Journal of Chromatography A 1711 (2023) 464437

Contents lists available at ScienceDirect

Journal of Chromatography A

journal homepage: www.elsevier.com/locate/chroma

Antibody sequence-based prediction of pH gradient elution in multimodal chromatography

Rudger Hess [a,b], Jan Faessler [b], Doil Yun [b], David Saleh [b], Jan-Hendrik Grosch [b], Thomas Schwab [b], Jürgen Hubbuch [a,*]

[a] Institute of Engineering in Life Sciences, Section IV: Biomolecular Separation Engineering, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
[b] DSP Development, Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach, Germany

ARTICLE INFO

Keywords:
Multispecific monoclonal antibody (mAb) formats
Structure-function analysis
Quantitative structure-activity/property relationship (QSAR/QSPR)
In silico process development
Downstream manufacturability assessment

ABSTRACT

Multimodal chromatography has emerged as a promising technique for antibody purification, owing to its capacity to selectively capture and separate target molecules. However, the optimization of chromatography parameters remains a challenge due to the intricate nature of protein-ligand interactions. To tackle this issue, efficient predictive tools are essential for the development and optimization of multimodal chromatography processes. In this study, we introduce a methodology that predicts the elution behavior of antibodies in multimodal chromatography based on their amino acid sequences. We analyzed a total of 64 full-length antibodies, including IgG1, IgG4, and IgG-like multispecific formats, which were eluted using linear pH gradients from pH 9.0 to 4.0 on the anionic mixed-mode resin Capto adhere. Homology models were constructed, and 1312 antibody-specific physicochemical descriptors were calculated for each molecule. Our analysis identified six key structural features of the multimodal antibody interaction, which were correlated with the elution behavior, emphasizing the antibody variable region. The results show that our methodology can predict pH gradient elution for a diverse range of antibodies and antibody formats, with a test set R^2 of 0.898. The developed model can inform process development by predicting initial conditions for multimodal elution, thereby reducing trial and error during process optimization. Furthermore, the model holds the potential to enable an *in silico* manufacturability assessment by screening target antibodies that adhere to standardized purification conditions. In conclusion, this study highlights the feasibility of using structure-based prediction to enhance antibody purification in the biopharmaceutical industry. This approach can lead to more efficient and cost-effective process development while increasing process understanding.

1. Introduction

At present, the monoclonal antibody (mAb) production relies on chromatographic purification, which is integrated into a templated platform process [1]. Multimodal chromatography has emerged as a highly selective separation method compared to using single-mode interaction resins [2,3]. Specifically, the application of multimodal chromatography in the primary capture from harvested cell culture fluid and subsequent polishing steps has demonstrated its effectiveness in separating process and product-related impurities [4–6]. The enhanced selectivity of multimodal resins stems from orthogonal physicochemical interactions with the molecule surface [7,8]. In this context, ligands functionalized with electrostatic, hydrophobic, aromatic, and/or hydrogen bonding groups are commonly used, as illustrated by the Capto adhere ligand in Fig. 1c [9].

Owing to the intricate multimodal interaction, a broad range of operating conditions must be assessed, as the purification is constraint to a narrow, molecule-specific parameter window of buffer conductivity, pH, modulator concentration, and temperature compared to unimodal chromatography, which can restrict molecule manufacturability [10, 11]. To support process development, extensive research has been conducted to enhance process understanding by examining multimodal protein-ligand interaction alongside efficient screening methodologies. Macroscopic effects have been explored through batch and dynamic-binding experiments, which were described using thermodynamic models [12–14]. To improve the resolution of macroscopic observations, domain contributions of multimeric proteins and the impact of amino acid substitutions in homologous protein libraries were

* Corresponding author.
E-mail address: juergen.hubbuch@kit.edu (J. Hubbuch).

https://doi.org/10.1016/j.chroma.2023.464437
Received 11 July 2023; Received in revised form 3 October 2023; Accepted 5 October 2023
Available online 11 October 2023

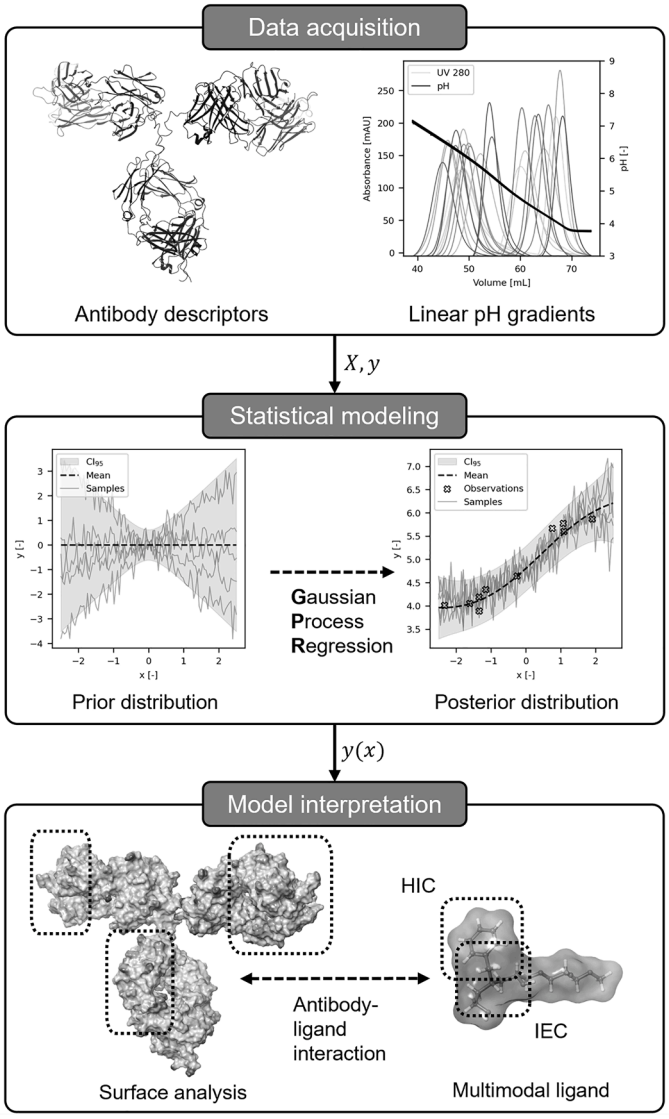

Fig. 1. QSPR modeling workflow. A three-step process is shown that includes a) the data acquisition for model training and testing, b) the statistical modeling using Gaussian Process Regression, and c) the model interpretation by evaluation of the identified descriptor-output correlations. The antibody structure depicts a modification of the PDB entry 1HZH [40].

investigated [15–18]. For a molecular level of detail, spectroscopic evaluation of protein-ligand pairs combined with protein labeling techniques such as atomic force microscopy, nuclear magnetic resonance spectroscopy, or mass spectrometry have been employed [19–22]. Additionally, molecular dynamics simulations and the calculation of theoretical physicochemical properties shed light on the complex protein-ligand interactions [23–26]. To bridge the gap between the molecular-level of detail and macroscopic observations, quantitative structure-property relationship (QSPR) models were developed, restricted by the amount of available data [15,27–33]. Conversely, automated screening setups using liquid-handling stations or controlled pH-gradients have been utilized to support process development or assess molecule manufacturability profiles [34,35].

