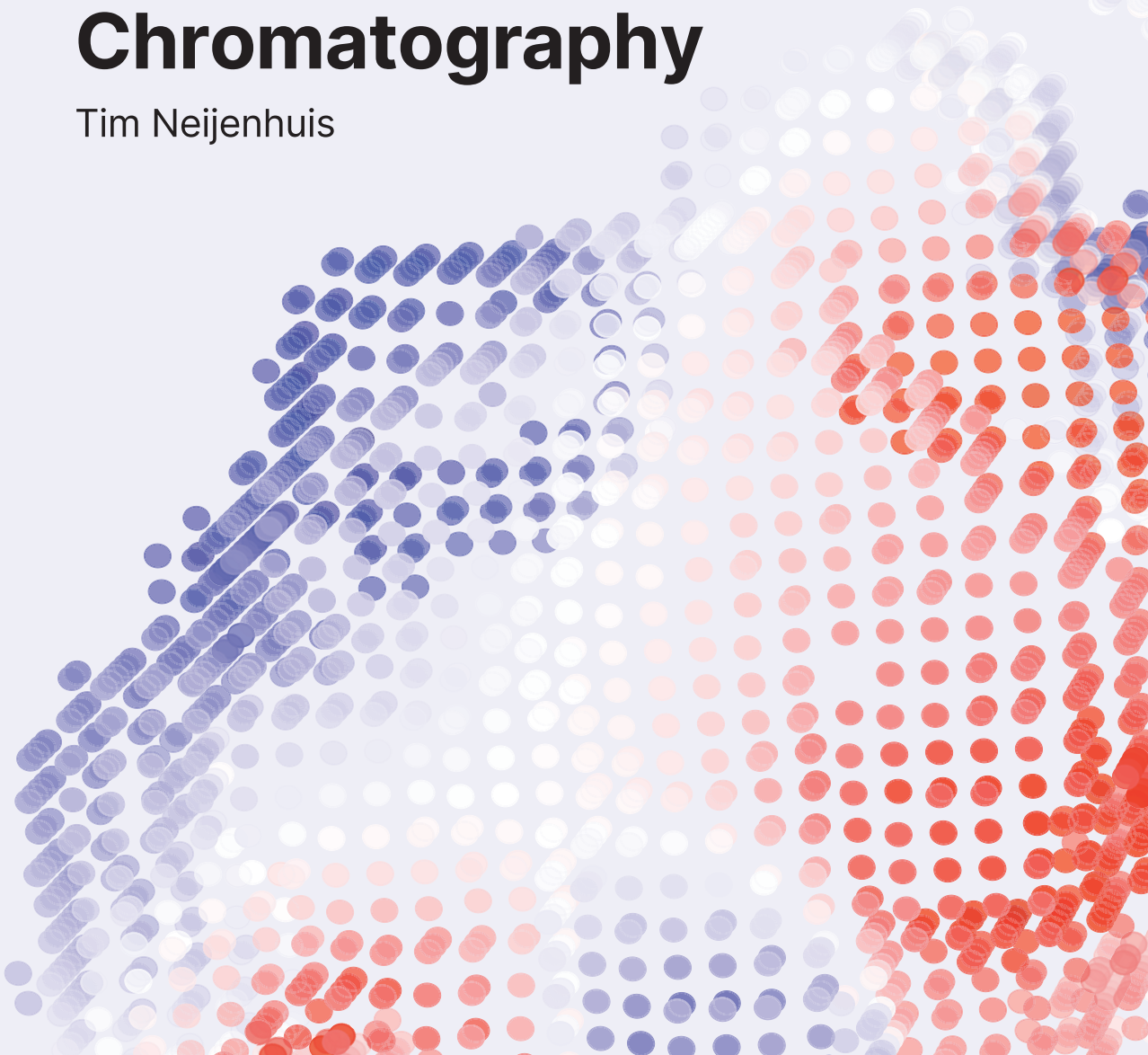
Structure-Based Prediction of Protein Behavior in Preparative Chromatography

Tim Neijenhuis



Structure-Based Prediction of Protein Behavior in Preparative Chromatography

Structure-Based Prediction of Protein Behavior in Preparative Chromatography

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. H. Bijl,
chair of the Board of Doctorates
to be defended publicly on
Friday 16 Januari 2026 at 12:30 o'clock

By

Tim NEIJENHUIS

Master of Science in Molecular Life Sciences
Radboud University, The Netherlands
born in Wageningen, The Netherlands

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus	Delft University of Technology
Prof.dr.ir. M. Ottens	Delft University of Technology, promotor
Dr.ir. M. E. Klijn	Delft University of Technology, copromotor

Independent members:

Prof.dr. F. Hollmann	Delft University of Technology
Prof.dr.ing. M.H.M. Eppink	Delft University of Technology
Prof.dr.ir. L.A.M. van der Wielen	Delft University of Technology / DTU, Denmark
Prof.dr. J. Buyel	Universität für Bodenkultur Wien, Austria
Dr. A. Azevedo	Instituto Superior Técnico, Portugal

This work was partly financed from PSS-allowance for Top consortiums for Knowledge and Innovation (TKI) of the ministry of Economic Affairs and partly sponsored by GlaxoSmithKline Biologicals S.A. under cooperative research and development agreement between Glaxo- SmithKline Biologicals S.A. (Belgium) and the Technical University of Delft (The Netherlands).

Printed by ZieZo Grafische Oplossingen

Cover design by Julia Gerritsen

ISBN: 978-94-6518-196-7

An electronic version of this dissertation is available at
https://repository.tudelft.nl

Table of contents

Summary

Vaccination plays a pivotal role in modern preventive healthcare and contributes to a global decline in infectious diseases. Efficient production of vaccines is essential to meet the growing demand which results from factors like a growing global population and increased international travel. Protein subunit vaccines are a vaccine modality that contains parts of the infectious pathogen as the active ingredients. These subunits are recognized by the immune system, which is trained to respond more effectively and reduce symptoms upon actual infection. Production of these vaccines is divided into upstream processing (USP), which involves fermentation using expression hosts, downstream processing (DSP) where the protein subunit is purified, and finally formulation where the vaccines are prepared for distribution. During the DSP, multiple chromatography modes are often used to reach the required purity. Selection of the optimal chromatographic resin types, as well as operating conditions can be expensive and time consuming. Model-based process development has the potential to speed up this selection by using computational methods to predict protein behavior. Especially in early phase development, models allow in silico screening of resins and conditions in tandem to classical experiments, reducing required material. These computational models can be divided into knowledge-driven, data-driven, or a combination thereof.

The focus of this thesis is the development of a data-driven modeling approach where protein behavior is predicted from its atomic structure. Specifically, quantitative structure property relationship (QSPR) models are used for this purpose. To achieve this, **chapter 2** introduces a Python tool that is developed to extract relevant information from the three-dimensional protein structure. This is done by sampling the

protein surface using grid representation that describes the distribution of different physicochemical properties. These are translated into numerical descriptors. Using literature data, descriptor relevance was shown by training two separate QSPR models for the prediction of retention times in ion exchange chromatography (IEX), resulting in cross validated R^2 of 0.87 and 0.95.

A limitation of the data-driven modeling approach is that these models are only valid for the experimental conditions for which they are trained. Knowledge-driven models use fundamental knowledge about, for example, mass transfer. In chromatography, adsorption isotherms are essential to describe the binding of a protein to a chromatographic resin. In **chapter 3** we developed a multiscale modeling approach by integrating QSPR with mechanistic modeling. Adsorption isotherm parameters predicted by QSPR were used in a mechanistic model. This multiscale model was validated with experimental data and showed only 0.2% difference between the retention peak values, relative to the salt gradient length. Subsequently, the validated mechanistic model was used to optimize a chromatographic capture step.

Commercially available model proteins provide a great basis for a proof of principle, however QSPR modeling becomes more powerful when applied to host cell proteins (HCPs). Therefore, we characterized the chromatographic behavior of the HCPs present in an *Escherichia coli* (*E. coli*) lysate in **chapter 4** by means of fractionation and subsequent analysis by mass spectrometry. Retention times of 816 and 908 HCPs were collected for hydrophobic interaction chromatography (HIC) and IEX, respectively. By dividing the HCPs into subsets based on cellular location, function, and interactions, basic trends were visualized. Next, we predicted the structures of each individual HCP which were used to train QSPR models. This was successful for IEX data resulting in a QSPR model with a cross validated R^2 of 0.70 when using the monomer HCP subset.

Obtaining high resolution HCP retention data still requires substantial experimental effort. Therefore, deployment of QSPR models for PD would benefit from the formulation of a list of widely available (commercial) proteins that can represent a host cell proteome. In **chapter 5** we analyze the transferability of a model trained on single protein solutions for HCP retention prediction. For this, retention times of 13 proteins were measured under the same conditions as used in **chapter 4** and used to train a QSPR model. This model was evaluated on 572 *E. Coli* HCPs and was able to predict retention behavior for 51% with sufficient accuracy (error ≤5%). Moreover, we identified the key attributes missing in the training dataset, which is important to increase model performance in the future.

Data quality is essential for successful training of QSPR models. Therefore, in **chapter 6** we compared the accuracy of three isotherm parameter determination methods for a HIC isotherm. Specifically, two correlation-based methods (Parente and Wetlaufer, and Yamamoto) and one simulation error minimization method (inverse method) were assessed for two proteins in different conditions. By comparing mechanistic modeling accuracies compared to the experimental data, the inverse method was found to produce most accurate results, followed by the Yamamoto method. Therefore, it provides practical guidance for method selection for isotherm determination, thereby enabling generation of high-quality data that can facilitate QSPR model training.

Overall, this thesis highlights the potential of QSPR for predicting the chromatographic behavior of proteins. Specifically for HCP prediction QSPR shows to be a valuable tool when paired with state-of-the-art structure prediction. Therefore, it contributes to a significant step towards in silico process development.

Samenvatting

Vaccinatie speelt een cruciale rol in de moderne preventieve gezondheidszorg en draagt bij aan een wereldwijde afname van infectieziekten. Efficiënte productie van deze vaccins is essentieel om te voldoen aan de groeiende vraag, die resulteert uit factoren zoals een groeiende wereldbevolking en toename in internationaal reisverkeer. Eiwit-subunitvaccins zijn een type vaccin dat delen van de infectieuze ziekteverwekker bevat als werkzame stof. Deze subunits worden herkend door het immuunsysteem, dat hierdoor wordt getraind om symptomen bij een daadwerkelijke infectie te verminderen.

De productie van deze vaccins is onderverdeeld in upstream processing (USP), waarbij fermentatie plaatsvindt met behulp van expressiehosts, downstream processing (DSP), waarbij het eiwit-subunit wordt gezuiverd, en tot slot formulering, waarbij de vaccins worden voorbereid voor distributie. Tijdens DSP worden vaak meerdere chromatografiemethoden gebruikt om de vereiste zuiverheid te bereiken. De selectie van de optimale chromatografische resins en de bijbehorende procescondities kan kostbaar en tijdrovend zijn. Modelgebaseerde procesontwikkeling heeft het potentieel om deze selectie te versnellen door gebruik te maken van computationele methoden om het gedrag van eiwitten te voorspellen. Vooral in de vroege ontwikkelingsfase maken modellen in silico screening van resin en condities mogelijk waardoor minder experimenten en daardoor minder materiaal nodig is. Deze computationele modellen kunnen worden onderverdeeld in kennisgedreven, datagedreven of een combinatie daarvan.

De focus van dit proefschrift ligt op de ontwikkeling van een datagedreven modelleringsaanpak waarbij het gedrag van eiwitten

wordt voorspeld op basis van hun atomaire structuur. Hiervoor worden kwantitatieve structuur-eigenschaprelatie (QSPR) modellen gebruikt. In **hoofdstuk 2** wordt een Python tool geïntroduceerd die is ontwikkeld om relevante informatie te extraheren uit de driedimensionale eiwitstructuur. Dit gebeurt door het oppervlak van het eiwit te beschrijven met behulp van een roosterrepresentatie die de verdeling van verschillende fysisch-chemische eigenschappen beschrijft. Deze worden vertaald naar numerieke descriptoren. Met behulp van literatuurdata werd de relevantie van deze descriptoren aangetoond door twee afzonderlijke QSPR-modellen te trainen voor de voorspelling van retentietijden in ionenuitwisselingschromatografie (IEX), wat resulteerde in een gevalideerde R^2 van 0,87 en 0,95.

Een beperking van de datagedreven modelleringsaanpak is dat deze modellen alleen geldig zijn voor de experimentele condities waarop ze zijn getraind. Kennisgedreven modellen maken gebruik van fundamentele kennis, bijvoorbeeld over massatransport. In chromatografie zijn adsorptie-isothermen essentieel om de binding van een eiwit aan een chromatografische resin te beschrijven. In **hoofdstuk 3** ontwikkelden we een multiscale modelleringsaanpak door QSPR te integreren met mechanistische modellering. Adsorptie-isothermparameters voorspeld door QSPR werden gebruikt in een mechanistisch model. Dit multiscale model werd gevalideerd met experimentele data en toonde slechts 0,2% verschil tussen de retentiepieken, relatief ten opzichte van de zoutgradiëntlengte. Vervolgens werd het gevalideerde mechanistische model gebruikt om een chromatografische vangststap te optimaliseren.

Commercieel beschikbare modeleiwitten vormen een goede basis voor een proof of principle, maar QSPR-modellering wordt krachtiger wanneer toegepast op hostcel-eiwitten (HCPs). Daarom karakteriseerden we in **hoofdstuk 4** het chromatografisch gedrag van de HCPs aanwezig in een Escherichia coli (*E. coli*) lysaat door middel

van fractionering en daaropvolgende analyse met massaspectrometrie. Retentietijden van respectievelijk 816 en 908 HCPs werden verzameld voor hydrofobe interactiechromatografie (HIC) en IEX. Door de HCPs op te splitsen in subsets op basis van cellulaire locatie, functie en interacties werden basistrends zichtbaar gemaakt. Vervolgens voorspelden we de structuur van elk individueel HCP, die werd gebruikt om QSPR-modellen te trainen. Dit was succesvol voor de IEX-data, wat resulteerde in een QSPR-model met een gevalideerde R^2 van 0,70 bij gebruik van de monomeer-HCP-subset.

Het verkrijgen van retentiegegevens met hoge resolutie voor HCPs vereist nog steeds aanzienlijke experimentele inspanning. Daarom zou de inzet van QSPR-modellen voor procesontwikkeling baat hebben bij het opstellen van een lijst van breed beschikbare (commerciële) eiwitten die een hostcelproteoom kunnen representeren. In **hoofdstuk 5** analyseren we de overdraagbaarheid van een model dat is getraind op oplossingen van enkele eiwitten voor de voorspelling van HCP-retentie. Hiervoor werden de retentietijden van 13 eiwitten gemeten onder dezelfde condities als in **hoofdstuk 4** en gebruikt om een QSPR-model te trainen. Dit model werd geëvalueerd op 572 E. coli HCPs en kon het retentiegedrag van 51% met voldoende nauwkeurigheid voorspellen (fout ≤5%). Bovendien identificeerden we de belangrijkste kenmerken die ontbraken in de trainingsdataset, wat belangrijk is om de modelprestaties in de toekomst te verbeteren.

Datakwaliteit is essentieel voor succesvolle training van QSPR-modellen. Daarom vergeleken we in **hoofdstuk 6** de nauwkeurigheid van drie methoden voor het bepalen van isothermparameters voor een HIC-isotherm. Specifiek werden twee correlatiegebaseerde methoden (Parente en Wetlaufer, en Yamamoto) en een simulatiegebaseerde foutminimalisatiemethode (inverse methode) geëvalueerd voor twee eiwitten onder verschillende condities. Door de nauwkeurigheid van de mechanistische modellering te vergelijken met experimentele data,

bleek de inverse methode de meest nauwkeurige resultaten te leveren, gevolgd door de Yamamoto-methode. Dit biedt praktische richtlijnen voor de keuze van een methode voor isothermbepaling en maakt het mogelijk om hoogwaardige data te genereren die QSPR-modeltraining kunnen ondersteunen.

Al met al benadrukt dit proefschrift het potentieel van QSPR voor het voorspellen van het chromatografisch gedrag van eiwitten. Vooral voor HCP-voorspelling blijkt QSPR een waardevol hulpmiddel te zijn in combinatie met geavanceerde structuurvoorspelling. Daarmee levert het een belangrijke bijdrage aan in silico procesontwikkeling.

Glossary

Abbreviation	Definition
AEX	Anion exchange chromatography
CEX	Cation exchange chromatography
CHO	Chinese hamster ovary
CV	Column volumes
DoE	Design of experiments
DRT	Dimensionless retention time
DSP	Downstream processing
EP	Electrostatic potential
HCP	Host cell protein
HIC	Hydrophobic interaction chromatography
HTS	High throughput screening
IEX	Ion exchange chromatography
IM	Inverse method
KS	Kolmogorov-Smirnov
LGE	Linear gradient experiments
mAbs	Monoclonal antibodies
MAE	Mean absolute error
MD	Molecular dynamics
MHP	Molecular hydrophobicity potential
MLR	Multi linear regression
MM	Mechanistic modelling
MS	Mass spectrometry
PD	Process development
pI	Isoelectric point
PLR	Partial least squares
PPI	Protein-protein interaction
PW	Parente and Wetlaufer
QbD	Quality by design
QSAR	Quantitative structure activity relationship
QSPR	Quantitative structure property relationship
RMSE	Root mean squared error
SASA	Solvent accessible surface area
SEC	Size exclusion chromatography
SFS	Sequential forward selection
UPLC	Ultra performance liquid chromatography
USP	Upstream processing

Chapter 1

General introduction and thesis outline

1.1 Background

Vaccine discovery has been pivotal in improving public health, and has contributed to extending the average life expectancy up to 30 years in most middle- to high-income countries over the last two centuries.[1] By probing the immune system by controlled pathogen exposure, a vaccine reduces severe symptoms upon wildtype infection.[2] As part of the biopharmaceutical industry, the total marked size for vaccines has grown to 77 billion dollars in 2023.[3]

The most important components of a vaccine are the active ingredients, which are the antigens that stimulate the immune system.[2] These active ingredients can be whole pathogens, as live attenuated or inactivated. Alternatively, specific parts of the pathogen that are recognized by the immune system can be used. The first SARS-CoV-2 vaccines are recent examples that use mRNA encoding for target antigens.[4] Upon vaccination, the mRNA transfects several host cells which will start producing the antigens, triggering a subsequent immune response.[5] These types of vaccines have proven to be a great success during the Covid-19 pandemic, as the established platform process allows for relatively fast process development (PD).

Alternatively, protein subunit vaccines already contain these specific antigens, and do therefore not require transfection and translation after vaccination. In contrast to mRNA vaccines, which are stable for up to 6 months when frozen (-20 to -80 °C)[6], protein subunit vaccines have been reported to be stable for multiple years when refrigerated (2 to 8 °C).[7,8] Therefore protein subunit vaccines currently have less distribution limitations.

These vaccines are also known as recombinant vaccines, meaning that they are produced during a fermentation process by host cells which are transformed/transfected with DNA aimed to express the antigen.[9]

Common host cells used for the production include bacteria, yeast, insect, and mammalian cells. The process of amplifying host cells and expressing the antigens is the upstream processing (USP), subsequently, active ingredients require purification from the crude mixture, which is important to ensure safety and efficacy.[10] This is done during the downstream processing (DSP), which precedes the vaccine formulation (Figure 1.1). During the DSP of protein subunits, removal of host cell proteins (HCPs) is most challenging, as these impurities might show similar behaviors as the antigens. For separation, chromatography often has a central role during DSP due to its versatility and specificity.[11]

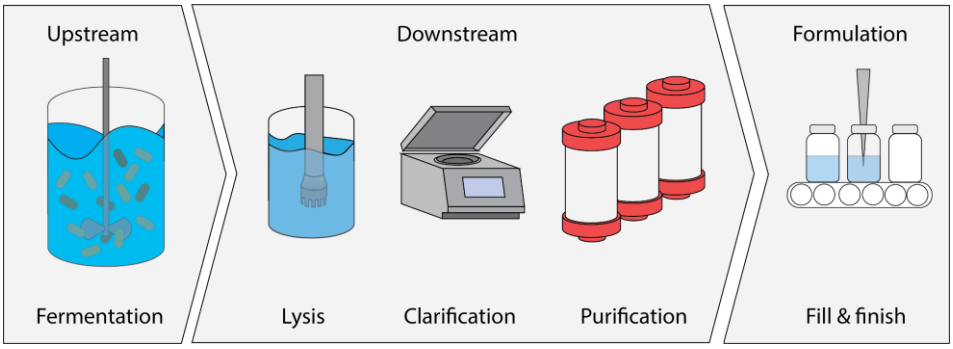


Figure 1.1: *General representation of a vaccine production pipeline.*

1.1.1 Chromatography

Packed bed chromatography is the most commonly used technique to achieve high resolution separation of proteins.[12] For this method a column is packed with porous beads, the resin, which will bind solutes based on their physicochemical properties (stationary phase). Solvent passes through the column, dragging along any dissolved proteins (mobile phase). The retention of a protein is determined by how strongly the protein binds to the chromatographic resin. Ion exchange chromatography (IEX) is one of the most used methods and DSP of pharmaceutical proteins often have one or more of these steps.[13] This type of chromatography separates based on charge; cation exchange

(CEX) or anion exchange (AEX) resins bind proteins based on positive or negative charge, respectively. Other alternative modes of separation include hydrophobic interaction chromatography (HIC), which separates based on hydrophobicity, size exclusion chromatography (SEC), separation based on size, or mixed mode chromatography, which is a combination of multiple modes (e.g., IEX and HIC).[14,15]

Typical vaccine purification consists of several orthogonal chromatography steps performing an initial capture, followed by an intermediate purification and final polishing.[11,16] During PD, appropriate resins are selected, and operating conditions are optimized to ensure a robust process. Optimization can be performed by heuristics or by experimental screening methods like design of experiments (DoE) or high throughput screening (HTS).[10] Alternatively, model-based PD uses data- and/or knowledge-driven methods to predict protein behavior in silico.[17] This reduces the required wet-lab experiments and thereby materials and has therefore the potential to significantly reduce development time and costs.

1.1.2 Model-based process development

Mechanistic modeling (MM) is a prime example of a knowledge-driven method that simulates the chromatographic behavior of proteins (Figure 1.2). In these models, partial differential equations describe the transport in the liquid phase, while partitioning between the solid and liquid phase is approximated by adsorption isotherms.[18] Process parameters such as column dimensions, operating conditions, and buffer compositions can be tested and optimized.[16,19,20] Successful deployment of MM is highly dependent on the model parameters, especially the adsorption parameters.[21] These parameters are determined experimentally from sets of dynamic (isocratic or linear gradient elution experiments) or static (batch adsorption studies).[22–24]

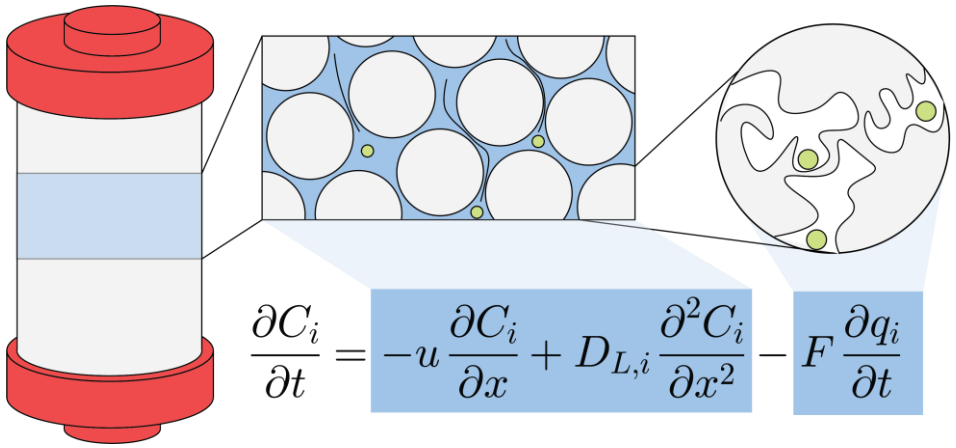


Figure 1.2: *Schematic representation of a mechanistic model. Different levels of a chromatography column are depicted left to right showing first the whole column, followed by the packed bed and finally a porous bead. The equation shows the general formulation of a lumped kinetic model where C_i and $q_i$represent the solute concentration in the mobile and stationary phase, respectively. u represents the superficial velocity, $D_{L,i}$ the axial dispersion, F the phase ratio, x the position in the column, and t represents the time. A more detailed description of the mechanistic model used in this thesis is documented in* ***Chapter 3****.*

Alternatively, chromatographic behavior can be predicted from the protein structure. The physicochemical properties of proteins are a product of the amino acid sequence and subsequent protein folding. Structure models contain the positions of every atom and can therefore be used to calculate properties relevant for different chromatographic modes (Figure 1.3). Recent breakthroughs in the field of structure prediction, primarily by AlphaFold, enable fast obtainment of high quality structure models.[25–27] These models can be used in molecular dynamics (MD) simulations that calculate the molecular forces and movement of each atom at femtosecond time scales.[28,29] For chromatography, these simulations have been used to predict the binding energies, isotherm parameters, and preferred binding orientations.[30–35] While these simulations provide immense detail, computational costs are a limiting factor as simulations in the nanosecond range can take days to calculate. Therefore, this method scales poorly for large molecules, like proteins, and is currently unsuitable for screening purposes.

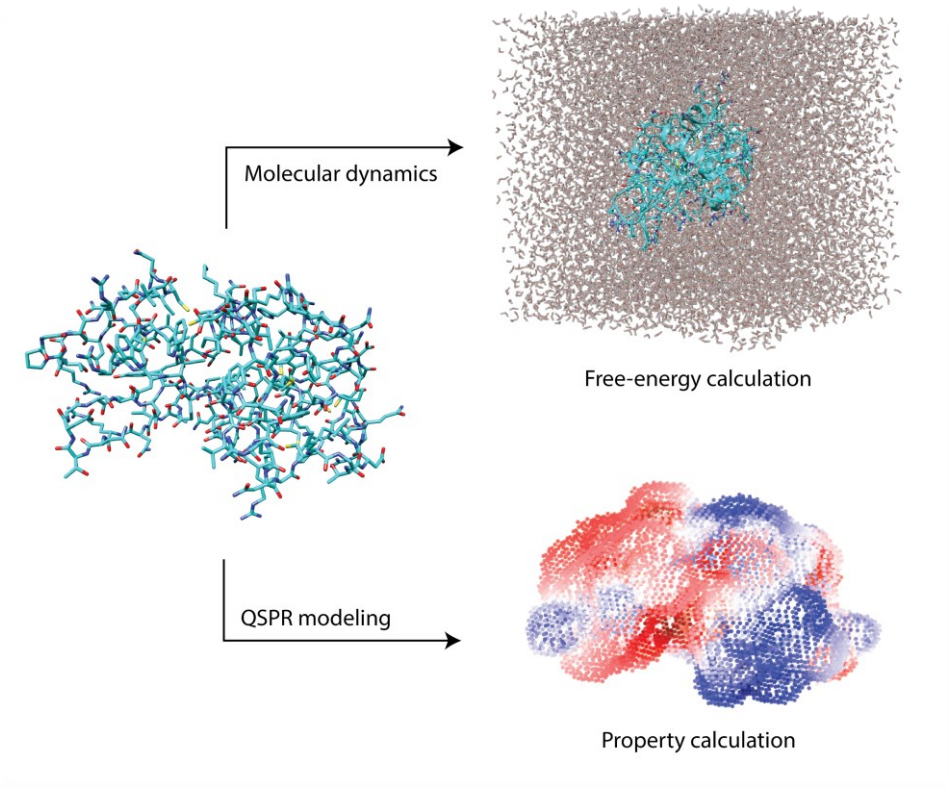


Figure 1.3: *Use of protein structure models. Top shows a solvated simulation box which is used during molecular dynamics simulations. Bottom shows surface charge projections that can be used to calculate protein properties for QSPR modeling*

Quantitative structure property relationship (QSPR) modeling is another method that uses the molecular structure to predict the chromatographic behavior.[11,36–38] This method is data-driven and uses fundamental knowledge derived from the structure to train predictive models. This method is most mature for the discovery of small molecule drugs where it carries the name quantitative structure activity relationships (QSAR).[39] Descriptors are calculated from the molecular structure which can range from number of double bonds to solvation energy.[40] Over 1000 distinct descriptors have been designed mainly focusing on one- or two-dimensional molecular representations. Proteins contain many more atoms, folded in complex structures.

Therefore, the descriptors designed for the small molecules are often not relevant. Different types of descriptors have been developed for protein chromatography prediction.[37,41–44] Specifically, surface descriptors that use the solvent accessible surface area of a protein onto which the hydrophobicity or charge can be distributed has shown to be effective. For the application of QSPR models, specific descriptors are selected and used to train regression or machine learning models that can recognize which descriptors are relevant to describe chromatographic retention. After training an accurate model, predicting the behavior of a new protein can be performed within seconds. QSPR is therefore an excellent method to screen different resin types in tandem with experimental characterization, limiting experimental efforts.

1.2 Project setting

The project Molecular Modeling for Protein Chromatography Prediction is a collaboration between GlaxoSmithKline Biologicals S.A. (Belgium) and Delft University of Technology (The Netherlands) and was partly funded by GlaxoSmithKline Biologicals S.A. (Belgium) and ChemistryNL (The Netherlands). The aim of this collaboration is to develop a model-based high throughput development platform for the DSP of protein subunit vaccines. This platform allows increased productivity and fundamental understanding. As such, two additional PhD projects are part of this collaboration. One of the projects focuses on the development of experimental methods to characterize HCPs which are applied to *Escherichia coli (E. coli)* lysates.[45] The other project aims to use MM to describe and optimize DSP.[46] The focus of this thesis is predicting the chromatographic behavior of proteins from their molecular structure. To support the goal of developing a high throughput development platform QSPR is used as the main modeling tool.

1.3 Thesis outline

The main content of this thesis is divided into 5 chapters which all focus on specific research questions (Figure 1.4).

In **chapter 2**, the feature calculation software that forms the basis of this thesis is introduced. It starts with an overview current state-of-the-art in QSPR modeling for protein chromatography, followed by a comprehensive explanation on how the features are calculated from protein structures. These features are subsequently used to train models capable of predicting IEX retention times obtained from literature.

This software is applied in **chapter 3** to predict the chromatographic behavior of model proteins in CEX using a multiscale modeling approach combining MM and QSPR. In this chapter, retention times as well as model parameters are predicted which are used to perform model-based optimization. To validate the impact of prediction uncertainty to the optimization, the parameters were varied using the 95% confidence interval.

In an actual process, an antigen often needs to be removed from a host cell lysate. Understanding this complex mixture of host cell material provides a great basis to guide PD. Therefore, in **chapter 4** the chromatographic behavior of HCPs from a host cell lysate is analyzed. Additionally, by using predicted HCP structures, QSPR models are trained to predict chromatographic retention.

As retention time determination of HCPs is experimentally demanding, general QSPR models are trained in **chapter 5**. Specifically, a set of widely available proteins are characterized for the same process conditions as used in **chapter 4**. The dimensionless retention times (DRTs) of these proteins are then used to train QSPR models which are applied to predict HCP DRTs. By analyzing feature distribution plots of

the training and HCP sets, concrete recommendations are made to improve training set selection.

While all previous chapters focus on IEX, **chapter 6** compares three adsorption isotherm parameter determination methods for HIC. This chapter contributes to the overall project as it describes different available methods and assesses the accuracy of all parameters. By providing practical guidance for method selection reliable HIC modeling is enabled, which can be extended to HCPs in the future.

The final **chapter 7** presents the overall conclusion of this thesis and summarizes all key findings. Using this information, prospects of the field are discussed and suggestions for future research are motivated.

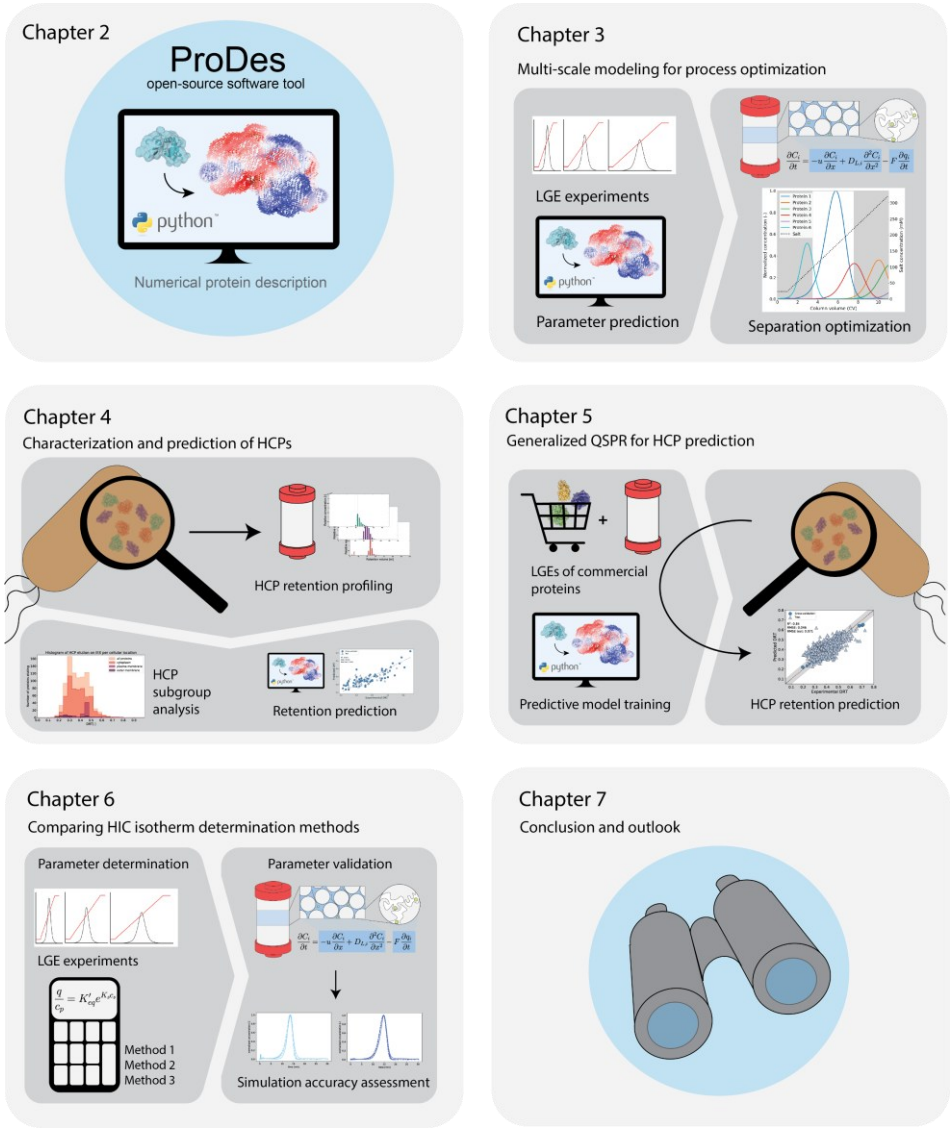


***Figure 1.4**: General overview of the Thesis*

1.4 References

1. Montero, D. A., Vidal, R. M., Velasco, J., Carreño, L. J., Torres, J. P., Benachi O, M. A., Tovar-Rosero, Y. Y., Oñate, A. A., & O'Ryan, M. (2023). Two centuries of vaccination: historical and conceptual approach and future perspectives. *Frontiers in Public Health*, *11*. https://doi.org/10.3389/fpubh.2023.1326154

2. Ghattas, M., Dwivedi, G., Lavertu, M., & Alameh, M. G. (2021). Vaccine technologies and platforms for infectious diseases: Current progress, challenges, and opportunities. *Vaccines*, *9*(12), 1–31. https://doi.org/10.3390/vaccines9121490

3. Accord, P., Union, A., Public, N., & Order, H. (2025). *Global vaccine market report 2024* (Issue December). World Health Organization. https://doi.org/10.2471/B09198

4. Park, H. J., Bang, Y. J., Kwon, S. P., Kwak, W., Park, S. I., Roh, G., Bae, S. H., Kim, J. Y., Kwak, H. W., Kim, Y., Yoo, S., Kim, D., Keum, G., Bang, E. K., Hong, S. H., & Nam, J. H. (2023). Analyzing immune responses to varied mRNA and protein vaccine sequences. *Npj Vaccines*, *8*(1). https://doi.org/10.1038/s41541-023-00684-0

5. Gote, V., Bolla, P. K., Kommineni, N., & Butreddy, A. (2023). *A Comprehensive Review of mRNA Vaccines | Enhanced Reader*.

6. Uddin, M. N., & Roni, M. A. (2021). Challenges of storage and stability of mrna-based covid-19 vaccines. *Vaccines*, *9*(9), 1–9. https://doi.org/10.3390/vaccines9091033

7. Flood, A., Estrada, M., Mcadams, D., Ji, Y., & Chen, D. (2016). *Development of Freeze Dried heat stable influenza vaccine*. 1–18. https://doi.org/10.17605/OSF.IO/CFK8Z

8. Sordo, Y., & Vargas, M. (2022). *Shelf Life and Accelerated Stability Studies of Porvac ® , a Marker Subunit Shelf Life and Accelerated Stability Studies of Porvac ® , a Marker Subunit Vaccine Against Classical Swine Fever*. *October*.

9. de Pinho Favaro, M. T., Atienza-Garriga, J., Martínez-Torró, C., Parladé, E., Vázquez, E., Corchero, J. L., Ferrer-Miralles, N., & Villaverde, A. (2022). Recombinant vaccines in 2022: a perspective from the cell factory. *Microbial Cell Factories*, *21*(1), 1–17. https://doi.org/10.1186/s12934-022-01929-8

10. Keulen, D., Geldhof, G., Bussy, O. Le, Pabst, M., & Ottens, M. (2022). Recent advances to accelerate purification process development: A review with a focus on vaccines. *Journal of Chromatography A*, *1676*, 463195. https://doi.org/10.1016/j.chroma.2022.463195

11. Hanke, A. T., & Ottens, M. (2014). Purifying biopharmaceuticals: Knowledge-based chromatographic process development. *Trends in Biotechnology*, *32*(4), 210–220. https://doi.org/10.1016/j.tibtech.2014.02.001

12. Boi, C., Malavasi, A., Carbonell, R. G., & Gilleskie, G. (2020). A direct comparison between membrane adsorber and packed column chromatography performance. *Journal of Chromatography A*, *1612*, 460629. https://doi.org/10.1016/j.chroma.2019.460629

13. Grönberg, A. (2018). Ion Exchange Chromatography. In *Biopharmaceutical Processing: Development, Design, and Implementation of Manufacturing Processes*. Elsevier Ltd. https://doi.org/10.1016/B978-0-08-100623-8.00018-9

14. Eriksson, K. O. (2018). Hydrophobic Interaction Chromatography. In *Biopharmaceutical Processing* (Vol. 130). Elsevier Ltd. https://doi.org/10.1016/B978-0-08-100623-8.00019-0

15. Hagel, L. (2011). Protein Purification- 3 - Gel filtration: size exclusion

chromatography. *Protein Purification: Principles, High Resolution Methods, and Applications*, 51–91.

16. Keulen, D., Apostolidi, M., Geldhof, G., Le Bussy, O., Pabst, M., & Ottens, M. (2024). Comparing in silico flowsheet optimization strategies in biopharmaceutical downstream processes. *Biotechnology Progress*, *August*, 1–16. https://doi.org/10.1002/btpr.3514

17. Wittkopp, F., Welsh, J., Todd, R., Staby, A., Roush, D., Lyall, J., Karkov, S., Hunt, S., Griesbach, J., Bertran, M. O., & Babi, D. (2024). Current state of implementation of in silico tools in the biopharmaceutical industry—Proceedings of the 5th modeling workshop. *Biotechnology and Bioengineering*, *May*, 2952–2973. https://doi.org/10.1002/bit.28768

18. Kumar, V., & Lenhoff, A. M. (2020). Mechanistic Modeling of Preparative Column Chromatography for Biotherapeutics. *Annual Review of Chemical and Biomolecular Engineering*, *11*, 235–255. https://doi.org/10.1146/annurev-chembioeng-102419-125430

19. Huuk, T. C., Hahn, T., Osberghaus, A., & Hubbuch, J. (2014). Model-based integrated optimization and evaluation of a multi-step ion exchange chromatography. *Separation and Purification Technology*, *136*, 207–222. https://doi.org/10.1016/j.seppur.2014.09.012

20. Pirrung, S. M., Berends, C., Backx, A. H., van Beckhoven, R. F. W. C., Eppink, M. H. M., & Ottens, M. (2019). Model-based optimization of integrated purification sequences for biopharmaceuticals. *Chemical Engineering Science: X*, *3*, 100025. https://doi.org/10.1016/j.cesx.2019.100025

21. Shekhawat, L. K., Tiwari, A., Yamamoto, S., & Rathore, A. S. (2022). An accelerated approach for mechanistic model based prediction of linear gradient elution ion-exchange chromatography of proteins. *Journal of Chromatography A*, *1680*, 463423. https://doi.org/10.1016/j.chroma.2022.463423

22. Kittelmann, J., Ottens, M., & Hubbuch, J. (2015). Robust high-throughput batch screening method in 384-well format with optical in-line resin quantification. *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences*, *988*, 98–105. https://doi.org/10.1016/j.jchromb.2015.02.028

23. Hess, R., Yun, D., Saleh, D., Briskot, T., Grosch, J. H., Wang, G., Schwab, T., & Hubbuch, J. (2023). Standardized method for mechanistic modeling of multimodal anion exchange chromatography in flow through operation. *Journal of Chromatography A*, *1690*, 463789. https://doi.org/10.1016/j.chroma.2023.463789

24. Saleh, D., Wang, G., Müller, B., Rischawy, F., Kluters, S., Studts, J., & Hubbuch, J. (2020). Straightforward method for calibration of mechanistic cation exchange chromatography models for industrial applications. *Biotechnology Progress*, *36*(4), 1–12. https://doi.org/10.1002/btpr.2984

25. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

26. Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., & Moult, J. (2023). Critical assessment of methods of protein structure prediction (CASP)—Round XV. *Proteins: Structure, Function and Bioinformatics*, *91*(12), 1539–1549. https://doi.org/10.1002/prot.26617

27. David, A., Islam, S., Tankhilevich, E., & Sternberg, M. J. E. (2022). The AlphaFold Database of Protein Structures: A Biologist's Guide. *Journal of Molecular Biology*,

434(2), 167336. https://doi.org/10.1016/j.jmb.2021.167336

28. Durrant, J. D., & McCammon, J. A. (2011). Molecular dynamics simulations and drug discovery. *BMC Biology*, *9*(1), 71. https://doi.org/10.1186/1741-7007-9-71

29. Hollingsworth, S. A., & Dror, R. O. (2018). Molecular Dynamics Simulation for All. *Neuron*, *99*(6), 1129–1143. https://doi.org/10.1016/j.neuron.2018.08.011

30. Ballweg, T., Liu, M., Grimm, J., Sedghamiz, E., Wenzel, W., & Franzreb, M. (2024). All-atom modeling of methacrylate-based multi-modal chromatography resins for Langmuir constant prediction of peptides. *Journal of Chromatography A*, *1730*(June), 465089. https://doi.org/10.1016/j.chroma.2024.465089

31. Parimal, S., Garde, S., & Cramer, S. M. (2015). Interactions of Multimodal Ligands with Proteins: Insights into Selectivity Using Molecular Dynamics Simulations. *Langmuir*, *31*(27), 7512–7523. https://doi.org/10.1021/acs.langmuir.5b00236

32. Liang, J., Fieg, G., & Jakobtorweihen, S. (2015). Molecular Dynamics Simulations of a Binary Protein Mixture Adsorption onto Ion-Exchange Adsorbent. *Industrial and Engineering Chemistry Research*, *54*(10), 2794–2802. https://doi.org/10.1021/ie504374x

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

34. Tournois, M., Mathé, S., André, I., Esque, J., & Fernández, M. A. (2020). Understanding adsorption behavior of α-chymotrypsin onto cation exchanger using all-atom molecular dynamics simulations. *Journal of Chromatography A*, *1614*, 460720. https://doi.org/10.1016/j.chroma.2019.460720

35. Jakobtorweihen, S., Heuer, J., & Waluga, T. (2020). A novel approach to calculate protein adsorption isotherms by molecular dynamics simulations. *Journal of Chromatography A*, *1620*, 460940. https://doi.org/10.1016/j.chroma.2020.460940

36. Mazza, C. B., Sukumar, N., Breneman, C. M., & Cramer, S. M. (2001). Prediction of protein retention in ion-exchange systems using molecular descriptors obtained from crystal structure. *Analytical Chemistry*, *73*(22), 5457–5461. https://doi.org/10.1021/ac010797s

37. Malmquist, G., Nilsson, U. H., Norrman, M., Skarp, U., Strömgren, M., & Carredano, E. (2006). Electrostatic calculations and quantitative protein retention models for ion exchange chromatography. *Journal of Chromatography A*, *1115*(1–2), 164–186. https://doi.org/10.1016/j.chroma.2006.02.097

38. Hou, Y., & Cramer, S. M. (2011). Evaluation of selectivity in multimodal anion exchange systems: A priori prediction of protein retention and examination of mobile phase modifier effects. *Journal of Chromatography A*, *1218*(43), 7813–7820. https://doi.org/10.1016/j.chroma.2011.08.080

39. Muratov, E. N., Bajorath, J., Sheridan, R. P., Tetko, I. V., Filimonov, D., Poroikov, V., Oprea, T. I., Baskin, I. I., Varnek, A., Roitberg, A., Isayev, O., Curtalolo, S., Fourches, D., Cohen, Y., Aspuru-Guzik, A., Winkler, D. A., Agrafiotis, D., Cherkasov, A., & Tropsha, A. (2020). QSAR without borders. *Chemical Society Reviews*, *49*(11), 3525–3564. https://doi.org/10.1039/d0cs00098a

40. Vilar, S., Cozza, G., & Moro, S. (2008). Medicinal Chemistry and the Molecular Operating Environment (MOE): Application of QSAR and Molecular Docking to Drug Discovery. *Current Topics in Medicinal Chemistry*, *8*(18), 1555–1572. https://doi.org/10.2174/156802608786786624

41. Emonts, J., & Buyel, J. F. (2023). An overview of descriptors to capture protein

properties – Tools and perspectives in the context of QSAR modeling. *Computational and Structural Biotechnology Journal*, *21*, 3234–3247. https://doi.org/10.1016/j.csbj.2023.05.022

42. Kittelmann, J., Lang, K. M. H., Ottens, M., & Hubbuch, J. (2017). Orientation of monoclonal antibodies in ion-exchange chromatography: A predictive quantitative structure–activity relationship modeling approach. *Journal of Chromatography A*, *1510*, 33–39. https://doi.org/10.1016/j.chroma.2017.06.047

43. Kittelmann, J., Lang, K. M. H., Ottens, M., & Hubbuch, J. (2017). An orientation sensitive approach in biomolecule interaction quantitative structure–activity relationship modeling and its application in ion-exchange chromatography. *Journal of Chromatography A*, *1482*, 48–56. https://doi.org/10.1016/j.chroma.2016.12.065

44. Sankar, K., Trainor, K., Blazer, L. L., Adams, J. J., Sidhu, S. S., Day, T., Meiering, E., & Maier, J. K. X. (2022). A Descriptor Set for Quantitative Structure-property Relationship Prediction in Biologics. *Molecular Informatics*, *41*(9), 2100240. https://doi.org/10.1002/minf.202100240

45. Disela, R. (2025). Chromatographic host cell protein removal in biopharmaceutical purification. In *TU Delft University*. https://doi.org/10.4233/uuid:5d0ce064-f5de-458f-92e7-7cd314ea9ae2

46. Keulen, D. (2024). Computational modeling and optimization of biopharmaceutical downstream processes. *TU Delft University*, 289. https://doi.org/10.4233/uuid:f55e8d73-d9c5-4e38-bb8b-43a87301ef82

Chapter 2

Predicting protein retention in ion-exchange chromatography using an open source QSPR workflow

Published as:

Neijenhuis, T., Le Bussy, O., Geldhof, G., Klijn, M. E., & Ottens, M. (2024). Predicting protein retention in ion-exchange chromatography using an open source QSPR workflow. Biotechnology Journal, 19(3), 2300708.

Abstract

Protein-based biopharmaceuticals require high purity before final formulation to ensure product safety, making process development time consuming. Implementation of computational approaches at the initial stages of process development offers a significant reduction in development efforts. By preselecting process conditions, experimental screening can be limited to only a subset. One such computational selection approach is the application of Quantitative Structure Property Relationship (QSPR) models that describe the properties exploited during purification. This work presents a novel open-source Python tool capable of extracting a range of features from protein 3D models on a local computer allowing total transparency of the calculations. As an open-source tool, it also impacts initial investments in constructing a QSPR workflow for protein property prediction for third parties, making it widely applicable within the field of bioprocess development. The focus of current calculated molecular features is projection onto the protein surface by constructing surface grid representations. Linear regression models were trained with the calculated features to predict chromatographic retention times/volumes. Model validation shows a high accuracy for anion and cation exchange chromatography data (cross-validated R^2 of 0.87 and 0.95). Hence, these models demonstrate the potential of the use of QSPR to accelerate process design.

2.1 Introduction

The market for protein-based biopharmaceuticals, such as protein subunit vaccines and therapeutic antibodies, developed rapidly over recent years.[1] Opposed to chemical synthesis to manufacture small molecule drugs, protein-based biopharmaceuticals are produced by living host cells. During downstream processing (DSP) the target product is separated from host cell impurities, which is of major importance to guarantee patient safety and drug efficacy. To attain sufficient purity, chromatography is a method of choice due to its specificity and versatility.[2–4] However, the vast variety of commercially available resin types (e.g., ion exchange (IEX) or hydrophobic interaction chromatography (HIC)) and experimental conditions (e.g., salt concentrations, buffers, and pH) results in extensive experimental screening to obtain optimal separation conditions, driving both cost and development time. In silico preselection of resins and conditions prior to experimentation would allow a decrease in costs and development time by narrowing the empirical screening space.

Chromatographic separation is based on the difference in physicochemical properties between the product and impurities. For proteins, physicochemical properties are determined by the amino acid sequence (1D) and the three-dimensional (3D) structure. Quantitative Structure Property Relationship (QSPR) aims to relate physicochemical properties to specific behavior (e.g., chromatographic retention time).[5] For QSPR, physicochemical properties are described as numerical features and subsequently used in predictive machine learning models as input variables. To build a QSPR workflow, experimental data of known proteins is split in a training and test set. Numerical features are calculated from the proteins in the training set and selected to train a machine learning model (e.g., linear regression, partial least squares (PLS), or neural networks) which predicts the

behavior of interest. The resulting model is tested using the numerical features obtained from the proteins in the test set, to assess the model accuracy for new data. When the model provides sufficiently accurate predictions, the property of proteins unknown to the model can be predicted (Figure 2.1).

