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

Disela, Roxana; Neijenhuis, Tim; Le Bussy, Olivier; Geldhof, Geoffroy; Klijn, Marieke; Pabst, Martin; Ottens, Marcel

DOI
10.1002/bit.28840

Publication date
2024

Document Version
Final published version

Published in
Biotechnology and Bioengineering

Citation (APA)
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*, *121*(12), 3848-3859. https://doi.org/10.1002/bit.28840

Received: 8 May 2024 | Revised: 29 August 2024 | Accepted: 31 August 2024

DOI: 10.1002/bit.28840

ARTICLE

Biotechnology and Bioengineering WILEY

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

Roxana Disela[1] | Tim Neijenhuis[1] | Olivier Le Bussy[2] | Geoffroy Geldhof[2] | Marieke Klijn[1] | Martin Pabst[1] | Marcel Ottens[1]

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

[2]Technical Research & Development, GSK, Rixensart, Belgium

Correspondence
Marcel Ottens
Email: m.ottens@tudelft.nl

Funding information
GlaxoSmithKline; ChemistryNL

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 of 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 data set 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 toward greater process understanding.

KEYWORDS

downstream process development, *E. coli*, host cell proteins (HCPs), mass spectrometry (MS), quantitative structure-property relationship (QSPR), recombinant biopharmaceuticals

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 (Bracewell et al., 2015). Especially for recombinant biopharmaceuticals, produced intracellularly or in the periplasm, where harvest requires cell lysis, resulting in a complex mixture (Bracewell et al., 2015; Tscheliessnig et al., 2013).

Roxana Disela and Tim Neijenhuis contributed equally to this study.

For the purification of protein-based pharmaceuticals, packed bed chromatography has been the industry standard due to its high versatility and specificity (Gottschalk et al., 2012). 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 (Hanke & Ottens, 2014; Keulen et al., 2022). In recent years, tools like high throughput experimentation and in-silico modeling have shown great potential to accelerate the design process (Bernau et al., 2022; Keulen et al., 2023; Nfor et al., 2012; Pirrung et al., 2018). 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, QSPR models aim to use specific features calculated from the protein structures (Bernau et al., 2022; Emonts & Buyel, 2023). Over the last 20 years, successful models have been trained for a variety of globular proteins or antibodies (Hanke et al., 2016; Hess et al., 2024; Kittelmann et al., 2017; Mazza et al., 2001; Saleh et al., 2023; Yang et al., 2007). 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 (Cai et al., 2024). These prediction methods become even more powerful in combination with mechanistic modeling, allowing full prediction of the elution profile (Hess et al., 2024; Saleh et al., 2023). 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 protein–protein interactions (PPIs) occur between HCPs (Arifuzzaman et al., 2006; Rajagopala et al., 2014). 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 (Kryshtafovych et al., 2023). 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 (Timmick et al., 2018; Tscheliessnig et al., 2013; Wang et al., 2009). Mass spectrometry (MS)-based proteomics has gained popularity for analyzing HCPs, enabling the sensitive detection of individual HCPs during process development (Bracewell et al., 2015; Jagschies et al., 2018; Rathore et al., 2018; Schenauer et al., 2012; Tscheliessnig et al., 2013). Advances in the field allow identification of specific proteins which are commonly remaining after the downstream processing (Molden et al., 2021). 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 (Jones et al., 2021; Migani et al., 2017; Panikulam et al., 2024; Vanderlaan et al., 2018). From these, high-risk HCPs have been identified for CHO, that have potential immunogenic responses or compromise product quality due to degradation (Jones et al., 2021). Studies showed that HCP aggregates with mAbs may promote the persistence of HCPs during the protein A capture step (Gagnon et al., 2014; Herman et al. 2023a, 2023b; Oh et al., 2023). A recent correlation analysis of HCPs identified co-elution of HCPs in groups that are associated with PPIs (Panikulam et al., 2024).

Less studies targeting *Escherichia coli* HCPs have been conducted. To identify HCP co-elution in immobilized metal affinity chromatography, Bartlow et al. analyzed 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 (2011). More recently, Lingg et al. investigated the effect of metal and chelator type on the HCPs found in the eluate of a similar process (2020). For cation- and anion-exchange chromatography, Swanson et al. studied *E. coli* HCP elution in a 5-step isocratic elution (2012, 2016). Using the experimentally determined molecular weight, isoelectric point (pI) and aqueous two-phase partitioning coefficients of the HCPs, random forest regressor models were trained to predict the protein retention. In a more fundamental study, Disela et al. performed MS analysis on *E. coli* BLR(DE3) and HMS174(DE3) HCPs and plotted proteome property maps using the physicochemical properties of around 2000 HCPs to showcase the selection of suitable purification strategies (2023).

Despite these efforts, knowledge of 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 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.

2 | MATERIALS AND METHODS

2.1 | Chromatographic experiments and proteomic analysis

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 (Disela et al., 2023).