Despite recent advancements in increasing the process understanding of multimodal mAb purification, the complexity and sensitivity of multidomain proteins towards multimodal interactions limits their widespread application in biopharmaceutical development [36,37]. Predictive tools to facilitate the integration of multimodal chromatography into the mAb purification platform and assess molecule manufacturability remain scarce, while demonstrating significant success in other areas of the development cycle, such as candidate screening or the prediction of agglomeration propensity during formulation development [38,39].

In this study, we developed a QSPR model to predict mAb retention in multimodal chromatography during linear pH gradient elution, as depicted in Fig. 1. Initially, a comprehensive dataset was acquired comprising antibody-specific descriptors calculated from homology models and chromatographic pH retention for 64 full-length mAbs, including multiple IgG-like derivatives. Subsequently, an empirical model was developed using Gaussian Process Regression (GPR) and thoroughly validated. Finally, the GPR model was interpreted, providing insights into the multimodal interaction mechanism. The validated model can be employed to support process development and enable a candidate manufacturability assessment based solely on sequence information. Moreover, the mechanistic insights can contribute to the development of advanced adsorption models, transitioning from a macroscopic process understanding to the molecular level.

2. Material and methods

2.1. Chromatography resin, buffers, and molecules

In this study, the multimodal strong anion exchanger Capto adhere (Cytiva, Marlborough, USA) was utilized during the chromatographic experiments. A prepacked Capto adhere HiScreen column (7.7 × 100 mm, Cytiva) with a column volume (CV) of 4.7 mL was employed, as detailed in Section 2.2. The resin surface is functionalized with the N-Benzyl-N-methyl ethanol amine ligand, as depicted in Fig. 1c. This ligand exhibits multimodal functionality due to its capacity for ionic interaction, hydrogen bond formation, and hydrophobic interactions [9].

All buffer substances were purchased from Sigma-Aldrich Co LLC (Saint Louis, USA), while ultrapure water was filtered with the Milli-Q Advantage A10 (Merck Millipore, Burlington, USA) water purification system. The linear pH gradients necessitated a multicomponent buffer system compatible with anion exchange chromatography. Consequently, an anionic multicomponent buffer was selected to avoid the introduction of unspecified counterions while providing a broad buffer capacity within the pH range of 9.0 to 4.0 [40]. The buffer system was adapted from Kröner and Hubbuch [41] and consists of 9.1 mM 1, 2-ethanediamine, 6.4 mM 1-methylpiperazine, 13.7 mM 1,4-dimethylpiperazine, 5.8 mM bis-tris, and 7.7 mM hydroxylamine. In addition, 125 mM sodium chloride and 75 mM hydrochloric acid were incorporated, resulting in a total of 200 mM chloride counterions and a conductivity of 20 mS/cm. The addition of sodium chloride was required to achieve mixed-mode behavior and increase protein solubility [42]. Furthermore, the increased conductivity values enabled the augmentation of cation exchange elution pool as load material, which is regularly employed prior to the salt tolerant Capto adhere resin within the antibody purification process [42]. Thereafter, the equilibration and the elution buffer were titrated to pH 9.0 and 4.0 using 1 M sodium hydroxide. Other buffers used in the chromatographic experiments included 1 M acetic acid for column regeneration, 1 M sodium hydroxide for column cleaning, and 20 % ethanol for column storage.

The study involved 64 full-length IgG derivatives (Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach, Germany), comprising 62 human-origin and 2 humanized murine-origin antibodies. These antibodies displayed an extensive range of physicochemical parameters, as evidenced by their widely distributed elution behavior, shown in Fig. 2a. The antibody set included 33 IgG1s, 20 IgG bispecifics with two single-chain fragment variables (scFv) appended to each heavy chain C-terminus (IgG(H)-scFv), 8 IgG4s, 2 Knob-in-Hole bispecifics (KiH), and 1 KiH trispecific with a single scFv attached to the C-terminus of the Hole chain (KiH-scFv). The antibody expression was achieved using a stably transfected Chinese hamster ovary cell line, followed by capture through

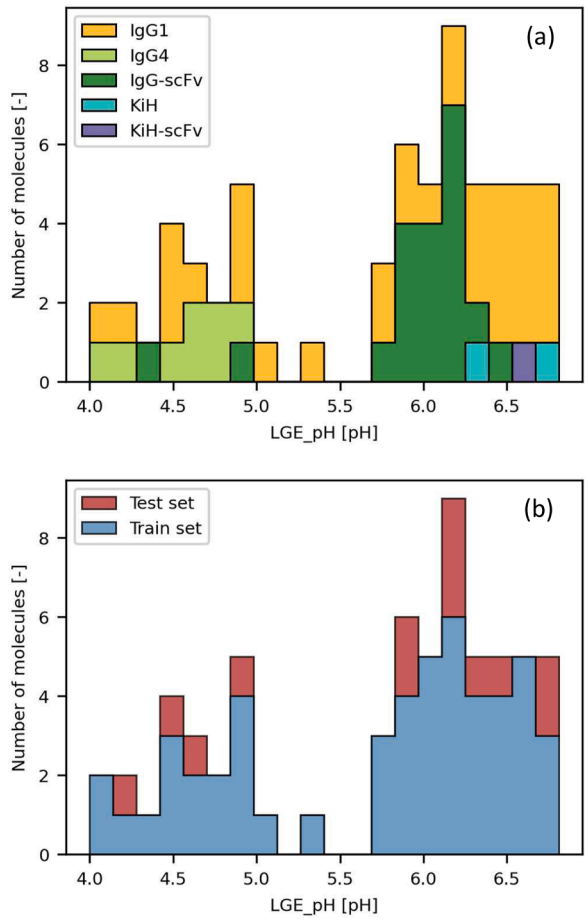

Fig. 2. Distribution of pH retentions derived from linear pH elution experiments. a) Histogram illustrating the antibody format-specific retention, and b) Histogram displaying the distribution of model training and testing split data.

protein A affinity chromatography. After neutralizing to pH 5.5 and sterile filtration using a 0.2 μm filter, the protein solutions were frozen at −70 °C. Upon thawing, the final load material was adjusted to a concentration of 5 g/L, as determined by a NanoDrop 2000c spectrophotometer (Thermo Fisher Scientific, Waltham, USA). Prior to sample application, the load material underwent buffer exchange into the equilibration buffer using a 5 mL HiTrap Desalting column (Cytiva) according to the manufacturer's instructions.

