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

Disela, Roxana; Keulen, Daphne; Fotou, Eleni; Neijenhuis, Tim; Le Bussy, Olivier; Geldhof, Geoffroy; Pabst, Martin; Ottens, Marcel

DOI
10.1002/btpr.3494

Publication date
2024

Document Version
Final published version

Published in
Biotechnology Progress

Received: 12 February 2024 | Revised: 30 May 2024 | Accepted: 1 July 2024

DOI: 10.1002/btpr.3494

RESEARCH ARTICLE

Bioseparations and Downstream Processing

BIOTECHNOLOGY PROGRESS

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

Roxana Disela[1] | Daphne Keulen[1] | Eleni Fotou[1] | Tim Neijenhuis[1] | Olivier Le Bussy[2] | Geoffroy Geldhof[2] | Martin Pabst[1] | Marcel Ottens[1]

[1]Department of Biotechnology, Delft University of Technology, Delft, The Netherlands

[2]GSK, Technical Research & Development, Rue de l'Institut 89, Rixensart, Belgium

Correspondence
Marcel Ottens, Department of Biotechnology, Delft University of Technology, Van der Maasweg 9, 2629 Delft, The Netherlands.
Email: m.ottens@tudelft.nl

Funding information
GlaxoSmithKline Biologicals S.A.

Abstract

Mechanistic models mostly focus on the target protein and some selected process- or product-related impurities. For a better process understanding, however, it is advantageous to describe also reoccurring host cell protein impurities. Within the purification of biopharmaceuticals, the binding of host cell proteins to a chromatographic resin is far from being described comprehensively. For a broader coverage of the binding characteristics, large-scale proteomic data and systems level knowledge on protein interactions are key. However, a method for determining binding parameters of the entire host cell proteome to selected chromatography resins is still lacking. In this work, we have developed a method to determine binding parameters of all detected individual host cell proteins in an *Escherichia coli* harvest sample from large-scale proteomics experiments. The developed method was demonstrated to model abundant and problematic proteins, which are crucial impurities to be removed. For these 15 proteins covering varying concentration ranges, the model predicts the independently measured retention time during the validation gradient well. Finally, we optimized the anion exchange chromatography capture step in silico using the determined isotherm parameters of the persistent host cell protein contaminants. From these results, strategies can be developed to separate abundant and problematic impurities from the target antigen.

KEYWORDS

downstream process development, *E. coli* BLR, host cell proteomics, ion exchange chromatography, isotherm parameter determination, mechanistic modeling, vaccine purification

1 | INTRODUCTION

Host cell protein (HCP) impurities, if present in final drug product, can pose risks to product stability and patient safety. These impurities are released together with DNA, RNA, and endotoxins when host cells are disrupted to obtain intracellular recombinant protein products. Compared to medications for chronic diseases, where HCP levels are typically kept below 100 ppm, vaccines allow for higher levels of tolerated HCPs.[1] Regulatory authorities determine acceptable levels of HCPs for vaccines on a case-by-case basis.[1] For instance, in the context of a malaria vaccine candidate produced in *Escherichia coli* and intended for administration at 80 μg of a protein antigen per dose,

Zhu et al.[2] proposed a limit of 1 μg/dose for every single HCP impurity. Tscheliessnig et al.[3] specified that the total HCP concentration should be 90 ng or <1100 ppm per dose in this particular case. Developing effective purification sequences to remove HCPs from diverse products often relying on expert knowledge or trial-and-error, emphasizes the crucial need for new, rational, and broadly applicable process development strategies.[4] To gain a higher level of process understanding, mechanistic models (MM) are employed in process development.[5–7] MMs describe the underlying physical phenomena during a chromatographic process by incorporating mass transfer correlations and binding kinetics. The binding kinetics are described by adsorption isotherm parameters, valid under the investigated conditions. A key challenge in applying these approaches to real purification problems is finding experimental techniques that are able to determine the necessary isotherm parameters for individual HCPs. This study aims to develop a method to determine isotherm parameters of the individual HCP impurities by coupling linear gradient experiments (LGE) with proteomic analysis. The developed method is applied to determine isotherm parameters of all detected HCPs present in an *E. coli* lysate and optimize a process step to separate an antigen from the HCP impurities.

Mass spectrometry (MS) is an increasingly popular analytical method for HCP analysis, allowing the identification of thousands of proteins within a biological sample.[8–10] Extensive research and development efforts have focused on the identification as well as the effective removal of HCP impurities from different hosts.[8,11–13] Specific Chinese hamster ovary cell (CHO) proteins, co-eluting with monoclonal antibodies (mAbs), are referred to as persistent HCP or post-protein A proteins.[14] The interaction between the product and production cell enzymes during cell disruption or enzyme release from dying cells is a potential source of significant damage to the intended native configuration.[15] This damage can lead to irreversible aggregation of the product, substantially reducing the overall yield and raising concerns like immunogenicity, as demonstrated in recent findings indicating HCP involvement in product aggregation.[16–18] Similarly, product stability can be impacted by low abundance HCPs such as host cell lipases able to degrade excipients Polysorbate 20 or Polysorbate 80.[19]

However, for products like vaccine antigens produced in *E. coli*, no general persistent proteins are known. *E. coli* lysates, characterized in previous work,[13] constitute a complex mixture of approximately 2000 detected proteins with diverse physicochemical properties (out of approximately 4300 possible gene products in *E. coli*). Especially proteolytic digestion poses a challenge when working with *E. coli* as a host.[20]

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

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

However, these studies employed chromatography as an analytical method. Wierling et al. approached the determination of HCP impurities from CHO cells during the purification of a mAb by combining high-throughput screening with mass spectrometric detection.[27] MS enables the detection of all individual HCPs down to 5 ppm.[28] Compared with anti-HCP enzyme-linked immunosorbent assays (ELISAs), that detect proteins against which they were developed, MS provides information on individual proteins present in the drug substance or product.[3]

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

2 | THEORY/CALCULATION

2.1 | Mass balance in mechanistic model

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

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

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

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

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

2.2 | Mass transfer in mechanistic model

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

$$\frac{\partial q_i}{\partial t}=k_{ov,i}\left(C_i-C^*_{eq,i}\right), \tag{3}$$

where equilibrium concentration in the bulk phase C_i^* is determined by the isotherm. The overall mass transfer coefficient $k_{ov,i}$ is defined as a summation of the separate film mass transfer resistance and the mass transfer resistance within the pores. Details of the mass transfer are described in Appendix A.1 (Table A1).

2.3 | Derivation of regression formula

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

$$A_i=K_{eq,i}\Lambda^{\nu_i}(z_s c_s)^{-\nu_i}, \tag{4}$$

where $K_{eq,i}$ is the equilibrium constant per protein, Λ is the ionic capacity of the resin skeleton, z_s is the charge on the salt counter-ion, c_s is the salt concentration and ν_i is the characteristic charge of the protein. The characteristic charge is described as $\nu_i=z_p/z_s$, where z_p is the effective binding charge of the protein. In this study, we set $z_s=1$ since the experiments are conducted using sodium chloride, which means that $z_p=\nu_i$. The protein specific constants $K_{eq,i}$ and ν_i are furthermore called isotherm parameters.

$$k'=\frac{V_{R,iso,i}-V_M}{V_M}=K_i c_s^{-m_i}=FK_{eq,i}\Lambda^{\nu_i}(z_s c_s)^{-\nu_i}. \tag{5}$$

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

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

This formula can be written in logarithmic form as.

$$log(k')=log(K_i)+m_i\,log(1/c_s)=log\left(FK_{eq,i}\Lambda^{\nu_i}\right)-\nu_i\,log(c_s). \tag{6}$$

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

$$K_i=FK_{eq,i}\Lambda^{\nu_i}, \tag{7}$$

$$m_i=\nu_i. \tag{8}$$

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

$$V_{R,g,i}=\left(\left({c_{s,0}}^{m_i+1}+\frac{V_m K(m_i+1)(c_{s,f}-c_{s,0})}{V_G}\right)^{1/(m_i+1)}-c_{s,0}\right)*\frac{V_G}{(c_{s,f}-c_{s,0})}, \tag{9}$$

where the gradient retention volume $V_{R,g,i}$ of a protein during gradient elution is described using additionally the initial and final salt concentration $c_{s,0}$ and $c_{s,f}$, and the length of the salt gradient V_G. When varying the gradient volume experimentally, this formula can be employed to regress K_i and m_i of the analyzed protein. Using Equation (7) and Equation (8), Equation (9) can be rewritten as

$$V_{R,g,i}=\left(\left({c_{s,0}}^{\nu_i+1}+\frac{V_m FK_{eq}\Lambda^{\nu_i}(\nu+1)(c_{s,f}-c_{s,0})}{V_G}\right)^{1/(\nu_i+1)}-c_{s,0}\right)*\frac{V_G}{(c_{s,f}-c_{s,0})} \tag{10}$$

