Research Article

Received: 23 April 2025 Revised: 3 July 2025 Published online in Wiley Online Library:

(wileyonlinelibrary.com) DOI 10.1002/jctb.70026

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

Tim Neijenhuis,[a] Olivier Le Bussy,[b] Geoffroy Geldhof,[b] Marieke E Klijn[a] and Marcel Ottens[a*]

Abstract

BACKGROUND: Selecting an optimal chromatography resin during biopharmaceutical downstream process development is a great challenge. This is especially the case for recombinant subunit vaccines, where product properties vary greatly and recovery often involves cell lysis, which yields a complex mixture of different host cell materials. Host cell protein (HCP) impurities may remain similar for platform processes, but their critical impact on separation efficiency is relative to specific product properties. Therefore, every process needs to be designed per product. Prior knowledge on the elution behavior of HCPs would support the identification of critical compounds. However, determining chromatographic behavior of HCPs experimentally is a time-consuming approach.

RESULTS: In this work, we leverage quantitative structure–property relationship (QSPR) models calibrated with retention data of 13 commercial proteins, collected at pH 7, 8, 9 and 10 to predict the anion-exchange retention of *Escherichia coli* HCPs. These models use features calculated from the molecular structure to describe protein behavior, like chromatographic retention. A multilinear regression model containing two features (isoelectric point and sum of negative surface electrostatics) was able to predict the retention times of 288 HCPs accurately (error ≤ 5%). Moreover, we identified the key attributes missing in the training dataset, which is important to increase model performance in the future.

CONCLUSION: This work showcases how chromatographic data obtained using commercial proteins can be translated to a clarified *E. coli* lysate to accelerate chromatography resin selection for new products.

Supporting information may be found in the online version of this article.

Keywords: chromatography; downstream; proteins; process development; modelling; biotechnology

INTRODUCTION

Recombinant proteins constitute approximately 80% of the global sales in the pharmaceutical industry.[1] To ensure safety and efficacy of these pharmaceuticals, sufficient product purity (reviewed case-by-case) is required.[2] This is achieved by downstream processing (DSP) that often involves a sequence of chromatographic steps separating the target protein from process and host cell impurities.[3-5] While product-related impurities are often most difficult to remove, host cell proteins (HCPs) are a class of impurities that are also challenging to eliminate sufficiently. The main reason for this is that conventionally, HCP impurities are treated as one entity, while these are actually individual entities with a wide variety of physicochemical properties. Therefore knowledge on persistent HCPs is valuable to guide the DSP design.[6] As co-purification is a risk, highly sensitive biochemical methods for detection of persistent HCPs have been developed,[7,8] including identification and quantification by LC–MS/MS proteomics.[9] The relevance of these techniques is reflected by a comprehensive list of high-risk HCPs for monoclonal antibody (mAb) production in Chinese hamster ovary cells.[10] This information can accelerate DSP design in platform processes as different mAb products have

* *Correspondence to: Marcel Ottens, Department of Biotechnology, Delft University of Technology, Van der Maasweg 9, Delft 2629 HZ, The Netherlands. E-mail: m.ottens@tudelft.nl*

a *Department of Biotechnology, Delft University of Technology, Delft, The Netherlands*

b *GSK, Technical Research & Development – Microbial Drug Substance, Rixensart, Belgium*

relatively similar properties that affect purification.[11] This means that the criticality of HCPs does not change for new products. Unfortunately, DSP design is less straightforward for other recombinant proteins, such as subunit vaccines.[12] Unlike for mAbs, affinity chromatography is rarely available for subunit vaccines, as their properties vary widely. Additionally, formulation of standardized sets of HCPs that are likely to co-elute during a chromatography step is impossible. To increase process understanding, Disela *et al*. analyzed the HCP content of *Escherichia coli* lysates from different strains and expression vectors.[13] The HCP content was found to be 80% to 90% similar between lysates, leading to the use of HCP property maps to guide DSP design. These property maps allow for the identification of potential critical HCPs by comparing their properties to the properties of the subunit vaccine.

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

Unfortunately, most QSPR models trained to predict protein chromatographic behavior have only been validated for purified proteins. This makes it challenging to assess their accuracy for complex mixtures, such as host cell lysates, where many interactions occur that potentially change protein retention behavior. An example of more complex mixtures is the study by Keulen *et al*., where QSPR models were successfully trained for the prediction of ion-exchange chromatography retention of proteins in three-component mixtures.[19] However, the total protein concentration of 2.5 g L^{-1} used in that study is considered insufficient for notable protein interactions. A more representative complex mixture was used by Buyel *et al*.[28] Here, QSPR models were trained on protein elution salt concentrations reported in the literature to predict the retention of tobacco HCPs in ion-exchange and mixed-mode chromatography. Estimated elution profiles of 67 HCPs were combined and compared to an experimental chromatogram of a clarified extract, where a good agreement for SP Sepharose FF was found. Unfortunately, accuracy of specific HCPs could not be quantified as the experimental data do not provide elution behavior of specific proteins. Disela *et al*. performed a more quantitative study on a clarified lysate of the *E. coli* expression host, where fractions were collected from linear gradient experiments and analyzed by LC–MS/MS.[20,33] Such detailed experimental characterization provides valuable data, but the studies are time- and resource-intensive. These efforts could be minimized by training QSPR models with data obtained for readily available (commercial) proteins and subsequently transfer the model for the prediction chromatographic behavior of HCPs in complex mixtures.