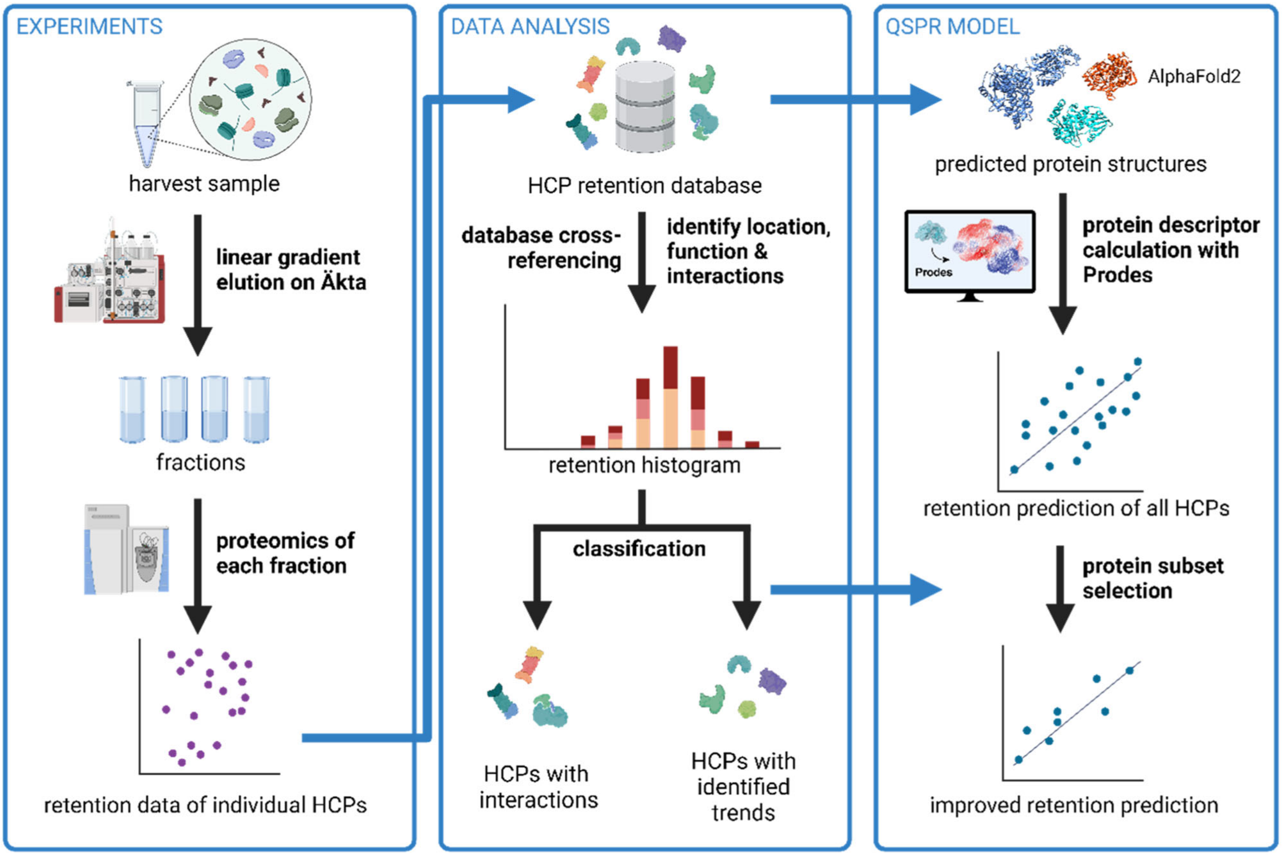
EXPERIMENTS
harvest sample
linear gradient elution on Äkta
fractions
proteomics of each fraction
retention data of individual HCPs
DATA ANALYSIS
HCP retention database
database cross-referencing
identify location, function & interactions
retention histogram
classification
HCPs with interactions
HCPs with identified trends
QSPR MODEL
AlphaFold2
predicted protein structures
Prodes
protein descriptor calculation with Prodes
retention prediction of all HCPs
protein subset selection
improved retention prediction

FIGURE 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 ion exchange and hydrophobic interaction chromatography 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 quantitative structure–property relationship model and investigate several variations using filters based on the deviating retention trends (Illustration created using BioRender.com).

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.

2.1.2 | LGE 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 afterward analyzed using MS. During the IEX experiment, 1 mL fractions were taken and every other fraction was analyzed, as described in more detail in Disela et al. (2024). For the HIC experiment, 2.5 mL fractions were taken and every fraction was analyzed.

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 Disela et al. (2024).

2.2 | Data processing

The retention profiles (in peak area) of the proteins eluting during the gradient were fitted to a Gaussian function. If the shape could be fitted with a R^2 above 0.7, the maximum of the fitted Gaussian function was used as the retention volume $V_{R,i}$ of each protein i as exemplified in (Disela et al., 2024). Since a constant flow rate was used in the experiments, the dimensionless retention time (DRT) could be calculated as

$$\mathrm{DRT}(i) = \frac{V_{R,i} - V_g}{V_G - V_g}, \quad (1)$$

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 (Hanke et al., 2016).

Abundance measures (for the common scatter plot) and theoretical physicochemical properties were retrieved from a previous study of the harvest sample (Disela et al., 2023). The cellular location and functions were retrieved from UniProt (Bateman et al., 2021). Hereby proteins that were exclusively located in the cytosol or cytoplasm, not in a membrane, were summarized as cytoplasm proteins. Comparable *E. coli* K-12 proteins were retrieved from Arifuzzaman et al. (2006) that show PPIs (Supporting Information S1: Table 1 in Arifuzzaman et al., 2006) and proteins without measured interactions (Supporting Information S1: Table 2 in Arifuzzaman et al., 2006).

