A review with a focus on vaccines," *J. Chromatogr. A*, vol. 1676, p. 463195, Aug. 2022.

[5] A. T. Hanke and M. Ottens, "Purifying biopharmaceuticals: Knowledge-based chromatographic process development," *Trends Biotechnol.*, vol. 32, no. 4, pp. 210–220, 2014.

[6] B. K. Nfor *et al.*, "Multi-dimensional fractionation and characterization of crude protein mixtures: Toward establishment of a database of protein purification process development parameters," *Biotechnol. Bioeng.*, vol. 109, no. 12, pp. 3070–3083, Dec. 2012.

[7] S. M. Pirrung *et al.*, "Chromatographic parameter determination for complex biological feedstocks," *Biotechnol. Prog.*, vol. 34, no. 4, pp. 1006–1018, 2018.

[8] C. R. Bernau, M. Knödler, J. Emonts, R. C. Jäpel, and J. F. Buyel, "The use of predictive models to develop chromatography-based purification processes," *Front. Bioeng. Biotechnol.*, vol. 10, no. October, pp. 1–24, 2022.

[9] D. Keulen, E. van der Hagen, G. Geldhof, O. Le Bussy, M. Pabst, and M. Ottens, "Using artificial neural networks to accelerate flowsheet optimization for downstream process development," *Biotechnol. Bioeng.*, no. February, pp. 1–14, May 2023.

[10] J. Emonts and J. F. Buyel, "An overview of descriptors to capture protein properties – Tools and perspectives in the context of QSAR modeling," *Comput. Struct. Biotechnol. J.*, vol. 21, pp. 3234–3247, 2023.

[11] T. Yang, M. C. Sundling, A. S. Freed, C. M. Breneman, and S. M. Cramer, "Prediction of pH-dependent chromatographic behavior in ion-exchange systems," *Anal. Chem.*, vol. 79, no. 23, pp. 8927–8939, 2007.

[12] R. Hess *et al.*, "Predicting multimodal chromatography of therapeutic antibodies using multiscale modeling," *J. Chromatogr. A*, vol. 1718, no. February, p. 464706, 2024.

[13] A. T. Hanke *et al.*, "Prediction of protein retention times in hydrophobic interaction chromatography by robust statistical characterization of their atomic-level surface properties," *Biotechnol. Prog.*, vol. 32, no. 2, pp. 372–381, 2016.

[14] C. B. Mazza, N. Sukumar, C. M. Breneman, and S. M. Cramer, "Prediction of protein retention in ion-exchange systems using molecular descriptors obtained from crystal structure," *Anal. Chem.*, vol. 73, no. 22, pp. 5457–5461, 2001.

[15] D. Saleh *et al.*, "A multiscale modeling method for therapeutic antibodies in ion exchange chromatography," *Biotechnol. Bioeng.*, vol. 120, no. 1, pp. 125–138, 2023.

[16] J. Kittelmann, K. M. H. Lang, M. Ottens, and J. Hubbuch, "Orientation of monoclonal antibodies in ion-exchange chromatography: A predictive quantitative structure–activity relationship modeling approach," *J. Chromatogr. A*, vol. 1510, pp. 33–39, 2017.

[17] Q. Y. Cai, L. Z. Qiao, S. J. Yao, and D. Q. Lin, "Machine learning assisted QSAR analysis to predict protein adsorption capacities on mixed-mode resins," *Sep. Purif. Technol.*, vol. 340, no. December 2023, p. 126762, 2024.

[18] S. V. Rajagopala *et al.*, "The binary protein-protein interaction landscape of escherichia coli," *Nat. Biotechnol.*, vol. 32, no. 3, pp. 285–290, 2014.

[19] M. Arifuzzaman *et al.*, "Large-scale identification of protein-protein interaction of Escherichia coli K-12," *Genome Res.*, vol. 16, no. 5, pp. 686–691, 2006.

[20] X. Wang, A. K. Hunter, and N. M. Mozier, "Host cell proteins in biologics development: Identification, quantitation and risk assessment," *Biotechnol. Bioeng.*, vol. 103, no. 3, pp. 446–458, 2009.

[21] S. M. Timmick *et al.*, "An impurity characterization based approach for the rapid development of integrated downstream purification processes," *Biotechnol. Bioeng.*, vol. 115, no. 8, pp. 2048–2060, 2018.

[22] M. R. Schenauer, G. C. Flynn, and A. M. Goetze, "Identification and quantification of host cell protein impurities in biotherapeutics using mass spectrometry," *Anal. Biochem.*, vol. 428, no. 2, pp. 150–157, 2012.

[23] D. Rathore, A. Faustino, J. Schiel, E. Pang, M. Boyne, and S. Rogstad, "The role of mass spectrometry in the characterization of biologic protein products," *Expert Rev. Proteomics*, vol. 15, no. 5, pp. 431–449, 2018.

[24] M. J. Traylor, P. Bernhardt, B. S. Tangarone, and J. Varghese, "Analytical Methods," in *Biopharmaceutical Processing*, G. Jagschies, E. Lindskog, K. Lacki, and P. Galliher, Eds. Elsevier, 2018, pp. 1001–1049.

[25] R. Molden *et al.*, "Host cell protein profiling of commercial therapeutic protein drugs as a benchmark for monoclonal antibody-based therapeutic protein development," *MAbs*, vol. 13, no. 1, 2021.

[26] D. Migani, C. M. Smales, and D. G. Bracewell, "Effects of lysosomal biotherapeutic recombinant protein expression on cell stress and protease and general host cell protein release in Chinese hamster ovary cells," *Biotechnol. Prog.*, vol. 33, no. 3, pp. 666–676, 2017.

[27] M. Jones *et al.*, "'High-risk' host cell proteins (HCPs): A multi-company collaborative view," *Biotechnol. Bioeng.*, vol. 118, no. 8, pp. 2870–2885, Aug. 2021.

[28] M. Vanderlaan, J. Zhu-Shimoni, S. Lin, F. Gunawan, T. Waerner, and K. E. Van Cott, "Experience with host cell protein impurities in biopharmaceuticals," *Biotechnol. Prog.*, vol. 34, no. 4, pp. 828–837, Jul. 2018.

[29] S. Panikulam *et al.*, "Host cell protein networks as a novel co-elution mechanism during protein A chromatography," *Biotechnol. Bioeng.*, Mar. 2024.

[30] Y. H. Oh *et al.*, "Characterization and implications of host-cell protein aggregates in biopharmaceutical processing," *Biotechnol. Bioeng.*, vol. 120, no. 4, pp. 1068–1080, Apr. 2023.

[31] C. E. Herman *et al.*, "Behavior of host-cell-protein-rich aggregates in antibody capture and polishing chromatography," *J. Chromatogr. A*, vol. 1702, p. 464081, 2023.

[32] C. E. Herman *et al.*, "Analytical characterization of host-cell-protein-rich aggregates in monoclonal antibody solutions," *Biotechnol. Prog.*, vol. 39, no. 4, pp. 1–16, 2023.

[33] P. Gagnon *et al.*, "Nonspecific interactions of chromatin with immunoglobulin G and protein A, and their impact on purification performance," *J. Chromatogr. A*, vol. 1340, pp. 68–78, May 2014.

[34] P. Bartlow *et al.*, "Identification of native Escherichia coli BL21 (DE3) proteins that bind to immobilized metal affinity chromatography under high imidazole conditions and use of 2D-DIGE to evaluate contamination pools with respect to recombinant protein expression level," *Protein Expr. Purif.*, vol. 78, no. 2, pp. 216–224, 2011.

[35] N. Lingg *et al.*, "Proteomics analysis of host cell proteins after immobilized metal affinity chromatography: Influence of ligand and metal ions," *J. Chromatogr. A*, vol. 1633, p. 461649, Dec. 2020.

[36] R. K. Swanson, R. Xu, D. S. Nettleton, and C. E. Glatz, "Accounting for host cell protein behavior in anion-exchange chromatography," *Biotechnol. Prog.*, vol. 32, no. 6, pp. 1453–1463, Nov. 2016.

[37] R. K. Swanson, R. Xu, D. Nettleton, and C. E. Glatz, "Proteomics-based, multivariate random forest method for prediction of protein separation behavior during cation-exchange chromatography," *J. Chromatogr. A*, vol. 1249, pp. 103–114, 2012.

[38] R. Disela, O. Le Bussy, G. Geldhof, M. Pabst, and M. Ottens, "Characterisation of the E. coli HMS174 and BLR host cell proteome to guide purification process development," *Biotechnol. J.*, vol. 18, no. 9, p. 2300068, Sep. 2023.

[39] R. Disela *et al.*, "Proteomics-based method to comprehensively model the removal of host cell protein impurities," *Biotechnol. Prog.*, 2024.

[40] A. Bateman *et al.*, "UniProt: the universal protein knowledgebase in 2021," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D480–D489, 2021.

[41] J. Jumper *et al.*, "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021.

[42] T. Neijenhuis, O. Le Bussy, G. Geldhof, M. E. Klijn, and M. Ottens, "Predicting protein retention in ion-exchange chromatography using an open source QSPR workflow," *Biotechnol. J.*, vol. 19, no. 3, p. e2300708, 2024.

[43] E. F. Pettersen *et al.*, "UCSF Chimera - A visualization system for exploratory research and analysis," *J. Comput. Chem.*, vol. 25, no. 13, pp. 1605–1612, 2004.

[44] J. Rappsilber, U. Ryder, A. I. Lamond, and M. Mann, "Large-scale proteomic analysis of the human spliceosome," *Genome Res.*, vol. 12, no. 8, pp. 1231–1245, 2002.

[45] G. Maddalo *et al.*, "Systematic analysis of native membrane protein complexes in Escherichia coli," *J. Proteome Res.*, vol. 10, no. 4, pp. 1848–1859, 2011.

[46] K. Kumazaki *et al.*, "Crystal structure of Escherichia coli YidC, a membrane protein chaperone and insertase," *Sci. Rep.*, vol. 4, pp. 1–6, 2014.

[47] T. Nagakubo, N. Nomura, and M. Toyofuku, "Cracking Open Bacterial Membrane Vesicles," *Front. Microbiol.*, vol. 10, no. January, 2020.

[48] L. A. Jakob, B. Beyer, C. Janeiro Ferreira, N. Lingg, A. Jungbauer, and R. Tscheließnig, "Protein-protein interactions and reduced excluded volume increase dynamic binding capacity of dual salt systems in hydrophobic interaction chromatography," *J. Chromatogr. A*, vol. 1649, p. 462231, 2021.

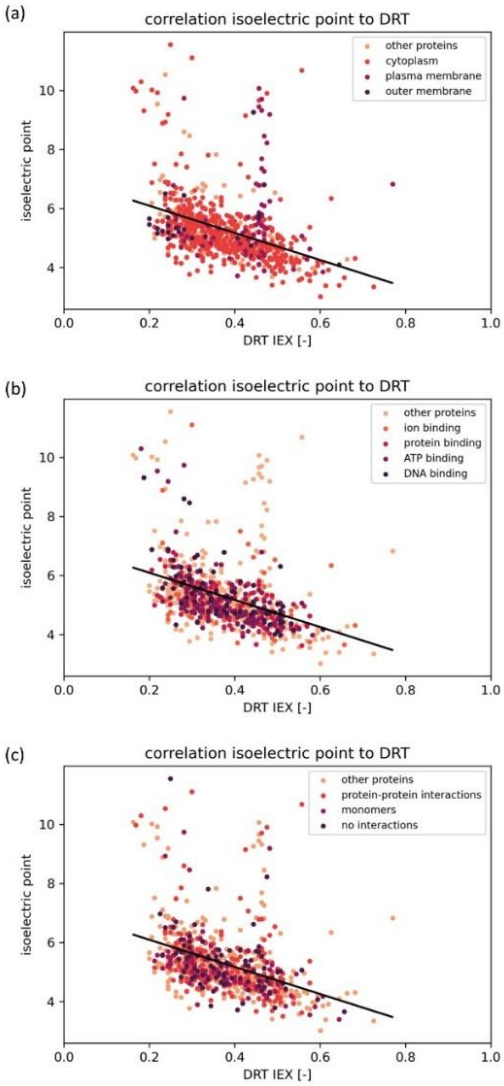
[49] P. A. O'Farrell, *Molecular Biomethods Handbook*. Totowa, NJ: Humana Press, 2008.

[50] F. Soleymani, E. Paquet, H. Viktor, W. Michalowski, and D. Spinello, "Protein–protein interaction prediction with deep learning: A comprehensive review," *Comput. Struct. Biotechnol. J.*, vol. 20, pp. 5316–5341, 2022.

Supplementary material

Model	Feature – feature filter	Feature – observation filter (%)
ALL	0.99	100
CYT	1	100
NI	1	100
MONO	1	50
CYT_NI	0.9	100

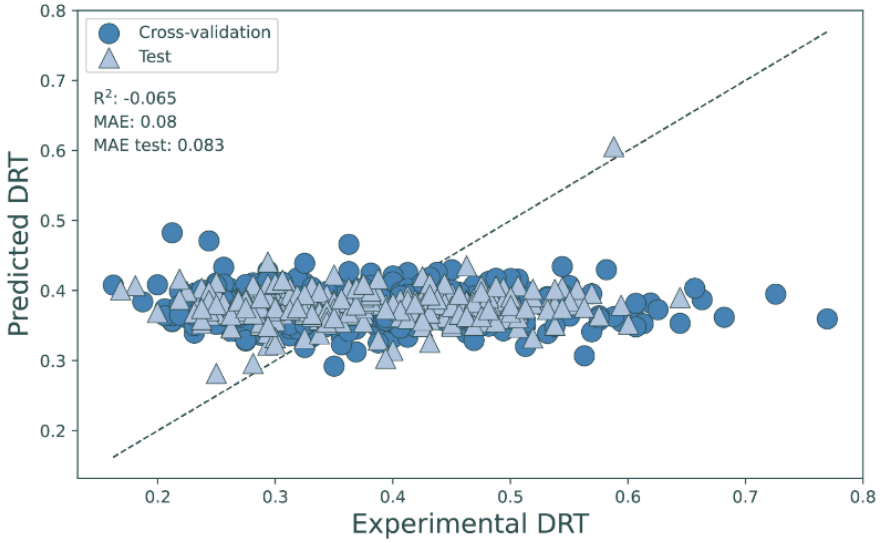
Supplemental Figure S1: Selected filtering thresholds selected for the different protein subsets. Protein subsets were generated based on all proteins (ALL), proteins present in the cytoplasm (CYT), proteins without PPIs (NI), proteins annotated as monomers (MONO) or combinations thereof. The feature – feature filter removes features with a Pearson correlation above the given threshold to other features. The feature – observation filter maintains a percentage of features with the highest Pearson correlation to the elution time.



Supplemental Figure S2: Correlation of the IEX DRT and the estimated isoelectric point. All plots contain all proteins identified for the IEX colored according to subsets based on the cellular location, function and interactions, for a, b and c respectively. The observed R^2: 0.1554, Pearson Correlation: -0.3942

Descriptor	Coefficient	Permutation R^2
SurfNegEpMeanAverage	0.0190	0.5602
SurfMhpMean	0.5221	0.5383
SurfNegEpStdFormal	0.1961	0.5270
SurfNegEpSumFormal	0.4755	0.5430
NSurfPosEpAverage	-0.2208	0.5610
Charge	-0.0277	0.5632
TYR surface fraction	0.0959	0.5344
SurfPosMhpTrimean	0.0917	0.5508
GLU surface fraction	0.1897	0.5299
LYS surface fraction	-0.1200	0.5488
SurfNegMhpMean	-0.2287	0.5450
GLY surface fraction	0.0707	0.5412
ShellEpPosSumFormal	0.5769	0.5135
ASP surface fraction	0.0852	0.5455
ARG surface fraction	-0.0561	0.5489
ShellEpPosMedianFormal	-0.1565	0.5519
ShellEpMaxFormal	0.2441	0.5524
ShellEpMedianFormal	-0.5183	0.5429
ShellEpNegMedianFormal	0.2582	0.5499
SurfPosEpSumFormal	-0.3920	0.5448
SurfMhpStd	-0.2029	0.5523
Isoelectric point	-0.1506	0.5536
NSurfPosEpAverage	0.6613	0.5554
Formal_Charge	-0.5314	0.5571
ShellEpPosStdFormal	-0.0680	0.5561
GLN surface fraction	0.0444	0.5525
Surface shape min	0.0233	0.5556
intercept	0.0552	

Supplemental Table S3: Regression coefficient and permutation performances for the linear regression model predicting DRT for all HCPs.



Supplemental Figure S4: Y-scrambled cross-validation and test of the QSPR model containing all protein retention times. The circles represent the 10-fold cross-validation and the triangles the test set.

Descriptor	Coefficient	Permutation R^2
SurfNegEpMeanAverage	-0.4008	0.4959
SurfMhpMean	0.1467	0.5926
SurfNegEpStdAverage	0.7528	0.6084
Avg. Mass	-0.3032	0.5969
LYS surface fraction	-0.1179	0.5838
SurfNegMhpMedian	-0.1301	0.6021
TYR surface fraction	0.0853	0.6049
NSurfNegMhp	0.1932	0.6122
SurfNegEpStdFormal	-0.5879	0.6134
GLY surface fraction	0.0494	0.6157
intercept	0.6464	

Supplemental Table S5: Regression coefficient and permutation performances for the linear regression model predicting DRT for the CYS subset.