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

METHODS

Materials and equipment

The retention experiments were performed on two separate Äkta pure systems (Cytiva, Marlborough, USA). Both systems were equipped with a prepacked HiTrap Q Sepharose XL 1 mL column (Cytiva, Marlborough, USA) (Table A1). All substances were purchased from Sigma-Aldrich (Saint Louis, USA) and buffers were prepared using ultrapure water filtered with a Milli-Q Advantage A10 (Merck Millipore, Burlington, USA). Buffer solutions at pH 7, 8, 9 and 10 were prepared with 20 mmol L^{-1} NaCl (buffer A) and 1000 mmol L^{-1} NaCl (buffer B) for running and elution. For pH 7 and 8, a 20 mmol L^{-1} Tris–HCl solution was made, while for pH 9 and 10, 20 mmol L^{-1} ethanolamine was used. The pH was adjusted by titration with 1 mol L^{-1} sodium hydroxide or 1 mol L^{-1} hydrochloric acid. All buffers were filtered using a 0.2 μm membrane disc filter (Pall Corporation, New York, USA) followed by 20 min of sonication.

Albumin (bovine), albumin (human), pepsin, trypsin inhibitor A, lipase, α-lactalbumin, β-lactoglobulin a, glucose oxidase, lipoxygenase, ovotransferrin, amyloglucosidase, urease and catalase were purchased from Sigma-Aldrich (Saint Louis, USA). Each protein was dissolved in buffer A to reach a concentration of 2 g L^{-1}, after which the solutions were filtered using a 0.22 μm Whatman Puradisc FP 30 mm (Cytiva, Marlborough, USA).

Linear gradient elution experiments and data processing

The retention times of the selected proteins were determined for a 10 column volume linear gradient elution from buffer A to buffer B. Each linear gradient elution was performed at a flowrate of 1 mL min^{-1} by injecting 200 μL of protein solution followed by a 5 column volume wash with buffer A and 10 column volume gradient to 100% buffer B. Columns were regenerated with 0.5 mol L^{-1} NaOH and stored in 20% ethanol. To normalize the protein retention for the two systems, the normalized retention times (V_R) were calculated as:

$$V_R = V_{R,0} - 0.5V_{inj} - V_d - V_m - V_{wash}$$

where $V_{R,0}$ is the initial retention time, V_{inj} is the injection volume, V_d is the dwell volume, V_m is the column void volume and V_{wash} is the volume of buffer A used between injection and start of the gradient.[19,33] Finally, to make the data column independent, and allowing the comparison of retention times obtained for the

5 mL HiTrap Q Sepharose XL column, the dimensionless retention time (DRT) was calculated as:

$$\mathrm{DRT} = \frac{V_R}{V_G}$$

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

QSPR modeling

Molecular structures of the commercial proteins were retrieved from the Protein Data Bank[34] with the exception of trypsin inhibitor A. The structure for this protein was retrieved from the AlphaFold database[35] as the experimental structures available missed the positions of some atoms. The full list of the structures used can be found in Table 1. For each protein the feature sets were calculated at pH 7, 8, 9 and 10 using the default settings of Prodes.[18] Feature redundancy was reduced by removing features with a Pearson correlation ≥0.9 to other features. Selection of which feature to remove was based on the cumulative cross-correlation to all other features, keeping the feature with the lowest score. The final feature set used for the multilinear regression model was selected by sequential forward selection. Model accuracy was evaluated by *k*-fold cross validation, leaving out all datapoints representing one protein at a time. This was done to reduce the risk of overfitting as pH-independent features would be constant for the same protein at different pH values. The final model was tested using a dataset of *E. coli* HCP DRTs described in a previous article.[20] To make sure that the test data are similar to the training data, HCPs with any features selected for the model that were outside the range (below the minimum or above the maximum) observed in the training data were removed from the test set.

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

$$r_i = y_i - \hat{y}_i$$

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

RESULTS AND DISCUSSION

For the purpose of training a transferable QSPR model, 13 proteins were selected with an isoelectric point (pI) ranging from 3 to 6.8, thereby ensuring chromatographic retention in anion-exchange chromatography. From the surface electrostatic potentials, it can be observed that the surface is predominantly negatively charged, except for lipoxygenase and ovotransferrin which also show positive patches (Fig. 1).