The simplest QSPR approach is to calculate protein features based on the amino acid sequence. From the amino acid sequences, one can derive properties such as residue counts, hydrophobicity scores, overall charge, and the isoelectric point. Although these properties are indicative, such features consider the contribution of each residue as equal since topological information on whether the residue is buried or accessible for resin ligands is lacking. This information can be obtained from 3D protein structure models. Developments in protein structure prediction allows accurate prediction of protein structures from amino acid sequences, the current state-of-the-art being Alphafold2.[6,7] PROFEAT[8] and ProtDCal[9] offer webserver interfaces where structure files can be analyzed to calculate protein features needed as input for QSPR model approaches. Both tools calculate a list of general numerical features based on the 1D and 3D protein structure. For feature calculations using a local machine, the drug discovery software platform Molecular Operating Environment (MOE) is widely applied.[10–16] An alternative package is Schrödinger's BioLuminate Suite, which has recently been expanded by including features based on the protein sequence, 3D structure, and surface patches.[17] A comprehensive overview can be found elsewhere.[18]

Using structural protein features to predict protein retention times was first described in 2001 by Mazza et al., who calculated protein features using the transferable atom equivalent method[5,19,20] and the proprietary software platform MOE. By applying a genetic algorithm for feature selection, a PLS model was trained, capable of predicting retention times for ion exchange chromatography from protein

structure models. Applying the same feature calculation methods, support vector machine regressions for both feature selection and the final predictive model have also been applied for successful protein retention prediction in ion exchange, hydrophobic interaction and mixed mode chromatography.[10–16] As the chromatographic resin interacts with the amino acid residues on the protein surface, Malmquist et al. implemented a grid representation of the protein surface to map protein properties.[21] By applying distance functions to project charge and hydrophobicity onto the surface grid points, protein features were calculated and used in a PLS model to predict retention times for anion and cation exchange columns. As charge and hydrophobicity are usually not uniformly distributed over the protein surface, binding orientations play important roles in protein-resin binding affinities.[22,23] To account for such orientations in QSPR models, Hanke et al. described a method to sample the surface in neighborhoods and uses this for HIC retention time predictions.[24] These neighborhoods are defined as the surface within a specific distance of a central surface point (7 Å and 14 Å distances were described). Alternatively, Kittelmann et al. used property projections on a plane, sampling different orientations.[25,26] By projecting the properties onto a plane, this method considers steric hindrance on the surface. This results in penalizing the area of surface cavities, which are located at a greater distance from the projection plane.

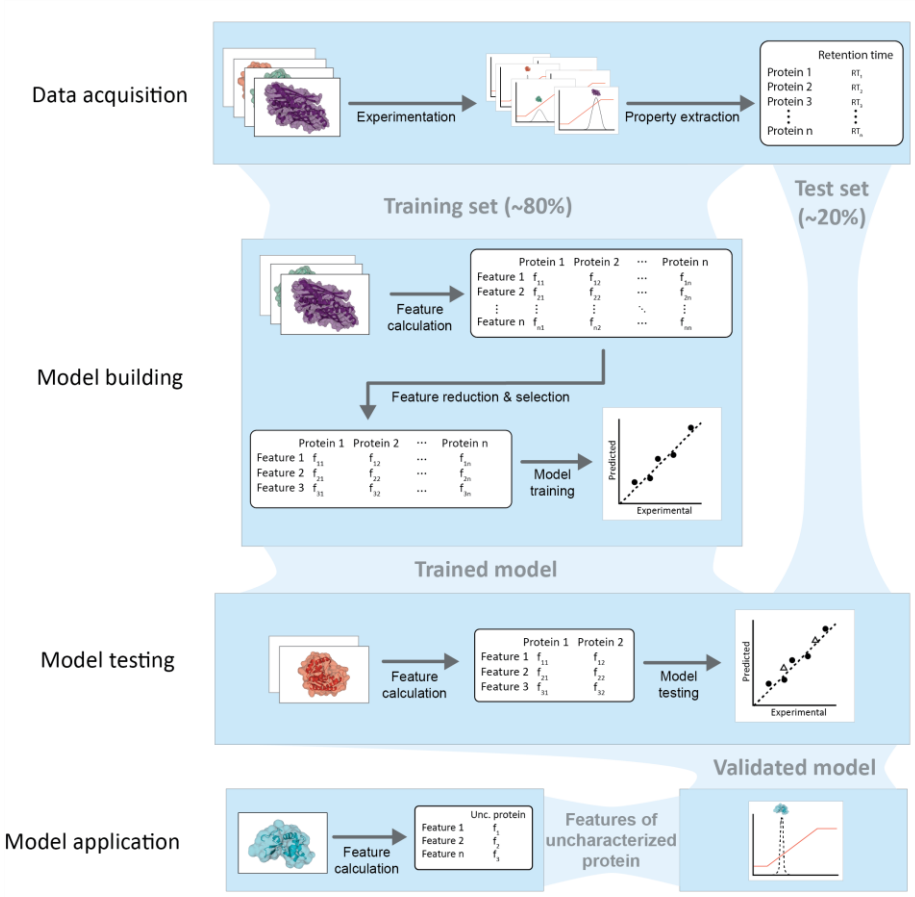


Figure 2.1: *Schematic representation of a Quantitative Structure Property Relationship (QSPR) workflow for chromatographic retention prediction. The first step to build a QSPR model is data acquisition. Here, a set of known proteins is used to construct a dataset containing experimentally determined properties (e.g., retention times). The experimental property dataset is split into a train and test set. The training set is used for model building. The physicochemical properties for each protein are calculated using the corresponding 3D structure. The physicochemical properties are expressed as numerical features. The number of features is reduced using dimension reduction methods such as principle component analysis or variance filtering, and the most descriptive features are selected by feature selection to train a predictive model. The resulting model is tested on the test set to assess the accuracy for unseen proteins. Predictive models with good accuracy can be applied to predict the properties of uncharacterized proteins.*

Most of the described studies use proprietary or in-house software to perform feature calculations and model training. As a result, reproducing these studies is near to impossible. Therefore, direct comparison between different approaches by minimizing the variables cannot be performed, hindering benchmarking opportunities and scientific progress. Additionally, the lack of open-source tools limits software availability for new users and customizability to solve a wide variety of development challenges. We aim to close this gap, and in this work, we provide an open-source Python tool that is able to calculate 3D protein features. The current implemented operations and features aim to consolidate the most often described protein features from literature.[13,21,25,26] The validity of the features for chromatographic process development was shown by training multiple linear regression (MLR) models predicting retention times/volumes for cation and anion chromatography resins obtained from literature. To promote transparency and scientific reproducibility, the software developed for this study is freely available open source at **https://dx.doi.org/10.5281/zenodo.10369949**.

2.2 Methods

2.2.1 Protein charge

Protein charge is the key property that governs separation in ion exchange chromatography. Protein charge is dependent on the protonation state of the titratable groups. Residues Arginine (Arg, R), Lysine (Lys, L) and Histidine (His, H) can have positively charged sidechains when fully protonated, while Aspartic acid (Asp, D), Glutamic acid (Glu, E), Cysteine (Cys, C) and Tyrosine (Tyr, T) can be negatively charged when deprotonated. Additionally, the C and N termini of the protein can also be negatively or positively charged, respectively. The protonation states of these residues can be described by the Henderson-Hasselbalch Equation[27]:

$$pH = pKa + \log\left(\frac{[A^-]}{[AH]}\right), \quad (2.1)$$

where AH is the protonated and A^- is the deprotonated form of the titratable group. Therefore, titratable residue sidechains are deprotonated when the pH is higher than their pKa and protonated when the pH is lower than their pKa resulting in charges of +1, 0 or -1. Alternatively, the overall charge can be calculated for negative and positive charges as follows:

$$Charge = \frac{-1}{1+10^{pKa-pH}} \text{ [e]}, \quad (2.2)$$

And

$$Charge = \frac{1}{1+10^{pH-pKa}} \text{ [e]}, \quad (2.3)$$

respectively. By default, pKa values are assigned based on a scale documented in Leninger Principles of Biochemistry[28] with the exception of Arginine, which is set to 14.[29] Alternatively, custom pKa values (predicted by e.g. PROPKA[30,31], H++[32,33], WHAT-IF[34]) can be assigned to specific residues using a json object, allowing improved description of the charge. To describe charge distribution, the dipole moment of the protein can be calculated which is defined as the magnitude of the dipole vector D, calculated as:

$$D = 4.803 * \sum_i (r_i - r_p) * q_i \text{ [D]}, \quad (2.4)$$

where r_p is the protein center and r_i is a vector containing the 3-dimentional coordinates of the atom.[35,36]

2.2.2 Surface definition

Interactions of proteins with their environment often take place at the protein surface. To rationalize these interactions using protein models, accurate representations of the surfaces are required. The Solvent Accessible Surface Area (SASA) is the most common for surface estimation that represents the protein surface which can be occupied by water molecules and was first described by Lee and Richards[37]

(Figure 2B). A number of tools specifically designed for the determination of the SASA are available.[38–40] A spherical probe, representing a solvent molecule, is rolled over the protein atoms tracing the accessible area using the center of the solvent. We adopted the method of Shrake and Rupley[41] where each surface sphere is represented by a set of sample points. The number of sample points are scaled according to the surface sphere radius and are distributed by a Fibonacci sphere[42], to obtain a distribution of 2 points per Å^2. The fraction of each amino acid occupying the surface can be calculated by dividing the number of surface points of a residue by the total number of surface points.

2.2.3 Property projection

Projection of properties onto the surface allows for assessing structural attributes where the interactions occur. A surface grid representation is composed by constructing grid cells of 1 Å^3 containing the surface. Using connected component labeling connecting the grid points occupied by the surface, a surface grid representation with a distribution of 1 point per Å^3 is composed (Figure 2C). Projection of charge, resulting in simplified electrostatic potential (EP), is performed by:

$$EP = \sum_i \frac{q_i}{\varepsilon d_i} \text{ [v]}, \tag{2.5}$$

where d represents the distance between atom i and the grid point, q is the charge of atom i and ε the dielectric constant of a protein, which is assumed to be 4.[43]

To represent a chromatographic resin, charges are mapped onto planes (Figure 2D). A total of 120 planes are equally distributed using a Fibonacci sphere and scaled to a distance of ≥1 Å to any of the protein atoms. Since the charge is now mapped through multiple media, ε is defined as:

$$\varepsilon = \frac{\varepsilon_p \times d_p + \varepsilon_w \times d_w}{d} \varepsilon_0 \quad [-], \tag{2.6}$$

Where subscript p indicates protein, w the solvent and 0 the conductivity in a vacuum. The distance through the protein and solvent is estimated using the solvent accessible surface.

Hydrophobicity of proteins is another important factor which governs interactions. Many different scales describing the contribution of each respective amino acid to hydrophobic phenomena have been published.[44] The Cowan–Whittaker[45] and the Miyazawa-Jernigan[46] scales have been reported to give highest correlation for hydrophobic interaction chromatography retention prediction.[47] In this work, we use the Miyazawa-Jernigan[46] scale, which was scaled using a min-max-scaler to values ranging from -1 to 1. Hydrophobicity values are projected onto the surface grid to obtain the molecular hydrophobic potential (MHP) using:

$$MHP = \sum_i f_i e^{-d_i} \quad [-], \tag{2.7}$$

where f_i indicates the hydrophobicity value of the residue, based on the work of Fauchére et al.[48] with a cut-off of 10 Å.

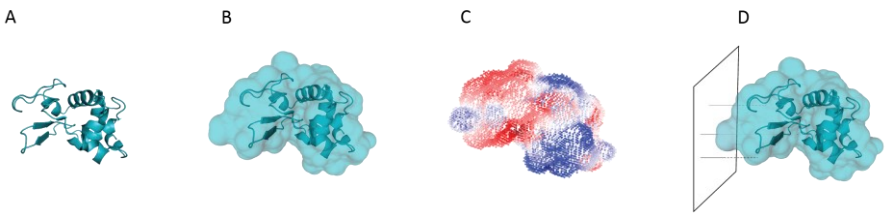


Figure 2.2: *Protein representation for feature calculation. A) shows all atom representation using the coordinates for each atom. B) shows the solvent accessible surface area. C) shows the surface grid representation with mapped electrostatic potentials. D) shows the plane projection of one orientation*

A list of all current supported features can be found in Supplemental Table S2.1.

2.2.4 Dataset composition and feature calculation

Two datasets with known retention behavior for Q Sepharose FF and SP Sepharose HP were required from literature, set 1[15] and set 2[13] respectively (Tables 2.1 and 2.2). For both datasets, structures were extracted from the PDB and used to generate homology models by SWISS-MODEL[49,50] to resolve missing atoms. Duplicate chains were removed for all protein models to obtain monomer structures which were used in the feature calculation. To calculate the protonation states, the default pKa values were used for the titratable residues. Building the surface grid was performed using a sphere radius of 1.4 Å to represent water.

Table 2.1: *Dataset 1, Retention times of specific proteins described by Hou and Cramer[15] for Q Sepharose Fast Flow. Superscript 1 indicates the protein models used as test set.*

Protein	PDB-ID	Retention time (min)
Lectin	2PEL	12.35
Phosphorylase	1GPB[1]	12.56
Conalbumin	1AIV	15.31
Transferrin	1A8E	15.63
Trypsin Inhibitor	1AVU	16.19
a-Lactalbumin	1F6R	18.63
Glutamic Dehydrogenase	1NR7	21.29
Ovalbumin	1OVA	21.47
Lipoxydase	1F8N	23.02
Human Serum Albumin	1AO6	23.19
Adenosine Deaminase	1VFL	25.00
B-Lactoglobulin B	1BSQ[1]	26.26
Lipase	3TGL	26.51
B-Lactoglobulin A	1BSO	29.16
Cellulase	1EG1	29.71
Amyloglucosidase	1LF6	36.61

Table 2.2: *Dataset 2, Retention volumes of specific proteins at different pHs described by Yang et al.[13] for sulfopropyl Sepharose high-performance. Superscript 1 indicates the pH used as test set (6).*

		Retention volume (mL)				
Protein	**PDB-ID**	**pH 4**	**pH 5**	**pH 6[1]**	**pH 7**	**pH 8**
Carbonic anhydrase	1V9E		7.86	3.51		
Conalbumin	1OVT		6.18	3.21	1.52	
Pyruvate kinase	1A49		7.48	2.37		
Bovine trypsin	1S81	6.94	3.82	2.37	2.14	1.15
Bee phospholipase A2	1POC	11.83	8.01	5.64	3.35	1.37
Elastase	1LVY	5.80	3.81	2.47	2.51	2.29
Trypsinogen	1TGB	7.17	4.27	3.34	3.34	2.90
Ribonuclease A	1RBX	13.12	9.23	5.72	4.96	3.66
α-Chymotrypsin	5CHA	8.93	6.87	5.95	5.87	5.19
α-Chymotrypsin A	2CGA	8.55	6.64	5.87	5.95	5.34
Bovine cytochrome C	2B4Z	17.55	10.91	8.39	8.47	7.86
Horse cytochrome C	1HRC	17.63	10.91	8.39	8.47	7.93
Lysozyme	1AKI	14.12	10.83	9.54	9.16	8.01
Avidin	1VYO	19.54	14.96	12.36	10.73	9.77
Aprotin	1PIT	14.35	11.29	10.68	10.68	10.53
Lactoferrin	1BKA	26.87	25.34	24.96	24.81	23.89

2.2.5 Linear regression modeling

After splitting the data in train and test sets, a correlation filter was applied for the removal of features with a high Pearson correlation coefficient (0.99). Deciding which features should remain was based on the Pearson correlation with the protein retention times/volumes, making this a supervised feature filter. Next the feature list was further reduced based on the Pearson correlation with the retention times, removing 30% and 10% of the features with lowest correlation for dataset 1 and dataset 2 respectively. Sequential forward feature selection was used for selecting the features for the linear regression model. Selected feature sets were validated using a repeated 2-fold cross-validation and leave-one-out cross-validation. Feature importance was assessed according to regression coefficients and by feature permutation.

2.3 Results and discussion

To evaluate the performance of the developed Python tool, two datasets were obtained from literature containing protein retention

times/volumes for ion-exchange chromatography columns. The first dataset contains protein retention for Q Sepharose FF, and the second for SP Sepharose HP. For both datasets, predictive models were trained relating protein structure to retention time or volume. To determine the validity of the selected features, the regression coefficient and cross-validated R^2 of a permutation model, where each feature is scrambled, are discussed.

2.3.1 Protein retention prediction for Q Sepharose FF

To develop a simple model with high interpretability, a MLR model was trained on protein retention times for the anion exchange resin Q Sepharose FF (Table 2.3). The dataset that was used (Table 2.1) was composed of 16 proteins, of which two were selected for testing while the remaining 14 were used for model training.[15] As overfitting can be an issue for linear regression models, a ratio of five datapoints per feature should be maintained, resulting in three features for this dataset [51]. The model's predictability was considered sufficient, with a cross-validated R^2 of 0.87, a RMSE of 2.23, and $RMSE_{test}$ of 2.50 (Figure 2.3). The two most important features are the median negative surface EP (regression coefficient of -31 and permutated CV R^2 of -0.352) and the number of positive electrostatic surface grid points (regression coefficient of 18.17 and a permutated CV R^2 of 0.563), both calculated using the formal charge (Table 2.3). A negative regression coefficient indicates an inverse correlation with the retention time of the protein and vice versa. In alignment with the mode of action of the anion exchange resin, the negative surface potential is the most important feature, as it has the highest regression coefficient and permutation of this feature yields a model uncapable of predicting retention times (Supplemental Figure 2.1A). The second feature, number of surface points with a positive EP, shows a positive correlation with protein retention time. This is not in line with the mode of action

as a positive protein surface should be repelled by the anion exchange resin. Permutation of this feature reduces the performance of the model to a cross-validated R^2 of 0.563 (Supplemental Figure 2.1B). The selection of this feature might be due to the current absence of local surface descriptors. The affected proteins might still contain areas on the surface which are negatively charged that could interact with the anion exchange ligands. The final feature, the valine surface fraction, is of the lowest importance, with a regression coefficient of -5.75. The permutation of this feature results in a model with a cross-validated R^2 of 0.733.

Table 2.3: *Overview of features selected for the linear regression model for Q Sepharose FF and the corresponding regression coefficient and cross-validated R2 of permutation models*

Feature	Coefficient	CV R^2 permutation
Intercept	36.76	-
Negative surface EP* median (formal)[a]	-31	-0.352
Number of surface points with positive EP* (formal)[a]	18.17	0.563
Valine surface fraction	-5.75	0.733

[a] Charge calculated using formal charge (+1, 0 or -1). * Electrostatic Potential

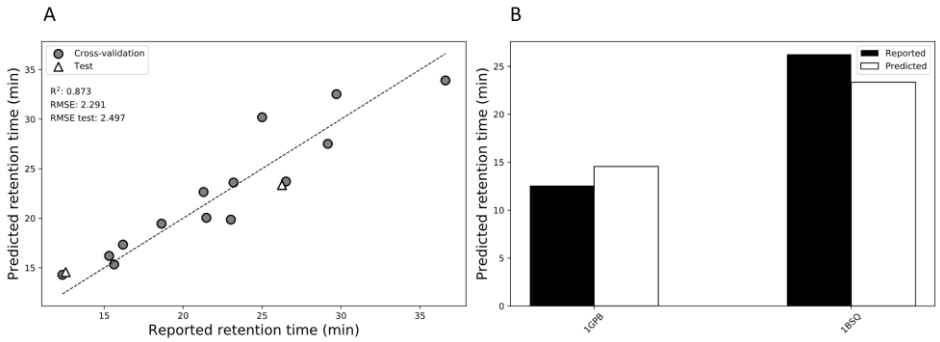


Figure 2.3: *Prediction of Q Sepharose FF retention times. A) shows the leave-one-out cross-validation (gray circles) and test set (white triangles) results of the model. B) shows the predicted retention times volumes for the external test set (Table 1).*

2.3.2 pH dependent protein retention prediction for SP Sepharose HP

The applicability of the Python tool for a different chromatography mode and varying process conditions was tested using a second set of protein retention volumes reported in literature.[13] The second set consists of retention volumes of 16 different proteins for the cation exchange resin SP Sepharose HP. In contrast to the previous dataset, the proteins were measured at a pH range from 4 to 8, yielding a total of 72 datapoints. The obtained numerical features were filtered and subsequently selected using forward feature selection, shown in Table 4. The final MLR model is composed of 10 features and has good predictability with a cross-validated R^2 of 0.95, a RMSE of 1.37, and $RMSE_{test}$ of 1.14 (Figure 2.4).

Six of the 10 selected features are directly related to the protein charge and are inherently interconnected. The feature with the highest regression coefficient of 31.24, and therefore deemed most important, is the minimum surface EP. The positive coefficient indicates that an increase in minimum surface EP leads to a higher retention volume, which is in line with the mode of action of the cation exchange resin. The total charge is the second most important feature with a regression coefficient of -27.77. This indicates that proteins with a higher total charge have lower retention volumes. Considering the dataset to be retention volumes for the cation exchange resin SP Sepharose HP, a negative correlation with the total charge is counter intuitive. This correlation might not indicate a direct causation with the retention volume, but rather that the total charge might compensate for other charge related features, as there is collinearity between the charge related features. To directly assess the importance of the feature, the permutation model results in a reduced cross-validated R^2 of 0.861. The permutation model for the minimum surface EP resulted in a

greater decrease in performance (cross-validated R^2 of 0.822). This indicates that the total charge is indeed less important for the final model compared to EP.

The dipole vector length has a regression coefficient of 20.72. The high positive regression coefficient indicates the importance of charge polarization, and that proteins elute later with more uneven charge distribution. The isoelectric point is the next charge-related feature with a regression coefficient of 12.02. This feature is unaffected by pH as it represents the pH at which the protein is neutrally charged. Interestingly, even though the feature has only the fourth highest coefficient, permutation of the feature results in a permutation model with the lowest R^2 of 0.769 (Supplemental Figure 2D). This feature has a low cross correlation with the other features, indicating that less compensation is possible with the remaining data. The importance of the remaining features is significantly lower compared to the first four features (Cross-validated R^2 of permutation > 0.888), a detailed discussion on these features can be found in the supplemental material.

Table 2.4: *Overview of features selected for the linear regression model for SP sepharose HP and the corresponding regression coefficient and cross-validated R^2 of permutation models*

Feature	Coefficient	CV R^2 permutation
Intercept	-3.78	-
Minimum surface EP* (average)[b]	31.24	0.822
Total charge (average)[b]	-27.77	0.861
Dipole vector length	20.72	0.842
Isoelectric point	12.02	0.769
Standard deviation of positive EP* shell projections	11.07	0.934
Lysine surface fraction	-7.42	0.919
Mean negative surface EP* (formal)[a]	-5.48	0.934
Standard deviation of negative surface hydrophobicity	5.46	0.934
Cysteine surface fraction	5.12	0.888
Surface shape max	-1.21	0.946

[a] Charge represented as formal charge (+1, 0 or -1). [b] Charge calculated using equations 2 and 3. * Electrostatic Potential.

While the QSPR model for the first dataset is trained to predict different proteins at similar conditions, the second model is trained to predict similar proteins for different pH conditions. The effect of different pH values is captured by five of the 10 selected features which are pH dependent (Minimum surface EP, Total charge, Dipole vector length, Standard deviation of positive shell projections and Mean negative surface EP). Thus, the remaining five features are pH independent, and therefore similar for different pH conditions. Therefore, a slight bias might have been introduced, indicated by clustering of identical proteins. The impact of this bias is considered minimal due to the greater regression coefficients and effect of permutation of the pH dependent features. The increased amount of available data for the second model is therefore thought to be the main factor driving greater accuracy compared to the first model.

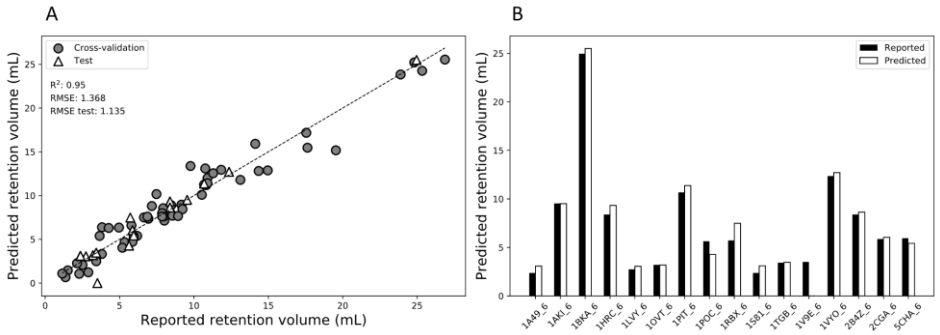


Figure 2.4: *Prediction of SP Sepharose HP retention volumes. A) shows model results of the leave-one-out cross-validation (gray circles) of the proteins at pH 4, 5, 7 and 8 as well as the test set (white triangles) which are the proteins at pH 6. B) shows the predicted retention volumes for the external test set which are all proteins measured at pH 6 (Table 2).*

The two QSPR models are cabable of the retention prediction for Q Sepharose FF and SP Sepharose HP. All physical phenomena are described implicitly, therefore these models would only be suitable for describing retention behavior for these specific resins. Extending these models to predict protein retention of other resins would require additional data. This data can subsequently be used in a similar model building approach as described here, yielding predictive models for the new conditions.

2.4 Conclusion

Physically relevant protein features are essential to achieve robust predictions of protein properties, like chromatographic retention behavior. To mature the field of protein QSPR, adaptable and transparent open-source software for the calculation of protein features is essential to directly benchmark between different tools and improve the current state-of-the-art. Using the open-source software presented here, we were able to train models that predict the retention times/volumes for two different ion-exchange chromatography datasets, showing applicability for unknown proteins and differences in pH (cross-validated R2 of 0.87 and 0.95, respectively). Most features

selected by the forward feature selection method have an apprehensible relation to protein retention for specific chromatographic conditions. However, collinearity between multiple features was observed. Model performance might therefore benefit from feature reduction techniques such as principal component analysis or PLS regression. Nevertheless, these models show good performance and would allow for pre-screening of chromatographic resins. Finally, it was shown that the amount of data available for model training is a major factor determining model accuracy. By increasing the available input data for protein properties like chromatographic retention time, the true impact of the 3D protein features and in silico property prediction for process design can be unlocked in the future.

2.5 References

1. Kesik-Brodacka, M. (2018). Progress in biopharmaceutical development. *Biotechnology and Applied Biochemistry*, *65*(3), 306–322. https://doi.org/10.1002/bab.1617

2. Gronemeyer, P., Ditz, R., & Strube, J. (2014). Trends in upstream and downstream process development for antibody manufacturing. *Bioengineering*, *1*(4), 188–212. https://doi.org/10.3390/bioengineering1040188

3. Hanke, A. T., & Ottens, M. (2014). Purifying biopharmaceuticals: Knowledge-based chromatographic process development. *Trends in Biotechnology*, *32*(4), 210–220. https://doi.org/10.1016/j.tibtech.2014.02.001

4. Keulen, D., Geldhof, G., Bussy, O. Le, Pabst, M., & Ottens, M. (2022). Recent advances to accelerate purification process development: A review with a focus on vaccines. *Journal of Chromatography A*, *1676*, 463195. https://doi.org/10.1016/j.chroma.2022.463195

5. Mazza, C. B., Sukumar, N., Breneman, C. M., & Cramer, S. M. (2001). Prediction of protein retention in ion-exchange systems using molecular descriptors obtained from crystal structure. *Analytical Chemistry*, *73*(22), 5457–5461. https://doi.org/10.1021/ac010797s

6. Masrati, G., Landau, M., Ben-Tal, N., Lupas, A., Kosloff, M., & Kosinski, J. (2021). Integrative Structural Biology in the Era of Accurate Structure Prediction. *Journal of Molecular Biology*, *433*(20), 167127. https://doi.org/10.1016/j.jmb.2021.167127

7. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

8. Rao, H. B., Zhu, F., Yang, G. B., Li, Z. R., & Chen, Y. Z. (2011). Update of PROFEAT: A web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Research*, *39*(SUPPL. 2), 385–390. https://doi.org/10.1093/nar/gkr284

9. Ruiz-Blanco, Y. B., Paz, W., Green, J., & Marrero-Ponce, Y. (2015). ProtDCal: A program to compute general-purpose-numerical descriptors for sequences and 3D-structures of proteins. *BMC Bioinformatics*, *16*(1), 1–15. https://doi.org/10.1186/s12859-015-0586-0

10. Song, M., Breneman, C. M., Bi, J., Sukumar, N., Bennett, K. P., Cramer, S., & Tugcu, N. (2002). Prediction of Protein Retention Times in Anion-Exchange Chromatography Systems Using Support Vector Regression. *Journal of Chemical Information and Computer Sciences*, *42*(6), 1347–1357. https://doi.org/10.1021/ci025580t

11. Ladiwala, A., Rege, K., Breneman, C. M., & Cramer, S. M. (2003). Investigation of Mobile Phase Salt Type Effects on Protein Retention and Selectivity in Cation-Exchange Systems Using Quantitative Structure Retention Relationship Models. *Langmuir*, *19*(20), 8443–8454. https://doi.org/10.1021/la0346651

12. Ladiwala, A., Rege, K., Breneman, C. M., & Cramer, S. M. (2005). A priori prediction of adsorption isotherm parameters and chromatographic behavior in ion-exchange systems. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(33), 11710–11715. https://doi.org/10.1073/pnas.0408769102

13. Yang, T., Sundling, M. C., Freed, A. S., Breneman, C. M., & Cramer, S. M. (2007). Prediction of pH-dependent chromatographic behavior in ion-exchange systems. *Analytical Chemistry*, *79*(23), 8927–8939. https://doi.org/10.1021/ac071101j

14. Chen, J., & Cramer, S. M. (2007). Protein adsorption isotherm behavior in hydrophobic interaction chromatography. *Journal of Chromatography A*, *1165*(1–2), 67–77. https://doi.org/10.1016/j.chroma.2007.07.038

15. Hou, Y., & Cramer, S. M. (2011). Evaluation of selectivity in multimodal anion exchange systems: A priori prediction of protein retention and examination of mobile phase modifier effects. *Journal of Chromatography A*, *1218*(43), 7813–7820. https://doi.org/10.1016/j.chroma.2011.08.080

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

17. Sankar, K., Trainor, K., Blazer, L. L., Adams, J. J., Sidhu, S. S., Day, T., Meiering, E., & Maier, J. K. X. (2022). A Descriptor Set for Quantitative Structure-property Relationship Prediction in Biologics. *Molecular Informatics*, *41*(9), 2100240. https://doi.org/10.1002/minf.202100240

18. Emonts, J., & Buyel, J. F. (2023). An overview of descriptors to capture protein properties – Tools and perspectives in the context of QSAR modeling. *Computational and Structural Biotechnology Journal*, *21*, 3234–3247. https://doi.org/10.1016/j.csbj.2023.05.022

19. Whitehead, C. E., Breneman, C. M., Sukumar, N., & Ryan, M. D. (2003). Transferable atom equivalent multicentered multipole expansion method. *Journal of Computational Chemistry*, *24*(4), 512–529. https://doi.org/10.1002/jcc.10240

20. Breneman, C. M., Thompson, T. R., Rhem, M., & Dung, M. (1995). Electron density modeling of large systems using the transferable atom equivalent method. *Computers & Chemistry*, *19*(3), 161–179. https://doi.org/10.1016/0097-8485(94)00052-G

21. Malmquist, G., Nilsson, U. H., Norrman, M., Skarp, U., Strömgren, M., & Carredano, E. (2006). Electrostatic calculations and quantitative protein retention models for ion exchange chromatography. *Journal of Chromatography A*, *1115*(1–2), 164–186. https://doi.org/10.1016/j.chroma.2006.02.097

22. Dismer, F., & Hubbuch, J. (2007). A novel approach to characterize the binding orientation of lysozyme on ion-exchange resins. *Journal of Chromatography A*, *1149*(2), 312–320. https://doi.org/10.1016/j.chroma.2007.03.074

23. Dismer, F., Petzold, M., & Hubbuch, J. (2008). Effects of ionic strength and mobile phase pH on the binding orientation of lysozyme on different ion-exchange adsorbents. *Journal of Chromatography A*, *1194*(1), 11–21. https://doi.org/10.1016/j.chroma.2007.12.085

24. Hanke, A. T., Klijn, M. E., Verhaert, P. D. E. M., van der Wielen, L. A. M., Ottens, M., Eppink, M. H. M., & van de Sandt, E. J. A. X. (2016). Prediction of protein retention times in hydrophobic interaction chromatography by robust statistical characterization of their atomic-level surface properties. *Biotechnology Progress*, *32*(2), 372–381. https://doi.org/10.1002/btpr.2219

25. Kittelmann, J., Lang, K. M. H., Ottens, M., & Hubbuch, J. (2017). Orientation of monoclonal antibodies in ion-exchange chromatography: A predictive quantitative structure–activity relationship modeling approach. *Journal of Chromatography A*, *1510*, 33–39. https://doi.org/10.1016/j.chroma.2017.06.047

26. Kittelmann, J., Lang, K. M. H., Ottens, M., & Hubbuch, J. (2017). An orientation sensitive approach in biomolecule interaction quantitative structure–activity relationship modeling and its application in ion-exchange chromatography. *Journal of Chromatography A*, *1482*, 48–56. https://doi.org/10.1016/j.chroma.2016.12.065

27. Henderson, L. J. (1908). CONCERNING THE RELATIONSHIP BETWEEN THE STRENGTH OF ACIDS AND THEIR CAPACITY TO PRESERVE NEUTRALITY. *American Journal of Physiology-Legacy Content*, *21*(2), 173–179. https://doi.org/10.1152/ajplegacy.1908.21.2.173

28. Nelson, D. L., & Cox, M. M. (2001). *Lehninger Principles of Biochemistry*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-08289-8

29. Fitch, C. A., Platzer, G., Okon, M., Garcia-Moreno, B. E., & McIntosh, L. P. (2015). Arginine: Its pKa value revisited. *Protein Science*, *24*(5), 752–761. https://doi.org/10.1002/pro.2647

30. Olsson, M. H. M., Søndergaard, C. R., Rostkowski, M., & Jensen, J. H. (2011). PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical p K a Predictions. *Journal of Chemical Theory and Computation*, *7*(2), 525–537. https://doi.org/10.1021/ct100578z

31. Bas, D. C., Rogers, D. M., & Jensen, J. H. (2008). Very fast prediction and rationalization of pKa values for protein-ligand complexes. *Proteins: Structure, Function and Genetics*, *73*(3), 765–783. https://doi.org/10.1002/prot.22102

32. Gordon, J. C., Myers, J. B., Folta, T., Shoja, V., Heath, L. S., & Onufriev, A. (2005). H++: A server for estimating pKas and adding missing hydrogens to macromolecules. *Nucleic Acids Research*, *33*(SUPPL. 2), 368–371. https://doi.org/10.1093/nar/gki464

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

34. Vriend, G. (1990). WHAT IF: A molecular modeling and drug design program. *Journal*

of Molecular Graphics, *8*(1), 52–56. https://doi.org/10.1016/0263-7855(90)80070-V

35. Antosiewicz, J. (1995). Computation of the dipole moments of proteins. *Biophysical Journal*, *69*(4), 1344–1354. https://doi.org/10.1016/S0006-3495(95)80001-9

36. Felder, C. E., Prilusky, J., Silman, I., & Sussman, J. L. (2007). A server and database for dipole moments of proteins. *Nucleic Acids Research*, *35*(SUPPL.2), 512–521. https://doi.org/10.1093/nar/gkm307

37. Lee, B., & Richards, F. M. (1971). The interpretation of protein structures: Estimation of static accessibility. *Journal of Molecular Biology*, *55*(3), 379–400. https://doi.org/10.1016/0022-2836(71)90324-X

38. Touw, W. G., Baakman, C., Black, J., Te Beek, T. A. H., Krieger, E., Joosten, R. P., & Vriend, G. (2015). A series of PDB-related databanks for everyday needs. *Nucleic Acids Research*, *43*(D1), D364–D368. https://doi.org/10.1093/nar/gku1028

39. Mitternacht, S. (2016). FreeSASA: An open source C library for solvent accessible surface area calculations. *F1000Research*, *5*, 1–11. https://doi.org/10.12688/f1000research.7931.1

40. Ali, S., Hassan, Md., Islam, A., & Ahmad, F. (2014). A Review of Methods Available to Estimate Solvent-Accessible Surface Areas of Soluble Proteins in the Folded and Unfolded States. *Current Protein & Peptide Science*, *15*(5), 456–476. https://doi.org/10.2174/1389203715666140327114232

41. Shrake, A., & Rupley, J. A. (1973). Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *Journal of Molecular Biology*, *79*(2), 361–371. https://doi.org/10.1016/0022-2836(73)90011-9

42. Swinbank, R., & Purser, R. J. (2006). Fibonacci grids: A novel approach to global modelling. *Quarterly Journal of the Royal Meteorological Society*, *132*(619), 1769–1793. https://doi.org/10.1256/qj.05.227

43. Schutz, C. N., & Warshel, A. (2001). What are the dielectric 'constants' of proteins and how to validate electrostatic models? *Proteins: Structure, Function and Genetics*, *44*(4), 400–417. https://doi.org/10.1002/prot.1106

44. Simm, S., Einloft, J., Mirus, O., & Schleiff, E. (2016). 50 years of amino acid hydrophobicity scales: Revisiting the capacity for peptide classification. *Biological Research*, *49*(1), 1–19. https://doi.org/10.1186/s40659-016-0092-5

45. Cowan, R., & Whittaker, R. G. (1990). Hydrophobicity indices for amino acid residues as determined by high-performance liquid chromatography. *Peptide Research*, *3*(2), 75–80.

46. Miyazawa, S., & Jernigan, R. L. (1985). Estimation of Effective Interresidue Contact Energies from Protein Crystal Structures: Quasi-Chemical Approximation. *Macromolecules*, *18*(3), 534–552. https://doi.org/10.1021/ma00145a039

47. Lienqueo, M. E., Mahn, A., & Asenjo, J. A. (2002). Mathematical correlations for predicting protein retention times in hydrophobic interaction chromatography. *Journal of Chromatography A*, *978*(1–2), 71–79. https://doi.org/10.1016/S0021-9673(02)01358-4

48. Fauchére, J. L., Quarendon, P., & Kaetterer, L. (1988). Estimating and representing hydrophobicity potential. *Journal of Molecular Graphics*, *6*(4), 203–206. https://doi.org/10.1016/S0263-7855(98)80004-0

49. Bienert, S., Waterhouse, A., De Beer, T. A. P., Tauriello, G., Studer, G., Bordoli, L., & Schwede, T. (2017). The SWISS-MODEL Repository-new features and functionality. *Nucleic Acids Research*, *45*(D1), D313–D319. https://doi.org/10.1093/nar/gkw1132

50. Kiefer, F., Arnold, K., Künzli, M., Bordoli, L., & Schwede, T. (2009). The SWISS-MODEL Repository and associated resources. *Nucleic Acids Research*, *37*(SUPPL. 1), 387–392. https://doi.org/10.1093/nar/gkn750

51. Topliss, J. G., & Costello, R. J. (1972). Chance correlations in structure-activity studies using multiple regression analysis. *Journal of Medicinal Chemistry*, *15*(10), 1066–1068. https://doi.org/10.1021/jm00280a017

2.6 Supplemental information

Supplemental Table S2.1: *list of all descriptors calculated*

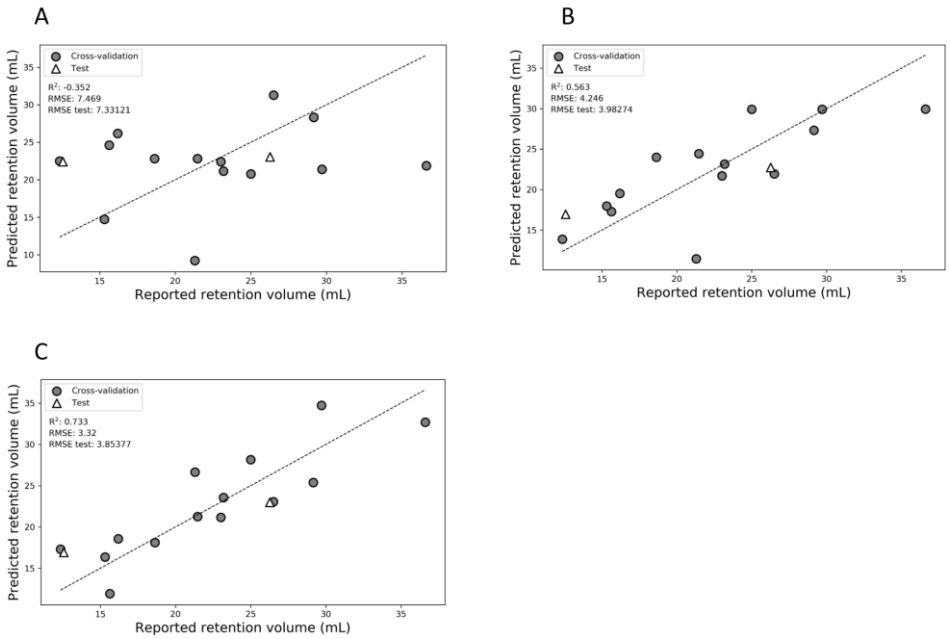
Name	unit	Description
Molecular weight	Da	Sum of the weight of each amino acid in the protein
Shape min	-	Shape of the protein surface, calculated by dividing the minimum distance by the average distance between the surface points and the protein centre of mass (COM)
Shape max	-	Shape of the protein surface, calculated by dividing the average distance by the maximum distance between the surface points and the protein COM
Area	$Å^2$	The calculated surface area based on the Shrake Rupley algorithm
Formal charge	-	The charge calculated based on the pH of the solution using binary charges of +1, 0 or -1
Average charge	-	The charge calculated based on the pH of the solution using charges ranging between 1 and -1
Isoelectric point	-	The estimated isoelectric point
Dipole	Å	The magnitude of the dipole vector
NsurfPoints	-	The number of surface grid points
AlaSurfFrac	-	Fraction of alanine on the surface calculated by dividing the alanine surface area by the total surface area
ArgSurfFrac	-	Fraction of arginine on the surface calculated by dividing the arginine surface area by the total surface area
AsnSurfFrac	-	Fraction of asparagine on the surface calculated by dividing the asparagine surface area by the total surface area
AspSurfFrac	-	Fraction of aspartic acid on the surface calculated by dividing the aspartic acid surface area by the total surface area
CysSurfFrac	-	Fraction of cysteine on the surface calculated by dividing the cysteine surface area by the total surface area
GlnSurfFrac	-	Fraction of glutamine on the surface calculated by dividing the glutamine surface area by the total surface area
GluSurfFrac	-	Fraction of glutamic acid on the surface calculated by dividing the glutamic acid surface area by the total surface area
GlySurfFrac	-	Fraction of glycine on the surface calculated by dividing the glycine surface area by the total surface area
HisSurfFrac	-	Fraction of histidine on the surface calculated by dividing the histidine surface area by the total surface area
IleSurfFrac	-	Fraction of isoleucine on the surface calculated by dividing the isoleucine surface area by the total surface area
LeuSurfFrac	-	Fraction of leucine on the surface calculated by dividing the leucine surface area by the total surface area

LysSurfFrac	-	Fraction of lysine on the surface calculated by dividing the lysine surface area by the total surface area
MetSurfFrac	-	Fraction of methionine on the surface calculated by dividing the methionine surface area by the total surface area
PheSurfFrac	-	Fraction of phenylalanine on the surface calculated by dividing the phenylalanine surface area by the total surface area
ProSurfFrac	-	Fraction of proline on the surface calculated by dividing the proline surface area by the total surface area
SerSurfFrac	-	Fraction of serine on the surface calculated by dividing the serine surface area by the total surface area
ThrSurfFrac	-	Fraction of threonine on the surface calculated by dividing the threonine surface area by the total surface area
TrpSurfFrac	-	Fraction of tryptophane on the surface calculated by dividing the tryptophane surface area by the total surface area
TyrSurfFrac	-	Fraction of tyrosine on the surface calculated by dividing the tyrosine surface area by the total surface area
ValSurfFrac	-	Fraction of valine on the surface calculated by dividing the valine surface area by the total surface area
SurfEpMaxFormal	v	The maximum observed electrostatic potential calculated using binary charges of +1, 0 or -1
SurfEpMeanFormal	v	The mean of all electrostatic potentials calculated using binary charges of +1, 0 or -1
SurfEpTrimeanFormal	v	The trimean of all electrostatic potentials calculated using binary charges of +1, 0 or -1
SurfEpminFormal	v	The minimum observed electrostatic potential calculated using binary charges of +1, 0 or -1
SurfEpMedianFormal	v	The median of all electrostatic potentials calculated using binary charges of +1, 0 or -1
SurfEpSumFormal	v	The sum of all electrostatic potentials calculated using binary charges of +1, 0 or -1
SurfEpStdFormal	v	The standard deviation of the electrostatic potentials calculated using binary charges of +1, 0 or -1
NSurfPosEpFormal	v	Number of points with a positive electrostatic potential calculated using binary charges of +1, 0 or -1
SurfPosEpMeanFormal	v	The mean of all positive electrostatic potentials calculated using binary charges of +1, 0 or -1
SurfPosEpTrimeanFormal	v	The trimean of all positive electrostatic potentials calculated using binary charges of +1, 0 or -1
SurfPosEpMedianFormal	v	The median of all positive electrostatic potentials calculated using binary charges of +1, 0 or -1
SurfPosEpsumFormal	v	The sum of all positive electrostatic potentials calculated using binary charges of +1, 0 or -1
SurfPosEpFracFormal	-	The fraction of points with a positive electrostatic potential, NsurfPosEp/NsurfPoints calculated using binary charges of +1, 0 or -1

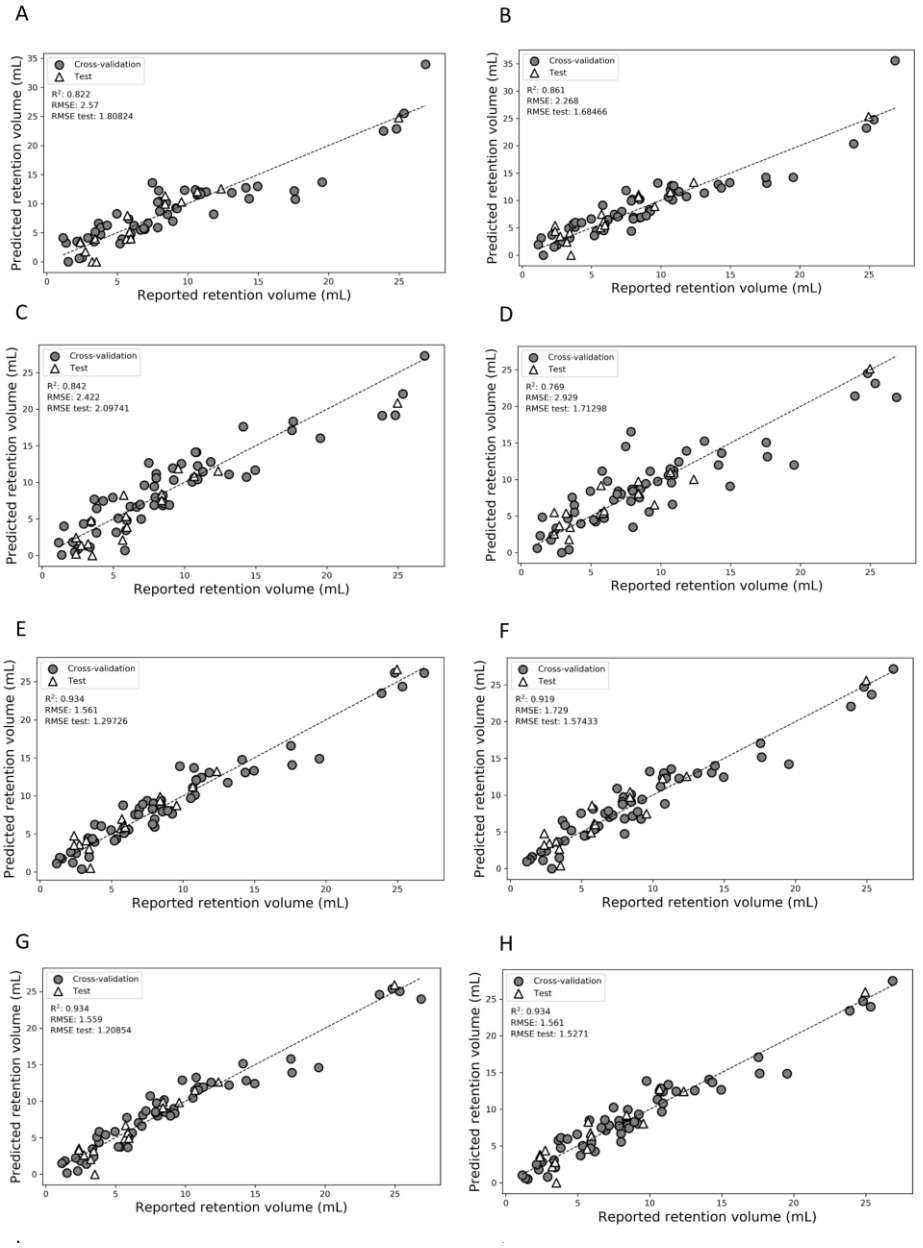
SurfPosEpStdFormal	v	The standard deviation of the positive electrostatic potentials calculated using binary charges of +1, 0 or -1
NSurfNegEpFormal	v	Number of points with a negative electrostatic potential calculated using binary charges of +1, 0 or -1
SurfNegEpMeanFormal	v	The mean of all negative electrostatic potentials calculated using binary charges of +1, 0 or -1
SurfNegEpTrimeanFormal	v	The trimean of all negative electrostatic potentials calculated using binary charges of +1, 0 or -1
SurfNegEpMedianFormal	v	The median of all negative electrostatic potentials calculated using binary charges of +1, 0 or -1
SurfNegEpsumFormal	v	The sum of all negative electrostatic potentials calculated using binary charges of +1, 0 or -1
SurfNegEpFracFormal	-	The fraction of points with a negative electrostatic potential, NsurfNegEp/NsurfPoints calculated using binary charges of +1, 0 or -1
SurfNegEpStdFormal	v	The standard deviation of the negative electrostatic potentials calculated using binary charges of +1, 0 or -1
SurfEpMaxAverage	v	The maximum observed electrostatic potential calculated using charges ranging between 1 and -1
SurfEpMeanAverage	v	The mean of all electrostatic potentials calculated using charges ranging between 1 and -1
SurfEpTrimeanAverage	v	The trimean of all electrostatic potentials calculated using charges ranging between 1 and -1
SurfEpminAverage	v	The minimum observed electrostatic potential calculated using charges ranging between 1 and -1
SurfEpMedianAverage	v	The median of all electrostatic potentials calculated using charges ranging between 1 and -1
SurfEpSumAverage	v	The sum of all electrostatic potentials calculated using charges ranging between 1 and -1
SurfEpStdAverage	v	The standard deviation of the electrostatic potentials calculated using charges ranging between 1 and -1
NSurfPosEpAverage	v	Number of points with a positive electrostatic potential calculated using charges ranging between 1 and -1
SurfPosEpMeanAverage	v	The mean of all positive electrostatic potentials calculated using charges ranging between 1 and -1
SurfPosEpTrimeanAverage	v	The trimean of all positive electrostatic potentials calculated using charges ranging between 1 and -1
SurfPosEpMedianAverage	v	The median of all positive electrostatic potentials calculated using charges ranging between 1 and -1
SurfPosEpsumAverage	v	The sum of all positive electrostatic potentials calculated using charges ranging between 1 and -1
SurfPosEpFracAverage	-	The fraction of points with a positive electrostatic potential, NsurfPosEp/NsurfPoints calculated using charges ranging between 1 and -1