Descriptor	Coefficient	Permutation R^2
SurfEpMinAverage	-0.2562	0.5351
SurfPosMhpsum	0.0747	0.6152
PRO surface fraction	-0.1570	0.4696
SurfMhpMax	0.0904	0.5024
SurfPosEpStdFormal	-0.1244	0.5940
TYR surface fraction	0.0969	0.5765
CYS surface fraction	0.0793	0.5528
Surface shape max	-0.0670	0.5700
LYS surface fraction	-0.0885	0.5658
SurfEpStdAverage	0.0730	0.5984
intercept	0.5735	

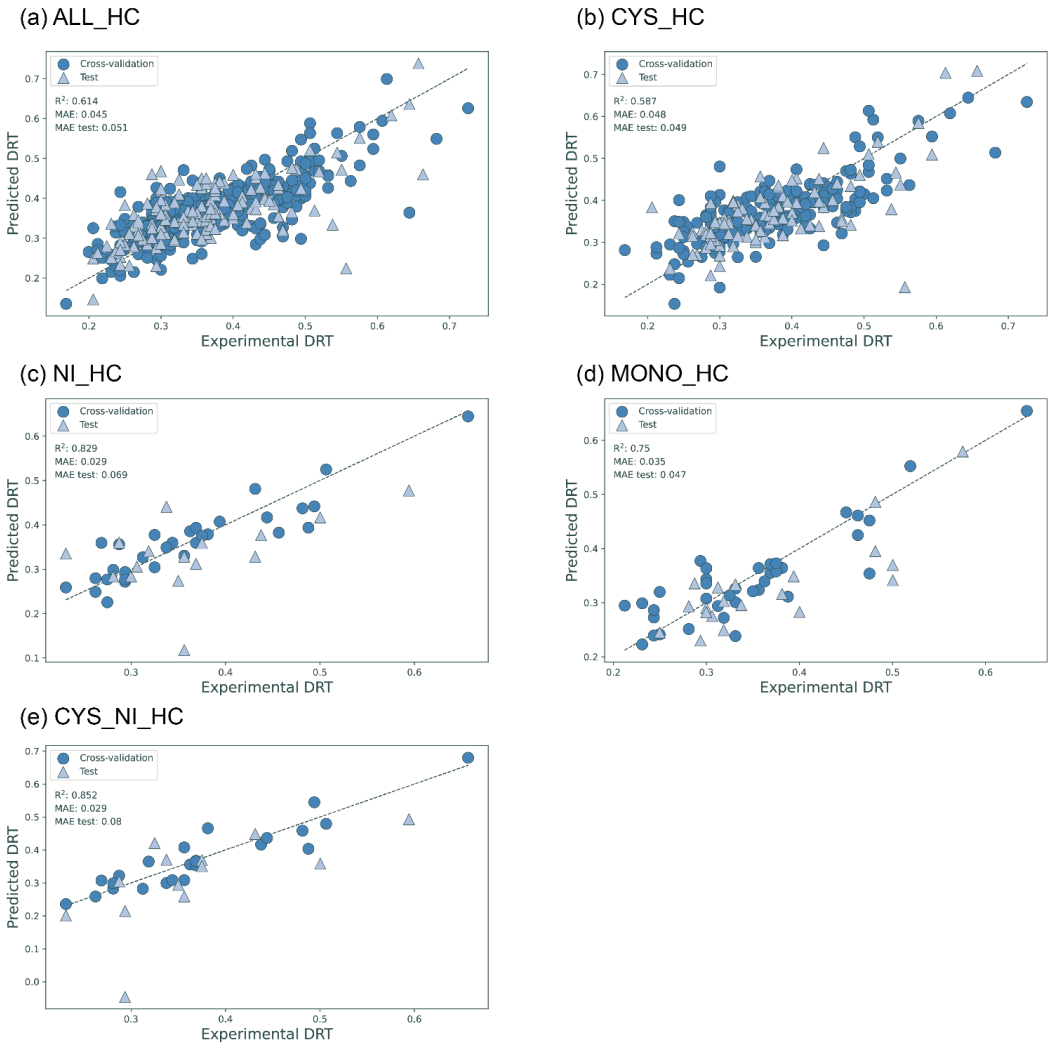
Supplemental Table S6: Regression coefficient and permutation performances for the linear regression model predicting DRT for the NI subset.

Descriptor	Coefficient	Permutation R^2
SurfNegEpMeanAverage	-0.6702	0.4642
SurfEpStdAverage	0.2387	0.6068
SurfNegEpsumAverage	0.3120	0.4672
SurfPosMhpsum	0.1692	0.6139
Dipole	-0.1435	0.6709
LYS surface fraction	-0.0600	0.6612
TYR surface fraction	0.0685	0.6585
ShellEpNegMedianFormal	0.1884	0.6728
CYS surface fraction	-0.0785	0.6836
SurfEpminFormal	0.0783	0.6959
intercept	0.3201	

Supplemental Table S7: Regression coefficient and permutation performances for the linear regression model predicting DRT for the MONO subset.

Descriptor	Coefficient	Permutation R^2
ShellEpNegMedianFormal	-0.1617	0.4965
NSurfPosEpFormal	-0.2318	0.1871
NSurfPosMhp	0.3078	0.4745
SurfMhpSum	0.2956	0.4208
SurfPosEpsumFormal	0.3028	0.5008
SurfNegEpStdFormal	0.1692	0.6861
CYS surface fraction	0.0365	0.6567
GLU surface fraction	0.1032	0.7012
intercept	0.0276	

Supplemental Table S8: Regression coefficient and permutation performances for the linear regression model predicting DRT for the CYS_NI subset.



Supplemental Figure S9: QSPR model results for the different protein subsets. Protein subsets were generated based on all proteins (ALL), proteins present in the cytoplasm (CYT), proteins without PPIs (NI), proteins annotated as monomers (MONO) and proteins with an average pLDDR > 0.95 (HC) or combinations thereof. The circles represent the 10-fold cross-validation and the triangles the test set.

Descriptor	Coefficient	Permutation R^2
SurfNegEpMeanAverage	-0.8810	0.5936
SurfMhpMean	0.5402	0.5738
SurfNegEpsumAverage	0.7996	0.5648
THR surface fraction	-0.0444	0.6111
Average charge	-1.2899	0.5717
SurfEpMaxFormal	0.3437	0.5965
ALA surface fraction	-0.0069	0.6140
SurfNegEpMedianAverage	0.8888	0.5902
ShellEpminFormal	-0.1162	0.6033
SurfEpStdFormal	-0.1678	0.6066
ShellEpPosSumFormal	0.2742	0.6035
Isoelectric point	-0.2400	0.5983
ShellEpPosTrimeanFormal	-0.1224	0.5959
ShellEpPosStdFormal	0.0797	0.6047
NShellPosEpFormal	-0.0774	0.6125
SurfMhpMedian	-0.5114	0.5891
SurfMhpMax	-0.0574	0.6086
TYR surface fraction	0.0617	0.6056
LYS surface fraction	-0.0754	0.6078
VAL surface fraction	0.0702	0.6032
NSurfPosEpFormal	0.1408	0.6104
HIS surface fraction	0.0557	0.6086
SurfMhpmin	-0.0126	0.6127
intercept	0.4896	

Supplemental Table S10: Regression coefficient and permutation performances for the linear regression model predicting DRT for the ALL_HC subset.

Descriptor	Coefficient	Permutation R^2
SurfNegEpMeanAverage	-0.3961	0.4525
SurfMhpMean	0.0696	0.5786
SurfEpSumFormal	0.4994	0.4904
THR surface fraction	-0.0219	0.5905
ShellEpminFormal	-0.2181	0.5746
SurfPosMhpMedian	0.0752	0.5743
LYS surface fraction	-0.0679	0.5846
ShellEpNegStdFormal	-0.1299	0.5783
SurfEpStdFormal	0.0975	0.5795
NSurfPosEpAverage	-0.1575	0.5317
intercept	0.4659	

Supplemental Table S11: Regression coefficient and permutation performances for the linear regression model predicting DRT for the CYS_HC subset.

Descriptor	Coefficient	Permutation R^2
ShellEpminFormal	-0.3764	0.1779
NSurfPosEpFormal	-0.0998	0.7549
SurfNegMhpMean	0.0722	0.7612
GLY surface fraction	0.0914	0.6853
SER surface fraction	0.0792	0.7727
SurfMhpmin	-0.0786	0.7753
intercept	0.6147	

Supplemental Table S12: Regression coefficient and permutation performances for the linear regression model predicting DRT for the NI_HC subset.

Descriptor	Coefficient	Permutation R^2
SurfNegEpMedianFormal	-0.5677	0.0910
SurfNegEpsumFormal	0.3621	0.3506
SurfNegMhpStd	-0.0714	0.6964
SurfNegEpStdAverage	0.0754	0.6861
GLN surface fraction	0.0444	0.7040
CYS surface fraction	0.0388	0.7400
SurfEpminFormal	0.1389	0.7231
intercept	0.3339	

Supplemental Table S13: Regression coefficient and permutation performances for the linear regression model predicting DRT for the MONO_HC subset.

Descriptor	Coefficient	Permutation R^2
SurfEpminFormal	-0.3609	0.1434
SurfPosEpMedianAverage	-0.1125	0.3420
ALA surface fraction	-0.0785	0.3290
GLN surface fraction	-0.0142	0.3588
intercept	0.6665	

Supplemental Table S14: Regression coefficient and permutation performances for the linear regression model predicting DRT for the CYS_NI_HC subset

Chapter 4

Proteomics-based method to comprehensively model the removal of host cell protein impurities

This chapter has been published as:

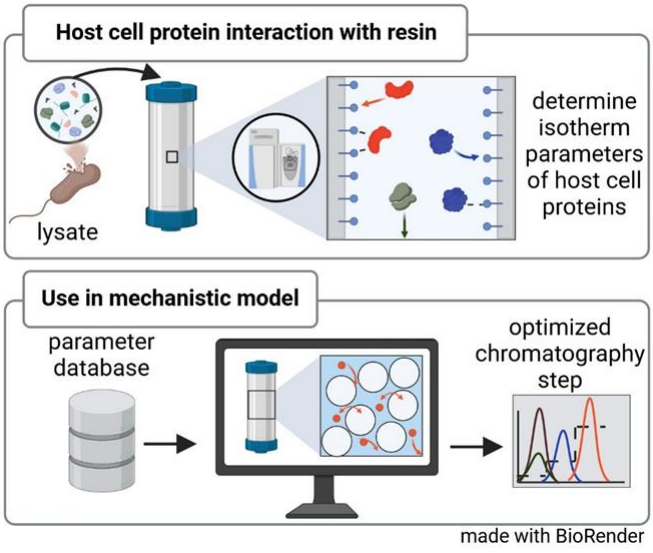
Disela, R., Keulen, D., Fotou, E., Neijenhuis, T., Le Bussy, O., Geldhof, G., Pabst, M. Ottens, M., Proteomics-based method to comprehensively model the removal of host cell protein impurities, *Biotechnology Progress*, 40(6):e3494.

https://doi.org/10.1002/btpr.3494

Abstract

Mechanistic models mostly focus on the target protein and some selected process- or product-related impurities. For a better process understanding, however, it is advantageous to describe also reoccurring host cell protein impurities. Within the purification of biopharmaceuticals, the binding of host cell proteins to a chromatographic resin is far from being described comprehensively. For a broader coverage of the binding characteristics, large-scale proteomic data and systems level knowledge on protein interactions are key.

However, a method for determining binding parameters of the entire host cell proteome to selected chromatography resins is still lacking. In this work, we have developed a method to determine binding parameters of all detected individual host cell proteins in an *E.coli* harvest sample from large-scale proteomics experiments. The developed method was demonstrated to model abundant and problematic proteins, which are crucial impurities to be removed. For these 15 proteins covering varying concentration ranges, the model predicts the independently measured retention time during the validation gradient well. Finally, we optimized the anion exchange chromatography capture step in silico using the determined isotherm parameters of the persistent host cell protein contaminants. From these results, strategies can be developed to separate abundant and problematic impurities from the target antigen.



4.1 Introduction

Host cell protein (HCP) impurities, if present in final drug product, can pose risks to product stability and patient safety. These impurities are released together with DNA, RNA and endotoxins when host cells are disrupted to obtain intracellular recombinant protein products. Compared to medications for chronic diseases, where HCP levels are typically kept below 100 ppm, vaccines allow for higher levels of tolerated HCPs [1]. Regulatory authorities determine acceptable levels of host cell proteins (HCPs) for vaccines on a case-by-case basis [1]. For instance, in the context of a malaria vaccine candidate produced in *E. coli* and intended for administration at 80 µg of a protein antigen per dose, Zhu et al. [2] proposed a limit of 1 µg/dose for every single HCP impurity. Tscheliessnig et al. [3] specified that the total HCP concentration should be 90 ng or < 1100 ppm per dose in this particular case. Developing effective purification sequences to remove HCPs from diverse products often relying on expert knowledge or trial-and-error, emphasizes the crucial need for new, rational, and broadly applicable process development strategies [4]. To gain a higher level of process understanding, mechanistic models (MM) are employed in process development [5]–[7]. MMs describe the underlying physical phenomena during a chromatographic process by incorporating mass transfer correlations and binding kinetics. The binding kinetics are described by adsorption isotherm parameters, valid under the investigated conditions. A key challenge in applying these approaches to real purification problems is finding experimental techniques that are able to determine the necessary isotherm parameters for individual HCPs. This study aims to develop a method to determine isotherm parameters of the individual HCP impurities by coupling linear gradient experiments (LGE) with proteomic analysis. The developed method is applied to determine isotherm parameters of all detected HCPs present in an *E. coli* lysate and optimize a process step to separate an antigen from the HCP impurities.

Mass spectrometry (MS) is an increasingly popular analytical method for HCP analysis, allowing the identification of thousands of proteins within a biological sample [8]–[10]. Extensive research and development efforts have focused on the identification as well as the effective removal of HCP impurities from different hosts [8], [11]–[13]. Specific Chinese hamster ovary cell (CHO) proteins, co-eluting with monoclonal antibodies (mAbs), are referred to as persistent HCP or post-protein A proteins [14].

The interaction between the product and production cell enzymes during cell disruption or enzyme release from dying cells is a potential source of significant damage to the intended native configuration [15]. This damage can lead to irreversible aggregation of the product, substantially reducing the overall yield and raising concerns like immunogenicity, as demonstrated in recent findings indicating HCP involvement in product aggregation [16]–[18]. Similarly, product stability can be impacted by low abundance HCPs such as host cell lipases able to degrade excipients Polysorbate 20 or Polysorbate 80 [19].

However, for products like vaccine antigens produced in *E. coli*, no general persistent proteins are known. *E. coli* lysates, characterized in previous work [13], constitute a complex mixture of approximately 2,000 detected proteins with diverse physicochemical properties (out of approximately 4300 possible gene products in *E.coli*). Especially proteolytic digestion poses a challenge when working with *E.coli* as a host [20].

Recognizing the importance of early removal, particularly of production cell enzymes such as proteases, proves advantageous in preserving product integrity [15]. Another critical group to eliminate is chaperones, proteins involved in correct folding and implicated in human diseases based on immunogenicity [21]. Although it is a high priority to remove these protein groups, they are not necessarily abundant in the cell lysate and are often not individually described.

Several approaches exist to determine isotherm parameters of the major protein impurities when producing mAbs or a therapeutic enzyme [22]–[24]. HCP identities are described according to their experimentally determined physicochemical properties. Fractionation was used to build multi-dimensional property maps, and isotherm parameters for these fractions of CHO HCP impurities were determined using orthogonal chromatographic methods. In a similar manner, a characterization of process-related impurities (including HCPs) in *Pichia pastoris* was conducted on a library of chromatographic resins to describe their affinities [25], [26].

However, these studies employed chromatography as an analytical method. Wierling et al. [27] approached the determination of HCP impurities from CHO cells during the purification of a mAb by combining high-throughput screening with mass spectrometric detection. Mass spectrometry enables the detection of all individual HCPs down to 5 ppm [28]. Compared to anti-

HCP enzyme-linked immunosorbent assays (ELISAs), that detect proteins against which they were developed, mass spectrometry provides information on individual proteins present in the drug substance or product [3].

In this study, we aim to address all detectable HCPs in an *E.coli* cell lysate regardless their abundance. To determine isotherm parameters for all these individual HCPs, proteomic-based mass spectrometry is coupled with LGEs. Fractions obtained from LGEs with varying gradient lengths are analyzed by shotgun proteomics to extract retention times of individual HCPs. From the extracted retention volumes per gradient, isotherm parameters were regressed for all individual HCPs detected in the harvest. Subsequently, a mechanistic model was validated using these isotherm parameters. This validated model was used to optimize a capture step using a two-step elution condition. The presented method can be used to build a comprehensive database with different resins and binding conditions. This one-time determination can be used to feed a mechanistic model used for flow sheet optimization in the future.

4.2 Theory/calculation

4.2.1 Mass balance in mechanistic model

The chromatographic column is modeled using a mechanistic model (in-house python software). This equilibrium transport dispersive model, also called lumped kinetic model, is described in detail elsewhere [29], [30]. In this model, near-equilibrium conditions are assumed, the mass balance equation within the pores is omitted, and the rate of change in stationary phase concentration is directly associated with the deviation of local concentrations from equilibrium [31]. In this context, the mobile phase is considered as the interstitial volume in between resin beads and the stationary phase is considered as the solid particles including the pore volumes. The phase ratio F between stationary phase volume V_s and mobile phase volume V_m is hence described using the bed porosity ε_b as

$$F = \frac{V_s}{V_m} = \frac{(1-\varepsilon_b)}{\varepsilon_b}. \tag{4.1}$$

The mass balance considers the concentration of each protein i in the bulk C_i and in the stationary phase q_i, these balances can be described over space x and time t as follows:

$$\frac{\partial C_i}{\partial t} + F\frac{\partial q_i}{\partial t} = -u\frac{\partial C_i}{\partial x} + D_{L,i}\frac{\partial^2 C_i}{\partial x^2}, \tag{4.2}$$

Where the interstitial velocity of the mobile phase u is determined by the superficial velocity v_0, and the bed porosity εb, expressed as $u = v_0/\varepsilon_b$. The coefficient $D_{L,i}$ characterizes the axial dispersion. To solve the ordinary differential equations (ODE's) the LSODA (Livermore Solver for Ordinary Differential Equations) algorithm was employed. This algorithm automatically switches between the nonstiff Adams method and the stiff BDF method [32].

4.2.2 Mass transfer in mechanistic model

For the mass transfer, a linear film driving force is assumed and the film surrounding the particle is assumed to be stagnant, described as

$$\frac{\partial q_i}{\partial t} = k_{ov,i}\left(C_i - C^*_{eq,i}\right), \tag{4.3}$$

where equilibrium concentration in the bulk phase C_i^* is determined by the isotherm. The overall mass transfer coefficient $k_{ov,i}$ is defined as a summation of the separate film mass transfer resistance and the mass transfer resistance within the pores. Details of the mass transfer are described in appendix 4.7.3.

4.2.3 Derivation of regression formula

To regress isotherm parameters from changes in elution volume according to changes in gradient length, a derivation of the formalism of Parente and Wetlaufer [33] was used adapted to the steric mass action (SMA) isotherm model [34]. The initial slope of this isotherm A_i is described as

$$A_i = K_{eq,i}\Lambda^{v_i}(z_s c_s)^{-v_i}, \tag{4.4}$$

where $K_{eq,i}$ is the equilibrium constant per protein, Λ is the ionic capacity of the resin skeleton, z_s is the charge on the salt counter-ion, c_s is the salt concentration and v_i is the characteristic charge of the protein. The

characteristic charge is described as described as $\nu_i = z_p/z_s$, where z_p is the effective binding charge of the protein. In this study we set $z_s = 1$ since the experiments are conducted using sodium chloride, which means that $z_p = \nu_i$. The protein specific constants $K_{eq,i}$ and ν_i are furthermore called isotherm parameters.

In literature [31] the retention factor k', also known as capacity factor, is described by the retention volume during an isocratic run $V_{R,iso,i}$ and the volume of the mobile phase V_M as

$$k' = \frac{V_{R,iso,i} - V_M}{V_M} = K_i {c_s}^{-m_i} = F K_{eq,i} \Lambda^{\nu_i} (z_s c_s)^{-\nu_i}. \quad (4.5)$$

Parente and Wetlaufer [33] describe the same retention factor as a function of the salt concentration c_s and the constants K_i and m_i. Brooks & Cramer describe the retention factor by using the SMA isotherm model parameter and the phase ratio [34].