Retention times for these proteins were determined for a 10 column volume gradient length (Table 2, Fig. S1), similar to the experimental conditions of the HCPs published elsewhere.[20] To maximize the value of this set of proteins, the retention time was measured at pH 7, 8, 9 and 10. Two datapoints are not reported, namely those for lipase at pH 10 (insufficient UV signal) and catalase at pH 8 (technical error). The results show a longer retention time for higher pH values, as would be expected due to deprotonation of titratable amino acids. However, this trend was not observed for urease and lipase, where chromatographic retention remained constant while varying the pH value. In other work it was reported that lysozyme displayed constant chromatographic retention for SP Sepharose resins at pH 7 and pH 9, which was attributed to a constant global charge.[37] However, in the case of urease and lipase, the global charge varies in the pH range of 7 to 10 when calculated from the molecular structure by Prodes (−15 to −28 and −18 to −24, respectively). Therefore, we hypothesize that these proteins have preferred binding orientations where the local charge does remain constant.

HCP retention prediction

Cross-validation was performed by iteratively removing the retention times of each protein at all pH values from the training set to ensure that prior knowledge about the specific protein was absent during model validation. The sequential forward selection

Table 1. Commercial proteins and the respective system used for linear gradient elution experiments

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

[a] Isoelectric point (pI) was calculated using Prodes.

Journal of Chemical Technology and Biotechnology published by John Wiley & Sons Ltd on behalf of Society of Chemical Industry (SCI).

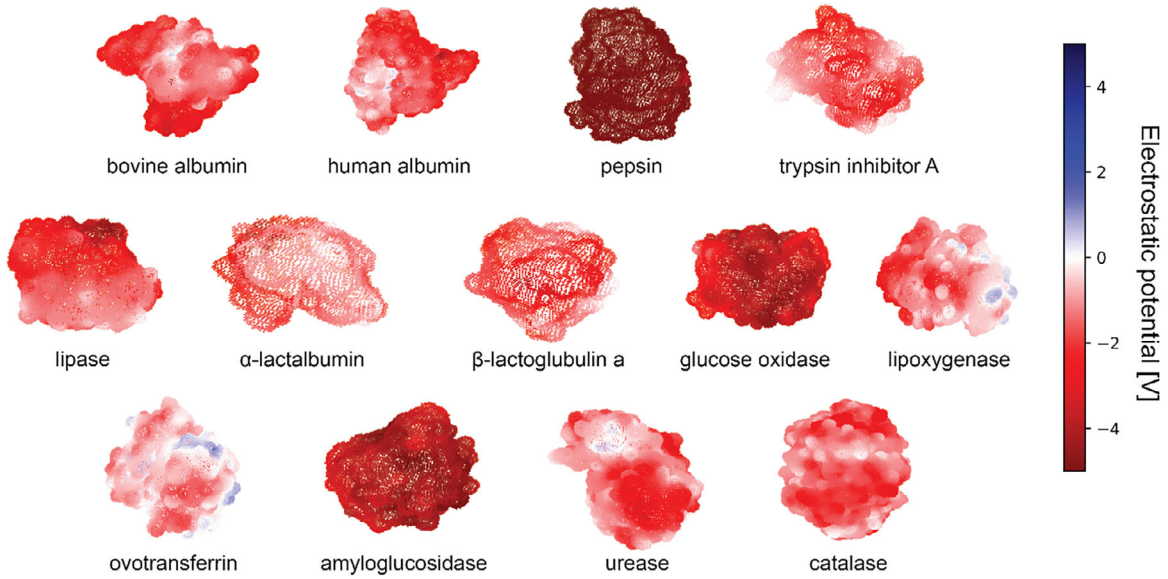

Figure 1. Surface electrostatic potential maps at pH 7 of 13 commercial proteins. The blue and red colors indicate positive and negative electrostatic potential (in volts), respectively.

Table 2. Experimental retention volumes of 13 commercial proteins at pH 7, 8, 9 and 10 on a HiTrap Q Sepharose XL 1 mL column with a 10 column volume gradient from 20 to 1000 mmol L^{-1} NaCl

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

method resulted in a model with four features and a cross-validated R^2 of 0.927 (Fig. S2). Of the four selected features, the protein's pI is most important for predicting the retention time. Permutating this feature has the greatest impact on cross-validation accuracy, diminishing all predictive capabilities (Table A2). However, this feature is not pH dependent and cannot describe any charge-specific behaviors. The second most important feature, the sum of all negative surface points, does capture retention changes by varying the pH. Permutation of this feature results in a significant accuracy reduction to a cross-validated R^2 of 0.76. The remaining two features, the proline surface fraction and median negative surface hydrophobicity potential, have similar permutation scores of 0.88 and 0.87, respectively.

To explore the transferability of the model trained with commercially available proteins, *E. coli* HCPs were used as a test set. This dataset consists of features for 836 HCPs, from which 481 HCPs (approximately 58%) have features that are within the range of the training set. Since QSPR models are only valid for the trained conditions, 481 HCPs were used for testing. With this approach, the retention time could be predicted with a root mean squared error (RMSE) of 0.085 using HCP structures predicted by AlphaFold2 (Fig. 2(A)). To identify HCPs that might co-elute with a target protein, we believe an error of ≤5% to be sufficient considering a DRT between 0 and 1. This takes into account that the DRT describes the retention as a single value, which in reality is a distribution. In practice, when a target protein has a DRT of 0.3, the HCPs with a DRT between 0.2 and 0.4 can be considered as potentially co-eluting. For the test set predictions, 207 (*ca* 43%) HCPs have an error of ≤5%.

