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

Keulen, Daphne; Neijenhuis, Tim; Lazopoulou, Adamantia; Disela, Roxana; Geldhof, Geoffroy; Le Bussy, Olivier; Klijn, Marieke E.; Ottens, Marcel

DOI
10.1002/btpr.3505

Publication date
2024

Document Version
Final published version

Published in
Biotechnology Progress

Received: 26 January 2024 | Revised: 21 June 2024 | Accepted: 3 September 2024

DOI: 10.1002/btpr.3505

RESEARCH ARTICLE

Bioseparations and Downstream Processing

BIOTECHNOLOGY PROGRESS

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

Daphne Keulen[1] | Tim Neijenhuis[1] | Adamantia Lazopoulou[1] | Roxana Disela[1] | Geoffroy Geldhof[2] | Olivier Le Bussy[2] | Marieke E. Klijn[1] | Marcel Ottens[1]

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

[2]GSK, Technical Research & Development – Microbial Drug Substance, Rixensart, Belgium

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

Funding information
Topconsortium voor Kennis en Innovatie; GlaxoSmithKline Biologicals

Abstract

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

Daphne Keulen and Tim Neijenhuis contributed equally to this work.

KEYWORDS
biopharmaceutical downstream processing, chromatography, mechanistic modeling, process optimization, quantitative structure–property relationships

1 | INTRODUCTION

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

In silico techniques, such as mechanistic or data-driven modeling, can be of great merit for process development. These methods allow for increased process understanding while reducing experimental effort and/or use of critical sample material, and decreasing process development times.[7,8] Within the next years, modeling techniques will become more essential for biopharmaceutical industry. Specifically for Industry 4.0 that aims to digitalize the entire manufacturing process.[9–12] Moreover, increased process understanding and process and product quality control are in agreement with the quality-by-design (QbD) guidelines.[13–16] Identifying the operating window of the critical process parameters (CPP) is an essential part to guarantee process' stability. Currently, these operating windows are determined with expensive and time-consuming wet-lab design-of-experiments (DoE). Chromatographic mechanistic models (MM) attempt to describe the chromatographic process in silico and could be an inexpensive and fast alternative to determine the CPP operating window. Over the past years, the industry has been gradually adopting chromatographic MM, with ongoing advancement being made in determining the essential input parameters.[17–20] In the future, the ultimate objective is to determine adsorption isotherm for complex mixtures more easily.[21,22] Progress in utilizing mass spectrometry data could play a crucial role in achieving this goal.[23] However, at this moment determining adsorption isotherm parameters for the MM remains a bottleneck for industrial application, mainly due to time and material limitations especially in the early phase of downstream process development.[24] Quantitative structure property relationships (QSPR) modeling could be an in silico alternative to experimentally determining the adsorption isotherm parameters. QSPR aims to correlate physicochemical properties with specific behavior, such as chromatographic retention time.[25] These physicochemical properties are calculated from protein structure models that describe the position of each atom. Combining MM with QSPR and optimization tools could pave the way for a holistic modeling approach/workflow.

In 2001, Mazza et al. introduced a QSPR model for predicting protein retention times for ion exchange chromatography.[25] Their approach involved feature calculation using the proprietary software platform Molecular Operating Environment (MOE), followed by a genetic algorithm for feature selection for the training of a partial least squares model.[26,27] As a result, several follow-up studies applied QSPR models to different modes of chromatography/type of chromatography resins, using support vector machine regression methods, and including pH effects.[28–33] Malmquist et al. developed an additional set of protein descriptors that are pH-dependent and based on electrostatic and hydrophobic properties.[34] Moreover, several studies considered the crucial binding orientations within protein-resin binding affinities in their QSPR models.[35–37] In recent years, QSPR has been applied to more complex proteins, such as Fabs and mAbs, showing the growing interest from industry and the added value of these models.[24,38,39] Robinson et al. showed the potential of QSPR models for in silico resin screening of six chromatographic systems applied to Fabs.[38] While Saleh et al. built QSPR models using 21 mAbs variants to predict the adsorption isotherm parameters, the equilibrium constant and the characteristic charge, which were subsequently applied to the MM and able to predict the cation exchange chromatography (CEX) step.[24] Their study shows promising capabilities of a multiscale model to simulate different process conditions without the need for wet-lab experiments. Several software packages are available to calculate the protein descriptors that are needed for QSPR modeling, an overview of these software packages has been provided elsewhere.[40,41] Most software tools are only available via webservers or commercially, lacking source code availability. Therefore, Neijenhuis et al. have recently published an open-source QSPR software tool, which has also been used in this work.[42]

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

This article presents a general multiscale modeling strategy that integrates QSPR and chromatographic MM to optimize a CEX capture step. We were able to simulate and validate a CEX step only using the protein structure. Subsequently, we compared the uncertainty of the experimentally determined and predicted parameters on the final

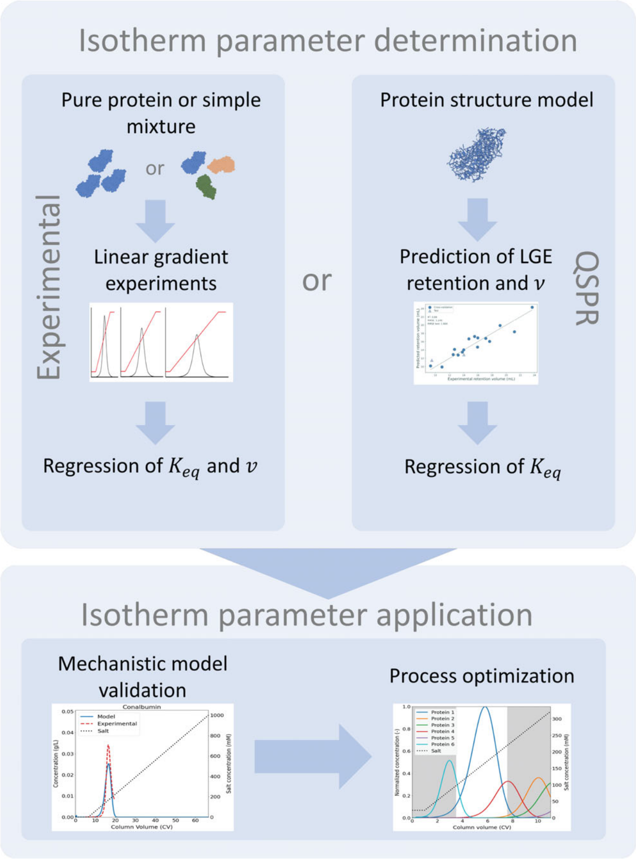


FIGURE 1 Overview of the experimental-based method and the QSPR-based method. Both methods can be used to determine the adsorption isotherm parameters that can be used in the mechanistic model for process optimization purposes. The equilibrium constant is denoted by K_{eq} and the stoichiometric coefficient of salt counter ions with v.

optimization outcome. An overview of the experimental-based and QSPR-based strategy is shown in Figure 1. This strategy can be used to determine the operating window of CPPs in early stage process development, showing the potential applicability for industry. Combining these modeling techniques together with an optimization software reduces the experimental effort for overall process development time significantly. Previous research mostly used pure components to perform the linear gradient experiments (LGE), however the availability of pure components is limited in biopharmaceutical industry. Therefore, performing LGE with complex protein mixtures would offer significant advantages. So far, only Buyel et al. applied QSPR modeling to a crude mixture of plant extracts to predict elution conditions for ion exchange and mixed mode chromatography separations.[33] Here, we performed LGE for five different gradient lengths and four pHs applied to two mixtures of each three proteins. Performing the experiments with protein mixtures instead of each protein individually, reduces the total LGE from 30 to 10 experiments. We developed QSPR models for predicting the retention volumes and characteristic charges. These predicted QSPR parameters were used to obtain the equilibrium constants. The multiscale model was validated for an unseen protein, which was excluded from the QSPR training and testing data. Finally, we compared the influence of parameter uncertainties on the optimization outcome by using experimental and QSPR predicted parameters.

2 | MATERIALS AND METHODS

2.1 | Experimental part

2.1.1 | Materials and equipment

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

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

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

2.1.2 | Linear gradient column experiments

LGE were conducted at various pH values (pH 3.5, 4.3, 5.0, and 7.0) for five gradient lengths: 20, 30, 40, 60, and 80 column volumes (CV). For every pH a different running buffer was needed, citric acid monohydrate (pH 3.5, 20 mM), sodium acetate trihydrate (pH 4.3 and 5.0, 50 mM), and sodium phosphate monobasic dihydrate (pH 7.0, 50 mM). The elution buffer is the same as the running buffer for that respective pH with the addition of 1 M sodium chloride. The pH-values were selected to theoretically favor a positive net charge for most proteins, and therefore anticipating their binding to the CEX resin. The chromatographic column experiments were performed on an ÄKTA pure system (Cytiva Life Sciences, USA) with UNICORN version 7.5 software, with a flowrate of 1 mL/min, and measuring UV absorbance at 230, 280, and 400 nm wavelength. The column characteristics are given in Table 1, more information on the characterization

TABLE 1 Column characteristics for HiTrap SP FF column.

Parameter	Value	Unit
Column volume	0.97	mL
Column diameter[a]	0.70	cm
Bed height[a]	2.50	cm
Maximum pressure[a]	2.0	MPa
Ionic capacity[44]	800	mM
Particle size[a]	90	µm
Pore diameter[45]	54	nm
Cross sectional area	0.39	cm^2
System dead volume (V_{dead})	0.34	mL
Total porosity (ε_t)	0.918	-
Extraparticle porosity (ε_b)	0.298	-
Intraparticle porosity (ε_p)	0.887	-
System dwell volume (V_{dwell})	1.09	mL

[a] Manufacturer.

methods can be found in Appendix A. During the chromatography runs, 1 mL samples were collected using a fraction collector. These samples were additionally analyzed with a Dionex UPLC system using Chromeleon Chromatography Data System version 7 software, measuring UV absorbance at 230, 280, and 400 nm wavelength. The UPLC-running buffer was a 100 mM sodium phosphate monobasic dihydrate with a pH of 6.8. A flowrate of 0.1 mL/min and analysis time of 40 min was applied. The SEC-UPLC analysis enabled the identification of the peaks obtained during the LGE's with their corresponding proteins. However, the protein mixture was divided into two groups, as some proteins with similar characteristics were indistinguishable in the SEC-UPLC analysis. Group 1 consisted of RNase, cytochrome c, conalbumin, and group 2 of chymotrypsinogen, lysozyme, and albumin. Both multi-component mixtures contained 0.8 mg/mL of each protein.

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

2.2 | Chromatographic MM

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

$$\frac{\partial C_i}{\partial t} + F\frac{\partial q_i}{\partial t} = -u\frac{\partial C_i}{\partial x} + D_{L,i}\frac{\partial^2 C_i}{\partial x^2}, \tag{1}$$

$$\frac{\partial q_i}{\partial t} = k_{ov,i}\left(C_i - C^*_{eq,i}\right), \tag{2}$$

$$k_{ov,i} = \left[\frac{d_p}{6k_{f,i}} + \frac{d_p^2}{60\varepsilon_p D_{p,i}}\right]^{-1}, \tag{3}$$

where the concentration in the liquid phase is represented by C_i and in the solid phase with q_i, in which subscript i denotes the protein component (Equation 1 and 2). The liquid phase concentration at equilibrium is denoted by $C^*_{eq,i}$. The phase ratio is equal to $F = (1-\varepsilon_b)/\varepsilon_b$, where ε_b is the bed porosity. Time and space are indicated by t and x respectively. u is the mobile phase interstitial velocity and D_L is the axial dispersion coefficient. The overall mass transfer coefficient, $k_{ov,i}$, is defined as the combined result of both the separate film mass transfer resistance and the mass transfer resistance within the pores.[47] In Equation 3, the particle diameter is denoted by d_p, the intraparticle porosity by ε_p, and the effective pore diffusivity coefficient by D_p. The effective pore diffusivity (Equation 4) is described according to Fick's law and calculated as

$$D_p = \frac{\varepsilon_p D_f}{\tau}\psi, \tag{4}$$

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

We employed the linear multicomponent mixed-mode isotherm (Equation 5), developed by Nfor et al., to determine the equilibrium liquid phase concentration as[52]

$$\frac{q_i}{C^*_{eq,i}} = K_{eq,i}\Lambda^{(v_i+n_i)}(z_s c_s)^{-v_i} c_v{}^{-n_i}\gamma_{i,} \tag{5}$$

where the equilibrium constant, $K_{eq,i}$, quantifies the strength of the interaction between the protein and the stationary phase. Λ is the ligand density or ionic capacity of the concerned resin, z_s is the charge of the salt counter ion, c_s is the salt concentration in the liquid phase, and c_v is the molarity of the solution in the pore volume. The stoichiometric coefficient of salt counter ions is denoted by v_i, determined by $v_i = z_p/z_s$, in which z_p is the effective binding charge of the protein. For monovalent counter-ions, the charge equals one ($z_s = 1$), for example Na^+ in the sodium chloride elution buffer. In this work, only the ion-exchange part of the mixed-mode isotherm is used, therefore hydrophobic interaction stoichiometric coefficient (n_i) will be equal to zero. The activity coefficient (γ) of the protein solution can be calculated via Equation 6 as

$$\gamma_i = e^{K_{s,i}c_s + K_{p,i}C_i}, \tag{6}$$

where K_s is the salt-protein interaction constant and K_p the protein–protein interaction constant. In the linear range of adsorption, the protein concentrations are low and protein–protein interactions are expected to be minimal, therefore K_p becomes insignificant and can be neglected.[53,54] Because of the low salting-out effects, the K_s also becomes negligible.[53] Subsequently, incorporating the assumptions for this work, the linear multicomponent mixed-mode isotherm is reformulated in Equation 7 as

$$\frac{q_i}{C^*_{eq,i}} = K_{eq,i}\Lambda^{v_i}(z_s c_s)^{-v_i}. \tag{7}$$

2.3 | Procedure to determine adsorption isotherm parameters

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

$$V_R = V_{R,0} - V_m - V_D - \frac{V_{inj}}{2}, \tag{8}$$

where V_R is the peak retention volume, V_m is the column void volume, determined by dextran pulse, and V_D is the system's dwell and dead volume (Equation 8), details can be found in Appendix A. The injection volume is denoted by V_{inj}, half of this volume needs to be subtracted.[55]

The regression formula of Shukla et al.,[56] (Equation 9) adapted from Parente and Wetlaufer,[57] was used to obtain the equilibrium constant (K_{eq}) and the characteristic charge (v) for each protein as follows:

$$V_R = \left(\left(C_{s,0}^{v+1} + \frac{V_m K_{eq} F \Lambda^v (v+1) * (C_{s,f} - C_{s,0})}{V_G}\right)^{\frac{1}{v+1}} - C_{s,0}\right) \times \frac{V_G}{C_{s,f} - C_{s,0}}, \tag{9}$$

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

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

Specifically at pH 5.0, Cytochrome c and RNase co-eluted. The absorbance and respective calibration lines of cytochrome c at 400 and 280 nm were used to trace back the RNase peak. Moreover, at pH 4.3, albumin and chymotrypsinogen co-eluted. However, from the SEC-UPLC analysis it was observed that albumin eluted later compared to the UV peak detected by the UNICORN software. Therefore, the peak retention volumes for albumin at pH 4.3 were determined by analyzing the concentrations by SEC-UPLC in the 1 mL fractions obtained from the LGE. Albumin peak areas obtained from the SEC-UPLC were used to fit a third degree polynomial function representing the retention volume as the maximum.