This formula can be written in logarithmic form as

$$\log(k') = \log(K_i) + m_i \log({}^{1}/_{c_s}) = \log(F K_{eq,i} \Lambda^{\nu_i}) - \nu_i \log(c_s). \quad (4.6)$$

Consequently, the parameters from Parente and Wetlaufer can be described with the parameters of the SMA isotherm as

$$K_i = F K_{eq,i} \Lambda^{\nu_i}, \quad (4.7)$$

$$m_i = \nu_i. \quad (4.8)$$

Parente and Wetlaufer [33] show that the isocratic elution parameters are transferable to gradient elution retention as

$$V_{R,g,i} = \left(\left({c_{s,0}}^{m_i+1} + \frac{V_m K (m_i + 1)(c_{s,f} - c_{s,0})}{V_G} \right)^{1/(m_i+1)} - c_{s,0} \right) * \frac{V_G}{(c_{s,f} - c_{s,0})}, \quad (4.9)$$

where the gradient retention volume $V_{R,g,i}$ of a protein during gradient elution is described using additionally the initial and final salt concentration $c_{s,0}$ and $c_{s,f}$, and the length of the salt gradient V_G. When varying the gradient volume experimentally, this formula can be employed to regress K_i and m_i of

the analysed protein. Using equation (4.7) and equation (4.8), equation (4.9) can be rewritten as

$$V_{R,g,i} = \left(\left({c_{s,0}}^{v_i+1} + \frac{V_m F K_{eq} \Lambda^{v_i}(v+1)(c_{s,f} - c_{s,0})}{V_G}\right)^{1/(v_i+1)} - c_{s,0}\right) * \frac{V_G}{(c_{s,f} - c_{s,0})} \quad (4.10)$$

as described by Shukla et al. [35]. Important to note is that in this formula the column phase ratio and mobile phase volume are used as defined earlier for the mechanistic model considering only the interstitial volume.

4.3 Material and methods

4.3.1 General method

For this study a new method was developed to determine isotherm parameters of individual HCPs (Figure 4.1). First, the harvest sample was injected into a chromatography column and linear gradient elution (LGE) experiments were employed. Through proteomic analysis of the fractions, the elution profile of individual HCPs were determined. The protein elution profiles were divided into 3 different categories according to their retention behavior. Category 1 shows single peak elution during the salt gradient and is fitted with a Gaussian function. However, some proteins showed multiple peak elution behavior (category 2) or an early elution before the gradient (category 3). Proteins of category 1 were further used to construct the isotherm parameter database. For each protein, the retention volumes during LGE experiments with different gradient lengths were extracted and used in a regression to determine the individual isotherm parameters. For 15 selected HCPs the isotherm parameters were validated in a mechanistic model. The model was furthermore used to optimize a chromatography step separating the antigen from the selected HCPs.

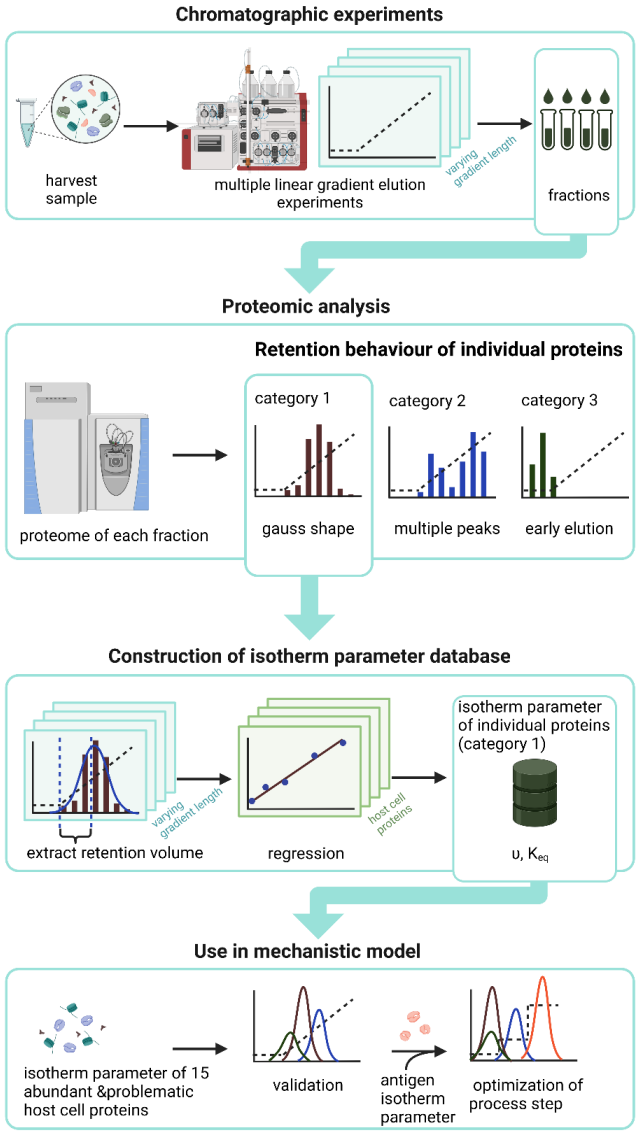


Figure 4.1: Schematic overview of applied method in this study. Chromatographic experiments are conducted using the harvest sample containing a mixture of host cell proteins. The protein mixture is injected to the Äkta chromatography system and linear gradient elution experiments with varying gradient lengths are conducted. From each of the gradient runs, fractions are taken and their proteome is analyzed via mass spectrometry. The majority of proteins, that show Gaussian function behavior, are used to build the isotherm parameter database. Their retention volumes during the varying gradient lengths are extracted and regressed using formula(4.10.). The fitted isotherm parameters for every individual protein are saved in the database. From this database, 15 critical host cell proteins were chosen for validation of the mechanistic model. The validated model was used together with the antigen isotherm data to optimize a capture step removing the 15 host cell protein impurities from the target antigen. (Illustration created using BioRender.com.).

4.3.2 Chromatographic experiments

Chromatographic experiments were conducted to observe the retention behavior of the HCPs and ultimately extract retention volumes used to regress isotherm parameters. The harvest sample was injected on the chromatography column in an Äkta system and several LGE experiments were conducted.

E. coli fermentation and harvest sample

The clarified disrupted harvest sample, used for the LGE experiments, is extensively characterized and described elsewhere [13]. This sample originated from the *E. coli* strain BLR(DE3), for which a fermentation was carried out with an empty plasmid cassette that lacked the gene for the antigen. The harvest material for the analysis of the host cell proteome was provided by GSK (Rixensart, Belgium). All harvest samples were dialyzed with the running buffer using the Slide-A-Lyzer™ G2 Dialysis Cassettes, 2K MWCO (No.10491945).

Materials and apparatus for chromatographic experiments

The chromatographic experiments were performed on an Äkta pure with a connected fraction collector F9-C from Cytiva (Uppsala, Sweden). The dwell volume of the Äkta system, describing the delay between the gradient initiation and the change in the mobile phase composition at the column inlet, was determined as 1.1 ml in a separate experiment described in appendix 4.7.2). A prepacked HiTrap Q Sepharose XL 5ml column from Cytiva (Uppsala, Sweden) was used for chromatographic experiments. The ionic capacity of the resin skeleton was measured by displacement experiments using HCl titration (appendix 4.7.2). It was determined to be 1.106 mmol/l. The running buffer for all experiments was 0.02 M Tris at pH 7.0 with 0.02 M NaCl added. The high salt buffer consisted of the same buffer components with 1 M NaCl added. Between experimental runs the chromatography column was cleaned using 1 M NaOH solution. All buffers were filtered with 0.22 μm pore size and sonicated before use.

Linear gradient elution experiments

LGE experiments were used to determine the retention behavior of the individual proteins and extract the retention volumes of category 1 HCPs (Gaussian function elution). When varying the gradient lengths the obtained retention times are used in a regression to determine the isotherm parameters of these proteins.

The LGE experiments were conducted at a flow rate of 5 ml/min. After injection of 1 ml of the dialyzed harvest sample the column was washed with 5 CV of running buffer. Then, the gradient elution was started by mixing the running buffer with the high salt buffer over varied gradient lengths (5, 10, 20, 30, and 50 CV) until 100% of high salt buffer was reached. The column was regenerated using high salt buffer and 1 M NaOH and then re-equilibrated with the running buffer. During the gradient elution runs, fractions were continuously taken with varied volumes (1, 1, 2, 3, and 5 ml) and afterwards analyzed using mass spectrometry. During the 5 CV gradient, 1 ml fractions were taken and all fractions were analyzed, while for the other gradient lengths 1, 2, 3, and 5 ml fractions were taken and every second fraction was analyzed. Only for the 20 CV gradient 1 ml fractions were collected during the isocratic conditions in the wash before the start of the elution gradient, since isocratic elution behavior was not expected to change under the same conditions.

4.3.3 Proteomic analysis

Shotgun proteomics was employed to identify *E.coli* proteins in each of the fractions taken during the LGE experiment runs and estimate their relative abundance compared to the other fractions collected in the same run. By treating all samples with the same procedure, it was possible to describe the retention behavior of individual HCPs from the relative abundance measurement, despite the unattainability of absolute quantification.

Shotgun host cell proteomics

Before the mass spectrometry analysis, the samples were prepared using the filter aided sample preparation (FASP) developed to simplify the preparation of samples [36], [37]. The applied method is further described in the appendix 4.7.1.

The SpeedVac dried peptide fractions were reconstituted in a solution comprising 3% acetonitrile and 0.01% trifluoroacetic acid (TFA) in LC-MS water. An aliquot, representing approximately 500 ng of the digested sample, was subjected to analysis using a nano-liquid chromatography separation system. This system featured an EASY-nLC 1200 instrument equipped with an Acclaim PepMap RSLC RP C18 separation column (50 µm x 150 mm, 2 µm particle size, and 100 Å pore size), coupled to a QE plus Orbitrap mass spectrometer (Thermo Scientific, Germany).

Reversed phase chromatography was performed at a flow rate of 350 nL/min before the mass spectrometry, with solvent A comprising LC-MS water and 0.1% formic acid, while solvent B consisted of 80% acetonitrile in water and 0.1% formic acid. The separation was achieved using a linear increase of solvent B from 2% to 40% over 60 minutes.

The Orbitrap mass spectrometer operated in data-dependent acquisition (DDA) mode, capturing spectra at a resolution of 70,000 over the m/z range of 385 to 1,150. The top 10 signals were selected for isolation with a window of 2.0 m/z and an isolation offset of 0.1 m/z, followed by fragmentation employing a normalized collision energy (NCE) of 28. Fragmentation spectra were acquired at a resolution of 17,000, with an automatic gain control (AGC) target of 5e5 and a maximum injection time (IT) of 75 ms. Unassigned, singly charged, and ions with 6 or more charges were excluded from fragmentation. Dynamic exclusion was set to 60 s.

Processing of mass spectrometric raw data

Mass spectrometric raw data was analyzed utilizing PEAKS Studio X, an application developed by Bioinformatics Solutions Inc., Canada. The analysis allowed for a 20 ppm tolerance for parent ion mass error and a 0.02 Da tolerance for fragment ion mass error. The analysis considered parameters such as the potential for 3 missed cleavages, carbamidomethylation as a fixed modification, and methionine oxidation, N/Q deamidation, and N-terminal acetylation as variable modifications.

To enhance the analysis, strain-specific proteome sequence databases were obtained from NCBI (BioProject PRJNA379778), and sequences of contaminant proteins were sourced from the Global Proteome Machine (GPM) database (https://www.thegpm.org/crap/). A decoy fusion strategy

was employed to estimate false discovery rates (FDRs). The filtering of peptide spectrum matches was carried out with a threshold of 1% FDR, and proteins with more than one unique peptide sequence were considered statistically significant.

To assess changes in protein abundance between the different fractions, label-free quantification was performed using the PEAKSQ module [38]. The abundance measure utilized in this analysis was the peak area obtained from the reversed-phase column prior to entering the mass spectrometer. Exclusively proteins that were identified with more than 3 peptides were used in the further analysis.

Processing of retention behavior of individual host cell proteins

Peak area was used as an abundance measure and plotted per fraction. The middle of the fraction was used as the value of volume during the chromatographic run. Retention volumes of every individual HCP during the five gradient runs were extracted using an in-house python script. The first fraction taken during the wash was excluded from the retention analysis, as these fractions most likely only contain digested peptides and the MS analysis did not distinguish between digested and undigested proteins.

To determine, which retention behavior was observed for individual proteins, the maximum value of abundance (in peak area) was determined. If this maximum was located before the start of the elution, proteins were assigned to category 3. The retention profiles of the remaining proteins were fitted to a Gaussian curve. If the shape was fitted with a R^2 below a set limit, the proteins were considered category 2, containing multiple peaks. The set limit for R^2 was 0.7 for the 10, 20, 30 and 50 CV runs and 0.5 for the 5 CV runs, since the abundance values occasionally reached saturation here. If the R^2 was above the limit, proteins were considered as category 1.

4.3.4 Construction of isotherm parameter database

For proteins in category 1, the maximum of the Gaussian function was extracted as retention volume of the raw data. Only for proteins that showed this retention behavior, it was possible to determine isotherm parameters with confidence.

Processing of retention volumes

The retention volumes of the varying gradient lengths used in the regression were calculated as

$$V_{R,g,i} = V_{R,g,i,raw} - 0.5\,V_{inj} - V_{dwell} - V_m - V_{wash}, \quad (4.11)$$

where $V_{R,g,i}$ is the corrected retention volume used in the regression. Half the injected volume V_{inj}, the dwell volume of the system V_{dwell}, the volume of the mobile phase V_m, and the volume of the wash before elution V_{wash} are subtracted from the raw data retention volume $V_{R,g,i,raw}$.

Regression of host cell protein isotherm parameters

The corrected retention volumes of 4 different gradient lengths were used in a weighted regression of the regression formula (Eq.(4.10)) utilizing an in-house python script with the optimize.curve_fit function from the scipy package. The 10 CV gradient elution experiment was left out for validation. Weights were assigned according to the fractionation scheme during the gradient elution runs, since a higher fractionation volume is associated with higher uncertainty of the exact retention volume. Less weight was given to the runs with higher fraction volumes by assigning the inversely dependent sigma values 0.1, 0.4, 0.6, and 1 to the 5 CV, 20 CV, 30 CV, and 50 CV gradient elution runs. From the employed weighted regression, isotherm parameters of individual HCPs (in category 1) were extracted.

Determination of antigen isotherm parameters

The isotherm parameters of the antigen (and the charge variant) were determined in a similar manner. LGE experiments with various gradient lengths (5, 10, 20, 40, and 60 CV) were conducted using purified antigen. The maximum of the main peak was extracted using the signal obtained from the UV spectrometer at 230 nm wavelength instead of employing mass spectrometry. This value was used as the raw data retention volume of the antigen, while an earlier eluting smaller peak was identified to be a charge variant. The corrected retention volumes were obtained with Eq. (4.11), and used to regress the isotherm parameters utilizing Eq. (4.10). Antigen isotherm parameters are then used in the mechanistic model as the

parameters of the target molecule that has to be separated from HCP impurities and the antigen charge variant.

4.3.5 Validation of host cell protein isotherm parameters in mechanistic model

For the validation of the HCP isotherm parameter in the mechanistic model, 15 proteins were selected and their retention behavior was modelled for the left out 10 CV gradient experiment. For the 15 proteins, the modeled retention volumes and elution peak shapes were compared with the experimentally determined data. As an input for the mechanistic model, a relative protein concentration was used (listed in Table 4.1). These concentrations were obtained from integration of the Gaussian functions that were fitted to the experimental data (of the 20 CV gradient). These values are given in percent of the peak area of the Gaussian function from each individual protein in relation to the total of all the proteins. The relative antigen concentration was calculated from the measured relation of the antigen to the total of all the proteins.

4.3.6 Optimization of chromatography step in mechanistic model

For this case study, an AEX capture step was optimized with the antigen as protein of interest. The optimization involved a two-step elution mode to mimic an industrial process. The global and local objective were formulated as follows:

$$minf(x) = 0.5 * \big(100 - yield(x)\big) + 0.4 * \big(100 - purity(x)\big) + 0.1 * buffer\ consumption\ (x) \tag{4.12}$$

$$s.t.\ \ h(x) = 0 \tag{4.13}$$

$$0 \leq x \leq 1, \tag{4.14}$$

where the objective is to minimize function f, in which the variables x were normalized between 0 and 1 for enhanced optimization purposes (Eq. (4.14)). Moreover, it is important to satisfy the mass balances and equilibrium relations as denoted in Eq. (4.13). A total of six variables were optimized: the salt concentration for the first and the second step, the gradient lengths for both steps, and the lower and upper cut points. The main objective for a capture step is obtaining a high yield, followed by the purity, and a low buffer consumption. The buffer consumption indirectly reflects the

costs, batch throughput, and productivity, as it minimizes the time needed to perform this purification step.

For the global optimization the differential evolution algorithm from the *scipy.optimize* package was utilized with 9 maximum number of iterations, a population size of 10 and Latin hypercube sampling to initialize the population. The Nelder-Mead algorithm was employed for the local optimization with a maximum number of iterations of 100. The relative and function tolerances for both global and local optimizations were set to 1e-2. The boundaries of both step lengths are between 0.1 – 9.99 CV. The salt concentration of the first step has to be between 5 – 499.5 mM, and of the second step between 300 – 999 mM. Lastly, the lower cut point is bound between 1 – 80% of the peak maximum on the left, while the upper cut point is between 20 – 99.9% of the peak maximum on the right.

4.4 Results and discussion

4.4.1 Retention behavior of individual host cell proteins

1,247 *E.coli* HCPs were identified via MS throughout all fractions in the 20 CV gradient run. The retention behavior of individual HCPs during the gradient elution was classified into 3 categories (sections 0 and 0). Most proteins fall into category 1 (898 proteins; 72 %), which shows a single peak elution during the salt gradient and therefore can be fitted well with a Gaussian function (Figure 4.2 (b)). However, 121 proteins (10 %) falling into category 2 showed multiple peaks or abundance in only a single fraction during the elution, so it was not possible to fit a Gaussian function (Figure 4.2 (c)). 215 proteins (17 %), falling into category 3, had their abundance maximum during the wash before the start of the salt gradient (Figure 4.2 (d)). The remaining 13 proteins were detected in the sample but were below the limit of quantification. The abundance of the detected proteins in the load sample is shown in 'peak area', as measured by mass spectrometry. It is plotted over the retention volume of the identified proteins in Figure 4.2 (a). Hereby, the maximum of the Gaussian function was considered the retention volume for proteins of category 1, while the maximum abundance value was used for proteins of category 2 and 3.