To assess the model's ability to generalize for new proteins, the ratio between the RMSE of the test and cross-validation should be analyzed. For the current model, the test set RMSE is three times the cross-validated RMSE. While this might indicate that the training set misses features which are essential to describe HCP retention, the model might also be overfitted. Therefore, a new model was trained using only the two most important features (pI and the sum of the negative surface electrostatics). For this model, the cross-validated R^2 was reduced to 0.840 (Fig. S2) while the test set was predicted with a RMSE of 0.07 (Fig. 2(B)). By eliminating the two least important features, overfitting was significantly reduced (test RMSE is 1.5 times the cross-validated RMSE). This also increased the number of accurately predicted HCPs to 246 (*ca* 51%), which is an 11% improvement. For this test set, the filtering criteria were based on the four feature ranges meaning that the same 481 HCPs were used despite the feature adjustment. Filtering based on the range of two features increases the test set size to 572 HCPs, of which 288 (*ca* 50%) can be predicted with an error of ≤5% (Fig. 2(C)).

HCP structural representation

It should be noted that DRTs of HCPs are predicted using monomer representations obtained from AlphaFold2. Therefore, the QSPR model does not take into account the complex dynamics

Journal of Chemical Technology and Biotechnology published by John Wiley & Sons Ltd on behalf of Society of Chemical Industry (SCI).