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

3 | MATERIALS AND METHODS

3.1 | General method

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

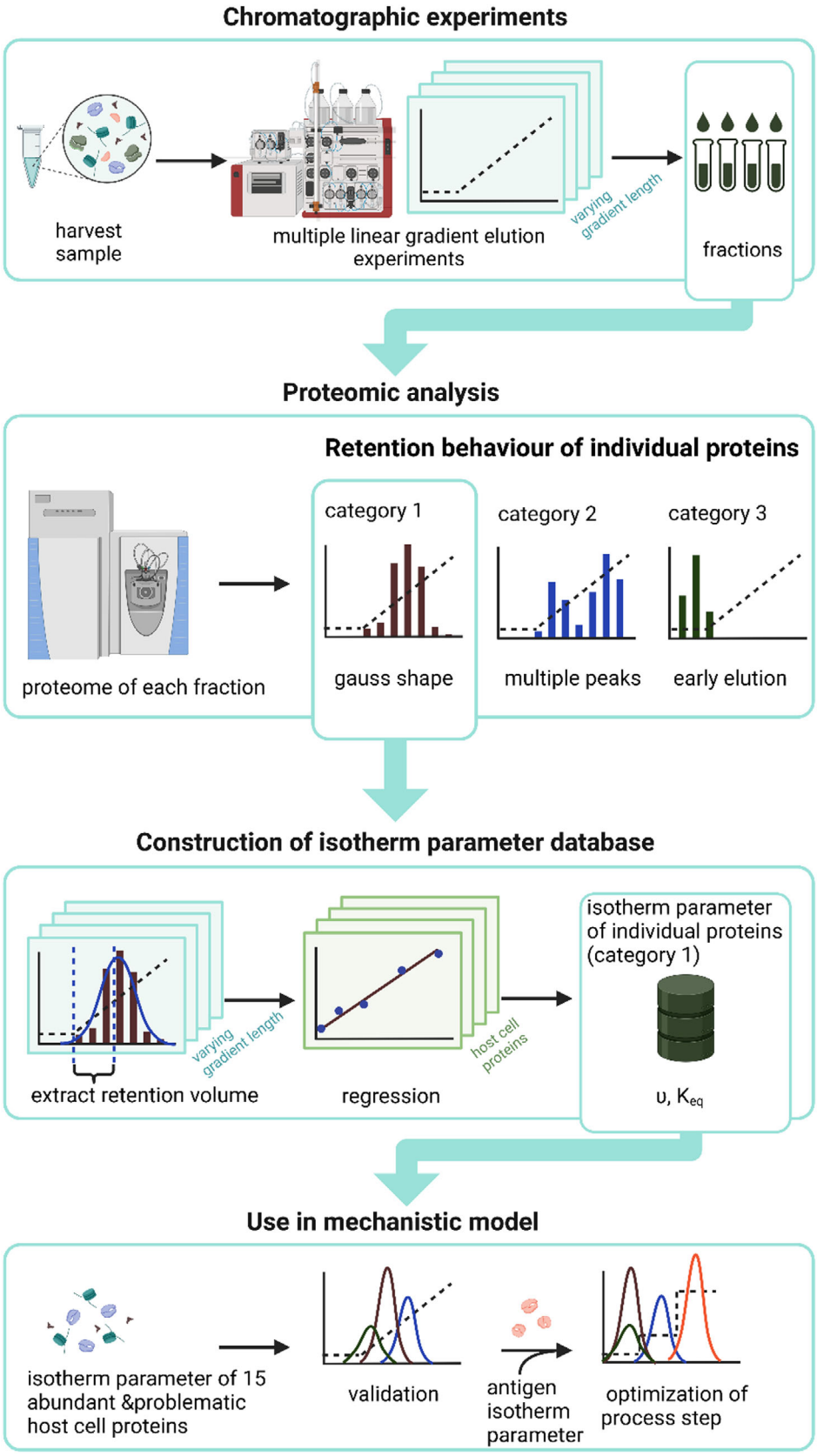


FIGURE 1 Schematic overview of applied method in this study. Chromatographic experiments are conducted using the harvest sample containing a mixture of host cell proteins. The protein mixture is injected to the Äkta chromatography system and linear gradient experiments with varying gradient lengths are conducted. From each of the gradient runs, fractions are taken and their proteome is analyzed via mass spectrometry. The majority of proteins, that show Gaussian function behavior, are used to build the isotherm parameter database. Their retention volumes during the varying gradient lengths are extracted and regressed using formula (12). The fitted isotherm parameters for every individual protein are saved in the database. From this database, 15 critical host cell proteins were chosen for validation of the mechanistic model. The validated model was used together with the antigen isotherm data to optimize a capture step removing the 15 host cell protein impurities from the target antigen (Illustration created using BioRender.com).

volumes during LGE experiments with different gradient lengths were extracted and used in a regression to determine the individual isotherm parameters. For 15 selected HCPs, the isotherm parameters were validated in a MM. The model was furthermore used to optimize a chromatography step separating the antigen from the selected HCPs.

3.2 | Chromatographic experiments

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

3.2.1 | *E. coli* fermentation and harvest sample

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

3.2.2 | Materials and apparatus for chromatographic experiments

The chromatographic experiments were performed on an Äkta pure with a connected fraction collector F9-C from Cytiva (Uppsala, Sweden). The dwell volume of the Äkta system, describing the delay between the gradient initiation and the change in the mobile phase composition at the column inlet, was determined as 1.1 mL in a separate experiment described in Appendix A.2.1. A prepacked HiTrap Q Sepharose XL 5 mL column from Cytiva (Uppsala, Sweden) was used for chromatographic experiments. The ionic capacity of the resin skeleton was measured by displacement experiments using HCl titration (Appendix A.2.2). It was determined to be 1.106 mmol/L. The running buffer for all experiments was 0.02 M Tris at pH 7.0 with 0.02 M NaCl added. The high salt buffer consisted of the same buffer components with 1 M NaCl added. Between experimental runs the chromatography column was cleaned using 1 M NaOH solution. All buffers were filtered with 0.22 µm pore size and sonicated before use.

3.2.3 | Linear gradient elution experiments

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

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

3.3 | Proteomic analysis

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

3.3.1 | Shotgun host cell proteomics

Before the MS analysis, the samples were prepared using the filter-aided sample preparation (FASP) developed to simplify the preparation of samples.[36,37] The applied method is further described in the Appendix A.1.

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

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

The Orbitrap mass spectrometer operated in data-dependent acquisition (DDA) mode, capturing spectra at a resolution of 70,000 over the m/z range of 385–1150. The top 10 signals were selected for isolation with a window of 2.0 m/z and an isolation offset of 0.1 m/z, followed by fragmentation employing a normalized collision energy (NCE) of 28. Fragmentation spectra were acquired at a resolution of 17,000, with an automatic gain control (AGC) target of 5e5 and a maximum injection time (IT) of 75 ms. Unassigned, singly

charged, and ions with six or more charges were excluded from fragmentation. Dynamic exclusion was set to 60 s.

3.3.2 | Processing of mass spectrometric raw data

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

To enhance the analysis, strain-specific proteome sequence databases were obtained from NCBI (BioProject PRJNA379778), and sequences of contaminant proteins were sourced from the global proteome machine (GPM) database (https://www.thegpm.org/crap/). A decoy fusion strategy was employed to estimate false discovery rates (FDRs). The filtering of peptide spectrum matches was carried out with a threshold of 1% FDR, and proteins with more than one unique peptide sequence were considered statistically significant.

To assess changes in protein abundance between the different fractions, label-free quantification was performed using the PEAKSQ module (den Ridder, Daran-Lapujade, & Pabst, 2020). The abundance measure utilized in this analysis was the peak area obtained from the reversed-phase column prior to entering the mass spectrometer. Exclusively proteins that were identified with more than peptides were used in the further analysis.

3.3.3 | Processing of retention behavior of individual host cell proteins

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

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

3.4 | Construction of isotherm parameter database

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

3.4.1 | Processing of retention volumes

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

$$V_{R,g,i} = V_{R,g,i,raw} - 0.5\,V_{inj} - V_{dwell} - V_m - V_{wash}, \quad (11)$$

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

3.4.2 | Regression of host cell protein isotherm parameters

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

3.4.3 | Determination of antigen isotherm parameters

The isotherm parameters of the antigen (and the charge variant) were determined in a similar manner. LGE experiments with various gradient lengths (5, 10, 20, 40, and 60 CV) were conducted using purified antigen. The maximum of the main peak was extracted using the signal obtained from the UV spectrometer at 230 nm wavelength instead of employing MS. This value was used as the raw data retention volume of the antigen, while an earlier eluting smaller peak was identified to be a charge variant. The corrected retention volumes were obtained with Equation (11), and used to regress the isotherm parameters utilizing Equation (10). Antigen isotherm parameters are then

TABLE 1 Fifteen abundant and problematic host cell proteins in focus of this study to be removed from the target antigen.