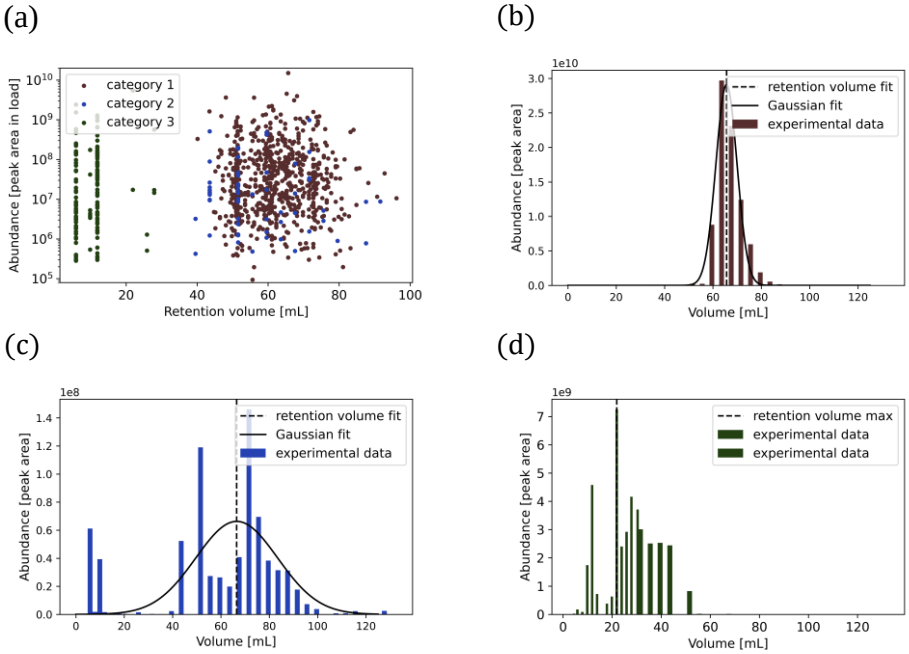


Figure 4.2: Categories of HCPs: (a) scatterplot showing the retention behaviour of the 3 categories of protein during a 100 ml (20 column volume) gradient on a 5 ml HiTrap Q Sepharose XL column, (b) example for category 1: single peak Gaussian function ($R^2 > 0.7$, here 0. 97) of 'translation elongation factor EF-Tu 1&2' (ARH99640.1), (c) example for category 2: multiple peaks eluting ($R^2 < 0.7$, here 0.41) in case of '30S ribosomal subunit protein S3' (ARH98930.1), (d) example of category 3: protein eluting before the gradient when observing 'RNA chaperone and antiterminator cold-inducible' (ARH99188.1).

The ideal elution behavior seen by proteins of category 1 makes it possible to extract retention volumes of the different gradients with confidence, illustrated in Figure 4.3 for the most abundant protein "translation elongation factor EF-Tu 1&2". Here small differences in absolute protein concentrations between experiments could be caused by fluctuations due to the dialysis. Especially for the 5 CV gradient, values close to saturation were observed, therefore it is advised to dilute the sample more, or to apply longer salt gradients. The retention volumes can further be used to extract isotherm parameters and mechanistically model the protein behavior on the tested column and conditions. This was possible for 721 proteins, for which Gaussian functions could be fitted with sufficient accuracy in all 5 gradients. The proteins from category 2 can be determined with less confidence, as the protein abundance is very low or the protein shows different isoforms. Different isoforms can be caused by charge variants or the formation of complexes with other protein species. Elution before the start of the gradient

described by proteins in category 3 could have several reasons. The proteins could simply have no affinity to the anion exchange resin because of a positive net-charge. Another reason can be that some proteins eluting during the first fractions might be digested by proteases in the harvest sample. Therefore, these proteins were at the time of the LGE gradient experiments only present as peptides. Peptides would more likely elute directly in the first fractions since less interaction with the resin is expected. No size differences were observed between the different protein categories and therefore size exclusion effects of proteins of category 3 can be disregarded.

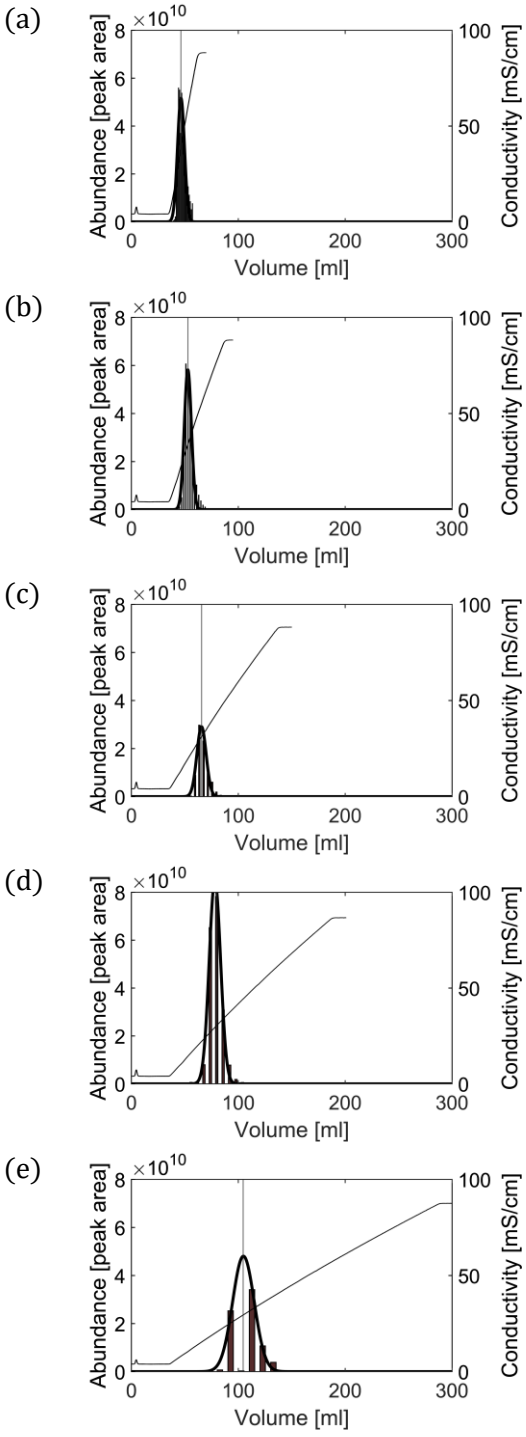


Figure 4.3: All 5 linear gradient elution experiments for "translation elongation factor EF-Tu 1&2" (ARH99640.1): (a) 25 ml (5 column volumes) gradient, (b) 50 ml (10 column volumes) gradient, (c) 100 ml (20 column volumes) gradient, (d) 150 ml (30 column volumes) gradient, (e) 250 ml (50 column volumes) gradient; The abundance in the fractions was determined using MS and displayed as peak area. Gaussian functions are fitted to the abundance data. The maxima of the fitted Gaussian functions are extracted as retention volumes and used in the isotherm parameter regression. Only the 50 ml gradient is left out for validation of the mechanistic model.

4.4.2 Selection of abundant and problematic proteins

While this big data lake of isotherm data is very insightful, it is not practical (and not necessary) to model every single protein as the mechanistic model would take days to perform one simulation, even more so a whole optimization that requires about 500 simulations. Therefore, we made a selection of proteins that are in our interest to be simulated in this study. Since the aim of our study is to find the optimal process to purify an antigen from HCP impurities, we choose the most relevant proteins to be removed in a capture step. As described in the introduction, abundant proteins, proteases, and chaperones are of high priority to be removed early on in the process. Hence, we choose to select the 5 most abundant proteins, the 5 most abundant proteases, and the 5 most abundant chaperones present in the dataset. These 15 HCPs together with their retention behavior, properties, and determined isotherm parameters are listed in Table 4.1.

The chosen proteins span a broad spectrum of abundances, demonstrating the applicability of our method to problematic proteins with varying concentrations. In comparison to abundant proteins, chaperones are present in a relative concentration reduced by a factor of 10, while proteases show a reduction by a factor of 100. Despite their lower abundance, proteins like proteases, often overlooked, can pose significant issues, such as protein degradation. Therefore, it is imperative to address and remove less abundant proteins early in the process to mitigate potential complications, as emphasized in literature [9].

Table 4.1: 15 abundant and problematic host cell proteins in focus of this study to be removed from the target antigen.

protein type	protein name	protein ID	relative concentration [% of area under Gaussian curve]	$K_{eq,i}$	stdev $K_{eq,i}$	v_i	stdev v_i	R^2	RSME	volume difference correlation to experiment [ml]	volume difference model to experiment [ml]
abundant	translation elongation factor EF-Tu 1&2	ARH99640.1	6.794	0.142	0.077	2.641	0.414	0.981	3.395	-1.484	-0.418
	protein chain elongation factor EF-G GTP-binding	ARH98956.1	2.753	0.276	0.212	2.274	0.605	0.942	6.378	-2.633	-1.465
	30S ribosomal protein S1	ARH96687.1	1.701	0.295	0.165	2.346	0.458	0.967	4.843	-2.717	-1.547
	glyceraldehyde-3-phosphate dehydrogenase A	ARH97514.1	1.607	0.009	0.007	2.620	0.363	0.977	1.535	-0.431	0.690
	isocitrate dehydrogenase (NADP(+))	ARH96911.1	1.543	0.007	0.005	3.664	0.501	0.989	1.825	-0.612	0.424
chaperone	Cpn60 chaperonin GroEL large subunit of GroESL	ARH99809.1	0.511	0.245	0.171	2.982	0.666	0.974	5.506	-4.791	-3.729
	molecular chaperone DnaK	ARH95794.1	0.543	0.520	0.284	2.063	0.468	0.944	6.755	-3.737	-2.544
	Cpn10 chaperonin GroES small subunit of GroESL	ARH99808.1	0.223	0.003	0.004	4.569	0.965	0.984	2.786	-0.639	0.464
	protein disaggregation chaperone	ARH98235.1	0.216	0.003	0.005	4.777	1.191	0.980	3.294	-0.809	0.255
	Fe/S biogenesis protein putative scaffold/chaperone protein	ARH99034.1	0.112	0.250	0.102	2.623	0.352	0.982	3.746	-1.723	-0.571

[redacted]

Table 4.1: 15 abundant and problematic host cell proteins in focus of this study to be removed from the target antigen.

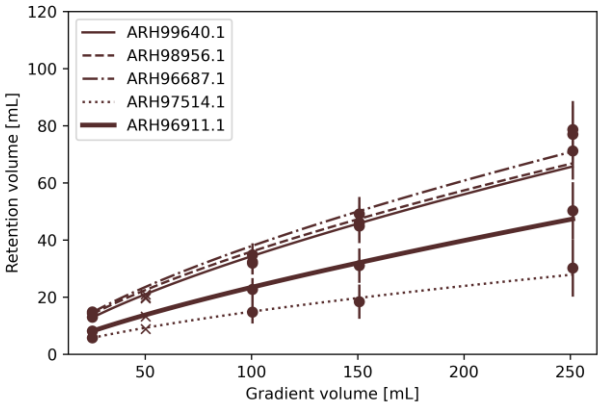
protein type	protein name	protein ID	relative concentration [% of area under Gaussian curve]	$K_{eq,i}$	stdev $K_{eq,i}$	v_i	stdev v_i	R^2	RSME	volume difference correlation to experiment [ml]	volume difference model to experiment [ml]
protease	carboxy-terminal protease for penicillin-binding protein 3	ARH97573.1	0.035	0.010	0.018	3.150	0.999	0.966	2.922	-0.601	0.478
	ATP-dependent Clp protease proteolytic subunit ClpP	ARH96177.1	0.084	0.259	0.196	1.996	0.531	0.919	6.316	-3.062	-1.844
	ATP-dependent Clp protease ATP-binding subunit ClpX	ARH96178.1	0.025	0.201	0.180	2.529	0.712	0.950	6.124	-2.190	-1.130
	modulator for HflB protease specific for phage lambda cII repressor	ARH99838.1	0.003	0.001	0.002	8.049	3.729	0.976	5.834	-1.231	-0.127
	molecular chaperone and ATPase component of HslUV protease	ARH99598.1	0.008	0.646	0.474	1.518	0.538	0.844	9.387	-4.598	-3.078

4.4.3 Isotherm parameter regression of individual host cell proteins

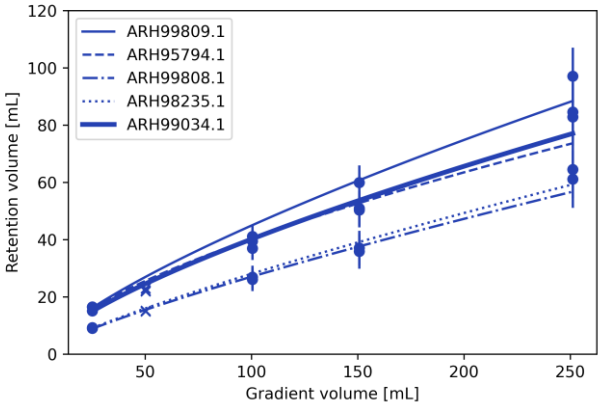
The retention volumes of the individual HCPs and the value of total gradient elution volume are related to each other via the formula from Parente and Wetlaufer as given in Eq. (4.10) and shown in Figure 4.4. Experimentally determined retention volumes (shown as dots) are fitted with the given formula (shown as lines) and compared to the 10 CV (50 ml) gradient that was left out for validation (shown as x). Fractionation size and frequency was considered to determine the experimental error (plotted error bars). The regressed values and their standard deviation obtained for the 15 selected abundant and problematic HCPs are added to Table 4.1.

The weighted regression (Figure 4.4) leads to a slight upwards bend in the fitted functions. The bend leads to a small overestimation of the shorter gradient lengths and slight underestimation of the longer gradient lengths compared to the experimental values. Though, these differences are still within the experimental error and the fitted function describes the data well. Non-weighted regression was also investigated, however, this provided $K_{eq,i}$ values close to the set boundaries with very high standard deviations caused by improper scaling of the data.

(a)



(b)



(c)

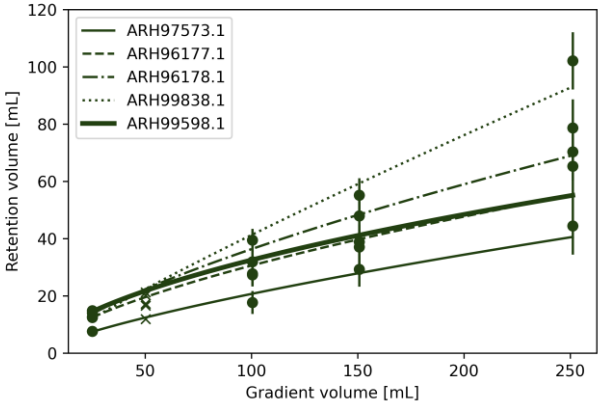


Figure 4.4: Regression of (a) 5 most abundant proteins, (b) 5 most abundant chaperones, (c) 5 most abundant proteases in harvest sample; The dots are the experimentally measured values with error bars according to the fractionation scheme. The lines connecting the dots show the regressed fit. Full protein names are listed in Table 4.1. The value for the 50 ml gradient run, marked with an x, was used as a validation run and not included in the fit.

An overview of the 721 values obtained for the isotherm parameters $K_{eq,i}$, v_i, and their standard deviations are given in form of boxplots in Figure 4.5(a)-(b). The values obtained for R^2 and RSME are shown in boxplots in Figure 4.5(c)-(d). Determined $K_{eq,i}$ values were between 0.0001 and 5.43. The standard deviation of this parameter was determined between 0.00015 and 1.46. Parameters determined for the characteristic charge varied between 0.47 and 13.93. For none of the proteins the regressed values were exactly at the given boundaries of $K_{eq,i}$ (0.00001 and 100) and v_i (0.1 and 15). The standard errors of the regressed parameters are on average 116 % of the parameters nominal value for $K_{eq,i}$ and 21 % for v_i. These relative high standard deviations for especially $K_{eq,i}$ might be caused by the relatively low absolute values. However, the R^2 varied between 0.47 and 1.00 with an average of 0.97, meaning that the fit with the regression formula described the experimental data well for the majority of the proteins. Likewise, the RSME values, varying between 0.11 and 41.12 ml with an average of 3.35 ml (6.7 % in a 50 ml gradient), indicate that the fitted regression formula describes the data well. More importantly, the differences for all HCPs in retention volume between the left out validation run and calculation from the correlation are low with an average of 1.23 ml (2.5 % in a 50 ml gradient) and a maximum value of 8.21 ml (16.4 % in a 50 ml gradient). Based on these results, we conclude that the regression function with the fitted isotherm parameters can describe the experimental data with high accuracy.

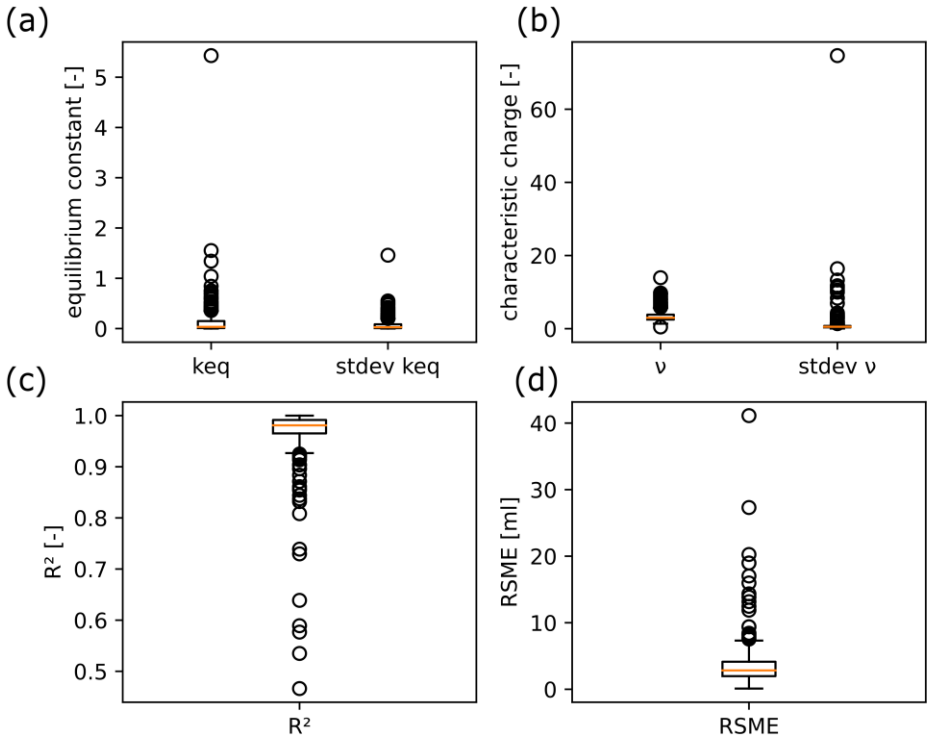


Figure 4.5: Overview of the regressed isotherm parameters of the complete host cell protein dataset. (a) equilibrium constant $K_{eq,i}$ and standard deviation of all HCPs, (b) characteristic charge ν_i and standard deviation of all HCPs, (c) R^2 of all HCPs, and (d) RSME of all HCPs.

4.4.4 Validation in mechanistic model

The average pore size of the Q Sepharose XL resin is described as 54 nm for the agarose skeleton [39] and 12 nm [40] including the dextran-graft that bind the ligands. Using a pore diameter of 12 nm for the resin in the mechanistic model lead to size exclusion effects and hence an early elution of HCPs. However, from the experimental data, it was concluded that no such size exclusion effects occurred for the HCPs. Hence a pore diameter of 54 nm was used in the mechanistic model assuming the flexible dextran-grafts inside the pores do not hinder the access of the HCPs into the pores.

For the validation, isotherm parameters of the 15 selected abundant and problematic HCPs, listed in Table 1, were used in the mechanistic model to simulate the left out 10 CV (50 ml) gradient run. These simulations were compared to the experimental result. In addition, the obtained isotherm

parameters of the antigen and its charge variant were simulated together with the 15 HCPs to compare their retention behavior.