2.3 | QSPR

2.3.1 | Protein model generation

Using the database presented in Disela et al. (2023), 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 (Jumper et al., 2021). 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 (Neijenhuis et al., 2024). Visualization of the protein structures was performed using UCSF Chimera (Pettersen et al., 2004).

2.3.2 | QSPR model training

Multilinear Regression (MLR) models were trained for the retention time prediction of the whole data set and specific subsets of HCPs (Supporting Information S1: Table 1). The selection of proteins for each subset was based on their presence in the cytoplasm, their multimeric state, described interactions and average per-residue model confidence score. Initially, the data sets 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%–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 (Supporting Information S2: Figure S1), were selected based on the R^2 of a 10-fold cross-validation.

3 | RESULTS AND DISCUSSION

3.1 | Retention behavior of individual HCPs

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 were 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 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 2a,b, 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 (Rappsilber et al., 2002). 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 2c,d, respectively).

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

3.1.2 | Influence of cellular location

To better understand the behavior of specific HCPs, the extensive proteome data set was explored regarding a variety of factors which

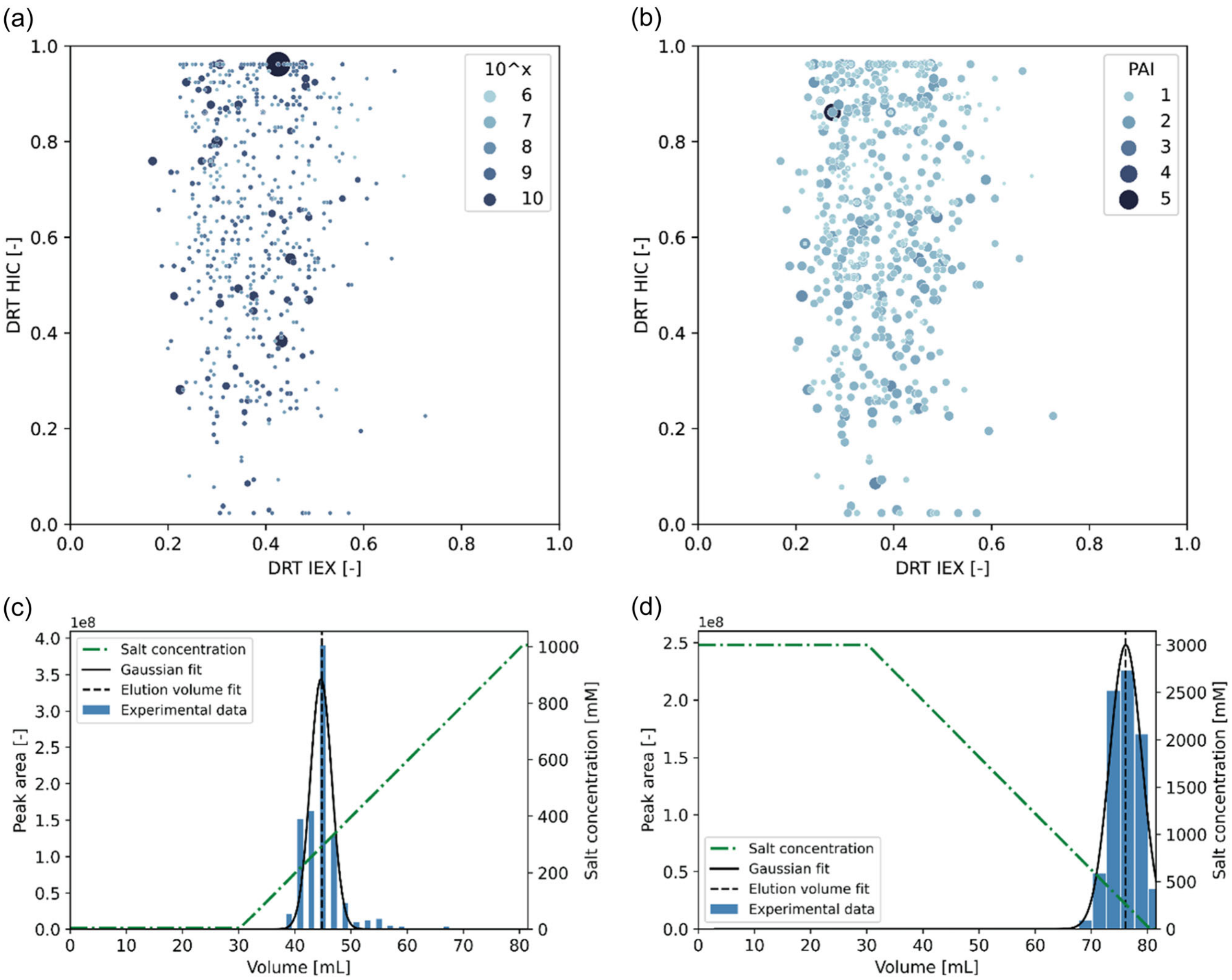

FIGURE 2 Host cell protein (HCP) retention map of individual HCPs in the *Escherichia coli* lysate. Dimensionless retention times were obtained from mass spectrometry analysis of fractions obtained from linear gradient experiments on Q Sepharose XL (ion exchange [IEX]) and Butyl FF (hydrophobic interaction chromatography [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 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.