Protein type	Protein name	Protein ID	Relative concentration [% of area under Gaussian curve]	$K_{eq,i}$	SD $K_{eq,i}$	v_i	SD v_i	R^2	RSME	Volume difference correlation to experiment [mL]	Volume difference model to experiment [mL]
Abundant	Translation elongation factor EF-Tu 1&2	ARH99640.1	6.794	0.142	0.077	2.641	0.414	0.981	3.395	−1.484	−0.418
	Protein chain elongation factor EF-G GTP-binding	ARH98956.1	2.753	0.276	0.212	2.274	0.605	0.942	6.378	−2.633	−1.465
	30S ribosomal protein S1	ARH96687.1	1.701	0.295	0.165	2.346	0.458	0.967	4.843	−2.717	−1.547
	Glyceraldehyde-3-phosphate dehydrogenase A	ARH97514.1	1.607	0.009	0.007	2.620	0.363	0.977	1.535	−0.431	0.690
	Isocitrate dehydrogenase (NADP(+))	ARH96911.1	1.543	0.007	0.005	3.664	0.501	0.989	1.825	−0.612	0.424
Chaperone	Cpn60 chaperonin GroEL large subunit of GroESL	ARH99809.1	0.511	0.245	0.171	2.982	0.666	0.974	5.506	−4.791	−3.729
	Molecular chaperone DnaK	ARH95794.1	0.543	0.520	0.284	2.063	0.468	0.944	6.755	−3.737	−2.544
	Cpn10 chaperonin GroES small subunit of GroESL	ARH99808.1	0.223	0.003	0.004	4.569	0.965	0.984	2.786	−0.639	0.464
	Protein disaggregation chaperone	ARH98235.1	0.216	0.003	0.005	4.777	1.191	0.980	3.294	−0.809	0.255
	Fe/S biogenesis protein putative scaffold/chaperone protein	ARH99034.1	0.112	0.250	0.102	2.623	0.352	0.982	3.746	−1.723	−0.571
Protease	Carboxy-terminal protease for penicillin-binding protein 3	ARH97573.1	0.035	0.010	0.018	3.150	0.999	0.966	2.922	−0.601	0.478
	ATP-dependent Clp protease proteolytic subunit ClpP	ARH96177.1	0.084	0.259	0.196	1.996	0.531	0.919	6.316	−3.062	−1.844
	ATP-dependent Clp protease ATP-binding subunit ClpX	ARH96178.1	0.025	0.201	0.180	2.529	0.712	0.950	6.124	−2.190	−1.130
	Modulator for HflB protease specific for phage lambda cII repressor	ARH99838.1	0.003	0.001	0.002	8.049	3.729	0.976	5.834	−1.231	−0.127
	Molecular chaperone and ATPase component of HslUV protease	ARH99598.1	0.008	0.646	0.474	1.518	0.538	0.844	9.387	−4.598	−3.078

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

3.5 | Validation of host cell protein isotherm parameters in mechanistic model

For the validation of the HCP isotherm parameter in the MM, 15 proteins were selected and their retention behavior was modeled for the left out 10 CV gradient experiment. For the 15 proteins, the modeled retention volumes and elution peak shapes were compared with the experimentally determined data.

As an input for the MM, a relative protein concentration was used (listed in Table 1). These concentrations were obtained from integration of the Gaussian functions that were fitted to the experimental data (of the 20 CV gradient). These values are given in percent of the peak area of the Gaussian function from each individual protein in relation to the total of all the proteins. The relative antigen concentration was calculated from the measured relation of the antigen to the total of all the proteins.

3.6 | Optimization of chromatography step in mechanistic model

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

$$minf(x) = 0.5 * (100 - yield(x)) + 0.4 * (100 - purity(x)) + 0.1 * buffer\ consumption(x) \quad (12)$$

$$s.t.\ \ h(x) = 0 \quad (13)$$

$$0 \le x \le 1, \quad (14)$$

where the objective is to minimize function f, in which the variables x were normalized between 0 and 1 for enhanced optimization purposes (Equation 14). Moreover, it is important to satisfy the mass balances and equilibrium relations as denoted in Equation (13). A total of six variables were optimized: the salt concentration for the first and the second step,

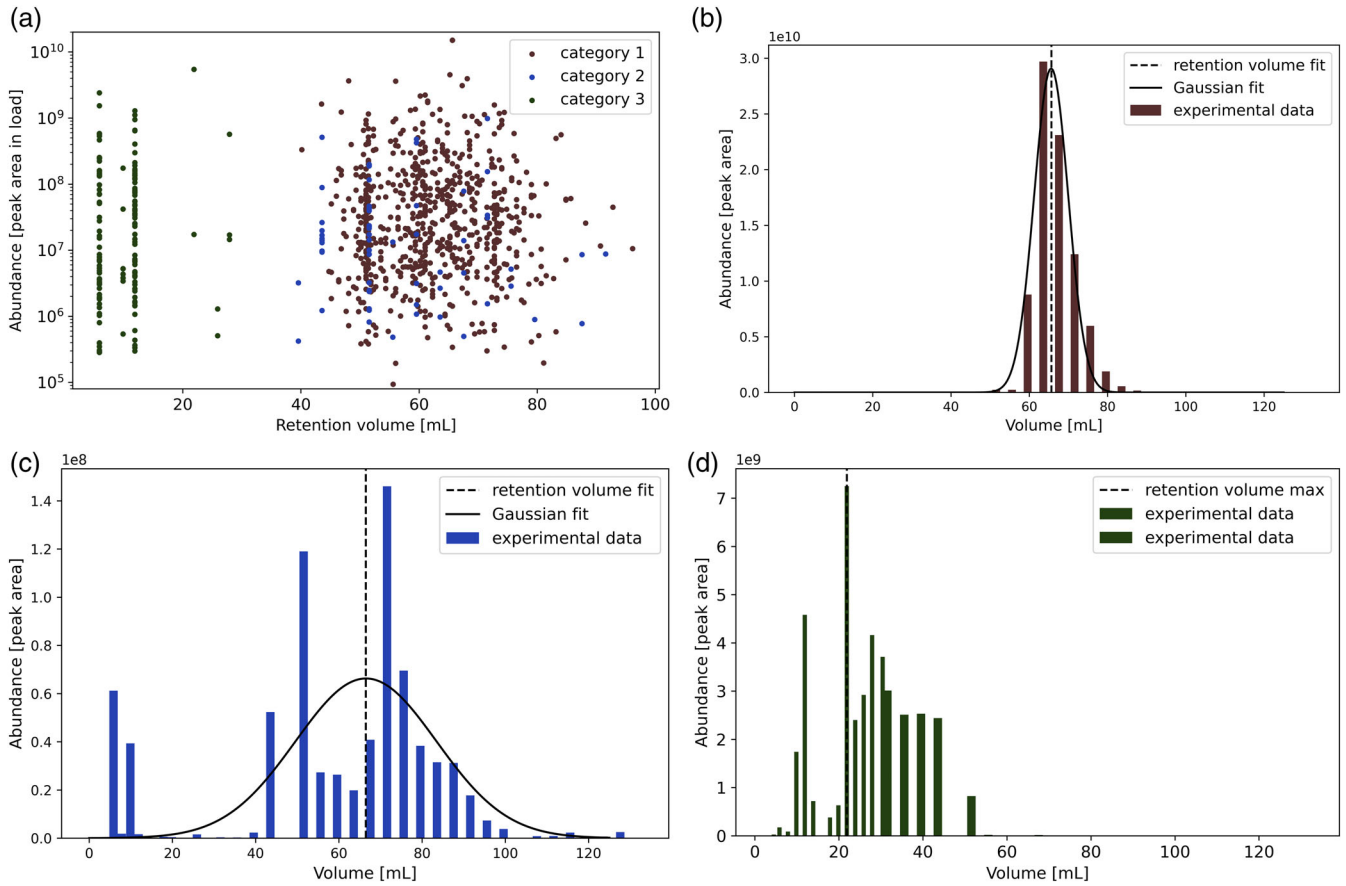
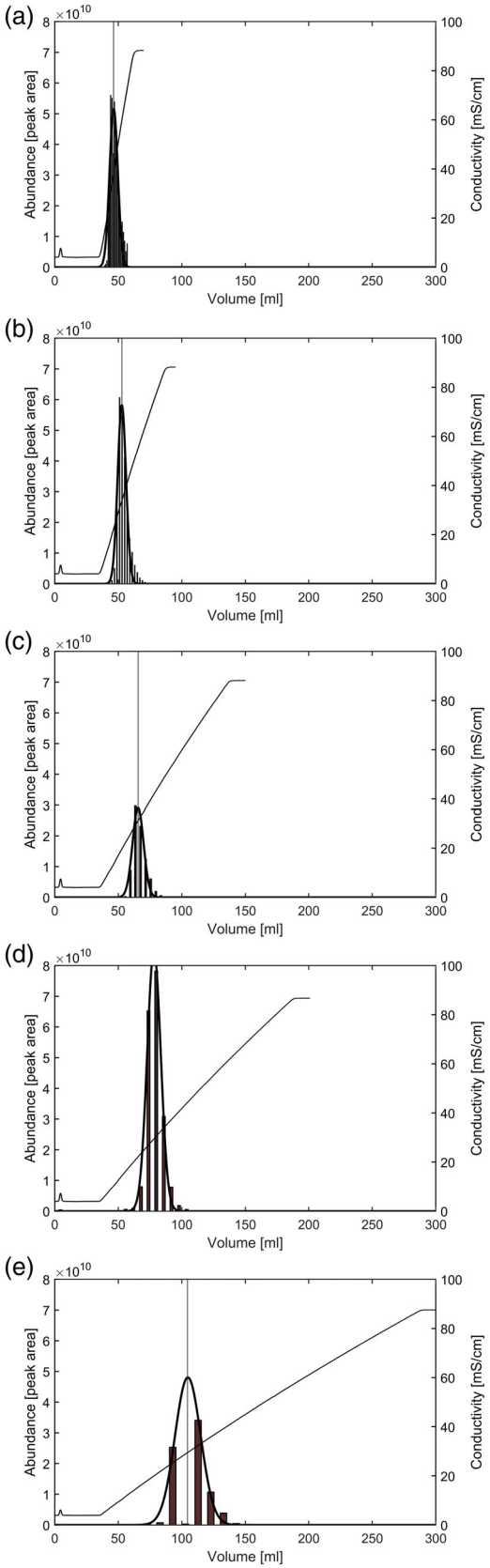