Volume differences between the modeled retention volumes and the experimental retention volumes are shown in Table 1. The average volume difference for the 15 HCPs is -0.94 ml (1.9 % in 10 CV gradient). This is lower than the average difference in volume from the correlation for the selected HCPs with -2.08 ml (4.2 % in 10 CV gradient). The maximum volume difference is reached by "Cpn60 chaperonin GroEL large subunit of GroESL" (ARH99809.1), further called Cpn 60 chaperonin, in both datasets with -3.73 ml (7.5 % in 10 CV gradient) by the model and -4.79 ml (9.6 % in 10 CV gradient) by correlation. While the differences in volume for the correlation are all negative, meaning the correlation predicts a later elution than experimentally measured, the model predicts 5 proteins to elute earlier than the experiment. Both the correlation and the model slightly overestimate the retention volume of the validation run. However, the differences are below 10 % and considered minor for the selected HCPs, that cover a big range of concentrations.

As explained previously in sub-chapter 4.3.4 the experimental data was fitted with a Gaussian curve. Figure 4.6 shows a side by side comparison of the experimental Gaussian curves (Figure 4.6(a)-(b)) and modeled curves (Figure 4.6 (c)-(d)). Thereby Figure 4.6(a) and (c) show the extended view, while Figure 4.6(b) and (d) are zoomed in to show low abundance peaks. Overall, similar peak shapes can be observed, despite their different abundance measures.

The height and width of the peaks are determined by a combination of regressed isotherm parameters and mass transfer correlations. Higher $K_{eq,i}$ values lead to a later retention with a more shallow, wide peak form. The width of the peaks in the middle of their height was determined for each protein. Compared to the experimental values, the modeled peaks had an average of 132 % width with the maximum for Cpn 60 chaperonin at 300 %. Overall the peaks are displayed well considering the chosen method to determine isotherm parameters solely based on their retention volume and without fitting any mass transfer or peak shapes. Slightly wider peaks in the model additionally calculate the worst case scenario and hence lead to a more robust process.

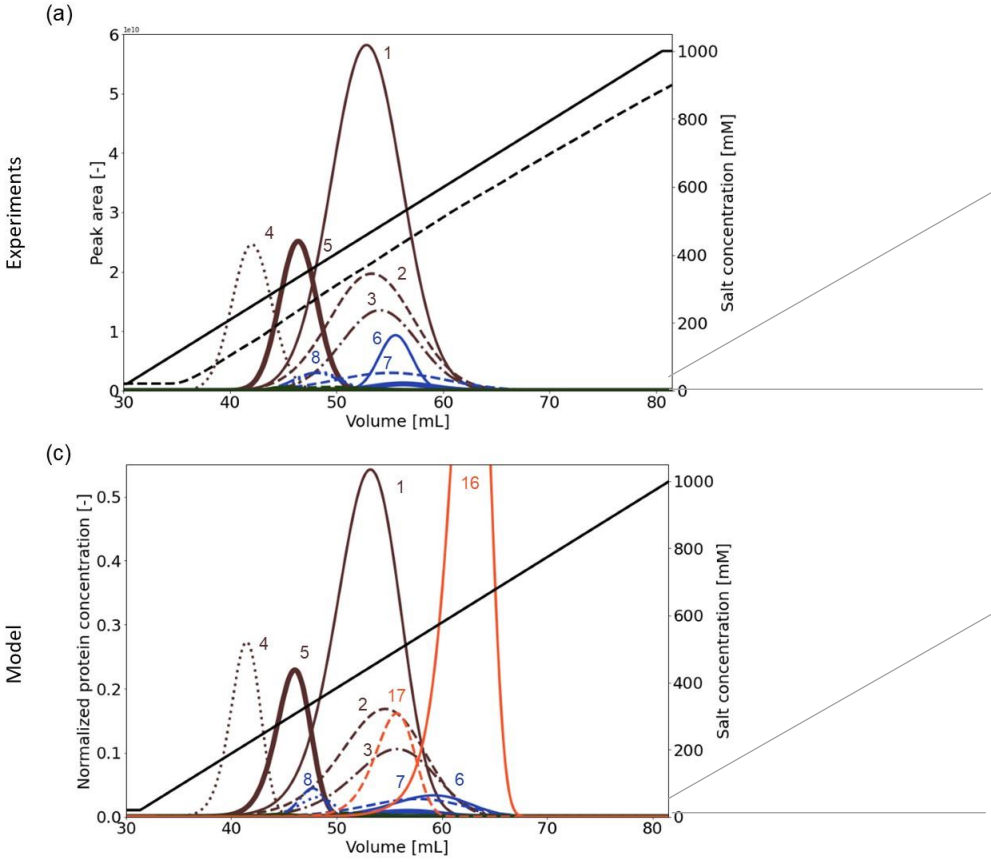
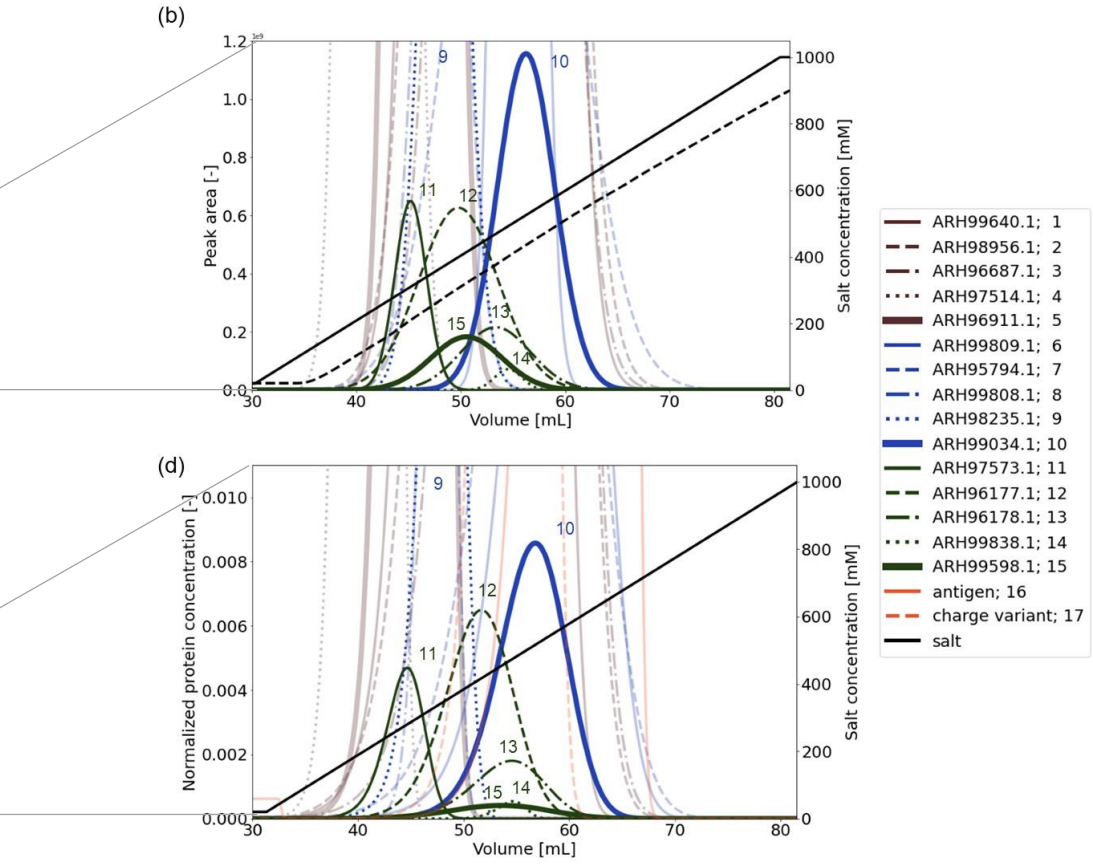


Figure 4.6: Comparison of experimentally determined and modeled retention behavior during the validation experiment of 15 abundant and problematic HCPs. In the graphs the retention during a 50 ml (10 column volume) gradient with a 5 ml Q Sepharose XL column is shown. Gaussian functions fit to the experimental raw data in peak area are shown in: (a) full view, (b) zoom in. Modeled elution of the components described by a the normalized protein concentration are shown in: (c) full view, (d) zoom in.

Even though the majority of proteins is displayed very well, Cpn 60 chaperonin and "molecular chaperone and ATPase component of HslUV protease" (ARH99598.1) stand out with the biggest difference between the model and experiment leading to a changed elution order. Both proteins elute later with a more shallow peak in the model compared to the experiment. The regression calculation shows a later expected elution compared to the experiment. Hence the regressed $K_{eq,i}$ is fitted to be higher, which leads to a later elution with a shallow peak shape. However, throughout the diverse concentration range of selected proteins, the model was able to simulate retention times and peak shapes well.



4.4.5 Optimization of capture step

The validated model was used to find the optimal capture step conditions on the 5 ml HiTrap Q Sepharose XL column to separate the antigen from the HCPs of interest with focus on the yield, purity and buffer consumption. In Figure 4.7, a chromatogram of the chosen optimized step elution is shown using 20 mM Tris buffer. First, the column is washed with 20 mM NaCl as a running buffer after the 1 ml injection of the load. The majority of HCPs are removed during the 362 mM NaCl wash (9.43 CV). Finally, the antigen elutes during the 634 mM NaCl step lasting 2.75 CV. The collected eluate, highlighted in white, has a yield of 98 % and purity of 99 %.

Impurities that co-elute with the antigen obtained in the product pool are discussed below in descending abundance. Cpn 60 chaperonin is expected to co-elute partially with the antigen. Since the $K_{eq,i}$ was slightly overestimated in the model compared to the experiment as discussed previously (chapter

4.4.4), less Cpn 60 chaperonin might co-elute in an experiment with the antigen than the model calculates. However, this protein is often detected as an impurity in the final drug product due to strong binding affinity to all proteins. Cpn 60 chaperonin has been identified to be very immunogenic and has a high priority to be removed as early as possible in the process [21]. In this case, the model predicts the worst case and this makes the actual process step more robust. Another protein, that co-elutes partially with the antigen, is its charge variant. The majority of the charge variant of the antigen is removed in the optimized capture step.

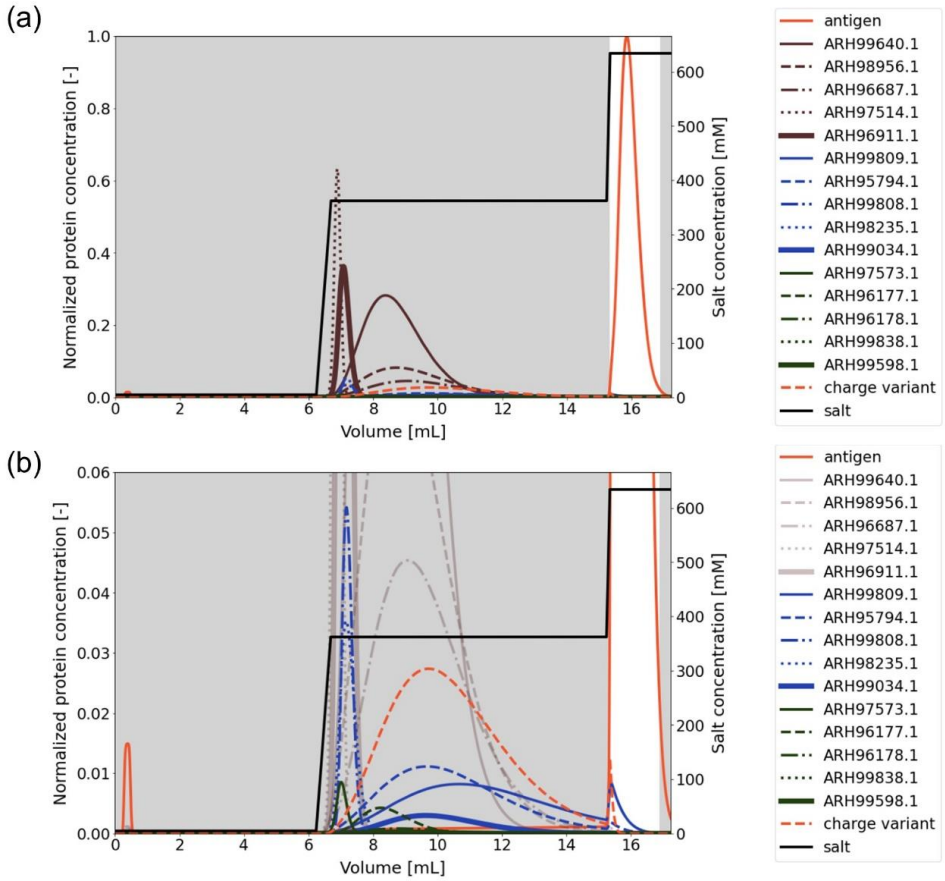


Figure 4.7: Model of optimized capture step for antigen on 5 ml HiTrap Q Sepharose XL column considering 15 selected host cell proteins to be removed (full names in table 1). First, the column is washed with 20 mM NaCl as a running buffer after the 1 ml injection of the load then with the 362 mM NaCl wash (9.43 CV). Finally, the antigen elutes during the 634 mM NaCl step lasting 2.75 CV. The collected eluate, highlighted in white, has a yield of 98 % and purity of 99 %. (a) full view, (b) zoom in.

However, the remaining charge variant could be removed in a consecutive orthogonal polishing step, if required. “molecular chaperone DnaK” (ARH95794.1) also co-elutes with the antigen. It shows a later retention in the model compared to the experiment and might be removed even more effective in reality, if it does not bind to the antigen itself. In literature [41], it was shown that this protein shows a high immunogenicity in mice. Here the model shows the worst case scenario and therefore finds a robust optimal process. We show that it is possible to use the validated model to find the optimal process step. This optimized step is used to separate the target antigen from the charge variant, and 15 abundant and problematic proteins.

4.5 Conclusions and outlook

In this work, we developed a method that combines gradient elution experiments with proteomic analysis. This method allows the determination of isotherm parameters for individual HCPs of a varying concentration range. Since the elution behavior of the individual proteins is measured while they are in a mixture, effects such as binding or co-elution of proteins in the feed sample were inherently described at the measured conditions. The different retention behaviors of the individual proteins were categorized. Only proteins with single Gaussian function elution were used to regress isotherm parameters, since in this case retention volumes could be determined with confidence. 15 abundant and problematic HCPs out of the isotherm parameter database were selected to validate the mechanistic model. In the mechanistic model the use of the isotherm parameters lead to an average volume difference of 7.5 % during a 10 CV gradient length compared to experiments. This accurate model was used to optimize a capture process step to remove the majority of the impurities from the antigen, achieving a yield of 98 % and purity of 99 %. This case study exemplifies, how the HCP database can be applied to fasten process development in the future.

In the future, this method might be applicable to design a new capture step for an unseen/new protein produced in *E.coli*. In this case, isotherm parameters of the new protein and product-related impurities are required. The existing database can also be used to describe other *E.coli* strains since abundances and protein concentrations are comparable for different strains [13]. In principle, the method can also be applied to other hosts such as *pichia pistoris* with a similar number of possible gen products. The number of proteins expressed by CHO cells might lead to increased analysis times and

efforts and requires some more attention on product-related impurities. Since hitchhiker proteins involved in aggregates pose a challenge for example for mAbs produced in CHO [17], the authors would suggest a joint measurement including the mAb. Present aggregate isotherms could be determined and treated as another impurity by the mechanistic model. Another approach would be to target proteins involved in protein-protein interactions [42] and use their isotherms in the mechanistic model to remove these early on in the process.

The accuracy of the isotherm parameters depend on the accuracy of the regression and the resolution used during the LGE experiments. An increased number of fractions collected during the LGE experiments results in improved accuracy of HCP isotherm parameters. However, since mass spectrometry is a costly and work-intensive/laborious analytical method, it is desirable to limit the number of samples. In the future, other fractionation schemes might be considered for example by keeping the fractionation volume constant throughout different gradient experiments.

This cutting-edge proteomics method enables to determine adsorption isotherm parameters for the entire proteome. The existing database can be expanded with HCP isotherm parameters for other resins or binding conditions. Once this universal impurity database exists, chromatography steps for new products can be developed mainly *in silico* with minimal experimental effort, characterizing only the binding behavior of the target protein and product-related impurities. The binding and elution behavior of the HCP impurities can be described by the mechanistic model using the isotherm database and knowledge can be transferred between different products. The experimental method for this one-time characterization of the host cell proteome binding behavior is providing the data needed for a computational led process development.

4.6 References

[1] K. Reiter, M. Suzuki, L. R. Olano, and D. L. Narum, “Host cell protein quantification of an optimized purification method by mass spectrometry,” *J. Pharm. Biomed. Anal.*, vol. 174, pp. 650–654, 2019.

[2] D. Zhu, A. J. Saul, and A. P. Miles, “A quantitative slot blot assay for host cell protein impurities in recombinant proteins expressed in E. coli,” *J. Immunol. Methods*, vol. 306, no. 1–2, pp. 40–50, Nov. 2005.

[3] A. L. Tscheliessnig, J. Konrath, R. Bates, and A. Jungbauer, "Host cell protein analysis in therapeutic protein bioprocessing - methods and applications," *Biotechnol. J.*, vol. 8, no. 6, pp. 655–670, 2013.

[4] A. T. Hanke and M. Ottens, "Purifying biopharmaceuticals: Knowledge-based chromatographic process development," *Trends Biotechnol.*, vol. 32, no. 4, pp. 210–220, 2014.

[5] D. Keulen, G. Geldhof, O. Le Bussy, M. Pabst, and M. Ottens, "Recent advances to accelerate purification process development: A review with a focus on vaccines," *J. Chromatogr. A*, vol. 1676, p. 463195, Aug. 2022.

[6] E. J. Close, J. R. Salm, D. G. Bracewell, and E. Sorensen, "A model based approach for identifying robust operating conditions for industrial chromatography with process variability," *Chem. Eng. Sci.*, vol. 116, pp. 284–295, 2014.

[7] D. Gétaz, G. Stroehlein, A. Butté, and M. Morbidelli, "Model-based design of peptide chromatographic purification processes," *J. Chromatogr. A*, vol. 1284, pp. 69–79, 2013.

[8] D. G. Bracewell, R. Francis, and C. M. Smales, "The future of host cell protein (HCP) identification during process development and manufacturing linked to a risk-based management for their control," *Biotechnol. Bioeng.*, vol. 112, no. 9, pp. 1727–1737, 2015.

[9] M. J. Traylor, P. Bernhardt, B. S. Tangarone, and J. Varghese, "Analytical Methods," in *Biopharmaceutical Processing*, G. Jagschies, E. Lindskog, K. Lacki, and P. Galliher, Eds. Elsevier, 2018, pp. 1001–1049.

[10] M. R. Schenauer, G. C. Flynn, and A. M. Goetze, "Identification and quantification of host cell protein impurities in biotherapeutics using mass spectrometry," *Anal. Biochem.*, vol. 428, no. 2, pp. 150–157, 2012.

[11] M. Jones *et al.*, "'High-risk' host cell proteins (HCPs): A multi-company collaborative view," *Biotechnol. Bioeng.*, vol. 118, no. 8, pp. 2870–2885, Aug. 2021.

[12] M. Vanderlaan, J. Zhu-Shimoni, S. Lin, F. Gunawan, T. Waerner, and K. E. Van Cott, "Experience with host cell protein impurities in biopharmaceuticals," *Biotechnol. Prog.*, vol. 34, no. 4, pp. 828–837, Jul. 2018.