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 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 data set, except for the plasma membrane HCPs (Supporting Information S2: Figure S2a), suggesting interactions of these proteins leads to concurrent elution. This indicates that forces causing these interactions are stronger compared to electrostatic forces that are the main interaction as shown by the IEX trendline of the majority of the proteins. Plasma membrane proteins might interact with each other directly forming parts of known (sdhB, secY) or unknown complexes (hflC, arnC) (Maddalo et al., 2011). We even observe the co-elution of yidC and secY, that are known to form a multiprotein complex for Sec-dependent membrane protein integration (Kumazaki et al., 2014). However, the joint elution of several plasma membrane proteins might indicate that they form liposomes or are parts of membrane vesicles (Nagakubo et al., 2020). Considering that HCPs are impurities, a concurrent elution could simplify the

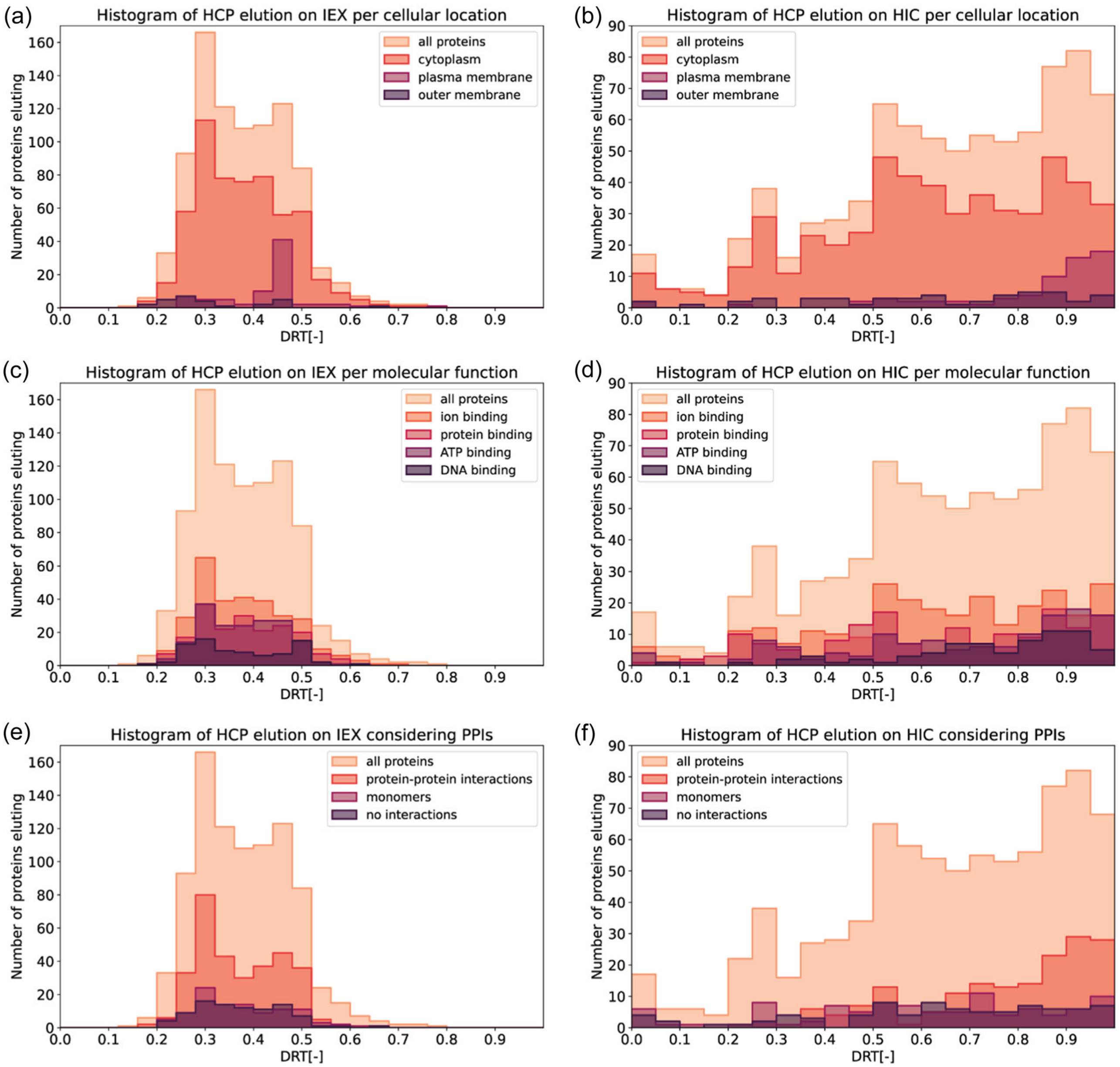

FIGURE 3 Histograms representing the elution of groups of host cell proteins (HCPs). The number of proteins with an elution maximum during a specific dimensionless retention time is listed for ion exchange (IEX) and hydrophobic interaction chromatography (HIC). (a) Histogram of cellular location groups 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.

development of the chromatography step. However, for a retention prediction model, joint elution hampers the prediction for these proteins, when using calculated protein features.