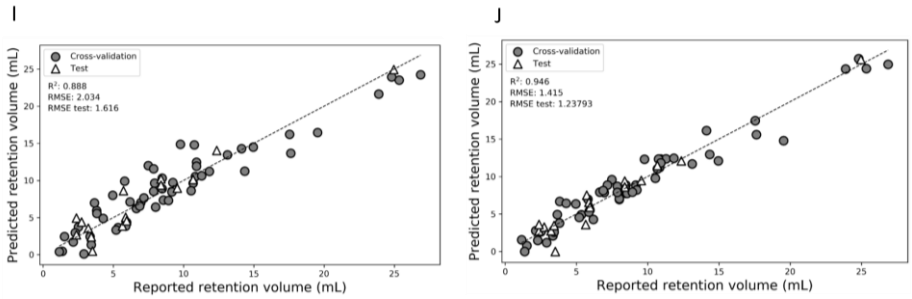
SurfPosEpStdAverage	v	The standard deviation of the positive electrostatic potentials calculated using charges ranging between 1 and -1
NSurfNegEpAverage	v	Number of points with a negative electrostatic potential calculated using charges ranging between 1 and -1
SurfNegEpMeanAverage	v	The mean of all negative electrostatic potentials calculated using charges ranging between 1 and -1
SurfNegEpTrimeanAverage	v	The trimean of all negative electrostatic potentials calculated using charges ranging between 1 and -1
SurfNegEpMedianAverage	v	The median of all negative electrostatic potentials calculated using charges ranging between 1 and -1
SurfNegEpsumAverage	v	The sum of all negative electrostatic potentials calculated using charges ranging between 1 and -1
SurfNegEpFracAverage	-	The fraction of points with a negative electrostatic potential, NsurfNegEp/NsurfPoints calculated using charges ranging between 1 and -1
SurfNegEpStdAverage	v	The standard deviation of the negative electrostatic potentials calculated using charges ranging between 1 and -1
SurfMhpMax	-	The maximum observed hydrophobicity potential
SurfMhpMean	-	The mean of all hydrophobicity potentials
SurfMhpTrimean	-	The trimean of all hydrophobicity potentials
SurfMhpmin	-	The minimum observed hydrophobicity potential
SurfMhpMedian	-	The median of all hydrophobicity potentials
SurfMhpSum	-	The sum of all hydrophobicity potentials
SurfMhpStd	-	The standard deviation of the hydrophobicity potentials
NSurfPosMhp	-	Number of points with a positive hydrophobicity potential
SurfPosMhpMean	-	The mean of all positive hydrophobicity potentials
SurfPosMhpTrimean	-	The trimean of positive hydrophobicity electrostatic potentials
SurfPosMhpsum	-	The sum of all positive hydrophobicity potentials
SurfPosMhpFrac	-	The fraction of points with a positive hydrophobicity potential, NsurfPosMhp/NsurfPoints
SurfPosMhpStd	-	The standard deviation of the positive hydrophobicity potentials
NSurfNegMhp	-	Number of points with a negative hydrophobicity potential
SurfNegMhpMean	-	The mean of all negative hydrophobicity potentials
SurfNegMhpTrimean	-	The trimean of all negative hydrophobicity potentials
SurfNegMhpsum	-	The sum of all negative hydrophobicity potentials
SurfNegMhpFrac	-	The fraction of points with a negative hydrophobicity potential, NsurfNegMhp/NsurfPoints
SurfNegMhpStd	-	The standard deviation of the negative hydrophobicity potentials

ShellEpMaxFormal	v	The maximum observed shell electrostatic potential calculated using binary charges of +1, 0 or -1
ShellEpMeanFormal	v	The mean of all shell electrostatic potentials calculated using binary charges of +1, 0 or -1
ShellEpTrimeanFormal	v	The trimean of all shell electrostatic potentials calculated using binary charges of +1, 0 or -1
ShellEpminFormal	v	The minimum observed shell electrostatic potential calculated using binary charges of +1, 0 or -1
ShellEpMedianFormal	v	The median of all shell electrostatic potentials calculated using binary charges of +1, 0 or -1
ShellEpSumFormal	v	The sum of all shell electrostatic potentials calculated using binary charges of +1, 0 or -1
ShellEpStdFormal	v	The standard deviation of the shell electrostatic potentials calculated using binary charges of +1, 0 or -1
NShellPosEpFormal	v	Number of points with a positive shell electrostatic potential calculated using binary charges of +1, 0 or -1
ShellPosEpMeanFormal	v	The mean of all positive shell electrostatic potentials calculated using binary charges of +1, 0 or -1
ShellPosEpTrimeanFormal	v	The trimean of all positive shell electrostatic potentials calculated using binary charges of +1, 0 or -1
ShellPosEpMedianFormal	v	The median of all positive shell electrostatic potentials calculated using binary charges of +1, 0 or -1
ShellPosEpsumFormal	v	The sum of all positive shell electrostatic potentials calculated using binary charges of +1, 0 or -1
ShellPosEpFracFormal	-	The fraction of points with a positive shell electrostatic potential, NshellPosEp/120 calculated using binary charges of +1, 0 or -1
ShellPosEpStdFormal	v	The standard deviation of the positive shell electrostatic potentials calculated using binary charges of +1, 0 or -1
NShellNegEpFormal	v	Number of points with a negative shell electrostatic potential calculated using binary charges of +1, 0 or -1
ShellNegEpMeanFormal	v	The mean of all negative shell electrostatic potentials calculated using binary charges of +1, 0 or -1
ShellNegEpTrimeanFormal	v	The trimean of all negative shell electrostatic potentials calculated using binary charges of +1, 0 or -1
ShellNegEpMedianFormal	v	The median of all negative shell electrostatic potentials calculated using binary charges of +1, 0 or -1
ShellNegEpsumFormal	v	The sum of all negative shell electrostatic potentials calculated using binary charges of +1, 0 or -1
ShellNegEpFracFormal	-	The fraction of points with a negative shellelectrostatic potential, NShellNegEp/120 calculated using binary charges of +1, 0 or -1
ShellNegEpStdFormal	v	The standard deviation of the negative electrostatic potentials calculated using binary charges of +1, 0 or -1



Supplemental Figure S2.1: *Model results of model 1 while removing each feature. A) contains the model results in absence of the negative surface EP median. B) contains the model results in absence of the number of surface points with positive EP. C) contains the model results in absence of the valine surface fraction.*


A
Cross-validation
Test
R²: 0.822
RMSE: 2.57
RMSE test: 1.80824
Predicted retention volume (mL)
Reported retention volume (mL)
B
Cross-validation
Test
R²: 0.861
RMSE: 2.268
RMSE test: 1.68466
Predicted retention volume (mL)
Reported retention volume (mL)
C
Cross-validation
Test
R²: 0.842
RMSE: 2.422
RMSE test: 2.09741
Predicted retention volume (mL)
Reported retention volume (mL)
D
Cross-validation
Test
R²: 0.769
RMSE: 2.929
RMSE test: 1.71298
Predicted retention volume (mL)
Reported retention volume (mL)
E
Cross-validation
Test
R²: 0.934
RMSE: 1.561
RMSE test: 1.29726
Predicted retention volume (mL)
Reported retention volume (mL)
F
Cross-validation
Test
R²: 0.919
RMSE: 1.729
RMSE test: 1.57433
Predicted retention volume (mL)
Reported retention volume (mL)
G
Cross-validation
Test
R²: 0.934
RMSE: 1.559
RMSE test: 1.20854
Predicted retention volume (mL)
Reported retention volume (mL)
H
Cross-validation
Test
R²: 0.934
RMSE: 1.561
RMSE test: 1.5271
Predicted retention volume (mL)
Reported retention volume (mL)



Supplemental Figure S2.2: *Model results of model 2 while removing each feature. A) contains the model results in absence of the minimum surface EP. B) contains the model results in absence of the total charge. C) contains the model results in absence of the dipole vector length. D) contains the model results in absence of the isoelectric point. E) contains the model results in absence of the standard deviation of positive shell projections. F) contains the model results in absence of the lysine surface fraction. G) contains the model results in absence of the mean negative surface EP. H) contains the model results in absence of the standard deviation of negative surface hydrophobicity. I) contains the model results in absence of the cysteine surface fraction. J) contains the model results in absence of the surface shape max.*

2.6.1 Supplemental discussion

Due to the increased number of features compared to the previous dataset, 10 opposed to three, interpreting the final model is more challenging. Six of the 10 selected features are directly related to the protein charge and are inherently interconnected. The feature with the highest regression coefficient of 31.24, and therefore deemed most important, is the minimum surface EP. The positive coefficient indicates that an increase in minimum surface EP leads to a higher retention volume, which is in line with the mode of action of the cation exchange resin. The total charge is the second most important feature with a regression coefficient of -27.77. This indicates that proteins with a higher total charge have lower retention volumes. Considering the dataset to be retention volumes for the cation exchange resin SP Sepharose HP, a negative correlation with the total charge is counter intuitive. This correlation might not indicate a direct causation with the retention volume, but rather that the total charge might compensate for other charge related features, as there is collinearity between the charge related features. To directly assess the importance of the feature, the permutation model results in a reduced cross-validated R^2 of 0.861. The permutation model for the minimum surface EP resulted in a greater decrease in performance (cross-validated R^2 of 0.822). This indicates that the total charge is indeed less important for the final model compared to EP.

The dipole vector length has a regression coefficient of 20.72. The high positive regression coefficient indicates the importance of charge polarization, and that proteins elute later with more uneven charge distribution. The isoelectric point is the next charge-related feature with a regression coefficient of 12.02. This feature is unaffected by pH as it represents the pH at which the protein is neutrally charged. Interestingly, even though the feature has only the fourth highest coefficient, permutation of the feature results in a permutation model

with the lowest R^2 of 0.769 (Supplemental Figure S2.2D). This feature has a low cross correlation with the other features, indicating that less compensation is possible with the remaining data. The fifth highest regression coefficient is 11.07 for the standard deviation of the shell projections with a positive value. This feature represents the spread of the plane projections with an overall positive EP. An increase in standard deviation indicates a greater spread of positive values. Although the correlation coefficient is similar to that of the isoelectric point, permutation of the feature results in only a 0.02 decrease in performance, resulting in a R^2 of 0.934 (Supplemental Figure S2.2E). The minor decrease in performance can be explained by the Pearson correlation coefficient of the fifth feature to the minimum surface EP, total charge, and the dipole vector length, which are 0.54, 0.76, and 0.69 respectively (data not shown). As a result of this relatively high correlation, the remaining features can compensate for the missing feature, minimizing the loss of model performance. Apart from the cysteine surface fraction, the permutation of the four remaining features with the lower regression coefficients results in a loss of predictive capability similar to the standard deviation of positive shell projections, resulting in a cross-validated R^2 range of 0.948 - 0.922 (Supplemental Figure S2.2F, G, H, J). These features are therefore important to finetune the model but are difficult to interpret due to the low level of correlation between the single feature and the protein retention volume. Permutation of the cysteine surface fraction feature from the model yielded a reduced cross-validated R^2 of 0.895 (Supplemental figure S2.2I). Since cysteine residues can act as a hydrogen bond donor, they can potentially interact with the sulphopropyl active groups on the resin, however no correlation was found for the single feature and retention volume.

The test set shows that 1V9E is predicted with low accuracy at the prediction limit of 0 mL. The inability to predict an accurate retention

volume is due to a 2.5 fold decrease in dipole moment for 1V9E when moving from pH 5 to pH 6 while the other proteins in the dataset only show a maximum decrease of 1.5 fold. Due to the importance of the dipole vector length in the model, the reduction in protein resin affinity is overestimated. Another feature which greatly affects the estimation of the retention volume of 1V9E_6 is the standard deviation of positive shell projections. This feature was found to be outside of the value range observed in the training set. This highlights the importance of outlier identification.

Chapter 3

From protein structure to an optimized chromatographic capture step using multiscale modeling

Published as:

Keulen, D., Neijenhuis, T.*, Lazopoulou, A., Disela, R., Geldhof, G., Le Bussy, O., Klijn, M., & Ottens, M. (2025). From protein structure to an optimized chromatographic capture step using multiscale modeling. Biotechnology Progress, 41(1), e3505.*

**Authors contributed equally*

Abstract

Optimizing a biopharmaceutical chromatographic purification process is currently the greatest challenge during process development. A lack of process understanding calls for extensive experimental efforts in pursuit of an optimal process. In silico techniques, such as mechanistic or data driven modeling, enhance the understanding, allowing more cost-effective and time efficient process optimization. This work presents a modeling strategy integrating quantitative structure property relationship (QSPR) models and chromatographic mechanistic models (MM) to optimize a cation exchange (CEX) capture step limiting experiments. In QSPR, structural characteristics obtained from the protein structure are used to describe physicochemical behavior. This QSPR information can be applied in MM to predict the chromatogram and optimize the entire process. To validate this approach, retention profiles of six proteins were determined experimentally from mixtures, at different pH (3.5, 4.3, 5.0, 7.0). Four proteins at different pH's were used to train QSPR models predicting the retention volumes and characteristic charge, subsequently the equilibrium constant was determined. For an unseen protein knowing only the protein structure, the retention peak difference between the modeled and experimental peaks was 0.2% relative to the gradient length (60 column volume). Next, the CEX capture step was optimized, demonstrating a consistent result in both the experimental and QSPR-based methods. The impact of model parameter confidence on the final optimization revealed two viable process conditions, one of which is similar to the optimization achieved using experimentally obtained parameters. The multiscale modeling approach reduces the required experimental effort by identification of initial process conditions which can be optimized.

3.1 Introduction

Over the past years, the biopharmaceutical industry has experienced substantial growth, with protein-based biopharmaceuticals (e.g., monoclonal antibodies (mAbs) and protein subunit vaccines) being a significant part of the industry.[1] As a consequence, the biopharmaceutical industry endeavors to accelerate process development with the primary goal to deliver biopharmaceuticals at the earliest possible time, pushing the competitive market.[2] Moreover, the competition even intensified more due to the emerging field of biosimilars.[3,4] The biopharmaceutical sector requires therefore innovative approaches to advance process development, while ensuring product quality and stability.[5] Especially the downstream process is the major cost driver of the overall manufacturing costs, demanding an efficient and cost-effective process. To achieve very high product purities, chromatography is currently the most essential but also the most costly technique.[6]

In silico techniques, such as mechanistic or data-driven modeling, can be of great merit for process development. These methods allow for increased process understanding while reducing experimental effort and/or use of critical sample material and decreasing process development times.[7,8] Within the next years, modeling techniques will become more essential for biopharmaceutical industry. Specifically for Industry 4.0 that aims to digitalize the entire manufacturing process.[9–12] Moreover, increased process understanding and process and product quality control are in agreement with the Quality-by-Design (QbD) guidelines.[13–16] Identifying the operating window of the critical process parameters (CPP) is an essential part to guarantee process' stability. Currently, these operating windows are determined with expensive and time-consuming wet-lab Design-of-Experiments (DoE). Chromatographic mechanistic models (MM) attempt to describe the

chromatographic process in silico and could be an inexpensive and fast alternative to determine the CPP operating window. Over the past years, the industry has been gradually adopting chromatographic MM, with ongoing advancement being made in determining the essential input parameters.[17–20] In the future, the ultimate objective is to determine adsorption isotherm for complex mixtures more easily.[21,22] Progress in utilizing mass spectrometry data could play a crucial role in achieving this goal.[23] However, at this moment determining adsorption isotherm parameters for the MM remains a bottleneck for industrial application, mainly due to time and material limitations especially in the early phase of downstream process development.[24] Quantitative Structure Property Relationships (QSPR) modeling could be an in silico alternative to experimentally determining the adsorption isotherm parameters. QSPR aims to correlate physicochemical properties with specific behavior, such as chromatographic retention time.[25] These physicochemical properties are calculated from protein structure models that describe the position of each atom. Combining MM with QSPR and optimization tools could pave the way for a holistic modeling approach/workflow.

In 2001, Mazza et al. introduced a QSPR model for predicting protein retention times for ion exchange chromatography.[25] Their approach involved feature calculation using the proprietary software platform MOE a genetic algorithm for feature selection for the training of a partial least squares model.[26,27] As a result, several follow-up studies applied QSPR models to different modes of chromatography/type of chromatography resins, using support vector machine regression methods, and including pH effects.[28–33] Malmquist et al. developed an additional set of protein descriptors that are pH-dependent and based on electrostatic and hydrophobic properties.[34] Moreover, several studies considered the crucial binding orientations within protein-resin binding affinities in their QSPR models.[35–37] In recent years, QSPR has

been applied to more complex proteins, such as Fabs and mAbs, showing the growing interest from industry and the added value of these models.[24,38,39] Robinson et al. showed the potential of QSPR models for in silico resin screening of six chromatographic systems applied to Fabs.[38] While Saleh et al. built QSPR models using 21 mAbs variants to predict the adsorption isotherm parameters, the equilibrium constant and the characteristic charge, which were subsequently applied to the MM and able to predict the cation exchange chromatography (CEX) step.[24] Their study shows promising capabilities of a multiscale model to simulate different process conditions without the need for wet-lab experiments. Several software packages are available to calculate the protein descriptors that are needed for QSPR modeling, an overview of these software packages has been provided elsewhere.[40,41] Most software tools are only available via webservers or commercially, lacking source code availability. Therefore, Neijenhuis et al. have recently published an open-source QSPR software tool, which has also been used in this work.[42]

Most research on QSPR modeling either developed protein descriptors or applied existing protein descriptors for their QSPR model with the aim to increase the protein-behavior understanding via retention prediction.[31,34,38,39,43] Additionally, other research also applied the predicted QSPR parameters to MM and validated the predicted chromatographic process from a protein structure/sequence.[24,30,32] So far, no research has shown the ability of QSPR models in combination with MM to optimize a chromatographic process step without any need for protein material. Moreover, the influence of the accuracy of the predicted QSPR-parameters on an optimized process has not yet been evaluated.

This study presents a general multiscale modeling strategy that integrates QSPR and chromatographic MM to optimize a CEX capture

step. We were able to simulate and validate a CEX step only using the protein structure. Subsequently, we compared the uncertainty of the experimentally determined and predicted parameters on the final optimization outcome. An overview of the experimental-based and QSPR-based strategy is shown in Figure 3.1. This strategy can be used to determine the operating window of CPPs in early-stage process development, showing the potential applicability for industry. Combining these modeling techniques together with optimization software reduces the experimentally effort overall process development time significantly. Previous research mostly used pure components to perform the linear gradient experiments (LGE), however the availability of pure components is limited in biopharmaceutical industry. Therefore, performing LGE with complex protein mixtures would offer significant advantages. So far, only Buyel et al. applied QSPR modeling to a crude mixture of plant extracts to predict elution conditions for ion exchange and mixed mode chromatography separations.[33] Here, we performed LGE for five different gradient lengths and four pHs applied to two mixtures of each three proteins. Performing the experiments with protein mixtures instead of each protein individually, reduces the total LGE from 30 to 10 experiments. We developed QSPR models for predicting the retention volumes and characteristic charges. These predicted QSPR parameters were used to obtain the equilibrium constants. The multiscale model was validated for an unseen protein, which was excluded from the QSPR training and testing data. Finally, we compared the influence of parameter uncertainties on the optimization outcome by using experimental and QSPR predicted parameters.

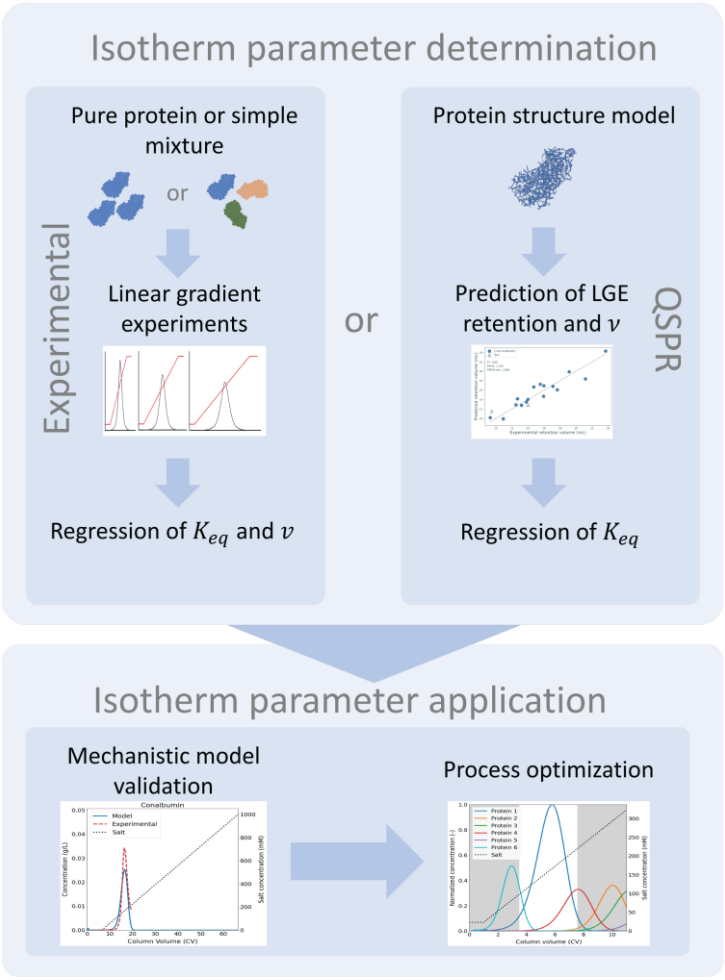


Figure 3.1: *Overview of the experimental-based method and the QSPR-based method. Both methods can be used to determine the adsorption isotherm parameters that can be used in the mechanistic model for process optimization purposes. The equilibrium constant is denoted by K_{eq} and the stoichiometric coefficient of salt counter ions with ν.*

3.2 Materials & Methods

3.2.1 Materials & Equipment

A 1-mL CEX column of HiTrap SP FF (Cytiva Life Sciences, USA) was used for the preparative column experiments. For the analytical size exclusion chromatography – ultra performance liquid chromatography (SEC-UPLC), an ACQUITY UPLC Protein BEH SEC 200 Å column (Waters

Corporation, USA) was used, protected with a prior/foregoing ACQUITY UPLC Protein BEH SEC guard 200 Å column (Waters Corporation, USA).

The following proteins were purchased from Sigma-Aldrich, USA: bovine serum albumin (BSA), lysozyme, cytochrome c, chymotrypsinogen A from bovine pancreas, and conalbumin. Ribonuclease pancreatic (RNAse) was purchased from Roche Diagnostics GmbH, Germany. Dextran (DXT1740K) (American Polymer Standards Corporation, USA) was used for column characterization.

The buffers were prepared with Milli-Q water and adjusted to the desired pH using either 0.5 M sodium hydroxide or 1 M hydrochloric acid. The buffers were filtered to remove undissolved salts, 0.2 μm pore-size hollow fiber MediaKap (Repligen, USA) filter for UPLC buffers and a 0.2 μm Membrane Disc Filter (Pall corporation, USA) for ÄKTA buffers. Moreover, all buffers were degassed for 20 minutes using an ultrasonic bath (Branson Ultrasonics, USA) to prevent introducing air bubbles into the column. The protein mixture was filtered using a 0.2 μm Whatman Puradisc FP 30 mm (GE Healthcare Life Sciences, USA).

3.2.2 Linear gradient column experiments

LGE were conducted at various pH values (pH 3.5, 4.3, 5.0, and 7.0) for five gradient lengths: 20, 30, 40, 60, and 80 column volumes (CV). For every pH a different running buffer was needed, citric acid monohydrate (pH 3.5, 20 mM), sodium acetate trihydrate (pH 4.3 and 5.0, 50 mM), and sodium phosphate monobasic dihydrate (pH 7.0, 50 mM). The elution buffer is the same as the running buffer for that respective pH with the addition of 1 M sodium chloride. The pH-values were selected to theoretically favor a positive net charge for most proteins and therefore anticipate their binding to the CEX resin. The chromatographic column experiments were performed on an ÄKTA pure system (Cytiva Life Sciences, USA) with UNICORN version 7.5 software, with a flowrate of 1 mL/min, and measuring UV absorbance

at 230, 280, and 400 nm wavelength. The column characteristics are given in Table 3.1, more information on the characterization methods can be found in the Supplemental Methods. During the chromatography runs, 1 mL samples were collected using a fraction collector. These samples were additionally analyzed with a Dionex UPLC system using Chromeleon Chromatography Data System version 7 software, measuring UV absorbance at 230, 280, and 400 nm wavelength. The UPLC-running buffer was a 100 mM sodium phosphate monobasic dihydrate with a pH of 6.8. A flowrate of 0.1 mL/min and analysis time of 40 minutes was applied. The SEC-UPLC analysis enabled the identification of the peaks obtained during the LGE's with their corresponding proteins. However, the protein mixture was divided into two groups, as some proteins with similar characteristics were indistinguishable in the SEC-UPLC analysis. Group one consisted of RNase, cytochrome c, conalbumin, and group two of chymotrypsinogen, lysozyme, and albumin. Both multi-component mixtures contained 0.8 mg/mL of each protein.

First, the column was equilibrated with 5 CV running buffer, followed by a 300 μL sample injection using a 10 mL Superloop (Cytiva Life Sciences, USA). After the sample injection, unretained proteins were removed by washing the column for 5 CV using the running buffer. Subsequently, a gradient elution was performed from 0 (running buffer) to 1 M sodium chloride (elution buffer). The proteins in the collected fractions were identified with the SEC-UPLC analytical method. Though, it is expected that the elution order of the proteins remains the same and therefore, only the fractions of two gradients for each pH were analyzed with SEC-UPLC. For each fraction analysis, 5 μL sample was injected.

Table 3.1: *Column characteristics for HiTrap SP FF column.*

Parameter	Value	Unit
Column volume	0.97	mL
Column diameter[a]	0.70	cm
Bed height[a]	2.50	cm
Maximum pressure[a]	2.0	MPa
Ionic capacity[b]	800	mM
Particle size[a]	90	μm
Pore diameter[c]	54	nm
Cross sectional area	0.39	cm^2
System dead volume (V_{dead})	0.34	mL
Total porosity (ε_t)	0.918	-
Extraparticle porosity (ε_b)	0.298	-
Intraparticle porosity (ε_p)	0.887	-
System dwell volume (V_{dwell})	1.09	mL

[a]Manufacturer, [b]Osberghaus et al.[44], [3]Hagemann et al.[45]

Table 3.2: *Overview of the protein characteristics and the protein data bank (PDB) entry used for calculations.*

Protein	PDB names	Mass (kDa)	Estimated Isoelectric point*
Conalbumin	1OVT	75.83	6.62
Albumin	6QS9	66.43	5.49
Chymotrypsinogen	2CGA	25.67	8.13
Lysozyme	1GWD	14.31	9.20
Ribonuclease	1RNC	13.69	8.29
Cytochrome C	6FF5	12.33	9.60

* Estimations were performed using the open-source QSPR tool

3.2.3 Chromatographic mechanistic model

The chromatographic MM from previous work was used to describe the dynamic adsorption behavior during the chromatographic separation process.[46] This employed MM is a combination of the equilibrium transport dispersive model combined with the linear driving force model as

$$\frac{\partial C_i}{\partial t} + F\frac{\partial q_i}{\partial t} = -u\frac{\partial C_i}{\partial x} + D_{L,i}\frac{\partial^2 C_i}{\partial x^2}, \tag{3.1}$$

$$\frac{\partial q_i}{\partial t} = k_{ov,i}\left(C_i - C^*_{eq,i}\right), \tag{3.2}$$

$$k_{ov,i} = \left[\frac{d_p}{6k_{f,i}} + \frac{d_p^2}{60\varepsilon_p D_{p,i}}\right]^{-1}, \tag{3.3}$$

where the concentration in the liquid phase is represented by C_i and in the solid phase with q_i, in which subscript i denotes the protein component. The liquid phase concentration at equilibrium is denoted by $C^*_{eq,i}$. The phase ratio is equal to $F = (1 - \varepsilon_b)/\varepsilon_b$, where ε_b is the bed porosity. Time and space are indicated by t and x respectively. u is the mobile phase interstitial velocity and D_L is the axial dispersion coefficient. The overall mass transfer coefficient, $k_{ov,i}$, is defined as the combined result of both the separate film mass transfer resistance and the mass transfer resistance within the pores.[47] In equation 3.3, the particle diameter is denoted by d_p, the intraparticle porosity by ε_p, and the effective pore diffusivity coefficient by D_p. The effective pore diffusivity is described according to Fick's law and calculated as

$$D_p = \frac{\varepsilon_p D_f}{\tau}\psi, \qquad (3.4)$$

where τ is the tortuosity and ψ the diffusional hindrance parameter determined by Brenner and Gaydos.[48] The free diffusivity (D_f) has been calculated using the Young correlation for globular proteins.[49] The film mass transfer resistance is $k_f = D_f Sh/d_p$, in which Sh is the Sherwood number. The Method of Lines was applied using a fourth-order central difference scheme for both first and second-order derivatives to spatially discretize the partial differential equation into a set of ordinary differential equations. The Livermore Solver for Ordinary Differential Equations (LSODA) algorithm, part of the scipy.integrate package, is employed to solve the Ordinary Differential Equations (ODEs), automatically transitioning between the nonstiff Adams method and the stiff BDF method.[50] Additional details regarding the MM can be found in a prior study.[51]

We employed the linear multicomponent mixed-mode isotherm, developed by Nfor et al., to determine the equilibrium liquid phase concentration as[52]

$$\frac{q_i}{C^*_{eq,i}} = K_{eq,i}\Lambda^{(v_i+n_i)}(z_s c_s)^{-v_i} c_v{}^{-n_i}\gamma_i, \tag{3.5}$$

where the equilibrium constant, $K_{eq,i}$, quantifies the strength of the interaction between the protein and the stationary phase. Λ is the ligand density or ionic capacity of the concerned resin, z_s is the charge of the salt counter ion, c_s is the salt concentration in the liquid phase, and c_v is the molarity of the solution in the pore volume. The stoichiometric coefficient of salt counter ions is denoted by v_i, determined by $v_i = z_p/z_s$, in which z_p is the effective binding charge of the protein. For monovalent counter-ions, the charge equals one ($z_s = 1$), for example Na^+ in the sodium chloride elution buffer. In this work, only the ion-exchange part of the mixed-mode isotherm is used, therefore hydrophobic interaction stoichiometric coefficient (n_i) will be equal to zero. The activity coefficient (γ) of the protein solution can be calculated as

$$\gamma_i = e^{K_{s,i}c_s + K_{p,i}C_i}, \tag{3.6}$$

where K_s is the salt-protein interaction constant and K_p the protein-protein interaction constant. In the linear range of adsorption, the protein concentrations are low and protein-protein interactions are expected to be minimal, therefore K_p becomes insignificant and can be neglected.[53,54] Because of the low salting-out effects, the K_s also becomes negligible.[53] Subsequently, incorporating the assumptions for this work, the linear multicomponent mixed-mode isotherm is reformulated as

$$\frac{q_i}{C^*_{eq,i}} = K_{eq,i}\Lambda^{v_i}(z_s c_s)^{-v_i}. \tag{3.7}$$

3.2.4 Procedure to determine adsorption isotherm parameters

The peak retention volumes were obtained from the LGE's for each gradient length and at each pH. The initial retention volumes ($V_{R,0}$) were corrected to be aligned with the elution gradients as follows:

$$V_R = V_{R,0} - V_m - V_D - \frac{V_{inj}}{2}, \tag{3.8}$$

where V_R is the peak retention volume, V_m is the column void volume, determined by dextran pulse, and V_D is the system's dwell and dead volume, details can be found in the Supplemental Methods.[55] The injection volume is denoted by V_{inj}, half of this volume needs to be subtracted.[56]

The regression formula of Shukla et al.[57], adapted from Parente and Wetlaufer[55], was used to obtain the equilibrium constant (K_{eq}) and the characteristic charge (v) for each protein as follows:

$$V_R = \left(\left(C_{s,0}^{v+1} + \frac{V_m K_{eq} F \Lambda^v (v+1) * (C_{s,f} - C_{s,0})}{V_G} \right)^{\frac{1}{v+1}} - C_{s,0} \right) * \frac{V_G}{C_{s,f} - C_{s,0}}, \tag{3.9}$$

where V_G is the gradient length. $C_{s,0}$ and $C_{s,f}$ are the initial and final salt concentration during the elution respectively. As no separate pore balance is considered in the chromatographic MM, the column phase ratio is considered the same $F = (1 - \varepsilon_b)/\varepsilon_b$. To validate the regression and accordingly the MM, the experimental data of 60 CV is left out during the regression.

The initial peak retention volumes ($V_{R,0}$) were determined using the function find_peaks of the signal module from the *SciPy* library. The regression was performed using the curve_fit function of the optimize module from the *SciPy* library.

Specifically at pH 5.0, Cytochrome c and RNase co-eluted. The absorbance and respective calibration lines of cytochrome c at 400 and

280 nm were used to trace back the RNase peak. Moreover, at pH 4.3, albumin and chymotrypsinogen co-eluted. However, from the SEC-UPLC analysis it was observed that albumin eluted later compared to the UV peak detected by the UNICORN software. Therefore, the peak retention volumes for albumin at pH 4.3 were determined by analyzing the concentrations by SEC-UPLC in the 1 mL fractions obtained from the LGE. Albumin peak areas obtained from the SEC-UPLC were used to fit a third degree polynomial function representing the retention volume as the maximum.

3.2.5 Structure preparation and descriptor calculation

For each protein, the respective models, listed in Table 2, were obtained from the protein data bank[58], specific entry selection was performed based on resolution and coverage. Duplicate chains were removed from each structural model using pdb-tools[59] to yield monomer representations. The side chain pKa of titratable residues were predicted using PROPKA3.0[60] allowing for more accurate charge calculations with respect to pH. Protein features at pH 3.5, 4.3, 5.0 and 7.0 were calculated using our open-source software package prodes, available at https://doi.org/10.5281/zenodo.10369949, using the default settings, only supplying the pKa estimations.[42] Visualization of protein structures was performed using UCSF-Chimera.[61]

3.2.6 QSPR model training

For predicting the protein retention volumes and adsorption isotherm parameters, Multi Linear Regression (MLR) models were trained. The prediction of conalbumin was removed from the dataset prior to train-test splitting to eliminate all bias. To find an accurate predictive MLR model, series of filter thresholds were screened by testing a range of feature-feature correlation filters (Pearson correlations of 0.8, 0.9 and 0.99). Followed by feature-observation correlations filtering, maintaining a predefined percentage of features (10% to 100% in 10%

increments). Feature selection was performed by sequential forward selection. Final models were selected based on the cross-validated R^2 and test set RMSE, which should be close to the cross-validation RMSE to ensure model robustness. Feature importance was assessed by analysis of the regression coefficient and the influence of feature permutation. For the prediction of the unknown conalbumin, the confidence interval was calculated as

$$\hat{y}_h \pm t_{\left(1-\frac{\alpha}{2},n-p\right)} \times \sqrt{MSE\ (1 + X_h^T (X^T X)^{-1} X_h)}, \tag{3.10}$$

where $\hat{y}_h$ is the predicted value, $t_{(1-\frac{\alpha}{2},n-p)}$ is the "t-multiplier", X and X_h are the feature matrixes of the training set and the value to be predicted. The mean squared error (MSE) is calculated as

$$MSE = \frac{1}{n}\sum_i^n (y_i - \hat{y}_i)^2, \tag{3.11}$$

3.2.7 Optimization

We evaluated the uncertainty-influence of the regressed and predicted QSPR adsorption isotherm parameters on the final optimization outcome. The equilibrium constant and characteristic charge values were varied between their standard deviation values for 100 samples. These samples were used in the optimization. First, the optimization was formulated and evaluated to be consistent when performing the same optimization multiple times. The global and local objectives were formulated as follows:

$$minf(x) = 2 * (100 - yield(x)) + 1 * \big(100 - purity(x)\big) \tag{3.12}$$

$$s.t.\ \ h(x) = 0 \tag{3.13}$$

$$0 \leq x \leq 1, \tag{3.14}$$

where the objective function, $f(x)$, is minimized. The equality equations, such as the mass balances and equilibrium relations, need to be satisfied (Eq. 3.12). Moreover, variables (x) were normalized for more efficient optimization purposes (Eq. 3.13). Four variables were

chosen namely the initial and final salt concentrations, and the lower and upper cut points. The weights of the objective function were chosen to reflect a capture step to be optimized, hence removing most of the bulk impurities and preventing losing product material.

For the global optimization, the differential_evolution algorithm from the scipy.optimize package was employed, using the Latin hypercube sampling to initialize the population and the maximum number of iterations was 10 with a population size of 23. For the local optimization the Nelder-Mead algorithm was used, with a maximum of 100 iterations. The relative and function tolerances for both global and local optimizations were set to 1e-2. The lower cut point ranges from 1 – 80% on the left of the peak maximum, and the upper cut point from 20 – 99% on the right of the peak maximum. The initial salt concentration varies between 1 – 150 mM, and the final salt concentration between 320 – 800 mM.

3.3 Results & Discussion

3.3.1 Determining the retention volume

LGE's were conducted for two protein mixtures at four pH values (pH 3.5, 4.3, 5.0, and 7.0) and various gradient lengths (20, 30, 40, 60, and 80 CV), as described in the experimental section 2.1. The elution order of the proteins was identified by SEC-UPLC analysis for each pH, to determine single peak retention volumes. The results for the 20 CV LGE are shown in Figure 3.2. As expected, a downward trend for the retention is observed when increasing the pH. No correlation between isoelectric point (PI) and retention was observed. Although cytochrome c, lysozyme, RNase and chymotrypsinogen elute in the order of descending pI (9.60, 9.20, 8.29, and 8.13 respectively) at pH 3.5. No retention volume for albumin and conalbumin (pI of 5.49 and 6.62, respectively) was determined as these proteins did not elute during the

salt gradient, showing greater affinity for the column, which is in accordance with Yang et al..[62]

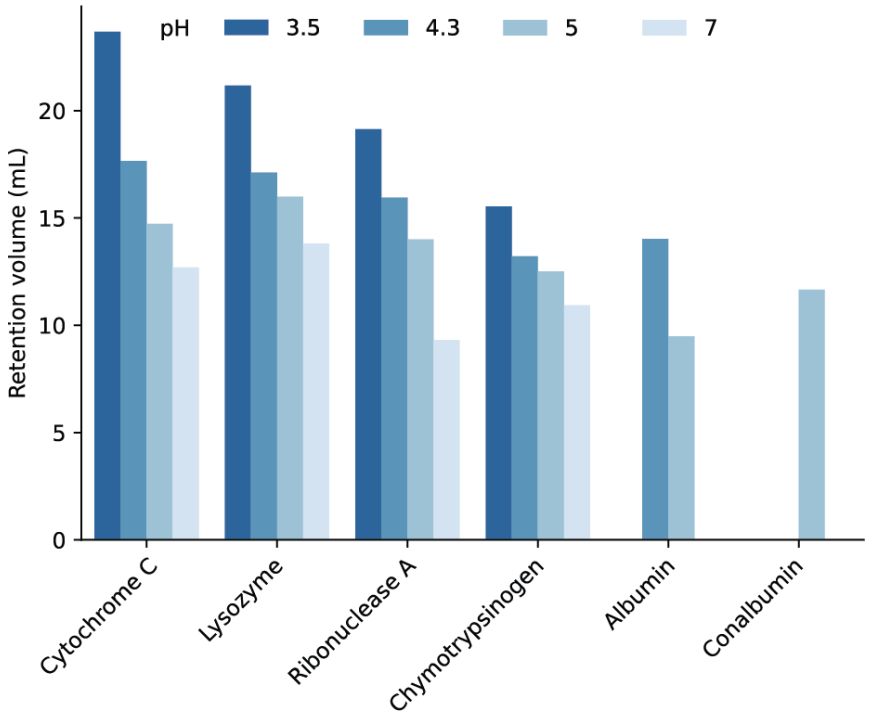


Figure 3.2: *Peak retention volumes (mL, y-axis) given for each protein (x-axis) at each pH (bars). These retention volumes are from the 20 CV gradient length using a HiTrap SP FF column, 1 CV is equal to 0.97 mL.*

3.3.2 Regression of adsorption isotherm parameters

The corrected retention volumes, according to equation 3.7, were used to regress K_{eq} and v using equation 3.8. The regression parameters for each protein at each pH are shown in Table 3.3. The regression plots of each protein at each pH are provided in Supplemental Figures S3.2-S3.5, all fits achieved an R^2 close to one and RMSE values varied between 0.002 and 0.22.A

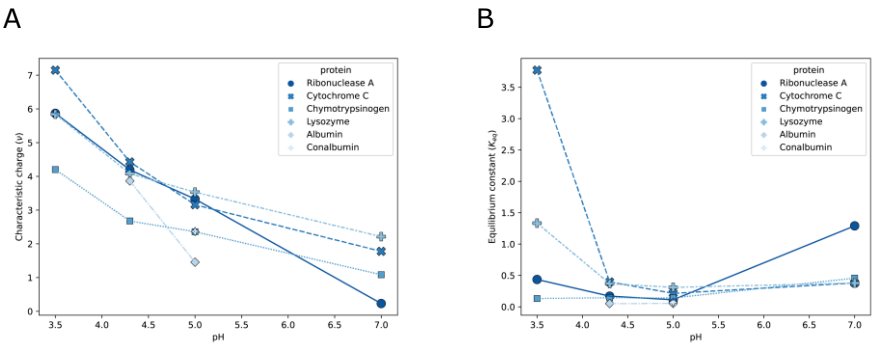


Figure 3.3: *Trendlines between the (a) characteristic charge (y-axis) and (b) the equilibrium constant (y-axis) and the pH value (x-axis) for each protein.*

From Table 3.3 it can observed that the characteristic charge, v, varied between 1% and 6% of the regressed parameter value and the standard deviation values of the equilibrium constant, K_{eq}, varied between 7% and 25%. Figure 3.3a shows that the characteristic charge decreases with increasing pH for all proteins with multiple data points. This is due to the protonation of amino acids, which results in a higher net protein charge at lower pH values. A higher net charge results in more available binding sites to interact with the resin. However, no general trend can be observed between the equilibrium constant and the pH (Figure 3.3b). The equilibrium constant of cytochrome c and lysozyme decreases rapidly from pH 3.5 to pH 4.3. However, at pH 7.0 K_{eq} increases again for RNAse, chymotrypsinogen, lysozyme, and cytochome c (increase of 1.19, 0.26, 0.23, and 0.23 respectively). Similar findings were reported by Yang et al.[62], and the regressed parameters are in the same order of magnitude as reported in literature.[44,62] In general, a higher equilibrium constant indicates a stronger binding affinity towards the resin and therefore eluting later during the salt gradient. The same trend can be observed for the majority of proteins, see Table 3.3 and Figure 3.3. Not all proteins follow this trend, such as chymotrypsinogen, cytochrome c, and lysozyme relative to RNAse (pH 7.0), and albumin relative to chymotrypsinogen (pH 4.3). These proteins elute at a later moment

while having a lower equilibrium constant than the proteins eluting at an earlier moment. Though, the characteristic charge value is higher for these proteins with a lower equilibrium constant. Eventually, it is the combination of these two parameter values that determines the protein's elution moment.

Table 3.3: *Regressed adsorption isotherm parameters, the characteristic charge and the equilibrium constant, for each protein at each pH. The standard deviation is indicated with number after ± sign.*

	pH 3.5	**pH 4.3**	**pH 5.0**	**pH 7.0**
Characteristic charge (ν)				
Conalbumin			2.37 ± 0.12	
Albumin		3.88 ± 0.66	1.46 ± 0.04	
Chymotrypsinogen	4.21 ± 0.22	2.68 ± 0.14	2.36 ± 0.11	1.09 ± 0.003
RNAse	5.88 ± 0.27	4.20 ± 0.26	3.30 ± 0.15	0.23 ± 0.05
Cytochrome C	7.16 ± 0.34	4.44 ± 0.21	3.16 ± 0.14	1.78 ± 0.04
Lysozyme	5.85 ± 0.28	4.09 ± 0.21	3.54 ± 0.15	2.22 ± 0.06
Equilibrium constant (K_{eq})				
Conalbumin			0.071 ± 0.02	
Albumin		0.05 ± 0.04	0.051 ± 0.01	
Chymotrypsinogen	0.13 ± 0.03	0.14 ± 0.03	0.14 ± 0.03	0.44 ± 0.003
RNAse	0.42 ± 0.07	0.16 ± 0.04	0.11 ± 0.02	1.26 ± 0.21
Cytochrome C	3.68 ± 0.28	0.39 ± 0.07	0.21 ± 0.04	0.37 ± 0.03
Lysozyme	1.30 ± 0.16	0.36 ± 0.07	0.30 ± 0.05	0.37 ± 0.04

3.3.3 Chromatographic mechanistic model validation

The chromatographic MM was validated for the gradient length of 60 CV, for pH 5.0 and 7.0. The results of pH 5.0 are shown in Figure 3.4, and of pH 7.0 in the supplemental discussion and Supplemental Figure S3.6. The calibration lines convert the UV absorbance to concentration, these can be found in Supplemental Figures S3.7 and S3.8. As the experiments were performed in two mixtures of each three proteins, only parts of the peaks corresponding to a certain protein were used to avoid pollution of the peak by another component. In this way, the validation of each protein with the MM could be clearly evaluated.

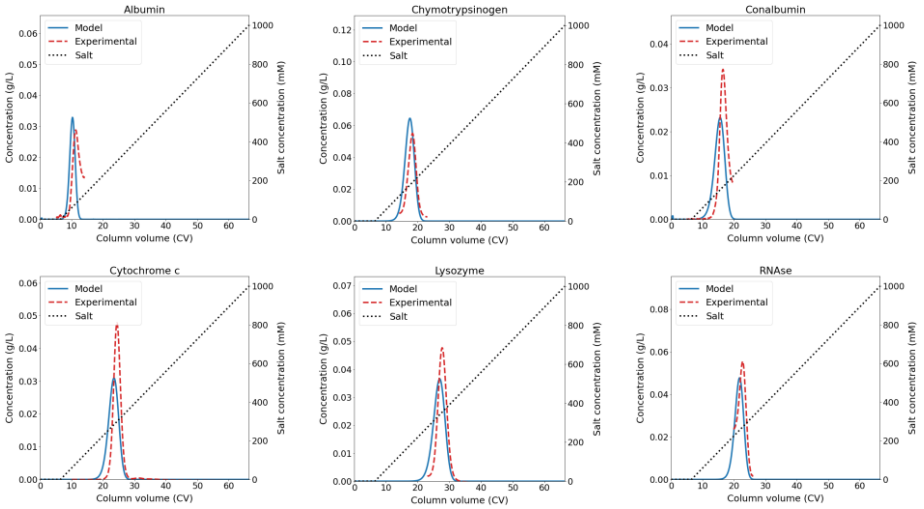


Figure 3.4: *Chromatographic mechanistic model validation for gradient length of 60 CV, equal to 58.2 mL, at a pH of 5.0. The blue line indicates the MM predicted concentration of the protein, while the red dotted line indicates the experimental concentration. The black dotted line indicates the salt concentration. The initial concentrations are albumin: 0.24 mg/mL, chymotrypsinogen: 0.80 mg/mL, conalbumin: 0.31 mg/mL, cytochrome C: 0.41 mg/mL, lysozyme: 0.55 mg/mL, and RNase: 0.56 mg/mL.*

For all proteins at pH 5.0, the maximum retention peak difference is 1.04 CV and the average retention peak difference is 0.92 CV, which is 1.73% and 1.53% with respect to the gradient length (60 CV). In all cases, except for RNAse, the model predicts the start of the elution and the peak maximum earlier than the experimental results. Even though it was not feasible to extract the entire experimental peak in all cases, it was observed that for conalbumin, cytochrome c, and lysozyme the experimental peak seems sharper than the modelled peak. To assess the concentration agreement between the modeled and experimental results, we compared the difference between the peak width at half of the peak maximum and the peak concentration. The maximum peak width difference is 1.14 CV, equal to 1.89% relative to the gradient length (60 CV). The average peak width difference is 0.81 CV, equal to 1.35% relative to the gradient length (60 CV). The average difference in the peak concentration is 0.04 mg/mL, equal to 7.36% relative to the initial concentration. Overall, the mechanistic model, using the regressed adsorption isotherm parameters, can predict the

experimental data sufficiently accurate with a maximum retention peak difference of 1.73%.

3.3.4 QSPR

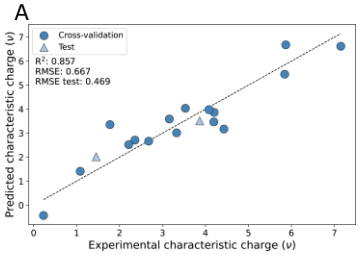
QSPR models relate specific descriptors, calculated from the protein structure, to behavior (e.g., retention). Prediction of the MM parameters, needed for simulation, starting from the protein structure allows for a full in silico optimization framework. From the dataset composed of the six different proteins, conalbumin at pH 5.0 was removed to be used for model verification. This protein and pH was selected because retention volumes for this protein were not obtained for any other pH value. This means that conalbumin at pH 5.0 would be truly unknown for the final predictive model. The remaining 18 datapoints were split into a train and test set, where the test set was comprised of albumin measured at pH 4.3 and 5.0. As retention volumes for albumin were only obtained for pH 4.3 and 5.0, these two data points will validate the models' ability to predict the effect of differences in pH and to predict unseen proteins. The features considered during the QSPR model training, ranging from protein shape to charge and hydrophobicity projections, were calculated using the open-source software prodes.