2.2. Linear pH gradient elution

In this study, all 64 antibodies were eluted by a linear pH gradient using the multicomponent buffer system described in Section 2.1. The chromatographic experiments were conducted with an ÄKTA Avant 25 (Cytiva) preparative chromatography system, controlled by the Unicorn 7.5 (Cytiva) control software, and maintained at a residence time of 5 min. Initially, the column was equilibrated for 3 CV with the pH 9.0 buffer. The equilibration was followed by column loading with the antibody solution up to a loading density of 1.0 g/L. Subsequently, the antibodies were eluted by linearly decreasing the pH from 9.0 to 4.0 within 10 CV using the elution buffer, while maintaining a constant conductivity of 20 mS/cm, as illustrated in Fig. 1a. The retention times of the molecules were determined by measuring the first moments of the elution peaks through the UV trace at a wavelength of 280 nm. Subsequently, the corresponding pH values at peak retention were determined by correcting the online pH trace of the chromatography system with the offline pH measurement of the equilibration and elution buffer, as well as accounting for the pH sensor dead volume. Following column elution, a 4 CV column regeneration step, a 5 CV cleaning in place procedure, and 4 CV column storage step were appended.

2.3. Antibody homology modeling

The prediction of the antibody structures was achieved through homology modeling, which was required for the subsequent calculation of physicochemical descriptors, as previously described by our group [43]. Hereto, the molecular modeling and visualization environment Maestro Bioluminate 4.9 (Schrödinger Inc., New York, USA) was employed for structure prediction, evaluation, and model refinement.

The initial homology modeling workflow was adapted from Zhu et al. [44] and comprises an automated five-stage process, which includes: (1) framework and complementary-determining region (CDR) template selection; (2) variable region model grafting; (3) CDR loop modeling and sidechain prediction; (4) full-length antibody modeling; (5) energy minimization. Within the workflow, the antibody numbering scheme Enhanced Chothia [45] was utilized. The modified crystal structure of a human IgG1, 1HZH [46], shown in the Fig. 1a/c, served as a full-length template for all molecules except for the IgG4 subtypes, which employed the human IgG4 crystal structure 5DK3 [47] instead. For structural prediction of the complex bi- and trispecific formats, intra- and intermolecular linkers were grafted using homology modeling and ab initio prediction to append the independently modeled scFv domains to the full-length mAb structures [48]. Hereafter, the initial homology models were further refined according to the protocol of Sastry et al. [49]. In brief, the protocol includes preprocessing steps to modify and validate the hydrogen network, bond order assignment, as well as atom naming and numbering. The preprocessing is followed by energetic optimization of terminal hydrogen atoms from the amino acid side chains and the assignment of protonation states of ionizable groups using PROPKA3 [50]. Lastly an all-atom energy minimization is conducted using the OPLS4 forcefield with nonhydrogen atoms being constrained to a root-mean-square deviation of 0.3 Å [51].

2.4. Antibody-specific descriptor calculation

Following the antibody homology modeling, physicochemical descriptors were calculated from the protein structures using the molecular modeling and visualization environment Bioluminate 4.9 (Schrödinger Inc., New York, USA). The utilized descriptor set comprises 165 unique features derived from first principal models, as well as parameterized empirical models [52]. Moreover, the descriptor set can be subdivided into sequence-based descriptors (n = 69) devised from bioinformatic scales, structural descriptors (n = 59) encoding for geometric and electrostatic properties of the molecule, and patch-specific descriptors (n = 37) calculated from the hydrophobic and electrostatic energy of the protein surface. Herein, the surface hydrophobicity is calculated by employing the atomistic Wildman and Crippen logP parameters [53], whereas the electrostatic surface potential is calculated from partial charges based on the OPLS4 [51] forcefield as described by Sankar et al. [54]. Furthermore, proximal hydrophobic and electrostatic surface characteristics are combined into aggregation propensity descriptors. The final patch descriptors are derived by binning the calculated surface properties into quantifiable features based on their interaction type (positive, negative, hydrophobic), size, intensity, and number.

To increase the resolution of the descriptor set further, a region-specific subset (n*=31) of the initial descriptor set is selected and calculated for 37 subdomains of the antibody structure. The antibody-specific subdomains comprise the light and heavy chain variable regions (VL, VL_Fv, VH, VH_Fv), emphasizing the complementarity-determining regions (CDR, CDRL, CDRH) and framework regions (FR, FRL, FRH), which consist of individual loops (L1, L2, L3, H1, H2, H3), and frameworks (LFR1, LFR2, LFR3, LFR4, HFR1, HFR2, HFR3, HFR4). On top of the variable region, the antibody constant regions (CL, CH1, CH2, CH3) and the hinge region (Hinge) are considered. Furthermore,

the regional descriptors are extended by seven custom regions to account for the fragment variable (Fv), the fragment antigen binding (Fab), the fragment crystallizable (Fc), as well as the sum of the constant regions (CR). Additionally, the single-chain fragment variable regions (scFv, VLscFv, VLscFv) of the bi- and trispecific formats are considered. The final descriptor set comprises 1312 features per molecule including the initial 165 global descriptors (All) and 1147 local descriptors, providing detailed information about the physicochemical topology of the IgG-like structures. Given the large quantity of the descriptors, the descriptor naming scheme combines the descriptor location, interaction type, and binning strategy, as exhibited in Section 3.2.