[13] R. Disela, O. Le Bussy, G. Geldhof, M. Pabst, and M. Ottens, "Characterisation of the E. coli HMS174 and BLR host cell proteome to guide purification process development," *Biotechnol. J.*, no. February, pp. 1–13, Jun. 2023.

[14] K. N. Valente, N. E. Levy, K. H. Lee, and A. M. Lenhoff, "Applications of proteomic methods for CHO host cell protein characterization in biopharmaceutical manufacturing," *Curr. Opin. Biotechnol.*, vol. 53, pp. 144–150, 2018.

[15] G. Jagschies and K. M. Łącki, "Process Capability Requirements," in *Biopharmaceutical Processing*, Elsevier, 2018, pp. 73–94.

[16] C. E. Herman *et al.*, "Analytical characterization of host-cell-protein-rich aggregates in monoclonal antibody solutions," *Biotechnol. Prog.*, vol. 39, no. 4, pp. 1–16, 2023.

[17] C. E. Herman *et al.*, "Behavior of host-cell-protein-rich aggregates in antibody capture and polishing chromatography," *J. Chromatogr. A*, vol. 1702, p. 464081, 2023.

[18] Y. H. Oh *et al.*, "Characterization and implications of host-cell protein aggregates in biopharmaceutical processing," *Biotechnol. Bioeng.*, vol. 120, no. 4, pp. 1068–1080, Apr. 2023.

[19] X. Li, F. Wang, H. Li, D. D. Richardson, and D. J. Roush, "The measurement and control of high-risk host cell proteins for polysorbate degradation in biologics formulation," *Antib. Ther.*, vol. 5, no. 1, pp. 42–54, 2022.

[20] E. K. Lindskog, S. Fischer, T. Wenger, and P. Schulz, "Host Cells," in *Biopharmaceutical Processing*, Elsevier, 2018, pp. 111–130.

[21] J. C. Ranford, A. R. M. Coates, and B. Henderson, "Chaperonins are cell-signalling proteins: The unfolding biology of molecular chaperones," *Expert Rev. Mol. Med.*, vol. 2, no. 8, pp. 1–17, 2000.

[22] B. K. Nfor *et al.*, "Multi-dimensional fractionation and characterization of crude protein mixtures: Toward establishment of a database of protein purification process development parameters," *Biotechnol. Bioeng.*, vol. 109, no. 12, pp. 3070–3083, Dec. 2012.

[23] A. T. Hanke *et al.*, "3D-liquid chromatography as a complex mixture characterization tool for knowledge-based downstream process development," *Biotechnol. Prog.*, vol. 32, no. 5, pp. 1283–1291, 2016.

[24] S. M. Pirrung *et al.*, "Chromatographic parameter determination for complex biological feedstocks," *Biotechnol. Prog.*, vol. 34, no. 4, pp. 1006–1018, 2018.

[25] S. M. Timmick *et al.*, "An impurity characterization based approach for the rapid development of integrated downstream purification processes," *Biotechnol. Bioeng.*, vol. 115, no. 8, pp. 2048–2060, 2018.

[26] N. Vecchiarello *et al.*, "A combined screening and in silico strategy for the rapid design of integrated downstream processes for process and product-related impurity removal," *Biotechnol. Bioeng.*, vol. 116, no. 9, pp. 2178–2190, 2019.

[27] P. S. Wierling, R. Bogumil, E. Knieps-Grünhagen, and J. Hubbuch, "High-throughput screening of packed-bed chromatography coupled with SELDI-TOF MS analysis: monoclonal antibodies versus host cell protein," *Biotechnol. Bioeng.*, vol. 98, no. 2, pp. 440–450, Oct. 2007.

[28] S. Eliuk and A. Makarov, "Evolution of Orbitrap Mass Spectrometry Instrumentation," *Annu. Rev. Anal. Chem.*, vol. 8, pp. 61–80, 2015.

[29] B. K. Nfor, D. S. Zuluaga, P. J. T. Verheijen, P. D. E. M. Verhaert, L. A. M. van der Wielen, and and M. Ottens, "Model-based rational strategy for chromatographic resin selection," *Biotechnol. Prog.*, vol. 27, no. 6, pp. 1629–1643, Nov. 2011.

[30] D. Keulen, E. van der Hagen, G. Geldhof, O. Le Bussy, M. Pabst, and M. Ottens, "Using artificial neural networks to accelerate flowsheet optimization for downstream process development," *Biotechnol. Bioeng.*, no. February, pp. 1–14, May 2023.

[31] G. Guiochon, D. G. Shirazi, A. Felinger, and A. M. Katti, *Fundamentals of Preparative and Nonlinear Chromatography*. Elsevier Science, 2006.

[32] L. Petzold, "Automatic Selection of Methods for Solving Stiff and Nonstiff Systems of Ordinary Differential Equations," *SIAM J. Sci. Stat. Comput.*, vol. 4, no. 1, pp. 136–148, Mar. 1983.

[33] E. S. Parente and D. B. Wetlaufer, “Relationship between isocratic and gradient retention times in the high-performance ion-exchange chromatography of proteins. Theory and experiment,” *J. Chromatogr. A*, vol. 355, no. C, pp. 29–40, 1986.

[34] C. a Brooks and S. M. Cramer, “Steric mass-action ion exchange: Displacement profiles and induced salt gradients,” *AIChE J.*, vol. 38, no. 12, pp. 1969–1978, 1992.

[35] A. A. Shukla, S. S. Bae, J. A. Moore, K. A. Barnthouse, and S. M. Cramer, “Synthesis and characterization of high-affinity, low molecular weight displacers for cation-exchange chromatography,” *Ind. Eng. Chem. Res.*, vol. 37, no. 10, pp. 4090–4098, 1998.

[36] J. R. Wiśniewski, A. Zougman, N. Nagaraj, and M. Mann, “Universal sample preparation method for proteome analysis,” *Nat. Methods*, vol. 6, no. 5, pp. 359–362, 2009.

[37] M. den Ridder, E. Knibbe, W. van den Brandeler, P. Daran-Lapujade, and M. Pabst, “A systematic evaluation of yeast sample preparation protocols for spectral identifications, proteome coverage and post-isolation modifications,” *J. Proteomics*, vol. 261, no. January, p. 104576, 2022.

[38] M. den Ridder, P. Daran-Lapujade, and M. Pabst, “Shot-gun proteomics: Why thousands of unidentified signals matter,” *FEMS Yeast Res.*, vol. 20, no. 1, pp. 1–9, 2020.

[39] C. Chen *et al.*, “Effect of pore structure on protein adsorption mechanism on ion exchange media: A preliminary study using low field nuclear magnetic resonance,” *J. Chromatogr. A*, vol. 1639, p. 461904, 2021.

[40] Y. Yao and A. M. Lenhoff, “Pore size distributions of ion exchangers and relation to protein binding capacity,” *J. Chromatogr. A*, vol. 1126, no. 1–2, pp. 107–119, 2006.

[41] K. D. Ratanji, J. P. Derrick, I. Kimber, R. Thorpe, M. Wadhwa, and R. J. Dearman, “Influence of Escherichia coli chaperone DnaK on protein immunogenicity,” *Immunology*, vol. 150, no. 3, pp. 343–355, 2017.

[42] S. Panikulam *et al.*, “Host cell protein networks as a novel co-elution mechanism during protein A chromatography,” *Biotechnol. Bioeng.*, Mar. 2024.

[43] H. Schmidt-Traub, M. Schulte, and A. Seidel-Morgenstern, *Preparative Chromatography*, Wiley-VCH, 2012.

[44] T. C. Huuk, T. Briskot, T. Hahn, and J. Hubbuch, "A versatile noninvasive method for adsorber quantification in batch and column chromatography based on the ionic capacity," *Biotechnol. Prog.*, 2016.

[45] D. M. Ruthven, Principles of adsorption and adsorption processes, Wiley, 1984.

[46] L. Hagel, "Chapter 3 - Gel filtration: Size exclusion chromatography," in *Protein Purification: Principles, High Resolution Methods, and Applications*, J.-C. Janson, Ed. 2011, pp. 51–92.

4.7 Appendix

4.7.1 Sample preparation for host cell proteomic analysis

Before the mass spectrometry analysis, the samples were prepared using the filter aided sample preparation (FASP) developed to simplify the preparation of samples [36], [37]. 200 µl of the protein samples were loaded onto a Merck-Millipore Microcon 10 kDa filter (Catalog No. MRCPRT010). These proteins were first reduced with the addition of 30 µl of 10 mM DTT and then alkylated using 30 µl of 20 mM iodoacetamide. After alkylation, the proteins underwent a wash with 100 µl of 6 M Urea and three consecutive washes with 100 µl of 200 mM Ammonium bicarbonate (ABC) buffer. Proteolytic digestion was carried out using Trypsin (Promega, Catalog No. V5111) at a 1:100 enzyme-to-protein ratio (v/v) and incubated overnight at 37°C. The peptides resulting from digestion were eluted from the filters using a sequence of ABC and 5% acetonitrile (ACN) / 0.1% formic acid (FA) buffers. Solid-phase extraction was performed employing an Oasis HLB 96-well µElution plate (Waters, Milford, USA, Catalog No. 186001828BA). The elution of peptide fractions was conducted in 2 steps using an 80 % MeOH buffer containing 2 % formic acid (FA) and a 80 % MeOH buffer with 10 mM ABC. The eluates were subsequently dried using a SpeedVac vacuum concentrator.

4.7.2 Methods to determine model parameter

For the development of the mechanistic model, various parameters were obtained experimentally. This included the determination of column

parameters like porosities and system dead volumes using pulse experiments. Furthermore, the ionic capacity was assessed by displacement experiments. Using these parameters, isotherm parameters are regressed from the retention volumes determined in LGE experiments with varying gradient length.

Pulse experiments

250 µl non-binding tracers were used to investigate the dead volumes and porosities in the system and chromatography column. 7.5 g/l dextran 2400K from the American Polymer Standards Corporation was used as a non-penetrating tracer. High salt buffer was used as penetrating tracer. Porosities were determined as described in literature [43].

The dwell volume of the Äkta system, describing the delay between the gradient initiation and the change in the mobile phase composition at the column inlet, was determined in a separate experiment. In this experiment, a pulse of high salt buffer was pumped into the purged system via the system pumps connected to the buffer reservoirs. The volume between the middle of the set pulse and the maximum of the measured conductivity minus the system volume to the conductivity sensor was determined as the system dwell volume (1.1 ml).

Displacement experiments

The ionic capacity of the absorber was measured by displacement experiments using HCl titration. First, the column was washed with 1 M NaOH and MilliQ. Subsequently, 0.05 M HCl was titrated until an increase of the in-line conductivity trace was observed. The HCl volume and the system dwell volumes were used to calculate the ionic capacity for the skeleton of the Q Sepharose XL resin in the column[44]. The ionic capacity was calculated as follows:

$$\Lambda = \frac{\left(V_{tit.HCl} - V_{dwell,system} - V_{dwell,cond} \right) * c_{HCl}}{V_{col} * (1 - \varepsilon_t)}, \tag{4.15}$$

where the titration Volume $V_{tit.HCl}$ is determined as the volume from the start of the titration until the start of the increase in measured conductivity signal. From this, the dwell volume of the system $V_{dwell,system}$ and the dwell

volume of the tubing until the conductivity sensor $V_{dwell,cond}$ are subtracted. The determined ionic capacity Λ for the skeleton of Q Sepharose XL resin was calculated using the HCl concentration, the column volume V_{col} and the total porosity of the resin and was determined to be 1.106 mmol/l.

4.7.3 Mass transfer correlation

The mass transfer is described with

$$k_{ov,i} = \left[\frac{d_p}{6k_{f,i}} + \frac{d_p^2}{60\varepsilon_p D_{p,i}}\right]^{-1}. \tag{4.16}$$

The overall mass transfer coefficient, represented as $k_{ov,i}$, is the composite outcome of both distinct film mass transfer resistance and mass transfer resistance within the pores [45]. Equation (4.16) incorporates parameters such as particle diameter d_p, intraparticle porosity ε_p, and effective pore diffusivity coefficient $D_{p,i}$. The film mass transfer resistance is expressed as $k_{f,i} = D_{f,i}Sh/d_p$, where $D_{f,i}$ describes free diffusivity, and Sh stands for the Sherwood number. Compared to previously mentioned mechanistic models, the empirical correlation from Sofer and Hagel [46] was employed to describe free diffusivity as a function of the molecular weight MW with

$$D_{f,i} = 260 * 10^{-11}\left(MW^{-\frac{1}{3}}\right). \tag{4.17}$$

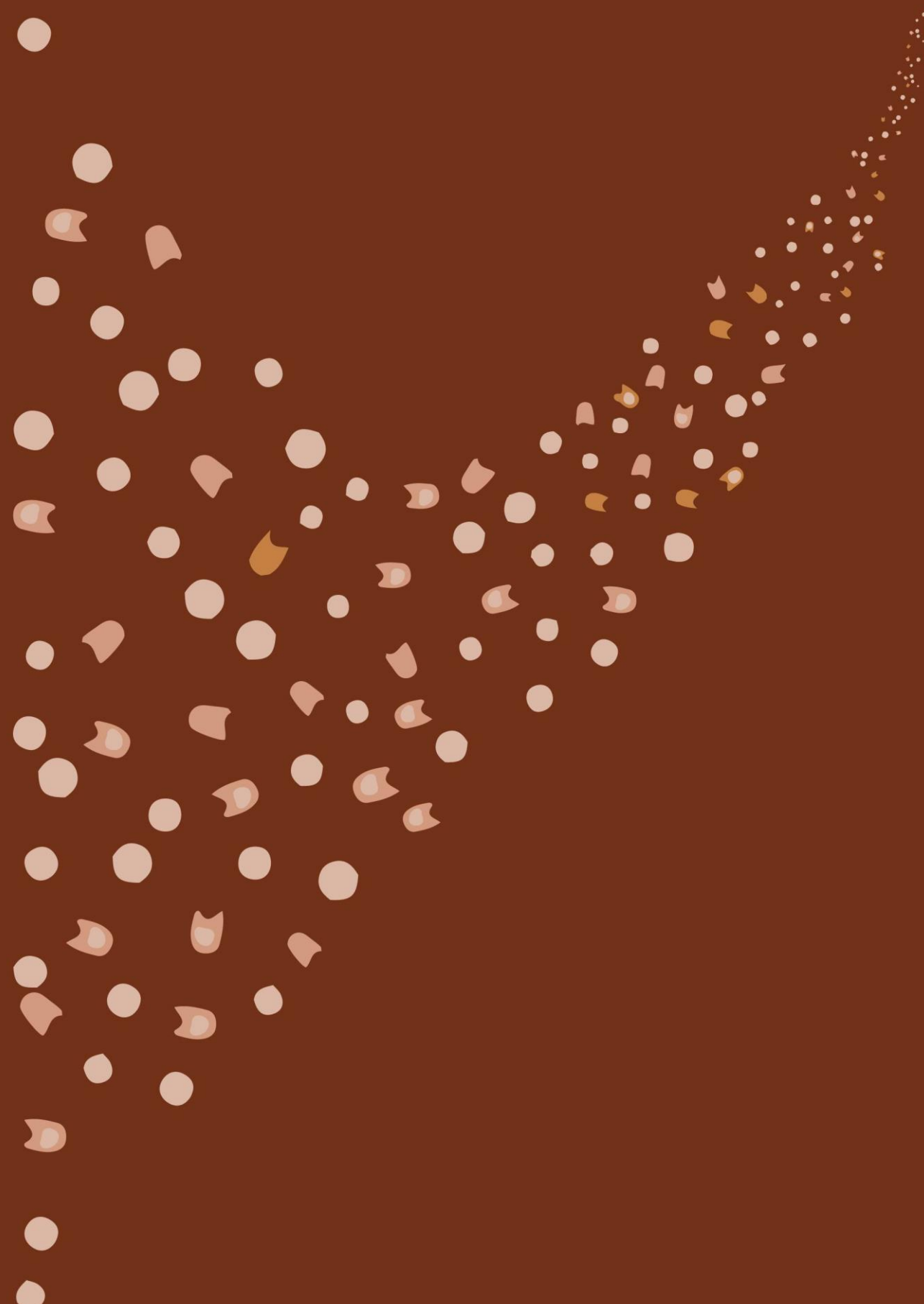
4.7.4 Column characteristics

Table 4.2.: Column characteristics for HiTrap Q XL column (5 ml).

Parameter	Value	Unit
Column volume	5.024	mL
Column diameter[1]	16e-3	m
Bed height[1]	25e-3	m
Ionic capacity (skeleton)	1.106	mmol/L
Particle size[1]	90e-6	m
Pore diameter[2]	54.36e-9	m
Mobile phase volume ($\boldsymbol{V_m}$)	1.50	mL
Total porosity ($\boldsymbol{\varepsilon_t}$)	0.82	-
Extraparticle porosity ($\boldsymbol{\varepsilon_b}$)	0.30	-
System dwell volume ($\boldsymbol{V_{dwell}}$)	1.1	mL
phase ratio (F)	2.35	-

[1]Manufacturer, [2]Reference [39]

4



Chapter 5

Towards High Throughput isotherm determination

A drawback of the conventional shotgun approach for determining isotherm parameters lies in the high time effort. In the chosen approach, the reversed phase chromatographic separation before mass spectrometric detection requires a 60 min separation gradient. However, additional time has to be considered for sample loading, column equilibration, blanks, instrument calibration and maintenance. Running some 100 samples, like measured for the isotherm determination method described in chapter 4, can already take several weeks of measurement time. However, recent developments in mass spectrometric instrumentation and measurement methods enable extremely short gradients (10-20 minutes) while maintaining the same sensitivity and protein identification rate [1]. Nevertheless, this requires dedicated and very expensive instrumentation.
Furthermore, all samples have to be processed before mass spectrometric analysis, including reduction, alkylation, and overnight proteolytic digestion. The employed filter assisted sample preparation (FASP) protocol requires 2 working days in the lab per sample batch handled by a scientist. Additionally, the time for buffer preparations, linear gradient elution (LGE) experiments, processing of mass spectrometric raw data, processing of retention behavior of individual host cell proteins (HCPs) and construction of the isotherm parameter database have to be considered.

In an attempt to reduce the experimental effort, several high throughput methods were investigated and reported in this chapter (Figure 5.1). The use of high throughput methods enables parallelization, automatization and miniaturization, which reduces overall experimental effort. Together with advanced proteomic instrumentation and methods, this could enable screening of larger numbers of chromatographic resins and conditions in a shorter amount of time.

