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

Neijenhuis, Tim; Le Bussy, Olivier; Geldhof, Geoffroy; Klijn, Marieke E.; Ottens, Marcel

DOI
10.1002/biot.202300708

Publication date
2024

Document Version
Final published version

Published in
Biotechnology Journal

Citation (APA)
Neijenhuis, T., Le Bussy, O., Geldhof, G., Klijn, M. E., & Ottens, M. (2024). Predicting protein retention in ion-exchange chromatography using an open source QSPR workflow. *Biotechnology Journal*, *19*(3), Article 2300708. https://doi.org/10.1002/biot.202300708

Received: 14 December 2023 | Revised: 1 February 2024 | Accepted: 15 February 2024

DOI: 10.1002/biot.202300708

Biotechnology Journal

RESEARCH ARTICLE

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

Tim Neijenhuis[1] | **Olivier Le Bussy**[2] | **Geoffroy Geldhof**[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, Delft, 2629 HZ, The Netherlands.
Email: M.Ottens@tudelft.nl

Funding information
ChemistryNL; GlaxoSmithKline Biologicals S.A.

Abstract

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

KEYWORDS
chromatography, protein features, Quantitative Structure Activity Relationship (QSAR), Quantitative Structure Property Relationship (QSPR), retention prediction

1 | INTRODUCTION

The market for protein-based biopharmaceuticals, such as protein subunit vaccines and therapeutic antibodies, developed rapidly over recent years.[1] Opposed to chemical synthesis to manufacture small-molecule drugs, protein-based biopharmaceuticals are produced by living host cells. During downstream processing (DSP) the target product is separated from host cell impurities, which is of major importance to guarantee patient safety and drug efficacy. To attain sufficient purity, chromatography is a method of choice due to its specificity and versatility.[2–4] However, the vast variety of commercially available resin types (e.g., ion exchange (IEX) or hydrophobic

interaction chromatography (HIC)) and experimental conditions (e.g., salt concentrations, buffers, and pH) results in extensive experimental screening to obtain optimal separation conditions, driving both cost and development time. In-silico preselection of resins and conditions prior to experimentation would allow a decrease in costs and development time by narrowing the empirical screening space.

Chromatographic separation is based on the difference in physicochemical properties between the product and impurities. For proteins, physicochemical properties are determined by the amino acid sequence (1D) and the 3D structure. Quantitative Structure Property Relationship (QSPR) aims to relate physicochemical properties to specific behavior (e.g., chromatographic retention time).[5] For QSPR, physicochemical properties are described as numerical features and subsequently used in predictive machine learning models as input variables. To build a QSPR workflow, experimental data of known proteins is split in a training and test set. Numerical features are calculated from the proteins in the training set and selected to train a machine learning model (e.g., linear regression, partial least squares (PLS), or neural networks) which predicts the behavior of interest. The resulting model is tested using the numerical features obtained from the proteins in the test set, to assess the model accuracy for new data. When the model provides sufficiently accurate predictions, the property of proteins unknown to the model can be predicted (Figure 1).

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

Using structural protein features to predict protein retention times was first described in 2001 by Mazza et al., who calculated protein features using the transferable atom equivalent method[5,19,20] and the proprietary software platform MOE. By applying a genetic algorithm for feature selection, a PLS model was trained, capable to predict retention times for ion exchange chromatography from protein structure models. Applying the same feature calculation methods, support vector machine regressions for both feature selection and the final predictive model have also been applied for successful protein retention prediction in ion exchange, hydrophobic interaction and mixed mode chromatography.[10–16] As the chromatographic resin interacts with the amino acid residues on the protein surface, Malmquist et al. implemented a grid representation of the protein surface to map protein properties.[21] By applying distance functions to project charge and hydrophobicity onto the surface grid points, protein features were calculated and used in a PLS model to predict retention times for anion and cation exchange columns. As charge and hydrophobicity are usually not uniformly distributed over the protein surface, binding orientations play important roles in protein-resin binding affinities.[22,23] To account for such orientations in QSPR models, Hanke et al. described a method to sample the surface in neighborhoods and uses this for HIC retention time predictions.[24] These neighborhoods are defined as the surface within a specific distance of a central surface point (7 and 14 Å distances were described). Alternatively, Kittelmann et al. used property projections on a plane, sampling different orientations.[25,26] By projecting the properties onto a plane, this method considers steric hindrance on the surface. This results in penalizing the area of surface cavities, which are located at a greater distance from the projection plane.

Most of the described studies use proprietary or in-house software to perform feature calculations and model training. As a result, reproducing these studies is near to impossible. Therefore, direct comparison between different approaches by minimizing the variables cannot be performed, hindering benchmarking opportunities and scientific progress. Additionally, the lack of open source tools limits software availability for new users and customizability to solve a wide variety of development challenges. We aim to close this gap, and in this work, we provide an open source Python tool that is able to calculate 3D protein features. The current implemented operations and features aim to consolidate the most often described protein features from literature.[16,21,25,26] The validity of the features for chromatographic process development was shown by training multiple linear regression (MLR) models predicting retention times/volumes for cation and anion chromatography resins obtained from literature. To promote transparency and scientific reproducibility, the software developed for this study is freely available open source at https://dx.doi.org/10.5281/zenodo.10369949.