2.5. QSPR model development and evaluation

A multivariate regression model was established to predict the linear pH gradient retention from the antibody-specific descriptor set. Due to the high dimensionality of the regression problem, with $N = 1312$ features per antibody, paired with the comparably small data set size of $M = 64$ observations, descriptor preprocessing, dimensionality reduction, and model evaluation were required. The QSPR workflow was developed with and evaluated by Python 3.9.12 in conjunction with the machine learning package scikit-learn 1.0.2 [55]. GPR was utilized to predict non-linear relationships between the target vector $y = \{y^{(m)}\}_{m=1}^{M}$ and the feature matrix $X = \{X^{(n)}\}_{n=1}^{N}$, while providing a heteroscedastic uncertainty estimation [56,57]. GPR is based on Bayesian inference and involves prior assumptions regarding the underlying target function $y(x)$ that can later be updated in course of the Bayesian update rule, shown in Eq. (1).

$$P(y(x)|\mathscr{D}) \propto P(y|y(x), X)P(y(x)) \tag{1}$$

Within the Bayesian framework, the model predictions are derived as the posterior $P(y(x)|\mathscr{D})$, which is a gaussian distribution of functions conditioned to fit the training data $\mathscr{D} = \{y, X\}$, as depicted in Fig. 1b. During model training, the prior $P(y(x))$ that defines the similarity and the smoothness between the observations is conditioned by maximizing the likelihood $P(y|y(x), X)$ of the mean and the variance from the posterior distribution to reflect all training data. In this study, the prior is derived as a mixed covariance function by multiplying a linear kernel with a Matérn class kernel and subsequent addition of a white noise kernel [55,57,58]. The addition of a noise kernel is necessary to avoid model overfitting by specifying the uncertainty of the measured data as visualized by the amplitude of the posterior distribution depicted in Fig. 1b. Subsequently, the model is conditioned by minimizing the log (marginal likelihood) (LML) of the posterior distribution using the L-BFGS-B algorithm [59].

The QSPR workflow is initiated by data preprocessing, where empty, positional, non-informative, and redundant descriptors are discarded. Additionally, several operations are performed on the descriptors to account for the structural diversity of the IgG-like molecules. The regional antibody descriptors are multiplied by the frequency of the given region within the multimer protein, sparse antibody regions imputed as zero, and the descriptor regions calculated for both KiH knob and hole chains, averaged based on the analysis of Parasnavis et al. [18]. The data set is then randomly split into 80 % training data and 20 % test data. Lastly, the descriptors are scaled by their standard deviation (SD) and centered based on the training data.

After data preprocessing, dimensionality reduction is performed by removing invariant descriptors, which decreases the risk of model overfitting and increases model interpretability [60]. Low variance ($cov(x,x) \leq 0.01$) features are discarded, and the remaining descriptors are sorted based on the results of a F-test from a univariate linear regression model with the target variable. Hereafter, collinear features are removed based on Pearson correlation ($\rho \geq 0.80$), and ten highest scoring features are selected following the F statistics. Lastly, recursive feature elimination (RFE) is conducted by iteratively removing the lowest-ranked features according to feature permutation importance [61]. Permutation importance is defined as the average increase in model deviation when accessing the model performance after shuffling a single feature one hundred times while keeping the remaining features constant. At each iteration of the RFE procedure, the LML of the current model and the mean absolute error (MAE) of leave-one-out cross-validation are calculated to identify the overall best model.

The last step of the QSPR workflow comprises model evaluation to increase the understanding of the underlying adsorption mechanisms of antibodies in multimodal chromatography, as visualized in Fig. 1c. The model evaluation includes an assessment of the overall model reliability and performance. On top of that, an investigation of the feature interdependence, sensitivity, and their contribution to the model predictions is conducted, to enable mechanistic interpretability of the model [62]. The model performance is assessed through inspection of goodness of fit to the training data and goodness of prediction of the test data, including an estimation of the model 95 % confidence interval per observation. Furthermore, fivefold cross-validation with ten repetitions is employed for internal validation of the training data. On the other hand, model reliability is analyzed by y-scrambling the full data set one hundred times with subsequent calculation of the MAE from leave-one-out cross-validation [63]. Feature interdependence is evaluated by investigating the pairwise relationships between model features, as well as the target variable. Finally, the feature sensitivity and contribution are assessed by means of feature permutation importance and partial dependence towards the model prediction [61,64,65].

3. Results and discussion

3.1. Elution behavior of antibody formats

In the course of this study, a large and structurally diverse set of IgG-like molecules was examined. All 64 full-length IgGs could be eluted from the anionic mixed-mode resin Capto adhere during linear pH gradients from pH 9.0 to 4.0 at a constant conductivity of 20 mS/cm. Fig. 2 depicts the first moments of the elution peaks from each molecule. Upon inspecting the elution distribution, a bimodal trend is apparent separating the molecules in two groups of antibodies. The first group elutes at a lower pH range (pH 4.00–4.99, $m = 20$) compared to the second group that elutes at a higher pH range (pH 5.71–6.81, $m = 43$). In contrast to the overall distribution mean of pH 5.72, the lower and the higher elution groups are centered around pH 4.59 and 6.26, with a single IgG1 laying in-between both groups at a pH of 5.33. To investigate the elution behavior further, Fig. 2a compares the retention of different antibody subclasses and formats that were analyzed in this study, as detailed in Section 2.1. The IgG1 subclass depicts the most abundant format within the data set and exhibits the broadest retention distribution across both pH groups. Interestingly, the distribution of the structurally homologue IgG1 antibodies follows the same trend as the full data set, being centered in the higher pH group and skewed towards lower pH values. This observation emphasizes that multimodal elution is not necessarily dependent on the overall size or shape of the molecule but physicochemical properties that are distributed within a structural homologue set of proteins, as observed in multiple studies [15,16,31]. The IgG-scFv formats exhibit a similar trend, while displaying a narrower pH distribution with only two molecules eluting in the lower pH group despite having two additional scFv regions attached to the Fc domain. Moving forward, the IgG4 antibodies exclusively elute in the lower pH group, whereas all KiH formats elute at the upper boundary of the observed pH values with a maximum elution pH of 6.79. As only three molecules are associated with the KiH format, no inference towards this antibody class can be conducted. On the other hand, the main structural distinction between the IgG4 subclass and the other antibodies is given by the Fc domain, as the remaining formats share similar IgG1 backbones. The difference in the elution behavior arising from deviating Fc regions leads to the assumption that multiple binding

domains on the whole antibody surface exist, as reported in numerous studies conducted by the Cramer lab involving cationic mixed-mode resins [17,18,20,22,24,26]. In general, IgG4 backbones exhibit fewer acidic residues leading to a lower pI, as well as increased surface hydrophobicity in comparison to IgG1 backbones [66]. These two characteristics align with the experimental observation of the IgG4 formats to elute in the lower pH group and the mechanism of hydrophobic charge induction chromatography (HCIC) [23,67]. According to the HCIC mechanism, molecules adsorb through hydrophobic attraction and desorb with increasing electrostatic repulsion in course of a pH modification. The lower pI of the IgG4 formats would result in a reduced positive surface charge at an acidic pH environment, which in turn would lead to a reduced repulsion towards the anionic multimodal ligand.