During the HIC gradient, the histogram of cytoplasm proteins (total 532) shows a similar shape to the histogram of all proteins with a slightly lower number of proteins eluting toward the end of the gradient (Figure 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 (Hanke et al., 2016). 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.

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 (Supporting Information S2: Figure S2b).

During the HIC gradient, the ion (272), protein (165), ATP (133), and DNA binding proteins (71) are distributed across all elution times with no clear elution points (Figure 3d).

3.1.4 | Influence of PPIs

In the complex mixture of a host cell lysate proteins can interact, forming functional or nonfunctional complexes. The different PPIs at physiological conditions between *E. coli* proteins were identified by Arifuzzaman et al. (2006). Out of the interactions identified by Arifuzzaman et al., 1270 were found in the IEX data set and 1225 in the HIC data set. 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 (Jakob et al., 2021). Nevertheless, these interactions pose an interesting effect on the DRTs of involved HCPs as indicated in a recent study for CHO cells (Panikulam et al., 2024).

To identify the effect of PPIs, proteins described to interact from protein pairs in close proximity were selected (Figure 3e,f). Proteins described to have no interactions in Arifuzzaman et al. were also plotted as one group. Additionally, proteins known to be present as monomers were grouped. During the IEX gradient, the proteins with PPIs (319) show a high abundance at 0.30 and 0.46 DRT and the surrounding fractions. This shape impacts the histogram with all proteins significantly. Monomers (104) and noninteracting 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 (O'Farrell, 2008). For the monomers (98) and noninteracting 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.

3.2 | Prediction of retention time of individual HCPs in IEX

3.2.1 | Descriptive QSPR model using the complete data set

Using the DRTs obtained from IEX LGE of all single peak proteins, a predictive QSPR model was trained, correlating specific physicochemical features to protein retention. A final MLR model composed of 27 features was build achieving a 10-fold cross-validated R^2 of 0.55 and a mean absolute error (MAE) of 0.049 (Figure 4 and Table 1 [ALL]). For the test set, data not involved during feature selection, an 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

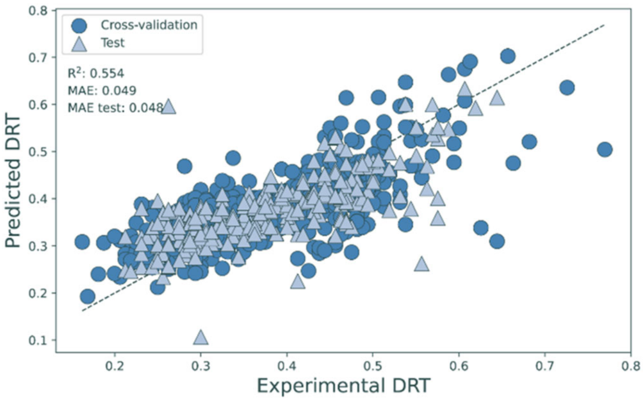

FIGURE 4 Quantitative structure–property relationship validation of the regression model trained to predict dimensionless retention time, where the circles represent the 10-fold cross-validation and the triangles the test set.

TABLE 1 Comparison of model performance for the different protein subsets.

	#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

Note: 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.
Abbreviation: MAE, mean absolute error.

protein with DRT > 0.3 (Figure 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 (Supporting Information S2: Figure S3). Additionally, the surface content of the four charged amino acids was found to be important. Due to the number of features and the inherent collinearity of the charge related features, specific feature importance cannot be identified. The remaining eight features describe the surface, hydrophobicity and the surface content of specific noncharged amino acids. Y-scrambling was performed before training as final validation (Supporting Information S2: Figure S4). The resulting model was not able to predict scrambled protein retention (R^2 of -0.065) proving physical validity.

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 (Jakob et al., 2021). As was mentioned in 3.1.2, no correlation was found with HIC elution and any of the hydrophobicity features for the full data set nor any subsets.

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 data set 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 1, Figure 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 1 and Figure 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 (Supporting Information S2: Figure S5a). In the contrary, the test set was predicted less accurately (MAE of 0.055) compared to the all HCP data set (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 3e), therefore, the bias at 0.3 DRT observed for the other data sets should not pose a problem. However, the test set accuracy (MAE of 0.058) was found to be lower than the all HCP data set (MAE 0.048) (Figure 5b, Table 1). Unlike the all HCP or cytoplasm data sets, 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 data set (data not shown). The loss of accuracy is also thought to be due to the smaller training data set, resulting in less general QSPR models. Therefore, complex behavior, such as oligomerization or complex formation, cannot be captured implicitly.

For the monomer subset a cross-validated R^2 of 0.697 was achieved and the accuracy of the test set was improved to an MAE of 0.043, 7.5% off the experimental error (Table 1, Figure 5c).