2 | METHODS

2.1 | Protein charge

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

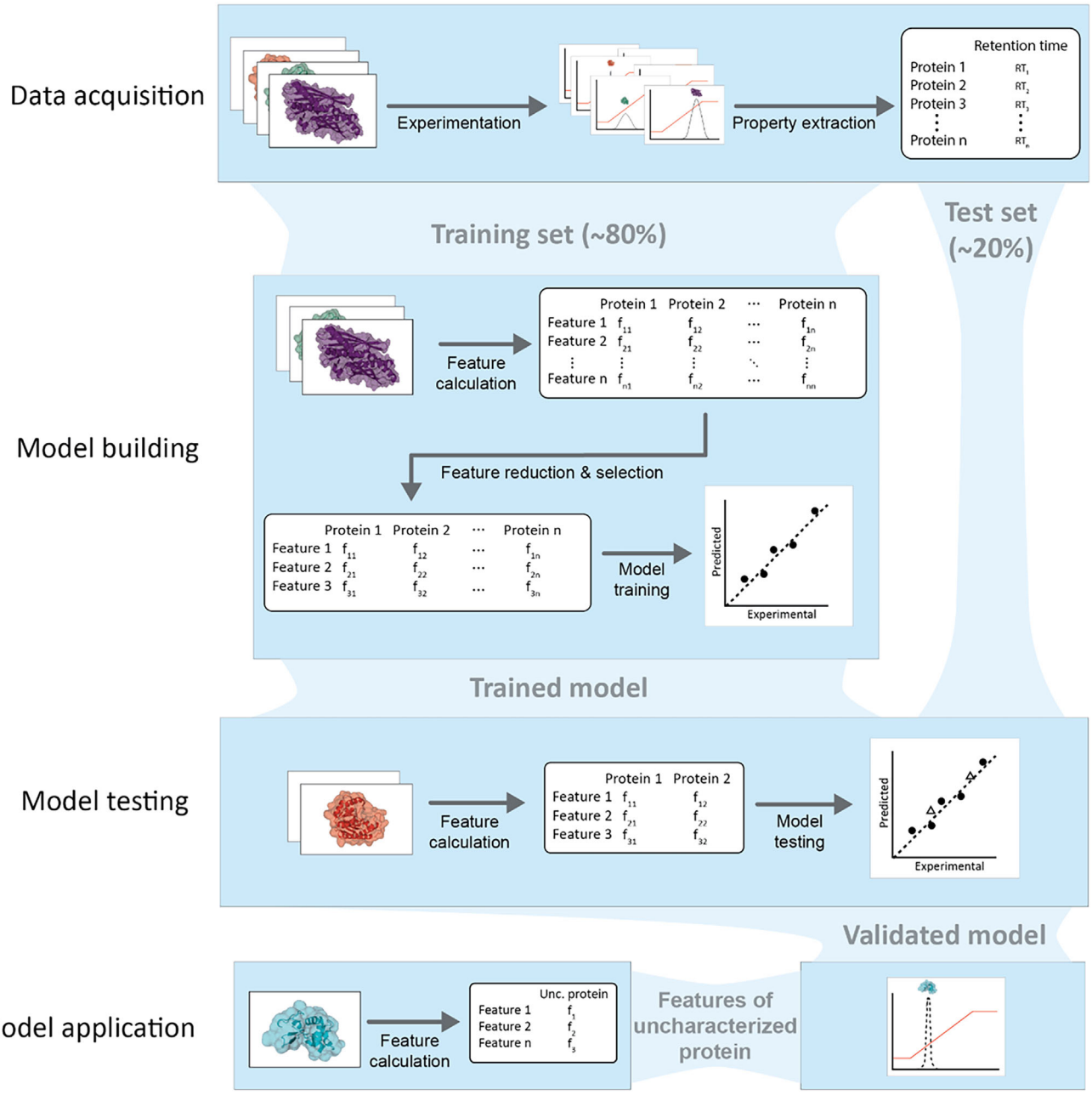

FIGURE 1 Schematic representation of a Quantitative Structure Property Relationship (QSPR) workflow for chromatographic retention prediction. The first step to build a QSPR model is data acquisition. Here, a set of known proteins is used to construct a dataset containing experimentally determined properties (e.g., retention times). The experimental property dataset is split into a train and test set. The training set is used for model building. The physicochemical properties for each protein are calculated using the corresponding 3D structure. The physicochemical properties are expressed as numerical features. The number of features is reduced using dimension reduction methods such as principal component analysis or variance filtering, and the most descriptive features are selected by feature selection to train a predictive model. The resulting model is tested on the test set to assess the accuracy for unseen proteins. Predictive models with good accuracy can be applied to predict the properties of uncharacterized proteins.