3.2. QSPR modeling of pH gradient retention

An integral part of statistical model development is the identification of a predictive feature set and a suitable mapping function followed by the assessment of model quality. In this study, a GPR model was used to regress the pH retention of a large antibody set ($M = 64$) to their physiochemical properties, which were encoded into 1312 descriptors per molecule, as detailed in Sections 2.4 and 2.5.

Initially, the data set was divided into training and testing data for the validation of the empirical model. Fig. 2b displays the distribution of randomly selected molecules into 20 % test set ($m = 13$) and 80 % training set ($m = 51$), ensuring a representative distribution of pH retentions and molecule formats. Within the test set, four molecules categorize to the lower pH group and nine molecules to the higher pH group. Furthermore, the test set comprises six IgG1, four IgG-scFv, two IgG4, and one KiH molecule yielding a robust test set selection, as listed in Table 1.

Thereafter, a two-staged feature selection was conducted using the training data to reduce 1102 preprocessed descriptors to six features. The first stage involved filter methods to efficiently discard the majority of uninformative or convoluted descriptors. Initially, low variance ($cov(x,x) \leq 0.01$) features were removed, reducing the feature number to 1083. The remaining features were sorted according to their unimodal interaction towards the target variable, using linear regression models. Multicollinear descriptors with significant Pearson correlation ($\rho \geq 0.80$) were removed, as suggested by Sankar et al. [52]. Although collinearity not necessarily diminishes model predictiveness, the removal of multicollinear descriptors was required to improve mechanistic interpretability. From the remaining 413 features, ten descriptors were selected based on their linear relationship towards the pH retention, as depicted in the bottom x-axis of Fig. 3. The selected features

Table 1
Overview of goodness of test set prediction and model uncertainty.

Molecule	Observed [pH]	Predicted [pH]	Residual [pH]	SD [pH]	CI_{95} [pH]
IgG1 (1)	4.18	4.23	0.05	0.44	3.36–5.09
IgG1 (2)	4.45	4.69	0.24	0.29	4.12–5.25
IgG4 (3)	4.7	4.75	0.05	0.53	3.72–5.79
IgG4 (4)	4.98	5.00	0.02	0.57	3.89–6.11
IgG-scFv (5)	5.83	6.17	0.34	0.27	5.65–6.70
IgG1 (6)	5.94	6.26	0.32	0.39	5.50–7.01
IgG-scFv (7)	6.12	6.48	0.36	0.25	5.98–6.97
IgG-scFv (8)	6.13	5.71	0.42	0.26	5.20–6.22
IgG-scFv (9)	6.22	6.15	0.07	0.27	5.61–6.68
IgG1 (10)	6.32	5.88	0.44	0.26	5.38–6.38
IgG1 (11)	6.51	6.70	0.19	0.34	6.05–7.36
IgG1 (12)	6.71	6.44	0.27	0.25	5.94–6.93
KiH (13)	6.79	6.56	0.23	0.36	5.84–7.27

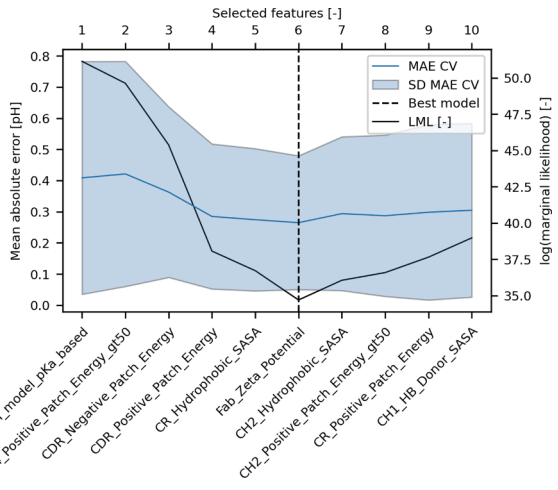

Fig. 3. Recursive feature elimination with cross-validation of training data. Starting with ten features on the right-hand side of the figure, a model is initially fitted using all available features. At each iteration, the model is evaluated by calculating the cross-validation mean absolute error (MAE CV), the standard deviation (SD) of cross-validation scores, and the log(marginal likelihood) (LML) of the given model. The feature with the lowest permutation importance is eliminated per iteration, as indicated on the bottom x-axis. Once all but one feature is removed, the best model with the lowest log(marginal likelihood) is selected.

provided insights into the multimodal binding mechanism of IgG-like molecules, as well as relevant antibody domains. Mostly charge-related descriptors were selected, with eight out of ten descriptors encoding for electrostatic interactions. The two remaining descriptors gave insight into hydrophobic contributions. The small representation of hydrophobic descriptors could be a result of the weak pH dependency of hydrophobic attraction compared to electrostatic interaction [24,68]. The strongest linear relationship was observed for the isoelectric point (pI) of the molecules termed “pI_model_pKA_based”, which categorizes as a global and structure-based descriptor. The other descriptors were either pointing towards the Fab domain or the constant region, as well as local descriptors within both regions. Furthermore, solely structure-based and patch descriptors were identified during the first stage of feature selection, while the relationship towards sequence-based descriptors deemed less significant. Comparable global descriptors have been identified in multiple studies, including the structure-based molecule pI [15,27,28,30–32]. Interestingly, no aggregation-propensity descriptors were selected, which in theory should be able to encode for adjacent electrostatic-hydrophobic interactions as observed for Capto adhere in a previous study [32]. On the other hand, proximal interactions have not been reported in a more recent study concerning Capto adhere, selecting from a significantly larger pool of initial descriptors compared to the previous study [69]. It remains an open question whether adjacent patch interactions are not as pronounced in Capto adhere binding as reported for multimodal cationic resins [16,31] or if the current class of proximal descriptors [30,54] is not sufficiently parameterized to describe multimodal anionic interactions. In addition, no descriptors from the custom scFv region were chosen during the filter process, which could indicate a minor role of this region in antibody binding. A possible explanation for the less significant role of the scFv domain compared to the Fv domain of the bispecific mAbs could be related to steric constraints due to the engineered inter- and intrachain linkers limiting the configurational flexibility of this region [70]. To answer these questions, more research is required to investigate local ligand interactions as by using recent labeling methods [22].