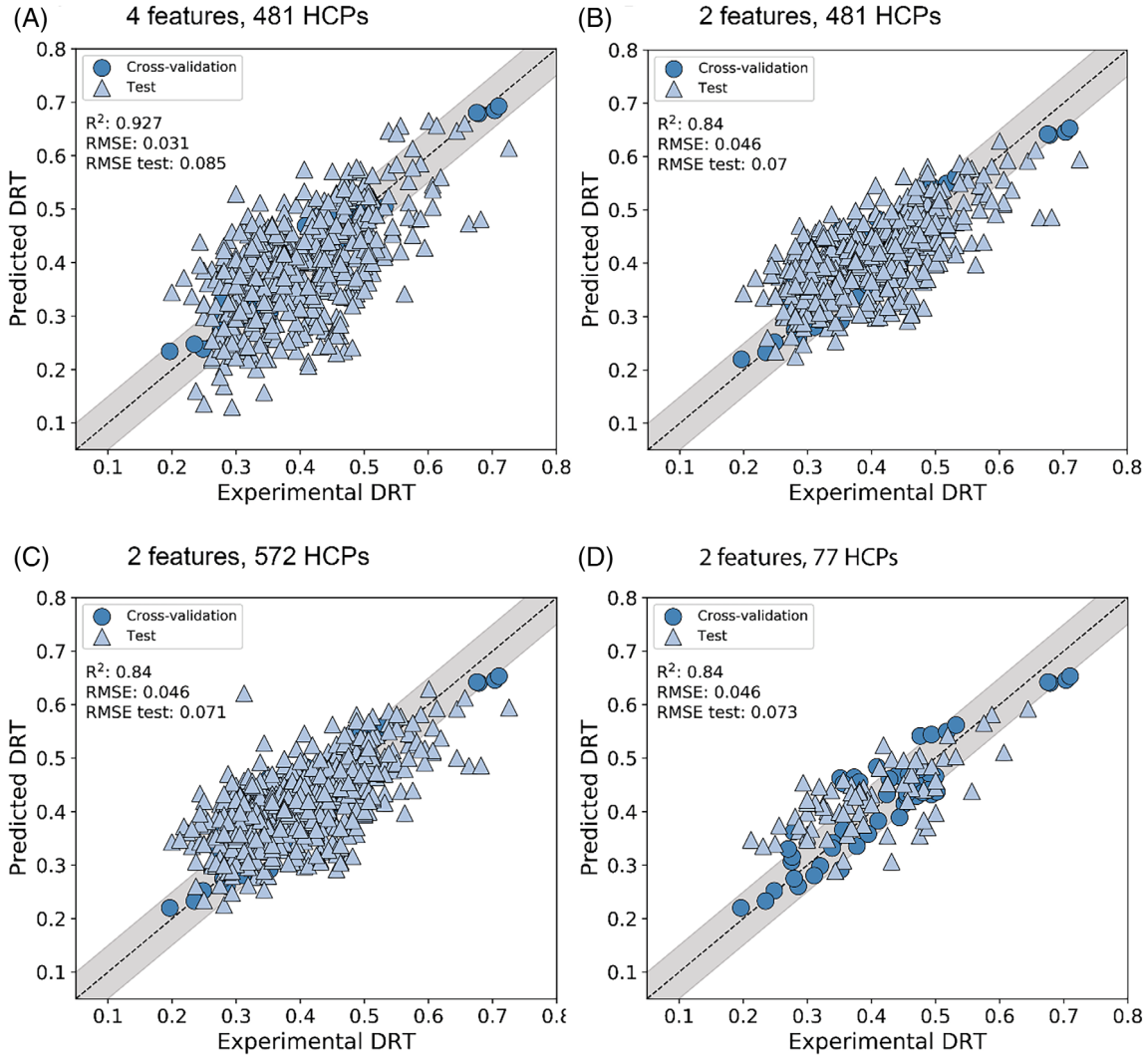

Figure 2. Measured (*x*-axis) *versus* predicted (*y*-axis) DRT of (A) four features and (B–D) two features. Models were validated with *k*-fold cross-validation (circles) and tested on HCP DRTs (triangles), The dotted line represents a perfect prediction and the gray area a 5% error. (A, B) HCP test set filtered for the four-feature model. (C, D) HCPs filtered on the two features. The test set in (D) is reduced to only include monomeric HCPs.

of a lysate mixture, in which many interactions may occur. Still, the model is capable of predicting the DRT of 288 HCPs. The structural representations of proteins that are actually monomeric are expected to be more representative. Therefore, the model with two features was also tested on 77 of the 572 HCPs that are annotated as monomer in Uniprot. Surprisingly, the subset performed similar to the complete HCP test set with a test RMSE of 0.073 and *ca* 43% predictions with ≤5% error (Fig. 2(D)). This suggests that the lack of interaction information about the HCPs does not limit the current model's accuracy. The two features used in the model describe the protein globally and might therefore not capture the required intricacies. A similar phenomenon was observed for the proteins presumed to be homodimers (Fig. S3, Table S1). For this subset of HCPs, predictions using monomer structures (RMSE: 0.071) performed similar to those of homodimer representations (RMSE: 0.068).

Model improvement strategies

We have shown that a QSPR model trained with 50 retention times obtained for 13 proteins at various pH values predicted 288 HCPs with an error of ≤5% using only two features. While this is a significant part of the available HCP retention times, application of QSPR modeling for *in silico* process design would require accurate prediction of all detectable HCPs. To identify possibilities to enhance model performance, the test set predictions were divided into over-predicted (181 HCPs), under-predicted (103 HCPs) and accurately predicted (288 HCPs). For these categories, feature value distributions were analyzed to identify potential biases in the model towards features that were not selected for the QSPR model (Table S2). For a feature that does not contribute to any bias, it can be expected that the distribution over the three sets is similar, which can be observed for the molecular weight (Fig. 3(A)). A feature that shows great differences in distribution is the standard deviation of the surface electrostatics (Fig. 3(B)), with Kolmogorov–Smirnov test values of 0.23 and 0.22 for under- and over-predicted HCPs, respectively. For under-predicted HCPs, a generally higher standard deviation is observed compared to the accurately predicted HCPs, while for over-predicted HCPs this feature tends to be lower. This indicates that the model is lacking information on deviations in surface electrostatics. For the training set, the feature range (0.6–1.2) is much smaller compared to the range in the test set (0.4–1.6) (Fig. S4).

Journal of Chemical Technology and Biotechnology published by John Wiley & Sons Ltd on behalf of Society of Chemical Industry (SCI).

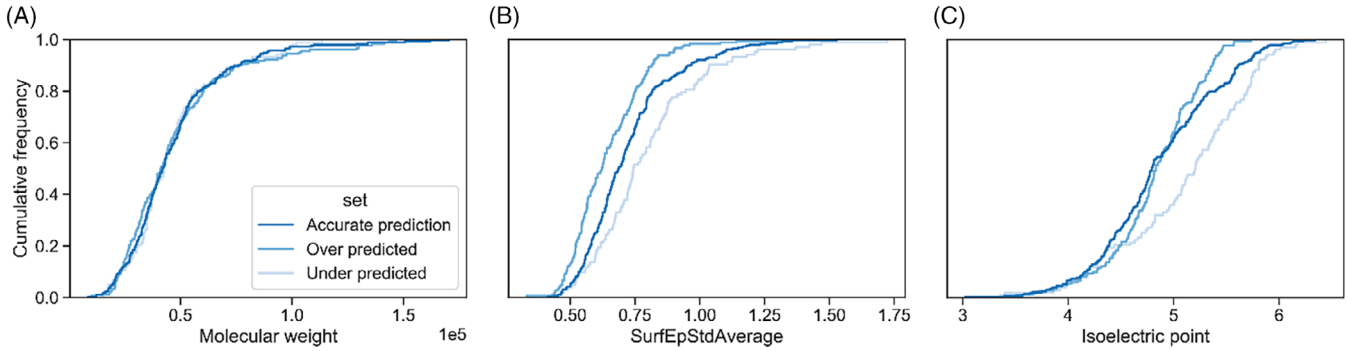

Figure 3. Cumulative distribution plots of the 572 HCPs for (A) molecular weight, (B) standard deviation of the surface electrostatics and (C) pI. The accurately predicted, over-predicted and under-predicted HCPs are represented by blue, orange and green, respectively.