The protonation states of these residues can be described by the Henderson–Hasselbalch Equation[27]:

$$pH = pKa + \log\left(\frac{[A^-]}{[AH]}\right), \quad (1)$$

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

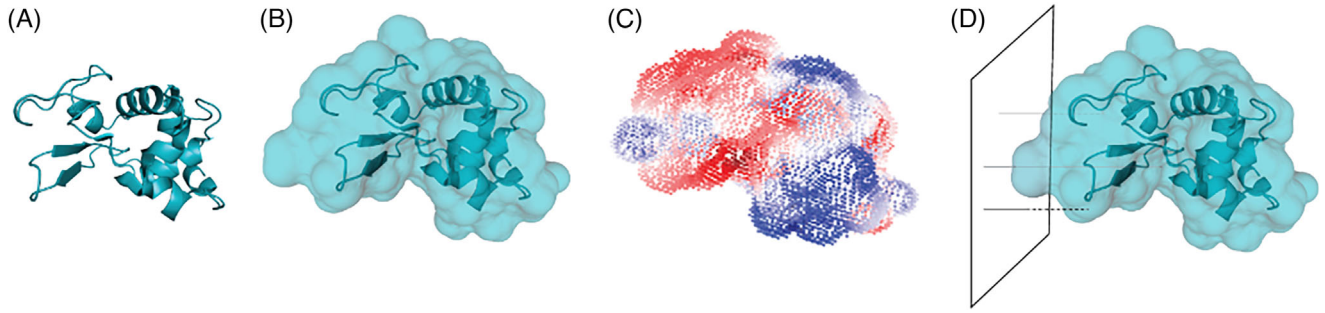

FIGURE 2 Protein representation for feature calculation. (A) shows all atom representation using the coordinates for each atom. (B) shows the Solvent Accessible Surface Area. (C) shows the surface grid representation with mapped electrostatic potentials. (D) shows the plane projection of one orientation.

charges as follows:

$$\text{Charge} = \frac{-1}{1+10^{pKa-pH}}\ [\text{e}] \tag{2}$$

and

$$\text{Charge} = \frac{1}{1+10^{pH-pKa}}\ [\text{e}], \tag{3}$$

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

$$D = 4.803 * \sum_i (r_i - r_p) * q_i\ [D], \tag{4}$$

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

2.2 | Surface definition

Interactions of proteins with their environment often take place at the protein surface. To rationalize these interactions using protein models, accurate representations of the surfaces are required. The Solvent Accessible Surface Area (SASA) is the most common for surface estimation that represents the protein surface which can be occupied by water molecules, and was first described by Lee and Richards[37] (Figure 2B). A number of tools specifically designed for the determination of the SASA are available.[38–40] A spherical probe, representing a solvent molecule, is rolled over the protein atoms tracing the accessible area using the center of the solvent. We adopted the method of Shrake and Rupley[41] where each surface sphere is represented by a set of sample points. The number of sample points is scaled according to the surface sphere radius and are distributed by a Fibonacci sphere,[42] to obtain a distribution of 2 points per Å^2. The fraction of each amino acid occupying the surface can be calculated by dividing the number of surface points of a residue by the total number of surface points.

2.3 | Property projection

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

$$EP = \sum_i \frac{q_i}{\varepsilon d_i}\ [\text{v}], \tag{5}$$

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

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

$$\varepsilon = \frac{\varepsilon_p \times d_p + \varepsilon_w \times d_w}{d}\varepsilon_0\ [-], \tag{6}$$

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

Hydrophobicity of proteins is another important factor which governs interactions. Many different scales describing the contribution of each respective amino acid to hydrophobic phenomena have

TABLE 1 Dataset 1, retention times of specific proteins described by Hou and Cramer [12] for Q Sepharose Fast Flow. Superscript 1 indicates the protein models used as test set.

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

been published.[44] The Cowan–Whittaker[45] and the Miyazawa–Jernigan[46] scales have been reported to give highest correlation for HIC retention prediction.[47] In this work, we use the Miyazawa–Jernigan[46] scale, which was scaled using a min-max-scaler to values ranging from −1 to 1. Hydrophobicity values are projected onto the surface grid to obtain the molecular hydrophobic potential (MHP) using:

$$\mathrm{MHP} = \sum_i f_i e^{-d_i}\ [-], \tag{7}$$

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

A list of all current supported features can be found in Table S1.

2.4 | Dataset composition and feature calculation

Two datasets with known retention behavior for Q Sepharose FF and SP Sepharose HP were required from literature, set 1[12] and set 2[16] respectively (Tables 1 and 2). For both datasets, structures were extracted from the PDB and used to generate homology models by SWISS-MODEL[49,50] to resolve missing atoms. Duplicate chains were removed for all protein models to obtain monomer structures which were used in the feature calculation. To calculate the protonation states, the default *pKa* values were used for the titratable residues. Building the surface grid was performed using a sphere radius of 1.4 Å to represent water.

TABLE 2 Dataset 2, retention volumes of specific proteins at different pHs described by Yang et al.[16] for sulfopropyl Sepharose high performance. Superscript 1 indicates the pH used as test set (6).