In the second stage of feature selection, recursive feature elimination (RFE) was employed based on model performance, as illustrated in

Fig. 3. A GPR model with a mixed covariance function was utilized to estimate the target function. The Bayesian method was chosen to account for both linear and nonlinear feature contributions in the multimodal binding mechanism, while accurately estimating heteroscedastic prediction uncertainty, as previously observed in cation exchange chromatography [43]. During the RFE process, GPR models were sequentially fitted to the training data, starting with the ten most significant descriptors on the right-hand side of Fig. 3 and ending with a single feature for the final model on the left-hand side of the figure. At each iteration, the feature with the lowest permutation importance or the smallest impact on model accuracy was removed. The calculation of feature permutation importance was required since no meaningful model weights could be extracted from the non-linear GPR model. Simultaneously, the log(marginal likelihood) of the model conditioning and the cross-validation mean absolute error of model prediction were recorded at each iteration.

Upon evaluating Fig. 3, a distinct trend for both the LML and the cross-validation MAE, along with its standard deviation is displayed, identifying the optimal GPR model as the one using six features. The LML of model conditioning ranged from 34.698 for the sixth to 51.133 for the final RFE iteration. Meanwhile, the cross-validation MAE fell within a pH range of 0.265 pH for the sixth to 0.421 pH for the penultimate iteration. Furthermore, the SD of cross-validation was lowest at the selected model at 0.215 pH, indicating increased model robustness compared to the largest SD observed for the final RFE iteration with a MAE SD of 0.373 pH. Overall, the cross-validation score appeared more susceptible to variation in identifying the most predictive model compared to the Bayesian likelihood, thus demonstrating the advantage of the Bayesian approach, not only for feature mapping, but also for feature selection.

The six selected features comprised the molecule pI as a global charge-related descriptor, three local charge descriptors within the antibody Fv domain, namely "LFR4_Positive_Patch_Energy_gt50", "CDR_Negative_Patch_Energy", and "CDR_Positive_Patch_Energy", as well as electrostatic and hydrophobic contributions from the constant region and the Fab domain through the descriptors "CR_Hydrophobic_SASA", and "Fab_Zeta_Potential". According to the descriptor naming scheme, "LFR4_Positive_Patch_Energy_gt50" is defined as the summation of all positive patches within the antibody LFR4 region with a patch area larger than 50 $Å^2$. Interestingly, the selected features included two custom descriptors, encoding for the antibody constant region and its Fab domain. In contrast, the local descriptors within the constant region were discarded, despite multiple studies highlighting the importance of the Hinge region and the CH2–CH3 interface in multimodal interaction [20-23,26]. Furthermore, the removal of the descriptors from the constant region and the Fab domain had less impact on the LML and the cross-validation MAE compared to the removal of local descriptors within the Fv domains. This finding emphasizes the significance of the Fv domain, particularly the CDRs, for antibody binding to multimodal chromatography resins, as observed by multiple authors for cationic MMC [16–18,24,31,33]. A comprehensive evaluation and interpretation of the model features will be provided in Section 3.3.

Following the feature selection, the predictiveness of the final QSPR model was evaluated *via* internal and external model validation. Accordingly, the goodness of fit to the training data and the goodness of prediction for the test data were assessed, as depicted in Fig. 4. Both the training set and the test set, consisting of 80 % ($m = 41$) and 20 % ($m = 13$) of the antibodies from the full data set are displayed. Furthermore, the pH observations and predictions of the test set, as well as molecule residuals and uncertainties are listed in Table 1. Again, an agglomeration of data points into a lower eluting and a higher eluting pH group is evident. The upper part of the figure compares the observed and predicted pH retention, while the lower part focuses on the distribution of the molecule residuals. Additionally, linear fits to the scattered training and test set molecules are displayed to enable a quick assessment of

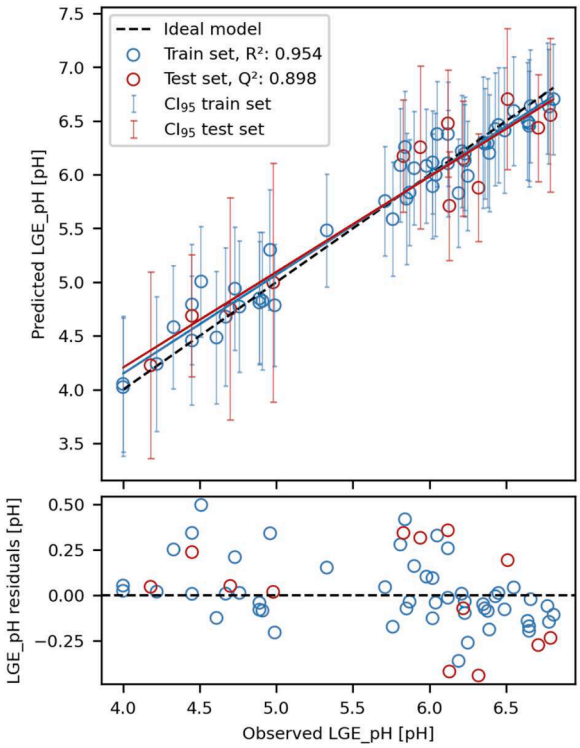

Fig. 4. Goodness of pH elution prediction (R^2, Q^2). The upper part of the figure compares the predicted with the experimental pH elution for the molecules in the random training and the testing data. An ideal model is represented by a straight line, where predicted and experimental observations have zero error. For each molecule, the 95 % confidence interval (CI_{95}) of prediction is calculated. The bottom part of the figure displays a residual plot, with the y-axis normalized to the absolute deviation of model prediction.

model performance against a theoretical ideal model with zero error in predictions. On the top figure, the heteroscedastic 95 % CI for each molecule is indicated, while the lower figure presents uniformly dispersed residuals, signifying the absence of systematic model errors. The model's uncertainty varies between one and two pH units, while the overall accuracy demonstrates a maximum deviation of 0.44 pH for IgG1 (10), as outlined in Table 1. This increased uncertainty, in comparison to the prediction accuracy, was necessary to avoid model overfitting, as discussed in Section 2.5. The source of this uncertainty might be attributed to the descriptor set's inaccuracy to fully capture the pH sensitivity of multimodal protein adsorption [3,7,13,24].

When evaluating the model performance metrics, the fit of the final GPR model to the training data achieved a coefficient of determination of $R^2 = 0.954$, with a MAE of 0.132 pH, and an average SD of model uncertainty of 0.281 pH. Internal model validation of the training data through five-fold cross-validation with ten repetitions resulted in a mean Q^2 of 0.780 and a MAE of 0.279 pH. Lastly, the pH retentions of the external test set could be predicted with a Q^2 of 0.898, a MAE of 0.231 pH, and an average SD of model uncertainty of 0.344 pH, as detailed in Table 1. Comparing the results of the model training and testing, similar scores were achieved for the R^2 and external Q^2, despite showing a decreased model accuracy of approximately 0.1 pH, as well as an increased uncertainty in the model predictions. The similarity of these quality metrics underlines the robustness of the empirical model, as significantly diverging training and test results would suggest model overdetermination. The model's robustness is further supported by the y-scrambling results in Appendix Fig. A1, which indicates a less than 1 % probability of achieving the model performance by chance. Continuing with the internal model validation, a divergence between the model fit and cross-validation scores is apparent. Larger divergence during cross-validation is a common phenomenon of empirical models and becomes

especially pronounced for small data sets, as the missing data used for subsampling can impair model performance.