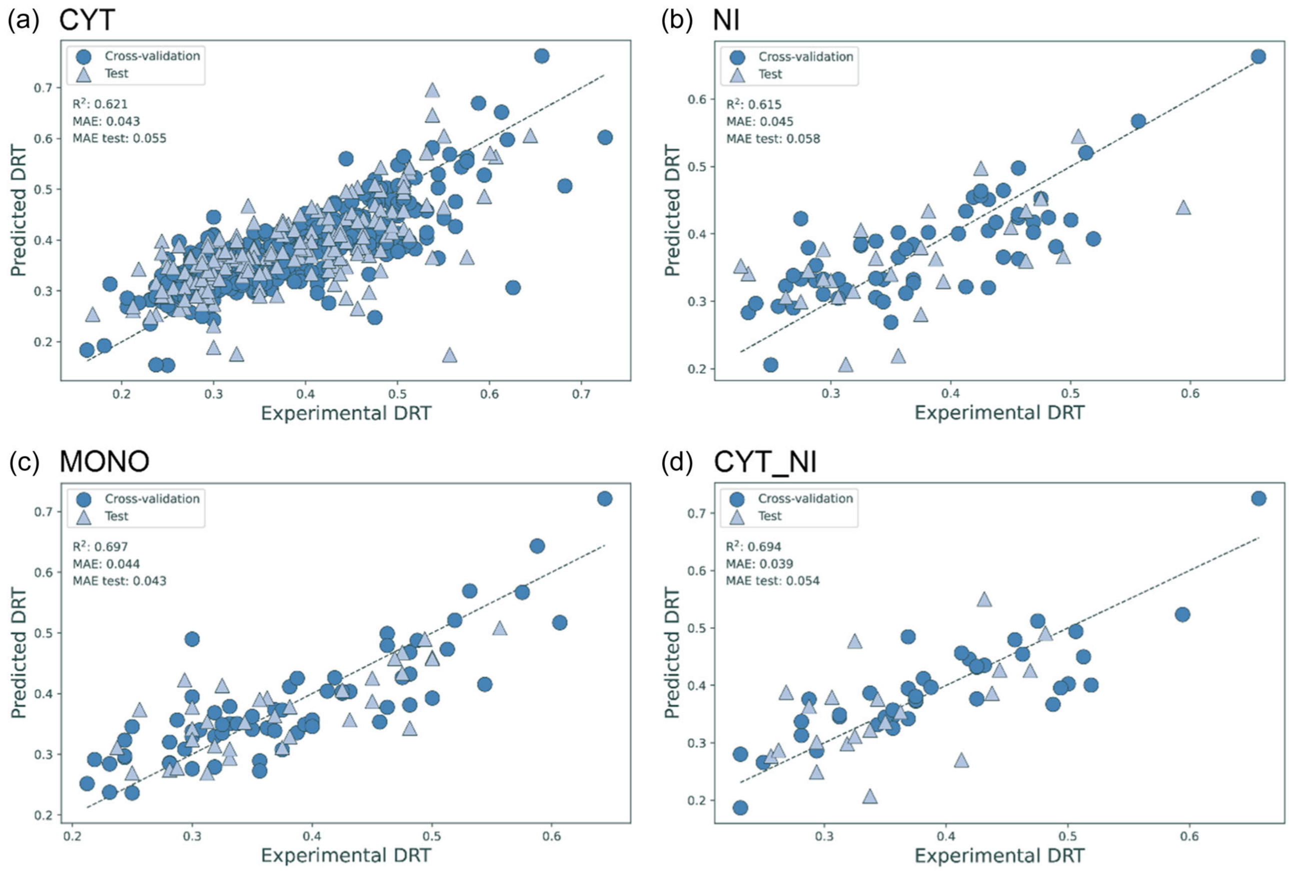

FIGURE 5 Quantitative structure–property relationship validation of the regression model trained to predict dimensionless retention time of protein subsets, where the circles represent the 10-fold cross-validation and the triangles the test set. The presented subsets are the cytosolic proteins (a), the proteins without interactions (b), proteins reported to be present as monomers (c), and proteins which are cytosolic and noninteracting (d).

Additionally, the residuals of the model are spread more evenly compared to the initial elution model allowing prediction of parts of the data set. 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 features resulted in a cross-validated R^2 of 0.47 (Supporting Information S2: Figure S7).

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 (Soleymani et al., 2022). 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 (Supporting Information S2: Figure S9).

Nevertheless, this work provides an important step toward 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 paired 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) (Cai et al., 2024; Hess et al., 2024; Saleh et al., 2023). It can, therefore, be expected that additional research on the algorithms and HCP understanding will allow for robust prediction of HCP retention and knowledge transfer between different processes.

4 | CONCLUSIONS AND OUTLOOK

The observed host cell proteome after lysis of the *E. coli* BLR(DE3) host covers the retention times of around 900 unique proteins on IEX and HIC. By selecting protein subsets based on location, function, and interactions, trends in retention behavior were examined.

For IEX, it was observed that proteins present in the plasma membrane would primarily co-elute, disregarding the general trend of the lower pI resulting in later retention. For HIC, an almost linear trend was observed for the number of proteins throughout the gradient. Only proteins located in the plasma membrane or that are known to engage in PPIs were found to deviate from this trend, primarily eluting at the end of the HIC gradient. Despite the complexity of the mixture, structure models predicted by AlphaFold2 were used to train a descriptive QSPR model (R^2 of 0.55) for IEX retention, approaching the experimental error. By selecting proteins annotated as monomer in UniProt, the accuracy of the QSPR model improved significantly (R^2 of 0.70). This work is the initial step toward 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, extensive 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 data sets are required. Ultimately, this research represents a significant step toward in-silico driven process development, increasing process understanding and reducing development times.