2.4 | QSPR model

2.4.1 | Structure preparation and descriptor calculation

For each protein, the respective models, listed in Table 2, were obtained from the protein data bank,[58] specific entry selection was performed based on resolution and coverage. Duplicate chains were removed from each structural model using pdb-tools[59] to yield monomer representations. The side chain pKa of titratable residues were predicted using PROPKA3.0[60] allowing for more accurate charge calculations with respect to pH. Protein features at pH 3.5, 4.3, 5.0, and 7.0 were calculated using our open-source software package prodes,

TABLE 2 Overview of the protein characteristics and the protein data bank (PDB) entry used for calculations.

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

[a] Estimations were performed using the open-source QSPR tool.

available at https://doi.org/10.5281/zenodo.10369949, using the default settings, only supplying the pKa estimations.[42] Visualization of protein structures was performed using UCSF-Chimera.[61]

2.4.2 | QSPR model training

For predicting the protein retention volumes and adsorption isotherm parameters, multi linear regression (MLR) models were trained. The prediction of conalbumin was removed from the dataset prior to train-test splitting to eliminate all bias. To find an accurate predictive MLR model, series of filter thresholds were screened by testing a range of feature-feature correlation filters (Pearson correlations of 0.8, 0.9, and 0.99). Followed by feature-observation correlations filtering, maintaining a predefined percentage of features (10% to 100% in 10% increments). Feature selection was performed by sequential forward selection. Final models were selected based on the cross-validated R^2 and test set Root Mean Square Error (RMSE), which should be close to the cross-validation RMSE to ensure model robustness. Feature importance was assessed by analysis of the regression coefficient and the influence of feature permutation. For the prediction of the unknown conalbumin, the confidence interval was calculated via Equation 10 as

$$\hat{y}_h \pm t_{\left(1-\frac{\alpha}{2},n-p\right)} \times \sqrt{MSE\left(1+X_h^T\left(X^TX\right)^{-1}X_h\right)}, \tag{10}$$

where $\hat{y}_h$ is the predicted value, $t_{\left(1-\frac{\alpha}{2},n-p\right)}$ is the "t-multiplier," X and X_h are the feature matrices of the training set and the value to be predicted. The mean squared error (MSE) is calculated via Equation 11 as

$$MSE = \frac{1}{n}\sum_{i}^{n}\left(y_i-\hat{y}_i\right)^2. \tag{11}$$

2.5 | Optimization

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

$$\min f(x) = 2\times(100-yield(x)) + 1\times(100-purity(x)). \tag{12}$$

$$s.t.\ h(x)=0, \tag{13}$$

$$0 \le x \le 1, \tag{14}$$

where the objective function, $f(x)$, is minimized (Equation 12). The equality equations, such as the mass balances and equilibrium relations, need to be satisfied (Equation 13). Moreover, variables (x) were normalized for more efficient optimization purposes (Equation 14). Four variables were chosen namely, the initial and final salt concentrations, and the lower and upper cut points. The weights of the objective function were chosen to reflect a capture step to be optimized, hence removing most of the bulk impurities and preventing losing product material.

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

3 | RESULTS AND DISCUSSION

3.1 | Linear gradient experiments

3.1.1 | Determining the retention volume

LGE's were conducted for two protein mixtures at four pH values (pH 3.5, 4.3, 5.0, and 7.0) and various gradient lengths (20, 30, 40, 60, and 80 CV), as described in the experimental Section 2.1. The elution order of the proteins was identified by SEC-UPLC analysis for each pH, to determine single peak retention volumes. The results for the 20 CV LGE are shown in Figure 2. As expected, a downward trend for

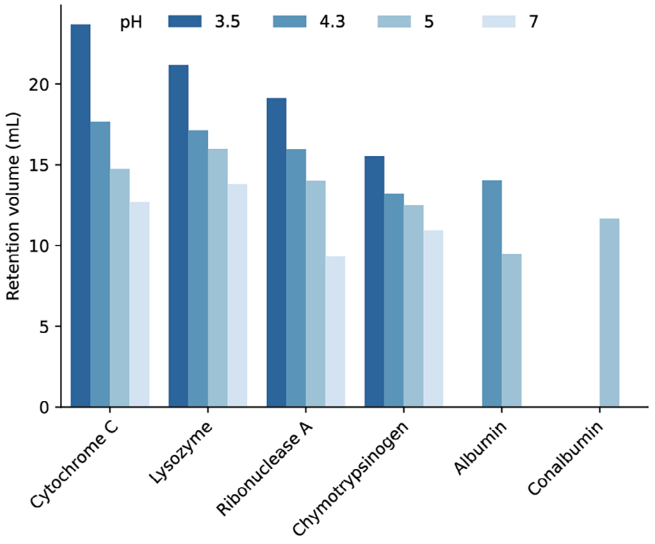


FIGURE 2 Peak retention volumes (mL, *y*-axis) given for each protein (*x*-axis) at each pH (bars). These retention volumes are from the 20 CV gradient length using a HiTrap SP FF column, 1 CV is equal to 0.97 mL.

TABLE 3 Regressed adsorption isotherm parameters, the characteristic charge and the equilibrium constant, for each protein at each pH. The standard deviation is indicated with number after ± sign.

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

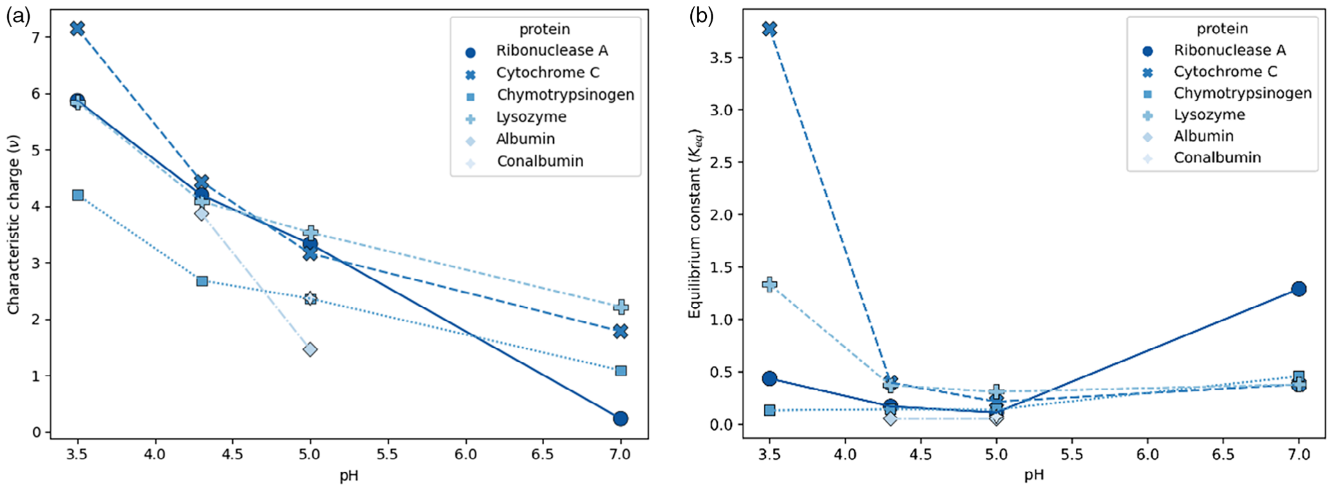


FIGURE 3 Trendlines between the (a) characteristic charge (y-axis) and (b) the equilibrium constant (y-axis), and the pH value (x-axis) for each protein.

the retention is observed when increasing the pH. No correlation between isoelectric point (PI) and retention was observed. Although cytochrome c, lysozyme, RNase, and chymotrypsinogen elute in the order of descending pI (9.60, 9.20, 8.29, and 8.13, respectively) at pH 3.5. No retention volume for albumin and conalbumin (pI of 5.49 and 6.62, respectively) was determined as these proteins did not elute during the salt gradient, showing greater affinity for the column, which is in accordance with Yang et al.[32]

3.1.2 | Regression of adsorption isotherm parameters

The corrected retention volumes, according to Equation 8, were used to regress K_{eq} and v using Equation 9. The regression parameters for each protein at each pH are shown in Table 3. The regression plots of each protein at each pH are provided in Appendix B, all fits achieved an R^2 close to one and RMSE values varied between 0.002 and 0.22.

From Table 3 it can observed that the characteristic charge, v, varied between 1% and 6% of the regressed parameter value and the standard deviation values of the equilibrium constant, K_{eq}, varied between 7% and 25%. Figure 3a shows that the characteristic charge decreases with increasing pH for all proteins with multiple data points. This is due to the protonation of amino acids, which results in a higher net protein charge at lower pH values. A higher net charge results in more available binding sites to interact with the resin. However, no general trend can be observed between the equilibrium constant and the pH (Figure 3b). The equilibrium constant of cytochrome c and lysozyme decreases rapidly from pH 3.5 to pH 4.3. However, at pH 7.0 K_{eq} increases again for RNase, chymotrypsinogen, lysozyme, and cytochome c (increase of 1.19, 0.26, 0.23, and 0.23, respectively). Similar findings were reported by Yang et al.,[32] and the regressed parameters are in the same order of magnitude as reported in literature.[32,44] In general, a higher equilibrium constant indicates a stronger binding affinity towards the resin, and therefore eluting later during the salt gradient. The same trend can be observed for the majority of proteins, see Table 3 and Figure 3. Not all proteins follow this trend, such as chymotrypsinogen, cytochrome c, and lysozyme relative to RNase (pH 7.0), and albumin relative to chymotrypsinogen (pH 4.3). These proteins elute at a later moment while having a lower equilibrium constant than the proteins eluting at an earlier moment. Though, the characteristic charge value is higher for these proteins with a lower equilibrium constant. Eventually, it is the combination of these two parameter values that determines the protein's elution moment.

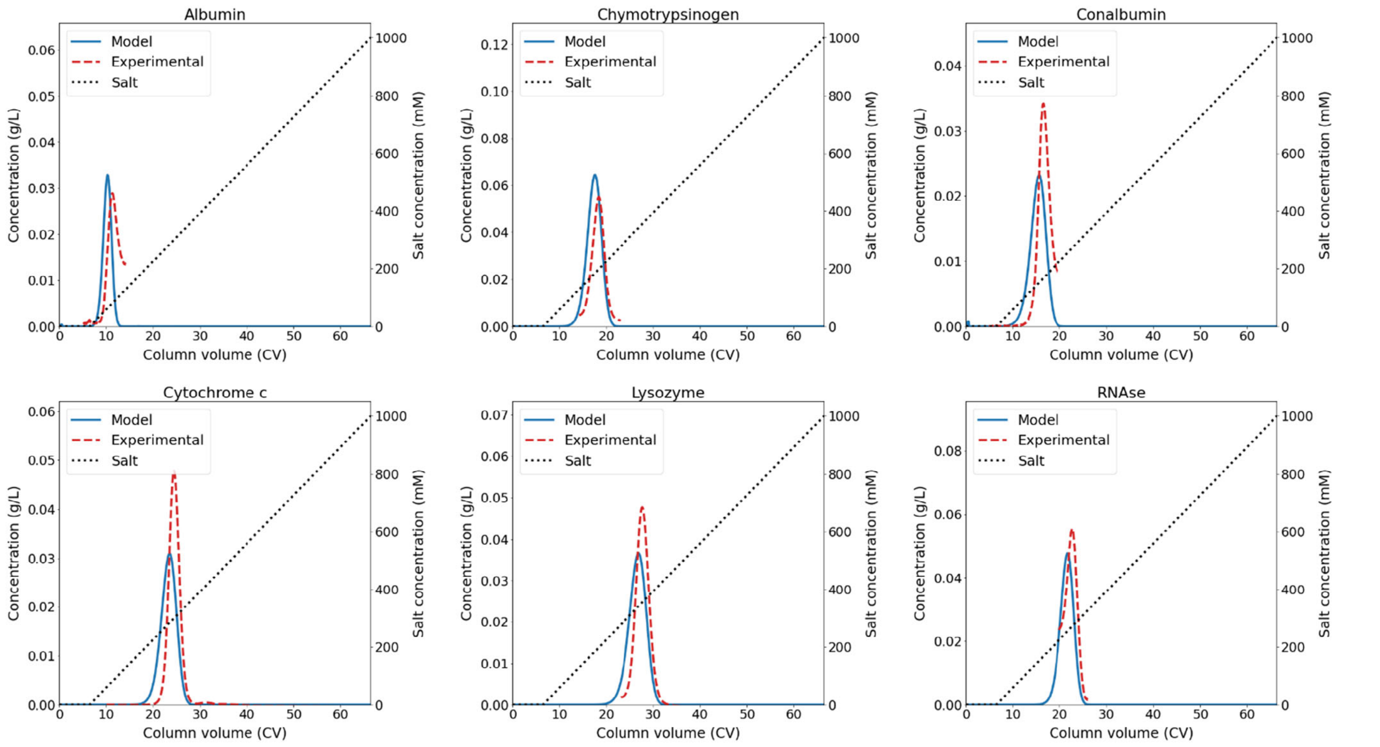


FIGURE 4 Chromatographic mechanistic model validation for gradient length of 60 CV, equal to 58.2 mL, at a pH of 5.0. The blue line indicates the MM predicted concentration of the protein, while the red dotted line indicates the experimental concentration. The black dotted line indicates the salt concentration. The initial concentrations are albumin: 0.24 mg/mL, chymotrypsinogen: 0.80 mg/mL, conalbumin: 0.31 mg/mL, cytochrome c: 0.41 mg/mL, lysozyme: 0.55 mg/mL, and RNase: 0.56 mg/mL.