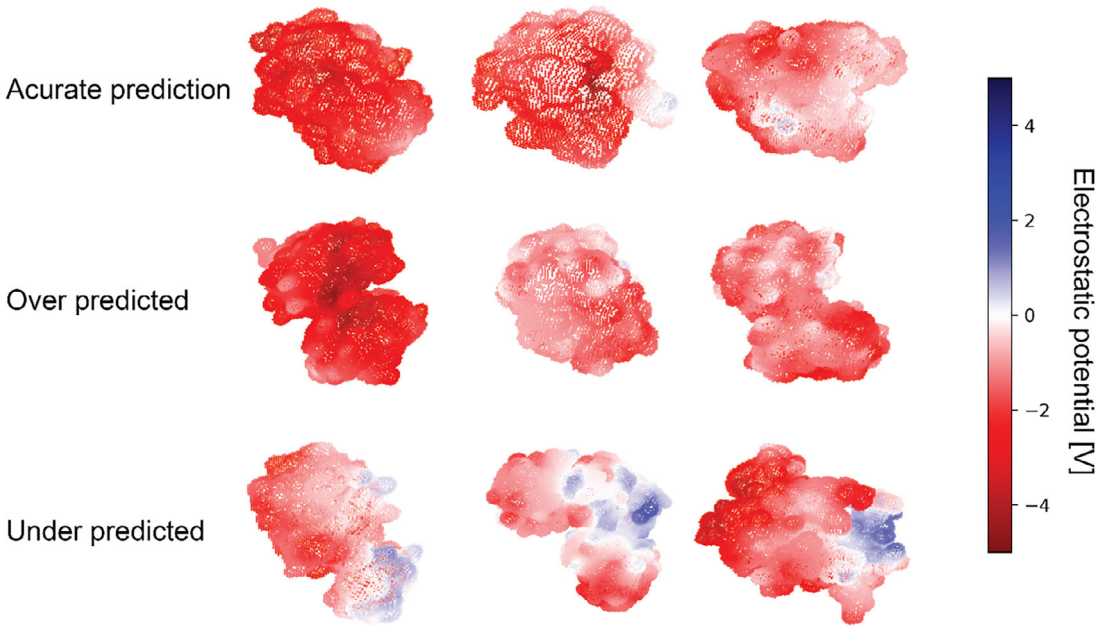

Figure 4. Surface electrostatics at pH 7 of monomer HCPs that are predicted most accurately (top), greatest over-prediction (middle) and greatest under-prediction (bottom). The blue and red colors indicate positive and negative electrostatic potential (in volts).

Therefore, expanding the training set with commercial proteins that have a wider range of this feature could improve model performance.

For the features that were selected for the model, pI showed a notable difference in the distributions (Fig. 3(C)). Especially for pI > 4.5 the feature distribution starts to differ, which indicates that there is a bias for proteins in this pI range. It is therefore important not only to extend the training set based on the surface electrostatics deviation, but also to select proteins with pI > 4.5.

While extending the training set is essential to improve model quality and robustness, design of novel features is considered equally important. Plotting the surface electrostatics of the three monomeric HCPs with the lowest and highest error reveals positively charged surface areas for the under-predicted HCPs (Fig. 4). Such positive patches are not found on the surface of the three accurately predicted HCPs. The presence of these patches contributes to an increase in the surface electrostatic potential standard distribution feature, as can be observed in Fig. 4. Additionally, favorable binding orientations might be more prevalent in the under-predicted HCPs, and these phenomena cannot be captured by the global features used in this study.[38,39] Therefore, designing specific local features representing binding orientations would be essential for improving model performance. For chromatography specifically, local surface features have been designed as either defining patches or projecting properties on a plane.[15,17,40] However, the contribution of preferred binding orientations on adsorption differs between proteins and pH.[37,41,42] This means each protein requires an individual assessment to identify possible binding orientations. This can be done with state-of-the-art molecular dynamics simulations coupled to advanced sampling methods.[39] Unfortunately, these methods are too computationally expensive to perform on the scale of a host cell proteome. As such, future research should focus on identifying computationally efficient methods to score surface patches based on interaction likelihood. This may also include combining information from patches distant from each other, as ligands with flexible linkers (e.g. XL resin used in this study) probably reach multiple binding sides of the protein.[37]

Journal of Chemical Technology and Biotechnology published by John Wiley & Sons Ltd on behalf of Society of Chemical Industry (SCI).