FIGURE 2 Categories of host cell proteins: (a) scatterplot showing the retention behavior of the three categories of protein during a 100 mL (20 column volume) gradient on a 5 mL HiTrap Q Sepharose XL column, (b) example for Category 1: Single peak Gaussian function (R^2 >0.7, here 0. 97) of "translation elongation factor EF-Tu 1&2" (ARH99640.1), (c) example for Category 2: Multiple peaks eluting (R^2 <0.7, here 0.41) in case of "30S ribosomal subunit protein S3" (ARH98930.1), (d) example of Category 3: Protein eluting before the gradient when observing "RNA chaperone and antiterminator cold-inducible" (ARH99188.1).

the gradient lengths for both steps, and the lower and upper cut points. The main objective for a capture step is obtaining a high yield, followed by the purity, and a low buffer consumption. The buffer consumption indirectly reflects the costs, batch throughput, and productivity, as it minimizes the time needed to perform this purification step.



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

4 | RESULTS AND DISCUSSION

4.1 | Retention behavior of individual host cell proteins

A total of 1247 *E. coli* HCPs were identified via MS throughout all fractions in the 20 CV gradient run. The retention behavior of individual HCPs during the gradient elution was classified into three categories (Sections 3.2.3 and 3.3). Most proteins fall into Category 1 (898 proteins; 72%), which shows a single peak elution during the salt gradient and therefore can be fitted well with a Gaussian function (Figure 2b). However, 121 proteins (10%) falling into Category 2 showed multiple peaks or abundance in only a single fraction during the elution, so it was not possible to fit a Gaussian function (Figure 2c). Two hundred and fifteen proteins (17%), falling into Category 3, had their abundance maximum during the wash before the start of the salt gradient (Figure 2d). The remaining 13 proteins were detected in the sample but were below the limit of quantification. The abundance of the detected proteins in the load sample is shown in "peak area," as measured by MS. It is plotted over the retention volume of the identified proteins in Figure 3a. Hereby, the maximum of the Gaussian function was considered the retention volume for proteins of Category 1, while the maximum abundance value was used for proteins of categories 2 and 3.

FIGURE 3 All five linear gradient elution experiments for "translation elongation factor EF-Tu 1&2" (ARH99640.1): (a) 25 mL (5 column volumes) gradient, (b) 50 mL (10 column volumes) gradient, (c) 100 mL (20 column volumes) gradient, (d) 150 mL (30 column volumes) gradient, (e) 250 mL (50 column volumes) gradient; The abundance in the fractions was determined using MS and displayed as peak area. Gaussian functions are fitted to the abundance data. The maxima of the fitted Gaussian functions are extracted as retention volumes and used in the isotherm parameter regression. Only the 50 mL gradient is left out for validation of the mechanistic model.

The ideal elution behavior seen by proteins of Category 1 makes it possible to extract retention volumes of the different gradients with confidence, illustrated in Figure 3 for the most abundant protein "translation elongation factor EF-Tu 1&2." Here, small differences in absolute protein concentrations between experiments could be caused by fluctuations due to the dialysis. Especially for the 5 CV gradient, values close to saturation were observed, therefore it is advised to dilute the sample more, or to apply longer salt gradients. The retention volumes can further be used to extract isotherm parameters and mechanistically model the protein behavior on the tested column and conditions. This was possible for 721 proteins, for which Gaussian functions could be fitted with sufficient accuracy in all five gradients. The proteins from Category 2 can be determined with less confidence, as the protein abundance is very low or the protein shows different isoforms. Different isoforms can be caused by charge variants or the formation of complexes with other protein species. Elution before the start of the gradient described by proteins in Category 3 could have several reasons. The proteins could simply have no affinity to the anion exchange resin because of a positive net-charge. Another reason can be that some proteins eluting during the first fractions might be digested by proteases in the harvest sample. Therefore, these proteins were at the time of the LGE gradient experiments only present as peptides. Peptides would more likely elute directly in the first fractions since less interaction with the resin is expected. No size differences were observed between the different protein categories and therefore size exclusion effects of proteins of Category 3 can be disregarded.

4.2 | Selection of abundant and problematic proteins

While this big data lake of isotherm data is very insightful, it is not practical (and not necessary) to model every single protein as the MM would take days to perform one simulation, even more so a whole optimization that requires about 500 simulations. Therefore, we made a selection of proteins that are in our interest to be simulated in this study. Since the aim of our study is to find the optimal process to purify an antigen from HCP impurities, we choose the most relevant proteins to be removed in a capture step. As described in the introduction, abundant proteins, proteases, and chaperones are of high priority to be removed early on in the process. Hence, we choose to select the five most abundant proteins, the five most abundant proteases, and the five most abundant chaperones present in the dataset. These 15 HCPs together with their retention behavior, properties, and determined isotherm parameters are listed in Table 1.

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

4.3 | Isotherm parameter regression of individual host cell proteins

The retention volumes of the individual HCPs and the value of total gradient elution volume are related to each other via the formula from

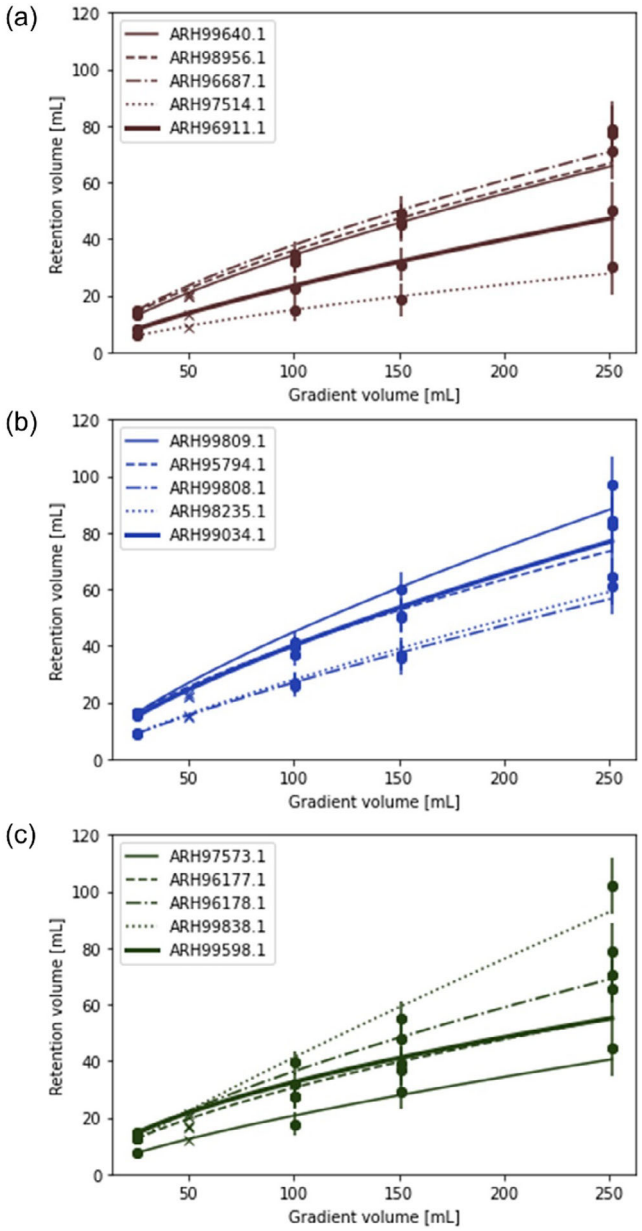


FIGURE 4 Regression of (a) five most abundant proteins, (b) five most abundant chaperones, (c) five most abundant proteases in harvest sample; the dots are the experimentally measured values with error bars according to the fractionation scheme. The lines connecting the dots show the regressed fit. Full protein names are listed in Table 1. The value for the 50 mL gradient run, marked with an x, was used as a validation run and not included in the fit.