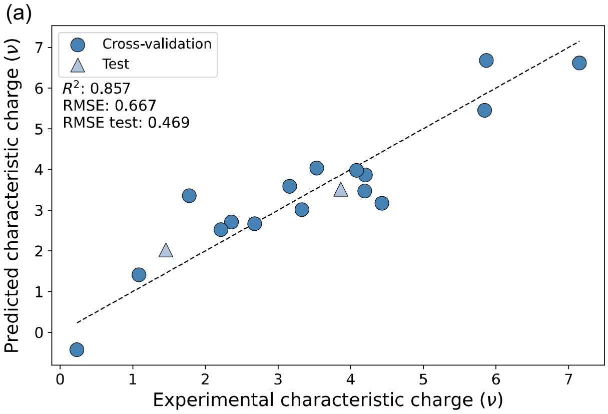
3.1.3 | Chromatographic MM validation

The chromatographic MM was validated for the gradient length of 60 CV, for pH 5.0 and 7.0. The results of pH 5.0 are shown in Figure 4, and of pH 7.0 in Appendix C. The calibration lines convert the UV absorbance to concentration, these can be found in Appendix D. As the experiments were performed in two mixtures of each three proteins, only parts of the peaks corresponding to a certain protein were used to avoid pollution of the peak by another component. In this way, the validation of each protein with the MM could be clearly evaluated.

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

3.2 | Quantitative structure property relationship modeling

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



(b)

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

FIGURE 5 Prediction of characteristic charge. (a) Model validation of the regression model trained to predict ν where the circles represent the leave-one-out cross-validation and the triangles the test set. (b) Overview of the selected features with the regression coefficient and the cross-validated R^2 after feature permutation.

hydrophobicity projections, were calculated using the open source software prodes.[42]

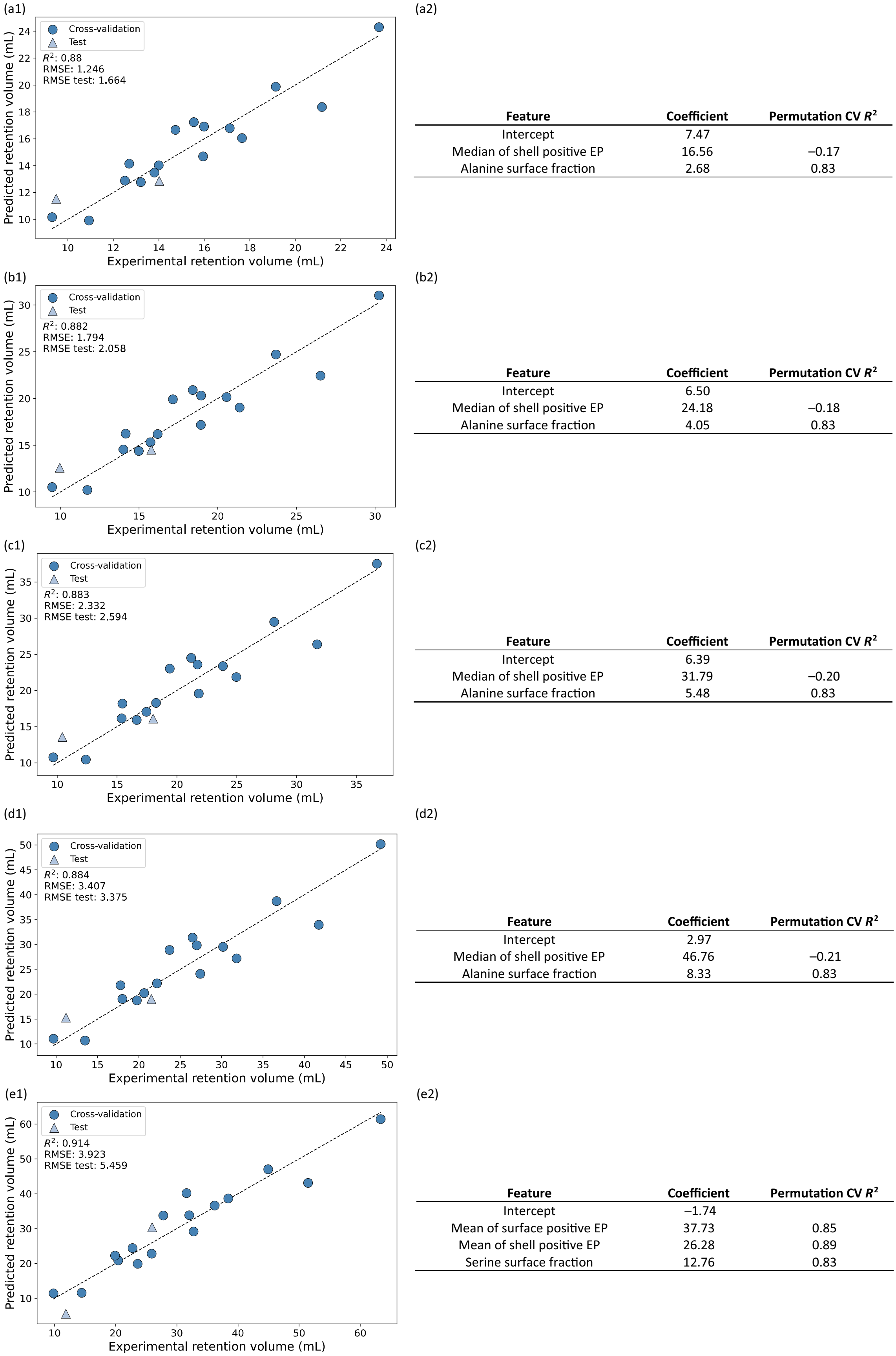
3.2.1 | Characteristic charge

For the prediction of the characteristic charge, a MLR was trained. To avoid overfitting, a ratio of five observations to one feature should be maintained.[62] Meaning only a maximum of three features should be used in the model. To select the specific features, a redundancy filter, removing features with a Pearson correlation of >0.99 to other features, was applied. A second filter step was performed removing 40% of the features with lowest correlation to the characteristic charge. From the remaining features, sequential forward selection was performed to select the best features. A model with high accuracy (cross-validated R^2 of 0.86 and RMSE of 0.67) was obtained using only two features (Figure 5). As would be expected, the most important feature was related to the electrostatic potential (EP) of the protein surface. More specifically, the maximal found surface EP. The regression coefficient of this feature was found to be 8 and permutation of the feature would result in a model not capable of predicting ν (Figure 5b). The second feature that was selected is the trimean of the negative hydrophobicity potential. This feature is less important as the regression coefficient is 1.5 and permutation results in a model with a cross-validated R^2 of 0.8. The positive regression coefficient for the second feature suggests that increasing the hydrophilicity reduces the characteristic charge. There is the possibility however, that this feature captures the titratable amino acid content on the surface, as amino acids contributing to a negative hydrophobicity are predominantly titratable. At this point, we have been unable to confirm this.

Applying the same approach to build a QSPR model for K_{eq} did not yield sufficiently accurate models. With the current dataset, the best performing models yielded only a R^2 of 0.58 (data not shown). While ν has direct physical implications, by representing the number of charge interactions between the resin and protein, K_{eq} is lacking this.[44,63] The equilibrium constant represents all phenomena contributing to adsorption. As observed in Figure 3, ν shows a clear negative trend with increasing pH, this trend is lacking for K_{eq}. It is thought that the current dataset-size is the main limitation as more features might be required to capture the complex relation. To overcome this challenge, increasing the dataset-size would result in a model trained over a greater range of property values, while also allowing an increase of the number of used features without loss of robustness.[24,32]

3.2.2 | Retention volumes

Alternatively, the K_{eq} can be obtained from the regression as performed in 3.1.2 for experimental data. To achieve this, a MLR model for each LGE was trained (Figure 6). The best performing models were obtained using a feature-property correlation filter, removing 40% of the features with the lowest correlation, prior to the feature selection. The trained MLR models, for each LGE, all achieved a cross-validated R^2 of at least 0.88. For all models, the most important feature relates to the EP. More specifically, the median shell positive EP was most important for the four lower gradient lengths (20, 30, 40, and 60 CV). This feature describes the positive EP on the exterior of the protein by projecting each charge onto a plane that represents the resin. For the calculation of the shell, a total of 120 planes surround the protein, in this way representing different binding orientations. Opposed to mapping the EP onto solvent accessible surface, this method considers the distance through the solvent, penalizing protein surface within pockets. The surface fraction of alanine was the second feature selected. Alanine is a small hydrophobic amino acid, therefore this feature implicitly describes the surface hydrophobicity. The positive regression coefficient fitted for this feature indicates that a greater alanine content, and thus higher surface hydrophobicity, results in a higher retention volume. This can be explained by the salting-out effect of the Na^+ ions used



(a2)

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

(b2)

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

(c2)

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

(d2)

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

(e2)

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

FIGURE 6 Legend on next page.

TABLE 4 Predicted properties for conalbumin at pH 5.0.

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

during the gradient elution, resulting in hydrophobic interactions with the resin material.[43]

For the 80 CV retention MLR model, the following features were selected: shell positive EP mean, solvent accessible surface positive EP mean, and the serine surface fraction. The feature combination yielded an accurate model with a cross-validated R^2 of 0.91 and a RMSE of 3.9 (Figure 6e). For the prediction of the test set, it is observed that the point at the lower end of the retention data is under predicted, compared with being over predicted in all other models. While the EP remains the most important in the model, different features were selected during the sequential feature selection. This is due to the fact that there is no exact linear relationship between gradient length and retention, as can be most notably observed at pH 7.0 in Appendix B. While the Mean and Median of the shell EP are similar, the slight differences in the features resulted in the selection of the mean. Both the mean of surface positive EP and mean of shell positive EP are important features, with regression coefficients of 37.73 and 26.28, respectively. This importance is not reflected by the permutation models, as both features describe the positive EP, collinearity allows for compensation for a loss of one of the features. However, it is essential to maintain both features to accurately predict the test set, as removing one of them results in less accurate retention estimates (data not shown). Surprisingly, the surface area fraction of serine has a positive regression coefficient, like the alanine surface fraction in the other four models. In contrast to alanine, serine is a hydrophilic residue. However, the positive regression coefficient indicates increasing retention with higher serine content on the surface, which contradicts the hypothesis for alanine selection for the previous four models. The reason behind the selection of serine in this model is currently unknown. While the models show difficulty in predicting the change of elution order switch of lysozyme and cytochrome c for pH 4.3 and 5, a sharper decrease in retention for cytochrome c compared with lysozyme is predicted (data not shown). Still all models show good accuracy during both cross-validation and model testing, providing high confidence in model robustness.

3.2.3 | Property prediction of conalbumin at pH 5

To demonstrate the true predictive capabilities of the trained QSPR models for the prediction of retention volumes and isotherm parameters, conalbumin was completely removed from the dataset prior to the train test splitting. This allowed to minimize the bias applied on the model selection. For the prediction of the retention volumes, the error of prediction increased with increasing gradient lengths (Table 4). The range of observed retention volumes rises along with the gradient lengths, likewise, the 95% confidence interval increases. Nevertheless, the effect of increasing the gradient length was captured correctly, having a maximal error of about 2 mL in retention volume, which falls within the 95% confidence interval. The characteristic charge was predicted with an error of 0.5, complying with the 95% confidence interval. Unfortunately, as no robust and accurate QSPR model for the K_{eq} could be trained with the current dataset, no direct prediction could be made. Therefore, we applied an alternative method, the predicted retention volumes and characteristic charge were used to regress the K_{eq} using the regression formula, similar to the experimental data method as shown in 3.1.2. regression of adsorption isotherm parameters. The K_{eq} obtained was 0.028 ± 0.006, which is lower than the K_{eq} of 0.078 ± 0.012 obtained by regression of the experimental data. This is due to the higher predicted ν by the QSPR model. Validation of the predicted parameters showed an accurate prediction of the conalbumin elution using a 60 CV gradient length (Figure 7). Both peak maximum and peak shape are simulated accurately. The difference in the peak retention volume is very small, 0.12 CV, which is 0.2% difference relative to the gradient length (60 CV). The peak concentration differs by 0.009 g/L, which is 2.85% relative to the initial concentration, and the difference in the peak width at half of the peak maximum is only 1.0% relative to the gradient length (60 CV). Interestingly, the predicted parameters seem to better describe the retention profile compared to the parameters obtained from the experimental LGE, which was an average peak retention difference of 1.53% and an average peak width difference of 1.35% with respect to the gradient length (60 CV).