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

2.5 | Linear regression modeling

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

3 | RESULTS AND DISCUSSION

To evaluate the performance of the developed Python tool, two datasets were obtained from literature containing protein retention times/volumes for ion-exchange chromatography columns. The first dataset contains protein retention for Q Sepharose FF, and the second for SP Sepharose HP. For both datasets, predictive models were trained relating protein structure to retention time or volume. To determine the validity of the selected features, the regression coefficient and

TABLE 3 Overview of features selected for the linear regression model for Q Sepharose FF and the corresponding regression coefficient and cross-validated R^2 of permutation models.

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

[a]Charge calculated using formal charge (+1, 0, or −1).
[b]Electrostatic potential.

cross-validated R^2 of a permutation model, where each feature is scrambled, are discussed.

vice versa. In alignment with the mode of action of the anion exchange resin, the negative surface potential is the most important feature, as it has the highest regression coefficient and permutation of this feature yields a model uncapable of predicting retention times (Figure S1A). The second feature, number of surface points with a positive EP, shows a positive correlation with protein retention time. This is not in line with the mode of action as a positive protein surface should be repelled by the anion exchange resin. Permutation of this feature reduces the performance of the model to a cross-validated R^2 of 0.563 (Figure S1B). The selection of this feature might be due to the current absence of local surface descriptors. The affected proteins might still contain areas on the surface which are negatively charged that could interact with the anion exchange ligands. The final feature, the valine surface fraction, is of the lowest importance, with a regression coefficient of −5.75. The permutation of this feature results in a model with a cross-validated R^2 of 0.733.

3.1 | Protein retention prediction for Q Sepharose FF

To develop a simple model with high interpretability, a MLR model was trained on protein retention times for the anion exchange resin Q Sepharose FF (Table 3). The dataset that was used (Table 1) was composed of 16 proteins, of which two were selected for testing while the remaining 14 were used for model training.[12] As overfitting can be an issue for linear regression models, a ratio of five datapoints per feature should be maintained, resulting in three features for this dataset.[51] The model's predictability was considered sufficient, with a cross-validated R^2 of 0.87, a RMSE of 2.23, and $RMSE_{test}$ of 2.50 (Figure 3). The two most important features are the median negative surface EP (regression coefficient of −31 and permutated CV R^2 of −0.352) and the number of positive electrostatic surface grid points (regression coefficient of 18.17 and a permutated CV R^2 of 0.563), both calculated using the formal charge (Table 3). A negative regression coefficient indicates an inverse correlation with the retention time of the protein and

3.2 | pH-dependent protein retention prediction for SP Sepharose HP

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

Six of the 10 selected features are directly related to the protein charge and are inherently interconnected. The feature with the highest regression coefficient of 31.24, and therefore deemed most important, is the minimum surface EP. The positive coefficient indicates that an

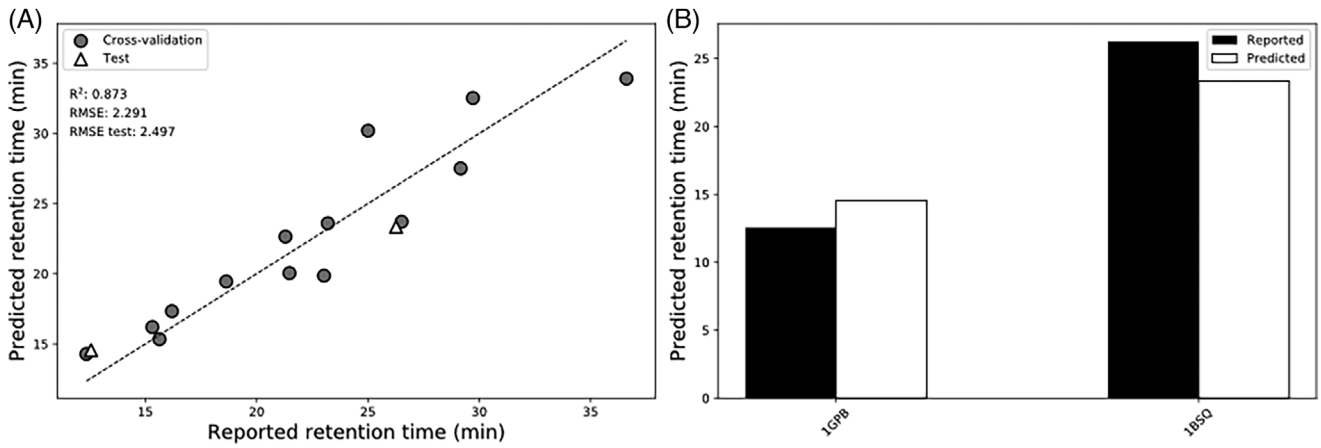

FIGURE 3 Prediction of Q Sepharose FF retention times. (A) shows the leave-one-out cross-validation (gray circles) and test set (white triangles) results of the model. (B) shows the predicted retention times volumes for the external test set (Table 1).