AUTHOR CONTRIBUTIONS

Roxana Disela: Conceptualization; data curation; formal analysis; investigation; methodology; visualization; writing—original draft; writing—review and editing. **Tim Neijenhuis**: Conceptualization; data curation; formal analysis; investigation; methodology; software visualization; writing—original draft; writing—review and editing. **Geoffroy Geldhof**: Resources; writing—review and editing. **Olivier Le Bussy**: Resources; writing—review and editing. **Marieke Klijn**: Supervision; writing—review and editing. **Martin Pabst**: Investigation; methodology; supervision; writing—review and editing. **Marcel Ottens**: Conceptualization; funding acquisition; supervision; writing—review and editing.

ACKNOWLEDGMENTS

This work was partly financed from PSS-allowance for Top consortia 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 GlaxoSmithKline Biologicals S.A. (Belgium) and the Technical University of Delft (The Netherlands). The authors thank the colleagues from GSK Vaccines and Technical University of Delft, especially Michel Eppink, for their valuable input.

CONFLICT OF INTEREST STATEMENT

Geoffroy Geldhof and Olivier Le Bussy are employees of the GSK group of companies. The remaining authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that supports the findings of this study are available in the supplementary material of this article.

ORCID

Roxana Disela http://orcid.org/0000-0002-8178-5684
Tim Neijenhuis http://orcid.org/0000-0002-6214-5438
Martin Pabst http://orcid.org/0000-0001-9897-0723

REFERENCES

Arifuzzaman, M., Maeda, M., Itoh, A., Nishikata, K., Takita, C., Saito, R., Ara, T., Nakahigashi, K., Huang, H. C., Hirai, A., Tsuzuki, K., Nakamura, S., Altaf-Ul-Amin, M., Oshima, T., Baba, T., Yamamoto, N., Kawamura, T., Ioka-Nakamichi, T., Kitagawa, M., … Mori, H. (2006). Large-scale identification of protein-protein interaction of *Escherichia coli* K-12. *Genome Research*, *16*(5), 686–691. https://doi.org/10.1101/gr.4527806

Bartlow, P., Uechi, G. T., Cardamone, J. r, Sultana, T., Fruchtl, M., Beitle, R. R., & Ataai, M. M. (2011). 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 Expression and Purification*, *78*(2), 216–224. https://doi.org/10.1016/j.pep.2011.04.021

Bateman, A., Martin, M. J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., Alpi, E., Bowler-Barnett, E. H., Britto, R., Bursteinas, B., Bye-A-Jee, H., Coetzee, R., Cukura, A., daSilva, A., Denny, P., Dogan, T., Ebenezer, T., Fan, J., Castro, L. G., … Teodoro, D. (2021). UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Research*, *49*(D1), D480–D489. https://doi.org/10.1093/nar/gkaa1100

Bernau, C. R., Knödler, M., Emonts, J., Jäpel, R. C., & Buyel, J. F. (2022). The use of predictive models to develop chromatography-based purification processes. *Frontiers in Bioengineering and Biotechnology*, *10*(October), 1009102. https://doi.org/10.3389/fbioe.2022.1009102

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

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

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

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*, e3494. https://doi.org/10.1002/btpr.3494

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

Gagnon, P., Nian, R., Lee, J., Tan, L., Latiff, S. M. A., Lim, C. L., Chuah, C., Bi, X., Yang, Y., Zhang, W., & Gan, H. T. (2014). Nonspecific interactions of chromatin with immunoglobulin G and protein A, and their impact on purification performance. *Journal of Chromatography A*, *1340*, 68–78. https://doi.org/10.1016/j.chroma.2014.03.010

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

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

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

Herman, C. E., Min, L., Choe, L. H., Maurer, R. W., Xu, X., Ghose, S., Lee, K. H., & Lenhoff, A. M. (2023a). Analytical characterization of host-cell-protein-rich aggregates in monoclonal antibody solutions. *Biotechnology Progress*, *39*(4), 1–16. https://doi.org/10.1002/btpr.3343

Herman, C. E., Min, L., Choe, L. H., Maurer, R. W., Xu, X., Ghose, S., Lee, K. H., & Lenhoff, A. M. (2023b). Behavior of host-cell-protein-rich aggregates in antibody capture and polishing chromatography. *Journal of Chromatography A*, *1702*, 464081. https://doi.org/10.1016/j.chroma.2023.464081

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

Jagschies, G., Lindskog, E., Lacki, K., & Galliher, P. (Eds.). (2018). *Development, design, and implementation of manufacturing processes*. John Fedor.