FIGURE 6 Prediction of protein retention at different salt gradient lengths where the circles represent the leave-one-out cross-validation and the triangles the test set. (a–e) show the validation and test of the prediction of the retention volume while applying a salt gradient of 20, 30, 40, 60, and 80 column volumes (CVs), respectively. One CV equals 0.97 mL (Table 1). The tables right of the plots show the feature coefficients and the effect of feature permutation on the cross validated R^2.

3.3 | Comparing optimization results between experimentally and QSPR-based methods

For the test protein, conalbumin at pH 5.0, both adsorption isotherm parameters, K_{eq} and v, were determined via two methods. The first method regressed the adsorption isotherm parameters from the LGE data directly, hence LGE are needed to perform this method. While the second method involved the QSPR approach, which, after being properly trained, requires the protein-structure to determine the v and the retention volumes. These two QSPR models were then used to regress the K_{eq} using the regression formula (Equation 9).

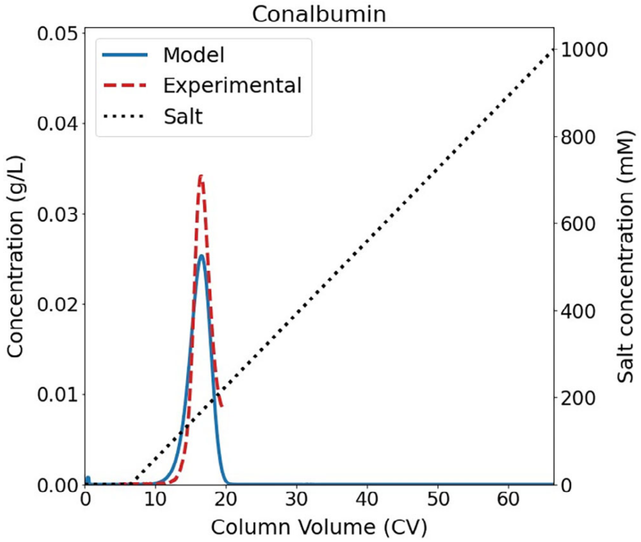


FIGURE 7 Chromatographic mechanistic model validation of conalbumin for gradient length of 60 CV, equal to 58.2 mL, at a pH of 5.0 using the predicted isotherm parameters. Blue line indicates the MM predicted concentration of the protein, while the red dotted line indicates the experimental concentration. The black dotted line indicates the salt concentration.

The capture step was optimized to separate conalbumin from the other proteins, prioritizing yield over purity, utilizing the adsorption isotherm parameters determined from both methods. This optimization aimed to assess the agreement between the optimized capture step and the parameters obtained from both methods. The resulting capture steps for both methods are depicted in Figure 8. The optimized variables (e.g., lower and uppercut points and the initial and final salt concentration) show comparability. The differences in both cut points are within 3.3%, and the deviation for both initial and final salt concentration is around 10 mM, approximately 3% relative to the final salt concentration (330 mM). The obtained purity only differs 0.3% and the yield 1.2% between both methods. These results demonstrate that, in this case study, it was viable to optimize the CEX capture step based solely on knowledge of the protein structure.

In the next part, we assessed the effect of the adsorption isotherm parameter uncertainties on the optimization outcome. We aimed to determine if variations within the standard deviation of the parameters would result in different optimal values. For both methods, numerous sample points were generated for each isotherm parameter, covering a range within their respective standard deviation. Subsequently, these sample points were used in the optimization case study. First, the consistency of the optimization case study was

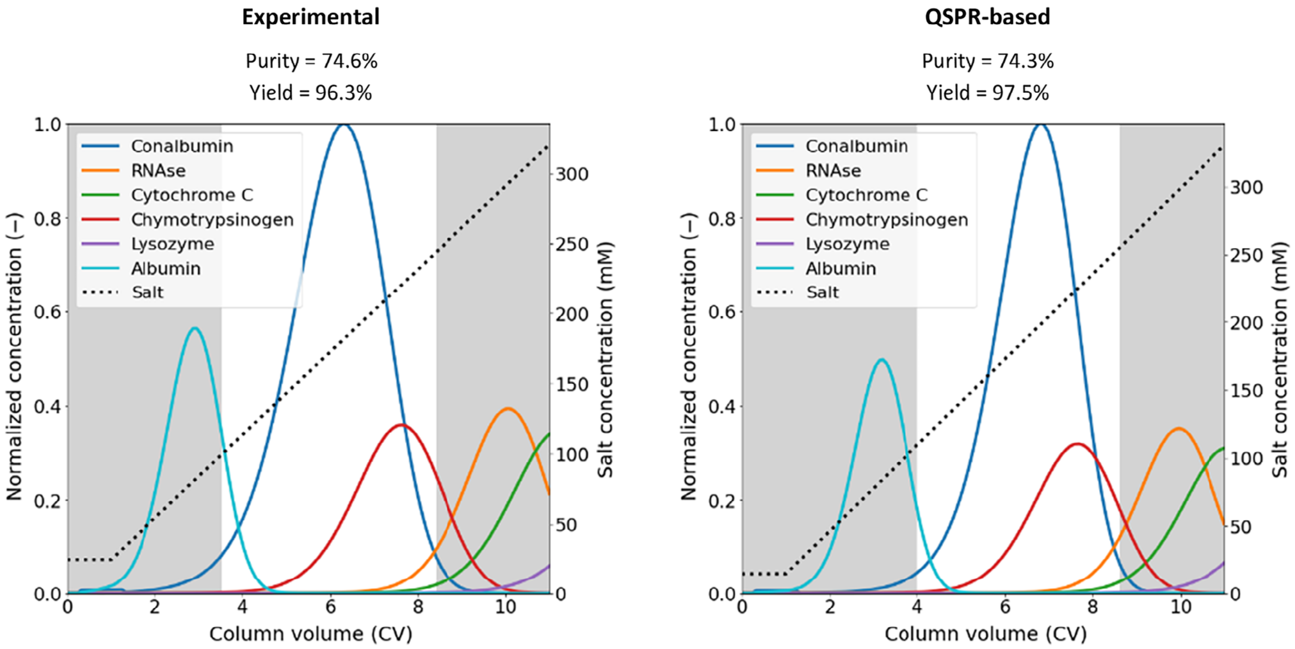


FIGURE 8 Optimized capture step using the mechanistic model, where the optimization results of the experimental-based (left) and QSPR-based (right) method are compared. Left: Experimental-based method, the adsorption isotherm parameters were regressed directly from the LGE. K_{eq} 0.071 and $v = 2.37$, lower and uppercut point are 7.7% and 91.2%, respectively. The initial and final salt concentration are 24.5 mM and 320.6 mM respectively. Right: QSPR-based method, the retention volumes and v are obtained from QSPR models, followed by using these QSPR models to regress the K_{eq} parameter. $K_{eq} = 0.028$ and $v = 3.05$, lower and uppercut points are 4.4% and 91.7%, respectively. The initial and final salt concentration are 14.8 and 330.4 mM, respectively.

evaluated by running the same optimization five times, these results for both methods can be found in Appendix E. This consistency evaluation aimed to ensure there were no major deviations in results within the same optimization using identical parameters. Additionally, the minor deviations could be attributed to the optimization process itself. The optimized results for various combinations of K_{eq} and v, ranging within their respective standard deviation, are shown in Figure 9 for both methods. This includes the optimized variables, such as the lower and uppercut points and the initial and final salt concentrations, as well as the purity, and the yield.

In the experimental-based method, the standard deviations for both K_{eq} (0.071 ± 0.012) and v (2.37 ± 0.12) are relatively small, resulting in minimal variance in the optimized variables (Figure 9a1–f1,a2–f2, for variations in K_{eq} and ν, respectively). The lower and upper cut points have a maximum difference of 7% (Figure 9a,b). The initial salt concentration varies between 15 and 40 mM (Figure 9c1,c2), and the final salt concentration is found between 320 and 327 mM (Figure 9d1,d2). These results suggest that despite variations in the isotherm parameters, a consistent optimum is identified, and the optimized variables exhibit only minor variations. The impact on the yield is minimal, with only a 2% variation (Figure 9f1,f2). On the contrary, the effect on purity is more pronounced, fluctuating between 70% and 81%. The decrease in purity is primarily attributed to an increase in the K_{eq} (Figure 9e1), which is due to the greater relative standard deviation compared to ν.

For the QSPR-based method, the standard deviation of K_{eq} is small (0.028 ± 0.006). The randomly spread data indicates that there is no clear correlation between K_{eq} and the optimized variables (Figure 9a3–f3). However, the standard deviation of v is significantly larger (3.05 ± 1.4), this standard deviation was defined by the 95% confidence interval calculated by Equation 9. The large variation in v resulted in two identified optima, which is clearly observed in the shift of the final salt concentration (Figure 6d4). The first solution finds an optimal final salt concentration between 320 and 400 mM. The shift to the second optimal solution occurs when v is greater than 3.6, finding the final salt concentration at around 800 mM. Remarkably, both optimal final salt concentrations are close to the set boundaries. As the characteristic charge increases, the component is expected to elute at a higher salt concentration and thus at a later moment during the gradient. This results in a greater overlap between conalbumin and the other impurities. Such a shift was not observed for the initial salt concentration, where most optimal conditions were found between 10 and 30 mM (Figure 9c4). The effect of v is also reflected in the purity and the yield (Figure 9e4 and 9f4 respectively). Until v is 2.2, the purity is around 75% and the yield almost 100%, while above this value of v, the purity increases rapidly and the yield drops to about 95%. From this point, increasing v results in a decreasing purity and increasing yield. However, the range of the purity is broader, 50%–85% than that of the yield, which only fluctuates between 95% and 99%. This broader range in the purity is probably due to a combination of the shift in retention volume resulting from variation of ν, and the optimization function Equation 11. In the function, the yield is prioritized, representing a capture step optimization. Therefore, during challenging separation processes, the compromise on the yield is always less compared with purity. Changes in the optimization weights would result in a shift in priority between purity and yield that would translate to the selection of different cut points rather than initial and final salt concentrations. Despite the greater uncertainty in the determined v in the QSPR-method, only two optima were identified, and one of them corresponds to the optimum found in the experimental-based method.

Furthermore, this optimization approach is applicable for defining the operating window of certain variables. The method employed for varying the adsorption isotherm parameters can also be used to vary other variables and assess the optimized result. In this way, the initial process design space for CPP can be defined, which is part of the QbD concept.[64] The mechanistic modeling outcomes provide knowledge on the process, therefore the number of wet-lab experiments to define the real process design space can be reduced in comparison to performing a wet-lab DoE from scratch. For the QSPR-based method, no wet-lab experiments are needed to determine the adsorption isotherm parameters and therefore the total number of experiments are even more reduced compared to the experimental-based method. For a new protein, only the protein-structure is needed to perform this optimization and make an estimation of the operating window for each optimizing variable. To illustrate, using the results from the QSPR-based method in this study, we can already narrow down the number of wet-lab DoE required to define the process design space. The final salt concentration only has to be evaluated around two main values (e.g., around 320 mM and 800 mM, see Figure 9d4), while only one point of the initial salt concentration has to be assessed (e.g., 20 mM). Ultimately, the QSPR-based method offers an added advantage by allowing the incorporation of additional data over time. This not only enhances the model's accuracy, but also enables the application to other process designs, provided that the same conditions are used.