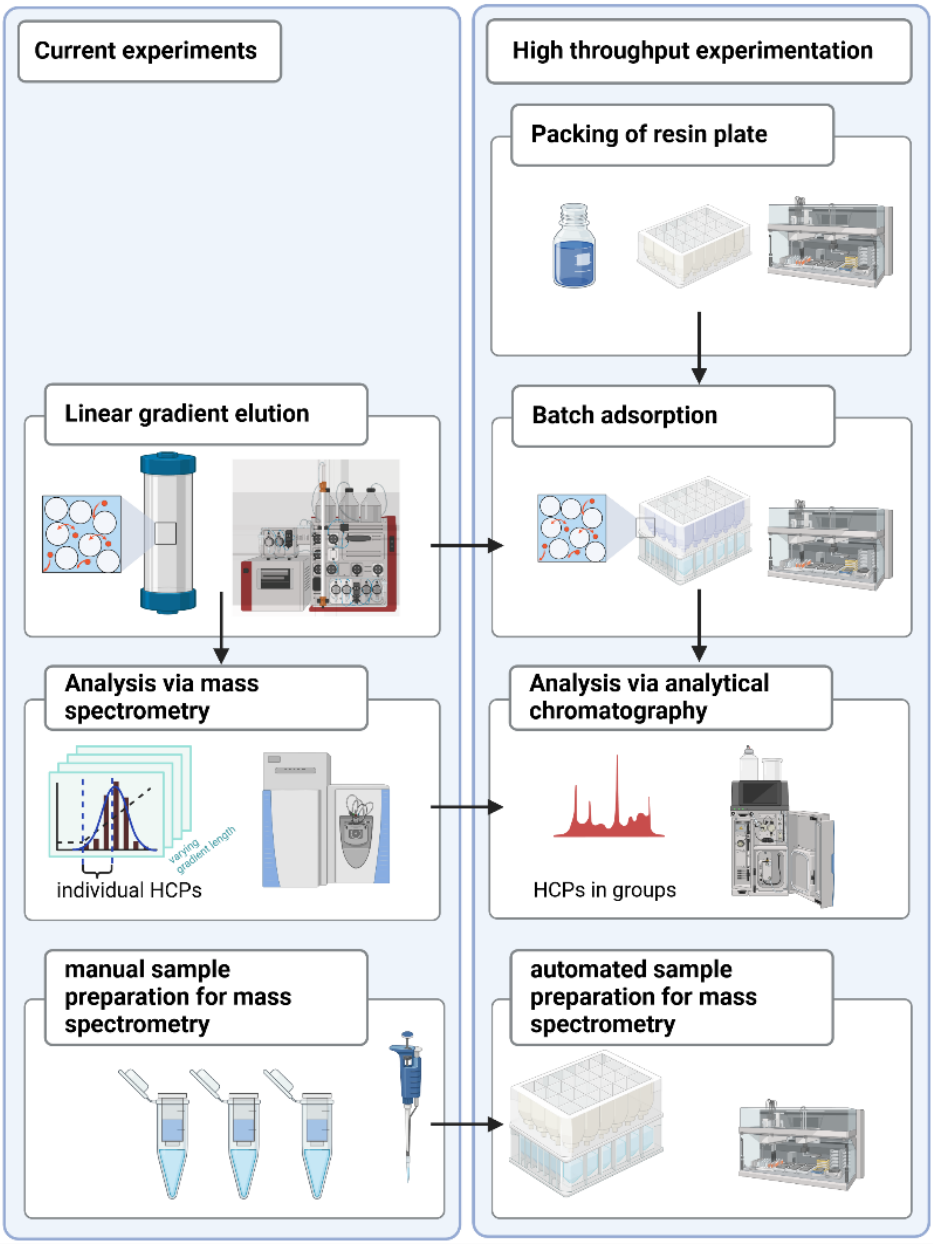


Figure 5.1: Overview of High throughput implementations.

5.1 High throughput screening techniques to determine isotherms

As indicated in the introduction (chapter 1.2.), high throughput screenings can be applied to determine isotherm parameters employing batch adsorption experiments as an alternative to the LGE experiments employed in chapter 4. In principle, this method reduces the experimental effort and used sample volumes. Therefore it was investigated in this chapter.

5.1.1 Packing of resin plate with Tecan

While pre-packed columns are available and used for LGE experiments, few prepacked filter plates are available and these are mostly custom made leading to higher costs and waiting times. Hence, the packing is often conducted in the lab. A significant challenge in implementing high-throughput experimentation (HTE) in batch experiments is to ensure an equal resin volume in every well of the filter plate. The ResiQuot system addresses this by using pressure to fill well-defined volumes with consistent amounts of resin [2]. Although this method can be highly accurate, variations introduced by different users and resins can affect the results. Additionally, the ResiQuot method requires manual operation, making it unsuitable for automation. To overcome these limitations, resin packing using the Tecan liquid handling system can be employed [3].

We implemented our own method in the laboratory. This alternative method involves pipetting resin slurry into the filter plate from a constantly shaken vessel, providing the advantages of automation, such as reduced variability and decreased experimental effort. Comparison of our method with the ResiQuot showed that the ResiQuot system produced individual resin plugs with larger errors due to bubbles or handling issues, which could be identified and excluded from the experiment. In contrast, our Tecan method exhibited errors related to extended pipetting times, resulting in a systematic error where the volume increased over time. In conclusion, while the Tecan system can be used for resin packing, further optimization is required to minimize the systematic error.

5.1.2 Batch adsorption

The use of batch adsorption in HTE presents several advantages, including the reduction of sample and resin volumes. In combination with robotic

liquid dispensing systems, it has shown to be a useful tool to be implemented in the development of new chromatographic steps [4].

Batch adsorption was furthermore implemented in process development to determine isotherm parameter for mechanistic models in past studies [5]–[7]. Additionally, it bears the possibility of automation with liquid handling systems like Tecan, which can integrate a microplate centrifuge. Determining isotherms in this setup is challenging due to the need to measure absolute protein concentrations accurately from small volumes. Especially for HCP mixtures, this requires accurate absolute high-resolution analysis strategies. While mass spectrometry is a high-resolution technique, it is challenging to extract absolute protein concentration values.

5.1.3 Analysis via analytical chromatography

Ultimately, mass spectrometry is a very insightful, but time consuming, analytical method. Therefore analytical ion exchange high-performance liquid chromatography (IEX-HPLC) was investigated as an alternative to supplement mass spectrometry. In the IEX-HPLC strategy, HCPs were investigated in groups according to their retention behavior in the analytical column rather than as individual. A method was developed and implemented using batch adsorption experiments and a IEX-HPLC gradient for the analysis of the host cell proteome. The HCP chromatographic profile was then divided into sub-peaks, which described a group of HCPs, sorted by their charge (Figure 5.2.). Each sub-peak in the analytical chromatogram was assigned a letter and the peak area was compared to the analytical chromatogram at other load concentration relative to a known cytochrome C spiked quantity. The mass balance of the flow through samples was used as an indication to calculate isotherms of the HCP groups.

It was possible to determine linear isotherm parameter of the HCP groups. During the measurement a change of the HCP profile of the load sample over time was observed for IEX-HPLC and SEC-UPLC measurements. This indicated that the charge and size of the sample vary. Degradation or aggregation could occur over time and influence the elution profile accordingly. The harvest sample contains not only proteins, but also other components such as endotoxins, DNA and peptides from the fermentation. It seems that these components have an influence on the retention behavior on analytical IEX or SEC-UPLC of the sample over time. Possible effects could be:

protein-protein interaction; interaction of the proteins with endotoxin or DNA; proteolytic activity; aggregation.

Additionally, the error of the determined isotherms was high and the fit for linear isotherms in some cases had a rather low R^2 indicating that another isotherm model might be more suited. Differences between flow through samples and load samples were hard to recognize in low abundant protein concentrations. In conclusion, the decision was taken not to use this approach in the remainder of the work and investigate the methodology described in chapter 4.

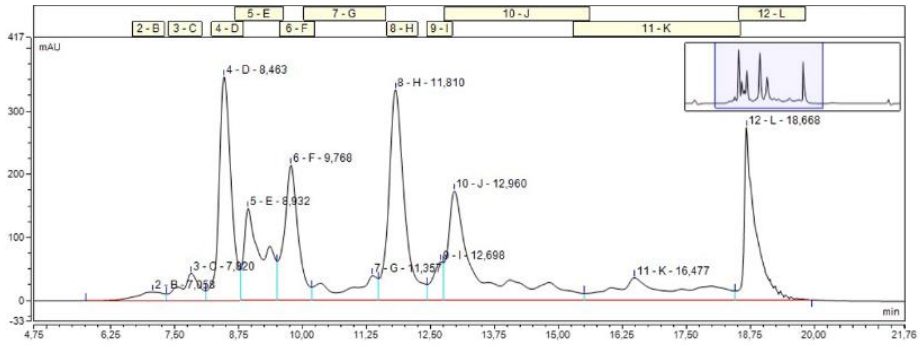


Figure 5.2: Example of analytical IEX-HPLC chromatogram used to determine isotherm parameters of host cell proteins in pseudo groups.

5.2 Automatization of sample preparation for MS

For the preparation of samples before the mass spectrometry measurements the FASP protocol employing 10 kDa MWCO filters was introduced by Wisniewkski et al. [8]. This universal sample preparation method for proteome analysis combines the advantages of in-gel and in-solution digestion for mass spectrometry–based proteomics. The filter unit thereby acts as a 'proteomic reactor' for detergent removal, buffer exchange, chemical modification and protein digestion. This simplifies the handling, which makes it a possible method to implement in high throughput.

The FASP method was implemented on 96-well filter plates [9], [10] and used in an automated settings before [11], [12]. These studies show that it is suited for a higher throughput experimentational set-up. Since the isotherm determination study involved the analysis of around 150 mass spectrometry samples, the automation of such a strategy seems advantageous and should be implemented as a next step to extend the database.

To automatize the full protocol, a Liquid Handling station (e.g. Tecan) is required that entails an integrated centrifuge, shaker and a vacuum station. The centrifuge is used to wash the 10 kDa MWCO filter plate and collect the supernatant. Here the bottleneck lies in the required centrifugal force that has to be reached by the integrated centrifuge. A shaker is needed to mix the solutions with the sample and the vacuum station is needed in the further sample wash. The overall protocol has the potential to be automated. However, the development of this automated protocol requires an intense development effort and attention on the multiple prone-to-error centrifugation steps.

In summary, this chapter explores various high throughput methods. It demonstrates that implementing high throughput techniques requires significant time and effort. Depending on the application of the high throughput pipeline, this investment may need to be carefully considered. In certain cases, implementing these methods proved disadvantageous due to the premature focus on high throughput without prior low-throughput testing. Conversely, the isotherm determination method described in chapter 4 presents a high-resolution analytical solution, highlighting its effectiveness within the context of biopharmaceutical purification process development by covering all detectable individual HCPs. Efforts to automate this method further and shorten measurement times using dedicated instrumentation to implement high throughput mass spectrometry should be taken.

5.3 References

[1] R. Zheng, M. Matzinger, R. L. Mayer, A. Valenta, X. Sun, and K. Mechtler, “A High-Sensitivity Low-Nanoflow LC-MS Configuration for High-Throughput Sample-Limited Proteomics,” *Anal. Chem.*, vol. 95, no. 51, pp. 18673–18678, Dec. 2023.

[2] T. Herrmann, M. Schröder, and J. Hubbuch, “Generation of equally sized particle plaques using solid-liquid suspensions,” *Biotechnol. Prog.*, vol. 22, no. 3, pp. 914–918, 2006.

[3] X. Li, G. de Roo, K. Burgers, M. Ottens, and M. Eppink, “Self-packed filter plates: A good alternative for pre-packed filter plates for developing purification processes for therapeutic proteins,” *Biotechnol. J.*, vol. 7, no. 10, pp. 1269–1276, Oct. 2012.

[4] J. L. Coffman, J. F. Kramarczyk, and B. D. Kelley, “High-throughput

screening of chromatographic separations: I. method development and column modeling," *Biotechnol. Bioeng.*, vol. 100, no. 4, pp. 605–618, 2008.

[5] M. Moreno-González, P. Chuekitkumchorn, M. Silva, R. Groenewoud, and M. Ottens, "High throughput process development for the purification of rapeseed proteins napin and cruciferin by ion exchange chromatography," *Food Bioprod. Process.*, vol. 125, pp. 228–241, 2021.

[6] B. K. Nfor, M. Noverraz, S. Chilamkurthi, P. D. E. M. Verhaert, L. A. M. Van Der Wielen, and M. Ottens, "High-throughput isotherm determination and thermodynamic modeling of protein adsorption on mixed mode adsorbents," *J. Chromatogr. A*, vol. 1217, no. 44, pp. 6829–6850, 2010.

[7] T. C. Silva, M. Eppink, and M. Ottens, "Digital twin in high throughput chromatographic process development for monoclonal antibodies," *J. Chromatogr. A*, vol. 1717, no. September 2023, p. 464672, 2024.

[8] J. R. Wiśniewski, A. Zougman, N. Nagaraj, and M. Mann, "Universal sample preparation method for proteome analysis," *Nat. Methods*, vol. 6, no. 5, pp. 359–362, 2009.

[9] L. Switzar, J. van Angeren, M. Pinkse, J. Kool, and W. M. A. Niessen, "A high-throughput sample preparation method for cellular proteomics using 96-well filter plates," *Proteomics*, vol. 13, no. 20, pp. 2980–2983, Oct. 2013.

[10] Y. Yu, M.-J. Suh, P. Sikorski, K. Kwon, K. E. Nelson, and R. Pieper, "Urine Sample Preparation in 96-Well Filter Plates for Quantitative Clinical Proteomics," *Anal. Chem.*, vol. 86, no. 11, pp. 5470–5477, Jun. 2014.

[11] J. Jeon, J. Yang, J.-M. Park, N.-Y. Han, Y.-B. Lee, and H. Lee, "Development of an automated high-throughput sample preparation protocol for LC-MS/MS analysis of glycated peptides," *J. Chromatogr. B*, vol. 1092, no. January, pp. 88–94, Aug. 2018.

[12] A.-B. Arul, M. Byambadorj, N.-Y. Han, J. M. Park, and H. Lee, "Development of an Automated, High-throughput Sample Preparation Protocol for Proteomics Analysis," *Bull. Korean Chem. Soc.*, vol. 36, no. 7, pp. 1791–1798, Jul. 2015.

Chapter 6

Conclusions and Outlook

This chapter summarizes the key findings of this thesis and offers insights into potential future research directions. The primary objectives of the work performed in this thesis were to characterize the *E. coli* host cell proteome and determine isotherm parameters for all detectable host cell proteins (HCPs). To achieve this, a comprehensive database was created, retention times of individual HCPs were measured and used to build a quantitative structure-property relationship (QSPR) model. Furthermore, a method to determine isotherms of the host cell proteome was established and its use to optimize a purification step *in silico* demonstrated.

Chapter 2:

- An extensive database of HCPs and their physicochemical properties was created.

- HCP profile findings were transferable between different *E. coli* strains (BLR & HMS174) and between BLR expressing the antigen and a null plasmid strain. This is particularly valuable for isotherm determination methods, as the HCP profile changes more in abundance than in identity, especially among low-abundance proteins.

- Protein property maps were generated, utilizing physicochemical properties to guide a suitable downstream processing strategy.

Chapter 3:

- An experimental retention map of the host cell proteome during a linear gradient elution (LGE) on hydrophobic interaction chromatography (HIC) and ion exchange chromatography (IEX) was developed.

- Patterns in retention behavior were identified according to cell location, molecular function, and protein-protein interactions (PPIs).

- A predictive QSPR model was built using IEX retention data relating protein sequence to elution behavior.

- Additional QSPR models were developed based on identified patterns, with monomers yielding the most accurate model.

Chapter 4:

- A method to determine the isotherm parameters of individual HCPs was provided, overcoming a significant hurdle in enabling the prediction of elution behavior of HCPs in mechanistic models.

- This method was exemplified using the BLR *E. coli* lysate, validated, and used to optimize a capture step *in silico*.

- The comprehensive database of HCP isotherm parameters captured the interactions of proteins in the harvest mixture, with jointly eluting proteins reflected by similar parameters.

Chapter 5:

- An overview of additionally investigated high-throughput methods to determine isotherm parameters was provided.

- It demonstrates that implementing high throughput techniques requires significant time and effort. Hence the development of a low-throughput method should be established before implementing HT techniques.

This research has advanced the field of biopharmaceutical purification through several pivotal contributions. It has characterized *E. coli* HCP impurities, providing a detailed database of these impurities and their physicochemical properties. By determining the retention behavior of HCPs and identifying patterns in their retention, the thesis has enhanced the understanding of how these impurities interact with chromatographic resins. Furthermore, a method to determine isotherm parameters for HCPs has been established, which is crucial for the development of more accurate and effective purification processes employing mechanistic models.

The isotherm parameter database can be extended with experimentally determined isotherm parameters to include orthogonal chromatography steps, such as HIC, cation exchange chromatography (CEX). The same resins

could also be investigated under varying mobile phase conditions. Although the mass spectrometry method used is time and cost-intensive, it only needs to be applied once per resin, and condition. For the investigated *E. coli* strains at least, findings are transferable between strains. In principle, this approach could facilitate the creation of a database containing information about impurities from various hosts, thereby accelerating the development of biopharmaceuticals.

One of the major advancements is the progress towards complete *in silico* process development. With this research, we are closer to a future where only molecular structures of proteins are needed as an input; their retention behavior and isotherm parameters can be predicted by a QSPR model trained with a host cell proteome database. The prediction of isotherm parameters from retention data was demonstrated in a study using standard proteins [1]. These isotherm parameters can then be utilized in mechanistic models to optimize the chromatography step. This capability could be used to predict unknown proteins, such as new antigens, HCPs under the detection limit, or HCPs from other hosts, provided their 3D structure is known or can be modeled using tools like AlphaFold.

Apart from the optimization of individual chromatography steps, a database with isotherm parameters of HCP impurities including orthogonal chromatography steps, can be used for flowsheet optimization. Flowsheet optimization is the most effective tool for identifying the optimal process sequence in the earliest stages of process development and can be applied to chromatographic purification sequences [2].

The thesis opened up new avenues, but also encountered new challenges:

1. The determination of the developed isotherm parameters is currently limited by the high effort required for mass spectrometric measurements.
2. PPIs in the lysate sample need further investigation. Additionally challenging here is the influence of the mobile phase on the formation of PPIs.
3. Other impurities than HCPs such as endotoxins, DNA, and RNA are not accounted for. Their removal should also be investigated. Interactions of these impurities with proteins were not considered, limiting the accuracy of the QSPR model.

4. The interactions between the target molecule and HCPs, such as observed in monoclonal antibody (mAb) aggregates [3]–[5], present additional challenges. Comprehensive mapping and understanding of these interactions are essential for developing more accurate predictive models and improving the overall purification process.

To address these limitations and to further advance the field, several areas for future research are recommended:

1. Recent developments in mass spectrometric instrumentation and measurement methods enable extremely short gradients (10-20 minutes) by maintaining the same sensitivity and protein identification rate [6]. However, these advancements require dedicated and expensive instrumentation. Efforts should be made to shorten measurement times by implementing high-throughput MS using such dedicated instruments. Additionally, enhancing data management pipelines and utilizing extended databases for whole-process flowsheet optimization is crucial. Establishing robust data management and processing systems will ensure efficient handling and meaningful analysis of extensive datasets. The potential of big data lakes for data mining and machine learning is significant, but it is essential to determine the optimal amount of data needed to answer specific questions without becoming overwhelmed by excessive data.
2. Future research should focus on modeling PPIs in the lysate comparable to computational approaches used to predicts PPIs within cells [7]. AlphaFold-Multimer could be employed to predict protein complex structures [8]. These artificial intelligence-assisted structural proteomics on the other hand could be employed in the QSPR model.
3. Future research should also focus on characterizing and determining isotherm parameters or similar for other impurities such as endotoxins, DNA, and RNA. Since chromatography is often used for endotoxin removal [9], it might make sense to investigate the persistence throughout chromatography in general using other analytical techniques such as e.g. the LAL-test for endotoxin.
4. When applying this method to mAbs, a joint measurement including the mAb is recommended. Isotherm parameters of present aggregates could be determined and treated as another impurity by

the mechanistic model. Another approach would be to target co-eluting HCPs involved in PPIs [10] and use their isotherms in the mechanistic model to remove these early on in the process. Such advancements will contribute to a more comprehensive understanding of the interactions and behaviors of various impurities, leading to improved predictive capabilities and process optimization.