Upon examining the 95 % CI of the model predictions, comparably large intervals are apparent for the molecules eluting in the lower pH group. When evaluating the SDs listed in Table 1, the three largest model uncertainties were observed for the molecules IgG4 (4), IgG4 (3), and IgG1 (1) in a descending order, with SDs of 0.57, 0.53, and 0.44 pH, respectively. The elution pH of both IgG4 formats within test set was predicted with the highest uncertainty among all molecules examined. The substantial 95 % CI of the two IgG4s may indicate an antibody format-dependent uncertainty captured by the empirical model, as all formats except the IgG4 type molecules shared IgG1 backbones, as previously discussed. Exemplarily, the KiH antibody was predicted more accurately than the IgG4 formats, despite its bispecific or trispecific functionality. Conversely, the molecule with the third largest uncertainty was a standard IgG1, eluting at the lower end of the pH spectrum. This observation could be attributed to the scarcity of data points in close vicinity to the IgG1 (1), compared to molecules in the higher eluting pH, potentially leading to impaired model performance.

In conclusion, the predictive power of the QSPR model relies on both the density and the overall number of molecules sharing similar physicochemical surface characteristics. In comparison to peer studies on (homologue) proteins libraries using a single resin system, our model demonstrates a superior performance [15,27,28,30,31]. However, previous models only employed a third of the number of molecules used in this study. Considering the performance and broad applicability of the validated model in accurately predicting pH retention for a wide range of commercially available antibodies based solely on their amino acid

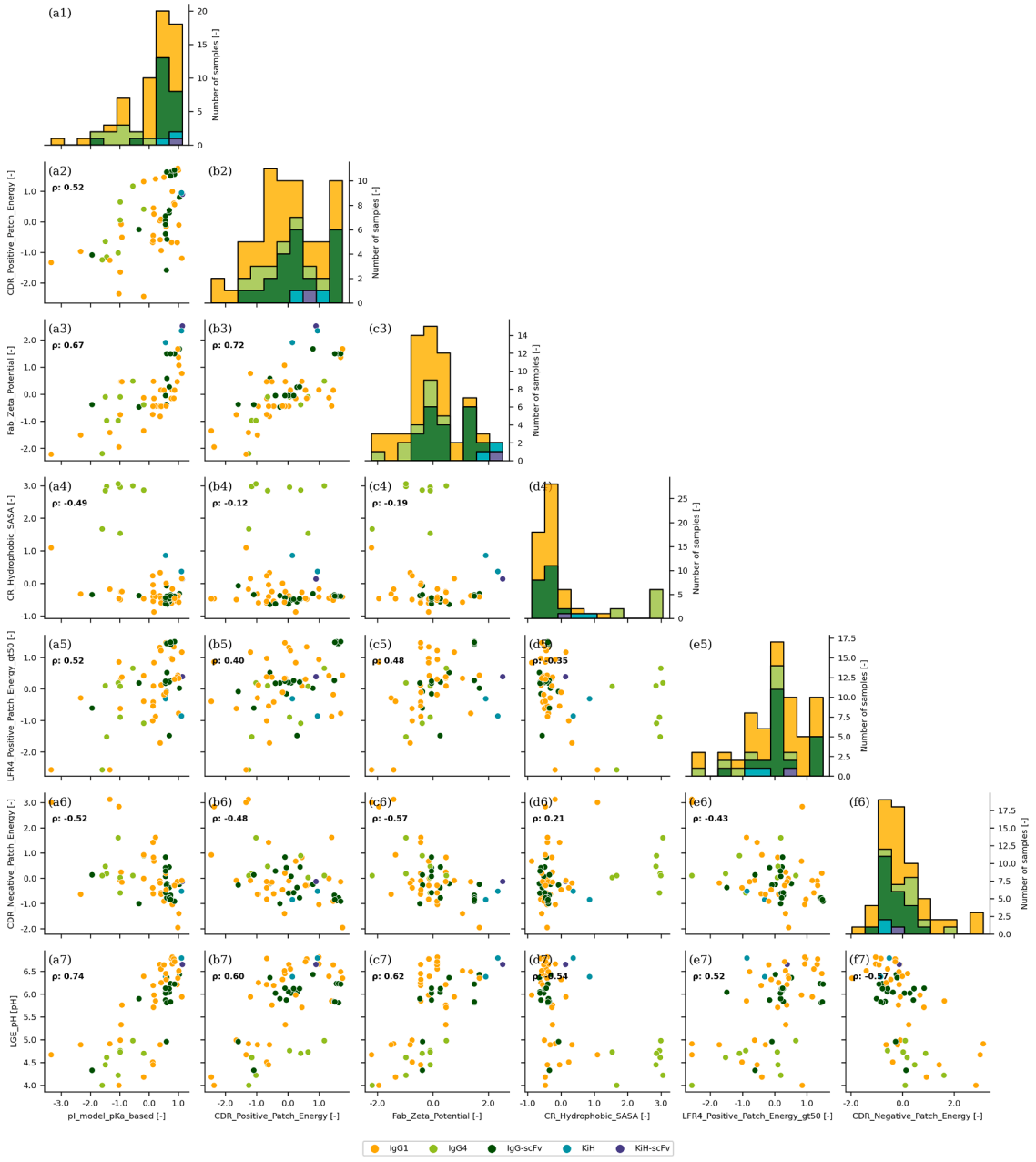

Fig. 5. Pairwise relationship and distribution of model features and pH elution. The diagonal subfigures display the univariate histograms of the model features, as shown on the bottom x-axis. The off-diagonal subfigures illustrate the bivariate relationship between model features, while the bottom row presents the relationship between the target variable and the model features. Features are sorted in descending order based on their absolute Pearson correlation coefficient (ρ) towards the target variable within the training data, indicated in the upper left corner of each subplot. All subfigures differentiate between antibody formats.

sequence, two potential applications emerge. First, the model can be employed to predict an initial pH set point for the anionic mixed-mode resin Capto adhere during early process development, which is especially advantageous for multi-domain proteins as mAbs concerning the pH-sensitivity of their domain contributions to chromatographic retention [24]. Second, it can serve as an in silico screening tool for identifying molecule manufacturability in regards to downstream purification.