3.3.4.1 Characteristic charge

For the prediction of the characteristic charge, a MLR was trained. To avoid overfitting, a ratio of five observations to one feature should be maintained.[63] Meaning only a maximum of three features should be used in the model. To select the specific features, a redundancy filter, removing features with a Pearson correlation of >0.99 to other features, was applied. A second filter step was performed removing 40% of the features with lowest correlation to the characteristic charge. From the remaining features, sequential forward selection was performed to select the best features. A model with high accuracy

(cross-validated R^2 of 0.86 and RMSE of 0.67) was obtained using only two features (Figure 3.5). As would be expected, the most important feature was related to the electrostatic potential (EP) of the protein surface. More specifically, the maximal found surface EP. The regression coefficient of this feature was found to be 8 and permutation of the feature would result in a model not capable of predicting ν (Figure 3.5B). The second feature that was selected is the trimean of the negative hydrophobicity potential. This feature is less important as the regression coefficient is 1.5 and permutation results in a model with a cross-validated R^2 of 0.8. The positive regression coefficient for the second feature suggests that increasing the hydrophilicity reduces the characteristic charge. There is the possibility however, that this feature captures the titratable amino acid content on the surface, as amino acids contributing to a negative hydrophobicity are predominantly titratable. At this point we have been unable to confirm this.

A



B

Feature	Coefficient	Permutation CV R^2
Intercept	-1.26	
Maximum surface EP	8.09	0.52
Trimean of negative hydrophobicity	1.54	0.80

Figure 3.5: *Prediction of characteristic charge. A: Model validation of the regression model trained to predict ν where the circles represent the leave-one-out cross-validation and the triangles the test set. B: Overview of the selected features with the regression coefficient and the cross-validated R^2 after feature permutation.*

Applying the same approach to build a QSPR model for K_{eq} did not yield sufficiently accurate models. With the current dataset, the best performing models yielded only a R^2 of 0.58 (data not shown). While ν has direct physical implications, by representing the number of charge interactions between the resin and protein, K_{eq} is lacking these physical implications.[44,64] The equilibrium constant represents all phenomena

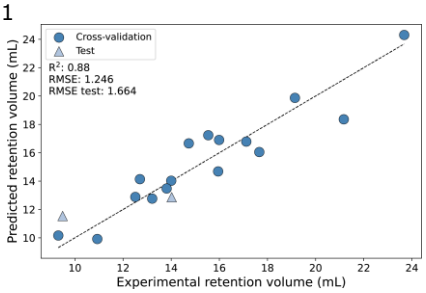
contributing to adsorption. As observed in Figure 3.3, ν shows a clear negative trend with increasing pH, this trend is lacking for K_{eq}. It is thought that the current dataset size is the main limitation as more features might be required to capture the complex relation. To overcome this challenge, increasing the dataset-size would result in a model trained over a greater range of property values, while also allowing an increase of the number of used features without loss of robustness.[24,62]

3.3.4.2 Retention volumes

Alternatively, the K_{eq} can be obtained from the regression as performed in 3.3.2 for experimental data. To achieve this, a MLR model for each LGE was trained (Figure 3.6). The best performing models were obtained using a feature - property correlation filter, removing 40% of the features with the lowest correlation, prior to the feature selection. The trained MLR models, for each LGE, all achieved a cross-validated R^2 of at least 0.88. For all models, the most important feature relates to the EP. More specifically, the median shell positive EP was most important for the four lower gradient lengths (20, 30, 40, and 60 CV). This feature describes the positive EP on the exterior of the protein by projecting each charge onto a plane that represents the resin. For the calculation of the shell, a total of 120 planes surround the protein, in this way representing different binding orientations. Opposed to mapping the EP onto solvent accessible surface, this method considers the distance through the solvent, penalizing protein surface within pockets. The surface fraction of alanine was the second feature selected. Alanine is a small hydrophobic amino acid, therefore this feature implicitly describes the surface hydrophobicity. The positive regression coefficient fitted for this feature indicates that a greater alanine content, and thus higher surface hydrophobicity, results in a higher retention volume. This can be explained by the salting-out effect

of the Na^+ ions used during the gradient elution, resulting in hydrophobic interactions with the resin material.[43]

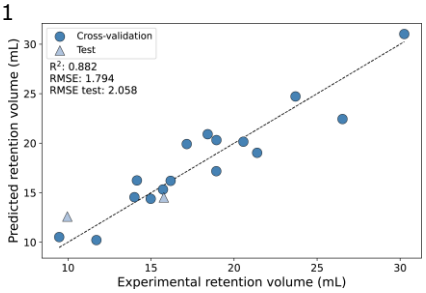
A.1



A.2

Feature	Coefficient	Permutation CV R^2
Intercept	7.47	
Median of shell positive EP	16.56	-0.17
Alanine surface fraction	2.68	0.83

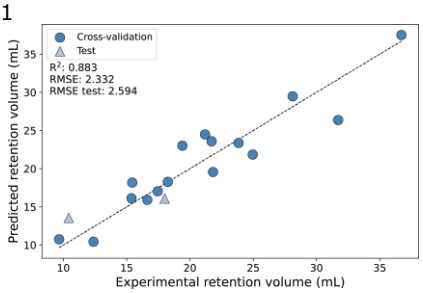
B.1



B.2

Feature	Coefficient	Permutation CV R^2
Intercept	6.50	
Median of shell positive EP	24.18	-0.18
Alanine surface fraction	4.05	0.83

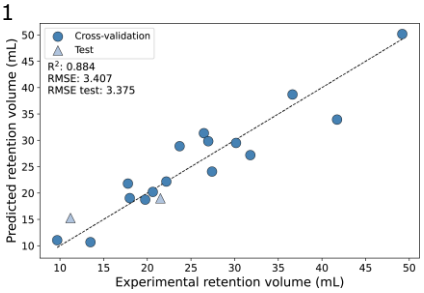
C.1



C.2

Feature	Coefficient	Permutation CV R^2
Intercept	6.39	
Median of shell positive EP	31.79	-0.20
Alanine surface fraction	5.48	0.83

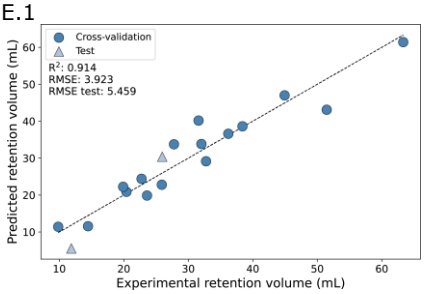
D.1



D.2

Feature	Coefficient	Permutation CV R^2
Intercept	2.97	
Median of shell positive EP	46.76	-0.21
Alanine surface fraction	8.33	0.83

E.1



E.2

Feature	Coefficient	Permutation CV R^2
Intercept	-1.74	
Mean of surface positive EP	37.73	0.85
Mean of shell positive EP	26.28	0.89
Serine surface fraction	12.76	0.83

Figure 3.6: *Prediction of protein retention at different salt gradient lengths where the circles represent the leave-one-out cross-validation and the triangles the test set. A to E show the validation and test of the prediction of the retention volume while applying a salt gradient of 20, 30, 40, 60 and 80 column volumes, respectively. One column volume equals 0.97 mL (Table 1). The tables right of the plots show the feature coefficients and the effect of feature permutation on the cross validated R^2.*

For the 80 CV retention MLR model, the following features were selected: shell positive EP mean, solvent accessible surface positive EP mean, and the serine surface fraction. The feature combination yielded an accurate model with a cross-validated R^2 of 0.91 and a RMSE of 3.9 (Figure 3.6E). For the prediction of the test set, it is observed that the point at the lower end of the retention data is under predicted, compared to being over predicted in all other models. While the EP remains the most important in the model, different features were selected during the sequential feature selection. This is due to the fact that there is no exact linear relationship between gradient length and retention, as can be most notably observed at pH 7.0 in Supplemental Figure S3.5. While the Mean and Median of the shell EP are similar, the slight differences in the features resulted in the selection of the mean. Both the mean of surface positive EP and mean of shell positive EP are important features, with regression coefficients of 37.73 and 26.28 respectively. This importance is not reflected by the permutation models, as both features describe the positive EP, collinearity allows for compensation for a loss of one of the features. However, it is essential to maintain both features to accurately predict the test set, as removing one of them results in less accurate retention estimates (data not shown). Surprisingly, the surface area fraction of serine has a

positive regression coefficient, like the alanine surface fraction in the other four models. In contrast to alanine, serine is a hydrophilic residue. However, the positive regression coefficient indicates increasing retention with higher serine content on the surface, which contradicts the hypothesis for alanine selection for the previous four models. The reason behind the selection of serine in this model is currently unknown. While the models show difficulty in predicting the change of elution order switch of lysozyme and cytochrome c for pH 4.3 and 5, a sharper decrease in retention for cytochrome c compared to lysozyme is predicted (data not shown). Still all models show good accuracy during both cross-validation and model testing, providing high confidence in model robustness.

3.3.4.3 Property prediction of conalbumin at pH 5

To demonstrate the true predictive capabilities of the trained QSPR models for the prediction of retention volumes and isotherm parameters, conalbumin was completely removed from the dataset prior to the train test splitting. This allowed to minimize the bias applied on the model selection. For the prediction of the retention volumes, the error of prediction increased with increasing gradient lengths (Table 3.4). The range of observed retention volumes rises along with the gradient lengths, likewise, the 95% confidence interval increases. Nevertheless, the effect of increasing the gradient length was captured correctly, having a maximal error of about 2 mL in retention volume, which falls within the 95% confidence interval. The characteristic charge was predicted with an error of 0.5, complying to the 95% confidence interval. Unfortunately, as no robust and accurate QSPR model for the K_{eq} could be trained with the current dataset, no direct prediction could be made. Therefore, we applied an alternative method, the predicted retention volumes and characteristic charge were used to regress the K_{eq} using the regression formula, similar to the experimental data method as shown in 3.3.2. regression of adsorption

isotherm parameters. The K_{eq} obtained was 0.028 ±0.006 which is lower than the K_{eq} of 0.078 ±0.012 obtained by regression of the experimental data. This is due to the higher predicted ν by the QSPR model. Validation of the predicted parameters showed an accurate prediction of the conalbumin elution using a 60 CV gradient length (Figure 3.7). Both peak maximum and peak shape are simulated accurately. The difference in the peak retention volume is very small, 0.12 CV, which is 0.2% difference relative to the gradient length (60 CV). The peak concentration differs by 0.009 g/L, which is 2.85% relative to the initial concentration, and the difference in the peak width at half of the peak maximum is only 1.0% relative to the gradient length (60 CV). Interestingly, the predicted parameters seem to better describe the retention profile compared to the parameters obtained from the experimental LGE, which was an average peak retention difference of 1.53% and an average peak width difference of 1.35% with respect to the gradient length (60 CV).

Table 3.4: *Predicted properties for conalbumin at pH 5.0.*

Property	**Experimental value (mL)**	**Predicted value (mL)**	**95% Confidence interval**
Retention volume 20 CV	11.66	11.89	2.56
Retention volume 30 CV	12.89	12.92	3.69
Retention volume 40 CV	14.02	13.76	4.80
Retention volume 60 CV	16.20	15.21	7.02
Retention volume 80 CV	18.19	20.23	8.98
Characteristic charge (ν)	2.36	3.05	1.40

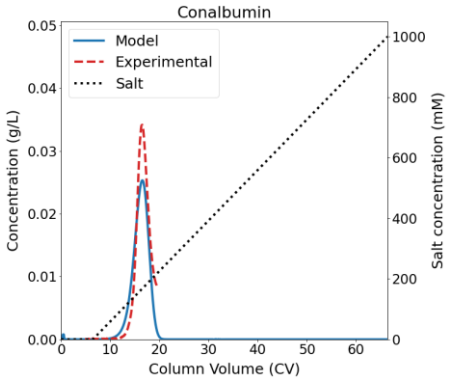


Figure 3.7: *Chromatographic mechanistic model validation of conalbumin for gradient length of 60 CV, equal to 58.2 mL, at a pH of 5.0 using the predicted isotherm parameters. Blue line indicates the MM predicted concentration of the protein, while the red dotted line indicates the experimental concentration. The black dotted line indicates the salt concentration.*

3.3.5 Comparing optimization results between experimentally and QSPR-based methods

For the test protein, conalbumin at pH 5.0, both adsorption isotherm parameters, K_{eq} and v, were determined via two methods. The first method regressed the adsorption isotherm parameters from the LGE data directly, hence LGE are needed to perform this method. While the second method involved the QSPR approach, which, after being properly trained, requires the protein-structure to determine the v and the retention volumes. These two QSPR models were then used to regress the K_{eq} using the regression formula (Eq. 8).

The capture step was optimized to separate conalbumin from the other proteins, prioritizing yield over purity, utilizing the adsorption isotherm parameters determined from both methods. This optimization aimed to assess the agreement between the optimized capture step and the parameters obtained from both methods. The resulting capture steps for both methods are depicted in Figure 8. The optimized variables (e.g., lower and upper cut points and the initial and final salt concentration) show comparability. The differences in both cut points are within 3.3%, and the deviation for both initial and final salt

concentration is around 10 mM, approximately 3% relative to the final salt concentration (330 mM). The obtained purity only differs 0.3% and the yield 1.2% between both methods. These results demonstrate that, in this case study, it was viable to optimize the CEX capture step based solely on knowledge of the protein structure.

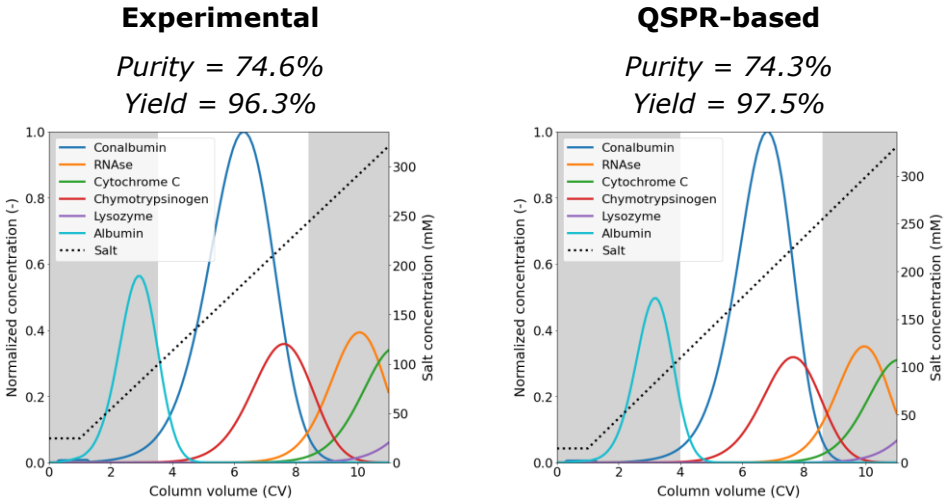


Figure 3.8: *Optimized capture step using the mechanistic model, where the optimization results of the experimental-based (left) and QSPR-based (right) method are compared. Left: experimental-based method, the adsorption isotherm parameters were regressed directly from the LGE.* K_{eq} *0.071 and* ν *= 2.37, lower and upper cut point are 7.7% and 91.2% respectively. The initial and final salt concentration are 24.5 mM and 320.6 mM respectively. Right: QSPR-based method, the retention volumes and* ν *are obtained from QSPR models, followed by using these QSPR models to regress the* K_{eq} *parameter.* K_{eq} *= 0.028 and* ν *= 3.05, lower and upper cut points are 4.4% and 91.7% respectively. The initial and final salt concentration are 14.8 mM and 330.4 mM respectively.*

In the next part, we assessed the effect of the adsorption isotherm parameter uncertainties on the optimization outcome. We aimed to determine if variations within the standard deviation of the parameters would result in different optimal values. For both methods, numerous sample points were generated for each isotherm parameter, covering a range within their respective standard deviation. Subsequently, these sample points were used in the optimization case study. First, the consistency of the optimization case study was evaluated by running the same optimization five times. These results for both methods can be found in Supplemental Tables S1,2. This consistency evaluation

aimed to ensure there were no major deviations in results within the same optimization using identical parameters. Additionally, the minor deviations could be attributed to the optimization process itself. The optimized results for various combinations of K_{eq} and ν, ranging within their respective standard deviation, are shown in Figure 9 for both methods. This includes optimized variables, such as the lower and upper cut points and the initial and final salt concentrations, as well as the purity, and the yield.

In the experimental-based method, the standard deviations for both K_{eq} (0.071 ±0.012) and ν (2.37 ±0.12) are relatively small, resulting in minimal variance in the optimized variables (Figure 3.9, A.1-F.1 and A.2-F.2, for variations in K_{eq} and ν respectively). The lower and upper cut points have a maximum difference of 7% (Figure 9A,B). The initial salt concentration varies between 15 and 40 mM (Figure 9C.1,2), and the final salt concentration is found between 320 and 327 mM (Figure 9D.1,2). These results suggest that despite variations in the isotherm parameters, a consistent optimum is identified, and the optimized variables exhibit only minor variations. The impact on the yield is minimal, with only a 2% variation (Figure 9F.1,2). On the contrary, the effect on purity is more pronounced, fluctuating between 70% and 81%. The decrease in purity is primarily attributed to an increase in the K_{eq} (Figure 9E.1), which is due to the greater relative standard deviation compared to ν.

For the QSPR-based method, the standard deviation of K_{eq} is small (0.028 ±0.006). The randomly spread data indicates that there is no clear correlation between K_{eq} and the optimized variables (Figure 3.9A.3-F3). However, the standard deviation of ν is significantly larger (3.05 ±1.4), this standard deviation was defined by the 95% confidence interval calculated by Eq. 9. The large variation in ν resulted in two identified optima, which is clearly observed in the shift of the

final salt concentration (Figure 3.9D.4). The first solution finds an optimal final salt concentration between 320 – 400 mM. The shift to the second optimal solution occurs when v is greater than 3.6, finding the final salt concentration at around 800 mM. Remarkably, both optimal final salt concentrations are close to the set boundaries. As the characteristic charge increases, the component is expected to elute at a higher salt concentration and thus at a later moment during the gradient. This results in a greater overlap between conalbumin and the other impurities. Such a shift was not observed for the initial salt concentration, where most optimal conditions were found between 10 and 30 mM (Figure 3.9C.4). The effect of v is also reflected in the purity and the yield (Figure 3.9E.4 and 3.9F.4 respectively). Until v is 2.2, the purity is around 75% and the yield is almost 100%, while above this value of v, the purity increases rapidly, and the yield drops to about 95%. From this point, increasing v results in a decreasing purity and increasing yield. However, the range of the purity is broader, 50 – 85% than that of the yield, which only fluctuates between 95% and 99%. This broader range in the purity is probably due to a combination of the shift in retention volume resulting from variation of v, and the optimization function (11). In the function, the yield is prioritized, representing a capture step optimization. Therefore, during challenging separation processes, the compromise on the yield is always less compared to purity. Changes in the optimization weights would result in a shift in priority between purity and yield that would translate to the selection of different cut points rather than initial and final salt concentrations. Despite the greater uncertainty in the determined v in the QSPR-method, only two optima were identified, and one of them corresponds to the optimum found in the experimental-based method.

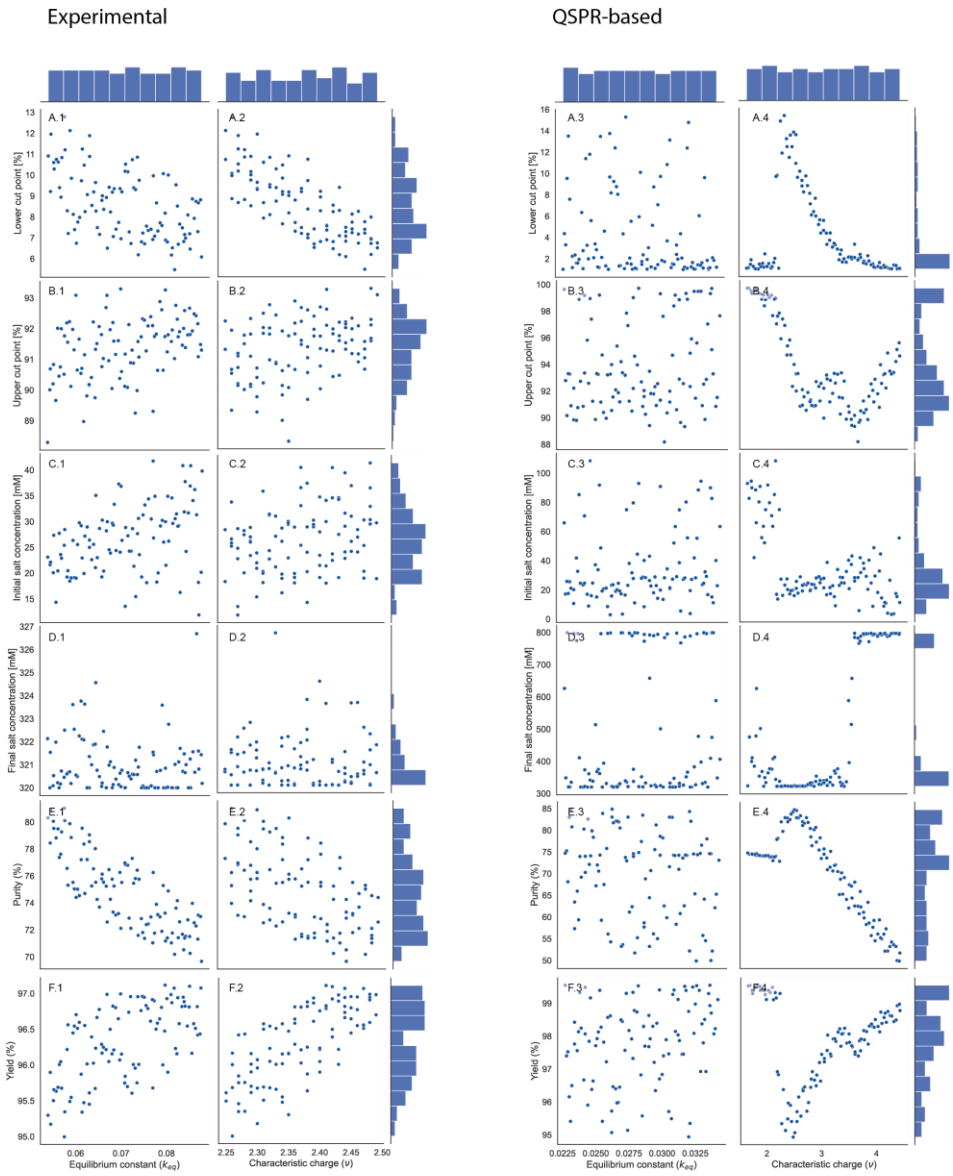


Figure 3.9: *Joint plots of scatter and hist plots between the adsorption isotherm parameters (e.g., the characteristic charge and the equilibrium constant) and the optimized variables (e.g., lower and upper cut point and the initial and final salt concentrations, and the purity and the yield). Left: experimental-based method results. Right: QSPR-based method results.*

Furthermore, this optimization approach is applicable for defining the operating window of certain variables. The method employed for varying the adsorption isotherm parameters can also be used to vary other variables and assess the optimized result. In this way, the initial

process design space for CPP can be defined, which is part of the QbD concept.[65] The mechanistic modeling outcomes provide knowledge on the process, therefore the number of wet-lab experiments to define the real process design space can be reduced in comparison to performing a wet-lab DoE from scratch. For the QSPR-based method, no wet-lab experiments are needed to determine the adsorption isotherm parameters and therefore the total number of experiments are even more reduced compared to the experimental-based method. For a new protein, only the protein-structure is needed to perform this optimization and make an estimation of the operating window for each optimizing variable. To illustrate, using the results from the QSPR-based method in this study, we can already narrow down the number of wet-lab DoE required to define the process design space. The final salt concentration only has to be evaluated around two main values (e.g., around 320 mM and 800 mM, see Figure 3.9D.4), while only one point of the initial salt concentration has to be assessed (e.g., 20 mM). Ultimately, the QSPR-based method offers an added advantage by allowing the incorporation of additional data over time. This not only enhances the model's accuracy, but also enables the application to other process designs, provided that the same conditions are used.

Currently, only the linear part of the isotherm is considered as only low loading conditions are investigated. Prediction of the parameters describing the non-linear part of the isotherm as well as competitive behavior would make the method more complete. Nevertheless, for the purpose of preselection of conditions for early-stage process design, considering only linear behavior should be sufficient. Additionally, the amount of available training data might pose a bottleneck, like the prediction of the K_{eq} presented in this work. Even though the predictions of the retention volumes and characteristic charge showed high accuracy, increasing the variety of proteins would make the models more robust. To extend this method to more complex mixtures,

such as host cell lysates, several challenges should be overcome. While a similar fractionation approach to convolute single peaks can be used for a complex mixture, more accurate analytical methods are required for protein identification. Potentially, mass spectrometry methods allow the required resolution providing relative protein abundances. Additionally, protein interactions and complex formation should be taken into account during the QSPR modeling. Co-elution has already been studied extensively, and recently Panikulam et al., published a novel method to describe co-elution mechanisms for protein A chromatography.[66] Further maturation and combination of these methods would allow better integration and application for complex mixtures.

3.4 Conclusion

In this work, we demonstrated a holistic modeling approach, where we combined QSPR and chromatographic MM to optimize a CEX capture step. For an unseen protein, only the protein structure was needed to determine the adsorption isotherm parameters and predict the chromatographic retention behavior with MM. We assessed that the uncertainties in the determined adsorption isotherm parameters have a minimal and nearly equal impact for both the experimental-based and QSPR-based method.

For the experimental-based method, we successfully regressed the adsorption isotherm parameters with an R^2 minimum of 0.95. The standard deviation for the characteristic charge is within 1 – 6% of the corresponding regressed parameter value, and for the equilibrium constant, it ranges between 7 – 25% of the regressed parameter value. Moreover, the MM validation showed to be accurate with an average retention peak difference of 1.53% with respect to the gradient length.

We successfully trained MLR-QSPR models with a minimum R^2 of 0.88, even with a limited dataset composed of only five different proteins

measured at four pH values. The MLR-QSPR models for predicting the characteristic charge and the retention volumes can be used to regress the equilibrium constant using the regression formula. A good agreement was obtained for the MM validation for an unseen protein, conalbumin, showing only 0.2% retention peak difference with respect to the gradient length.

Both the experimental-based and the QSPR-based methods demonstrated a consistent optimized CEX capture step. The same optimum was found by both methods, and an additional optimum was identified using the QSPR-based method, due to the larger standard deviation in v (3.05 ±1.4) compared to the experimentally predicted v (2.37 ±0.12). Using in silico optimization results as a guide can substantially reduce experimental effort, requiring experimental validation only for promising conditions. Moreover, increasing dataset sizes enhances the QSPR model accuracy, diminishing uncertainty in adsorption isotherm parameters and therefore minimizing the variance in the identified operating window.

This work highlights the value and applicability of multiscale modeling, capable of optimizing a CEX capture step with only knowing the protein structure. Integrating QSPR, chromatographic MM, and optimization tools creates a versatile workflow relevant to industrial case studies. The specific case study presented aims to provide a workflow which should be expanded using larger datasets to enable more accurate predictions. This approach ultimately enables determining initial optimal process conditions without preliminary experiments, which is especially beneficial for early phase process development when limited material and resources are available. Future applications involve extending this strategy to complex protein mixtures and broader type of chromatographic resins, offering a cost-effective and time-saving alternative that enhances overall process understanding and efficiency.

3.5 References

1. Birch, J. R., & Onakunle, Y. (2005). Biopharmaceutical proteins: opportunities and challenges. *Methods Mol Biol*, *308*, 1–16. https://doi.org/10.1385/1-59259-922-2:001

2. Wen, E. P., Ellis, R., & Pujar, N. S. (2014). *Vaccine development and manufacturing* (E. P. Wen, R. Ellis, & N. S. Pujar, Eds.). John Wiley & Sons.

3. Jagschies, G., Lindskog, E., Łacki, K., & Galliher, P. (2018). *Biopharmaceutical Processing: Development, Design, and Implementation of Manufacturing Processes*.

4. Kesik-Brodacka, M. (2018). Progress in biopharmaceutical development. *Biotechnology and Applied Biochemistry*, *65*(3), 306–322. https://doi.org/10.1002/bab.1617

5. Kelley, B. (2020). Developing therapeutic monoclonal antibodies at pandemic pace. *Nature Biotechnology*, *38*(5), 540–545. https://doi.org/10.1038/s41587-020-0512-5

6. Łącki, K. M. (2018). Chapter 16 - Introduction to Preparative Protein Chromatography. In G. Jagschies, E. Lindskog, K. Łącki, & P. Galliher (Eds.), *Biopharmaceutical Processing* (pp. 319–366). Elsevier. https://doi.org/https://doi.org/10.1016/B978-0-08-100623-8.00016-5

7. Hanke, A. T., & Ottens, M. (2014). Purifying biopharmaceuticals: Knowledge-based chromatographic process development. *Trends in Biotechnology*, *32*(4), 210–220. https://doi.org/10.1016/j.tibtech.2014.02.001

8. Keulen, D., Geldhof, G., Bussy, O. Le, Pabst, M., & Ottens, M. (2022). Recent advances to accelerate purification process development: A review with a focus on vaccines. *Journal of Chromatography A*, *1676*, 463195. https://doi.org/https://doi.org/10.1016/j.chroma.2022.463195

9. Reinhardt, I. C., Oliveira, D. J. C., & Ring, D. D. T. (2020). Current Perspectives on the Development of Industry 4.0 in the Pharmaceutical Sector. *Journal of Industrial Information Integration*, *18*, 100131. https://doi.org/https://doi.org/10.1016/j.jii.2020.100131

10. von Stosch, M., Portela, R. M. C., & Varsakelis, C. (2021). A roadmap to AI-driven in silico process development: bioprocessing 4.0 in practice. *Current Opinion in Chemical Engineering*, *33*, 100692. https://doi.org/https://doi.org/10.1016/j.coche.2021.100692

11. Alosert, H., Savery, J., Rheaume, J., Cheeks, M., Turner, R., Spencer, C., S. Farid, S., & Goldrick, S. (2022). Data integrity within the biopharmaceutical sector in the era of Industry 4.0. *Biotechnology Journal*, *17*(6), 2100609. https://doi.org/https://doi.org/10.1002/biot.202100609

12. Narayanan, H., Luna, M. F., von Stosch, M., Cruz Bournazou, M. N., Polotti, G., Morbidelli, M., Butté, A., & Sokolov, M. (2020). Bioprocessing in the Digital Age: The Role of Process Models. *Biotechnology Journal*, *15*(1), 1900172. https://doi.org/https://doi.org/10.1002/biot.201900172

13. Rathore, A. S. (2016). Quality by Design (QbD)-Based Process Development for Purification of a Biotherapeutic. *Trends in Biotechnology*, *34*(5), 358–370. https://doi.org/https://doi.org/10.1016/j.tibtech.2016.01.003

14. FDA. (2004). *PAT Guidance for Industry - A Framework for innovative Pharmaceutical Development, Manufacturing and Quality Assurance* (F. and D. A. (FDA) US Department of Health and Human Services Center for Drug Evaluation and Research (CDER), Center for Veterinary Medicine (CVM), Office of Regulatory

Affairs (ORA), Ed.). http://www.fda.gov/regulatory-information/search-fda-guidance-documents/pat-framework-innovative-pharmaceutical-development-manufacturing-and-quality-assurance

15. ICH. (2009). ICH Harmonised Tripartite Guideline: Pharmaceutical Development Q8 (R2). In *ICH*.

16. Mollerup, J. M., Hansen, T. B., Kidal, S., & Staby, A. (2008). Quality by design-Thermodynamic modelling of chromatographic separation of proteins. *Journal of Chromatography A*, *1177*(2), 200–206. https://doi.org/10.1016/j.chroma.2007.08.059

17. Saleh, D., Wang, G., Müller, B., Rischawy, F., Kluters, S., Studts, J., & Hubbuch, J. (2020). Straightforward method for calibration of mechanistic cation exchange chromatography models for industrial applications. *Biotechnology Progress*, *36*(4), 1–12. https://doi.org/10.1002/btpr.2984

18. Kumar, V., & Lenhoff, A. M. (2020). Mechanistic Modeling of Preparative Column Chromatography for Biotherapeutics. *Annual Review of Chemical and Biomolecular Engineering*, *11*(1), 235–255. https://doi.org/https://doi.org/10.1146/annurev-chembioeng-102419-125430

19. Shekhawat, L. K., Tiwari, A., Yamamoto, S., & Rathore, A. S. (2022). An accelerated approach for mechanistic model based prediction of linear gradient elution ion-exchange chromatography of proteins. *Journal of Chromatography A*, *1680*, 463423. https://doi.org/10.1016/j.chroma.2022.463423

20. Rischawy, F., Saleh, D., Hahn, T., Oelmeier, S., Spitz, J., & Kluters, S. (2019). Good modeling practice for industrial chromatography: Mechanistic modeling of ion exchange chromatography of a bispecific antibody. *Computers and Chemical Engineering*, *130*, 106532. https://doi.org/10.1016/j.compchemeng.2019.106532

21. Nfor, B. K., Ahamed, T., Pinkse, M. W. H., van der Wielen, L. A. M., Verhaert, P. D. E. M., van Dedem, G. W. K., Eppink, M. H. M., van de Sandt, E. J. A. X., & Ottens, M. (2012). Multi-dimensional fractionation and characterization of crude protein mixtures: Toward establishment of a database of protein purification process development parameters. *Biotechnology and Bioengineering*, *109*(12), 3070–3083. https://doi.org/10.1002/bit.24576

22. Close, E. J., Salm, J. R., Bracewell, D. G., & Sorensen, E. (2014). A model based approach for identifying robust operating conditions for industrial chromatography with process variability. *Chemical Engineering Science*, *116*, 284–295. https://doi.org/https://doi.org/10.1016/j.ces.2014.03.010

23. Disela, R., Le Bussy, O., Geldhof, G., Pabst, M., & Ottens, M. (2023). Characterisation of the E. coli HMS174 and BLR host cell proteome to guide purification process development. *Biotechnology Journal*, *18*(9), 2300068. https://doi.org/10.1002/biot.202300068

24. Saleh, D., Hess, R., Ahlers-Hesse, M., Rischawy, F., Wang, G., Grosch, J.-H., Schwab, T., Kluters, S., Studts, J., & Hubbuch, J. (2023). A multiscale modeling method for therapeutic antibodies in ion exchange chromatography. *Biotechnology and Bioengineering*, *120*(1), 125–138. https://doi.org/https://doi.org/10.1002/bit.28258

25. Mazza, C. B., Sukumar, N., Breneman, C. M., & Cramer, S. M. (2001). Prediction of protein retention in ion-exchange systems using molecular descriptors obtained from crystal structure. *Analytical Chemistry*, *73*(22), 5457–5461. https://doi.org/10.1021/ac010797s

26. Breneman, C. M., Thompson, T. R., Rhem, M., & Dung, M. (1995). Electron density modeling of large systems using the transferable atom equivalent method. *Computers & Chemistry*, *19*(3), 161–179. https://doi.org/10.1016/0097-8485(94)00052-G

27. Whitehead, C. E., Breneman, C. M., Sukumar, N., & Ryan, M. D. (2003). Transferable atom equivalent multicentered multipole expansion method. *Journal of Computational Chemistry*, *24*(4), 512–529. https://doi.org/10.1002/jcc.10240

28. Song, M., Breneman, C. M., Bi, J., Sukumar, N., Bennett, K. P., Cramer, S., & Tugcu, N. (2002). Prediction of Protein Retention Times in Anion-Exchange Chromatography Systems Using Support Vector Regression. *Journal of Chemical Information and Computer Sciences*, *42*(6), 1347–1357. https://doi.org/https://doi.org/10.1021/ci025580t

29. Ladiwala, A., Xia, F., Luo, Q., Breneman, C. M., & Cramer, S. M. (2006). Investigation of protein retention and selectivity in HIC systems using quantitative structure retention relationship models. *Biotechnology and Bioengineering*, *93*(5), 836–850. https://doi.org/10.1002/bit.20771

30. Ladiwala, A., Rege, K., Breneman, C. M., & Cramer, S. M. (2005). Prediction of adsorption isotherm parameters and chromatographic behavior in ion-exchange systems. *Proceedings of the National Academy of Sciences*, *102*(33), 11710–11715. https://doi.org/https://www.pnas.org/doi/abs/10.1073/pnas.0408769102

31. Chen, J., & Cramer, S. M. (2007). Protein adsorption isotherm behavior in hydrophobic interaction chromatography. *Journal of Chromatography A*, *1165*(1), 67–77. https://doi.org/https://doi.org/10.1016/j.chroma.2007.07.038

32. Yang, T., Sundling, M. C., Freed, A. S., Breneman, C. M., & Cramer, S. M. (2007). Prediction of pH-dependent chromatographic behavior in ion-exchange systems. *Analytical Chemistry*, *79*(23), 8927–8939. https://doi.org/10.1021/ac071101j

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

34. Malmquist, G., Nilsson, U. H., Norrman, M., Skarp, U., Strömgren, M., & Carredano, E. (2006). Electrostatic calculations and quantitative protein retention models for ion exchange chromatography. *Journal of Chromatography A*, *1115*(1–2), 164–186. https://doi.org/10.1016/j.chroma.2006.02.097

35. Hanke, A. T., Klijn, M. E., Verhaert, P. D. E. M., van der Wielen, L. A. M., Ottens, M., Eppink, M. H. M., & van de Sandt, E. J. A. X. (2016). Prediction of protein retention times in hydrophobic interaction chromatography by robust statistical characterization of their atomic-level surface properties. *Biotechnology Progress*, *32*(2), 372–381. https://doi.org/10.1002/btpr.2219

36. Kittelmann, J., Lang, K. M. H., Ottens, M., & Hubbuch, J. (2017). Orientation of monoclonal antibodies in ion-exchange chromatography: A predictive quantitative structure–activity relationship modeling approach. *Journal of Chromatography A*, *1510*, 33–39. https://doi.org/10.1016/j.chroma.2017.06.047

37. Kittelmann, J., Lang, K. M. H., Ottens, M., & Hubbuch, J. (2017). An orientation sensitive approach in biomolecule interaction quantitative structure–activity relationship modeling and its application in ion-exchange chromatography. *Journal of Chromatography A*, *1482*, 48–56. https://doi.org/10.1016/j.chroma.2016.12.065

38. Robinson, J. R., Karkov, H. S., Woo, J. A., Krogh, B. O., & Cramer, S. M. (2017). QSAR models for prediction of chromatographic behavior of homologous Fab variants. *Biotechnology and Bioengineering*, *114*(6), 1231–1240. https://doi.org/10.1002/bit.26236

39. Hess, R., Faessler, J., Yun, D., Saleh, D., Grosch, J. H., Schwab, T., & Hubbuch, J. (2023). Antibody sequence-based prediction of pH gradient elution in multimodal chromatography. *Journal of Chromatography A*, *1711*(October), 464437.

https://doi.org/10.1016/j.chroma.2023.464437

40. Emonts, J., & Buyel, J. F. (2023). An overview of descriptors to capture protein properties – Tools and perspectives in the context of QSAR modeling. *Computational and Structural Biotechnology Journal*, *21*, 3234–3247. https://doi.org/10.1016/j.csbj.2023.05.022

41. Danishuddin, & Khan, A. U. (2016). Descriptors and their selection methods in QSAR analysis: paradigm for drug design. *Drug Discovery Today*, *21*(8), 1291–1302. https://doi.org/https://doi.org/10.1016/j.drudis.2016.06.013

42. Neijenhuis, T., Le Bussy, O., Geldhof, G., Klijn, M. E., & Ottens, M. (2024). Predicting protein retention in ion-exchange chromatography using an open source QSPR workflow. *Biotechnology Journal*, *19*(3), e2300708. https://doi.org/10.1002/biot.202300708

43. Hou, Y., & Cramer, S. M. (2011). Evaluation of selectivity in multimodal anion exchange systems: A priori prediction of protein retention and examination of mobile phase modifier effects. *Journal of Chromatography A*, *1218*(43), 7813–7820. https://doi.org/10.1016/j.chroma.2011.08.080

44. Osberghaus, A., Hepbildikler, S., Nath, S., Haindl, M., von Lieres, E., & Hubbuch, J. (2012). Determination of parameters for the steric mass action model-A comparison between two approaches. *Journal of Chromatography A*, *1233*, 54–65. https://doi.org/10.1016/j.chroma.2012.02.004

45. Hagemann, F., Adametz, P., Wessling, M., & Thom, V. (2020). Modeling hindered diffusion of antibodies in agarose beads considering pore size reduction due to adsorption. *Journal of Chromatography A*, *1626*, 461319. https://doi.org/https://doi.org/10.1016/j.chroma.2020.461319

46. Keulen, D., van der Hagen, E., Geldhof, G., Le Bussy, O., Pabst, M., & Ottens, M. (2023). Using artificial neural networks to accelerate flowsheet optimization for downstream process development. *Biotechnology and Bioengineering*, 1–14. https://doi.org/https://doi.org/10.1002/bit.28454

47. Ruthven, D. M. (1984). *Principles of adsorption and adsorption processes*. John Wiley & Sons.

48. Brenner, H., & Gaydos, L. J. (1977). The constrained brownian movement of spherical particles in cylindrical pores of comparable radius. *Journal of Colloid and Interface Science*, *58*(2), 312–356. https://doi.org/10.1016/0021-9797(77)90147-3

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

50. Petzold, L. (1983). Automatic Selection of Methods for Solving Stiff and Nonstiff Systems of Ordinary Differential Equations. *SIAM Journal on Scientific and Statistical Computing*, *4*(1), 136–148. https://doi.org/https://doi.org/10.1137/0904010

51. Nfor, B. K., Zuluaga, D. S., Verheijen, P. J. T., Verhaert, P. D. E. M., van der Wielen, L. A. M., & Ottens, M. (2011). Model-based rational strategy for chromatographic resin selection. *Biotechnology Progress*, *27*(6), 1629–1643. https://doi.org/https://doi.org/10.1002/btpr.691

52. Nfor, B. K., Noverraz, M., Chilamkurthi, S., Verhaert, P. D. E. M., van der Wielen, L. A. M., & Ottens, M. (2010). High-throughput isotherm determination and thermodynamic modeling of protein adsorption on mixed mode adsorbents. *Journal of Chromatography A*, *1217*(44), 6829–6850. https://doi.org/https://10.1016/j.chroma.2010.07.069

53. Pirrung, S. M., Parruca da Cruz, D., Hanke, A. T., Berends, C., Van Beckhoven, R. F. W. C., Eppink, M. H. M., & Ottens, M. (2018). Chromatographic parameter determination for complex biological feedstocks. *Biotechnology Progress*, *34*(4), 1006–1018. https://doi.org/10.1002/btpr.2642

54. Hahn, T., Geng, N., Petrushevska-Seebach, K., Dolan, M. E., Scheindel, M., Graf, P., Takenaka, K., Izumida, K., Li, L., Ma, Z., & Schuelke, N. (2023). Mechanistic modeling, simulation, and optimization of mixed-mode chromatography for an antibody polishing step. *Biotechnology Progress*, *39*(2), e3316. https://doi.org/https://doi.org/10.1002/btpr.3316

55. Parente, E. S., & Wetlaufer, D. B. (1986). Relationship between isocratic and gradient retention times in the high-performance ion-exchange chromatography of proteins. Theory and experiment. *Journal of Chromatography A*, *355*(C), 29–40. https://doi.org/10.1016/S0021-9673(01)97301-7

56. Schmidt-Traub, H., Schulte, M., Seidel-Morgenstern, A., & Schmidt-Traub, H. (2012). *Preparative chromatography*. Wiley Online Library.

57. Shukla, A. A., Bae, S. S., Moore, J. A., Barnthouse, K. A., & Cramer, S. M. (1998). Synthesis and characterization of high-affinity, low molecular weight displacers for cation-exchange chromatography. *Industrial and Engineering Chemistry Research*, *37*(10), 4090–4098. https://doi.org/10.1021/ie9801756

58. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, *28*(1), 235–242. https://doi.org/10.1093/nar/28.1.235

59. Rodrigues, J., Teixeira, J. M. C., Trellet, M., & Bonvin, A. (2018). pdb-tools: a swiss army knife for molecular structures [version 1; peer review: 2 approved]. *F1000Research*, *7*(1961). https://doi.org/https://doi.org/10.12688/f1000research.17456.1

60. Olsson, M. H. M., Søndergaard, C. R., Rostkowski, M., & Jensen, J. H. (2011). PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical p K a Predictions. *Journal of Chemical Theory and Computation*, *7*(2), 525–537. https://doi.org/10.1021/ct100578z

61. Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF Chimera - A visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, *25*(13), 1605–1612. https://doi.org/10.1002/jcc.20084

62. Yang, T., Sundling, M. C., Freed, A. S., Breneman, C. M., & Cramer, S. M. (2007). Prediction of pH-dependent chromatographic behavior in ion-exchange systems. *Analytical Chemistry*, *79*(23), 8927–8939. https://doi.org/10.1021/ac071101j

63. Topliss, J. G., & Costello, R. J. (1972). Chance correlations in structure-activity studies using multiple regression analysis. *Journal of Medicinal Chemistry*, *15*(10), 1066–1068. https://doi.org/10.1021/jm00280a017

64. Brooks, C. A., & Cramer, S. M. (1992). Steric mass-action ion exchange: Displacement profiles and induced salt gradients. *AIChE Journal*, *38*(12), 1969–1978. https://doi.org/https://doi.org/10.1002/aic.690381212

65. Rathore, A. S. (2009). Roadmap for implementation of quality by design (QbD) for biotechnology products. *Trends in Biotechnology*, *27*(9), 546–553. https://doi.org/https://doi.org/10.1016/j.tibtech.2009.06.006

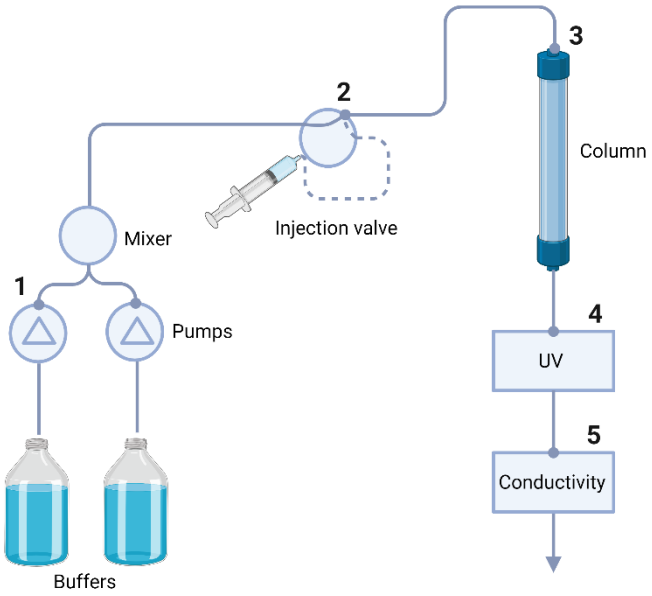
66. Panikulam, S., Hanke, A., Kroener, F., Karle, A., Anderka, O., Villiger, T. K., & Lebesgue, N. (2024). Host cell protein networks as a novel co-elution mechanism during protein A chromatography. *Biotechnology and Bioengineering*. https://doi.org/10.1002/bit.28678

3.6 Supplemental material

3.6.1 Supplemental Methods

Dead volume and dwell volume

The volume of the tubing was determined by excluding the column and using 1 M sodium chloride with a 100 µL sample loop. A schematic overview of the tubing in the Äkta system is shown in Figure S3.1, in which the dead volume is indicated from the numbers 2 to 4 and the dwell volume from 1 to 3.



Supplemental Figure S3.1: *Schematic representation of the Äkta system, the dead volume is defined from point 2 to 4 and the dwell volume from point 1 to 3. The injection valve is indicated with the dashed line and not considered in the dead volume and dwell volume. Created with Biorender.com*

The dead volume (V_{dead}), tubing 3 and 4, is calculated according to Schmidt-Traub et al. (2012) as follows[128]:

$$V_{dead} = V_{R,0} - \frac{V_{inj}}{2} - V_5, \qquad \text{(S3.1)}$$

where $V_{R,0}$ is the retention volume measured including the injection volume (V_{inj}), which is therefore subtracted to only obtain the dead volume. V_5 is the tubing between the UV-detector and the conductivity (indicated with number 5), from the internal diameter, 0.50 mm, and the length, 170 mm, it was calculated to be 0.033 mL.

The dwell volume is needed for the calculations in the regression formula and is equal to the volume from point 1 to 3 (Figure S3.1). The tubing before point 1 is already filled prior to elution. The dwell volume was determined by introducing buffer B, containing 1 M sodium chloride as a pulse for 5 CV, followed by subtracting the V_{dead} and V_5.