TABLE 4 Overview of features selected for the linear regression model for SP sepharose HP and the corresponding regression coefficient and cross-validated R^2 of permutation models.

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

[a]Charge represented as formal charge (+1, 0, or −1).
[b]Charge calculated using Equations (2) and (3).
[c]Electrostatic potential.

increase in minimum surface EP leads to a higher retention volumes, which is in line with the mode of action of the cation exchange resin. The total charge is the second most important feature with a regression coefficient of −27.77. This indicates that proteins with a higher total charge to have lower retention volumes. Considering the dataset to be retention volumes for the cation exchange resin SP Sepharose HP, a negative correlation with the total charge is counter intuitive. This correlation might not indicate a direct causation with the retention volume, but rather that the total charge might compensate for other charge related features, as there is collinearity between the charge related features. To directly assess the importance of the feature, the permutation model results in a reduced cross-validated R^2 of 0.861. The permutation model for the minimum surface EP resulted in a greater decrease in performance (cross-validated R^2 of 0.822). This indicates that the total charge is indeed less important for the final model compared to EP.

The dipole vector length has a regression coefficient of 20.72. The high positive regression coefficient indicates the importance of charge polarization, and that proteins elute later with more uneven charge distribution. The isoelectric point is the next charge-related feature with a regression coefficient of 12.02. This feature is unaffected by pH as it represents the pH at which the protein is neutrally charged. Interestingly, even though the feature has only the fourth highest coefficient, permutation of the feature results in a permutation model with the lowest R^2 of 0.769 (Figure S2D). As this feature has a low cross correlation with the other features, indicating that less compensation is possible with the remaining data. The importance of the remaining features is significantly lower compared to the first four features (Cross-validated R^2 of permutation > 0.888), a detailed discussion on these features can be found in the Supplemental material.

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

The two QSPR models are capable of the retention prediction for Q Sepharose FF and SP Sepharose HP. All physical phenomena are described implicitly, therefore these models would only be suitable

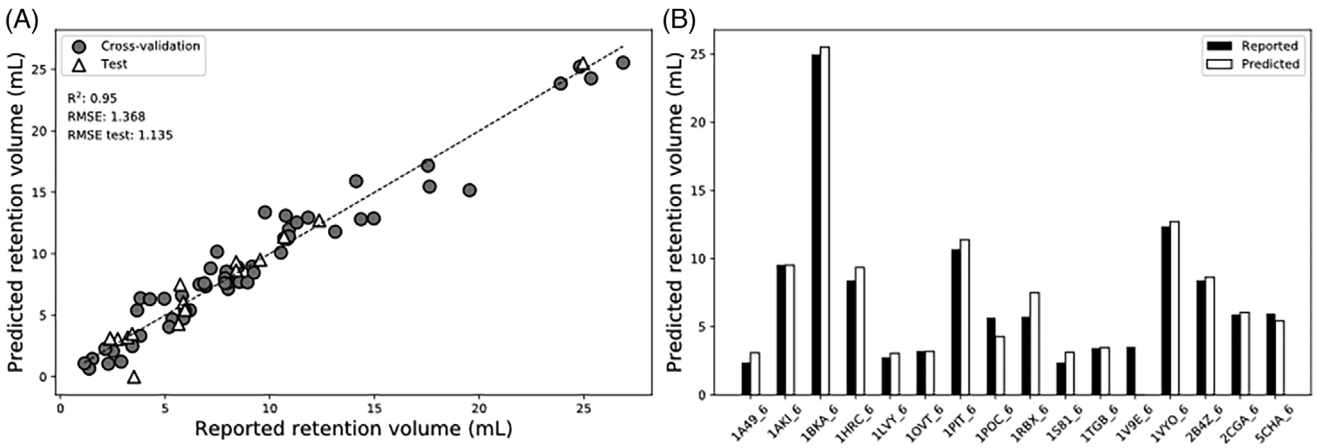

FIGURE 4 Prediction of SP Sepharose HP retention volumes. (A) shows model results of the leave-one-out cross-validation (gray circles) of the proteins at pH 4, 5, 7, and 8 as well as the test set (white triangles) which are the proteins at pH 6. (B) shows the predicted retention volumes for the external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

for describing retention behavior for these specific resins. Extending these models to predict protein retention of other resins would require additional data. This data can subsequently be used in a similar model building approach as described here, yielding predictive models for the new conditions.

4 | CONCLUSION

Physically relevant protein features are essential to achieve robust predictions of protein properties, like chromatographic retention behavior. To mature the field of protein QSPR, adaptable and transparent open source software for the calculation of protein features is essential to directly benchmark between different tools and improve the current state-of-the-art. Using the open source software presented here, we were able to train models that predict the retention times/volumes for two different ion-exchange chromatography datasets, showing applicability for unknown proteins and differences in pH (cross-validated R^2 of 0.87 and 0.95, respectively). Most features selected by the forward feature selection method have an apprehensible relation to protein retention for specific chromatographic conditions. However, collinearity between multiple features was observed. Model performance might therefore benefit from feature reduction techniques such as principal component analysis or PLS regression. Nevertheless, these models show good performance and would allow for prescreening of chromatographic resins. Finally, it was showed that the amount of data available for model training is a major factor determining model accuracy. By increasing the available input data for protein properties like chromatographic retention time, the true impact of the 3D protein features and in silico property prediction for process design can be unlocked in the future.