Parente and Wetlaufer as given in Equation (10) and shown in Figure 4. Experimentally determined retention volumes (shown as dots) are fitted with the given formula (shown as lines) and compared to the 10 CV (50 mL) gradient that was left out for validation (shown as x). Fractionation size and frequency was considered to determine the experimental error (plotted error bars). The regressed values and their standard deviation obtained for the 15 selected abundant and problematic HCPs are added to Table 1.

The weighted regression (Figure 4) leads to a slight upwards bend in the fitted functions. The bend leads to a small overestimation of the shorter gradient lengths and slight underestimation of the longer gradient lengths compared to the experimental values. Although, these differences are still within the experimental error and the fitted function describes the data well. Nonweighted regression was also investigated, however, this provided $K_{eq,i}$ values close to the set boundaries with very high standard deviations caused by improper scaling of the data.

An overview of the 721 values obtained for the isotherm parameters $K_{eq,i}$, ν_i, and their standard deviations are given in form of boxplots in Figure 5a,b. The values obtained for R^2 and RSME are shown in boxplots in Figure 5c,d. Determined $K_{eq,i}$ values were between 0.0001 and 5.43. The standard deviation of this parameter was determined between 0.00015 and 1.46. Parameters determined for the characteristic charge varied between 0.47 and 13.93. For none of the proteins the regressed values were exactly at the given boundaries of $K_{eq,i}$ (0.00001 and 100) and ν_i (0.1 and 15). The standard errors of the regressed parameters are on average 116% of the parameters nominal value for $K_{eq,i}$ and 21% for ν_i. These relative high standard deviations for especially $K_{eq,i}$ might be caused by the relatively low absolute values. However, the R^2 varied between 0.47 and 1.00 with an average of 0.97, meaning that the fit with the regression formula described the experimental data well for the majority of the proteins. Likewise, the RSME values, varying between 0.11 and 41.12 mL with an average of 3.35 mL (6.7% in a 50 mL gradient), indicate that the fitted regression formula describes the data well. More importantly, the differences for all HCPs in retention volume between the left-out validation run and calculation from the correlation are low

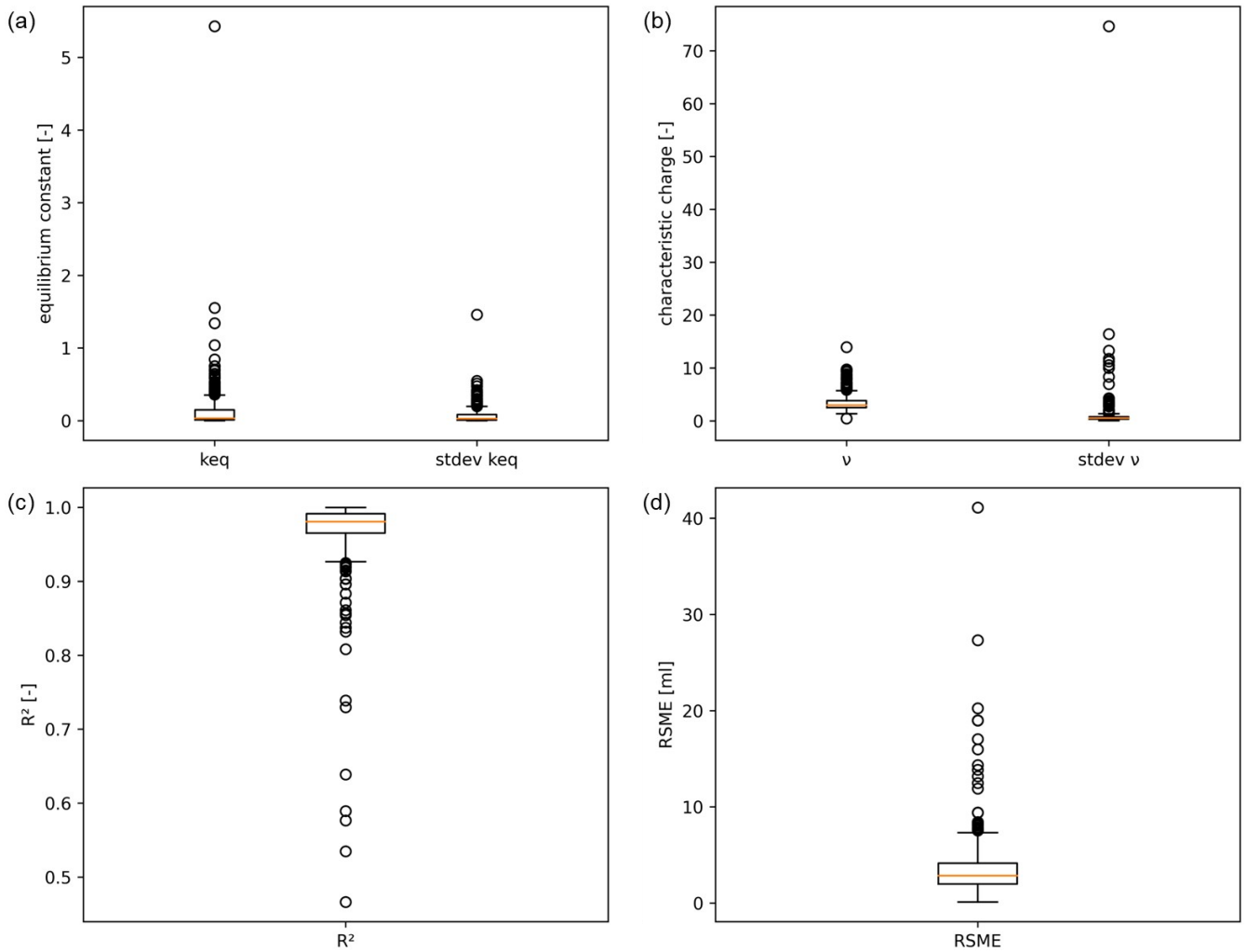


FIGURE 5 Overview of the regressed isotherm parameters of the complete host cell protein dataset. (a) equilibrium constant $K_{eq,i}$ and standard deviation of all Host cell proteins (HCPs), (b) characteristic charge ν_i and standard deviation of all HCPs, (c) R^2 of all HCPs, and (d) RSME of all HCPs.

with an average of 1.23 mL (2.5% in a 50 mL gradient) and a maximum value of 8.21 mL (16.4% in a 50 mL gradient). Based on these results, we conclude that the regression function with the fitted isotherm parameters can describe the experimental data with high accuracy.

4.4 | Validation in mechanistic model

The average pore size of the Q Sepharose XL resin is described as 54 nm for the agarose skeleton[38] and 12 nm[39] including the dextran-graft that bind the ligands. Using a pore diameter of 12 nm for the resin in the MM lead to size exclusion effects and hence an early elution of HCPs. However, from the experimental data, it was concluded that no such size exclusion effects occurred for the HCPs. Hence a pore diameter of 54 nm was used in the MM assuming the flexible dextran-grafts inside the pores do not hinder the access of the HCPs into the pores.

For the validation, isotherm parameters of the 15 selected abundant and problematic HCPs, listed in Table 1, were used in the MM to simulate the left out 10 CV (50 mL) gradient run. These simulations were compared with the experimental result. In addition, the obtained isotherm parameters of the antigen and its charge variant were simulated together with the 15 HCPs to compare their retention behavior.

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

As explained previously in Section 3.4.1 the experimental data was fitted with a Gaussian curve. Figure 6 shows a side by side comparison of the experimental Gaussian curves (Figure 6a,b) and modeled curves (Figure 6c,d). Thereby Figure 6a,c shows the extended view, while Figure 6b,d is zoomed in to show low abundance peaks. Overall, similar peak shapes can be observed, despite their different abundance measures.