Porosity calculations

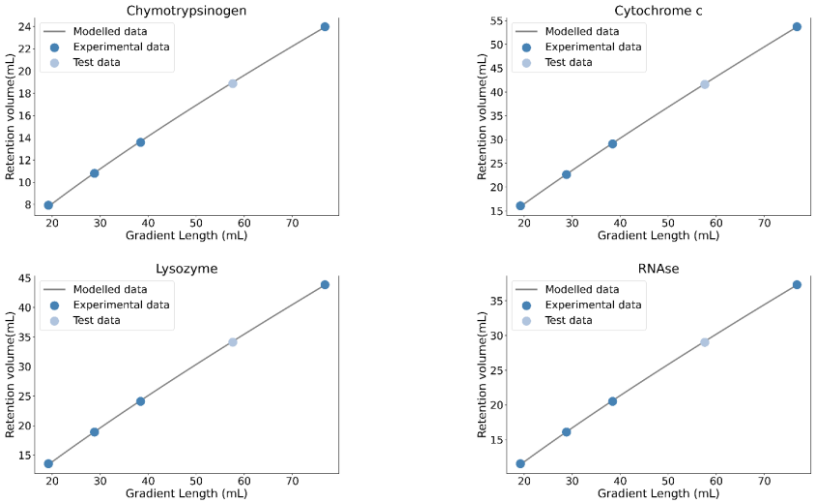
The total porosity (ε_t) was determined using 1 M sodium chloride, as salt can enter the pores, and calculated as follows:

$$\varepsilon_t = \frac{V_m + V_{pore}}{V_C}, \tag{S3.2}$$

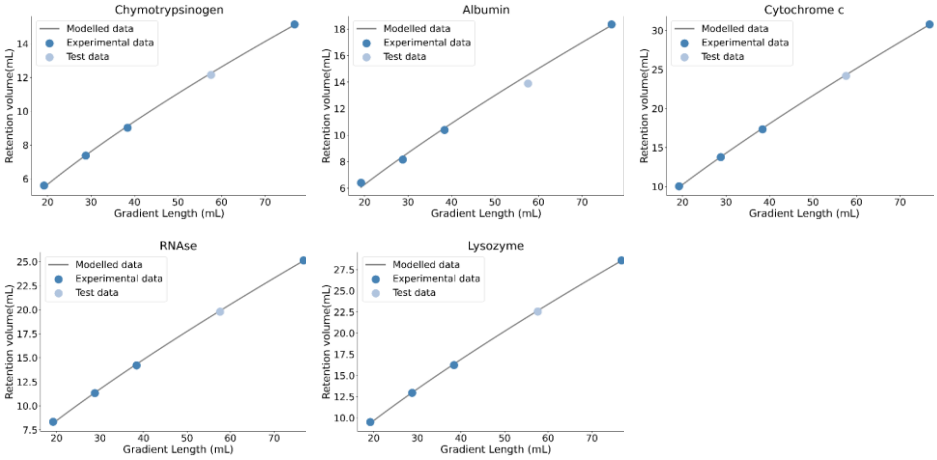
$$V_m + V_{pore} = V_{0,ret} - V_{dead}, \tag{S3.3}$$

where V_m is the interstitial volume of the fluid phase also known as the column void volume, V_{pore} is the volume of the pore system, and V_C is the total volume of the packed column. $V_{0,ret}$ is the measured retention volume from which the dead volume is subtracted to only consider the retention volume in the column. The external porosity, $\varepsilon_b = V_m/V_C$, was determined using a solution of 10 mg/mL Dextran (DXT1740K, American Polymer Standards Corporation, USA) with a volume of 250 µL. V_m was determined using Eq. 3. Subsequently, the total and external porosity are used to determine the internal porosity (ε_p) as

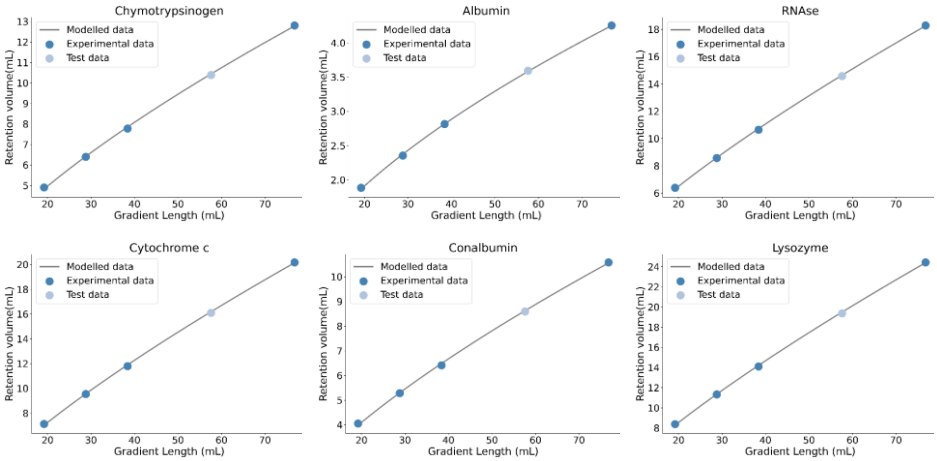
$$\varepsilon_p = \frac{\varepsilon_t - \varepsilon_b}{1 - \varepsilon_b}. \tag{S3.4}$$



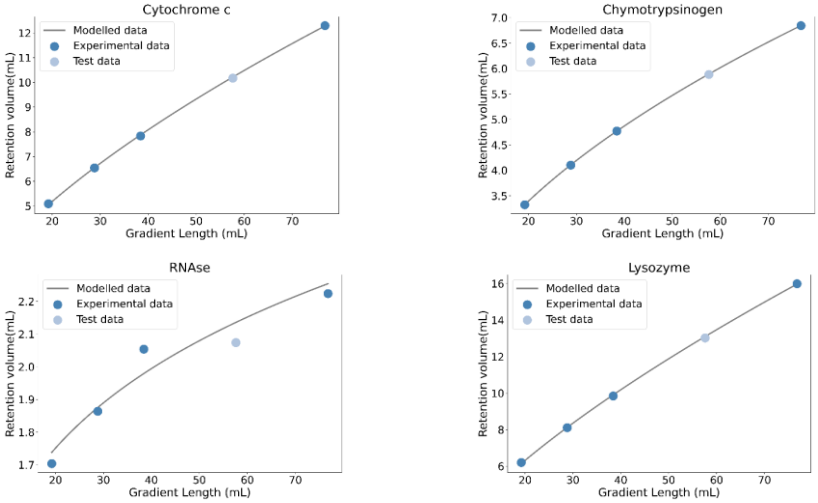
Supplemental Figure S3.2: *Fitted regression curves at pH 3.5 (grey line) of the experimental data (dark blue dots) and the test data point (light blue dot) at 58.2 mL, equal to 60 CV as 1 CV is 0.97 mL. All fits obtained an R^2 of 0.999 and an RMSE of 0.08, 0.11, 0.11, and 0.09 for chymtrypsinogen, cytochrome C, lysozyme, and RNase respectively.*



Supplemental Figure S3.3: *Fitted regression curves at pH 4.3 (grey line) of the experimental data (dark blue dots) and the test data point (light blue dot) at 58.2 mL, equal to 60 CV as 1 CV is 0.97 mL. All fits obtained an R^2 of 0.999 and an RMSE of 0.07, 0.22, 0.10, 0.10, and 0.09 for albumin, chymtrypsinogen, cytochrome C, lysozyme, and RNase respectively.*



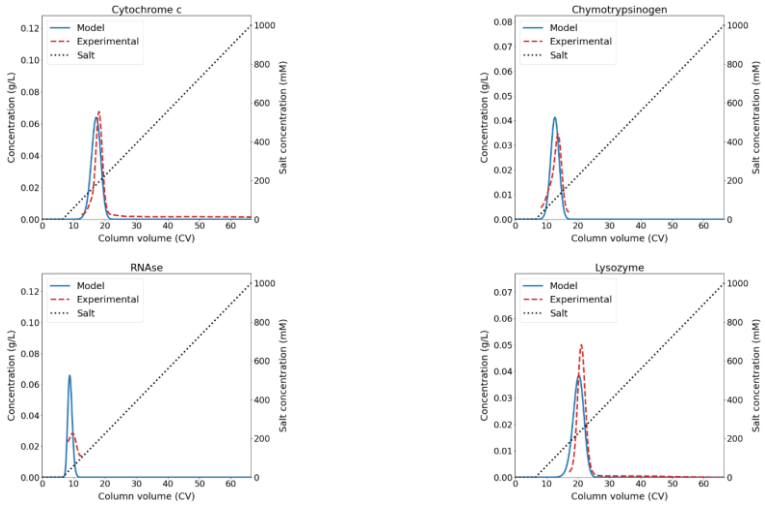
***Supplemental Figure S3.4**: Fitted regression curves at pH 5.0 (grey line) of the experimental data (dark blue dots) and the test data point (light blue dot) at 58.2 mL, equal to 60 CV as 1 CV is 0.97 mL. All fits obtained an R^2 of 0.999 and an RMSE of 0.01, 0.05, 0.06, 0.06, 0.07, and 0.08 for albumin, chymotrypsinogen, cytochrome C, lysozyme, RNase, and conalbumin respectively.*



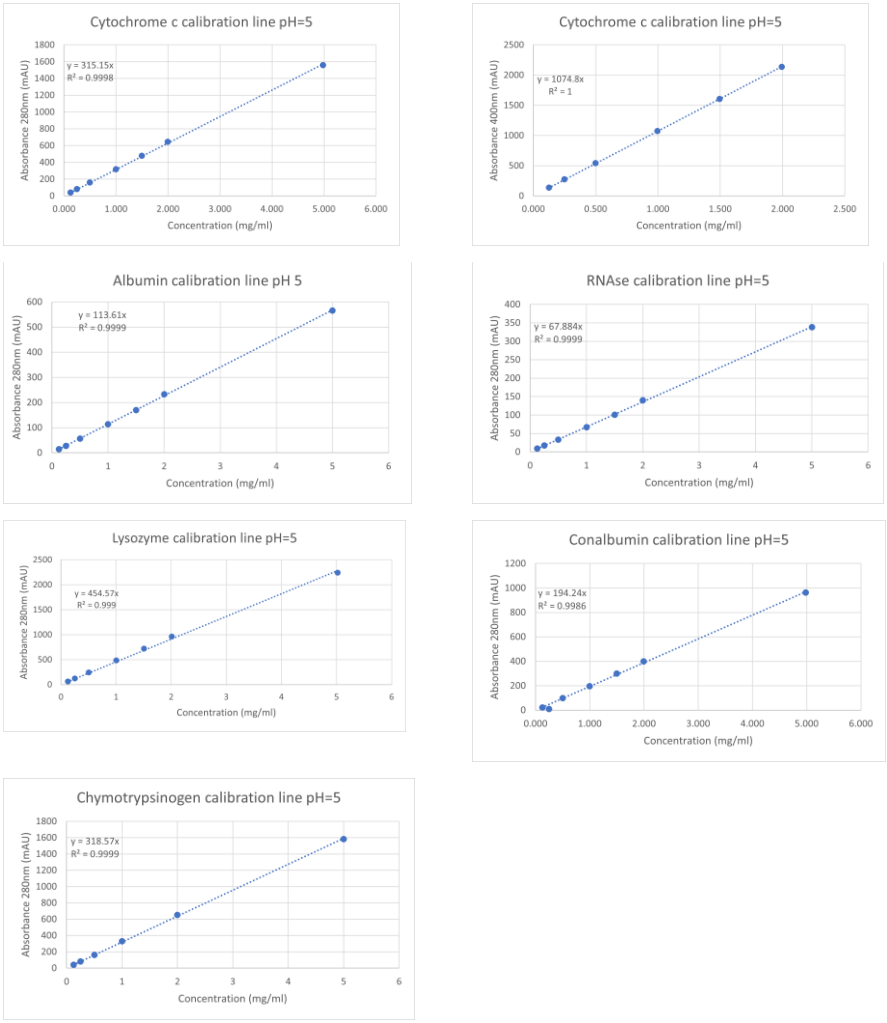
***Supplemental Figure S3.5:** Fitted regression curves at pH 7.0 (grey line) of the experimental data (dark blue dots) and the test data point (ligth blue dot) at 58.2 mL, equal to 60 CV as 1 CV is 0.97 mL. All fits obtained an R^2 of 0.999, except for RNAse that has an R^2 of 0.95. The RMSE values are 0.03, 0.002, 0.04, and 0.04 for cytochrome C, chymtrypsinogen, RNAse, and lysozyme respectively.*

3.6.2 Supplemental Discussion

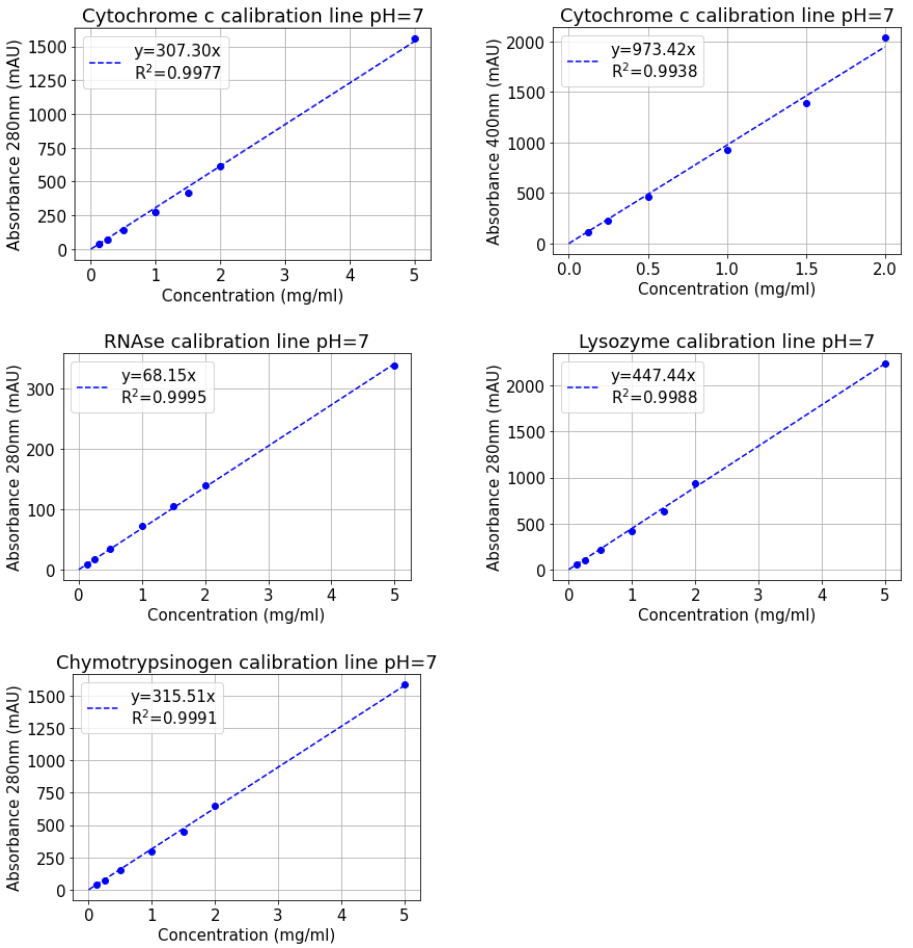
Additional data for the mechanistic model validated at pH 7.0. For all proteins at pH 7.0, the maximum retention peak difference is 1.01 CV and the average difference is 0.86 CV, which is 1.68% and 1.43% with respect to the gradient length (60 CV). To assess the concentration agreement between the modeled and experimental results, we compared the difference between the peak width at half of the peak maximum and the peak concentration. RNAse was left out of this comparison for the peak width difference, as determining half of the peak maximum is not possible for the experimental data. The maximum peak width difference is 2.07 CV, equal to 2.23% relative to the gradient length (60 CV). The average peak width difference is 0.81 CV, equal to 1.35% relative to the gradient length (60 CV). The peak concentration differs maximally by 0.04 mg/mL, which deviates about 7.8% to the initial concentration. The average difference in the peak concentration is 0.01 mg/mL, equal to 3.1% relative to the initial concentration.



supplemental Figure S3.6: *Chromatographic mechanistic model validation for gradient length of 60 CV, equal to 58.2 mL, at a pH of 7.0. Blue line indicate the MM predicted concentration of the protein, while the red dotted line indicates the experimental concentration. The black dotted line indicates the salt concentration. The initial concentrations are Chymotrypsinogen: 0.46 mg/mL, Cytochrome C: 0.80 mg/mL, Lysozyme: 0.55 mg/mL, and RNAse: 0.39 mg/mL.*



Supplemental Figure S3.7: *Calibration lines (blue dotted line) for each protein at pH = 5, the blue dots indicate the experimental data. The concentrations are measured at an Absorbance of 280 and 400 nm. 400 nm absorbance is specifically needed to quantify cytochrome C.*



Supplemental Figure S3.8: *Calibration lines (blue dotted line) for each protein at pH = 7.0, the blue dots indicate the experimental data. The concentrations are measured at an Absorbance of 280 and 400 nm. 400 nm absorbance is specifically needed to quantify cytochrome C.*

Table S3.1: *Optimization results using the QSPR-based method, showing the performance measurements and obtained optimized variables.* K_{eq} *= 0.028 and* v *= 3.05.*

	Purity (%)	Yield (%)	HCP clearance (%)	Product concentration (g/L)	Lower cut point (%)	Upper cut point (%)	Initial salt concentration (mM)	Final salt concentration (mM)
1	74.33	97.50	79.79	0.32	4.4	91.7	14.8	330.4
2	73.66	97.81	79.01	0.30	3.7	92.8	19.8	324.5
3	73.91	97.68	79.31	0.30	4.2	93.0	24.4	327.9
4	74.23	97.48	79.69	0.34	4.7	92.3	17.7	354.7
5	74.44	97.40	79.93	0.31	4.4	90.9	18.0	325.9
Maximum difference	0.78	0.41	0.92	0.03	0.9	2.1	9.6	30.2

Table S3.2: *Optimization results using the experimental-based method, showing the performance measurements and obtained optimized variables.* K_{eq} *= 0.071 and* v *= 2.37.*

	Purity (%)	Yield (%)	HCP clearance (%)	Product concentration (g/L)	Lower cut point (%)	Upper cut point (%)	Initial salt concentration (mM)	Final salt concentration (mM)
1	74.63	96.30	80.36	0.30	7.69	91.21	24.54	320.58
2	74.09	96.62	79.72	0.29	8.54	91.78	22.14	320.00
3	74.22	96.50	79.88	0.29	8.32	91.91	36.47	321.72
4	74.45	96.44	80.15	0.30	8.54	90.80	23.90	320.85
5	74.59	96.38	80.30	0.30	7.99	91.94	28.55	320.13
Maximum difference	0.50	0.23	0.58	0.005	0.85	1.14	14.33	1.72

Chapter 4

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

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,*

**Authors contributed equally*

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.

4.1 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.[2,3]

For the purification of protein-based pharmaceuticals, packed bed chromatography has been the industry standard due to its high versatility and specificity.[4] 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.[5,6] In recent years, tools like high throughput experimentation and *in silico* modeling have shown great potential to accelerate the design process.[7–10] 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.[11,12] Over the last 20 years, successful models have been trained for a variety of globular proteins or antibodies.[13–18] 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.[19] These prediction methods become even more powerful in combination with mechanistic modeling,

allowing full prediction of the elution profile.[17,20] 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.[21,22] Additionally, QSPR requires accurate structures of the HCPs, which are not always available. Recent advances in protein structure prediction by tools like AlphaFold allow for construction of missing HCP structures [23]. While promising, the accuracy and confidence of HCPs which are poorly annotated can be problematic and should therefore be assessed critically.

Describing the HCP content of various expression host has been of interest in the last two decades.[24–26] Mass spectrometry-based proteomics (MS) has gained popularity for analyzing HCPs, enabling the sensitive detection of individual HCPs during process development.[25,27–30] Advances in the field allow identification of specific proteins which are commonly remaining after the downstream processing [31]. 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.[32–35] From these, high-risk HCPs have been identified for CHO, that have potential immunogenic responses or compromise product quality due to degradation.[33] Studies showed that HCP aggregates with mAbs may promote the persistence of HCPs during the protein A capture step.[36–39] A recent correlation analysis of HCPs identified co-elution of HCPs in groups that are associated with PPIs.[35]

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.[40] More recently, Lingg et al., investigated the effect of metal and chelator type on the HCPs found in the eluate of a similar process.[41] For cation- and anion-exchange chromatography, Swanson et al., studied *E. coli* HCP elution in a 5-step isocratic elution.[42,43] 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 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.[44]

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 4.1). 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 retrieved 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.

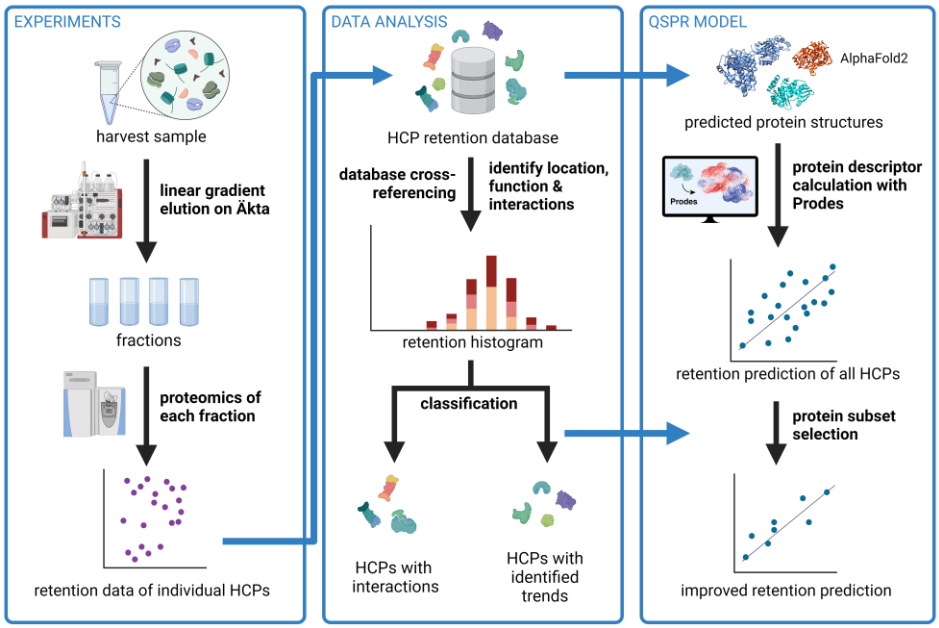


Figure 4.1: *Schematic overview of this study. Chromatographic experiments are conducted using the lysate containing a mixture of host cell proteins (HCPs). The protein mixture is injected to the Äkta chromatography system and linear gradient elution experiments on IEX and HIC are conducted. From each 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.).*

4.2 Materials and methods

4.2.1 Chromatographic experiments and proteomic analysis

4.2.1.1 *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.[44] 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.

4.2.1.2 Linear gradient elution experiments

After injection of 1 ml of the dialyzed clarified harvest sample the column was washed with 5 column volumes 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 volumes (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 [40]. For the HIC experiment, 2.5 ml fractions were taken and every fraction was analyzed.

4.2.1.3 Proteomic analysis

Shotgun proteomics to identify individual *E. coli* proteins in each of the analyzed fractions from the LGE experiments was performed using LC-MS as described in [40].

4.2.1.4 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 [45]. Since a constant flow rate was used in the experiments, the dimensionless retention time (DRT) could be calculated as

$$DRT(i) = \frac{V_{R,i} - V_g}{V_G - V_g}, \tag{4.1}$$

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 [46].

Abundance measures (for the common scatter plot) and theoretical physicochemical properties were retrieved from a previous study of the harvest sample [44]. The cellular location and functions were retrieved from UniProt [47]. 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 [19] that show PPIs (Supplemental Table 1 in [19]) and proteins without measured interactions (Supplemental Table 2[19]).

4.2.2 QSPR

4.2.2.1 Protein model generation

Using the database presented in[44] the amino acid sequence for each identified protein was retrieved. From the sequences, protein structures were predicted using AlphaFold2 to ensure full sequence coverage in the structure.[50] 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.[51] Visualization of the protein structures was performed using UCSF Chimera [52].

4.2.2.2 QSPR model training

Multi Linear Regression (MLR) models were trained for the retention time prediction of the whole dataset and specific subsets of HCPs. 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 Table S4.1), were selected based on the R^2 of a 10-fold cross-validation.

4.3 Results and discussion

4.3.1 Retention behavior of individual host cell proteins

4.3.1.1 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 4.2). As determination of protein abundance remains cumbersome using shotgun proteomics, relative abundance using peak area and the protein abundance index (PAI) were used (Figure 4.2a and Figure 4.2b, 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 [53]. 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 4.2c and Figure 4.2d, 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 [39]. 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.

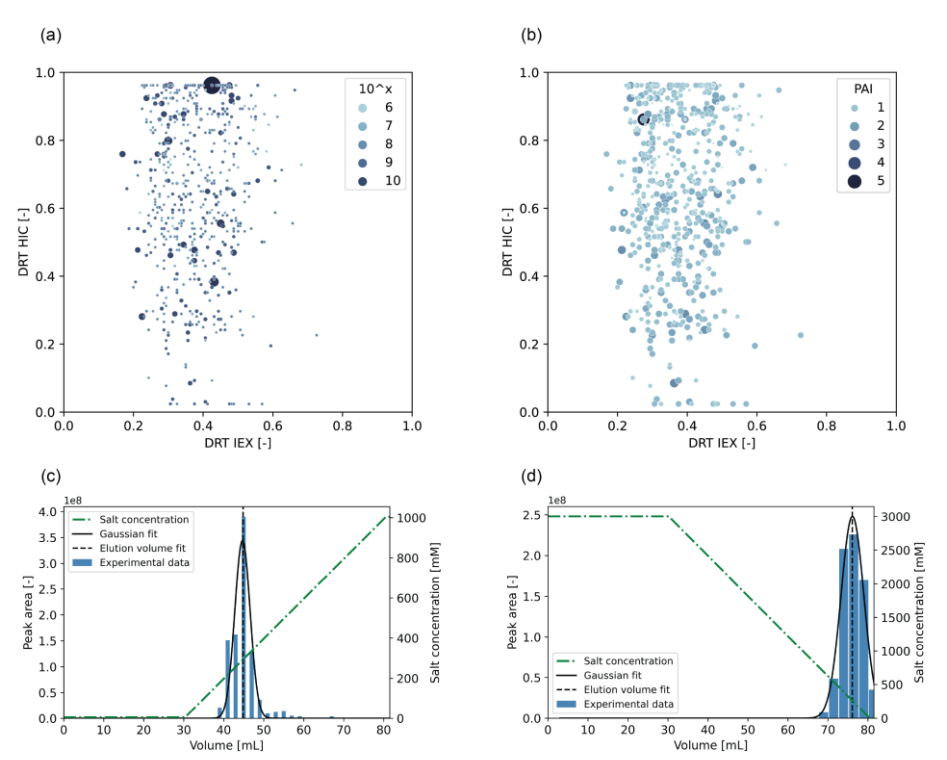


Figure 4.2: *Host cell protein (HCP) retention map of individual HCPs in the E. coli lysate. Dimensionless retention times (DRTs) were obtained from MS analysis of fractions obtained from linear gradient experiments on Q Sepharose XL (IEX) and Butyl FF (HIC) HiTrap 5 ml columns at pH 7 using NaCl as 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.*

4.3.1.2 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 4.3a&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 S4.1a), 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) [54]. 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.[55] However, the joint elution of several plasma membrane proteins might indicate that they form liposomes or are parts of membrane vesicles [56]. 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.

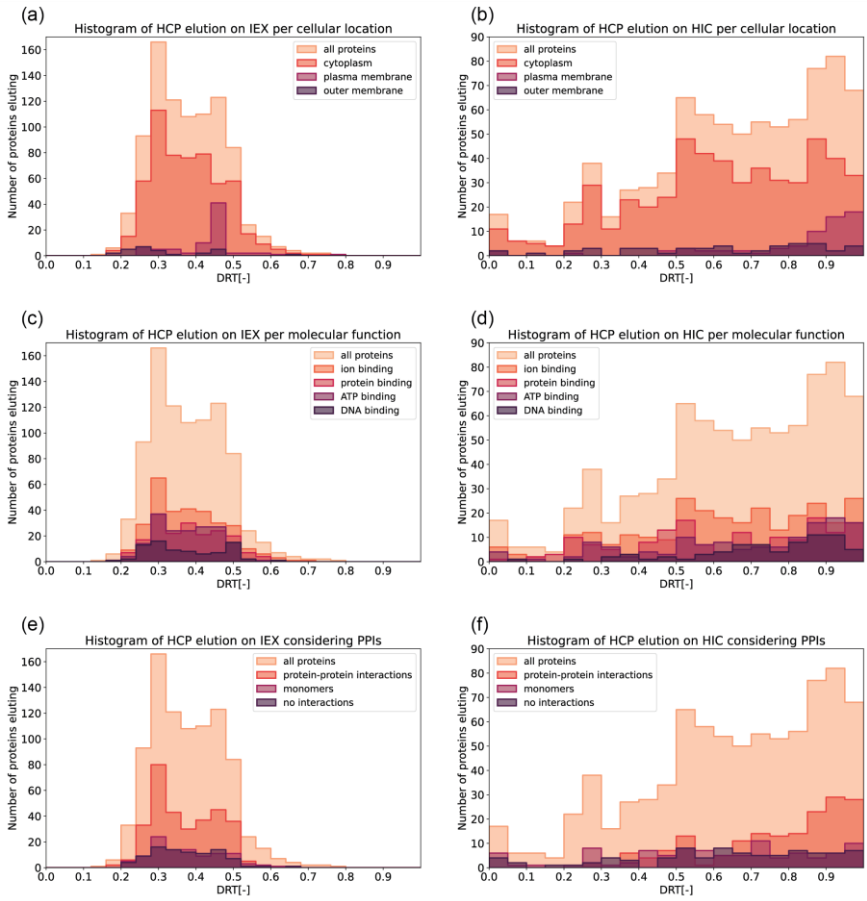


Figure 4.3: *Histograms representing the elution of groups of host cell proteins (HCPs). The number of proteins with an elution maximum during a specific dimensionless retention time (DRT) is listed for ion exchange (IEX) and hydrophobic interaction chromatography (HIC). (a) histogram of cellular location 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.*

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 4.3b). 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 to be 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.[57] This complicates the retention prediction substantially, leaving room for future studies to develop new features to describe flexibility and local aggregation propensities, influencing protein retention in HIC.

4.3.1.3 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 S1b).

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 3d).

4.3.1.4 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..[49] 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.[58] Nevertheless, these interactions pose an interesting effect on the DRTs of involved HCPs as indicated in a recent study for CHO cells [35].

To identify the effect of PPIs, proteins described to interact from protein pairs in proximity were selected (Figure 4.3e&f). Proteins described as having 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.[59] For the monomers (98) and

non-interacting proteins (80) no such trend was observed as these elute throughout the gradient.

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.

4.3.2 Prediction of retention time of individual HCPs in IEX

4.3.2.1 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 built achieving a 10-fold cross validated R^2 of 0.55 and a mean absolute error (MAE) of 0.049 (Figure 4.4 and Table 4.1 [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 4.4). 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 Table S4.2). 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 S4.2). The resulting model was not able to predict scrambled protein retention (R^2 of -0.065) proving physical validity.

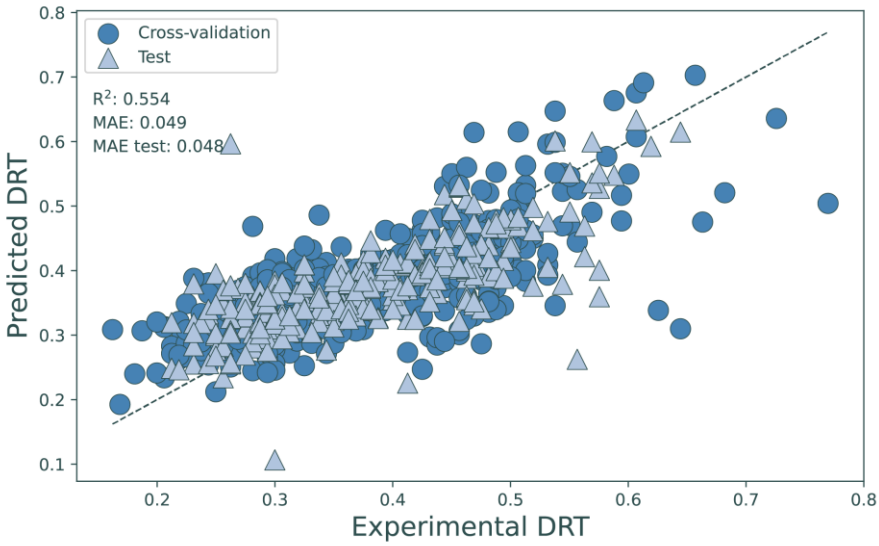


Figure 4.4: *QSPR validation of the regression model trained to predict DRT, where the circles represent the 10-fold cross-validation and the triangles the test set.*

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 to be due to the nonspecific protein interactions at high salt conditions and partial unfolding upon binding which often occur [60]. As was mentioned

in 4.3.1.2, no correlation was found with HIC elution and any of the hydrophobicity features for the full dataset nor any subsets.

4.3.2.2 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 4.1, Figure 4.5). 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 4.1 and Figure 4.5a) 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 4.1a). In contrast, 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 2a).

The subset containing the proteins without PPIs were found to elute according to a normal distribution (Figure 4.3e), 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 4.5b, Table 4.1). 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.

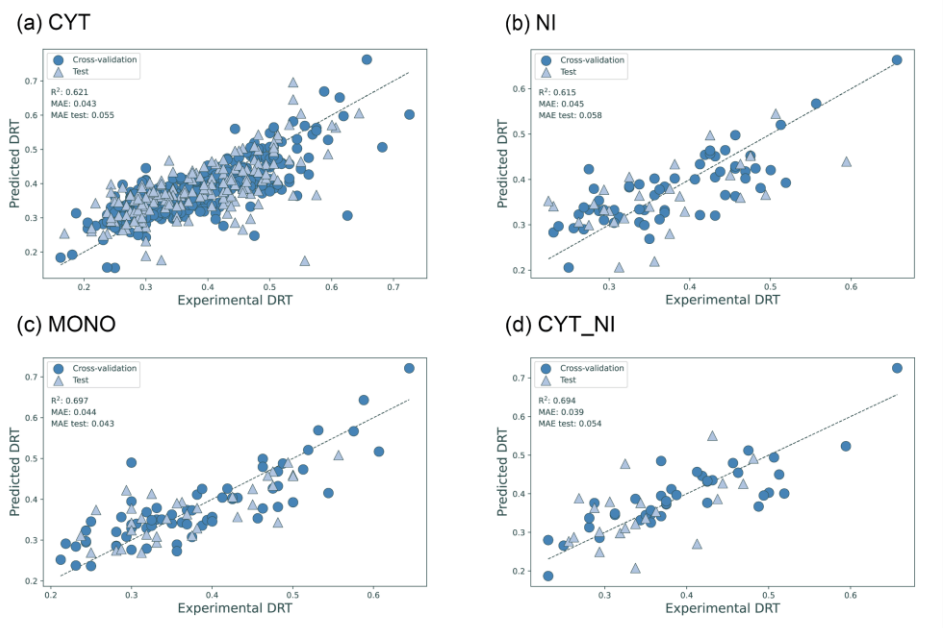


Figure 4.5: *QSPR validation of the regression model trained to predict DRT of protein subsets, where the circles represent the 10-fold cross-validation and the triangles the test set. The presented subsets are the cytosolic proteins (a), the proteins without interactions (b), proteins reported to be present as monomers (c) and proteins which are cytosolic and non-interacting (d).*

Table 4.1: *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

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 4.1, Figure 4.5c). 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, no major influence was observed. For this model, the average and sum of the negative electrostatic potential were found to be most important, as removing either feature resulted in a cross validated R^2 of 0.47 (Supplemental Table S4.5). The increased accuracy of the 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.[61] 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, current efforts in setting confident thresholds for the predicted structures did not yield more accurate retention prediction models (Supplemental Figure S4.4).

Nevertheless, this work provides an important step towards holistic in silico process design. In contrast to recent literature, the retention data used in this work is obtained from a clarified lysate. The increased uncertainty pared with the heterogeneity results complicates the predictive modeling compared to the use of model proteins. The

achieved cross validated R^2 of 0.697 for the monomer subset approaches recent work on the retention prediction of mAbs (0.780-0.835) and model proteins for a range of ligands (0.79-0.82).[62–64] It can therefore be expected that additional research on the algorithms and HCP understanding will allow for robust prediction of HCP retention prediction and knowledge transfer between different processes.

4.4 Conclusion 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, extensivelots 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 are required. Ultimately, this research represents a significant step towards in silico driven process development, increasing process understanding and reducing development times.

4.5 References

1. Bracewell, D. G., Francis, R. & Smales, C. M. The future of host cell protein (HCP) identification during process development and manufacturing linked to a risk-based management for their control. *Biotechnol Bioeng* 112, 1727–1737 (2015).
2. Tscheliessnig, A. L., Konrath, J., Bates, R. & Jungbauer, A. Host cell protein analysis in therapeutic protein bioprocessing - methods and applications. *Biotechnol J* 8, 655–670 (2013).
3. Bracewell, D. G., Francis, R. & Smales, C. M. The future of host cell protein (HCP) identification during process development and manufacturing linked to a risk-based management for their control. *Biotechnol Bioeng* 112, 1727–1737 (2015).
4. Gottschalk, U., Brorson, K. & Shukla, A. A. The need for innovation in biomanufacturing. *Nat Biotechnol* 30, 489–492 (2012).
5. Keulen, D., Geldhof, G., Bussy, O. Le, Pabst, M. & Ottens, M. Recent advances to accelerate purification process development: A review with a focus on vaccines. *J Chromatogr A* 1676, 463195 (2022).
6. Hanke, A. T. & Ottens, M. Purifying biopharmaceuticals: Knowledge-based chromatographic process development. *Trends Biotechnol* 32, 210–220 (2014).
7. Nfor, B. K. *et al.* Multi-dimensional fractionation and characterization of crude protein mixtures: Toward establishment of a database of protein purification process development parameters. *Biotechnol Bioeng* 109, 3070–3083 (2012).
8. Pirrung, S. M. *et al.* Chromatographic parameter determination for complex biological feedstocks. *Biotechnol Prog* 34, 1006–1018 (2018).
9. Bernau, C. R., Knödler, M., Emonts, J., Jäpel, R. C. & Buyel, J. F. The use of predictive models to develop chromatography-based purification processes. *Front Bioeng Biotechnol* 10, 1–24 (2022).
10. Keulen, D. *et al.* Using artificial neural networks to accelerate flowsheet optimization for downstream process development. *Biotechnol Bioeng* 121, 2318–2331 (2024).
11. Emonts, J. & Buyel, J. F. An overview of descriptors to capture protein properties – Tools and perspectives in the context of QSAR modeling. *Comput Struct Biotechnol J* 21, 3234–3247 (2023).
12. Bernau, C. R., Knödler, M., Emonts, J., Jäpel, R. C. & Buyel, J. F. The use of predictive models to develop chromatography-based purification processes. *Front Bioeng Biotechnol* 10, 1–24 (2022).

13. Yang, T., Sundling, M. C., Freed, A. S., Breneman, C. M. & Cramer, S. M. Prediction of pH-dependent chromatographic behavior in ion-exchange systems. *Anal Chem* 79, 8927–8939 (2007).

14. Hess, R. *et al.* Predicting multimodal chromatography of therapeutic antibodies using multiscale modeling. *J Chromatogr A* 1718, 464706 (2024).

15. Hanke, A. T. *et al.* Prediction of protein retention times in hydrophobic interaction chromatography by robust statistical characterization of their atomic-level surface properties. *Biotechnol Prog* 32, 372–381 (2016).

16. Mazza, C. B., Sukumar, N., Breneman, C. M. & Cramer, S. M. Prediction of protein retention in ion-exchange systems using molecular descriptors obtained from crystal structure. *Anal Chem* 73, 5457–5461 (2001).

17. Saleh, D. *et al.* A multiscale modeling method for therapeutic antibodies in ion exchange chromatography. *Biotechnol Bioeng* 120, 125–138 (2023).

18. Kittelmann, J., Lang, K. M. H., Ottens, M. & Hubbuch, J. Orientation of monoclonal antibodies in ion-exchange chromatography: A predictive quantitative structure–activity relationship modeling approach. *J Chromatogr A* 1510, 33–39 (2017).

19. Cai, Q. Y., Qiao, L. Z., Yao, S. J. & Lin, D. Q. Machine learning assisted QSAR analysis to predict protein adsorption capacities on mixed-mode resins. *Sep Purif Technol* 340, 126762 (2024).

20. Hess, R. *et al.* Predicting multimodal chromatography of therapeutic antibodies using multiscale modeling. *J Chromatogr A* 1718, 464706 (2024).

21. Rajagopala, S. V. *et al.* The binary protein-protein interaction landscape of escherichia coli. *Nat Biotechnol* 32, 285–290 (2014).

22. Arifuzzaman, M. *et al.* Large-scale identification of protein-protein interaction of Escherichia coli K-12. *Genome Res* 16, 686–691 (2006).

23. Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. & Moult, J. Critical assessment of methods of protein structure prediction (CASP)—Round XV. *Proteins: Structure, Function and Bioinformatics* 91, 1539–1549 (2023).

24. Wang, X., Hunter, A. K. & Mozier, N. M. Host cell proteins in biologics development: Identification, quantitation and risk assessment. *Biotechnol Bioeng* 103, 446–458 (2009).

25. Tscheliessnig, A. L., Konrath, J., Bates, R. & Jungbauer, A. Host cell protein analysis in therapeutic protein bioprocessing - methods and applications. *Biotechnol J* 8, 655–670 (2013).

26. Timmick, S. M. *et al.* An impurity characterization based approach for the rapid development of integrated downstream purification processes. *Biotechnol Bioeng* 115, 2048–2060 (2018).

27. Bracewell, D. G., Francis, R. & Smales, C. M. The future of host cell protein (HCP) identification during process development and manufacturing linked to a risk-based management for their control. *Biotechnol Bioeng* 112, 1727–1737 (2015).

28. Schenauer, M. R., Flynn, G. C. & Goetze, A. M. Identification and quantification of host cell protein impurities in biotherapeutics using mass spectrometry. *Anal Biochem* 428, 150–157 (2012).

29. Rathore, D. *et al.* The role of mass spectrometry in the characterization of biologic protein products. *Expert Rev Proteomics* 15, 431–449 (2018).

30. *Development, Design, and Implementation of Manufacturing Processes*. (John Fedor, 2018).

31. Molden, R. *et al.* Host cell protein profiling of commercial therapeutic protein drugs as a benchmark for monoclonal antibody-based therapeutic protein development. *MAbs* 13, (2021).

32. Migani, D., Smales, C. M. & Bracewell, D. G. 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* 33, 666–676 (2017).

33. Jones, M. *et al.* "High-risk" host cell proteins (HCPs): A multi-company collaborative view. *Biotechnol Bioeng* 118, 2870–2885 (2021).

34. Vanderlaan, M. *et al.* Experience with host cell protein impurities in biopharmaceuticals. *Biotechnol Prog* 34, 828–837 (2018).

35. Panikulam, S. *et al.* Host cell protein networks as a novel co-elution mechanism during protein A chromatography. *Biotechnol Bioeng* https://doi.org/10.1002/bit.28678 (2024) doi:10.1002/bit.28678.

36. Oh, Y. H. *et al.* Characterization and implications of host-cell protein aggregates in biopharmaceutical processing. *Biotechnol Bioeng* 120, 1068–1080 (2023).

37. Herman, C. E. *et al.* Behavior of host-cell-protein-rich aggregates in antibody capture and polishing chromatography. *J Chromatogr A* 1702, 464081 (2023).

38. Herman, C. E. *et al.* Analytical characterization of host-cell-protein-rich aggregates in monoclonal antibody solutions. *Biotechnol Prog* 39, 1–16 (2023).

39. Gagnon, P. *et al.* Nonspecific interactions of chromatin with immunoglobulin G and protein A, and their impact on purification performance. *J Chromatogr A* 1340, 68–78 (2014).

40. Bartlow, P. *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* 78, 216–224 (2011).

41. Lingg, N. *et al.* Proteomics analysis of host cell proteins after immobilized metal affinity chromatography: Influence of ligand and metal ions. *J Chromatogr A* 1633, 461649 (2020).

42. Swanson, R. K., Xu, R., Nettleton, D. S. & Glatz, C. E. Accounting for host cell protein behavior in anion-exchange chromatography. *Biotechnol Prog* 32, 1453–1463 (2016).

43. Swanson, R. K., Xu, R., Nettleton, D. & Glatz, C. E. Proteomics-based, multivariate random forest method for prediction of protein separation behavior during cation-exchange chromatography. *J Chromatogr A* 1249, 103–114 (2012).

44. 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. *Biotechnol J* 18, 2300068 (2023).

45. Disela, R. *et al.* Proteomics-based method to comprehensively model the removal of host cell protein impurities. *Biotechnol Prog* https://doi.org/10.1002/btpr.3494 (2024) doi:10.1002/btpr.3494.

46. Hanke, A. T. *et al.* Prediction of protein retention times in hydrophobic interaction chromatography by robust statistical characterization of their atomic-level surface properties. *Biotechnol Prog* 32, 372–381 (2016).

47. Bateman, A. *et al.* UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* 49, D480–D489 (2021).

48. Arifuzzaman, M. *et al.* Large-scale identification of protein-protein interaction of Escherichia coli K-12. *Genome Res* 16, 686–691 (2006).

49. Arifuzzaman, M. *et al.* Large-scale identification of protein-protein interaction of Escherichia coli K-12. *Genome Res* 16, 686–691 (2006).

50. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589 (2021).

51. Neijenhuis, T., Le Bussy, O., Geldhof, G., Klijn, M. E. & Ottens, M. Predicting protein retention in ion-exchange chromatography using an open source QSPR workflow. *Biotechnol J* 19, e2300708 (2024).

52. Pettersen, E. F. *et al.* UCSF Chimera - A visualization system for exploratory research and analysis. *J Comput Chem* 25, 1605–1612 (2004).

53. Rappsilber, J., Ryder, U., Lamond, A. I. & Mann, M. Large-scale proteomic analysis of the human spliceosome. *Genome Res* 12, 1231–1245 (2002).

54. Maddalo, G. *et al.* Systematic analysis of native membrane protein complexes in Escherichia coli. *J Proteome Res* 10, 1848–1859 (2011).

55. Kumazaki, K. *et al.* Crystal structure of Escherichia coli YidC, a membrane protein chaperone and insertase. *Sci Rep* 4, 1–6 (2014).

56. Nagakubo, T., Nomura, N. & Toyofuku, M. Cracking Open Bacterial Membrane Vesicles. *Front Microbiol* 10, (2020).

57. Hanke, A. T. *et al.* Prediction of protein retention times in hydrophobic interaction chromatography by robust statistical characterization of their atomic-level surface properties. *Biotechnol Prog* 32, 372–381 (2016).

58. Jakob, L. A. *et al.* Protein-protein interactions and reduced excluded volume increase dynamic binding capacity of dual salt systems in hydrophobic interaction chromatography. *J Chromatogr A* 1649, 462231 (2021).

59. O'Farrell, P. A. *Molecular Biomethods Handbook*. *Molecular biomethods handbook* (Humana Press, Totowa, NJ, 2008). doi:10.1007/978-1-60327-375-6.

60. Jakob, L. A. *et al.* Protein-protein interactions and reduced excluded volume increase dynamic binding capacity of dual salt systems in hydrophobic interaction chromatography. *J Chromatogr A* 1649, 462231 (2021).

61. Soleymani, F., Paquet, E., Viktor, H., Michalowski, W. & Spinello, D. Protein–protein interaction prediction with deep learning: A comprehensive review. *Comput Struct Biotechnol J* 20, 5316–5341 (2022).

62. Hess, R. *et al.* Predicting multimodal chromatography of therapeutic antibodies using multiscale modeling. *J Chromatogr A* 1718, 464706 (2024).

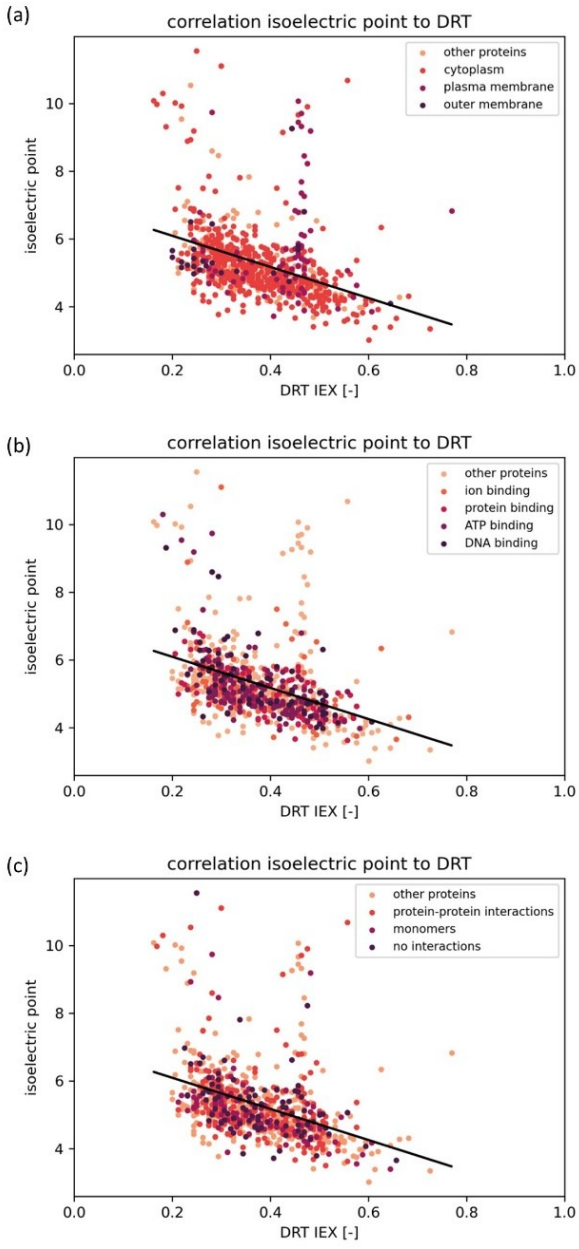
63. Cai, Q. Y., Qiao, L. Z., Yao, S. J. & Lin, D. Q. Machine learning assisted QSAR analysis to predict protein adsorption capacities on mixed-mode resins. *Sep Purif Technol* 340, 126762 (2024).

64. Hess, R. *et al.* Antibody sequence-based prediction of pH gradient elution in multimodal chromatography. *J Chromatogr A* 1711, 464437 (2023).