Currently, only the linear part of the isotherm is considered as only low loading conditions are investigated. Prediction of the parameters describing the non-linear part of the isotherm as well as competitive behavior would make the method more complete. Nevertheless, for the purpose of preselection of conditions for early stage process design, considering only the linear behavior should be sufficient. Additionally, the amount of available training data might pose a bottleneck, like the prediction of the K_{eq} presented in this work. Even though the predictions of the retention volumes and characteristic charge showed high accuracy, increasing the variety of proteins would make the models more robust. To extend this method to more complex mixtures, such as host cell lysates, several challenges should be overcome. While a similar fractionation approach to convolute single peaks can be used for a complex mixture, more accurate analytical methods are required for protein identification. Potentially, mass spectrometry methods allow the required resolution providing relative protein abundances. Additionally, protein interactions and complex formation should be taken into account during the QSPR modeling. Co-elution has already been studied extensively, and recently Panikulam et al., published a novel method to describe co-elution mechanisms for

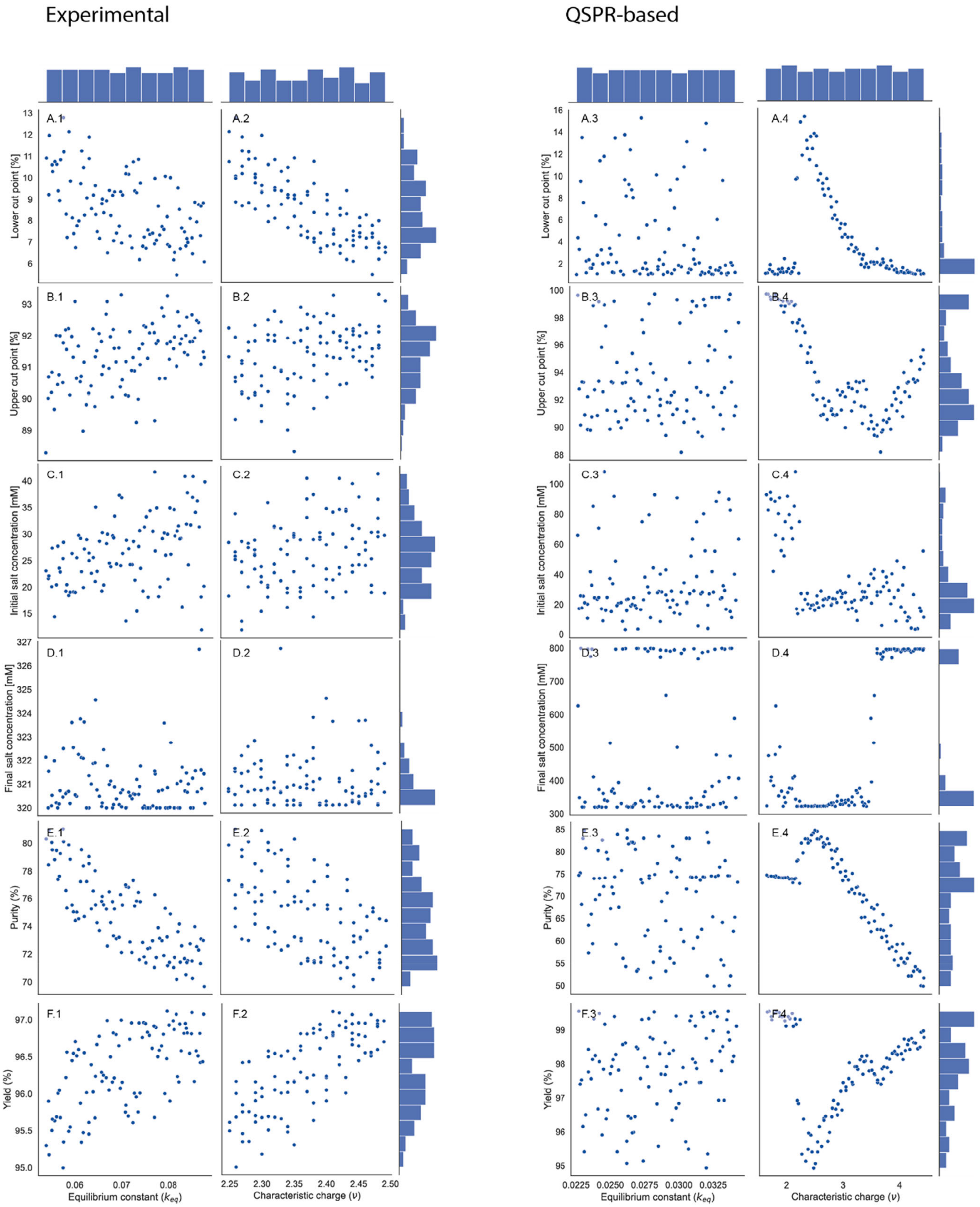


FIGURE 9 Joint plots of scatter and histogram plots between the adsorption isotherm parameters (e.g., the characteristic charge and the equilibrium constant) and the optimized variables (e.g., lower and upper cut point and the initial and final salt concentrations, and the purity and the yield). Left: Experimental-based method results. Right: QSPR-based method results.

protein A chromatography.[65] Further maturation and combination of these methods would allow better integration and application for complex mixtures.

4 | CONCLUSION

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

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

We successfully trained MLR-QSPR models with a minimum cross-validated R^2 of 0.88, even with a limited dataset composed of only five different proteins measured at four pH values. The MLR-QSPR models for predicting the characteristic charge and the retention volumes can be used to regress the equilibrium constant using the regression formula. A good agreement was obtained for the MM validation for an unseen protein, conalbumin, showing only 0.2% retention peak difference with respect to the gradient length.

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

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

AUTHOR CONTRIBUTIONS

Daphne Keulen: Conceptualization; methodology; software; writing – original draft; writing – review and editing; visualization; validation; investigation. **Tim Neijenhuis:** Conceptualization; investigation; writing – original draft; writing – review and editing; visualization; validation; methodology; software. **Adamantia Lazopoulou:** Data curation; methodology; investigation; writing – review and editing. **Roxana Disela:** Writing – review and editing; methodology. **Geoffroy Geldhof:** Supervision; writing – review and editing. **Olivier Le Bussy:** Supervision; writing – review and editing. **Marieke E. Klijn:** Writing – review and editing; supervision. **Marcel Ottens:** Conceptualization; funding acquisition; writing – review and editing; supervision.

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 for their valuable input.

Adamantia-Maria Lazopoulou was supported by the Onassis Foundation – Scholarship ID: F ZR 031-1/2021-2022.

CONFLICT OF INTEREST STATEMENT

All authors have declared the following interests: Geoffroy Geldhof and Olivier Le Bussy are employees of the GSK group of companies. The other authors declare no conflict of interests.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

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

REFERENCES

1. Birch JR, Onakunle Y. *Biopharmaceutical Proteins: Opportunities and Challenges*. Methods and protocols; 2005:1-16.
2. Wen EP, Ellis R, Pujar NS. *Vaccine Development and Manufacturing*. Wiley; 2014.
3. Jagschies G, Lindskog E, Łacki K, Galliher P. Biopharmaceutical processing: development. *Design, and Implementation of Manufacturing Processes*. Elsevier; 2018:1-1275.
4. Kesik-Brodacka M. Progress in biopharmaceutical development. *Biotechnol Appl Biochem*. 2018;65(3):306-322. doi:10.1002/bab.1617

5. Kelley B. Developing therapeutic monoclonal antibodies at pandemic pace. *Nat Biotechnol*. 2020;38(5):540-545. doi:10.1038/s41587-020-0512-5
6. Łącki KM. Chapter 16 - introduction to preparative protein chromatography. In: Jagschies G, Lindskog E, Łącki K, Galliher P, eds. *Biopharmaceutical Processing*. Elsevier; 2018:319-366.
7. Keulen D, Geldhof G, Bussy OL, Pabst M, Ottens M. Recent advances to accelerate purification process development: a review with a focus on vaccines. *J Chromatogr A*. 2022;1676:463195. doi:10.1016/j.chroma.2022.463195
8. Hanke AT, Ottens M. Purifying biopharmaceuticals: knowledge-based chromatographic process development. *Trends Biotechnol*. 2014;32(4):210-220. doi:10.1016/j.tibtech.2014.02.001
9. Reinhardt IC, Oliveira DJC, Ring DDT. Current perspectives on the development of industry 4.0 in the pharmaceutical sector. *J Ind Inform Dermatol Int*. 2020;18:100131. doi:10.1016/j.jii.2020.100131
10. von Stosch M, Portela RMC, Varsakelis C. A roadmap to AI-driven in silico process development: bioprocessing 4.0 in practice. *Curr Opin Chem Eng*. 2021;33:100692. doi:10.1016/j.coche.2021.100692
11. Alosert H, Savery J, Rheaume J, et al. Data integrity within the biopharmaceutical sector in the era of industry 4.0. *Biotechnol J*. 2022;17(6):e2100609. doi:10.1002/biot.202100609
12. Narayanan H, Luna MF, von Stosch M, et al. Bioprocessing in the digital age: the role of process models. *Biotechnol J*. 2020;15(1):1900172. doi:10.1002/biot.201900172
13. Rathore AS. Quality by design (QbD)-based process development for purification of a biotherapeutic. *Trends Biotechnol*. 2016;34(5):358-370. doi:10.1016/j.tibtech.2016.01.003
14. FDA. PAT Guidance for Industry - A Framework for innovative Pharmaceutical Development, Manufacturing and Quality Assurance. www.fda.gov/regulatory-information/search-fda-guidance-documents/pat-framework-innovative-pharmaceutical-development-manufacturing-and-quality-assurance
15. ICH. ICH Harmonised Tripartite Guideline: Pharmaceutical Development Q8 (R2). presented at: ICH; 2009. https://www.ema.europa.eu/en/ich-q8-r2-pharmaceutical-development-scientific-guideline
16. Mollerup JM, Hansen TB, Kidal S, Staby A. Quality by design—thermodynamic modelling of chromatographic separation of proteins. *J Chromatogr A*. 2008;1177(2):200-206. doi:10.1016/j.chroma.2007.08.059
17. Shekhawat LK, Tiwari A, Yamamoto S, Rathore AS. An accelerated approach for mechanistic model based prediction of linear gradient elution ion-exchange chromatography of proteins. *J Chromatogr A*. 2022;1680:463423. doi:10.1016/j.chroma.2022.463423
18. Saleh D, Wang G, Müller B, et al. Straightforward method for calibration of mechanistic cation exchange chromatography models for industrial applications. *Biotechnol Prog*. 2020;36(4):e2984. doi:10.1002/btpr.2984
19. Kumar V, Lenhoff AM. Mechanistic modeling of preparative column chromatography for biotherapeutics. *Annu Rev Chem Biomol Eng*. 2020;11(1):235-255. doi:10.1146/annurev-chembioeng-102419-125430
20. Rischawy F, Saleh D, Hahn T, Oelmeier S, Spitz J, Kluters S. Good modeling practice for industrial chromatography: mechanistic modeling of ion exchange chromatography of a bispecific antibody. *Comput Chem Eng*. 2019;130:106532. doi:10.1016/j.compchemeng.2019.106532
21. Nfor BK, Ahamed T, Pinkse MWH, et al. Multi-dimensional fractionation and characterization of crude protein mixtures: toward establishment of a database of protein purification process development parameters. *Biotechnol Bioeng*. 2012;109(12):3070-3083. doi:10.1002/bit.24576
22. Close EJ, Salm JR, Bracewell DG, Sorensen E. A model based approach for identifying robust operating conditions for industrial chromatography with process variability. *Chem Eng Sci*. 2014;116:284-295. doi:10.1016/j.ces.2014.03.010
23. Disela R, Bussy OL, Geldhof G, Pabst M, Ottens M. Characterisation of the E. Coli HMS174 and BLR host cell proteome to guide purification process development. *Biotechnol J*. 2023;18(9):2300068. doi:10.1002/biot.202300068
24. Saleh D, Hess R, Ahlers-Hesse M, et al. A multiscale modeling method for therapeutic antibodies in ion exchange chromatography. *Biotechnol Bioeng*. 2023;120(1):125-138. doi:10.1002/bit.28258
25. Mazza CB, Sukumar N, Breneman CM, Cramer SM. Prediction of protein retention in ion-exchange systems using molecular descriptors obtained from crystal structure. *Anal Chem*. 2001;73(22):5457-5461. doi:10.1021/ac010797s
26. Breneman CM, Thompson TR, Rhem M, Dung M. Electron density modeling of large systems using the transferable atom equivalent method. *Comput Chem*. 1995;19(3):161-179. doi:10.1016/0097-8485(94)00052-G
27. Whitehead CE, Breneman CM, Sukumar N, Ryan MD. Transferable atom equivalent multicentered multipole expansion method. *J Comput Chem*. 2003;24(4):512-529. doi:10.1002/jcc.10240
28. Song M, Breneman CM, Bi J, et al. Prediction of protein retention times in anion-exchange chromatography systems using support vector regression. *J Chem Inf Comput Sci*. 2002;42(6):1347-1357. doi:10.1021/ci025580t
29. Ladiwala A, Rege K, Breneman CM, Cramer SM. Investigation of Mobile phase salt type effects on protein retention and selectivity in cation-exchange systems using quantitative structure retention relationship models. *Langmuir*. 2003;19(20):8443-8454. doi:10.1021/la0346651
30. Ladiwala A, Rege K, Breneman CM, Cramer SM. Prediction of adsorption isotherm parameters and chromatographic behavior in ion-exchange systems. *Proc Natl Acad Sci*. 2005;102(33):11710-11715. doi:10.1073/pnas.0408769102
31. Chen J, Cramer SM. Protein adsorption isotherm behavior in hydrophobic interaction chromatography. *J Chromatogr A*. 2007;1165(1):67-77. doi:10.1016/j.chroma.2007.07.038
32. Yang T, Sundling MC, Freed AS, Breneman CM, Cramer SM. Prediction of pH-dependent chromatographic behavior in ion-exchange systems. *Anal Chem*. 2007;79(23):8927-8939. doi:10.1021/ac071101j
33. Buyel JF, Woo JA, Cramer SM, 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*. 2013;1322:18-28. doi:10.1016/j.chroma.2013.10.076
34. Malmquist G, Nilsson UH, Norrman M, Skarp U, Strömgren M, Carredano E. Electrostatic calculations and quantitative protein retention models for ion exchange chromatography. *J Chromatogr A*. 2006;1115(1):164-186. doi:10.1016/j.chroma.2006.02.097
35. Hanke AT, Klijn ME, Verhaert PDEM, et al. Prediction of protein retention times in hydrophobic interaction chromatography by robust statistical characterization of their atomic-level surface properties. *Biotechnol Prog*. 2016;32(2):372-381. doi:10.1002/btpr.2219
36. Kittelmann J, Lang KMH, Ottens M, Hubbuch J. Orientation of monoclonal antibodies in ion-exchange chromatography: a predictive quantitative structure–activity relationship modeling approach. *J Chromatogr A*. 2017;1510:33-39. doi:10.1016/j.chroma.2017.06.047
37. Kittelmann J, Lang KMH, Ottens M, Hubbuch J. An orientation sensitive approach in biomolecule interaction quantitative structure–activity relationship modeling and its application in ion-exchange chromatography. *J Chromatogr A*. 2017;1482:48-56. doi:10.1016/j.chroma.2016.12.065
38. Robinson JR, Karkov HS, Woo JA, Krogh BO, Cramer SM. QSAR models for prediction of chromatographic behavior of homologous