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

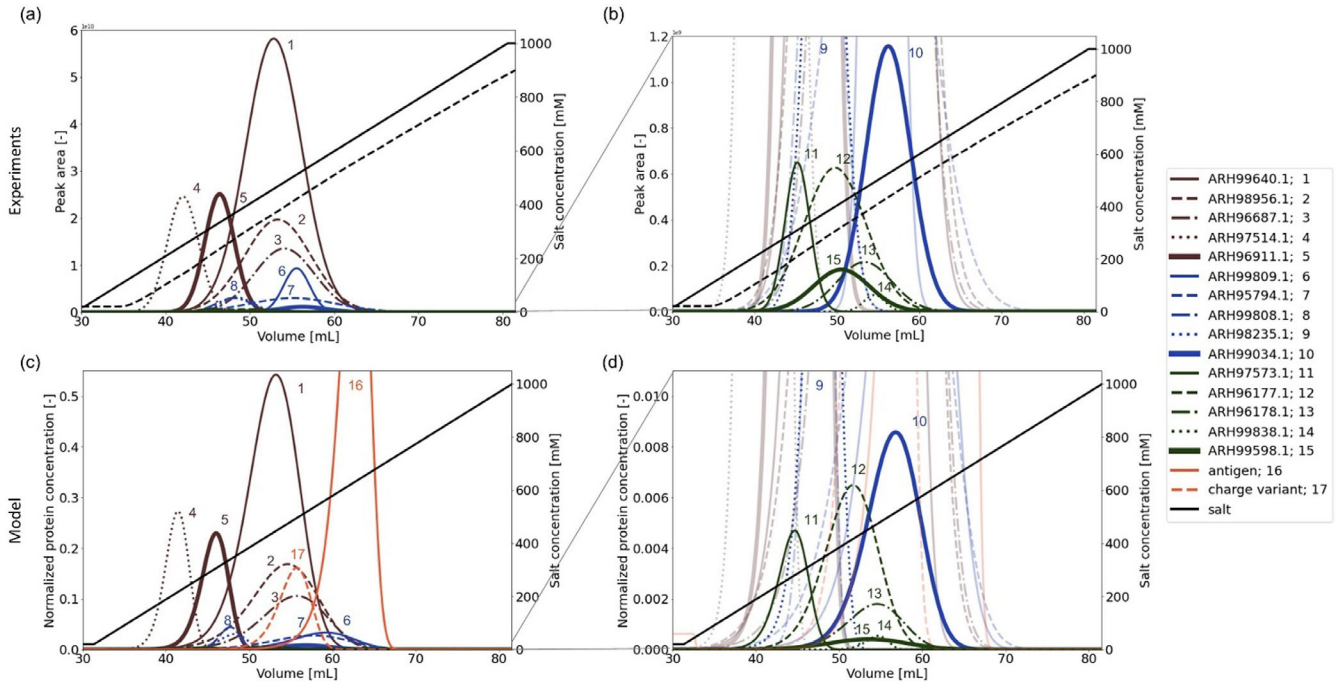


FIGURE 6 Comparison of experimentally determined and modeled retention behavior during the validation experiment of 15 abundant and problematic Host cell proteins. In the graphs, the retention during a 50 mL (10 column volume) gradient with a 5 mL Q Sepharose XL column is shown. Gaussian functions fit to the experimental raw data in peak area are shown in: (a) full view, (b) zoom in. Modeled elution of the components described by (a) the normalized protein concentration is shown in: (c) full view, (d) zoom in.

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

4.5 | Optimization of capture step

The validated model was used to find the optimal capture step conditions on the 5 mL HiTrap Q Sepharose XL column to separate the antigen from the HCPs of interest with focus on the yield, purity, and buffer consumption. In Figure 7, a chromatogram of the chosen optimized step elution is shown using 20 mM Tris buffer. First, the column is washed with 20 mM NaCl as a running buffer after the 1 mL injection of the load. The majority of HCPs are removed during the 362 mM NaCl wash (9.43 CV). Finally, the antigen elutes during the 634 mM NaCl step lasting 2.75 CV. The collected eluate, highlighted in white, has a yield of 98% and purity of 99%.

Impurities that co-elute with the antigen obtained in the product pool are discussed below in descending abundance. Cpn 60 chaperonin is expected to co-elute partially with the antigen. Since the $K_{eq,i}$ was slightly overestimated in the model compared to the experiment as discussed previously (Section 4.4), less Cpn 60 chaperonin might co-elute in an experiment with the antigen than the model calculates. However, this protein is often detected as an impurity in the final drug product due to strong binding affinity to all proteins. Cpn

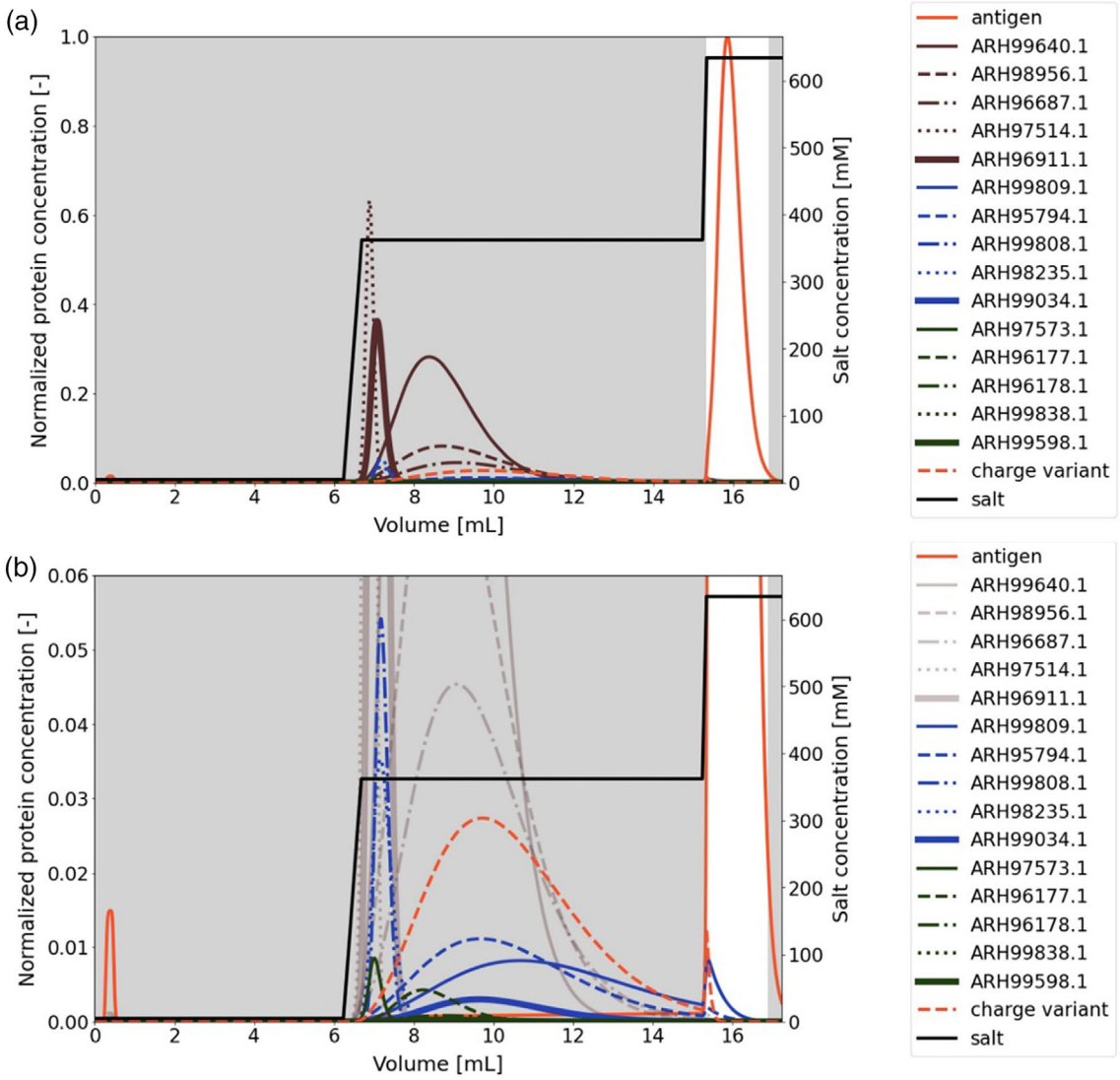


FIGURE 7 Model of optimized capture step for antigen on 5 mL HiTrap Q Sepharose XL column considering 15 selected host cell proteins to be removed (full names in Table 1). First, the column is washed with 20 mM NaCl as a running buffer after the 1 mL injection of the load then with the 362 mM NaCl wash (9.43 CV). Finally, the antigen elutes during the 634 mM NaCl step lasting 2.75 CV. The collected eluate, highlighted in white, has a yield of 98% and purity of 99%. (a) Full view, (b) zoom in.