Jakob, L. A., Beyer, B., Janeiro Ferreira, C., Lingg, N., Jungbauer, A., & Tscheließnig, R. (2021). Protein-protein interactions and reduced excluded volume increase dynamic binding capacity of dual salt systems in hydrophobic interaction chromatography. *Journal of Chromatography A*, *1649*, 462231. https://doi.org/10.1016/j.chroma.2021.462231

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

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., 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

Keulen, D., Geldhof, G., Bussy, O. L., 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

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*, *121*, 2318–2331. https://doi.org/10.1002/bit.28454

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

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

Kumazaki, K., Kishimoto, T., Furukawa, A., Mori, H., Tanaka, Y., Dohmae, N., Ishitani, R., Tsukazaki, T., & Nureki, O. (2014). Crystal structure of *Escherichia coli* YidC, a membrane protein chaperone and insertase. *Scientific Reports*, *4*, 7299. https://doi.org/10.1038/srep07299

Lingg, N., Öhlknecht, C., Fischer, A., Mozgovicz, M., Scharl, T., Oostenbrink, C., & Jungbauer, A. (2020). Proteomics analysis of host cell proteins after immobilized metal affinity chromatography: Influence of ligand and metal ions. *Journal of Chromatography A*, *1633*:461649. https://doi.org/10.1016/j.chroma.2020.461649

Maddalo, G., Stenberg-Bruzell, F., Götzke, H., Toddo, S., Björkholm, P., Eriksson, H., Chovanec, P., Genevaux, P., Lehtiö, J., Ilag, L. L., & Daley, D. O. (2011). Systematic analysis of native membrane protein complexes in *Escherichia coli*. *Journal of Proteome Research*, *10*(4), 1848–1859. https://doi.org/10.1021/pr101105c

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

Migani, D., Smales, C. M., & Bracewell, D. G. (2017). Effects of lysosomal biotherapeutic recombinant protein expression on cell stress and protease and general host cell protein release in Chinese hamster ovary cells. *Biotechnology Progress*, *33*(3), 666–676. https://doi.org/10.1002/btpr.2455

Molden, R., Hu, M., Yen, E. S., Saggese, D., Reilly, J., Mattila, J., Qiu, H., Chen, G., Bak, H., & Li, N. (2021). Host cell protein profiling of commercial therapeutic protein drugs as a benchmark for monoclonal antibody-based therapeutic protein development. *mAbs*, *13*(1), e1955811-1. https://doi.org/10.1080/19420862.2021.1955811

Nagakubo, T., Nomura, N., & Toyofuku, M. (2020). Cracking open bacterial membrane vesicles. *Frontiers in Microbiology*, *10*(January). https://doi.org/10.3389/fmicb.2019.03026

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

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

O'Farrell, P. A. (2008). Molecular biomethods handbook. In I. J. M. Walker, & R. Rapley, *Molecular biomethods handbook*. Humana Press. https://doi.org/10.1007/978-1-60327-375-6

Oh, Y. H., Becker, M. L., Mendola, K. M., Choe, L. H., Min, L., Lee, K. H., Yigzaw, Y., Seay, A., Bill, J., Li, X., Roush, D. J., Cramer, S. M., Menegatti, S., & Lenhoff, A. M. (2023). Characterization and implications of host-cell protein aggregates in biopharmaceutical processing. *Biotechnology and Bioengineering*, *120*(4), 1068–1080. https://doi.org/10.1002/bit.28325

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*, *121*, 1716–1728. https://doi.org/10.1002/bit.28678

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

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

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

Rappsilber, J., Ryder, U., Lamond, A. I., & Mann, M. (2002). Large-scale proteomic analysis of the human spliceosome. *Genome Research*, *12*(8), 1231–1245. https://doi.org/10.1101/gr.473902

Rathore, D., Faustino, A., Schiel, J., Pang, E., Boyne, M., & Rogstad, S. (2018). The role of mass spectrometry in the characterization of biologic protein products. *Expert Review of Proteomics*, *15*(5), 431–449. https://doi.org/10.1080/14789450.2018.1469982

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

Schenauer, M. R., Flynn, G. C., & Goetze, A. M. (2012). Identification and quantification of host cell protein impurities in biotherapeutics using mass spectrometry. *Analytical Biochemistry*, *428*(2), 150–157. https://doi.org/10.1016/j.ab.2012.05.018

Soleymani, F., Paquet, E., Viktor, H., Michalowski, W., & Spinello, D. (2022). Protein–protein interaction prediction with deep learning: A comprehensive review. *Computational and Structural Biotechnology Journal*, *20*, 5316–5341. https://doi.org/10.1016/j.csbj.2022.08.070

Swanson, R. K., Xu, R., Nettleton, D., & Glatz, C. E. (2012). Proteomics-based, multivariate random forest method for prediction of protein separation behavior during cation-exchange chromatography. *Journal of Chromatography A*, *1249*, 103–114. https://doi.org/10.1016/j.chroma.2012.06.009

Swanson, R. K., Xu, R., Nettleton, D. S., & Glatz, C. E. (2016). Accounting for host cell protein behavior in anion-exchange chromatography. *Biotechnology Progress*, *32*(6), 1453–1463. https://doi.org/10.1002/btpr.2342

Timmick, S. M., Vecchiarello, N., Goodwine, C., Crowell, L. E., Love, K. R., Love, J. C., & Cramer, S. M. (2018). An impurity characterization based approach for the rapid development of integrated downstream purification processes. *Biotechnology and Bioengineering*, *115*(8), 2048–2060. https://doi.org/10.1002/bit.26718

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

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

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

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

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: 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*, 1–12. https://doi.org/10.1002/bit.28840