fab variants. *Biotechnol Bioeng*. 2017;114(6):1231-1240. doi:10.1002/bit.26236

39. Hess R, Faessler J, Yun D, et al. Antibody sequence-based prediction of pH gradient elution in multimodal chromatography. *J Chromatogr A*. 2023;1711:464437. doi:10.1016/j.chroma.2023.464437
40. Emonts J, Buyel JF. An overview of descriptors to capture protein properties – tools and perspectives in the context of QSAR modeling. *Comput Struct Biotechnol J*. 2023;21:3234-3247. doi:10.1016/j.csbj.2023.05.022
41. Danishuddin KAU. Descriptors and their selection methods in QSAR analysis: paradigm for drug design. *Drug Discov Today*. 2016;21(8):1291-1302. doi:10.1016/j.drudis.2016.06.013
42. Neijenhuis T, Le Bussy O, Geldhof G, Klijn ME, Ottens M. Predicting protein retention in ion-exchange chromatography using an open source QSPR workflow. *Biotechnol J*. 2024;19(3):2300708. doi:10.1002/biot.202300708
43. Hou Y, 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*. 2011;1218(43):7813-7820. doi:10.1016/j.chroma.2011.08.080
44. Osberghaus A, Hepbildikler S, Nath S, Haindl M, von Lieres E, Hubbuch J. Determination of parameters for the steric mass action model—a comparison between two approaches. *J Chromatogr A*. 2012;1233:54-65. doi:10.1016/j.chroma.2012.02.004
45. Hagemann F, Adametz P, Wessling M, Thom V. Modeling hindered diffusion of antibodies in agarose beads considering pore size reduction due to adsorption. *J Chromatogr A*. 2020;1626:461319. doi:10.1016/j.chroma.2020.461319
46. Keulen D, van der Hagen E, Geldhof G, Le Bussy O, Pabst M, Ottens M. Using artificial neural networks to accelerate flowsheet optimization for downstream process development. *Biotechnol Bioeng*. 2023;121:2318-2331. doi:10.1002/bit.28454
47. Ruthven DM. *Principles of Adsorption and Adsorption Processes*. John Wiley & Sons; 1984.
48. Brenner H, Gaydos LJ. The constrained brownian movement of spherical particles in cylindrical pores of comparable radius: models of the diffusive and convective transport of solute molecules in membranes and porous media. *J Colloid Interface Sci*. 1977;58(2):312-356. doi:10.1016/0021-9797(77)90147-3
49. Young ME, Carroad PA, Bell RL. Estimation of diffusion coefficients of proteins. *Biotechnol Bioeng*. 1980;22(5):947-955. doi:10.1002/bit.260220504
50. Petzold L. Automatic selection of methods for solving stiff and nonstiff Systems of Ordinary Differential Equations. *SIAM J Sci Stat Comput*. 1983;4(1):136-148. doi:10.1137/0904010
51. Nfor BK, Zuluaga DS, Verheijen PJT, Verhaert PDEM, van der Wielen LAM, Ottens M. Model-based rational strategy for chromatographic resin selection. *Biotechnol Prog*. 2011;27(6):1629-1643. doi:10.1002/btpr.691
52. Nfor BK, Noverraz M, Chilamkurthi S, Verhaert PDEM, van der Wielen LAM, Ottens M. High-throughput isotherm determination and thermodynamic modeling of protein adsorption on mixed mode adsorbents. *J Chromatogr A*. 2010;1217(44):6829-6850. doi:10.1016/j.chroma.2010.07.069
53. Pirrung SM, da Cruz DP, Hanke AT, et al. Chromatographic parameter determination for complex biological feedstocks. *Biotechnol Prog*. 2018;34(4):1006-1018. doi:10.1002/btpr.2642
54. Hahn T, Geng N, Petrushevska-Seebach K, et al. Mechanistic modeling, simulation, and optimization of mixed-mode chromatography for an antibody polishing step. *Biotechnol Prog*. 2023;39(2):e3316. doi:10.1002/btpr.3316
55. Schmidt-Traub H, Schulte M, Seidel-Morgenstern A, Schmidt-Traub H. *Preparative Chromatography*. Wiley Online Library; 2012.
56. Parente ES, Wetlaufer DB. Relationship between isocratic and gradient retention times in the high-performance ion-exchange chromatography of proteins: theory and experiment. *J Chromatogr A*. 1986;355:29-40. doi:10.1016/S0021-9673(01)97301-7
57. Shukla AA, Bae SS, Moore JA, Barnthouse KA, Cramer SM. Synthesis and characterization of high-affinity, low molecular weight displacers for cation-exchange chromatography. *Ind Eng Chem Res*. 1998;37(10):4090-4098. doi:10.1021/ie9801756
58. Berman HM, Westbrook J, Feng Z, et al. The Protein Data Bank. *Nucleic Acids Res*. 2000;28(1):235-242. doi:10.1093/nar/28.1.235
59. Rodrigues J, Teixeira J, Trellet M, Bonvin A. Pdb-tools: a swiss army knife for molecular structures [version 1; peer review: 2 approved]. *F1000Research*. 2018;7(1961):1-9. doi:10.12688/f1000research.17456.1
60. Olsson MHM, Søndergaard CR, Rostkowski M, Jensen JH. PROPKA3: consistent treatment of internal and surface residues in empirical pKa predictions. *J Chem Theory Comput*. 2011;7(2):525-537. doi:10.1021/ct100578z
61. Pettersen EF, Goddard TD, Huang CC, et al. UCSF chimera—a visualization system for exploratory research and analysis. *J Comput Chem*. 2004;25(13):1605-1612. doi:10.1002/jcc.20084
62. Topliss JG, Costello RJ. Chance correlations in structure-activity studies using multiple regression analysis. *J Med Chem*. 1972;15(10):1066-1068. doi:10.1021/jm00280a017
63. Brooks CA, Cramer SM. Steric mass-action ion exchange: displacement profiles and induced salt gradients. *AIChE Journal*. 1992;38(12):1969-1978. doi:10.1002/aic.690381212
64. Rathore AS. Roadmap for implementation of quality by design (QbD) for biotechnology products. *Trends Biotechnol*. 2009;27(9):546-553. doi:10.1016/j.tibtech.2009.06.006
65. Panikulam S, Hanke A, Kroener F, et al. Host cell protein networks as a novel co-elution mechanism during protein a chromatography. *Biotechnol Bioeng*. 2024;121(5):1716-1728. doi:10.1002/bit.28678

How to cite this article: Keulen D, Neijenhuis T, Lazopoulou A, et al. From protein structure to an optimized chromatographic capture step using multiscale modeling. *Biotechnol. Prog*. 2024;e3505. doi:10.1002/btpr.3505

APPENDIX A

A.1 | Dead volume and dwell volume

The volume of the tubing was determined by excluding the column and using 1 M sodium chloride with a 100 µL sample loop. A schematic overview of the tubing in the Äkta system is shown in Figure A1, in which the dead volume is indicated from the numbers 2 to 4 and the dwell volume from 1 to 3.

The dead volume (V_{dead}), tubing 3 and 4, is calculated according to Schmidt-Traub et al. (2012) (Equation A1) as follows[1]:

$$V_{\text{dead}} = V_{R,0} - \frac{V_{inj}}{2} - V_5, \tag{A1}$$

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

The dwell volume is needed for the calculations in the regression formula and is equal to the volume from point 1 to 3 (Figure A1). The tubing before point 1 is already filled prior to elution. The dwell volume was determined by introducing buffer B, containing 1 M sodium chloride as a pulse for 5 CV, followed by subtracting the V_{dead} and V_5.

A.2 | Porosity calculations

The total porosity (ε_t) was determined using 1 M sodium chloride, as salt can enter the pores, and calculated using Equation A2 as follows

$$\varepsilon_t = \frac{V_m + V_{\text{pore}}}{V_C}, \tag{A2}$$

$$V_m + V_{\text{pore}} = V_{0,\text{ret}} - V_{\text{dead}} \tag{A3}$$

where V_m is the interstitial volume of the fluid phase also known as the column void volume, V_{pore} is the volume of the pore system, and V_C is the total volume of the packed column. $V_{0,\text{ret}}$ is the measured retention volume from which the dead volume is subtracted to only consider the retention volume in the column. The external porosity, $\varepsilon_b = V_m / V_C$, was determined using a solution of 10 mg/mL Dextran (DXT1740K, American Polymer Standards Corporation, USA) with a volume of 250 µL. V_m was determined using Equation A3. Subsequently, the total and external porosity are used to determine the internal porosity (ε_p) via Equation A4 as

$$\varepsilon_p = \frac{\varepsilon_t - \varepsilon_b}{1 - \varepsilon_b}. \tag{A4}$$

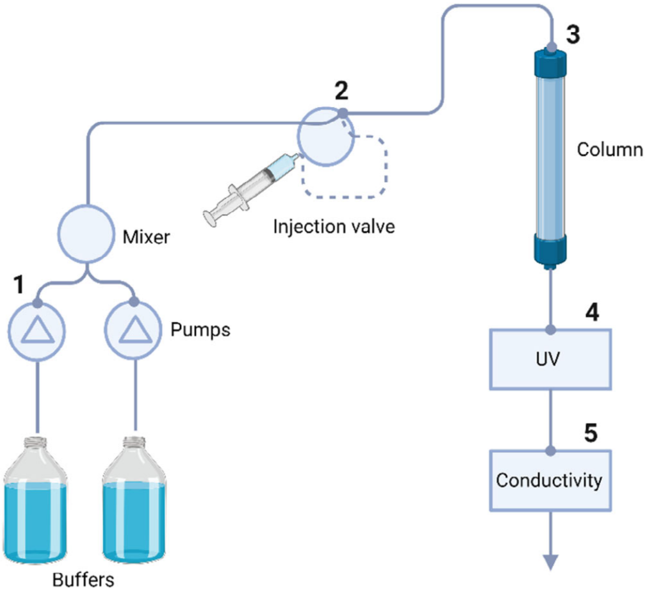


FIGURE A1 Schematic representation of the Äkta system, the dead volume is defined from point 2 to 4 and the dwell volume from point 1 to 3. The injection valve is indicated with the dashed line and not considered in the dead volume and dwell volume. Created with biorender.com.

[1] Schmidt-Traub H, Schulte M, Seidel-Morgenstern A, Schmidt-Traub H. Preparative chromatography. Wiley Online Library; 2012.

APPENDIX B

Regression plots of each protein at each pH, 3.5, 4.3, 5.0, and 7.0 corresponding to the Figures A2–A5, respectively.

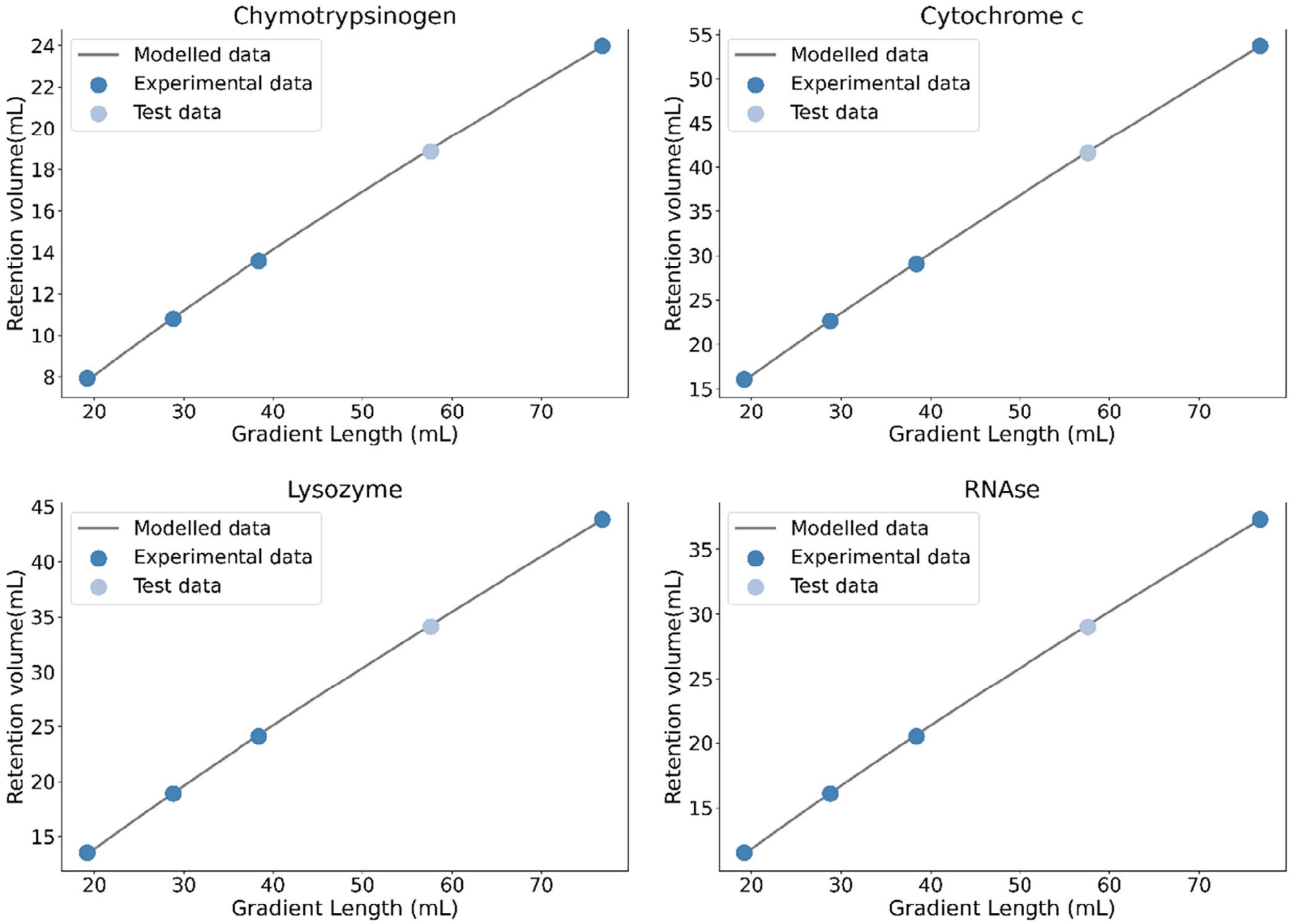


FIGURE A2 Fitted regression curves at pH 3.5 (gray line) of the experimental data (dark blue dots) and the test data point (light blue dot) at 58.2 mL, equal to 60 CV as 1 CV is 0.97 mL. All fits obtained an R^2 of 0.999 and an RMSE of 0.08, 0.11, 0.11, and 0.09 for chymtrypsinogen, cytochrome C, lysozyme, and RNase, respectively.