4.6 Supplemental material

Supplemental Table S4.1: *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.*

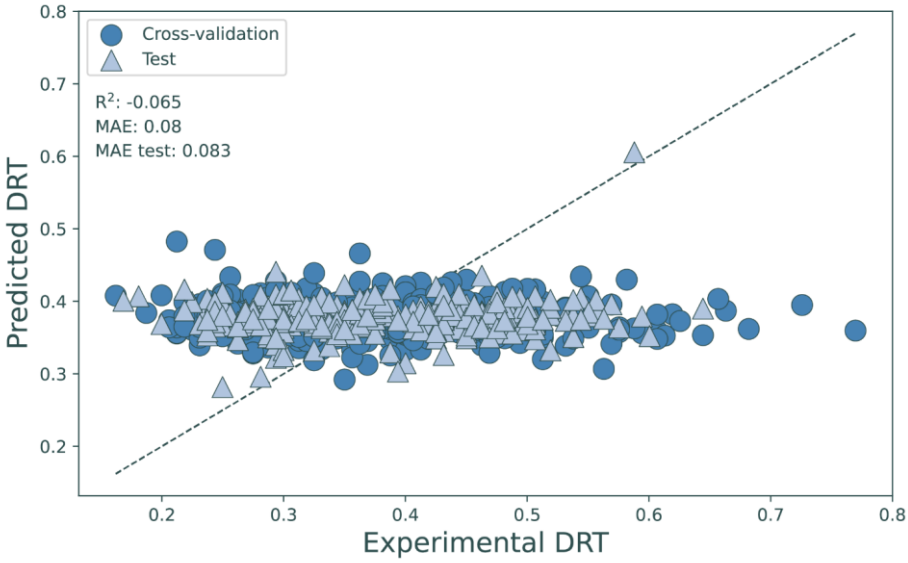
Model	Feature – feature filter	Feature – observation filter (%)
ALL	0.99	100
CYT	1	100
NI	1	100
MONO	1	50
CYT_NI	0.9	100



Supplemental Figure S4.1: *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*

Supplemental Table S4.2: *Regression coefficient and permutation performances for the linear regression model predicting DRT for all HCPs.*

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 Figure S4.2: *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.*

Supplemental Table S4.3: *Regression coefficient and permutation performances for the linear regression model predicting DRT for the CYS subset.*

Descriptor	Coefficient	Permutation R2
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 S4.4: *Regression coefficient and permutation performances for the linear regression model predicting DRT for the NI subset.*

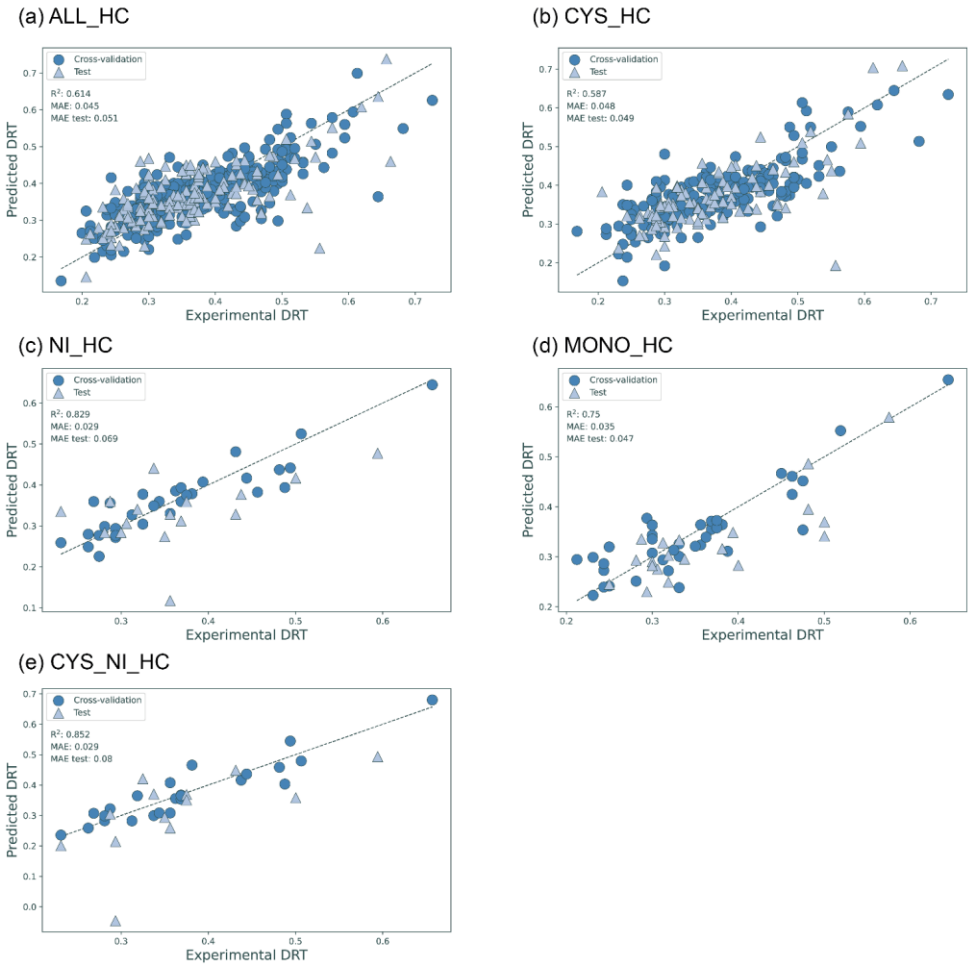
Descriptor	Coefficient	Permutation R2
SurfEpMinAverage	-0.2562	0.5351
SurfPosMhpsum	0.0747	0.6152
PRO surface fraction	-0.1570	0.4696
SurfMhpMax	0.0904	0.5024
SurfPosEpStdFormal	-0.1244	0.5940
TYR surface fraction	0.0969	0.5765
CYS surface fraction	0.0793	0.5528
Surface shape max	-0.0670	0.5700
LYS surface fraction	-0.0885	0.5658
SurfEpStdAverage	0.0730	0.5984
intercept	0.5735	

Supplemental Table S4.5: *Regression coefficient and permutation performances for the linear regression model predicting DRT for the MONO subset.*

Descriptor	Coefficient	Permutation R2
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 S4.6: *Regression coefficient and permutation performances for the linear regression model predicting DRT for the CYS_NI subset.*

Descriptor	Coefficient	Permutation R2
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 Figure S4.3: *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.*

Supplemental Table S4.7: *Regression coefficient and permutation performances for the linear regression model predicting DRT for the ALL_HC subset.*

Descriptor	Coefficient	Permutation R2
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 S4.8: *Regression coefficient and permutation performances for the linear regression model predicting DRT for the CYS_HC subset.*

Descriptor	Coefficient	Permutation R2
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 S4.9: *Regression coefficient and permutation performances for the linear regression model predicting DRT for the NI_HC subset.*

Descriptor	Coefficient	Permutation R2
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 S4.10: *Regression coefficient and permutation performances for the linear regression model predicting DRT for the MONO_HC subset.*

Descriptor	Coefficient	Permutation R2
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 S4.11: *Regression coefficient and permutation performances for the linear regression model predicting DRT for the CYS_NI_HC subset.*

Descriptor	Coefficient	Permutation R2
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	

Chapter 5

Using generalized quantitative structure property relationship (QSPR) models to predict host cell protein retention in ion-exchange chromatography

Published as:

Neijenhuis, T., Le Bussy, O., Geldhof, G., Klijn, M. E., & Ottens, M. (2025). Using generalized quantitative structure–property relationship (QSPR) models to predict host cell protein retention in ion-exchange chromatography. Journal of Chemical Technology & Biotechnology.

Abstract

Selecting the optimal chromatography resin during biopharmaceutical downstream process development is a great challenge. Especially for recombinant sub-unit vaccines, where product properties vary greatly and recovery often involves cell lysis, which yields a complex mixture of different host cell materials. Host cell protein (HCP) impurities may remain similar for platform processes, but their critical impact on separation efficiency is relative to specific product properties. Therefore, every process needs to be designed per product. Prior knowledge on the elution behavior of HCPs would support the identification of critical compounds. However, determining chromatographic behavior of HCPs experimentally is a time-consuming approach. In this work, we leverage quantitative structure property relationship (QSPR) models calibrated with retention data of 13 commercial proteins, collected at pH 7, 8, 9, and 10 to predict the anion exchange (AEX) retention of *Escherichia coli* HCPs. These models use features calculated from the molecular structure to describe protein behavior, like chromatographic retention. A multi linear regression model containing two features (Isoelectric point and sum of negative surface electrostatics) was able to predict the retention times of 288 HCPs accurately (error ≤5%). Moreover, we identified the key attributes missing in the training dataset, which is important to increase model performance in the future. This work showcases how chromatographic data obtained using commercial proteins can be translated to a clarified *E. coli* lysate to accelerate chromatography resin selection for new products.

5.1 Introduction

Recombinant proteins constitute approximately 80% of the global sales in pharmaceutical industry.[1] To ensure safety and efficacy of these pharmaceuticals, sufficient product purity (reviewed case-by-case) is required.[2] This is achieved by the downstream processing (DSP) that often involves a sequence of chromatographic steps separating the target protein from process and host cell impurities.[3–5] While product related impurities are often most difficult to remove, host cell proteins (HCPs) are a class of impurities that are also challenging to eliminate sufficiently. The main reason for this is that conventionally, HCP impurities are treated as one entity, while these are actually individual entities with a wide variety in physicochemical properties. Therefore knowledge on persistent HCPs is valuable to guide the DSP design.[6] As co-purification is a risk, highly sensitive biochemical methods for detection of persistent HCPs have been developed [7,8], including identification and quantification by LC-MS/MS proteomics.[9] The relevance of these techniques is reflected by a comprehensive list of high-risk HCPs for monoclonal antibody (mAb) production in Chinese hamster ovary cells.[10] This information can accelerate DSP design in platform processes as different mAbs products have relatively similar properties that affect purification.[11] This means the criticality of HCPs does not change for new products. Unfortunately, DSP design is less straightforward for other recombinant proteins, such as subunit vaccines.[12] Unlike for mAbs, affinity chromatography is rarely available for subunit vaccines, as their properties vary widely. Additionally, formulation of standardized sets of HCPs that are likely to coelute during a chromatography step is impossible. To increase process understanding, Disela et al. analyzed the HCP content of *Escherichia coli* lysates from different strains and expression vectors [13]. The HCP content was found to be 80% to 90% similar between lysates, leading to the use of HCP property maps to guide DSP design.

These property maps allow for the identification of potential critical HCPs by comparing their properties to the properties of the subunit vaccine.

An alternative to the property maps are quantitative structure property relationship (QSPR) models that correlate protein properties to behavior under specific conditions. These models use features calculated from the molecular structure in regression or classification algorithms [14]. In In the last 25 years, a wide range of regression methods have been applied to predict the chromatographic behavior of proteins, including multi linear regression (MLR)[15–20], partial least squares[21,22], support vector machines[23–26], random forests[27,28], and Gaussian process regressions[29–31]. While traditional QSPR models predict chromatographic behavior of proteins for a specific resin, Cai et al. demonstrated a QSPR analysis combining both protein and ligand features to predict the protein adsorption on different mixed-mode resins reaching a cross validated R^2 of 0.8.[27] More recently, Hartmann et al. trained QSPR models for predicting the partition coefficient by including protein, resin (ion-exchange, hydrophobic interaction, and mixed-mode), and mobile phase features.[32] Their models were trained for therapeutic proteins in their native and high molecular weight form, and were able to predict low, medium, and high binding conditions with 93-95% accuracy.

Unfortunately, most QSPR models trained to predict protein chromatographic behavior have only been validated for purified proteins. This makes it challenging to assess their accuracy for complex mixtures, such as host cell lysates, where many interactions occur that potentially change protein retention behavior. An example of more complex mixtures is the study by Keulen et al., where QSPR models were successfully trained for the prediction of ion exchange chromatography (IEX) retention of proteins in three component mixtures.[19] However, the total protein concentration of 2.5 g/L used

in this study is considered insufficient for notable protein interactions. A more representative complex mixture was used by Buyel et al..[28] Here, QSPR models were trained on protein elution salt concentrations reported in literature to predict the retention of tobacco HCPs in IEX and mixed mode chromatography. Estimated elution profiles of 67 HCPs were combined and compared to an experimental chromatogram of a clarified extract, where a good agreement for SP Sepharose FF was found. Unfortunately, accuracy of specific HCPs could not be quantified as the experimental data does not provide elution behavior of specific proteins. Disela et al. performed a more quantitative study on a clarified lysate of the *E. coli* expression host, where fractions were collected from linear gradient experiments and analyzed by LC-MS/MS.[20,33] Such detailed experimental characterization provides valuable data, but the studies are time and resource intensive. These efforts could be minimized by training QSPR models with data obtained for readily available (commercial) proteins and subsequently transfer the model for the prediction chromatographic behavior of HCPs in complex mixtures.

To this end, there is limited knowledge on translating models trained on purified proteins towards complex host cell lysates. Therefore, we explored the transferability of a QSPR model trained on commercial proteins for the prediction of HCPs retention in anion exchange chromatography (AEX). A QSPR model was trained using linear gradient elution data for 13 proteins on a Q Sepharose XL column as used by Disela et. al..[20] We defined the performance of these models by testing different subsets of HCPs (including all or only monomeric HCPs) to identify the current limits of this approach. The work described in this study is a significant step towards generalizability in QSPR model application, thereby contributing to faster model deployment and cost-effective process development.

5.2 Methods

5.2.1 Materials and Equipment

The retention experiments were performed on two separate Äkta pure systems (Cytiva, Marlborough*, USA*). Both systems were equipped with a prepacked HiTrap Q Sepharose XL 1 mL column (Cytiva, Marlborough*, USA*) (Supplemental Table S1). All substances were purchased from Sigma-Aldrich (Saint Louis, USA) and buffers were prepared using ultrapure water filtered with the Milli-Q Advantage A10 (Merck Millipore, Burlington, USA). Buffer solutions at pH 7, 8, 9, and 10 were prepared with 20 mM NaCl (Buffer A) and 1000 mM NaCl (Buffer B) for running and elution. For pH 7 and 8, a 20 mM Tris-HCl solution was made, while for pH 9 and 10, 20 mM Ethanolamine was used. pH was adjusted by titration with 1 M sodium hydroxide or 1 M hydrochloric acid. All buffers were filtered using a 0.2 μm Membrane Disc Filter (Pall corporation, New York, USA) followed by 20 minutes of sonication.

Albumin (Bovine), albumin (Human), pepsin, trypsin inhibitor A, lipase, α-lactalbumin, β-lactoglubulin a, glucose oxidase, lipoxygenase, ovotransferrin, amyloglucosidase, urease and catalase were purchased from Sigma-Aldrich (Saint Louis, USA). Each protein was dissolved in buffer A to reach a concentration of 2 g/L, after which the solutions were filtered using a 0.22 μm Whatman Puradisc FP 30 mm (*Cytiva,* Marlborough*, USA*).

5.2.2 Linear gradient elution experiments and data processing

The retention times of the selected proteins were determined for a 10 column volume linear gradient elution from buffer A to buffer B. Each LGE was performed at a flowrate of 1 mL/min by injecting 200 µL protein solution followed by a 5 column volume wash with buffer A and 10 column volume gradient to 100% buffer B. Columns were

regenerated with 0.5 M NaOH and stored in 20% Ethanol. To normalize the protein retention for the two systems, the normalized retention times (V_R) were calculated as

$$V_R = V_{R,0} - 0.5V_{inj} - V_d - V_m - V_{wash}, \quad (5.1)$$

Where $V_{R,0}$ is the initial retention time, V_{inj} is the injection volume , V_d is the dwell volume, V_m is the column void volume and V_{wash} is the volume of buffer A used between injection and start of the gradient [19,33]. Finally, to make the data column independent, and allowing the comparison of retention times obtained for 5 mL HiTrap Q Sepharose XL column, the dimensionless retention time (DRT) was calculated as

$$DRT = \frac{V_R}{V_G}, \quad (5.2)$$

Where V_G is the gradient length, which is 10 column volumes for these experiments.

5.2.3 Quantitative structure property relationship modeling

Molecular structures of the commercial proteins were retrieved from the protein data bank [34] with the exception of trypsin inhibitor A. The structure for this protein was retrieved from the AlphaFold database[35] as the experimental structures available missed the positions of some atoms. The full list of the structures used can be found in Table 5.1. For each protein the feature sets were calculated at pH 7, 8, 9, and 10 using the default settings of Prodes.[18] Feature redundancy was reduced by removing features with a Pearson correlation ≥0.9 to other features. Selection of which feature to remove was based on the cumulative cross-correlation to all other features, keeping the feature with the lowest score. The final feature set used for the multilinear regression (MLR) model was selected by sequential forward selection (SFS). Model accuracy was evaluated by k-fold cross validation, leaving out all datapoints representing one protein at a time. This was done to

reduce the risk of overfitting as pH independent features would be constant for the same protein at different pH values. The final model was tested using a dataset of *E. coli* HCP DRTs described in a previous article.[20] To make sure that the test data is similar to the training data, HCPs with any features selected for the model that were outside the range (below the minimum or above the maximum) of observed in the training data were removed from the test set.

For the purpose of identifying areas of improvement for the QSPR model, feature value distributions were compared using the Kolmogorov Smirnov test for proteins that were over predicted, under predicted, or accurately predicted.[36] These HCP groups were made depending on the residuals, calculated by:

$$r_i = y_i - \hat{y}_i, \tag{5.3}$$

where r is the residual value, y and $\hat{y}$ are the experimental and predicted value respectively. Over predicted proteins are defined as r_i <-0.1 DRT, under prediction as r_i >0.1 DRT and all other HCPs are accurately predicted. Visualization of the surface electrostatics was performed using Prodes.[18]

For the purpose of training a transferable QSPR model, 13 proteins were selected with a pI ranging from 3 to 6.8, thereby ensuring chromatographic retention in AEX. From the surface electrostatic potentials (EP), it can be observed that the surface is predominantly negatively charged, except for lipoxygenase and ovotransferrin which also show positive patches (Figure 5.1).

Table 5.1: *commercial proteins and the respective system used for LGE experiments*

Name	PDB/AF model	Molecular mass (kDa)	pI (theoretical)*	System
Bovine albumin	4F5S	66.4	5.5	2
Human albumin	1AO6	66.5	5.6	2
Pepsin	4PEP	34.5	3.0	1
Trypsin inhibitor A	AF-P01070-F1-model_v4	20.1	4.4	1
Lipase	1TRH	57.1	4.5	2
α-lactalbumin	1F6R	14.2	4.6	2
β-lactoglubulin a	1BSQ	18.3	4.6	1
Glucose oxidase	1CF3	64.1	4.9	1
Lipoxygenase	1F8N	94.4	5.9	1
Ovotransferrin	1OVT	75.8	6.6	2
Amyloglucosidase	6FRV	65.8	4.0	1
Urease	3LA4	90.7	6.0	1
Catalase	6PO0	59.8	6.8	2

*pI was calculated using Prodes

5.3 Results and Discussion

Retention times for these proteins were determined for a 10 CV gradient length (Table 5.2, Supplemental figure S5.1), similar to the experimental conditions of the HCPs published elsewhere [20]. To maximize the value of this set of proteins, the retention time was measured at pH 7, 8, 9, and 10. Two datapoints are not reported, namely lipase at pH 10 (insufficient UV signal) and catalase at pH 8 (technical error). The results show a longer retention time for higher pH values, as would be expected due to deprotonation of titratable amino acids. However, this trend was not observed for urease and lipase, where chromatographic retention remained constant while varying the pH value. In other work is reported that lysozyme displayed constant chromatographic retention for SP Sepharose resins at pH 7 and pH 9, which was attributed to a constant global charge.[37] However, in the case of urease and lipase, the global charge varies in the pH range of 7 to 10 when calculated from the molecular structure by Prodes (-15 to -28 and -18 to -24, respectively). Therefore, we hypothesize that these proteins have preferred binding orientations where the local charge does remain constant.

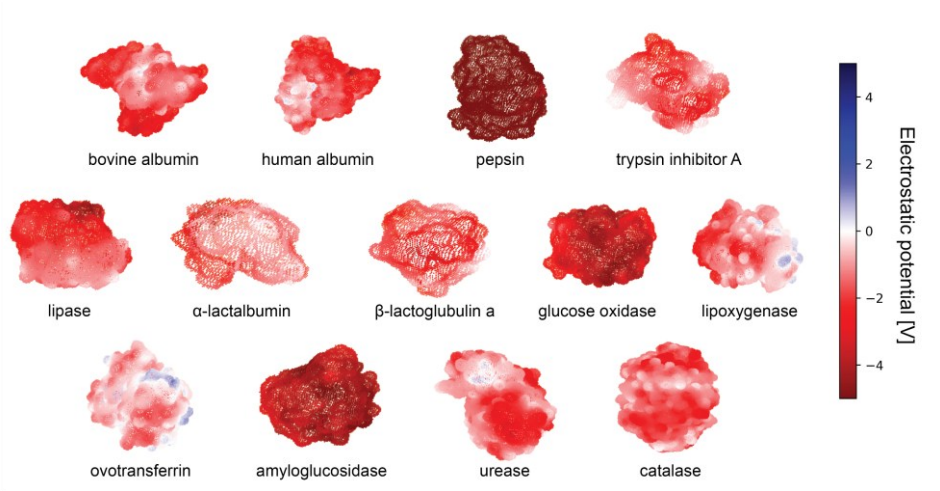


Figure 5.1: *Surface electrostatic potential maps at pH 7 of 13 commercial proteins. The blue and red color indicate positive and negative electrostatic potential (in volts), respectively.*

Table 5.2: *Experimental retention volumes (in mL) of 13 commercial proteins at pH 7, 8, 9, and 10 on a HiTrap Q Sepharose XL 1 mL column with a 10 column volume gradient from 20 to 1000 mM NaCl.*

	Retention volume [mL]			
Protein	**pH 7**	**pH 8**	**pH 9**	**pH 10**
Bovine albumin	3.42	3.95	4.34	4.51
Human albumin	3.27	3.80	4.27	4.43
Pepsin	6.53	6.50	6.77	6.83
Trypsin inhibitor A	4.38	4.53	4.75	4.83
Lipase	4.80	4.81	4.72	
α-lactalbumin	3.38	3.59	4.23	4.41
β-lactoglobulin a	4.08	4.38	4.62	4.70
Glucose oxidase	3.43	3.67	4.12	4.57
Lipoxygenase	2.69	2.99	3.39	3.63
Ovotransferrin	1.89	2.26	2.75	3.08
Amyloglucosidase	4.58	4.75	4.98	5.12
Urease	2.65	2.66	2.60	2.68
Catalase	2.39		3.26	3.93

5.3.1 Host cell protein retention prediction

Cross validation was performed by iteratively removing the retention times of each protein at all pH values from the training set to ensure that prior knowledge about the specific protein was absent during model validation. The SFS selection method resulted in a model with four features and a cross-validated R^2 of 0.927 (Supplemental figure S5.2). Of the four selected features, the protein's isoelectric point (pI)

is most important for predicting the retention time. Permutating this feature has the greatest impact on cross validation accuracy, diminishing all predictive capabilities (Supplemental Table S5.2). However, this feature is not pH dependent cannot describe any charge specific behaviors. The second most important feature, the sum of all negative surface points does capture retention changes by varying the pH. Permutation of this feature results in a significant accuracy reduction to a cross validated R^2 of 0.76. The remaining two features, the proline surface fraction and median negative surface hydrophobicity potential, have similar permutation scores of 0.88 and 0.87, respectively.

To explore the transferability of the model trained with commercially available proteins, *E. coli* HCPs were used as a test set. This data set consists of features for 836 HCPs, from which 481 HCPs (approximately 58%) have features that are within range of the training set. Since QSPR models are only valid for the trained conditions, 481 HCPs were used for testing. With this approach, the retention time could be predicted with a root mean squared error (RMSE) of 0.085 using HCP structures predicted by AlphaFold2 (Figure 5.2A). To identifying HCPs that might coelute with a target protein, we believe an error of ≤5% to be sufficient considering a DRT between 0 and 1. This takes into account that the DRT describes the retention as a single value, which in reality is a distribution. In practice, when a target protein has a DRT of 0.3, the HCPs with a DRT between 0.2 and 0.4 can be considered as potentially coeluting. For the test set predictions, 207 (~43%) HCPs have an error of ≤5%.

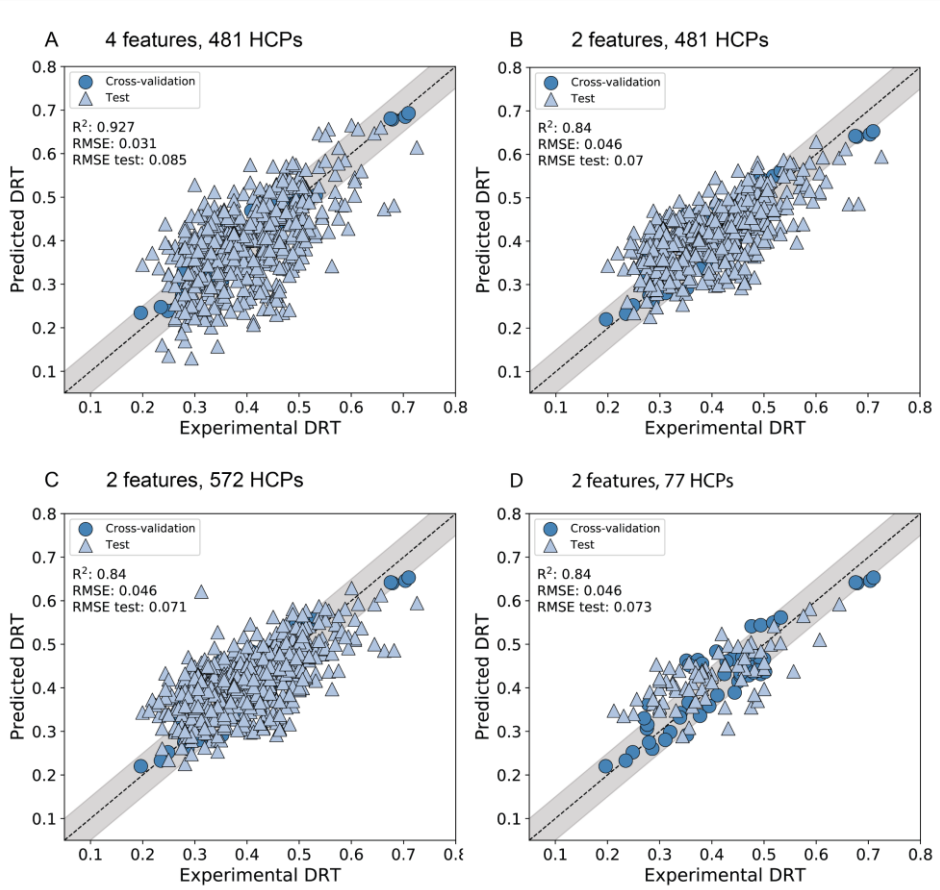


Figure 5.2: *Measured (x-axis) versus predicted (y-axis) dimensionless retention time (DRT) of A) four features and B-D) two features. Models were validated with kfold-cross validation (circles) and tested on HCP DRTs (triangles), The dotted line represents a perfect prediction and the gray area a 5% error. A and B show the HCPs test set filtered for the 4 features model while C and D show the HCPs filtered on the two features. The test set in D is reduced to only include monomeric HCPs.*

To assess the model's ability to generalize for new proteins, the ratio between the RMSE of the test and cross validation should be analyzed. For the current model, the test set RMSE is 3 times the cross validated RMSE. While this might indicate that the training set misses features which are essential to describe HCP retention, the model might also be overfitted. Therefore, a new model was trained using only the two most important features (isoelectric point and the sum of the negative surface electrostatics). For this model, the cross validated R^2 was reduced to 0.840 (Supplemental figure S5.2) while the test set was

predicted with a RMSE of 0.07 (Figure 2B). By eliminating the two least important features, overfitting was significantly reduced (test RMSE is 1.5 times the cross validated RMSE). This also increased the number of accurately predicted HCPs to 246 (~51%) HCPs, which is a 11 percent point improvement. For this test set, the filtering criteria were based on the four feature ranges meaning that the same 481 HCPs were used despite the feature adjustment. Filtering based on the range of two features increases the test set size to 572 HCPs, of which 288 (~50%) can be predicted with an error of ≤5% (Figure 5.2C).

5.3.2 HCP structural representation

It should be noted that DRTs of HCPs are predicted using monomer representations obtained from AlphaFold2. Therefore, the QSPR model does not take into account the complex dynamics of a lysate mixture, in which many interactions may occur. Still, the model is capable of predicting the DRT of 288 HCPs. The structural representations of proteins that are actually monomeric are expected be more representative. Therefore, the model with two features was also tested on 77 of the 572 HCPs that are annotated as monomer in Uniprot. Surprisingly, the subset performed similar to the complete HCP test set with a test RMSE of 0.073 and ~43% predictions with ≤5% error (Figure 5.2D). This suggests that the lack of interaction information about the HCPs does not limit the current model's accuracy. The two features used in the model describe the protein globally and might therefore not capture the required intricacies. A similar phenomenon was observed for the proteins presumed to be homodimers (Supplemental Figure S5.3, Supplemental Table S5.3). For this subset of HCPs, predictions using monomer structures (RMSE: 0.071) performed similar to homodimer representations (RMSE: 0.068).

5.3.3 Model improvement strategies

We have shown that a QSPR model trained with 50 retention times obtained for 13 proteins at various pH values predicted 288 HCPs with an error ≤5% using only two features. While this is a significant part of the available HCP retention times, application of QSPR modeling for in silico process design would require accurate prediction of all detectable HCPs. To identify possibilities to enhance model performance, the test set predictions were divided into overpredicted (181 HCPs), underpredicted (103 HCPs), and accurately predicted (288 HCPs). For these categories, feature value distributions were analyzed to identify potential biases in the model towards features that were not selected for the QSPR model (Supplemental Table S5.4). For a feature that does not contribute to any bias, it can be expected that the distribution over the three sets is similar, which can be observed for the molecular weight (Figure 5.3A). A feature that shows great differences in distribution is the standard deviation of the surface electrostatics (Figure 5.3B), with Kolmogorov-Smirnov (KS) test values of 0.23 and 0.22 for under- and overpredicted HCPs, respectively. For underpredicted HCPs, a generally higher standard deviation is observed compared to the accurately predicted HCPs, while for overpredicted HCPs this feature tends to be lower. This indicates that the model is lacking information on deviations in surface electrostatics. For the training set, the feature range (0.6 to 1.2) is much smaller compared to the range in the test set (0.4 to 1.6) (Supplemental Figure S5.4). Therefore, expanding the training set with commercial proteins that have a wider range of this feature could improve model performance.

For the features that were selected for the model, the pI showed a notable difference in the distributions (Figure 5.3C). Especially for pI >4.5 the feature distribution starts to differ, which indicates that there is a bias for proteins in this pI range. It is therefore not only important

to extend the training set based on the surface electrostatics deviation, but also selecting proteins with a pI >4.5.

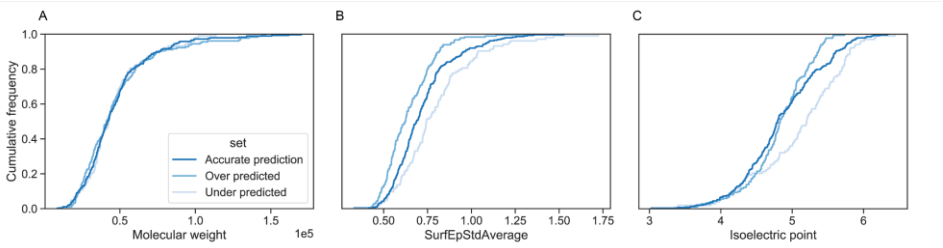


Figure 5.3: *Cumulative distribution plots of the 572 HCPs for A) molecular weight, B) standard deviation of the surface electrostatics, and C) the isoelectric point. The accurate, over and under predicted HCPs are represented by blue, orange and green respectively.*

While extending the training set is essential to improve model quality and robustness, design of novel features is considered equally important. Plotting the surface electrostatics of the three monomeric HCPs with the lowest and highest error reveals positively charged surface areas for the under predicted HCPs (Figure 5.4). Such positive patches are not found on the surface of the three accurately predicted HCPs. The presence of these patches contribute to an increase in the surface electrostatic potential standard distribution feature, as can be observed in Figure 4. Additionally, favorable binding orientations might be more prevalent in the underpredicted HCPs, and these phenomena cannot be captured by the global features used in this study.[38,39] Therefore, designing specific local features representing binding orientations would be essential to improve model performance. For chromatography specifically, local surface features have been designed as either defining patches or projecting properties on a plane.[15,17,40] However, the contribution of preferred binding orientations on adsorption differs between proteins and pH.[37,41,42] This means each protein requires an individual assessment to identify possible binding orientations. This can be done with state-of-the-art molecular dynamics simulations coupled to advanced sampling methods.[39]

Unfortunately, these methods are too computationally expensive to perform on the scale of a host cell proteome. As such, future research should focus on identifying computationally efficient methods to score surface patches based on interaction likelihood. This may also include combining information from patches distant from each other, as ligands with flexible linkers (e.g., XL resin used in this study) probably reach multiple binding sides of the protein.[37]

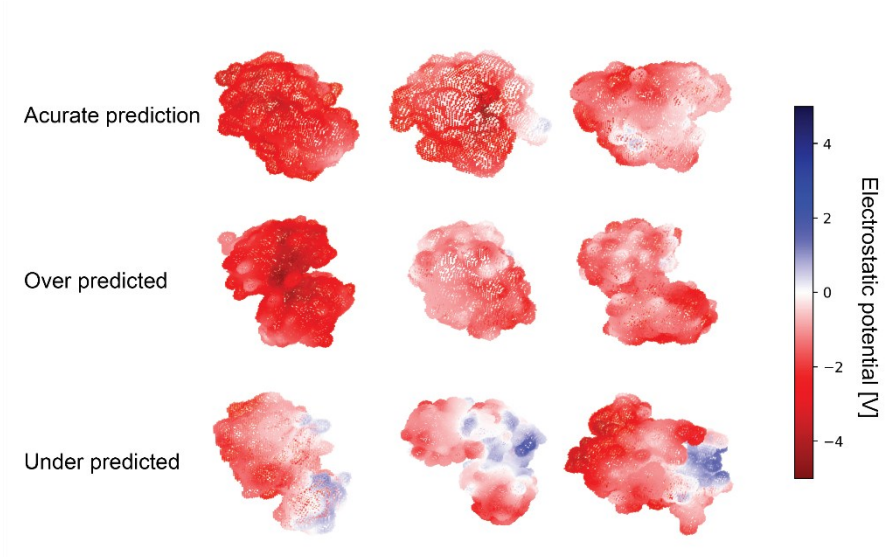


Figure 5.4: *Surface electrostatics at pH 7 of monomer HCPS that are predicted most accurate (top), greatest over prediction (middle) and greatest under prediction (bottom). The blue and red color indicate positive and negative electrostatic potential (in volt).*

Finally, the choice of regression method could also influence the accuracy. Even though the validation on the training data was satisfactory with a cross validated R^2 of 0.84, assumptions associated with a MLR model might limit the accuracy.[43] Especially the assumption that protein retention has a linear dependency on the features. Alternative non-linear regression methods might be a solution to capture non-linear dependencies between protein properties and retention behavior. In recent literature, algorithms such as random forest regression, support vector regression, or Gaussian process regression, have been applied for accurate prediction (R^2 > 0.85) of

different attributes corresponding to chromatographic behavior.[25,27–30,32] Unfortunately, increasing model complexity comes with a risk of overfitting, especially when using small training datasets.[44]

5.4 Conclusion and outlook

In this work, we showcased a workflow to predict retention behavior of 572 *E. coli* HCPs for a Q Sepharose XL column using experimental data obtained for 13 commercial proteins under similar experimental conditions. The described QSPR model with two molecular features (isoelectric point and standard deviation of the surface electrostatics) can predict a total of 288 (~50% of the total test set) HCPs with an error of ≤5% DTR. Interestingly, predictions of the monomer HCP subset did not yield greater accuracy than the complete dataset, which includes proteins that may form multimers. This suggests that the model handles 3D structural inaccuracies regarding multimerization well.

We identified significant differences for the features representing electrostatic deviations on the surface by comparing the feature value distributions for HCPs with an error of ≤5% and >5%. Additionally, it was observed that for proteins with a pI higher than 4.5, HCP retention time is more likely to be underpredicted. Therefore, it is suggested to extend the current training set with proteins that have a pI >4.5 and that contribute to a wider range of surface electrostatic deviations. Additionally, novel features representing preferred binding orientations are required to better describe charge distributions and further increase model accuracy. Despite these proposed improvements, this work provides insight into the use of a small dataset for the prediction of HCP retention behavior, thereby accelerating chromatography resin selection for new products.

5.5 References

1. Walsh, G., & Walsh, E. (2022). Biopharmaceutical benchmarks 2022. *Nature Biotechnology*, *40*(12), 1722–1760. https://doi.org/10.1038/s41587-022-01582-x

2. Reiter, K., Suzuki, M., Olano, L. R., & Narum, D. L. (2019). Host cell protein quantification of an optimized purification method by mass spectrometry. *Journal of Pharmaceutical and Biomedical Analysis*, *174*, 650–654. https://doi.org/10.1016/j.jpba.2019.06.038

3. Keulen, D., Geldhof, G., Bussy, O. Le, Pabst, M., & Ottens, M. (2022). Recent advances to accelerate purification process development: A review with a focus on vaccines. *Journal of Chromatography A*, *1676*, 463195. https://doi.org/10.1016/j.chroma.2022.463195

4. Hanke, A. T. (2016). Technologies to accelerate protein purification process development. *TU Delft University*, 289. https://doi.org/10.4233/uuid:0dc8a46c-1963-4f53-a3b3-6d1dc67202c7

5. Baumann, P., & Hubbuch, J. (2017). Downstream process development strategies for effective bioprocesses: Trends, progress, and combinatorial approaches. *Engineering in Life Sciences*, *17*(11), 1142–1158. https://doi.org/10.1002/elsc.201600033

6. Bracewell, D. G., Francis, R., & Smales, C. M. (2015). The future of host cell protein (HCP) identification during process development and manufacturing linked to a risk-based management for their control. *Biotechnology and Bioengineering*, *112*(9), 1727–1737. https://doi.org/10.1002/bit.25628

7. Wang, X., Hunter, A. K., & Mozier, N. M. (2009). Host cell proteins in biologics development: Identification, quantitation and risk assessment. *Biotechnology and Bioengineering*, *103*(3), 446–458. https://doi.org/10.1002/bit.22304

8. Tscheliessnig, A. L., Konrath, J., Bates, R., & Jungbauer, A. (2013). Host cell protein analysis in therapeutic protein bioprocessing - methods and applications. *Biotechnology Journal*, *8*(6), 655–670. https://doi.org/10.1002/biot.201200018

9. Vanderlaan, M., Zhu-Shimoni, J., Lin, S., Gunawan, F., Waerner, T., & Van Cott, K. E. (2018). Experience with host cell protein impurities in biopharmaceuticals. *Biotechnology Progress*, *34*(4), 828–837. https://doi.org/10.1002/btpr.2640

10. Jones, M., Palackal, N., Wang, F., Gaza-Bulseco, G., Hurkmans, K., Zhao, Y., Chitikila, C., Clavier, S., Liu, S., Menesale, E., Schonenbach, N. S., Sharma, S., Valax, P., Waerner, T., Zhang, L., & Connolly, T. (2021). "High-risk" host cell proteins (HCPs): A multi-company collaborative view. *Biotechnology and Bioengineering*, *118*(8), 2870–2885. https://doi.org/10.1002/bit.27808

11. Shukla, A. A., Hubbard, B., Tressel, T., Guhan, S., & Low, D. (2007). Downstream processing of monoclonal antibodies-Application of platform approaches. *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences*, *848*(1), 28–39. https://doi.org/10.1016/j.jchromb.2006.09.026

12. Keulen, D., Apostolidi, M., Geldhof, G., Le Bussy, O., Pabst, M., & Ottens, M. (2024). Comparing in silico flowsheet optimization strategies in biopharmaceutical downstream processes. *Biotechnology Progress*, *August*, 1–16. https://doi.org/10.1002/btpr.3514

13. Disela, R., Le Bussy, O., Geldhof, G., Pabst, M., & Ottens, M. (2023). Characterisation of the E. coli HMS174 and BLR host cell proteome to guide purification process development. *Biotechnology Journal*, *18*(9), 2300068. https://doi.org/10.1002/biot.202300068

14. Emonts, J., & Buyel, J. F. (2023). An overview of descriptors to capture protein properties – Tools and perspectives in the context of QSAR modeling. *Computational and Structural Biotechnology Journal*, *21*, 3234–3247. https://doi.org/10.1016/j.csbj.2023.05.022

15. Hanke, A. T., Klijn, M. E., Verhaert, P. D. E. M., van der Wielen, L. A. M., Ottens, M., Eppink, M. H. M., & van de Sandt, E. J. A. X. (2016). Prediction of protein retention times in hydrophobic interaction chromatography by robust statistical characterization of their atomic-level surface properties. *Biotechnology Progress*, *32*(2), 372–381. https://doi.org/10.1002/btpr.2219

16. Kittelmann, J., Lang, K. M. H., Ottens, M., & Hubbuch, J. (2017). An orientation sensitive approach in biomolecule interaction quantitative structure–activity relationship modeling and its application in ion-exchange chromatography. *Journal of Chromatography A*, *1482*, 48–56. https://doi.org/10.1016/j.chroma.2016.12.065

17. Kittelmann, J., Lang, K. M. H., Ottens, M., & Hubbuch, J. (2017). Orientation of monoclonal antibodies in ion-exchange chromatography: A predictive quantitative structure–activity relationship modeling approach. *Journal of Chromatography A*, *1510*, 33–39. https://doi.org/10.1016/j.chroma.2017.06.047

18. Neijenhuis, T., Le Bussy, O., Geldhof, G., Klijn, M. E., & Ottens, M. (2024). Predicting protein retention in ion-exchange chromatography using an open source QSPR workflow. *Biotechnology Journal*, *19*(3), e2300708. https://doi.org/10.1002/biot.202300708

19. Keulen, D., Neijenhuis, T., Lazopoulou, A., Disela, R., Geldhof, G., Le Bussy, O., Klijn, M. E., & Ottens, M. (2024). From protein structure to an optimized chromatographic capture step using multiscale modeling. *Biotechnology Progress*, *June*, 1–26. https://doi.org/10.1002/btpr.3505

20. Disela, R., Neijenhuis, T., Le Bussy, O., Geldhof, G., Klijn, M., Pabst, M., & Ottens, M. (2024). Experimental characterization and prediction of Escherichia coli host cell proteome retention during preparative chromatography. *Biotechnology and Bioengineering*, *May*, 3848–3859. https://doi.org/10.1002/bit.28840

21. Mazza, C. B., Sukumar, N., Breneman, C. M., & Cramer, S. M. (2001). Prediction of protein retention in ion-exchange systems using molecular descriptors obtained from crystal structure. *Analytical Chemistry*, *73*(22), 5457–5461. https://doi.org/10.1021/ac010797s

22. Malmquist, G., Nilsson, U. H., Norrman, M., Skarp, U., Strömgren, M., & Carredano, E. (2006). Electrostatic calculations and quantitative protein retention models for ion exchange chromatography. *Journal of Chromatography A*, *1115*(1–2), 164–186. https://doi.org/10.1016/j.chroma.2006.02.097

23. Yang, T., Sundling, M. C., Freed, A. S., Breneman, C. M., & Cramer, S. M. (2007). Prediction of pH-dependent chromatographic behavior in ion-exchange systems. *Analytical Chemistry*, *79*(23), 8927–8939. https://doi.org/10.1021/ac071101j

24. Chen, J., Yang, T., & Cramer, S. M. (2008). Prediction of protein retention times in gradient hydrophobic interaction chromatographic systems. *Journal of Chromatography A*, *1177*(2), 207–214. https://doi.org/10.1016/j.chroma.2007.11.003

25. Hou, Y., & Cramer, S. M. (2011). Evaluation of selectivity in multimodal anion exchange systems: A priori prediction of protein retention and examination of mobile phase modifier effects. *Journal of Chromatography A*, *1218*(43), 7813–7820. https://doi.org/10.1016/j.chroma.2011.08.080

26. Song, M., Breneman, C. M., Bi, J., Sukumar, N., Bennett, K. P., Cramer, S., & Tugcu, N. (2002). Prediction of Protein Retention Times in Anion-Exchange Chromatography Systems Using Support Vector Regression. *Journal of Chemical*

Information and Computer Sciences, *42*(6), 1347–1357. https://doi.org/10.1021/ci025580t

27. Cai, Q. Y., Qiao, L. Z., Yao, S. J., & Lin, D. Q. (2024). Machine learning assisted QSAR analysis to predict protein adsorption capacities on mixed-mode resins. *Separation and Purification Technology*, *340*(December 2023), 126762. https://doi.org/10.1016/j.seppur.2024.126762

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

29. Hess, R., Faessler, J., Yun, D., Mama, A., Saleh, D., Grosch, J. H., Wang, G., Schwab, T., & Hubbuch, J. (2024). Predicting multimodal chromatography of therapeutic antibodies using multiscale modeling. *Journal of Chromatography A*, *1718*(February), 464706. https://doi.org/10.1016/j.chroma.2024.464706

30. Saleh, D., Hess, R., Ahlers-Hesse, M., Rischawy, F., Wang, G., Grosch, J. H., Schwab, T., Kluters, S., Studts, J., & Hubbuch, J. (2023). A multiscale modeling method for therapeutic antibodies in ion exchange chromatography. *Biotechnology and Bioengineering*, *120*(1), 125–138. https://doi.org/10.1002/bit.28258

31. Hess, R., Faessler, J., Yun, D., Saleh, D., Grosch, J. H., Schwab, T., & Hubbuch, J. (2023). Antibody sequence-based prediction of pH gradient elution in multimodal chromatography. *Journal of Chromatography A*, *1711*(October), 464437. https://doi.org/10.1016/j.chroma.2023.464437

32. Hartmann, M., Rauscher, M., Robinson, J., Welsh, J., & Roush, D. (2025). Integration of QSAR models with high throughput screening to accelerate the development of polishing chromatography unit operations. *Journal of Chromatography A*, *1747*(December 2024), 465818. https://doi.org/10.1016/j.chroma.2025.465818

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

34. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, *28*(1), 235–242. https://doi.org/10.1093/nar/28.1.235

35. David, A., Islam, S., Tankhilevich, E., & Sternberg, M. J. E. (2022). The AlphaFold Database of Protein Structures: A Biologist's Guide. *Journal of Molecular Biology*, *434*(2), 167336. https://doi.org/10.1016/j.jmb.2021.167336

36. Berger, V. W., & Zhou, Y. (2014). Kolmogorov–Smirnov Test: Overview. In *Wiley StatsRef: Statistics Reference Online* (pp. 1–5). Wiley. https://doi.org/10.1002/9781118445112.stat06558

37. Dismer, F., Petzold, M., & Hubbuch, J. (2008). Effects of ionic strength and mobile phase pH on the binding orientation of lysozyme on different ion-exchange adsorbents. *Journal of Chromatography A*, *1194*(1), 11–21. https://doi.org/10.1016/j.chroma.2007.12.085

38. Rabe, M., Verdes, D., & Seeger, S. (2011). Understanding protein adsorption phenomena at solid surfaces. *Advances in Colloid and Interface Science*, *162*(1–2), 87–106. https://doi.org/10.1016/j.cis.2010.12.007

39. Quan, X., Liu, J., & Zhou, J. (2019). Multiscale modeling and simulations of protein adsorption: progresses and perspectives. *Current Opinion in Colloid and Interface Science*, *41*, 74–85. https://doi.org/10.1016/j.cocis.2018.12.004

40. Sankar, K., Trainor, K., Blazer, L. L., Adams, J. J., Sidhu, S. S., Day, T., Meiering, E., & Maier, J. K. X. (2022). A Descriptor Set for Quantitative Structure-property Relationship Prediction in Biologics. *Molecular Informatics*, *41*(9), 2100240. https://doi.org/10.1002/minf.202100240

41. Aguilar, M.-I., Clayton, D. J., Holt, P., Kronina, V., Boysen, R. I., Purcell, A. W., & Hearn, M. T. W. (1998). RP-HPLC Binding Domains of Proteins. *Analytical Chemistry*, *70*(23), 5010–5018. https://doi.org/10.1021/ac980473c

42. Yao, Y., & Lenhoff, A. M. (2004). Electrostatic Contributions to Protein Retention in Ion-Exchange Chromatography. 1. Cytochrome c Variants. *Analytical Chemistry*, *76*(22), 6743–6752. https://doi.org/10.1021/ac049327z

43. Osborne, J. W., & Waters, E. (2002). Four Assumptions of Multiple Regression That Researchers Should Always Test. - Practical Assessment, Research & Evaluation. *Practical Assessment, Research and Evaluation*, *8*(2), 1–5.