Looking ahead, future trends in this research area may include the widespread adoption of comprehensive databases for HCPs and use for *in silico* strategies. Using an extended database for flowsheet optimization promises to streamline the entire process further. The broader implications of this research include faster and more cost-effective development of biopharmaceuticals, benefiting society and advancing technological and academic knowledge.

The knowledge about re-occurring impurities that co-elute with the product could furthermore be combined with the knowledge of the functional proteome of *E. coli* [11] with the purpose of novel strain development. Non-functional and difficult-to-remove HCPs can be knocked out as demonstrated for CHO [12], and *E. coli* [13].

In summary, this research has significantly enhanced our understanding of HCP impurities. The comprehensive characterization and determination of parameters have enabled the development of new predictive models for purification. Consequently, the findings presented and the methodologies developed contribute to a more efficient and accurate process development. This work as part of the existing collaboration with GSK lays the groundwork for future innovations in biopharmaceutical purification, promising substantial improvements in the field.

References

[1] D. Keulen *et al.*, "From protein structure to an optimized chromatographic capture step using multiscale modeling," *Biotechnol. Prog.*, e3505, 2024.

[2] D. Keulen, E. van der Hagen, G. Geldhof, O. Le Bussy, M. Pabst, and M. Ottens, "Using artificial neural networks to accelerate flowsheet optimization for downstream process development," *Biotechnol. Bioeng.*, no. February, pp. 1–14, May 2023.

[3] C. E. Herman *et al.*, “Behavior of host-cell-protein-rich aggregates in antibody capture and polishing chromatography,” *J. Chromatogr. A*, vol. 1702, p. 464081, 2023.

[4] C. E. Herman *et al.*, “Analytical characterization of host-cell-protein-rich aggregates in monoclonal antibody solutions,” *Biotechnol. Prog.*, vol. 39, no. 4, pp. 1–16, 2023.

[5] Y. H. Oh *et al.*, “Characterization and implications of host-cell protein aggregates in biopharmaceutical processing,” *Biotechnol. Bioeng.*, vol. 120, no. 4, pp. 1068–1080, Apr. 2023.

[6] R. Zheng, M. Matzinger, R. L. Mayer, A. Valenta, X. Sun, and K. Mechtler, “A High-Sensitivity Low-Nanoflow LC-MS Configuration for High-Throughput Sample-Limited Proteomics,” *Anal. Chem.*, vol. 95, no. 51, pp. 18673–18678, Dec. 2023.

[7] G. Grassmann *et al.*, “Computational Approaches to Predict Protein–Protein Interactions in Crowded Cellular Environments,” *Chem. Rev.*, vol. 124, no. 7, pp. 3932–3977, Apr. 2024.

[8] F. J. O’Reilly *et al.*, “ Protein complexes in cells by AI -assisted structural proteomics ,” *Mol. Syst. Biol.*, vol. 19, no. 4, pp. 1–20, 2023.

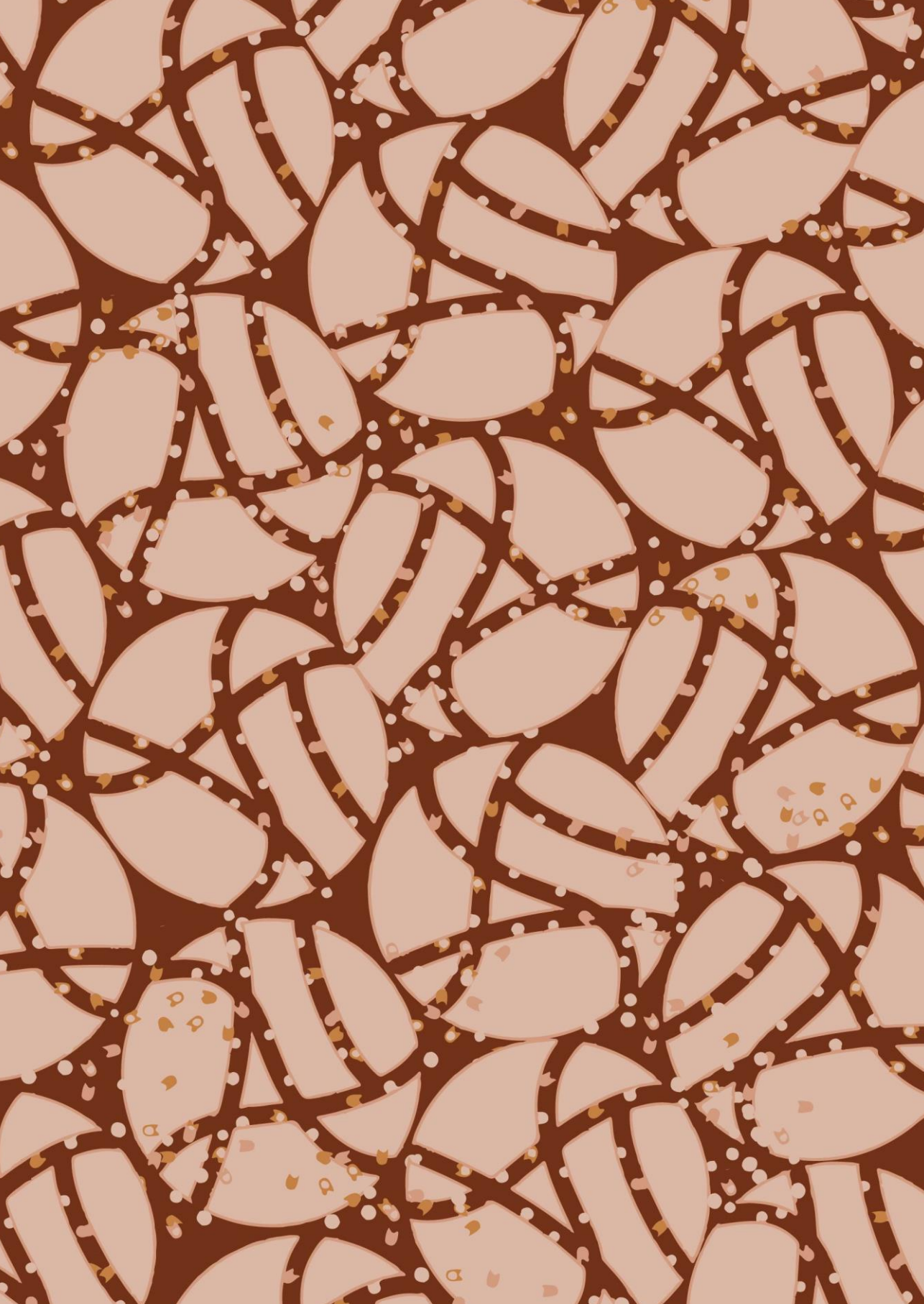
[9] C. M. Ongkudon, J. H. Chew, B. Liu, and M. K. Danquah, “Chromatographic Removal of Endotoxins: A Bioprocess Engineer’s Perspective,” *ISRN Chromatogr.*, vol. 2012, no. Figure 1, pp. 1–9, 2012.

[10] S. Panikulam *et al.*, “Host cell protein networks as a novel co-elution mechanism during protein A chromatography,” *Biotechnol. Bioeng.*, Mar. 2024.

[11] A. Mateus *et al.*, “The functional proteome landscape of Escherichia coli,” *Nature*, vol. 588, no. 7838, pp. 473–478, Dec. 2020.

[12] J. Chiu, K. N. Valente, N. E. Levy, L. Min, A. M. Lenhoff, and K. H. Lee, “Knockout of a difficult-to-remove CHO host cell protein, lipoprotein lipase, for improved polysorbate stability in monoclonal antibody formulations,” *Biotechnol. Bioeng.*, vol. 114, no. 5, pp. 1006–1015, May 2017.

[13] M. H. Caparon *et al.*, “Integrated solution to purification challenges in the manufacture of a soluble recombinant protein in E. coli,” *Biotechnol. Bioeng.*, vol. 105, no. 2, pp. 239–249, Feb. 2010.



Acknowledgments

I am deeply grateful to **Marcel** for teaching me what it means to be a scientist. It has been a joy to be part of your group, where I learned not only how to manage scientific projects but also how to approach challenges with curiosity and confidence. Thank you for encouraging me to attend numerous conferences, which enriched my knowledge of the field and helped shape me into a better scientist. I also greatly appreciated the time you took to navigate through challenges in the project like our dedicated brainstorming sessions.

A heartfelt thank you to **Martin** for your supervision and guidance. I thoroughly enjoyed our critical discussions. Your insights and expertise in analytical work were indispensable. I am grateful for the time you took to demonstrate the sample preparation protocols in the MS lab and assist with experiments and troubleshooting. Your efforts in processing MS data and enabling the high-throughput runs in your lab made this work possible.

To **Geoffroy and Olivier**, thank you for welcoming us to GSK and offering valuable insights into the biopharmaceutical industry. I greatly appreciated our engaging discussions, and your challenging questions throughout this collaboration.

Daphne and Tim: I am happy to be able to work in a team with you two in our project. I am so grateful as well to have you as officemates and truly enjoyed our coffee break discussions. From trips to the Portugal, Spain, US, Berlin, and Belgium to our brainstorming meetings, it has been an absolute pleasure working with you both. **Daphne**, I cherished our times at conferences, working on the mechanistic model software, visiting GSK, supervising students together, and celebrating milestones. We truly saw each other grow in the project. **Tim**, I greatly valued our collaboration on the retention prediction paper and the countless spontaneous discussions in the office. I learned so many details about coffee and cooking from you. Thank you also, Tim, for standing by me as my paranymph.

I want to thank all students involved in the project: **Floor, Shawn, Inés, Anne-Marijn, Eleni**. I am thankful for the contributions you made and honored to be your supervisor. Floor, thanks for your enthusiasm in developing a new resin packing strategy with the Tecan. Shawn, thank you for implementing

the first method to determine HCP isotherms in groups. Inés, thank you for your work on the determination of antigen isotherms. Anne-Marijn, thank you for your work determining isotherms and modeling membrane adsorbers. I regret that this work did not end up in the PhD thesis. Thank you, Eleni for developing a workflow and initial method to determine isotherm parameter for individual HCPs from MS experiments.

To all **members of BPE**, thank you for fostering a collaborative and supportive working atmosphere, enriched by coffee breaks and cake-filled gatherings. I hope that this friendly atmosphere will live on forever.

Special thanks go to the PhDs that started around the same time as me and with who I shared the highs and lows of this journey—from a flooded lab to navigating a pandemic. **Oriol and Marijn,** thank you for your support outside the lab and our fun evenings together. **Tiago**, I deeply appreciated our spirited debates on science and society, which challenged and shaped my perspectives. **Mariana**, thank you for your open-door kindness during work and personal conversations. I enjoyed working alongside you in the lab. **Joan**, I enjoyed the short overlap we had in the lab and discussions on the bike ride home. **Marina**, I'll always treasure our late bike rides, ice skating outings, and stargazing adventures. **Lars**, your enthusiasm for all kind of activities (ice skating, Oktoberfest, star gazing or just Fridy borrels) and friendship brightened this journey, and I am grateful for your support (also with the propositions). **Maarten**, it was great traveling together in the US.

To the earlier members of BPE, **Monica, Debbie, Joana, Chema, Rita, and Bianca**, thank you for your warm welcome in the section and advice along the way. **Monica**, our early discussions of the projects were very helpful, and our online coffee breaks during Covid were always refreshing. **Debbie**, thanks for our discussions early on and advice.

Thanks to the PostDocs,: **Tim Nijjsen, Mona, Zulhaj and Eduardo! Tim Nijssen**, thank you for your motivational presence and thoughtful advice, which ranged from scientific challenges to life's curveballs. **Mona**, I am so happy, that we met . Thanks for the private activities and discussions.

Thanks to all the new generations of PhDs for keeping up the BPE spirit, being curious and organizing more group building events. **Hector, Rik, Tamara, Marika, Brenda, Mariana (Carvalhi), Ramon, Meryl, Miki, Dimitri, Ben, Jelle, Pieter**!

To the BPE staff: **Adrie, Ludo, Cees, Marieke, Christiaan, Stef, and Max. Christiaan, Kawieta,, Jeroen, Josh**. Thank you for taking care of all of us! **Kawieta**, your prompt responses and organizational support made managing student projects, business trips, and contracts so much easier. **Adrie**, I appreciated our (online) coffee breaks back in 2020. **Marieke**, thank you for promoting this PhD project in Karlsruhe, which led me to this incredible journey (for example the conference in San Francisco together). To the **technicians** I express a big thank you. Your expertise and support were essential for this project. **Song**, your patience and guidance with the Tecan and other machines were invaluable. I learned a lot about the green lab from you! **Max Zomerdijk**, thank you for making teaching enjoyable and for being a source of practical advice (in the same office). **Stef**, I appreciated your perseverance in troubleshooting experiments with me last minute and that you are always up for a coffee break. **Christiaan**, your enthusiasm for the Tecan arrived at just the right moment and helped me move forward.

To my **friends**, thank you for believing in me and providing support, whether in the Netherlands or Germany.

To **my parents**, your unwavering support throughout my studies and this PhD means the world to me. **Nessi**, you are the best sister and my greatest cheerleader/fan. Your contributions, including designing the cover art and being a paranymph, were invaluable.

Finally, to **Max**, my partner/fiancée, thank you for your steadfast encouragement and support throughout this journey. You were always there to listen, and celebrate my successes.

List of publications

Journal articles

Disela, R., Neijenhuis, T., Le Bussy, O., Geldhof, G., Klijn, M., Pabst, M., Ottens, M., submitted as "Experimental characterization and prediction of Escherichia coli host cell proteome retention during preparative chromatography", *Biotechnology and Bioengineering*, 2024, doi: 10.1002/bit.28840

Keulen, D., Neijenhuis, T., Lazopoulou, A., **Disela, R.**, Geldhof, G., Le Bussy, O., Klijn, M. E., Ottens, M., "From protein structure to an optimized chromatographic capture step using multiscale modeling", *Biotechnology Progress*, 2024, doi: 10.1002/btpr.3505

Disela, R., Keulen, D., Fotou, E., Neijenhuis, T., Le Bussy, O., Geldhof, G., Pabst, M., Ottens, M., "Proteomics-based method to comprehensively model the removal of host cell protein impurities", *Biotechnology Progress*, 2024, doi: 10.1002/btpr.3494

Disela, R., Le Bussy, O., Geldhof, G., Pabst, M., Ottens, M., "Characterisation of the E. coli HMS174 and BLR host cell proteome to guide purification process development", *Biotechnology Journal*, (2023), 1-13. doi: 10.1002/biot.202300068

Oral presentations

Disela, R., Keulen, D., Le Bussy, O., Geldhof, G., Pabst, M., Ottens, M., (2023) "Proteomic analysis of the retention behavior of host cell protein impurities in the purification of vaccines", 14th European Congress of Chemical Engineering and 7th European Congress of Applied Biotechnology, Berlin, Germany

Disela, R., Keulen, D., Fotou, E., Le Bussy, O., Geldhof, G., Pabst, M., Ottens, M., (2023) "Novel method to extract isotherm parameters of the host cell proteome in the purification of biopharmaceuticals", American Chemical Society Fall meeting, San Francisco, CA, United States of America

Disela, R., Keulen, D., Le Bussy, O., Geldhof, G., Pabst, M., Ottens, M., (2022) "Acceleration of vaccine development by increasing process understanding – Comprehensive analysis of the *E.coli* host cell proteome", Biopartitioning & Purification conference, Aveiro, Portugal (speaker M. Ottens)

Disela, R., Keulen, D., Geldhof, G., Le Bussy, O., Pabst, M., Ottens, M. (2022), "Acceleration of vaccine development by improvement of process understanding - Analysis of the host cell proteome", 17th international seminar on chromatographic separation science, Karlsruhe, Germany

Disela, R., Keulen, D., Geldhof, G., Le Bussy, O., Pabst, M., Ottens, M. (2021), "Acceleration of subunit vaccine development – Analyzing the E.coli Host cell proteins in the harvest", 16th international seminar on chromatographic separation science, Vienna, Austria

Disela, R., Le Bussy, O., Geldhof, G., Pabst, M., Ottens, M., (2021) "A novel strategy to determine chromatographic parameters of host cell proteins to accelerate vaccine process development", 13th European Congress of Chemical Engineering and 6th European Congress of Applied Biotechnology, online

Poster presentations

Disela, R., Keulen, D., Geldhof, G., Le Bussy, O., Pabst, M., Ottens, M. (2022), "Acceleration of vaccine development by improvement of process understanding - Analysis of the host cell proteome", Vaccine Technology Conference VIII, Sitges, Spain

Disela, R., Keulen, D., Geldhof, G., Le Bussy, O., Pabst, M., Ottens, M. (2022), "Model-based High Throughput Process Development for vaccine purification – Experimental implementations", 17th edition of the Netherlands Process technology Symposium, Delft, The Netherlands

Disela, R., Keulen, D., Geldhof, G., Le Bussy, O., Pabst, M., Ottens, M. (2021), "Model-based High Throughput Process Development for vaccine purification – Experimental implementations", 13th European Symposium on Biochemical Engineering Sciences, online

Curriculum vitae

Roxana Clarissa Disela was born on the 25th of August 1994 in Basel, Switzerland and grew up in Lörrach, Germany. In 2012, she started studying Bioengineering/Life Science engineering at the Karlsruhe Institute of Technology. In her Bachelor thesis, she worked on "*High throughput calibration of adsorption isotherms and transfer to a chromatography column model*" in the Biomolecular Separation Engineering group of Prof. Hubbuch, where her interest for robotic handling systems, mechanistic modeling and protein purification was sparked.

Afterwards, she went to Melbourne, Australia for an 6-month internship at the Commonwealth Scientific and Industrial Research Organisation (CSIRO) in the food and nutrition department. During her master she conduct another 6-month internship in the pharmaceutical industry at Roche in Penzberg, Germany where she deepened her expertise in high throughput screenings of (mixed mode) chromatographic media. For her master thesis she joined the Institute of Functional Surfaces in Karlsruhe and investigated "*Production of smart-actuating polymer fibers by electrojetting for the use in 4D micro-objects*" under supervision of Prof. Lahann and Prof. Franzreb.

To pursue her PhD, she became part of the bioprocess engineering group at TU Delft under the supervision of Marcel Ottens and Martin Pabst in September 2019. Her work on "Chromatographic host cell protein removal in biopharmaceutical purification" is described in this thesis.

In November 2024 she started working as a Mechanistic Modeling specialist at Cytiva in Uppsala, Sweden.