Finally, the choice of regression method could also influence the accuracy. Even though the validation on the training data was satisfactory with a cross-validated R^2 of 0.84, assumptions associated with a multilinear regression model might limit the accuracy.[43] The is especially the case for the assumption that protein retention has a linear dependency on the features. Alternative nonlinear regression methods might be a solution to capture nonlinear dependencies between protein properties and retention behavior. In recent literature, algorithms such as random forest regression, support vector regression or Gaussian process regression have been applied for accurate prediction ($R^2 > 0.85$) of different attributes corresponding to chromatographic behavior.[25,27-30,32] Unfortunately, increasing model complexity comes with a risk of overfitting, especially when using small training datasets.[44]

CONCLUSION AND OUTLOOK

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

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

AUTHOR CONTRIBUTIONS

Tim Neijenhuis: Conceptualization, methodology, investigation, validation, data curation, data analysis, writing – original draft, writing – review & editing, visualization. **Olivier Le Bussy**: Supervision and writing – review & editing. **Geoffroy Geldhof**: Supervision and writing – review & editing. **Marieke E Klijn**: Conceptualization, supervision, writing – review & editing. **Marcel Ottens**: Funding acquisition, conceptualization, supervision, writing – review & editing.

ACKNOWLEDGEMENTS

This work was partly financed from PSS-allowance for Top consortiums for Knowledge and Innovation (TKI) of the Ministry of Economic Affairs and partly sponsored by GlaxoSmithKline Biologicals SA under cooperative research and development agreement between GlaxoSmithKline Biologicals SA (Belgium) and the Technical University of Delft (The Netherlands). The authors thank colleagues from GSK and Technical University of Delft for their valuable input.

DATA AVAILABILITY STATEMENT

Research data are not shared.

CONFLICT OF INTEREST

GG and OLB are employees of the GSK group of companies. The remaining authors declare no conflicts of interest.

SUPPORTING INFORMATION

Supporting information may be found in the online version of this article.

REFERENCES

1 Walsh G and Walsh E, Biopharmaceutical benchmarks 2022. *Nat Biotechnol* **40**:1722–1760 (2022). https://doi.org/10.1038/s41587-022-01582-x.
2 Reiter K, Suzuki M, Olano LR and Narum DL, Host cell protein quantification of an optimized purification method by mass spectrometry. *J Pharm Biomed Anal* **174**:650–654 (2019). https://doi.org/10.1016/j.jpba.2019.06.038.
3 Keulen D, Geldhof G, Le Bussy O, Pabst M and Ottens M, Recent advances to accelerate purification process development: a review with a focus on vaccines. *J Chromatogr A* **1676**:463195 (2022). https://doi.org/10.1016/j.chroma.2022.463195.
4 Hanke AT, *Technologies to Accelerate Protein Purification Process Development*, Vol. **289**. Delft: TU Delft University, (2016).
5 Baumann P and Hubbuch J, Downstream process development strategies for effective bioprocesses: trends, progress, and combinatorial approaches. *Eng Life Sci* **17**:1142–1158 (2017). https://doi.org/10.1002/elsc.201600033.
6 Bracewell DG, Francis R and 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* **112**:1727–1737 (2015). https://doi.org/10.1002/bit.25628.
7 Wang X, Hunter AK and Mozier NM, Host cell proteins in biologics development: identification, quantitation and risk assessment. *Biotechnol Bioeng* **103**:446–458 (2009). https://doi.org/10.1002/bit.22304.
8 Tscheliessnig AL, Konrath J, Bates R and Jungbauer A, Host cell protein analysis in therapeutic protein bioprocessing: methods and applications. *Biotechnol J* **8**:655–670 (2013). https://doi.org/10.1002/biot.201200018.
9 Vanderlaan M, Zhu-Shimoni J, Lin S, Gunawan F, Waerner T and Van Cott KE, Experience with host cell protein impurities in biopharmaceuticals. *Biotechnol Prog* **34**:828–837 (2018). https://doi.org/10.1002/btpr.2640.
10 Jones M, Palackal N, Wang F, Gaza-Bulseco G, Hurkmans K, Zhao Y *et al.*, 'High-risk' host cell proteins (HCPs): a multi-company collaborative view. *Biotechnol Bioeng* **118**:2870–2885 (2021). https://doi.org/10.1002/bit.27808.
11 Shukla AA, Hubbard B, Tressel T, Guhan S and Low D, Downstream processing of monoclonal antibodies: application of platform approaches. *J Chromatogr B* **848**:28–39 (2007). https://doi.org/10.1016/j.jchromb.2006.09.026.
12 Keulen D, Apostolidi M, Geldhof G, Le Bussy O, Pabst M and Ottens M, Comparing in silico flowsheet optimization strategies in biopharmaceutical downstream processes. *Biotechnol Prog* **41**:p.e3514 (2024). https://doi.org/10.1002/btpr.3514.
13 Disela R, Le Bussy O, Geldhof G, Pabst M and Ottens M, Characterisation of the *E. coli* HMS174 and BLR host cell proteome to guide purification process development. *Biotechnol J* **18**:2300068 (2023). https://doi.org/10.1002/biot.202300068.
14 Emonts J and Buyel JF, An overview of descriptors to capture protein properties – tools and perspectives in the context of QSAR modeling. *Comput Struct Biotechnol J* **21**:3234–3247 (2023). https://doi.org/10.1016/j.csbj.2023.05.022.

Journal of Chemical Technology and Biotechnology published by John Wiley & Sons Ltd on behalf of Society of Chemical Industry (SCI).