44. Ying, X. (2019). An Overview of Overfitting and its Solutions. *Journal of Physics: Conference Series*, *1168*(2), 022022. https://doi.org/10.1088/1742-6596/1168/2/022022

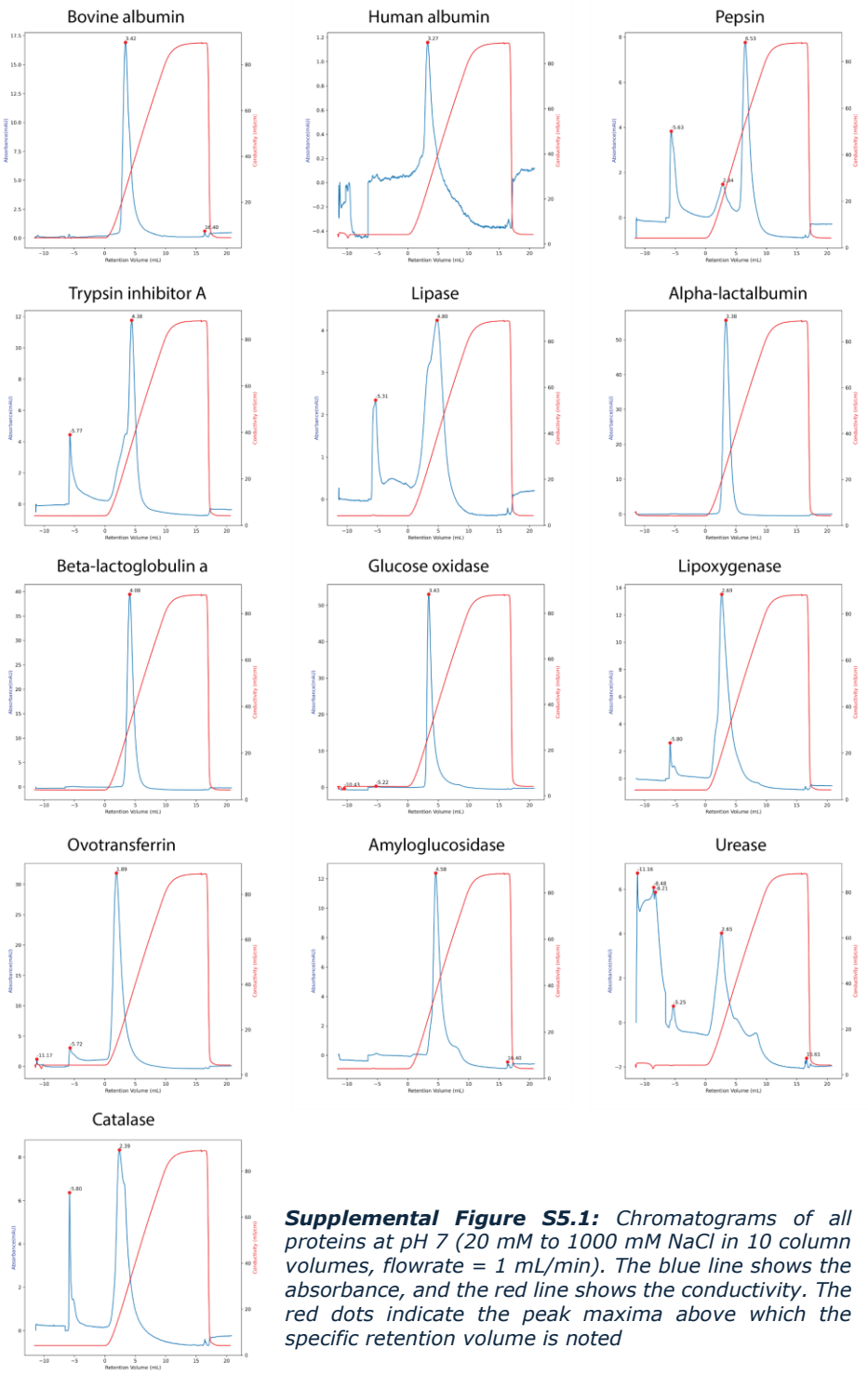
5.6 Supporting Information

Supplemental Table S5.1: *System properties*

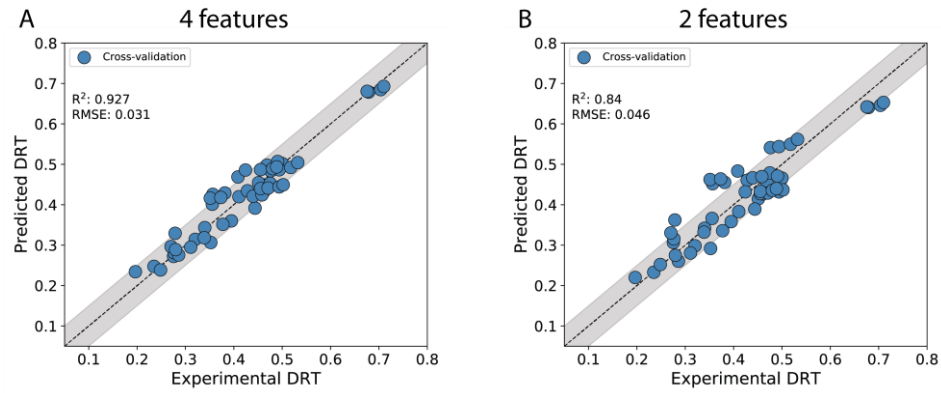
System	1	2
Dead volume [mL]	0.246	0.239
Dwell volume [mL]	1.109	1.109
Void volume [mL]	0.253	0.249
Column length [mm]	7	
Column diameter [mm]	25	

Supplemental Table S5.2: *Model parameters for the QSPR model with four features.*

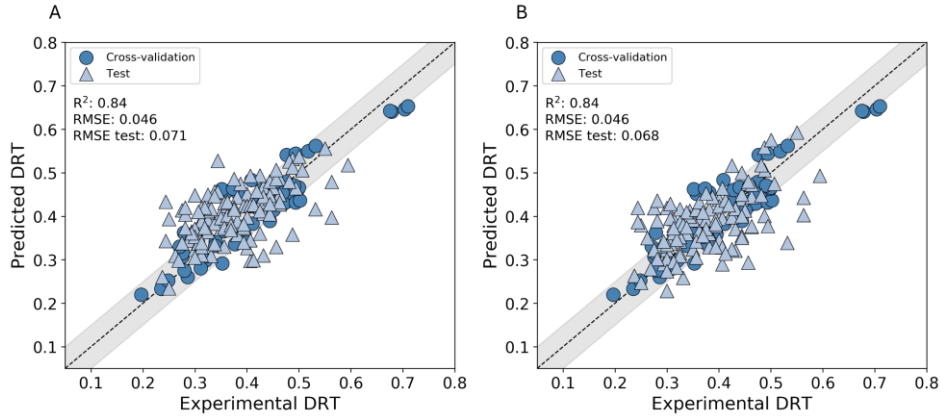
	Coefficient	Permutation R^2
Isoelectric point	-0.539	-0.27
SurfEpNegSumAverage	-0.231	0.76
PROSurfFrac	0.089	0.88
SurfNegMhpMean	-0.123	0.87
intercept	0.813	



Supplemental Figure S5.1: *Chromatograms of all proteins at pH 7 (20 mM to 1000 mM NaCl in 10 column volumes, flowrate = 1 mL/min). The blue line shows the absorbance, and the red line shows the conductivity. The red dots indicate the peak maxima above which the specific retention volume is noted*



Supplemental figure S5.2: *QSPR model cross validation results of the training set. Measured (x-axis) versus predicted (y-axis) dimensionless retention time (DRT) of a QSPR model trained on four (A) or two (B) features. The model was validated with kfold-cross validation.*



Supplemental figure S5.3: *Homodimer HCP predictions using A) monomer structure and B) predicted homodimer structure. Measured (x-axis) versus predicted (y-axis) dimensionless retention time (DRT) of a QSPR model trained on two features. The model was validated with kfold-cross validation (circles) and tested on HCP DRTs (triangles).*

Supplemental Table S5.3: *model features for homodimer HCPs calculated from a monomer or homodimer structure.*

	Monomer structure		Homodimer structure	
ID	pI	SurfEpNegSumAverage	pI	SurfEpNegSumAverage
ARH97593	5.03	-19378.1	5.363	-54988.1
ARH96837	5.215	-19917.7	5.738	-52527.1
ARH96219	5.251	-29950.9	5.403	-87292.2
ARH96022	5.024	-20139.4	5.279	-55287.8
ARH98242	6.133	-14919.8	6.218	-43961.7
ARH98543	4.504	-11312.7	5.057	-31176.7
ARH99203	4.906	-22561.4	5.569	-49525.7
ARH97151	3.859	-12130.3	4.882	-25314.9
ARH98083	5.312	-12210.7	5.926	-27508.5
ARH97695	4.756	-19305.6	4.815	-45812.2
ARH96704	4.769	-15453.2	5.419	-34216.3
ARH99670	4.345	-30868.3	4.875	-88303.6
ARH98250	5.324	-18458.5	5.405	-49684.4
ARH97968	5.05	-21129.7	5.818	-54097.8
ARH97818	4.476	-17344.4	5.325	-41168.1
ARH97190	5.117	-21365.5	5.419	-47288.2
ARH99778	5.228	-10979.7	5.596	-24431.9
ARH97386	4.965	-12408.7	5.039	-43649.2
ARH95908	5.711	-11634.3	6.177	-21775.1
ARH98111	5.327	-19589.1	5.174	-74270.2
ARH97701	4.547	-20670	5.268	-43418.9
ARH95833	5.534	-13658.4	5.851	-39497.5
ARH96141	5.817	-14686.5	5.943	-29570
ARH97716	5.702	-11011.5	5.445	-33338.1
ARH97717	4.463	-31109.7	5.378	-75546
ARH98656	4.572	-24523.5	4.8	-63261.4
ARH98155	5.576	-17202	5.545	-46233
ARH97457	4.779	-17535.9	5.086	-46810.2
ARH97452	5.258	-47150	5.749	-132326
ARH97703	4.315	-39709.6	4.854	-87131.2
ARH95854	4.969	-37135.4	5.363	-117947
ARH96954	3.943	-95116.2	4.418	-297729
ARH99014	4.281	-16082.2	4.651	-42222.3
ARH99185	4.951	-28103.8	5.415	-71048.4
ARH98432	4.825	-18669.4	5.571	-41188.3
ARH95944	4.583	-16556.3	4.928	-44904.7
ARH98173	5.034	-12588.2	5.604	-29113.7
ARH99034	3.884	-27690.8	4.326	-85279.6
ARH95789	4.832	-15249.3	4.915	-46949.2
ARH95876	4.031	-53997.7	4.439	-140268
ARH96669	4.739	-32529.6	5.188	-93616.6
ARH95939	4.079	-47555.9	4.541	-112189
ARH96404	4.866	-46565.5	5.463	-124958
ARH98978	5.723	-17631.3	5.767	-53215.5
ARH97914	4.571	-22863.8	5.051	-82901.9
ARH97435	4.366	-108507	4.747	-307185
ARH99054	4.754	-16038	5.219	-37074.7
ARH99823	4.807	-29248.7	4.904	-87391.2
ARH99236	5.592	-27242.8	6.029	-60161.9
ARH97789	4.898	-43657.1	5.465	-134692
ARH98367	4.602	-13063.4	5.38	-29087.4
ARH97611	5.065	-35322.2	5.329	-93082.4
ARH96870	5.429	-18006.2	5.659	-41534.4
ARH99681	5.35	-31069.8	5.825	-71970.3

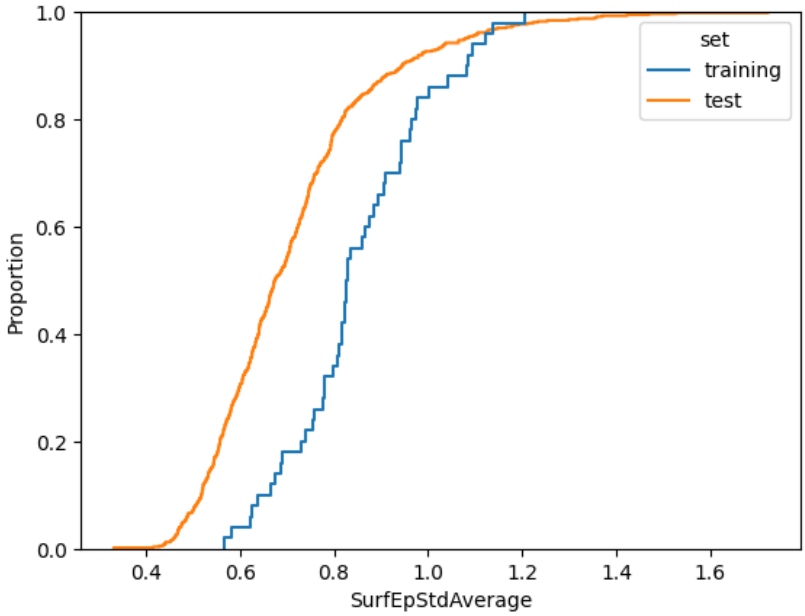
ARH99115	4.395	-14894.1	4.921	-35612.9
ARH96780	5.432	-15651	5.236	-59268.5
ARH96098	4.876	-21853.3	5.82	-40398.2
ARH97261	5.058	-19510.8	5.305	-59681.8
ARH98634	5.426	-16596.5	5.675	-52380.6
ARI00054	4.568	-23742.5	5.024	-68950.9
ARH95959	5.301	-13205.1	5.6	-46407
ARH98295	4.479	-15914.7	5.032	-34093.8
ARH97371	4.815	-17410.2	5.465	-44325.3
ARH97928	5.583	-11288.2	6.42	-19109.5
ARH96965	4.394	-18618.2	4.954	-55026.2
ARH98514	4.968	-23414.1	5.447	-67909.5
ARH97394	6.339	-11407.5	6.286	-26101.6
ARH99841	4.774	-24142.1	5.151	-55434.2
ARH99712	4.802	-21576.2	5.16	-49661.3
ARH99585	5.18	-12451.1	5.567	-30734.5
ARH99442	5.05	-36027.7	5.067	-95771.2
ARH97497	5.479	-15038.5	5.14	-51817.7
ARH99358	4.674	-16572.2	4.935	-53669.9
ARH98673	4.61	-28557.7	5.12	-79624.2
ARH97996	5.213	-25358.9	5.203	-61874.6
ARH99658	5.332	-11282.8	5.849	-29860.3
ARH96706	4.826	-39849.9	4.997	-107115
ARH98524	4.833	-50854.5	5.309	-140219
ARH99426	5.737	-22505.1	5.486	-69686.4
ARH99592	5.244	-12799.5	5.158	-38905.6
ARH98103	5.392	-43513.7	5.836	-116498
ARH97557	5.271	-11849.7	5.643	-27593
ARH98612	4.552	-87285	4.967	-236979
ARH98399	5.463	-12813.3	5.677	-33845.8
ARH99828	4.663	-12813.9	4.779	-41634.1
ARH99628	5.548	-15148.5	5.943	-41601.1
ARH99392	5.017	-57800.4	5.478	-146690
ARH98148	4.706	-56375.4	5.073	-156174
ARH96683	4.814	-16366.8	5.224	-41373.5
ARH97345	4.352	-47627.7	4.821	-145974
ARH99626	4.803	-11576.2	5.253	-29441.6
ARH96265	4.79	-16881.2	5.524	-44887.5
ARH95810	4.604	-24506.5	5.009	-68341.8
ARH99655	5.555	-37215.3	5.637	-131533
ARH97615	5.115	-16519	5.237	-51034.6
ARH96215	4.428	-84757.1	4.907	-257712
ARH98664	4.316	-14932.3	4.71	-48865.2
ARH98193	5.18	-17889.8	6.016	-31891.2
ARH96414	5.461	-14774	5.652	-40725.6
ARH99407	4.774	-29559.8	5.356	-69585.5
ARH95981	4.358	-65018.7	4.941	-153146
ARH99258	5.171	-14991.2	5.561	-46578.1
ARH99121	5.175	-28339.8	5.573	-77386.4
ARH96866	4.491	-21356.9	4.91	-59684.8
ARH96902	5.277	-29660.4	5.235	-107112
ARH99877	4.239	-18056.2	4.656	-53798.7
ARH99441	5.208	-21168.5	5.899	-45069.7
ARH99624	5.532	-25967	5.473	-93404.3
ARH96155	5.846	-24498.1	6.158	-60816
ARH98443	5.808	-16133.1	5.807	-48766.6
ARH96911	4.91	-27304.9	4.968	-81314.4
ARH96956	4.137	-36967.3	4.63	-95611.4
ARH98479	4.815	-50718.9	4.93	-168448

ARH99443	4.943	-23290.8	5.567	-66256.8
ARH96347	4.981	-31162.6	5.338	-90789.8
ARH97343	5.761	-16723.6	6.094	-33816.2
ARH96664	4.762	-27572.8	5.161	-69301.9
ARH98064	4.114	-98919.9	4.581	-302197
ARH97694	4.815	-24038.7	4.872	-72766.9
ARH98256	4.094	-26668.3	4.484	-69073.5

Supplemental Table S5.4: *Kolmogorov-Smirnov (KS) test results for under and over predicted HCPs against accurately predicted HCPs.*

	Under predicted		Over predicted	
feature	KS value	p value	KS value	p value
Molecular weight	0.061	0.921	0.072	0.582
Isoelectric point	0.262	0	0.155	0.008
Dipole	0.212	0.002	0.156	0.008
Formal charge	0.201	0.004	0.188	0.001
Average charge	0.205	0.003	0.185	0.001
Area	0.086	0.591	0.123	0.062
ALASurfFrac	0.09	0.535	0.19	0.001
ARGSurfFrac	0.121	0.199	0.151	0.011
ASNSurfFrac	0.112	0.269	0.137	0.027
ASPSurfFrac	0.125	0.169	0.043	0.977
CYSSurfFrac	0.054	0.97	0.15	0.012
GLNSurfFrac	0.073	0.784	0.046	0.963
GLUSurfFrac	0.201	0.004	0.181	0.001
GLYSurfFrac	0.08	0.685	0.157	0.007
HISSurfFrac	0.149	0.061	0.164	0.004
ILESurfFrac	0.12	0.204	0.067	0.668
LEUSurfFrac	0.123	0.182	0.046	0.964
LYSSurfFrac	0.109	0.304	0.234	0
METSurfFrac	0.06	0.932	0.228	0
PHESurfFrac	0.162	0.033	0.103	0.174
PROSurfFrac	0.071	0.802	0.111	0.118
SERSurfFrac	0.107	0.319	0.051	0.912
THRSurfFrac	0.066	0.865	0.068	0.654
TRPSurfFrac	0.11	0.294	0.101	0.188
TYRSurfFrac	0.141	0.087	0.133	0.035
VALSurfFrac	0.126	0.163	0.068	0.651
NSurfPoints	0.093	0.493	0.118	0.081
Shape max	0.052	0.978	0.166	0.004
Shape min	0.06	0.927	0.058	0.824
SurfEpMaxFormal	0.275	0	0.112	0.114
SurfEpMinFormal	0.108	0.311	0.208	0
SurfEpMeanFormal	0.241	0	0.194	0
SurfEpTrimeanFormal	0.251	0	0.194	0
SurfEpMedianFormal	0.238	0	0.193	0
SurfEpSumFormal	0.137	0.103	0.149	0.013
SurfEpStdFormal	0.183	0.011	0.218	0
NSurfPosEpFormal	0.278	0	0.133	0.035
SurfEpPosMeanFormal	0.284	0	0.11	0.124
SurfEpPosTrimeanFormal	0.273	0	0.124	0.06
SurfEpPosMedianFormal	0.268	0	0.103	0.171
SurfEpPosSumFormal	0.28	0	0.12	0.074
SurfEpPosStdFormal	0.269	0	0.104	0.163
NSurfNegEpFormal	0.061	0.923	0.116	0.092
SurfEpNegMeanFormal	0.208	0.002	0.196	0
SurfEpNegTrimeanFormal	0.222	0.001	0.192	0
SurfEpNegMedianFormal	0.226	0.001	0.193	0
SurfEpNegSumFormal	0.138	0.1	0.156	0.008
SurfEpNegStdFormal	0.115	0.244	0.242	0
SurfMhpMax	0.126	0.162	0.117	0.088
SurfMhpMin	0.076	0.742	0.055	0.864
SurfMhpMean	0.189	0.007	0.116	0.093
SurfMhpTrimean	0.192	0.006	0.1	0.196
SurfMhpMedian	0.187	0.008	0.099	0.208
SurfMhpSum	0.064	0.886	0.081	0.435

SurfMhpStd	0.145	0.074	0.16	0.006
NSurfPosMhp	0.13	0.138	0.145	0.016
SurfPosMhpMean	0.156	0.043	0.139	0.025
SurfPosMhpTrimean	0.137	0.103	0.167	0.003
SurfPosMhpMedian	0.156	0.044	0.179	0.001
SurfPosMhpSum	0.119	0.213	0.189	0.001
SurfPosMhpStd	0.099	0.411	0.112	0.112
NSurfNegMhp	0.057	0.948	0.104	0.163
SurfNegMhpMean	0.153	0.051	0.139	0.024
SurfNegMhpTrimean	0.185	0.01	0.112	0.113
SurfNegMhpMedian	0.156	0.044	0.101	0.194
SurfNegMhpSum	0.045	0.995	0.084	0.39
SurfNegMhpStd	0.117	0.225	0.047	0.958
SurfEpMaxAverage	0.26	0	0.104	0.163
SurfEpMinAverage	0.098	0.431	0.192	0
SurfEpMeanAverage	0.232	0	0.179	0.001
SurfEpTrimeanAverage	0.216	0.001	0.185	0.001
SurfEpMedianAverage	0.226	0.001	0.182	0.001
SurfEpSumAverage	0.138	0.099	0.128	0.046
SurfEpStdAverage	0.208	0.002	0.225	0
NSurfPosEpAverage	0.257	0	0.123	0.063
SurfEpPosMeanAverage	0.274	0	0.107	0.14
SurfEpPosTrimeanAverage	0.261	0	0.102	0.182
SurfEpPosMedianAverage	0.266	0	0.103	0.172
SurfEpPosSumAverage	0.276	0	0.129	0.044
SurfEpPosStdAverage	0.266	0	0.105	0.156
NSurfNegEpAverage	0.073	0.778	0.107	0.141
SurfEpNegMeanAverage	0.188	0.008	0.189	0.001
SurfEpNegTrimeanAverage	0.194	0.006	0.189	0.001
SurfEpNegMedianAverage	0.184	0.01	0.191	0
SurfEpNegSumAverage	0.118	0.222	0.127	0.05
SurfEpNegStdAverage	0.112	0.269	0.219	0
ShellEpMaxFormal	0.201	0.004	0.065	0.703
ShellEpminFormal	0.091	0.515	0.213	0
ShellEpMeanFormal	0.206	0.003	0.18	0.001
ShellEpTrimeanFormal	0.224	0.001	0.179	0.001
ShellEpMedianFormal	0.212	0.002	0.184	0.001
ShellEpSumFormal	0.212	0.002	0.18	0.001
ShellEpStdFormal	0.121	0.196	0.183	0.001
NShellPosEpFormal	0.251	0	0.06	0.79
ShellEpPosMeanFormal	0.146	0.069	0.093	0.271
ShellEpPosTrimeanFormal	0.111	0.282	0.071	0.592
ShellEpPosMedianFormal	0.134	0.115	0.073	0.556
ShellEpPosSumFormal	0.233	0	0.062	0.764
ShellEpPosStdFormal	0.203	0.003	0.084	0.384
NShellNegEpFormal	0.242	0	0.083	0.405
ShellEpNegMeanFormal	0.178	0.014	0.202	0
ShellEpNegTrimeanFormal	0.176	0.015	0.19	0.001
ShellEpNegMedianFormal	0.183	0.011	0.182	0.001
ShellEpNegSumFormal	0.191	0.007	0.177	0.002
ShellEpNegStdFormal	0.091	0.514	0.176	0.002



Supplemental figure S5.4: *Cumulative distribution plot of the SurfEpStdAverage for the training (Blue) and test (Orange) sets.*

Chapter 6

Comparing isotherm parameter determination methods for hydrophobic interaction chromatography

Submitted for publication:

Neijenhuis, T., Vale, T. Le Bussy, O., Geldhof, G., Klijn, M., Ottens, M..

Abstract

Hydrophobic interaction chromatography (HIC) is a widely used separation method in biopharmaceutical downstream processing. For process development (PD), silico mechanistic modeling can be used to reduce timelines by simulating protein transport and adsorption during chromatography. Accuracy of the parameters used in the model is essential for successful deployment. This work compares three isotherm parameter determination methods for a simplified linear HIC isotherm. Specifically, the Parente and Wetlaufer method, the Yamamoto method, and the inverse method. These methods were tested for two proteins, using the same linear gradient elution (LGE) experiments. Accuracy of the obtained parameters was determined via cross-validation using three LGEs. Finally, the obtained parameters were tested for alternative linear gradients with varying initial and final salt concentrations. While all results were comparable, parameters obtained by the inverse method showed the greatest accuracy. This method does require high quality chromatograms, while the other methods only need retention volumes. Therefore, it is less suitable when signal quality is compromised. The Yamamoto method showed similar robustness as the inverse method outperforming the Parente and Wetlaufer method. Therefore, the Yamamoto method is a good alternative for parameter determination. This comparison offers practical guidance for method selection for isotherm determination, thereby enabling reliable mechanistic modeling of HIC processes.

6.1 Introduction

Hydrophobic interaction chromatography (HIC) is a separation method widely used at different stages biopharmaceutical downstream processing (DSP).[1,2] It is specifically applied as an orthogonal method for ion exchange chromatography (IEX), as HIC separates based on surface hydrophobicity rather than surface charge.[2] Protein affinity is driven by solvophobic effects, which can be enhanced by anti-chaotropic ions or reduced by chaotropic ions. These effects need to be optimized to establish a robust separation process. Therefore, process development involves an elaborate screening of operation conditions.

To accelerate the design of a chromatographic purification step, in silico tools, in combination with high throughput experimentation can be deployed.[3–5] Recently, mechanistic models (MMs) have proven to be valuable by increasing process understanding.[6–13] These models can describe the flow and mass transfer of proteins through a chromatography column. The dynamic adsorption of proteins is captured by adsorption isotherms that describe the equilibrium between the protein concentration in the solid and liquid phase.[14] For HIC, the isotherm developed by Mollerup[15–17] is commonly applied to simulate protein adsorption under varying salt concentrations.[18–20] This isotherm is based on the stochiometric displacement model and uses an activity coefficient to incorporate salt dependency.[17] To apply this isotherm, several parameters require to be determined, which can be done using a set of linear gradient elution (LGE) experiments or batch adsorption experiments. The accuracy of these parameters is essential to ensure successful protein adsorption modeling.

The inverse method (IM) is a common method to estimate isotherm parameters and has been proven to provide accurate simulations for different chromatographic modes.[12,21–25] IM fits the result of the mechanistic model to experimental chromatograms and updates the

isotherm parameters to minimize the difference between model and experiment. Therefore, high quality chromatograms of pure components are required for accurate parameter estimation. Alternatively, isotherm parameters can be estimated from protein retention volumes. The Parente and Wetlaufer (PW) method (non-linear) and the Yamamoto method (linear) are correlations that relate LGE conditions to retention volumes.[26,27] While both methods are developed for IEX, they have been adapted for HIC in recent literature.[20,28] As no iterative simulations are required, using the correlations is more computational efficient which is beneficial when large datasets are analyzed.[29] However, the correlations only allow determination of the linear part of the isotherm, therefore it can only be used under low loading conditions.

In this work we compare the accuracy of isotherms obtained using IM, PW, and Yamamoto using the same LGE experiments. For this we apply the transport dispersive model and the linear part of the isotherm developed by Mullerup to model the adsorption behavior of two proteins under dilute conditions. The model parameters are subsequently validated via cross validation and compared to experimental chromatograms. Quantitative analysis is performed based on differences in peak maxima and peak widths. Finaly, the robustness of the estimated isotherm parameters are tested under alternative salt gradient conditions. Consequently, this work enables informed method selection, enhancing reliability of mechanistic modeling of HIC processes.

6.2 Methods

6.2.1 Materials and Equipment

The retention experiments were performed on an Äkta pure system (Cytiva, Marlborough*, USA*), equipped with a prepacked HiTrap Butyl FF 1 mL column (Cytiva, Marlborough*, USA*) (Appendix A1). All

substances were purchased from Sigma-Aldrich (Saint Louis, USA) and buffers were prepared using ultrapure water filtered with the Milli-Q Advantage A10 (Merck Millipore, Burlington, USA). Buffer solutions were prepared using 50 mM sodium phosphate and a range of ammonium sulfate concentrations (2.0 M, 1.5 M, 1.3 M, 1.1 M, 0.8 M and 0M) to be adjusted to pH 7 using 1 M sodium hydroxide. All buffers were filtered using a 0.2 μm Membrane Disc Filter (Pall corporation, New York, USA) followed by 20 minutes of sonication.

Chymotrypsinogen A and glucoamylase were purchased from Sigma-Aldrich (Saint Louis, USA). For each experiment, proteins were dissolved in the respective high salt buffer to reach a concentration of 2 mg/mL, after which the solutions were filtered using a 0.22 μm Whatman Puradisc FP 30 mm (*Cytiva, Marlborough, USA*).

6.2.2 System and column characterization

To determine relevant system and column properties, a set of pulse experiments with a flowrate of 1 mL/min were performed using a set of nonbinding tracers as described by Schmidt-Traub et al..[30] Dextran DXT180 (Agilent, Santa Clara, USA) and dextran DXT2000k (Toronto Research Chemicals, Toronto, Canada) were used as penetrating and non-penetrating tracers, respectively. The system dwell volume, describing the volume between the mixing chamber and the column inlet was determined as described by Keulen et. al..[11] A complete list of the determined properties can be found in Supplemental Table S1.

6.2.3 Linear gradient elution experiments

A set of LGE experiments were performed with 10, 15, 20, 30 and 40 column volume (CV) gradient lengths with a flowrate of 1 mL/min. After equilibration with the high salt buffer, 200 µL protein solution was injected followed by a 5 CV wash and the start of the salt gradient. Upon reaching the end of the gradient, the column was washed with 10 CV low salt buffer. During the experiments, UV absorbance was

measured at 280 nm and the system was operated using UNICORN version 7.5 software. For the chymotrypsinogen LGEs from 1.5 M to 0 M ammonium sulfate, the chromatograms were deconvoluted in python using two gaussians that were parameterized by scipy.minimize.

6.2.4 Mechanistic model

Simulation of the adsorption behavior of the proteins during the chromatographic experiments was performed using equilibrium transport dispersive model combined with the linear driving force (Equation 6.1), as described in chapter 3.[31]

$$\frac{\partial C_i}{\partial t} + F\frac{\partial q_i}{\partial t} = -u\frac{\partial C_i}{\partial x} + D_{L,i}\frac{\partial^2 C_i}{\partial x^2} \tag{6.1}$$

$$\frac{\partial q_i}{\partial t} = K_{ov,i}(C_i - C^*_{eq,i}), \tag{6.2}$$

$$K_{ov,i} = \left[\frac{d_p}{6k_{f,i}} + \frac{d_p^2}{60\varepsilon_p D_{p,i}}\right]^{-1} \tag{6.3}$$

Here, the protein concentration in the liquid and solid phase are denoted as C and q, respectively, while C^*_{eq} is the liquid phase concentration at equilibrium. The phase ratio is defined as $F = (1 - \varepsilon_b)/\varepsilon_b$, where ε_b is the bed porosity, u is the interstitial velocity of the mobile phase and D_L is the axial dispersion coefficient. Time and space are represented by t and x, respectively. The overall mass transfer coefficient (K_{ov}) is defined as the summation of the mass transfer resistance in the film and within the pores. Here, d_p is the particle diameter, D_p is the effective pore diffusivity and ε_p is the intraparticle porosity. The film mass transfer coefficient is defined as $k_f = D_f Sh/d_p$ where Sh is the Sherwood number and D_f is the free diffusivity which is calculated using empirical correlation (equation 6.4) based on the molecular mass (MW). [32]

$$D_f = 260 * 10^{-11}(MW^{-1/3}). \tag{6.4}$$

6.2.5 Hydrophobic interaction isotherm

In this work, the commonly used HIC isotherm developed by Mullerup[15–17] is used. This isotherm is defined as:

$$\frac{q}{c_p} = K_{eq}\left(\frac{\Lambda}{c}\right)^n \left(1 - \frac{q}{q_{max}}\right)^n \exp\left(K_s c_s + K_p c_P\right), \tag{6.5}$$

where Λ is the ligand concentration, n is the stochiometric coefficient, c is the molar concentration in the pores. K_{eq}, K_s and K_p are the equilibrium constant, salt and protein interaction parameters, respectively. Finally, q and q_{max} are the current and maximum concentration in the solid phase.

This isotherm allows for some simplifications, since K_p has been proven to have minor impact, this parameter can be assumed to be zero.[20,33,34] Additionally, assuming that c remains constant[19,28,35] allows for an alternative definition of the equilibrium constant as $K'_{eq} \approx K_{eq}(\Lambda/\mathrm{c})^n$ resulting in the following:

$$\frac{q}{c_p} = K'_{eq}\left(1 - \frac{q}{q_{max}}\right)^n \exp\left(K_s c_s\right). \tag{6.6}$$

Finally, for low loading conditions, it can be assumed that $q \ll q_{max}$, resulting in $(1 - q/q_{max})^n \approx 1$. Applying this to equation 6.6 yields the final form of the linear isotherm used in this study, which is defined as:

$$\frac{q}{c_p} = K'_{eq}\exp\left(K_s c_s\right).. \tag{6.7}$$

6.2.6 Isotherm parameter determination

To apply isotherm equation 6.7, accurate determination of K'_{eq} and K_s are essential. In this work, we will compare three methods which require a set of linear gradient elution (LGE) experiments.

6.2.6.1 Parente and Wetlaufer

The first approach is based on the Parente and Wetlaufer regression formula, originally developed for ion-exchange chromatography.[26] Chen et al.[36] adapted this formula for HIC:

$$V_{R,g} = \frac{V_G}{-\beta \cdot (c_{s,f} - c_{s,i})} ln\,(1 + V_m \frac{-\beta \cdot (c_{s,f} - c_{s,i})}{V_G} e^{\alpha + \beta \cdot c_{s,i}}), \quad (6.8)$$

where $V_{R,g}$ is the corrected retention volume ($V_{R,g} = V_R - V_m - 0.5V_{inj} - V_{dwell}$), V_G is the gradient length and V_m is the void volume. α and β are fitted using the retention volumes at different gradient lengths. The fitted parameters relate to the retention factor by:

$$\ln(k') = \alpha + \beta c_s, \quad (6.9)$$

As

$$k' = \frac{t_R - t_0}{t_0} = F A_i, \quad (6.10)$$

where t_R and t_0 are the time of retention and start of the gradient, respectively. A_i is the initial slope of the isotherm, which is equal to equation 6.7. As such, equation 8 can be rewritten as follows:

$$V_{R,g} = \frac{V_G}{-K_s(c_{s,f} - c_{s,i})} * \ln(1 + V_m F \frac{-K_s(c_{s,f} - c_{s,i})}{V_G} K'_{eq} e^{K_s c_{s,i}}), \quad (6.11)$$

6.2.6.2 Yamamoto

The second method is based on the Yamamoto approach, which is like the previous method originally developed for ion-exchange[10,27,37]. Recently, Hess et al. adapted this method for the regression of HIC isotherm parameters.[37] It relates the normalized gradient slope (GH) to the salt concentration at which the peak maximum is observed ($c_{s,R}$) using a linear formula. The normalized gradient slope is defined as follows:

$$GH = g(1 - \epsilon_b)V_{col}, \quad (6.12)$$

where g is the gradient slope, defined as:

$$g = \frac{c_{s,f} - c_{s,i}}{V_G} \quad (6.13)$$

and V_{col} is the total volume of the column. When size exclusion effects are neglected, the normalized gradient length can be related to $c_{s,R}$ by the following linear equation[20]:

$$\ln(-GH) = -K_s c_{s,R} - \ln(-K_s K'_{eq}) \quad (6.14)$$

The salt concentration at which the peak maximum is determined by $c_{s,R} = c_{s,i} + gV_{R,g}$. The isotherm parameters are obtained from fitting a linear regression model (sklearn.linear_model.LinearRegression) using $\ln(-GH)$ and $c_{s,R}$ as y and x variables respectively. By combining the regression model and equation 6.14, K_s can be identified as the negative slope of the linear fit, while $K'_{eq} = \exp(-intercept)\, K_s^{-1}$.

6.2.6.3 Inverse method

The final approach tested in this study is the inverse method, where the parameters are tuned by running simulations and fitting the results to the experimental data. This is performed by minimizing the sum of squared errors (SSR) calculated by:

$$SSR = \sum_i \sum_{t_0}^{t_{end}} \left(\hat{c}_i(t) - c_i(t)\right)^2 \quad (6.15)$$

where $c(t)$ and $\hat{c}(t)$ are protein concentrations at the outlet of the columns at time t determined experimentally and computationally, respectively. This function is minimized using Scipy.minimize with the Nelder-Mead method and initial guesses of 0.01 and 10 for K'_{eq} and K_s respectively. Both the experimental and simulated chromatograms are scaled using a minmax scaler to normalised values between 0 and 1.

6.2.7 Error calculation

Comparing the accuracy of the isotherm parameters is performed based on retention volume and peak width at 50% intensity using the scaled chromatograms that results from mechanistic model. The simulations are performed using a near identical protocol as the experimental. As the system dwell volume is not modeled explicitly, the wash prior the gradient start is extended with this volume to a total of 6.025 mL (4.85 mL wash + 1.175 mL dwell volume). The retention volume was determined from the peak maximum and the absolute

error was calculated by subtracting the experimental retention volume, normalized errors are calculated by dividing the absolute error with the gradient length. Relative peak width is calculated by dividing the modeled peak width by the experimental peak width, both determined at 50% intensity.

6.3 Results and discussion

6.3.1 *Linear gradient elution experiments*

Linear gradient elution (LGE) experiments are required for all three HIC parameter estimation methods. The chromatographic retention of chymotrypsinogen and glucoamylase were measured for 5 gradient lengths (Figure 6.1). For both proteins it is observed that the retention shifts towards the beginning of the gradient. In the chromatogram of chymotrypsinogen at gradient lengths 15 to 40 CV, a shoulder is observed prior to the main peak. This is considered to be a result of a conformational shift, as the high salt concentrations during HIC can cause conformational changes, leading to more than one peak.[2,38,39] When the initial ammonium sulfate concentration was lowered from 1.3 M to 1.1M, this shoulder was not observed (Supplemental figure S6.1A). By increasing the initial concentration to 1.5 M, the shoulder moves towards the back of the main peak (Supplemental Figure S6.1B). This suggests that the dominant conformation shifts to the weaker binding orientation for an increasing salt concentration. While glucoamylase eluted as a symmetrical peak for all gradient lengths, an initial isocratic elution is observed during the wash. This is most notable for the 40 CV gradient length where glucoamylase elutes over the greatest volume, resulting in lower peak intensity. For both proteins the corrected retention volume is reported in table 6.1.

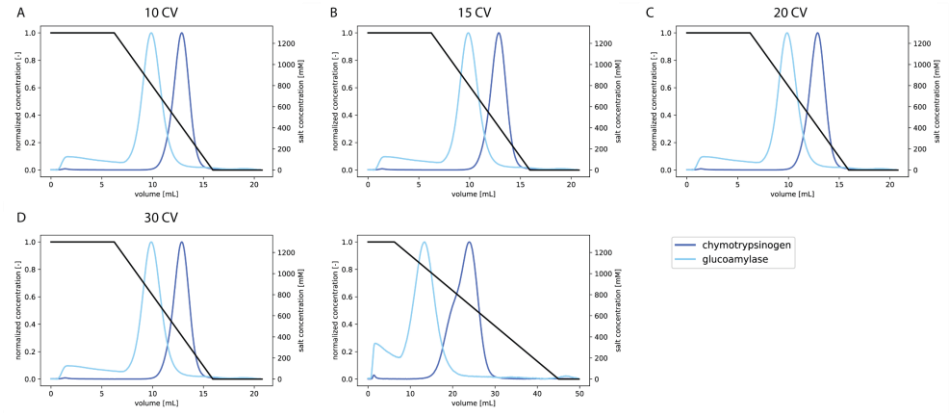


Figure 6.1: *Superimposed normalized LGE chromatograms for chymotrypsinogen (dark blue) and glucoamylase (light blue) (Hitrap Butyl FF 1 mL, flowrate 1 ml/min) with varying gradient lengths (black).*

Table 6.1: *Corrected experimental retention volumes in mL of chymotrypsinogen and glucoamylase for LGEs with five gradient lengths.*

	V_R [mL]	
gradient length	chymotrypsinogen	glucoamylase
10 CV	6.25	3.24
15 CV	8.25	4.03
20 CV	10.32	4.66
30 CV	13.91	5.7
40 CV	17.34	6.71

6.3.2 Parente and Wetlaufer method

The PW method fits isotherm parameters K'_{eq} and K_s simultaneously to the experimental data. Cross-validation by leaving out individual gradient lengths provided an accurate representation of regression accuracy for data not used in the fit (Figure 6.2A, Supplemental Figure S6.2). The cross-validation shows that the retention times of gradient lengths 15, 20 and 30 CV are predicted with high accuracy (errors <0.1 mL). For 10 and 40 CV a greater error is observed (errors >0.15 mL and >0.35 mL respectively). The difference in accuracy highlights that the PW method is less accurate when extrapolation is required. For the three intermediate gradient lengths, K'_{eq} values of 0.134 and 0.013 and K_s values of 5.681 and 5.682 chymotrypsinogen and glucoamylase were obtained, respectively.

6.3.3 Yamamoto method

When using the Yamamoto method the parameters are obtained using linear regression. Isotherm parameter K_s is directly obtained from the intercept while K'_{eq} is derived from the slope (Figure 6.2B). As observed for the PW method, the Yamamoto method estimates the retention at gradient lengths 15, 20 and 30 CV more accurately compared (errors <0.21 mL) to the gradient lengths at the bounds, especially for the 40 CV LGE, resulting in absolute errors >0.66 mL (Supplemental Figure S3). For the 15, 20 and 30 CV gradient lengths, fits with R^2 >0.97 were achieved, providing K'_{eq} values of 0.181 and 0.006 and K_s values of 5.350 and 6.488 chymotrypsinogen and glucoamylase, respectively.

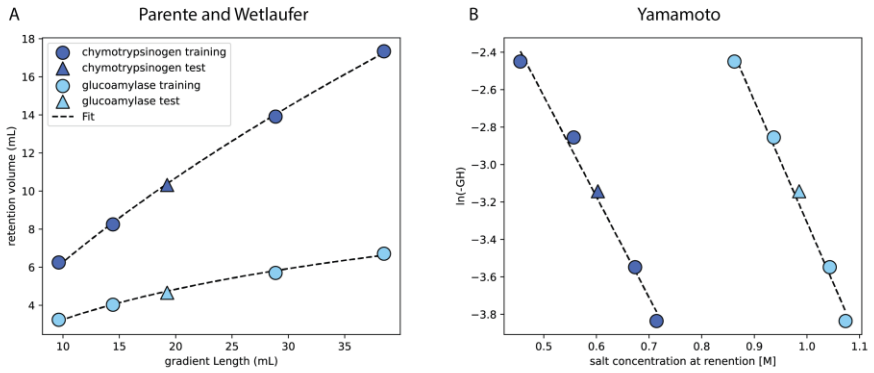


Figure 6.2: *Isotherm parameter fitting results of chymotrypsinogen (dark) and glucoamylase (light) for the 20 CV gradient length as test (triangle) and the remaining gradient lengths as fitting data (circles). A) shows the Parente and Wetlaufer method results with gradient length on the x-axis and retention volume on the y-axis. B) shows the Yamamoto method results with salt concentration of the peak maximum on the x-axis and the natural log of the normalized gradient slope on the y-axis*

6.3.4 Inverse method

While the previous two methods only require the retention volume, the inverse method uses the full chromatograms as presented in Figure 6.3. Because of this, the inverse method does not only optimize for retention volume, but also for peak shape, which comes at the cost of increased computational time (minutes compared to seconds). This

method provides average K'_{eq} values of 0.094 and 0.003 and K_s values of 6.137 and 7.331 chymotrypsinogen and glucoamylase, respectively. As observed for the PW and Yamamoto method, the cross validation shows difficulty to extrapolate gradient lengths, especially for the shorter gradient lengths (Supplemental Figure S6.4).

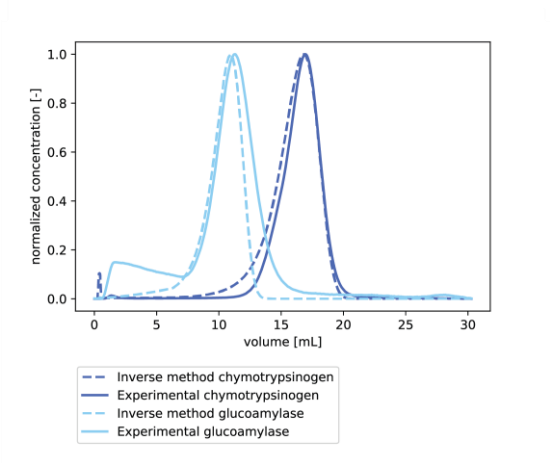


Figure 6.3: *Superimposed normalized inverse method results of chymotrypsinogen (dark) and glucoamylase (light) for the 20 CV gradient length test. The dashed line depicts the model results while the continuous line shows the experimental chromatogram.*

6.3.5 *Comparing predictive accuracy*

Given the reduced accuracy observed at the shortest (10 CV) and longest (40 CV) gradient lengths, comparisons between methods focus exclusively on the intermediate gradients of 15, 20, and 30 CV. Table 6.2 presents an overview of the isotherm parameters estimated using the three methods.

For both proteins, IM determines a lower K'_{eq} and a higher K_s compared to the other two methods. A higher K_s indicates a greater salt dependence, resulting in sharper peaks. In contrast to the standard deviation of K'_{eq} which is similar for all methods, the standard deviation for K_s is highest for IM (0.158 and 1.127), while the other two methods provide more similar deviations (0.054 to 0.073). The high standard

deviation is considered to be a result of the extensive fitting effort by IM which might amplify differences in the data being used.

Table 6.2: *Isotherm parameters obtained from cross-validation excluding 15, 20 and 30 CV gradient lengths iteratively.*

	K'_{eq} [-]	K_s [M^{-1}]
Chymotrypsinogen		
Parente and Wetlaufer	0.134±0.006	5.681±0.054
Yamamoto	0.181±0.010	5.35±0.058
Inverse method	0.094±0.011	6.137±0.158
Glucoamylase		
Parente and Wetlaufer	0.013±0.001	5.682±0.073
Yamamoto	0.006±0.001	6.488±0.063
Inverse method	0.003±0.000	7.331±1.127

To determine the actual accuracy of the different parameter combinations, simulations were performed for the 15, 20 and 30 CV gradient length experiments comparing the results to the experimental data (Figure 6.4, Supplemental Figure S6.5). Table 6.3 shows the average absolute and normalized peak maximum errors as well as the relative peak width for the different parameters. For chymotrypsinogen, peak maxima were predicted with an average error of close to 0.1 mL by all methods. For peak width, parameters obtained with IM resulted in the best agreement with the experimental data (relative peak width of 1.018). This is to be expected since this method considers chromatogram shape during the fitting. Simulations using the parameters obtained from the PW and Yamamoto method resulted in broader peaks (relative width of 1.095 and 1.164, respectively), which can be attributed to the lower K_s and higher K'_{eq} compared to the IM.

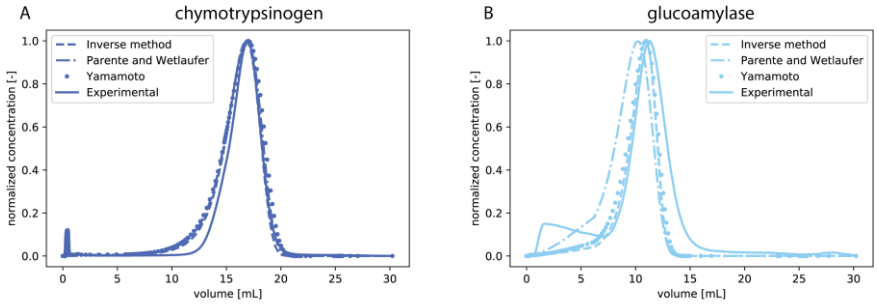


Figure 6.4: *Normalized modeled and experimental chromatogram of chymotrypsinogen (A) and glucoamylase (B) with a 20 CV linear gradient starting at 1.3 M ammonium sulfate.*

Simulations of glucoamylase retention showed to be more challenging, resulting in higher absolute errors compared to chymotrypsinogen. Especially parameters estimated using PW let to an average retention offset of 1.1 mL, while the Yamamoto method and IM achieved offsets below 0.4 mL. This might be attributed to the fact that glucoamylase elutes early in the gradient, even displaying minor isocratic elution. Interestingly, while IM yielded the most accurate retention times overall, it produced the largest deviations in peak width. As shown in Figure 4B, the simulations capture the initial slope of the chromatogram accurately but predict a too steep decent after reaching the peak maximum.

To verify whether the early elution of glucoamylase limits the accuracy of the different methods, gradients starting at 1.5 M ammonium sulfate were used for parameter determination (Supplemental Table S6.2). These parameters were subsequently used to predict the behavior of glucoamylase at a gradient running from 1.3 M to 0 M ammonium sulfate (Table 6.3). For the PW parameters, the simulation accuracy was significantly improved to an average retention error of 0.54 mL. For the Yamamoto and inverse method, accuracy was slightly reduced to an average error of 0.44 mL, while the normalized error remained constant.

Table 6.3: *Average quantitative modeling accuracy measures for the different parameter sets. *Parameters determined using the 1.5 M to 0 M ammonium sulfate LGEs.*

	chymotrypsinogen	glucoamylase	glucoamylase*
Absolute retention error [mL]			
PW	0.119	1.096	0.539
Yamamoto	0.095	0.389	0.440
IM	0.119	0.273	0.440
Normalized retention error [-]			
PW	0.0056	0.0525	0.0267
Yamamoto	0.0043	0.0192	0.0211
IM	0.0057	0.0154	0.0143
Relative peak width [-]			
PW	1.095	1.084	0.946
Yamamoto	1.164	0.917	1.023
IM	1.018	0.803	0.938

6.3.6 Predicting alternative salt gradients.

To assess the quality of the isotherm and the parameters, obtained parameter sets were tested for other salt concentrations. For chymotrypsinogen, two additional gradients were measured starting at 1.1 M and 1.5 M ammonium sulfate, both reaching a final concentration of 0 M. The additional salt gradients for glucoamylase were measured at 1.5 M to 0 M and 2.0 to 0.8 M ammonium sulfate (Figure 6.5, Table 6.4). Simulations were performed using the parameters determined based on the 1.3 M to 0 M gradients for chymotrypsinogen, while for glucoamylase the gradients starting at 1.5 M were used.

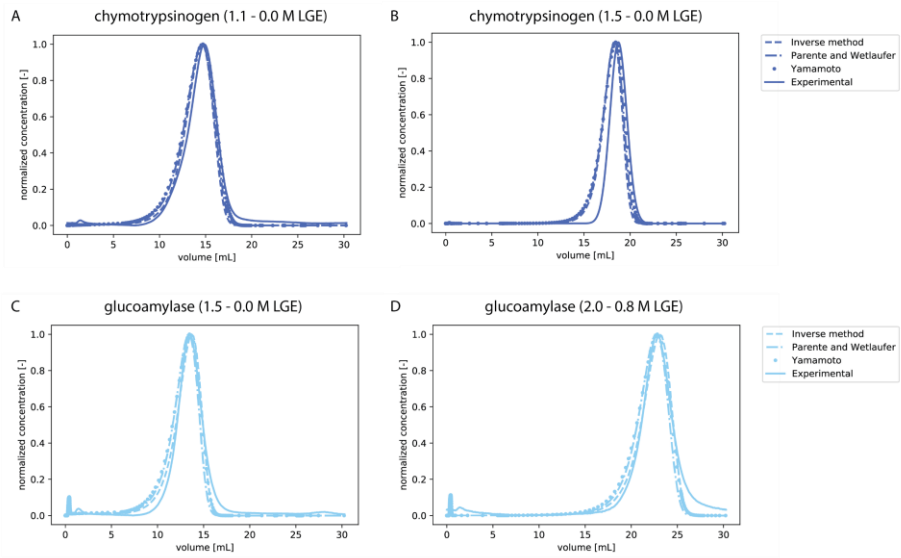


Figure 6.5: *normalized predicted (dashed, dash-doted, and dotted) and experimental (continuous) chromatograms of chymotrypsinogen (dark) and glucoamylase (light) for the 20 CV LGE at alternative buffer compositions. A) and B) show chymotrypsinogen ammonium sulfate LGEs 1.1 M to 0 M and 1.5 M to 0 M, respectively. C) and D) show glucoamylase LGEs from 1.5 M to 0 M and 2.0 M to 0.8 M, respectively.*

For all three methods, retention could be predicted with high accuracy, resulting in an average retention offset <0.44 mL. While the parameters obtained from the Yamamoto method resulted in simulated chromatograms with the smallest error in peak maxima (0.03 to 0.33 mL), the relative widths are highest (1.16 to 1.28). Simulations using the parameters obtained from the inverse method result in chromatograms with peak widths closest to the experimental peaks (1.03 to 1.12 relative widths). Peak widths for chymotrypsinogen starting at 1.5 M ammonium sulfate were estimated with the greatest deviation from the experimental data (1.12 to 1.28 relative widths). For these conditions, the chymotrypsinogen peak was deconvoluted from experimental data using two gaussians (Supplemental Figure S6.6). Therefore, the chromatogram (Figure 5B) is an estimation of the elution which might occur less symmetrically, like observed for chymotrypsinogen starting the LGE at 1.1 M ammonium sulfate (Figure 6.5A).