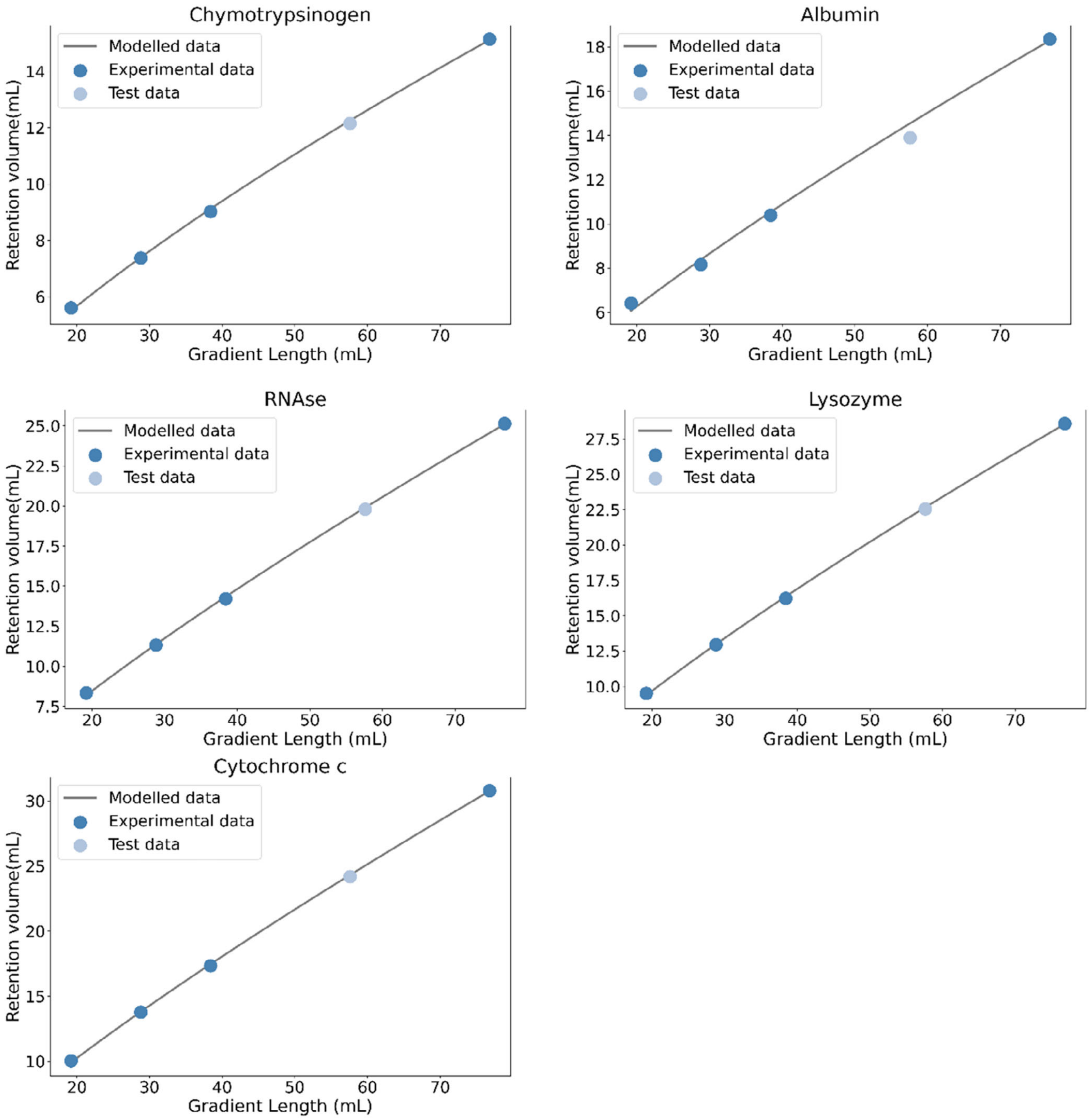


FIGURE A3 Fitted regression curves at pH 4.3 (gray line) of the experimental data (dark blue dots) and the test data point (light blue dot) at 58.2 mL, equal to 60 CV as 1 CV is 0.97 mL. All fits obtained an R^2 of 0.999 and an RMSE of 0.07, 0.22, 0.10, 0.10, and 0.09 for albumin, chymtrypsinogen, cytochrome c, lysozyme, and RNase, respectively.

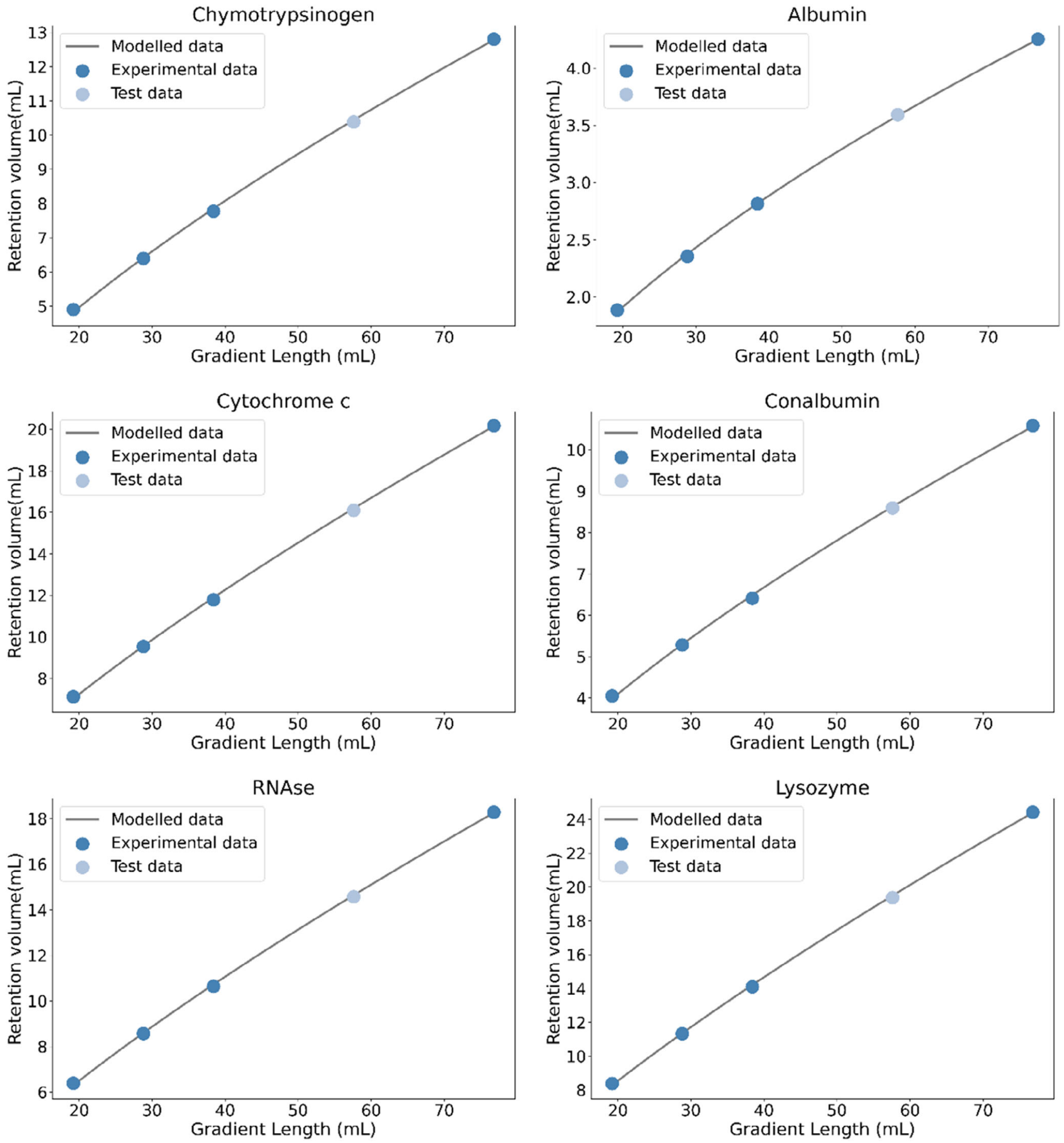


FIGURE A4 Fitted regression curves at pH 5.0 (gray line) of the experimental data (dark blue dots) and the test data point (light blue dot) at 58.2 mL, equal to 60 CV as 1 CV is 0.97 mL. All fits obtained an R^2 of 0.999 and an RMSE of 0.01, 0.05, 0.06, 0.06, 0.07, and 0.08 for albumin, chymotrypsinogen, cytochrome c, lysozyme, RNase, and conalbumin, respectively.

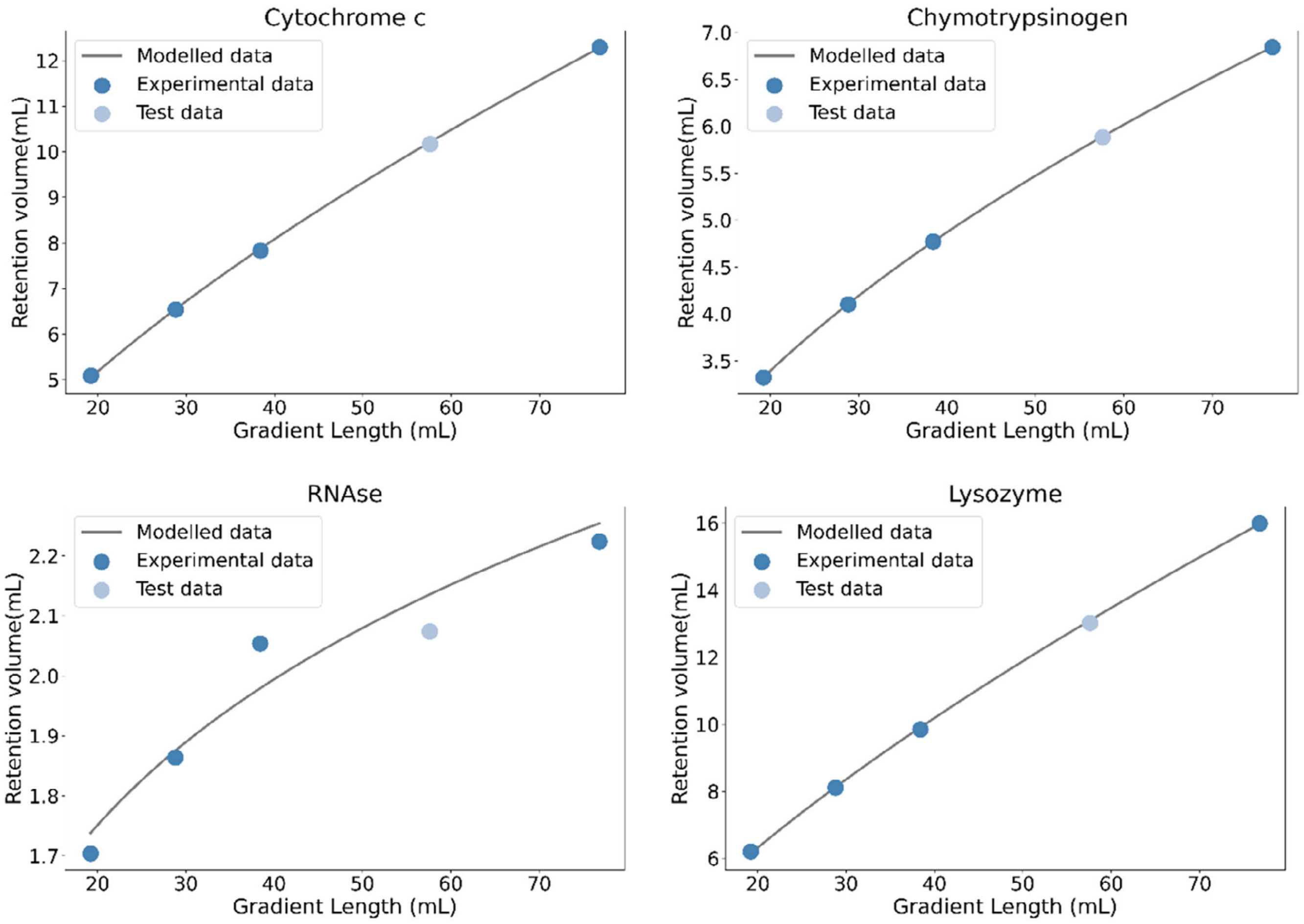


FIGURE A5 Fitted regression curves at pH 7.0 (gray line) of the experimental data (dark blue dots) and the test data point (ligth blue dot) at 58.2 mL, equal to 60 CV as 1 CV is 0.97 mL. All fits obtained an R^2 of 0.999, except for RNAse that has an R^2 of 0.95. The RMSE values are 0.03, 0.002, 0.04, and 0.04 for cytochrome c, chymtrypsinogen, RNAse, and lysozyme, respectively.