15 Hanke AT, Klijn ME, Verhaert PDEM, van der Wielen LAM, Ottens M, Eppink MHM *et al.*, Prediction of protein retention times in hydrophobic interaction chromatography by robust statistical characterization of their atomic-level surface properties. *Biotechnol Prog* **32**: 372–381 (2016). https://doi.org/10.1002/btpr.2219.

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

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

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

19 Keulen D, Neijenhuis T, Lazopoulou A, Disela R, Geldhof G, Le Bussy O *et al.*, From protein structure to an optimized chromatographic capture step using multiscale modeling. *Biotechnol Prog* **41**:p.e3505 (2024). https://doi.org/10.1002/btpr.3505.

20 Disela R, Neijenhuis T, Le Bussy O, Geldhof G, Klijn M, Pabst M *et al.*, Experimental characterization and prediction of *Escherichia coli* host cell proteome retention during preparative chromatography. *Biotechnol Bioeng* **121**:3848–3859 (2024). https://doi.org/10.1002/bit.28840.

21 Mazza CB, Sukumar N, Breneman CM and Cramer SM, Prediction of protein retention in ion-exchange systems using molecular descriptors obtained from crystal structure. *Anal Chem* **73**:5457–5461 (2001). https://doi.org/10.1021/ac010797s.

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

23 Yang T, Sundling MC, Freed AS, Breneman CM and Cramer SM, Prediction of pH-dependent chromatographic behavior in ion-exchange systems. *Anal Chem* **79**:8927–8939 (2007). https://doi.org/10.1021/ac071101j.

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

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

26 Song M, Breneman CM, Bi J, Sukumar N, Bennett KP, Cramer S *et al.*, Prediction of protein retention times in anion-exchange chromatography systems using support vector regression. *J Chem Inf Comput Sci* **42**:1347–1357 (2002). https://doi.org/10.1021/ci025580t.

27 Cai QY, Qiao LZ, Yao SJ and Lin DQ, Machine learning assisted QSAR analysis to predict protein adsorption capacities on mixed-mode resins. *Sep Purif Technol* **340**:126762 (2023). https://doi.org/10.1016/j.seppur.2024.1267622024.

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

29 Hess R, Faessler J, Yun D, Mama A, Saleh D, Grosch JH *et al.*, Predicting multimodal chromatography of therapeutic antibodies using multiscale modeling. *J Chromatogr A* **1718**:464706 (2024). https://doi.org/10.1016/j.chroma.2024.464706.

30 Saleh D, Hess R, Ahlers-Hesse M, Rischawy F, Wang G, Grosch JH *et al.*, A multiscale modeling method for therapeutic antibodies in ion exchange chromatography. *Biotechnol Bioeng* **120**:125–138 (2023). https://doi.org/10.1002/bit.28258.

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

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

33 Disela R, Keulen D, Fotou E, Neijenhuis T, Le Bussy O, Geldhof G *et al.*, Proteomics-based method to comprehensively model the removal of host cell protein impurities. *Biotechnol Prog* **40**:e3494 (2024). https://doi.org/10.1002/btpr.3494.

34 Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H *et al.*, The Protein Data Bank. *Nucleic Acids Res* **28**:235–242 (2000). https://doi.org/10.1093/nar/28.1.235.

35 David A, Islam S, Tankhilevich E and Sternberg MJE, The AlphaFold database of protein structures: a biologist's guide. *J Mol Biol* **434**: 167336 (2022). https://doi.org/10.1016/j.jmb.2021.167336.

36 Berger VW and Zhou Y, Kolmogorov–Smirnov test: overview, in *Wiley StatsRef: Statistics Reference Online*. Wiley, pp. 1–5 (2014).

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

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

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

40 Sankar K, Trainor K, Blazer LL, Adams JJ, Sidhu SS, Day T *et al.*, A descriptor set for quantitative structure-property relationship prediction in biologics. *Mol Inform* **41**:2100240 (2022). https://doi.org/10.1002/minf.202100240.

41 Aguilar M-I, Clayton DJ, Holt P, Kronina V, Boysen RI, Purcell AW *et al.*, RP-HPLC binding domains of proteins. *Anal Chem* **70**:5010–5018 (1998). https://doi.org/10.1021/ac980473c.

42 Yao Y and Lenhoff AM, Electrostatic contributions to protein retention in ion-exchange chromatography. 1. Cytochrome c variants. *Anal Chem* **76**:6743–6752 (2004). https://doi.org/10.1021/ac049327z.

43 Osborne JW and Waters E, Four assumptions of multiple regression that researchers should always test.-practical assessment, research & evaluation. *Pract Assess Res Eval* **8**:1–5 (2002).

44 Ying X, An overview of overfitting and its solutions. *J Phys Conf Ser* **1168**:022022 (2019). https://doi.org/10.1088/1742-6596/1168/2/022022.

Journal of Chemical Technology and Biotechnology published by John Wiley & Sons Ltd on behalf of Society of Chemical Industry (SCI).

APPENDIX A

Table A1. System properties

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

Table A2. Model parameters for the QSPR model with four features

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