Table 6.4: *Average quantitative modeling accuracies for the alternative LGEs.*

	chymotrypsinogen		glucoamylase	
Ammonium sulfate [M]	1.1 – 0	1.5 - 0	1.5 - 0	2.0 – 0.8
Absolute retention error [mL]				
PW	0.357	0.431	0.070	0.143
Yamamoto	0.206	0.328	0.031	0.077
IM	0.311	0.431	0.206	0.288
Normalized retention error [-]				
PW	0.0163	0.0216	0.0039	0.0070
Yamamoto	0.0103	0.0156	0.0014	0.0028
IM	0.0162	0.0216	0.0080	0.0128
Relative peak width [-]				
PW	1.111	1.207	1.066	1.074
Yamamoto	1.179	1.283	1.164	1.171
IM	1.027	1.116	1.055	1.070

Overall, all three methods provided parameters to model the retention of both proteins accurately. Even though both correlation methods only use peak maxima information, peak shape can be predicted with a maximum over estimation of 30%. Therefore, for the simplified Mullerup isotherm, the correlations are good alternatives to the more computationally expensive inverse method. This is useful when data quality is compromised, as was seen for the LGE starting at 1.5 M ammonium sulfate for the elution of chymotrypsinogen. Comparing the two correlations, the Yamamoto method provides the most accurate peak maxima for the two proteins, while the PW method results in more accurate peak widths. Additionally, the linear representation of the Yamamoto method showed to be more robust compared to the PW method for the early eluting glucoamylase for the 1.3 M to 0 M ammonium sulfate LGE.

6.4 Conclusion

In this study, we have compared the Parente and Wetlaufer method, the Yamamoto method, and the inverse method to obtain isotherm parameters for a simplified Mullerup isotherm for HIC. The different methods applied on five LGE experiments (10, 15, 20, 30, and 40 CV gradient lengths) for chymotrypsinogen and glucoamylase. While the

different methods estimated parameters within the same order of magnitudes, the early elution of glucoamylase resulted in systemic under prediction using the parameters estimated by the PW method, which was not observed for the other methods. Overall the inverse method performed best, but it is most computationally expensive and requires high quality chromatograms. Therefore, the Yamamoto method is a good alternative for the inverse method when data quality is compromised, or computational resources are limited. This comparison offers practical guidance for isotherm determination method selection, thereby enabling reliable mechanistic modeling of HIC processes.

6.5 References

1. Eriksson, K. O. (2018). Hydrophobic Interaction Chromatography. In *Biopharmaceutical Processing* (Vol. 130). Elsevier Ltd. https://doi.org/10.1016/B978-0-08-100623-8.00019-0

2. To, B. C. S., & Lenhoff, A. M. (2007). Hydrophobic interaction chromatography of proteins. I. The effects of protein and adsorbent properties on retention and recovery. *Journal of Chromatography A*, *1141*(2), 191–205. https://doi.org/10.1016/j.chroma.2006.12.020

3. Keulen, D., Geldhof, G., Bussy, O. Le, Pabst, M., & Ottens, M. (2022). Recent advances to accelerate purification process development: A review with a focus on vaccines. *Journal of Chromatography A*, *1676*, 463195. https://doi.org/10.1016/j.chroma.2022.463195

4. Hanke, A. T., & Ottens, M. (2014). Purifying biopharmaceuticals: Knowledge-based chromatographic process development. *Trends in Biotechnology*, *32*(4), 210–220. https://doi.org/10.1016/j.tibtech.2014.02.001

5. Wittkopp, F., Welsh, J., Todd, R., Staby, A., Roush, D., Lyall, J., Karkov, S., Hunt, S., Griesbach, J., Bertran, M. O., & Babi, D. (2024). Current state of implementation of in silico tools in the biopharmaceutical industry—Proceedings of the 5th modeling workshop. *Biotechnology and Bioengineering*, *May*, 2952–2973. https://doi.org/10.1002/bit.28768

6. Nfor, B. K., Ahamed, T., Pinkse, M. W. H., van der Wielen, L. A. M., Verhaert, P. D. E. M., van Dedem, G. W. K., Eppink, M. H. M., van de Sandt, E. J. A. X., & Ottens, M. (2012). Multi-dimensional fractionation and characterization of crude protein mixtures: Toward establishment of a database of protein purification process development parameters. *Biotechnology and Bioengineering*, *109*(12), 3070–3083. https://doi.org/10.1002/bit.24576

7. Pirrung, S. M., Berends, C., Backx, A. H., van Beckhoven, R. F. W. C., Eppink, M. H. M., & Ottens, M. (2019). Model-based optimization of integrated purification sequences for biopharmaceuticals. *Chemical Engineering Science: X*, *3*, 100025. https://doi.org/10.1016/j.cesx.2019.100025

8. Vecchiarello, N., Timmick, S. M., Goodwine, C., Crowell, L. E., Love, K. R., Love, J. C.,

& Cramer, S. M. (2019). A combined screening and in silico strategy for the rapid design of integrated downstream processes for process and product-related impurity removal. *Biotechnology and Bioengineering*, *116*(9), 2178–2190. https://doi.org/10.1002/bit.27018

9. Keulen, D., Apostolidi, M., Geldhof, G., Le Bussy, O., Pabst, M., & Ottens, M. (2024). Comparing in silico flowsheet optimization strategies in biopharmaceutical downstream processes. *Biotechnology Progress*, *August*, 1–16. https://doi.org/10.1002/btpr.3514

10. Saleh, D., Wang, G., Müller, B., Rischawy, F., Kluters, S., Studts, J., & Hubbuch, J. (2020). Straightforward method for calibration of mechanistic cation exchange chromatography models for industrial applications. *Biotechnology Progress*, *36*(4), 1–12. https://doi.org/10.1002/btpr.2984

11. Keulen, D., Neijenhuis, T., Lazopoulou, A., Disela, R., Geldhof, G., Le Bussy, O., Klijn, M. E., & Ottens, M. (2024). From protein structure to an optimized chromatographic capture step using multiscale modeling. *Biotechnology Progress*, *June*, 1–26. https://doi.org/10.1002/btpr.3505

12. Kumar, V., & Lenhoff, A. M. (2020). Mechanistic Modeling of Preparative Column Chromatography for Biotherapeutics. *Annual Review of Chemical and Biomolecular Engineering*, *11*, 235–255. https://doi.org/10.1146/annurev-chembioeng-102419-125430

13. Shekhawat, L. K., Tiwari, A., Yamamoto, S., & Rathore, A. S. (2022). An accelerated approach for mechanistic model based prediction of linear gradient elution ion-exchange chromatography of proteins. *Journal of Chromatography A*, *1680*, 463423. https://doi.org/10.1016/j.chroma.2022.463423

14. Al-Ghouti, M. A., & Da'ana, D. A. (2020). Guidelines for the use and interpretation of adsorption isotherm models: A review. *Journal of Hazardous Materials*, *393*(January), 122383. https://doi.org/10.1016/j.jhazmat.2020.122383

15. Mollerup, J. M. (2006). Applied thermodynamics: A new frontier for biotechnology. *Fluid Phase Equilibria*, *241*(1–2), 205–215. https://doi.org/10.1016/j.fluid.2005.12.037

16. Mollerup, J. M. (2007). The thermodynamic principles of ligand binding in chromatography and biology. *Journal of Biotechnology*, *132*(2), 187–195. https://doi.org/10.1016/j.jbiotec.2007.05.036

17. Mollerup, J. M. (2008). A review of the thermodynamics of protein association to ligands, protein adsorption, and adsorption isotherms. *Chemical Engineering and Technology*, *31*(6), 864–874. https://doi.org/10.1002/ceat.200800082

18. Lietta, E., Pieri, A., Cardillo, A. G., Vanni, M., Pisano, R., & Barresi, A. A. (2022). An Experimental and Modeling Combined Approach in Preparative Hydrophobic Interaction Chromatography. *Processes*, *10*(5). https://doi.org/10.3390/pr10051027

19. Andris, S., & Hubbuch, J. (2020). Modeling of hydrophobic interaction chromatography for the separation of antibody-drug conjugates and its application towards quality by design. *Journal of Biotechnology*, *317*(April), 48–58. https://doi.org/10.1016/j.jbiotec.2020.04.018

20. Yang, Y. X., Chen, Y. C., Yao, S. J., & Lin, D. Q. (2024). Parameter-by-parameter estimation method for adsorption isotherm in hydrophobic interaction chromatography. *Journal of Chromatography A*, *1716*(November 2023), 464638. https://doi.org/10.1016/j.chroma.2024.464638

21. Hahn, T., Baumann, P., Huuk, T., Heuveline, V., & Hubbuch, J. (2016). UV absorption-based inverse modeling of protein chromatography. *Engineering in Life Sciences*, *16*(2), 99–106. https://doi.org/10.1002/elsc.201400247

22. Osberghaus, A., Hepbildikler, S., Nath, S., Haindl, M., von Lieres, E., & Hubbuch, J. (2012). Determination of parameters for the steric mass action model-A comparison between two approaches. *Journal of Chromatography A*, *1233*, 54–65. https://doi.org/10.1016/j.chroma.2012.02.004

23. Osberghaus, A., Hepbildikler, S., Nath, S., Haindl, M., von Lieres, E., & Hubbuch, J. (2012). Optimizing a chromatographic three component separation: A comparison of mechanistic and empiric modeling approaches. *Journal of Chromatography A*, *1237*, 86–95. https://doi.org/10.1016/j.chroma.2012.03.029

24. Rajagopala, S. V., Sikorski, P., Kumar, A., Mosca, R., Vlasblom, J., Arnold, R., Franca-Koh, J., Pakala, S. B., Phanse, S., Ceol, A., Häuser, R., Siszler, G., Wuchty, S., Emili, A., Babu, M., Aloy, P., Pieper, R., & Uetz, P. (2014). The binary protein-protein interaction landscape of escherichia coli. *Nature Biotechnology*, *32*(3), 285–290. https://doi.org/10.1038/nbt.2831

25. Kumar, V., Leweke, S., von Lieres, E., & Rathore, A. S. (2015). Mechanistic modeling of ion-exchange process chromatography of charge variants of monoclonal antibody products. *Journal of Chromatography A*, *1426*, 140–153. https://doi.org/10.1016/j.chroma.2015.11.062

26. Parente, E. S., & Wetlaufer, D. B. (1986). Relationship between isocratic and gradient retention times in the high-performance ion-exchange chromatography of proteins. Theory and experiment. *Journal of Chromatography A*, *355*(C), 29–40. https://doi.org/10.1016/S0021-9673(01)97301-7

27. Yamamoto, S. (2005). Electrostatic interaction chromatography process for protein separations: Impact of engineering analysis of biorecognition mechanism on process optimization. *Chemical Engineering and Technology*, *28*(11), 1387–1393. https://doi.org/10.1002/ceat.200500199

28. Hanke, A. T., Tsintavi, E., Ramirez Vazquez, M. del P., van der Wielen, L. A. M., Verhaert, P. D. E. M., Eppink, M. H. M., van de Sandt, E. J. A. X., & Ottens, M. (2016). 3D-liquid chromatography as a complex mixture characterization tool for knowledge-based downstream process development. *Biotechnology Progress*, *32*(5), 1283–1291. https://doi.org/10.1002/btpr.2320

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

30. Schmidt-Traub, H., Schulte, M., & Seidel-Morgenstern, A. (2020). Preparative chromatography. In *Preparative Chromatography: Third Edition* (3rd ed.). John Wiley & Sons. https://doi.org/10.1002/9783527816347

31. Keulen, D., Hagen, E. van der, Geldhof, G., Le Bussy, O., Pabst, M., & Ottens, M. (2024). Using artificial neural networks to accelerate flowsheet optimization for downstream process development. *Biotechnology and Bioengineering*, *121*(8), 2318–2331. https://doi.org/10.1002/bit.28454

32. Hagel, L. (2011). Protein Purification- 3 - Gel filtration: size exclusion chromatography. *Protein Purification: Principles, High Resolution Methods, and Applications*, 51–91.

33. Nfor, B. K., Noverraz, M., Chilamkurthi, S., Verhaert, P. D. E. M., van der Wielen, L. A. M., & Ottens, M. (2010). High-throughput isotherm determination and thermodynamic modeling of protein adsorption on mixed mode adsorbents. *Journal of Chromatography A*, *1217*(44), 6829–6850. https://doi.org/10.1016/j.chroma.2010.07.069

34. Deitcher, R. W., Rome, J. E., Gildea, P. A., O'Connell, J. P., & Fernandez, E. J. (2010). A new thermodynamic model describes the effects of ligand density and type, salt

concentration and protein species in hydrophobic interaction chromatography. *Journal of Chromatography A*, *1217*(2), 199–208. https://doi.org/10.1016/j.chroma.2009.07.068

35. Hess, R., Yun, D., Saleh, D., Briskot, T., Grosch, J. H., Wang, G., Schwab, T., & Hubbuch, J. (2023). Standardized method for mechanistic modeling of multimodal anion exchange chromatography in flow through operation. *Journal of Chromatography A*, *1690*, 463789. https://doi.org/10.1016/j.chroma.2023.463789

36. Chen, J., Yang, T., & Cramer, S. M. (2008). Prediction of protein retention times in gradient hydrophobic interaction chromatographic systems. *Journal of Chromatography A*, *1177*(2), 207–214. https://doi.org/10.1016/j.chroma.2007.11.003

37. Rüdt, M., Gillet, F., Heege, S., Hitzler, J., Kalbfuss, B., & Guélat, B. (2015). Combined Yamamoto approach for simultaneous estimation of adsorption isotherm and kinetic parameters in ion-exchange chromatography. *Journal of Chromatography A*, *1413*, 68–76. https://doi.org/10.1016/j.chroma.2015.08.025

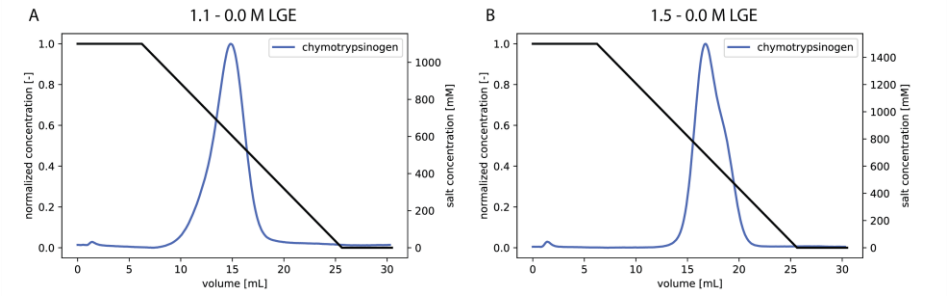
38. Ueberbacher, R., Haimer, E., Hahn, R., & Jungbauer, A. (2008). Hydrophobic interaction chromatography of proteins. V. Quantitative assessment of conformational changes. *Journal of Chromatography A*, *1198–1199*(1–2), 154–163. https://doi.org/10.1016/j.chroma.2008.05.062

39. Beyer, B., & Jungbauer, A. (2018). Conformational changes of antibodies upon adsorption onto hydrophobic interaction chromatography surfaces. *Journal of Chromatography A*, *1552*, 60–66. https://doi.org/10.1016/j.chroma.2018.04.009

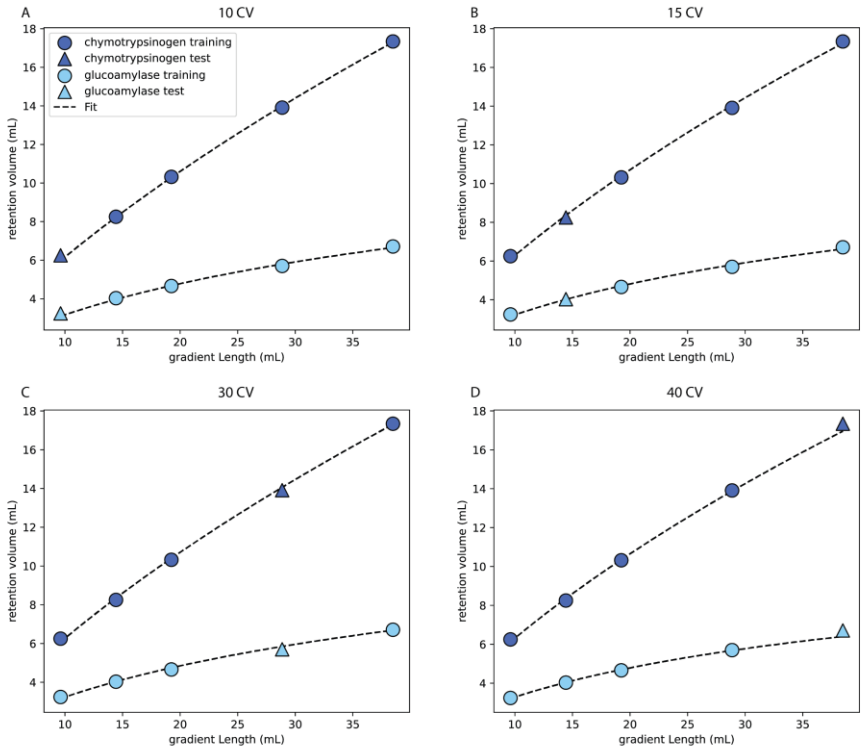
6.6 Supplemental information

Supplemental Table S6.1: *system and column parameters*

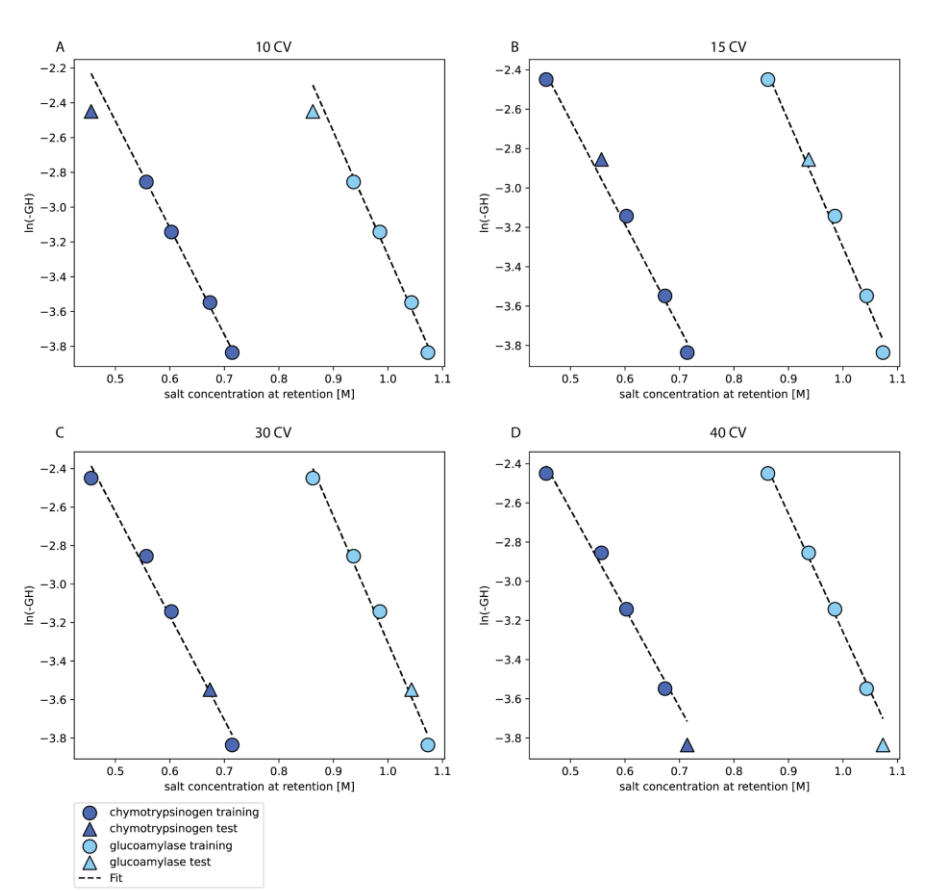
Parameter	Value	Unit
Column diameter	0.70	cm
Column height	2.50	cm
Particle size	90	µm
Total porosity (ε_t)	0.914	-
Extraparticle porosity (ε_b)	0.336	-
Intraparticle porosity (ε_p)	0.870	-
System dead volume	0.281	mL
System dwell volume	1.175	mL



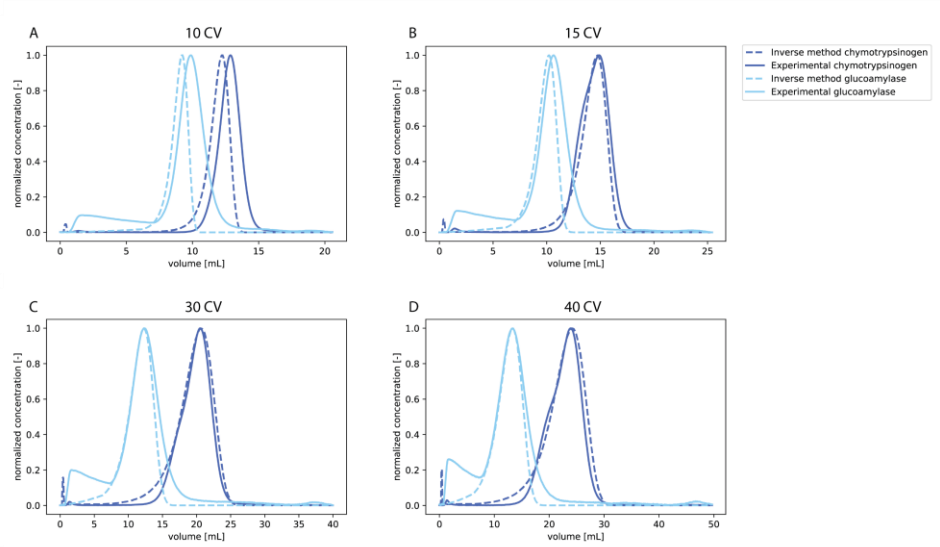
Supplemental Figure S6.1: *Linear gradient elution chromatograms of chymotrypsinogen with a 20 CV gradient length starting at 1.1 M (A) and 1.5 M (B) $(NH_4)_2SO_4$*



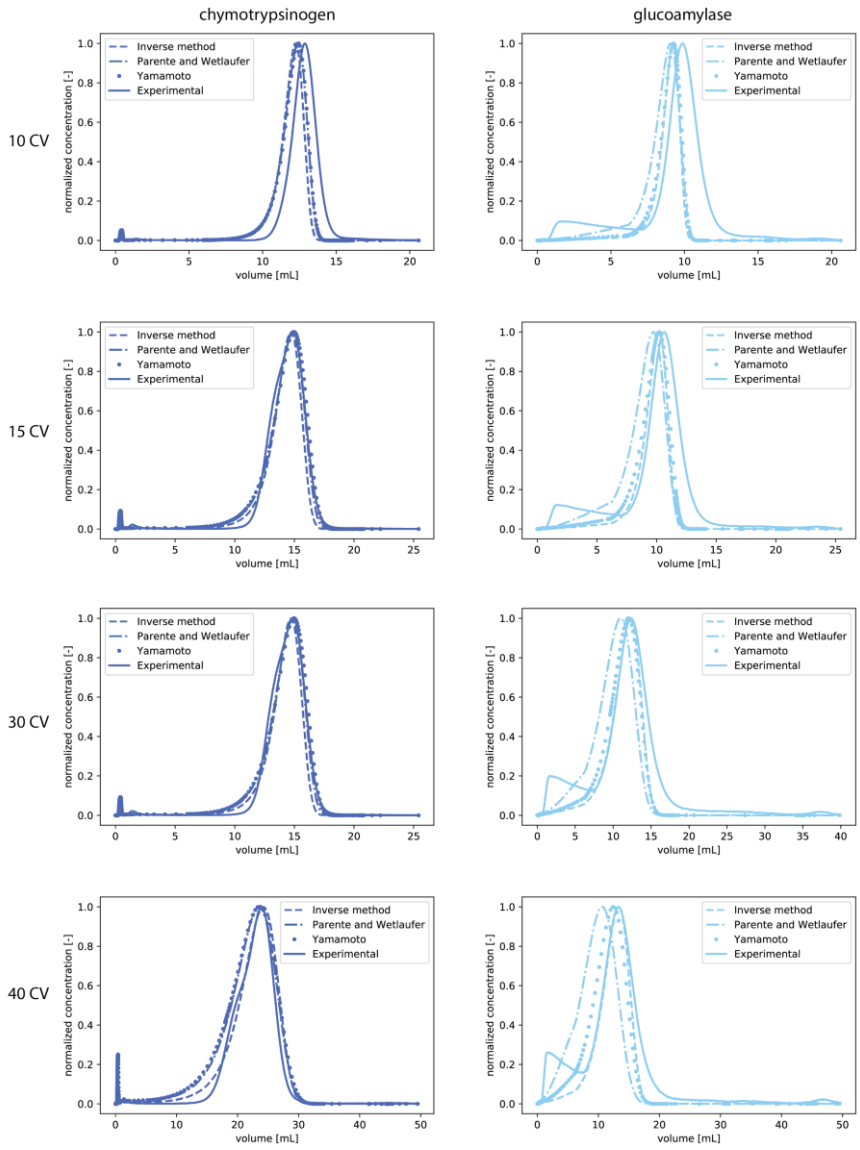
Supplemental Figure S6.2: *Parente and Wetlaufer Isotherm parameter fitting results of chymotrypsinogen (dark) and glucoamylase (light) for the 10, 15, 30 and 40 CV gradient length tests (triangle) and the remaining gradient lengths as fitting data (circles) with gradient length on the x-axis and retention volume on the y-axis.*



Supplemental Figure S6.3: *Yamamoto isotherm parameter fitting results of chymotrypsinogen (dark) and glucoamylase (light) for the 10, 15, 30 and 40 CV gradient length as test (triangle) and the remaining gradient lengths as fitting data (circles), with salt concentration of the peak maximum on the x-axis and the natural log of the normalized gradient slope on the y-axis*



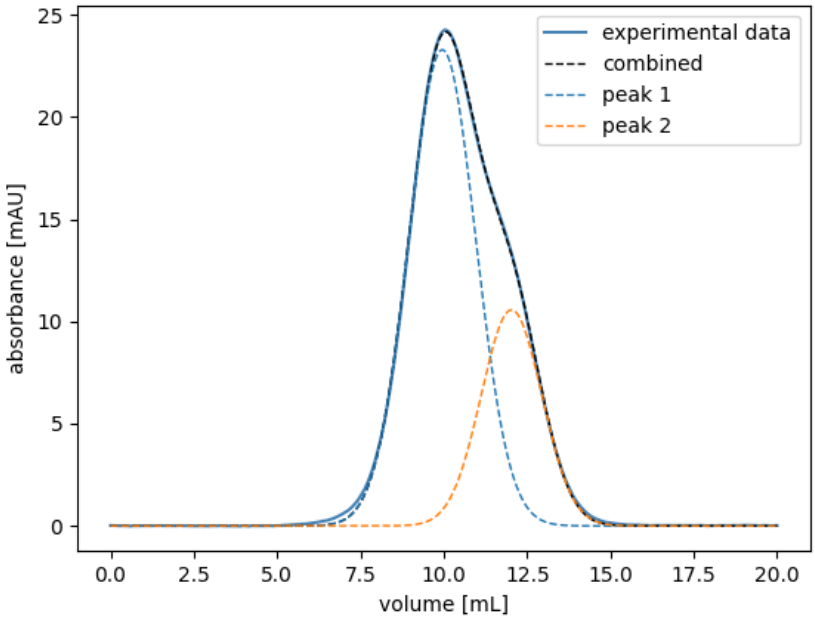
Supplemental Figure S6.4: *Superimposed normalized inverse method results of chymotrypsinogen (dark) and glucoamylase (light) for the 10, 15, 30 and 40 CV gradient length tests. The dashed line depicts the model results while the continuous line shows the experimental chromatogram.*



Supplemental Figure S6.5: *Normalized modeled and experimental chromatograms of chymotrypsinogen and glucoamylase at 10, 15, 30, and 40 CV elution gradient lengths.*

Supplemental Table S6.2: *Isotherm parameters for glucoamylase determined using 1.5 to 0 M ammonium sulfate LGEs*

	Keq [-]	Ks [M-1]
Parente and Wetlaufer	0.007±0.000	6.34±0.088
Yamamoto	0.014±0.001	5.819±0.069
Inverse method	0.008±0.003	6.365±0.417



Supplemental Figure S6.6: *Deconvolution of the chymotrypsinogen chromatogram of a 20 CV LGE from 1.5 M to 0 M ammonium sulfate.*

Chapter 7

Conclusion and outlook

7.1 Conclusion

Model-based process development (PD) of biopharmaceuticals can significantly increase productivity and thereby shorten the lab to market timelines. Understanding fundamental phenomena allows a reduction in experimental effort and costs. For protein chromatography modeling, predicting the interaction behavior between the protein and resin is the most challenging. These interactions are governed by physicochemical properties which are a result of the protein amino acid sequence and subsequent 3D structure. Structure models contain the coordinates of every atom. Therefore, they have the potential to provide all information.

This thesis shows how protein structures can be used for prediction of their chromatographic behavior. Specifically, a comprehensive quantitative structure property relationship (QSPR) workflow is presented. For this, an open-source software package was developed (**chapter 2**) that was validated for literature data on ion exchange chromatography (IEX). This package allows wide deployment of QSPR using the Python programming language that is available for the whole scientific community. This contributes to general progress in the field by providing transparency, thereby lowering the initial investment for beginners. Additionally, the distribution of the source code allows for customizability by experts.

This software package was successfully used in a multiscale modeling approach (**chapter 3**). For a total of six model proteins, QSPR models were trained that accurately predict retention times for different salt gradient conditions, which were subsequently used in a regression formula to obtain isotherm parameters. These parameters were used for process optimization resulting in similar optimal operation conditions when using experimentally determined parameters. Additionally, it was observed that the estimation uncertainties have a

minimal effect on the operation condition selection. This highlights the potential of QSPR for PD.

Extensive knowledge of crude mixture is essential to be able to apply QSPR for the development of a capture step. Therefore, in **chapter 4** retention profiles of Escherichia coli BLR(DE3) lysates were determined for hydrophobic interaction chromatography (HIC) and IEX chromatography. Retention times of around 900 unique host cell proteins (HCPs) could be determined. By analyzing protein subsets based on location, function, and interactions, it was observed that proteins located in the plasma membrane or that are participating in protein-protein interactions deviate from general elution trends. Using predicted protein structures, QSPR models could be trained to predict HCP retention times, which was most successful for monomeric proteins.

As experimental characterization remains expensive and time consuming, the ability of a set of widely available proteins to represent the HCPs was assessed in **chapter 5**. For IEX, a QSPR model could be trained that predicts part of the HCP data also presented in **chapter 4**. No difference in accuracy was observed when predicting HCP subsets. This shows that pure protein data can be used for predicting HCP behavior in a mixture. Key differences in feature distributions were found in the training and testing data, indicating areas of improvements.

Finally, a study of different isotherm parameter determination methods for HIC was performed in **chapter 6**. This study showed that both correlation methods and inverse fitting methods results in mechanistic model simulations with similar accuracies. Additionally, the robustness and computational complexity of the different methods are discussed. As such, it provides practical guidance for method selection, thereby enabling reliable mechanistic modeling of HIC processes.

These findings are a significant step towards more accessible model-based PD. They provide evidence that protein structures can be used to predict chromatographic behavior, specifically for IEX. The recent advances in protein structure prediction methods, like AlphaFold, have shown to be highly valuable for partly elucidating the chromatographic behavior of the host cell proteome. Still, there remain some challenges that need to be overcome in order to fully implement these workflows for a broad range of resin types.

7.2 Outlook

7.2.1 Surface hydrophobics

While this thesis shows successful QSPR modeling of IEX, it does not present any models for HIC or mixed mode chromatography. While hydrophobicity features can be calculated by the software in **chapter 2**, attempts in training models to predict HIC retention were unsuccessful. These prediction challenges are thought to be the result of the complex adsorption mechanism, which is driven by the entropy of water.[1] Extending the calculated protein features might enable prediction of HIC processes and improve prediction of other modes. Spatial aggregation propensity maps are an alternative method to map local hydrophobicity by describing regions which are likely to aggregate. Therefore, these maps might be better at capturing the forces that drive HIC adsorption.[2–4]

7.2.2 Protein binding conformations

Another challenge that needs to be tackled for HIC is alternative binding conformations resulting from the high salt conditions, as observed in **chapter 6**. Implementation of local features, like surface patches, might partially solve this issue. As preferred binding orientations are not only present for HIC, local features could also greatly improve predictions for other chromatography modes.

Additionally, proteins might be susceptible to aggregation or unfolding upon binding, which is currently not described by the static structures used for the QSPR. Calculating local flexibility scores can potentially provide indications of surface areas that might undergo conformational changes.[5] These additional features may indicate which proteins will present alternative binding conformations. Alternative protein structures can be generated by constrained molecular dynamics (MD) introducing additional rigidity in inflexible regions coupled with advanced sampling methods like simulated annealing.[6] Unfortunately, these methods are often paired with high computational costs and should therefore be reserved for high interest targets.

7.2.3 Protein docking

Protein docking is another method that can provide detailed information of the interactions.[7–9] This method uses the molecular structure representing the protein and resin to minimize the binding energies. Accurate description of the chromatographic resin is currently a limiting factor. Often the resin is modeled as a plane of ligand molecules, which lacks the three-dimensional pore structure.[10] Ballweg et al. proposed a method to simulate polymerization reactions that form the resin beads, resulting in a complex resin structure.[9] They revealed that not only the ligand, but also the backbone of the resin is essential to accurately estimate affinity of peptide. While this is very relevant, application on larger scales like complete proteins is limited due to the high associated computational costs. Recent developments in graphical processing unit (GPU) acceleration and advances in hardware allowed for a reduction in computational times for protein docking and MD simulations.[11] These advances are mainly attributable to the increased power budgets and improved thermal regulation.[12,13] Therefore, advancement might be limited due to foreseeable issues regarding sustainability.

7.2.4 Hybrid QSPR-MM models

Considering model-based PD, accurate description of protein adsorption behavior remains challenging. Hybrid modeling provides a solution by using black box models to predict phenomena that are ill understood.[14,15] Currently, these hybrid models have been limited to the use of machine learning models trained on experimental data. Therefore, they provide no knowledge that can be transferred to other targets. Implementation of QSPR into the hybrid modeling framework could bring new possibilities. Specifically, by training QSPR models that predict the ratio between bound and unbound proteins as a function of protein and salt concentration, the dynamic binding behavior can be predicted directly from the protein structure. This approach would solve limitations with the adsorption isotherms that require multiple parameters to be fit and only describe a single isotherm shape.

7.2.5 Model proteomes

In **chapter 5**, we attempt to describe the *E. coli* proteome using a set of 13 model proteins, which was successful for over 200 HCPs. Additionally, a set of improvement strategies are described in this chapter. Specifically, extending the model protein dataset to capture most relevant protein features is relevant for straight-forward QSPR deployment. For initial screening, a general proteome would be suitable to provide retention predictions to support early phase resin selections. For more accurate predictions, model proteomes could be tuned to represent specific hosts as post-translational modifications might vary. Currently, pure protein solutions provide the highest quality retention data and have been shown to represent close to 25% of the observer *E. coli* proteins accurately.

While there have been major breakthroughs in the field of chromatography modeling as well as protein structure prediction, significant investments are required before the field reaches maturity.

The suggestions for follow-up research as well as an increased effort in data collection and curation will enable further advancements. Ultimately, progress in model-based PD will allow for faster and more cost-effective processing.

7.3 References

1. Eriksson, K. O. (2018). Hydrophobic Interaction Chromatography. In *Biopharmaceutical Processing* (Vol. 130). Elsevier Ltd. https://doi.org/10.1016/B978-0-08-100623-8.00019-0

2. Robinson, J. R., Karkov, H. S., Woo, J. A., Krogh, B. O., & Cramer, S. M. (2017). QSAR models for prediction of chromatographic behavior of homologous Fab variants. *Biotechnology and Bioengineering*, *114*(6), 1231–1240. https://doi.org/10.1002/bit.26236

3. Banerjee, S., Parimal, S., & Cramer, S. M. (2017). A molecular modeling based method to predict elution behavior and binding patches of proteins in multimodal chromatography. *Journal of Chromatography A*, *1511*, 45–58. https://doi.org/10.1016/j.chroma.2017.06.059

4. Sankar, K., Trainor, K., Blazer, L. L., Adams, J. J., Sidhu, S. S., Day, T., Meiering, E., & Maier, J. K. X. (2022). A Descriptor Set for Quantitative Structure-property Relationship Prediction in Biologics. *Molecular Informatics*, *41*(9), 2100240. https://doi.org/10.1002/minf.202100240

5. Narwani, T. J., Etchebest, C., Craveur, P., Léonard, S., Rebehmed, J., Srinivasan, N., Bornot, A., Gelly, J. C., & de Brevern, A. G. (2019). In silico prediction of protein flexibility with local structure approach. *Biochimie*, *165*, 150–155. https://doi.org/10.1016/j.biochi.2019.07.025

6. Bernardi, R. C., Melo, M. C. R., & Schulten, K. (2015). Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochimica et Biophysica Acta - General Subjects*, *1850*(5), 872–877. https://doi.org/10.1016/j.bbagen.2014.10.019

7. Kallberg, K., Johansson, H. O., & Bulow, L. (2012). Multimodal chromatography: An efficient tool in downstream processing of proteins. *Biotechnology Journal*, *7*(12), 1485–1495. https://doi.org/10.1002/biot.201200074

8. Salha, D., Andaç, M., & Denizli, A. (2021). Molecular docking of metal ion immobilized ligands to proteins in affinity chromatography. *Journal of Molecular Recognition*, *34*(2), 1–11. https://doi.org/10.1002/jmr.2875

9. Ballweg, T., Liu, M., Grimm, J., Sedghamiz, E., Wenzel, W., & Franzreb, M. (2024). All-atom modeling of methacrylate-based multi-modal chromatography resins for Langmuir constant prediction of peptides. *Journal of Chromatography A*, *1730*(June), 465089. https://doi.org/10.1016/j.chroma.2024.465089

10. Jakobtorweihen, S., Heuer, J., & Waluga, T. (2020). A novel approach to calculate protein adsorption isotherms by molecular dynamics simulations. *Journal of Chromatography A*, *1620*, 460940. https://doi.org/10.1016/j.chroma.2020.460940

11. Pandey, M., Fernandez, M., Gentile, F., Isayev, O., Tropsha, A., Stern, A. C., & Cherkasov, A. (2022). The transformational role of GPU computing and deep learning in drug discovery. *Nature Machine Intelligence*, *4*(3), 211–221. https://doi.org/10.1038/s42256-022-00463-x

12. Patel, P., Gong, Z., Rizvi, S., Choukse, E., Misra, P., Anderson, T., & Sriraman, A. (2023). Towards Improved Power Management in Cloud GPUs. *IEEE Computer Architecture Letters*, *22*(2), 141–144. https://doi.org/10.1109/LCA.2023.3278652

13. Bridges, R. A., Imam, N., & Mintz, T. M. (2016). Understanding GPU power: A survey of profiling, modeling, and simulation methods. *ACM Computing Surveys*, *49*(3). https://doi.org/10.1145/2962131

14. Narayanan, H., Seidler, T., Luna, M. F., Sokolov, M., Morbidelli, M., & Butté, A. (2021). Hybrid Models for the simulation and prediction of chromatographic processes for protein capture. *Journal of Chromatography A*, *1650*, 462248. https://doi.org/10.1016/j.chroma.2021.462248

15. Narayanan, H., Luna, M., Sokolov, M., Arosio, P., Butté, A., & Morbidelli, M. (2021). Hybrid Models Based on Machine Learning and an Increasing Degree of Process Knowledge: Application to Capture Chromatographic Step. *Industrial and Engineering Chemistry Research*, *60*(29), 10466–10478. https://doi.org/10.1021/acs.iecr.1c01317

Acknowledgements

First of all, I would like to thank **Marcel** as my supervisor and promotor for his guidance throughout the project. You enabled me to grow as an independent researcher and gave me the confidence to pursue my own ideas. I enjoyed our often-brief discussions, and I am grateful for all opportunities you provided.

Marieke, my second supervisor and copromotor. Thank you for your critical view and desire for deep understanding on all relevant topics. Your feedback on my writing has been invaluable, even though some reviewers deemed the supplemental discussion of chapter 2 too detailed.

During this project, I have had the privilege of working together with GSK, providing invaluable support. I want to thank **Geoffroy and Olivier**. Our periodic meetings resulted in additional direction and motivation, fueled by your enthusiasm.

Daphne and Roxana, who were stuck with me in our collaborative project. The two of you started about two years before I arrived, yet you made me feel welcome from the start. **Daphne**, sitting across from me obscured by our many monitors, I would just stand up to get your attention. Thank you for our countless joy-filled conversations, from which you would occasionally zone out. I admire your sense of direction and confidence throughout our collaboration and beyond. **Roxana**, your laughter can be recognized from thousands and could often be heard many offices further. I enjoyed working with you, especially spending time locked away while working on joined paper/chapter. Your eye for detail sets you apart, fixating on tiny details for which you are willing to empty the library. I am very grateful for both of you, and I believe that our mix of serious and less serious discussion sessions was key to our collaboration.

The three Master students to who I had the privilege of supervising: **Tijn, Adamantia and Tomás**. Seeing all of you grow within your projects made me proud and provided me with great new insights.

Starting just two months before me, **Maarten** has been my companion in navigating all that PhD at TU Delft has to offer. Now standing next to me as my paranymph. While your height might be intimidating, your kindness is radiant. Thank you for defusing any situation involving drunk students in the Kurk. I knew that whenever my office was fed up with my banter, I could always knock on your door. The memories we made on our US road trip will be with me forever. But please remind me to book a bed larger than 140 cm whenever we plan for something similar.

My other wonderful paranymph **Ramon**. While your height is less impressive than Maarten's, your knowledge of computational methods makes you equally intimidating. I enjoyed our sparring sessions about statistical methods and listening to your monologues on CFD analysis, to which I could only smile and nod. Equally so, our occasional Friday afternoon chess match or taking a glass from the bierfabriek and getting caught 100m down the road while chugging for dear life. You also introduced me to the Wim Hoff philosophy on cold showers, I can't believe you have such a warm heart while exposing yourself to those temperatures.

Everyone else I shared an office with. **Tim**, you filled the desk next to me that was empty for months due to covid regulations. You showed what it is like to be a real scientist. I enjoyed the time of "us Tim" and the confusion of everyone while introducing ourselves. I cherish our trip to Portugal which we closed off by eating that monstrosity of a meal in Porto and finishing with a port older than both of us. **Max**, I enjoyed reading every new fokke en sukke you would hang on the wall. **Rob,** every Thursday you would join, I knew I would leave that day

with new perspectives. **Ben**, your passion and enthusiasm are radiant. **Jelle**, I very much enjoyed our conversation on peculiar cooking techniques, and I will be rooting for your 1:45 half marathon, I know you can do it.

Next, I would like to thank all my other BPE colleagues. **Meryl**, to me you are the embodiment of Gen Z, and at first it seemed like we were speaking another language. But teaching with you has been one of the most laughter filled weeks of my PhD. Now I feel like we can level on a lot of things, and if you are ever in Breda, remember that dancing is free. **Mariana**, my chromatography buddy, when everyone else left. I am sure I will say your name correctly one day. **Rik**, I enjoyed our discussions on random topics to which you would bring interesting perspectives. Also, thanks for our trip to Zurich where we walked God knows where to find that one specific cordon bleu. **Miki**, the third person to join the course in Zurich. I cherish the conversations on the real priorities in life. **Marika, Dimitri, Hector, Pieter, Tamara, Joanna, Nicole, Brenda, Moumita, and Mungyu** for all great conversations.

Tiago, thank you for every time I crashed at your place. As well as the friendly competition during all BPE activities for which you were the most important person to beat. Also, for joining me to the concert of While She Sleeps, twice, during which we would only briefly meet due to the constant crowd surfing. **Mariana**, while your poker face is often difficult to read, I am thankful for your kindness as a fellow non-engineer. **Marina**, your enthusiasm is infectious. **Lars**, thank you for brightening most Friday evenings with your cheerful laugh. **Oriol**, for being there when it counts, but completely forgetting about concert tickets. **Marijn**, for organizing bowling activities in which you can't participate yourself. **Joan, Mona, Zulhaj,** and **Eduardo**, thanks for all our insightful discussions.

The BPE staff **Cees, Ludo, Adri, Simon, Stef, Jeroen** and **Kawieta,** thank you for being indispensable for the stable foundation of the section. **Christiaan**, thanks for challenging me and providing my most humbling experience in recent years. **Song**, thank you for your patience when teaching me how to operate the Äkta.

To the people that started during the last few months of my PhD: **Gianmarco, Lorin, and Tom.** Thanks for our brief interactions, I am sure you will enjoy the department as much as I have. **Ester**, you have taken on the challenge of continuing what Daphne, Roxana, and I started. I am sure you will bring interesting advances, and I am curious which direction you will go.

My Parents, thank you for providing everything I needed to thrive. I am grateful for your patience and the time you spend on helping me improve my reading abilities. Also, for stimulating me take on numerous hobbies and teaching me that anything can be achieved with hard work and dedication.

And finally, my partner **Dewi**, what an online video game can bring. Thank you for being there when I need you, even late at night when no trains can take me home. With you it is easy to let go of worries, whether through baking and cooking, gardening, walking Guusje or simply relaxing. I cannot wait to find out what the future will bring during this exciting next chapter in our lives.

Curriculum vitae

Tim Neijenhuis was born on 13 February 1995 in Wageningen, The Netherlands. After completing high school, where he obtained his VMBO-TL diploma in 2011 and his HAVO diploma in 2013, he enrolled in the Bachelor Biomedical Laboratory Research at the Hogeschool Arnhem Nijmegen. Driven by his fascination with biology and chemistry, he graduated with a major in biochemistry and a minor in organic chemistry.

In 2017, Tim began a one-year pre-master program, followed by a two-year Master's in Molecular Life Sciences at Radboud University, Nijmegen. During his specialization in *Chemistry of Life*, he developed a strong interest in protein chemistry. This curiosity led him to undertake a six-month research project at the Centre of Informatics and Bioinformatics at Radboud University, followed by another project at the Computational Structural Biology group at Utrecht University.

After completing his MSc in 2020, Tim joined the Bioprocess Engineering section of the Biotechnology Department at Delft University of Technology in 2021, under the supervision of Marcel Ottens. His four-year PhD project focused on extracting meaningful insights from protein structural models to predict their behavior during preparative chromatography. The findings of this research are presented in this thesis.

List of publications

Journal articles

Neijenhuis, T., Le Bussy, O., Geldhof, G., Klijn, M. E., & Ottens, M. (2025). Using generalized quantitative structure–property relationship (QSPR) models to predict host cell protein retention in ion-exchange chromatography. *Journal of Chemical Technology & Biotechnology*.

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

Disela, R., Keulen, D., Fotou, E., **Neijenhuis, T.,** Le Bussy, O., Geldhof, G., Pabst, M., & Ottens, M. (2024) Proteomics-based method to comprehensively model the removal of host cell protein impurities. *Biotechnology Progress*, *40*(6), e3494.

Keulen, D., **Neijenhuis, T.,** Lazopoulou, A., Disela, R., Le Bussy, O., Geldhof, G., Klijn, M. E., & Ottens, M. (2024) From protein structure to an optimized chromatographic capture step using multiscale modeling. *Biotechnology Progress*, *41*(1), e3505.

Neijenhuis, T., Le Bussy, O., Geldhof, G., Klijn, M. E., & Ottens, M. (2024). Predicting protein retention in ion-exchange chromatography using an open source QSPR workflow. *Biotechnology Journal,* 19, e2300708.

Neijenhuis, T., van Keulen, S. C, & Bonvin, A. M. (2022). Interface refinement of low-to medium-resolution cryo-EM complexes using HADDOCK2. 4. *Structure*, *30*(4), 476-484.

Conference contributions

Neijenhuis, T., Disela, R., Le Bussy, O., Geldhof, G., Klijn, M. E., & Ottens, M., Description and prediction of E. coli host cell protein behavior in anion exchange chromatography. 19th International Symposium on Preparative and Industrial Chromatography and Allied Techniques, Milan, Italy, October 2024, Science slam presentation

Neijenhuis, T., Keulen, D., Lazopoulou, A., Disela, R., Le Bussy, O., Geldhof, G., Klijn, M. E., & Ottens, M., Multiscale modeling for protein retention prediction in cation exchange chromatography. The 19th International PhD Seminar on Chromatographic Separation Science, Lund, Sweden, May 2024, Oral presentation

Neijenhuis, T., Le Bussy, O., Geldhof, G., Klijn, M. E., & Ottens, M., Protein modeling as a tool for chromatographic separation prediction in early downstream process design. 14th European Congress of Chemical Engineering and 7th European Congress of Applied Biotechnology, Berlin, Germany, September 2023, Oral presentation

Neijenhuis, T., Le Bussy, O., Geldhof, G., Klijn, M. E., & Ottens, M,. Protein modeling as a tool for chromatographic separation prediction in early downstream process design. American Chemical Society (ACS), San Fransisco, United States of America, August 2023, Poster presentation

Neijenhuis, T., Le Bussy, O., Geldhof, G., Klijn, M. E., & Ottens, M., Protein modeling as a tool for chromatographic separation prediction. The 18th International PhD Seminar on Chromatographic Separation Science, Düren, Germany, June 2023, Oral presentation

Neijenhuis, T., Pabst, M., Klijn, M. E., & Ottens, M., Protein Quantitative Structure-Property Relationships (QSPR) for improved chromatographic separation. Biopartitioning and Purification conference (BPP), Aveiro, Portugal, September 2022, poster presentation.

Neijenhuis, T., Pabst, M., Klijn, M. E., & Ottens, M., Protein quantitative structure property relationships for improved chromatographic separation. Netherlands Process technology Symposium (NPS), Delft, The Netherlands, April 2022, poster presentation.

Neijenhuis, T., Pabst, M., Klijn, M. E., & Ottens, M., Protein quantitative structure property relationships for improved chromatographic separation. American Chemical Society (ACS), Online, March 2022, Flash presentation.