APPENDIX C

Additional data for the mechanistic model validated at pH 7.0. For all proteins at pH 7.0, the maximum retention peak difference is 1.01 CV and the average difference is 0.86 CV, which is 1.68% and 1.43% with respect to the gradient length (60 CV). To assess the concentration agreement between the modeled and experimental results, we compared the difference between the peak width at half of the peak maximum and the peak concentration. RNAse was left out of this comparison for the peak width difference, as determining half of the peak maximum is not possible for the experimental data. The maximum peak width difference is 2.07 CV, equal to 2.23% relative to the gradient length (60 CV). The average peak width difference is 0.81 CV, equal to 1.35% relative to the gradient length (60 CV). The peak concentration differs maximally by 0.04 mg/mL, which deviates about 7.8% to the initial concentration. The average difference in the peak concentration is 0.01 mg/mL, equal to 3.1% relative to the initial concentration (Figure A6).

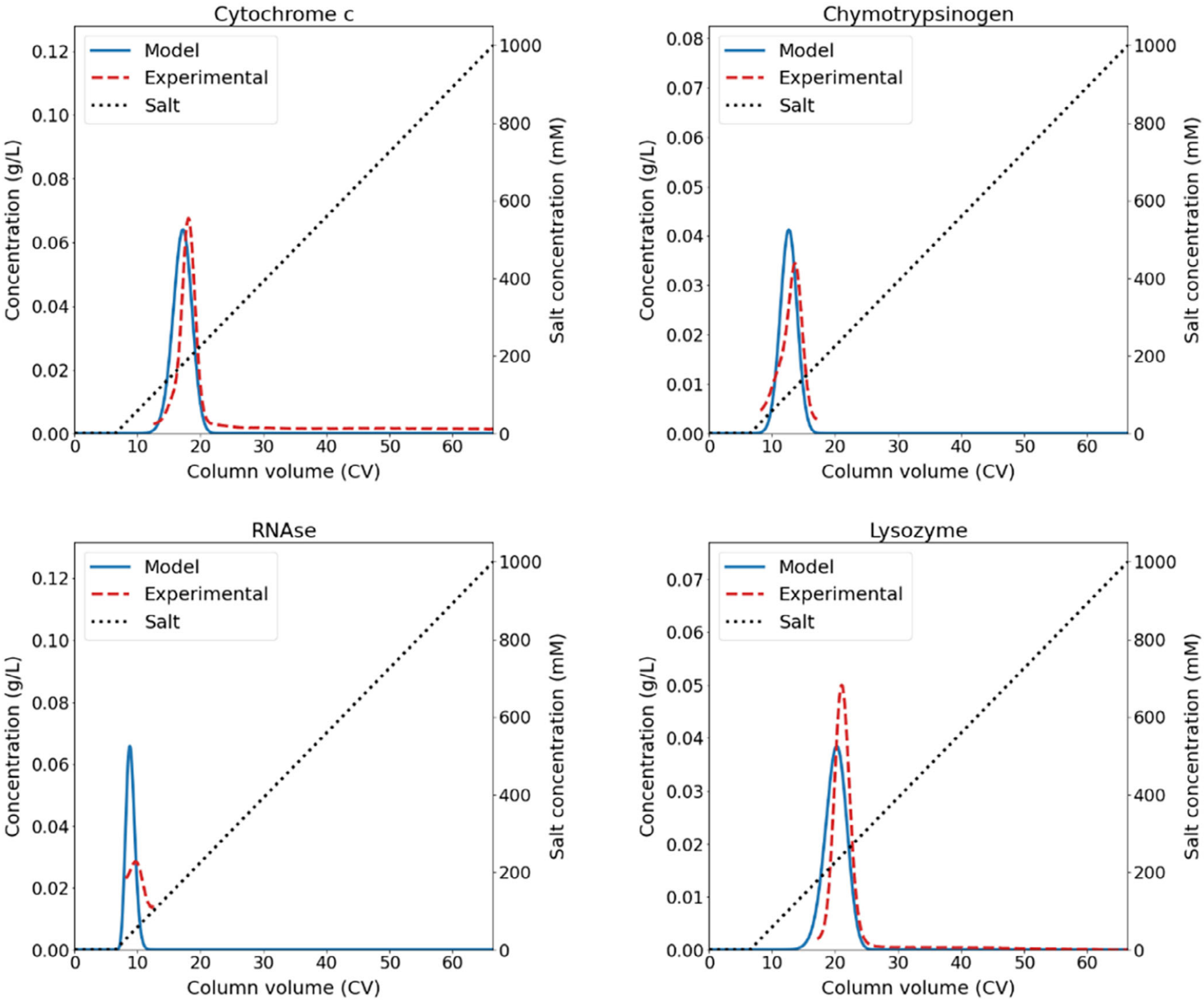


FIGURE A6 Chromatographic mechanistic model validation for gradient length of 60 CV, equal to 58.2 mL, at a pH of 7.0. Blue line indicate the MM predicted concentration of the protein, while the red dotted line indicates the experimental concentration. The black dotted line indicates the salt concentration. The initial concentrations are Chymotrypsinogen: 0.46 mg/mL, Cytochrome c: 0.80 mg/mL, Lysozyme: 0.55 mg/mL, and RNAse: 0.39 mg/mL.

APPENDIX D

Calibration lines for each protein at pH 5.0 and 7.0, shown in Figures A7 and A8, respectively.

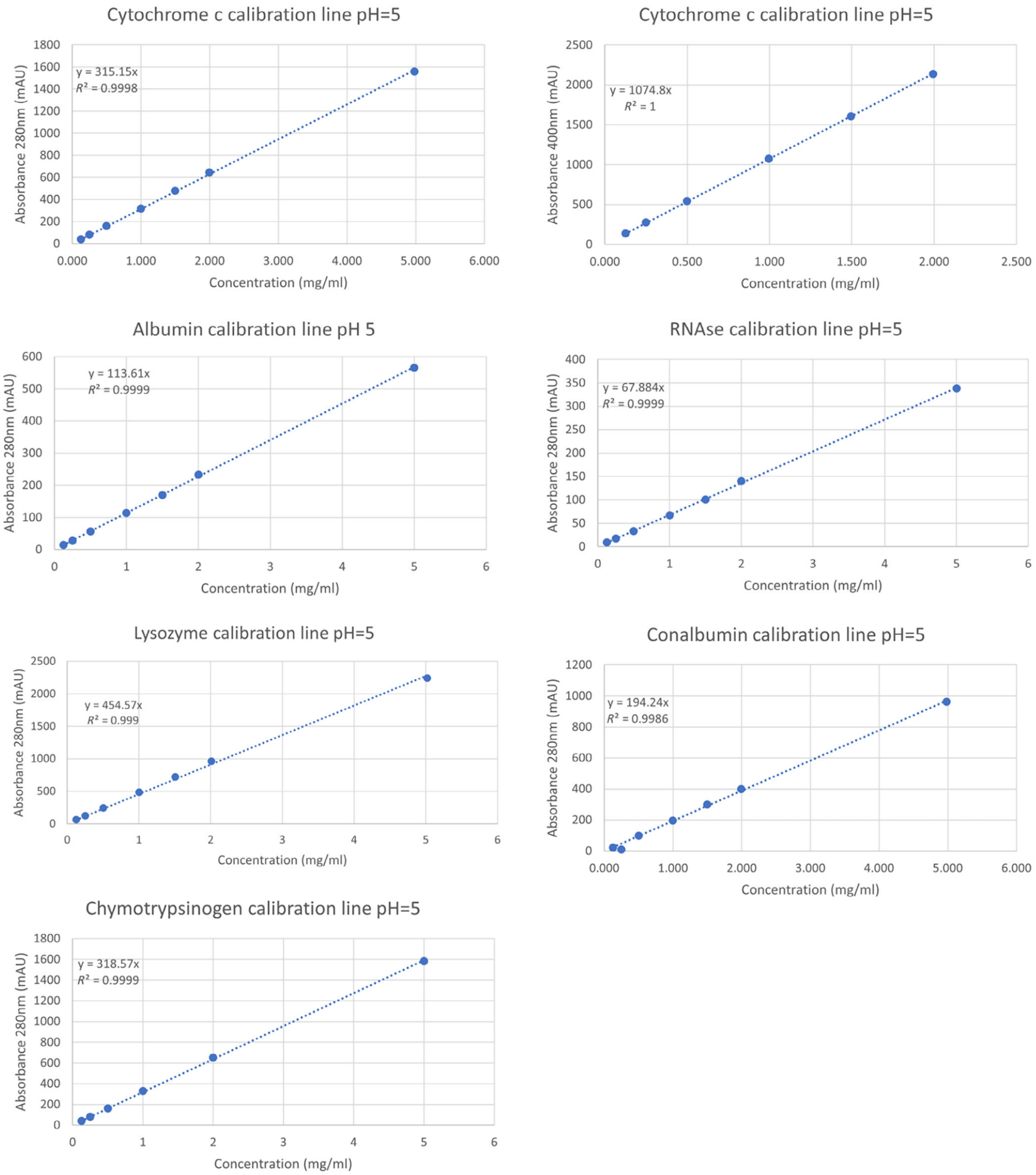


FIGURE A7 Calibration lines (blue dotted line) for each protein at pH = 5, the blue dots indicate the experimental data. The concentrations are measured at an Absorbance of 280 and 400 nm. 400 nm absorbance is specifically needed to quantify cytochrome C.

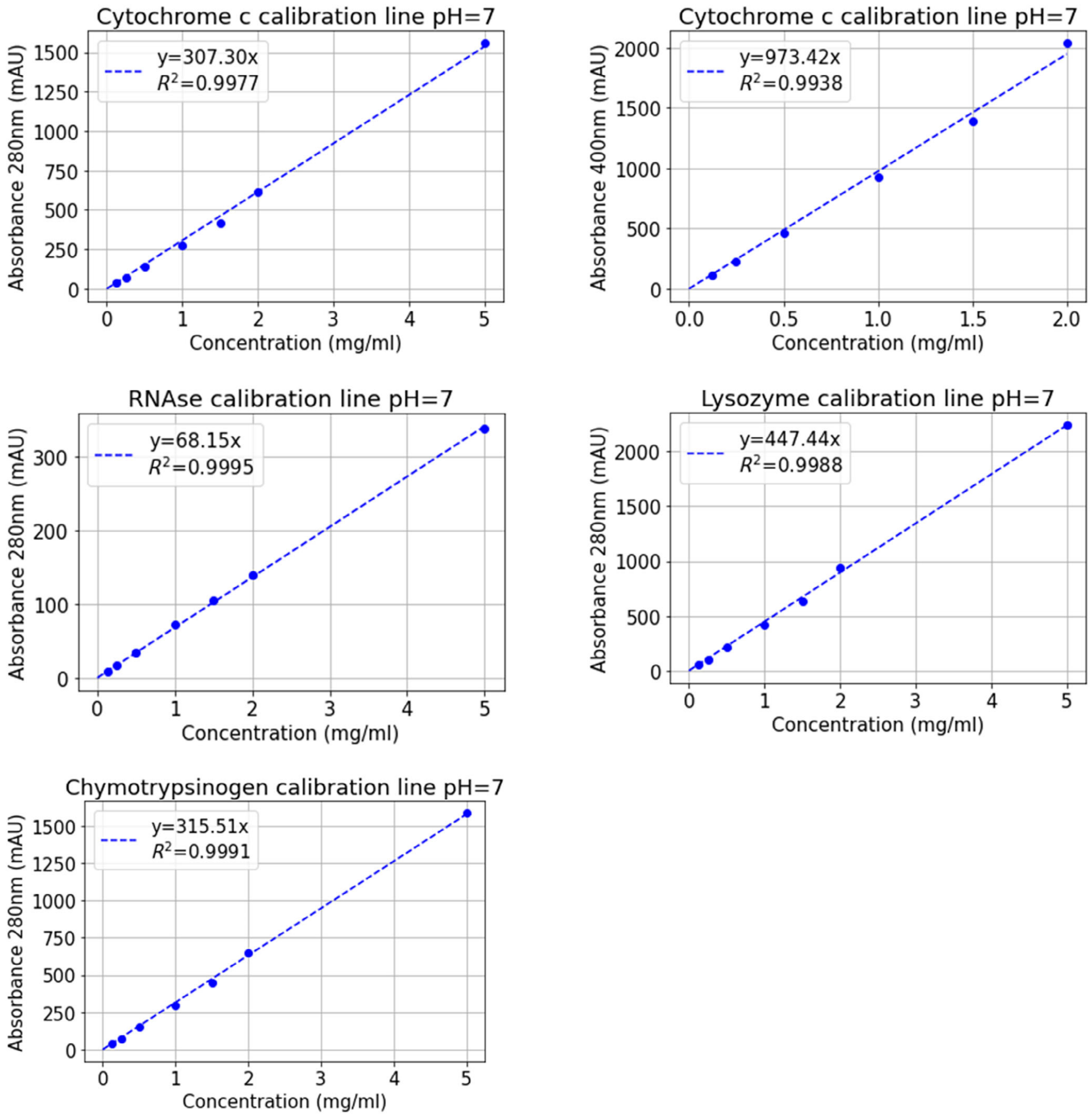


FIGURE A8 Calibration lines (blue dotted line) for each protein at pH = 7.0, the blue dots indicate the experimental data. The concentrations are measured at an Absorbance of 280 and 400 nm. 400 nm absorbance is specifically needed to quantify cytochrome C.

APPENDIX E

The consistency of the optimization case study was evaluated by running the same optimization five times. The QSPR-based and experimental-based method results are shown in Tables A1 and A2 respectively.

TABLE A1 Optimization results using the QSPR-based method, showing the performance measurements and obtained optimized variables. $K_{eq} = 0.028$ and $v = 3.05$.

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

TABLE A2 Optimization results using the experimental-based method, showing the performance measurements and obtained optimized variables. $K_{eq} = 0.071$ and $v = 2.37$.

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