AUTHOR CONTRIBUTIONS

Tim Neijenhuis: Conceptualization, methodology, investigation, software, 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 Klijn: Conceptualization, supervision, and writing–review & editing. Marcel Ottens: Funding acquisition, conceptualization, supervision, and writing–review & editing.

ACKNOWLEDGMENTS

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

CONFLICT OF INTEREST STATEMENT

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

DATA AVAILABILITY STATEMENT

The open source python tool used in this work will be available upon publication at https://dx.doi.org/10.5281/zenodo.10369949.

ORCID

Tim Neijenhuis https://orcid.org/0000-0002-6214-5438

REFERENCES

1. Kesik-Brodacka, M. (2018). Progress in biopharmaceutical development. *Biotechnology and Applied Biochemistry*, *65*(3), 306–322. https://doi.org/10.1002/bab.1617
2. Gronemeyer, P., Ditz, R., & Strube, J. (2014). Trends in upstream and downstream process development for antibody manufacturing. *Bioengineering*, *1*(4), 188–212. https://doi.org/10.3390/bioengineering1040188
3. Hanke, A. T., & Ottens, M. (2014). Purifying biopharmaceuticals: Knowledge-based chromatographic process development. *Trends in Biotechnology*, *32*(4), 210–220. https://doi.org/10.1016/j.tibtech.2014.02.001
4. Keulen, D., Geldhof, G., Bussy, O. L., Pabst, M., & Ottens, M. (2022). Recent advances to accelerate purification process development: A review with a focus on vaccines. *Journal of Chromatography A*, *1676*, 463195. https://doi.org/10.1016/j.chroma.2022.463195
5. Mazza, C. B., Sukumar, N., Breneman, C. M., & Cramer, S. M. (2001). Prediction of protein retention in ion-exchange systems using molecular descriptors obtained from crystal structure. *Analytical Chemistry*, *73*(22), 5457–5461. https://doi.org/10.1021/ac010797s
6. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2
7. Masrati, G., Landau, M., Ben-Tal, N., Lupas, A., Kosloff, M., & Kosinski, J. (2021). Integrative structural biology in the era of accurate structure prediction. *Journal of Molecular Biology*, *433*(20), 167127. https://doi.org/10.1016/j.jmb.2021.167127
8. Rao, H. B., Zhu, F., Yang, G. B., Li, Z. R., & Chen, Y. Z. (2011). Update of PROFEAT: A web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Research*, *39*(2), 385–390. https://doi.org/10.1093/nar/gkr284
9. Ruiz-Blanco, Y. B., Paz, W., Green, J., & Marrero-Ponce, Y. (2015). ProtDCal: A program to compute general-purpose-numerical descriptors for sequences and 3D-structures of proteins. *BMC Bioinformatics [Electronic Resource]*, *16*(1), 1–15. https://doi.org/10.1186/s12859-015-0586-0
10. Buyel, J. F., Woo, J. A., Cramer, S. M., & Fischer, R. (2013). The use of quantitative structure–activity relationship models to develop optimized processes for the removal of tobacco host cell proteins during biopharmaceutical production. *Journal of Chromatography A*, *1322*, 18–28. https://doi.org/10.1016/j.chroma.2013.10.076
11. Chen, J., & Cramer, S. M. (2007). Protein adsorption isotherm behavior in hydrophobic interaction chromatography. *Journal of Chromatography A*, *1165*(1–2), 67–77. https://doi.org/10.1016/j.chroma.2007.07.038
12. Hou, Y., & Cramer, S. M. (2011). Evaluation of selectivity in multimodal anion exchange systems: A priori prediction of protein retention and examination of mobile phase modifier effects. *Journal of Chromatography A*, *1218*(43), 7813–7820. https://doi.org/10.1016/j.chroma.2011.08.080
13. Ladiwala, A., Rege, K., Breneman, C. M., & Cramer, S. M. (2003). Investigation of mobile phase salt type effects on protein retention and selectivity in cation-exchange systems using quantitative structure

retention relationship models. *Langmuir*, *19*(20), 8443–8454. https://doi.org/10.1021/la0346651