60 chaperonin has been identified to be very immunogenic and has a high priority to be removed as early as possible in the process.[21] In this case, the model predicts the worst case and this makes the actual process step more robust. Another protein, that co-elutes partially with the antigen, is its charge variant. The majority of the charge variant of the antigen is removed in the optimized capture step. However, the remaining charge variant could be removed in a consecutive orthogonal polishing step, if required. "molecular chaperone DnaK" (ARH95794.1) also co-elutes with the antigen. It shows a later retention in the model compared with the experiment and might be removed even more effective in reality, if it does not bind to the antigen itself. In literature,[40] it was shown that this protein shows a high immunogenicity in mice. Here, the model shows the worst-case scenario and therefore finds a robust optimal process. We show that it is possible to use the validated model to find the optimal process step. This optimized step is used to separate the target antigen from the charge variant, and 15 abundant and problematic proteins.

5 | CONCLUSIONS AND OUTLOOK

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

In the future, this method might be applicable to design a new capture step for an unseen/new protein produced in *E. coli*. In this case, isotherm parameters of the new protein and product-related impurities are required. The existing database can also be used to describe other *E. coli* strains since abundances and protein concentrations are comparable for different strains.[13] In principle, the method can also be applied to other hosts such as *pichia pistoris* with a similar number of possible gen products. The number of proteins expressed by CHO cells might lead to increased analysis times and efforts and requires some more attention on product-related impurities. Since hitchhiker proteins involved in aggregates pose a challenge for example for mAbs produced in CHO,[17] the authors would suggest a joint measurement including the mAb. Present aggregate isotherms could be determined and treated as another impurity by the MM. Another approach would be to target proteins involved in protein–protein interactions[41] and use their isotherms in the MM to remove these early on in the process.

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

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

AUTHOR CONTRIBUTIONS

Roxana Disela: Conceptualization; data curation; formal analysis; investigation; methodology; software; validation; visualization; writing – original draft; writing – review and editing. **Daphne Keulen:** Conceptualization; data curation; methodology; software; validation; writing – review and editing. **Eleni Fotou:** Formal analysis; software; writing – review and editing. **Tim Neijenhuis:** Conceptualization; writing – review and editing. **Olivier Le Bussy:** Writing – review and editing; resources. **Geoffroy Geldhof:** Writing – review and editing; resources. **Martin Pabst:** Conceptualization; data curation; methodology; project administration; supervision; writing – review and editing. **Marcel Ottens:** Conceptualization; funding acquisition; project administration; resources; supervision; writing – review and editing.

ACKNOWLEDGMENTS

The authors thank Ines Ribeiro Juvandes da Silva for conducting the LGE experiments with the antigen sample. This study was funded by GlaxoSmithKline Biologicals S.A. under a cooperative research and development agreement between GlaxoSmithKline Biologicals S.A. (GSK, Belgium) and Delft University of Technology (TUD, The Netherlands). The authors are grateful to colleagues from GSK and the department of Biotechnology from TUD for valuable discussions.

CONFLICT OF INTEREST STATEMENT

All authors have declared the following interests: Geoffroy Geldhof and Olivier Le Bussy are employees of the GSK group of companies.

Geoffroy Geldhof reports ownership of shares from the GSK group of companies. Roxana Disela and Daphne Keulen report that their PhD was financed by the GSK group of companies. The other authors declare no conflict of interests.

DATA AVAILABILITY STATEMENT

Research data are not shared.

ORCID

Roxana Disela https://orcid.org/0000-0002-8178-5684
Daphne Keulen https://orcid.org/0000-0001-8086-333X
Tim Neijenhuis https://orcid.org/0000-0002-6214-5438

REFERENCES

1. Reiter K, Suzuki M, Olano LR, Narum DL. Host cell protein quantification of an optimized purification method by mass spectrometry. *J Pharm Biomed Anal*. 2019;174:650-654. doi:10.1016/j.jpba.2019.06.038
2. Zhu D, Saul AJ, Miles AP. A quantitative slot blot assay for host cell protein impurities in recombinant proteins expressed in *E. coli*. *J Immunol Methods*. 2005;306(1–2):40-50. doi:10.1016/j.jim.2005.07.021
3. Tscheliessnig AL, Konrath J, Bates R, Jungbauer A. Host cell protein analysis in therapeutic protein bioprocessing—methods and applications. *Biotechnol J*. 2013;8(6):655-670. doi:10.1002/biot.201200018
4. Hanke AT, Ottens M. Purifying biopharmaceuticals: knowledge-based chromatographic process development. *Trends Biotechnol*. 2014;32(4):210-220. doi:10.1016/j.tibtech.2014.02.001
5. Keulen D, Geldhof G, Le Bussy O, Pabst M, Ottens M. Recent advances to accelerate purification process development: a review with a focus on vaccines. *J Chromatogr A*. 2022;1676:463195. doi:10.1016/j.chroma.2022.463195
6. Close EJ, Salm JR, Bracewell DG, Sorensen E. A model based approach for identifying robust operating conditions for industrial chromatography with process variability. *Chem Eng Sci*. 2014;116:284-295. doi:10.1016/j.ces.2014.03.010
7. Gétaz D, Stroehlein G, Butté A, Morbidelli M. Model-based design of peptide chromatographic purification processes. *J Chromatogr A*. 2013;1284:69-79. doi:10.1016/j.chroma.2013.01.118
8. Bracewell DG, Francis R, Smales CM. The future of host cell protein (HCP) identification during process development and manufacturing linked to a risk-based management for their control. *Biotechnol Bioeng*. 2015;112(9):1727-1737. doi:10.1002/bit.25628
9. Traylor MJ, Bernhardt P, Tangarone BS, Varghese J. Analytical methods. In: Jagschies G, Lindskog E, Lacki K, Galliher P, eds. *Biopharmaceutical Processing*. Elsevier; 2018:1001-1049. doi:10.1016/B978-0-08-100623-8.00047-5
10. Schenauer MR, Flynn GC, Goetze AM. Identification and quantification of host cell protein impurities in biotherapeutics using mass spectrometry. *Anal Biochem*. 2012;428(2):150-157. doi:10.1016/j.ab.2012.05.018
11. Jones M, Palackal N, Wang F, et al. "High-risk" host cell proteins (HCPs): a multi-company collaborative view. *Biotechnol Bioeng*. 2021;118(8):2870-2885. doi:10.1002/bit.27808
12. Vanderlaan M, Zhu-Shimoni J, Lin S, Gunawan F, Waerner T, Van Cott KE. Experience with host cell protein impurities in biopharmaceuticals. *Biotechnol Prog*. 2018;34(4):828-837. doi:10.1002/btpr.2640
13. 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*. 2023;18(9):1-13. doi:10.1002/biot.202300068
14. Valente KN, Levy NE, Lee KH, Lenhoff AM. Applications of proteomic methods for CHO host cell protein characterization in biopharmaceutical manufacturing. *Curr Opin Biotechnol*. 2018;53:144-150. doi:10.1016/j.copbio.2018.01.004
15. Jagschies G, Łącki KM. Chapter 4 – Process capability requirements. In: Jagschies G, Lindskog E, Łącki K, Galliher P, eds. *Biopharmaceutical Processing*. Elsevier; 2018:73-94. doi:10.1016/B978-0-08-100623-8.00004-9
16. Herman CE, Min L, Choe LH, et al. Analytical characterization of host-cell-protein-rich aggregates in monoclonal antibody solutions. *Biotechnol Prog*. 2023;39(4):1-16. doi:10.1002/btpr.3343
17. Herman CE, Min L, Choe LH, et al. Behavior of host-cell-protein-rich aggregates in antibody capture and polishing chromatography. *J Chromatogr A*. 2023;1702:464081. doi:10.1016/j.chroma.2023.464081
18. Oh YH, Becker ML, Mendola KM, et al. Characterization and implications of host-cell protein aggregates in biopharmaceutical processing. *Biotechnol Bioeng*. 2023;120(4):1068-1080. doi:10.1002/bit.28325
19. Li X, Wang F, Li H, Richardson DD, Roush DJ. The measurement and control of high-risk host cell proteins for polysorbate degradation in biologics formulation. *Antib Ther*. 2022;5(1):42-54. doi:10.1093/abt/tbac002
20. Lindskog EK, Fischer S, Wenger T, Schulz P. Chapter 6 – Host cells. In: Jagschies G, Lindskog E, Łącki K, Galliher P, eds. *Biopharmaceutical Processing*. Elsevier; 2018:111-130. doi:10.1016/B978-0-08-100623-8.00006-2
21. Ranford JC, Coates ARM, Henderson B. Chaperonins are cell-signalling proteins: the unfolding biology of molecular chaperones. *Expert Rev Mol Med*. 2000;2(8):1-17. doi:10.1017/S1462399400002015
22. Nfor BK, Ahamed T, Pinkse MWH, et al. Multi-dimensional fractionation and characterization of crude protein mixtures: toward establishment of a database of protein purification process development parameters. *Biotechnol Bioeng*. 2012;109(12):3070-3083. doi:10.1002/bit.24576
23. Hanke AT, Tsintavi E, del Ramirez Vazquez MP, et al. 3D-liquid chromatography as a complex mixture characterization tool for knowledge-based downstream process development. *Biotechnol Prog*. 2016;32(5):1283-1291. doi:10.1002/btpr.2320
24. Pirrung SM, Parruca da Cruz D, Hanke AT, et al. Chromatographic parameter determination for complex biological feedstocks. *Biotechnol Prog*. 2018;34(4):1006-1018. doi:10.1002/btpr.2642
25. Timmick SM, Vecchiarello N, Goodwine C, et al. An impurity characterization based approach for the rapid development of integrated downstream purification processes. *Biotechnol Bioeng*. 2018;115(8):2048-2060. doi:10.1002/bit.26718
26. Vecchiarello N, Timmick SM, Goodwine C, et al. A combined screening and in silico strategy for the rapid design of integrated downstream processes for process and product-related impurity removal. *Biotechnol Bioeng*. 2019;116(9):2178-2190. doi:10.1002/bit.27018
27. Eliuk S, Makarov A. Evolution of orbitrap mass spectrometry instrumentation. *Annu Rev Anal Chem*. 2015;8:61-80. doi:10.1146/annurev-anchem-071114-040325
28. Nfor BK, Zuluaga DS, Verheijen PJT, Verhaert PDEM, van der Wielen LAM, Ottens and M. Model-based rational strategy for chromatographic resin selection. *Biotechnol Prog*. 2011;27(6):1629-1643. doi:10.1002/btpr.691
29. Keulen D, van der Hagen E, Geldhof G, Le Bussy O, Pabst M, Ottens M. Using artificial neural networks to accelerate flowsheet optimization for downstream process development. *Biotechnol Bioeng*. 2023;1-14. doi:10.1002/bit.28454
30. Guiochon G, Shirazi DG, Felinger A, Katti AM. *Fundamentals of Preparative and Nonlinear Chromatography*. Elsevier Science; 2006.
31. Petzold L. Automatic selection of methods for solving stiff and nonstiff systems of ordinary differential equations. *SIAM J Sci Stat Comput*. 1983;4(1):136-148. doi:10.1137/0904010