14. Ladiwala, A., Rege, K., Breneman, C. M., & Cramer, S. M. (2005). A priori prediction of adsorption isotherm parameters and chromatographic behavior in ion-exchange systems. *Proceedings of the National Academy of Sciences of the USA*, *102*(33), 11710–11715. https://doi.org/10.1073/pnas.0408769102
15. Song, M., Breneman, C. M., Bi, J., Sukumar, N., Bennett, K. P., Cramer, S., & Tugcu, N. (2002). Prediction of protein retention times in anion-exchange chromatography systems using support vector regression. *Journal of Chemical Information and Computer Sciences*, *42*(6), 1347–1357. https://doi.org/10.1021/ci025580t
16. Yang, T., Sundling, M. C., Freed, A. S., Breneman, C. M., & Cramer, S. M. (2007). Prediction of pH-dependent chromatographic behavior in ion-exchange systems. *Analytical Chemistry*, *79*(23), 8927–8939. https://doi.org/10.1021/ac071101j
17. Sankar, K., Trainor, K., Blazer, L. L., Adams, J. J., Sidhu, S. S., Day, T., Meiering, E., & Maier, J. K. X. (2022). A descriptor set for quantitative structure-property relationship prediction in biologics. *Molecular Informatics*, *41*(9), 2100240. https://doi.org/10.1002/minf.202100240
18. Emonts, J., & Buyel, J. F. (2023). An overview of descriptors to capture protein properties—Tools and perspectives in the context of QSAR modeling. *Comput. Struct. Biotechnol. J*, *21*, 3234–3247. https://doi.org/10.1016/j.csbj.2023.05.022
19. Breneman, C. M., Thompson, T. R., Rhem, M., & Dung, M. (1995). Electron density modeling of large systems using the transferable atom equivalent method. *Computers & Chemistry*, *19*(3), 161–179. https://doi.org/10.1016/0097-8485(94)00052-G
20. Whitehead, C. E., Breneman, C. M., Sukumar, N., & Ryan, M. D. (2003). Transferable atom equivalent multicentered multipole expansion method. *Journal of Computational Chemistry*, *24*(4), 512–529. https://doi.org/10.1002/jcc.10240
21. Malmquist, G., Nilsson, U. H., Norrman, M., Skarp, U., Strömgren, M., & Carredano, E. (2006). Electrostatic calculations and quantitative protein retention models for ion exchange chromatography. *Journal of Chromatography A*, *1115*(1–2), 164–186. https://doi.org/10.1016/j.chroma.2006.02.097
22. Dismer, F., & Hubbuch, J. (2007). A novel approach to characterize the binding orientation of lysozyme on ion-exchange resins. *Journal of Chromatography A*, *1149*(2), 312–320. https://doi.org/10.1016/j.chroma.2007.03.074
23. Dismer, F., Petzold, M., & Hubbuch, J. (2008). Effects of ionic strength and mobile phase pH on the binding orientation of lysozyme on different ion-exchange adsorbents. *Journal of Chromatography A*, *1194*(1), 11–21. https://doi.org/10.1016/j.chroma.2007.12.085
24. Hanke, A. T., Klijn, M. E., Verhaert, P. D. E. M., van der Wielen, L. A. M., Ottens, M., Eppink, M. H. M., & van de Sandt, E. J. A. X. (2016). Prediction of protein retention times in hydrophobic interaction chromatography by robust statistical characterization of their atomic-level surface properties. *Biotechnology Progress*, *32*(2), 372–381. https://doi.org/10.1002/btpr.2219
25. Kittelmann, J., Lang, K. M. H., Ottens, M., & Hubbuch, J. (2017a). An orientation sensitive approach in biomolecule interaction quantitative structure–activity relationship modeling and its application in ion-exchange chromatography. *Journal of Chromatography A*, *1482*, 48–56. https://doi.org/10.1016/j.chroma.2016.12.065
26. Kittelmann, J., Lang, K. M. H., Ottens, M., & Hubbuch, J. (2017b). Orientation of monoclonal antibodies in ion-exchange chromatography: A predictive quantitative structure–activity relationship modeling approach. *Journal of Chromatography A*, *1510*, 33–39. https://doi.org/10.1016/j.chroma.2017.06.047
27. Henderson, L. J. (1908). Concerning the relationship between the strength of acids and their capacity to preserve neutrality. *American Journal of Physiology-Legacy Content*, *21*(2), 173–179. https://doi.org/10.1152/ajplegacy.1908.21.2.173
28. Nelson, D. L., & Cox, M. M. (2001). *Lehninger principles of biochemistry*. Springer Berlin Heidelberg (Springer-Lehrbuch). https://doi.org/10.1007/978-3-662-08289-8
29. Fitch, C. A., Platzer, G., Okon, M., Garcia-Moreno, B. E., & McIntosh, L. P. (2015). Arginine: Its pKa value revisited. *Protein Science*, *24*(5), 752–761. https://doi.org/10.1002/pro.2647
30. Bas, D. C., Rogers, D. M., & Jensen, J. H. (2008). Very fast prediction and rationalization of p Ka values for protein–ligand complexes. *Proteins: Structure, Function and Genetics*, *73*(3), 765–783. https://doi.org/10.1002/prot.22102
31. Olsson, M. H. M., Søndergaard, C. R., Rostkowski, M., & Jensen, J. H. (2011). PROPKA3: Consistent treatment of internal and surface residues in empirical p K a predictions. *Journal of Chemical Theory and Computation*, *7*(2), 525–537. https://doi.org/10.1021/ct100578z
32. Anandakrishnan, R., Aguilar, B., & Onufriev, A. V. (2012). H++ 3.0: Automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic Acids Research*, *40*(W1), 537–541. https://doi.org/10.1093/nar/gks375
33. Gordon, J. C., Myers, J. B., Folta, T., Shoja, V., Heath, L. S., & Onufriev, A. (2005). H++: A server for estimating pKas and adding missing hydrogens to macromolecules. *Nucleic Acids Research*, *33*(2), 368–371. https://doi.org/10.1093/nar/gki464
34. Vriend, G. (1990). WHAT IF: A molecular modeling and drug design program. *Journal of Molecular Graphics*, *8*(1), 52–56. https://doi.org/10.1016/0263-7855(90)80070-V
35. Antosiewicz, J. (1995). Computation of the dipole moments of proteins. *Biophysical Journal*, *69*(4), 1344–1354. https://doi.org/10.1016/S0006-3495(95)80001-9
36. Felder, C. E., Prilusky, J., Silman, I., & Sussman, J. L. (2007). A server and database for dipole moments of proteins. *Nucleic Acids Research*, *35*(2), 512–521. https://doi.org/10.1093/nar/gkm307
37. Lee, B., & Richards, F. M. (1971). The interpretation of protein structures: Estimation of static accessibility. *Journal of Molecular Biology*, *55*(3), 379–400. https://doi.org/10.1016/0022-2836(71)90324-X
38. Ali, S., Hassan, M., Islam, A., & Ahmad, F. (2014). A review of methods available to estimate solvent-accessible surface areas of soluble proteins in the folded and Unfolded States. *Current Protein & Peptide Science*, *15*(5), 456–476. https://doi.org/10.2174/1389203715666140327114232
39. Mitternacht, S. (2016). FreeSASA: An open source C library for solvent accessible surface area calculations. *F1000Research*, *5*, 1–11. https://doi.org/10.12688/f1000research.7931.1
40. Touw, W. G., Baakman, C., Black, J., Te Beek, T. A. H., Krieger, E., Joosten, R. P., & Vriend, G. (2015). A series of PDB-related databanks for everyday needs. *Nucleic Acids Research*, *43*(D1), D364–D368. https://doi.org/10.1093/nar/gku1028
41. Shrake, A., & Rupley, J. A. (1973). Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *Journal of Molecular Biology*, *79*(2), 361–371. https://doi.org/10.1016/0022-2836(73)90011-9
42. Swinbank, R., & Purser, R. J. (2006). Fibonacci grids: A novel approach to global modelling. *Quarterly Journal of the Royal Meteorological Society*, *132*(619), 1769–1793. https://doi.org/10.1256/qj.05.227
43. Schutz, C. N., & Warshel, A. (2001). What are the dielectric "constants" of proteins and how to validate electrostatic models? *Proteins: Structure, Function and Genetics*, *44*(4), 400–417. https://doi.org/10.1002/prot.1106
44. Simm, S., Einloft, J., Mirus, O., & Schleiff, E. (2016). 50 years of amino acid hydrophobicity scales: revisiting the capacity for peptide classification. *Biological Research*, *49*(1), 1–19. https://doi.org/10.1186/s40659-016-0092-5
45. Cowan, R., & Whittaker, R. G. (1990). Hydrophobicity indices for amino acid residues as determined by high-performance liquid chromatography. *Peptide Research*, *3*(2), 75–80. http://www.ncbi.nlm.nih.gov/pubmed/2134053