32. Parente ES, Wetlaufer DB. Relationship between isocratic and gradient retention times in the high-performance ion-exchange chromatography of proteins. Theory and experiment. *J Chromatogr A*. 1986; 355:29-40. doi:10.1016/S0021-9673(01)97301-7
33. Brooks CA, Cramer SM. Steric mass-action ion exchange: displacement profiles and induced salt gradients. *AIChE J*. 1992;38(12):1969-1978. doi:10.1002/aic.690381212
34. Shukla AA, Bae SS, Moore JA, Barnthouse KA, Cramer SM. Synthesis and characterization of high-affinity, low molecular weight displacers for cation-exchange chromatography. *Ind Eng Chem Res*. 1998;37(10): 4090-4098. doi:10.1021/ie9801756
35. Wiśniewski JR, Zougman A, Nagaraj N, Mann M. Universal sample preparation method for proteome analysis. *Nat Methods*. 2009;6(5): 359-362. doi:10.1038/nmeth.1322
36. den Ridder M, Knibbe E, van den Brandeler W, Daran-Lapujade P, Pabst M. A systematic evaluation of yeast sample preparation protocols for spectral identifications, proteome coverage and post-isolation modifications. *J Proteomics*. 2022;261:104576. doi:10.1016/j.jprot.2022.104576
37. Chen C, Zhao D, Su Z, et al. Effect of pore structure on protein adsorption mechanism on ion exchange media: a preliminary study using low field nuclear magnetic resonance. *J Chromatogr A*. 2021; 1639:461904. doi:10.1016/j.chroma.2021.461904
38. Yao Y, Lenhoff AM. Pore size distributions of ion exchangers and relation to protein binding capacity. *J Chromatogr A*. 2006;1126(1–2): 107-119. doi:10.1016/j.chroma.2006.06.057
39. Ratanji KD, Derrick JP, Kimber I, Thorpe R, Wadhwa M, Dearman RJ. Influence of *Escherichia coli* chaperone DnaK on protein immunogenicity. *Immunology*. 2017;150(3):343-355. doi:10.1111/imm.12689
40. Panikulam S, Hanke A, Kroener F, et al. Host cell protein networks as a novel co-elution mechanism during protein A chromatography. *Biotechnol Bioeng*. 2024;121:1716-1728. doi:10.1002/bit.28678
41. Schmidt-Traub H, Schulte M, Seidel-Morgenstern A. Preparative Chromatography. 2012.
42. Huuk TC, Briskot T, Hahn T, Hubbuch J. A versatile noninvasive method for adsorber quantification in batch and column chromatography based on the ionic capacity. *Biotechnol Prog*. 2016;32:666-677. doi:10.1002/btpr.2228
43. Ruthven DM. *Principles of Adsorption and Adsorption Processes*. 1984.
44. Hagel L. Chapter 3 - Gel filtration: Size exclusion chromatography. In: Jan-Christer J, ed. *Protein Purification: Principles, High Resolution Methods, and Applications*. 3rd ed. Wiley; 2011:57.
45. Wierling PS, Bogumil R, Knieps-Grünhagen E, Hubbuch J. High-throughput screening of packed-bed chromatography coupled with SELDI-TOF MS analysis: monoclonal antibodies versus host cell protein. *Biotechnology and Bioengineering*. 2007;98(2):440–450. doi:10.1002/bit.21399

How to cite this article: Disela R, Keulen D, Fotou E, et al. Proteomics-based method to comprehensively model the removal of host cell protein impurities. *Biotechnol. Prog*. 2024; e3494. doi:10.1002/btpr.3494

APPENDIX A

A.1 | SAMPLE PREPARATION FOR HOST CELL PROTEOMIC ANALYSIS

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

A.2 | METHODS TO DETERMINE MODEL PARAMETER

For the development of the mechanistic model, various parameters were obtained experimentally. This included the determination of column parameters like porosities and system dead volumes using pulse experiments. Furthermore, the ionic capacity was assessed by displacement experiments. Using these parameters, isotherm parameters are regressed from the retention volumes determined in LGE experiments with varying gradient length.

A.2.1. | Pulse experiments

250 µL nonbinding tracers were used to investigate the dead volumes and porosities in the system and chromatography column. 7.5 g/L dextran 2400 K from the American Polymer Standards Corporation was used as a nonpenetrating tracer. High salt buffer was used as penetrating tracer. Porosities were determined as described in literature.[42]

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

A.2.2. | Displacement experiments

The ionic capacity of the absorber was measured by displacement experiments using HCl titration. First, the column was washed with 1 M NaOH and MilliQ. Subsequently, 0.05 M HCl was titrated until

an increase of the inline conductivity trace was observed. The HCl volume and the system dwell volumes were used to calculate the ionic capacity for the skeleton of the Q Sepharose XL resin in the column.[43] The ionic capacity was calculated as follows:

$$\Lambda = \frac{\left(V_{tit.HCl} - V_{dwell,system} - V_{dwell,cond}\right) * c_{HCl}}{V_{col} * (1 - \varepsilon_t)}, \quad \text{(A1)}$$

where the titration volume $V_{tit.HCl}$ is determined as the volume from the start of the titration until the start of the increase in measured conductivity signal. From this, the dwell volume of the system $V_{dwell,system}$ and the dwell volume of the tubing until the conductivity sensor $V_{dwell,cond}$ are subtracted. The determined ionic capacity Λ for the skeleton of Q Sepharose XL resin was calculated using the HCl concentration, the column volume V_{col} and the total porosity of the resin and was determined to be 1.106 mmol/L.

A.3 | MASS TRANSFER CORRELATION

The mass transfer is described with

$$k_{ov,i} = \left[\frac{d_p}{6k_{f,i}} + \frac{d_p^2}{60\varepsilon_p D_{p,i}}\right]^{-1}. \quad \text{(A2)}$$

The overall mass transfer coefficient, represented as $k_{ov,i}$, is the composite outcome of both distinct film mass transfer resistance and mass transfer resistance within the pores.[44] Equation (A2) incorporates parameters such as particle diameter d_p, intraparticle porosity ε_p, and effective pore diffusivity coefficient $D_{p,i}$. The film mass transfer resistance is expressed as $k_{f,i} = D_{f,i}Sh/d_p$, where $D_{f,i}$ describes free diffusivity, and Sh stands for the Sherwood number. Compared with previously mentioned mechanistic models, the empirical correlation from Sofer and Hagel[45] was employed to describe free diffusivity as a function of the molecular weight MW with

$$D_{f,i} = 260 * 10^{-11}\left(MW^{-\frac{1}{3}}\right). \quad \text{(A3)}$$

A.4 | COLUMN CHARACTERISTICS

TABLE A1 Column characteristics for HiTrap Q XL column (5 mL).

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

[a]Manufacturer.
[b]Reference 38.