46. Miyazawa, S., & Jernigan, R. L. (1985). Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. *Macromolecules*, *18*(3), 534–552. https://doi.org/10.1021/ma00145a039
47. Lienqueo, M. E., Mahn, A., & Asenjo, J. A. (2002). Mathematical correlations for predicting protein retention times in hydrophobic interaction chromatography. *Journal of Chromatography A*, *978*(1–2), 71–79. https://doi.org/10.1016/S0021-9673(02)01358-4
48. Fauchére, J. L., Quarendon, P., & Kaetterer, L. (1988). Estimating and representing hydrophobicity potential. *Journal of Molecular Graphics*, *6*(4), 203–206. https://doi.org/10.1016/S0263-7855(98)80004-0
49. Bienert, S., Waterhouse, A., De Beer, T. A. P., Tauriello, G., Studer, G., Bordoli, L., & Schwede, T. (2017). The SWISS-MODEL Repository—new features and functionality. *Nucleic Acids Research*, *45*(D1), D313–D319. https://doi.org/10.1093/nar/gkw1132
50. Kiefer, F., Arnold, K., Künzli, M., Bordoli, L., & Schwede, T. (2009). The SWISS-MODEL Repository and associated resources. *Nucleic Acids Research*, *37*(1), 387–392. https://doi.org/10.1093/nar/gkn750
51. Topliss, J. G., & Costello, R. J. (1972). Chance correlations in structure-activity studies using multiple regression analysis. *Journal of Medicinal Chemistry*, *15*(10), 1066–1068. https://doi.org/10.1021/jm00280a017

SUPPORTING INFORMATION

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

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