3.3. Model inspection and interpretation

After the final QSPR model was established, the feature contributions within the model were further analyzed to investigate the interaction mechanism of the antibodies to the anionic multimodal resin illustrated in Fig. 1c. Throughout this study, in particular charged-based descriptors within the variable region appeared correlated to the multimodal interaction. Therefore, a model inspection was conducted, identifying interdependencies between the selected features, to support the mechanistic interpretability of the final model. Fig. 5. depicts the histogram of the standardized and mean centered feature values for the entire data set on the top diagonal axes (column a – f), as well as the pairwise relationship of model features on the off-diagonal axes (row 2 – 6). Additionally, the pairwise relationship of the model features to the target variable is appended to the figure bottom (row 7) and sorted according to the linearity of the training set features to the pH retention. Furthermore, the different molecule formats are indicated in each subplot and the Pearson correlation coefficient p, calculated from the entire data set.

When evaluating the individual feature distributions shown in the top diagonal axes of Fig. 5, similar trends to the pH retention introduced in Fig. 2a are evident. In both figures, the IgG1 and IgG-scFv formats display the broadest distribution throughout the parameter space, whereas the IgG4 and KiH formats show a clustered behavior. Notably, the distribution of the molecule pI, shown in subfigure (a1) of Fig. 5 is comparable to the distribution of the target variable, shown in Fig. 2a. The similarity between the molecule pI distribution and the pH retention distribution is supported by their strong positive linear correlation of $p = 0.74$ shown in Fig. 5, subfigure (a7). The skewed and bimodal distribution of the pH retention shown in Fig. 2a appears to be primarily influenced by the molecule pI. In this context, the abundance of elevated pI values within the full data set, in contrast to the significantly lower pI values of the IgG4 formats, could explain the bimodal appearance of the pH retention distribution. The global pI, however, depends on individual contributions from the antibody subdomains. Consequently, further insight into the antibody interactions can be obtained by inspecting the zeta potential distribution of the Fab domain in Fig. 5, subfigure (c3), as well as the hydrophobic accessible surface area (SASA) distribution of the constant region shown in subfigure (d4). The IgG4 formats exhibit a reduced zeta potential of their Fab regions, while the three KiH formats define the upper limit of zeta potential. The surface charge of the Fab not only translates to an increased pI, as assessed by their positive correlation in subfigure (a3), but also correlates positively to the pH retention, as depicted in subfigure (c7). Moreover, the Fab zeta potential is correlated to the electrostatic surface potential of the CDRs encoded *via* their positive and negative patch energy, shown in subfigure (b3) and (c6). Upon inspecting the distribution of hydrophobic SASA from the antibody constant region in subfigure (d4), again, two groups are apparent. The first group represents less hydrophobic molecules with a IgG1 backbone, while the second group is comprised of IgG4 antibodies exhibiting increased surface hydrophobicity, which it is expected considering the conserved nature of the Fc domain. In this context, the hydrophobic clustering behavior might obscure the true importance of the constant region to the multimodal binding by weakening the linear relationship of its contribution.

In conclusion, the strong correlation of the pH retention to Fab domain, and particularly, the antibody CDRs does not explicitly imply a Fab-first binding orientation but suggests a complex interaction mechanism that could depend on multiple binding domains on the entire antibody surface. This assumption is supported by the strong correlation of the global pI to the pH retention and the results published by Robinson et al., studying domain contributions and pH dependency of the multimodal antibody interaction [17,24].

Intriguingly, the energy of large positive patches in the LFR4 region shown in Fig. 5, subfigure (e7) exhibited no significant correlation with the target variable, other features, or clustering behavior. However, it was removed second to last during the RFE process depicted in Fig. 3, indicating the potential relevance of this region.

To further investigate the descriptor contributions and finalize the mechanistic interpretation, the partial dependences of the features within the GPR model were analyzed, as depicted in Fig. 6. Partial dependence allows for examining the strength and form of non-linear feature contributions to a multivariate model, as well as identifying feature interdependencies. In brief, a single factor perturbation is performed by marginalizing all but one feature and permutating it within the full feature range. Subsequent recording of output predictions for a single molecule result in its individual conditional expectation (ICE), while averaging over all ICEs yields the partial dependence of the feature. Consequently, heterogeneous behavior of the ICEs can reveal feature interdependencies. However, during the interpretation of partial dependence, it is crucial to consider multicollinearity within the inspected feature set, as implausible parameter pairs can form, such as the simultaneous occurrence of strong positive and strong negative patch energies in a specific region [71].

The partial dependences of the model features shown in Fig. 6 share the same order as displayed in the bottom row of Fig. 5 and follow similar relationships suggested by the scattered data. The feature distributions of each descriptor are indicated by deciles lines on the x-axis. Additionally, comparable permutation importance's are calculated in a range in between 0.12 and 0.18 pH for each feature and are displayed on the upper left corner of the subplots already employed during the RFE. Furthermore, a distinction between the training and test set is made, which implies robust model performance based on comparable functions of the partial dependences throughout the full feature space. Upon inspecting the ICE lines, all but the molecule pI and the hydrophobic SASA of the constant region shown in subfigure (d) display pronounced heterogeneous behavior. This behavior can be explained for the Fab and the CDR descriptors based on their mutual linearity, as discussed earlier. Moreover, the feature collinearity leads to underestimation of permutation importance, which implies a dominating role of the variable region during the multimodal interaction when adding up their individual permutation importance's. Only the heterogeneous ICE lines of the energy from large positive patches in the LFR4 region shown in subfigure (e) provide further insight into a potential interaction mechanism. It appears that for early eluting molecules during the pH gradient, increasing the positive patch energy in the LFR4 region, an area in-between the L3 loop and the VL-CL interface, has a negligible effect on the pH retention. Conversely, an increase of positive patch energy in the LFR4 region shifts the retention of strong binding molecules more than one pH unit. In conclusion, the ICE lines of the large positive patches in the LFR4 region suggest that the LFR4 region can contribute to a significant binding domain, as recently identified by Parasnavis et al. [18]. Furthermore, the contribution of the LFR4 region to the antibody binding is dependent from the LFR4 surface charge, leading to an increased adsorption in the absence of strongly repelling positive patches.

Considering all feature contributions, the multimodal antibody-ligand binding appears to be driven by linear contributions of electrostatic attraction and repulsion from the CDRs but further depends on significant binding domains throughout the entire antibody surface, as implied by the strong relationship to the global pI and the constant region. A significant role of adjacent electrostatic-hydrophobic patches in Capto adhere binding was not observed but might be attributed to

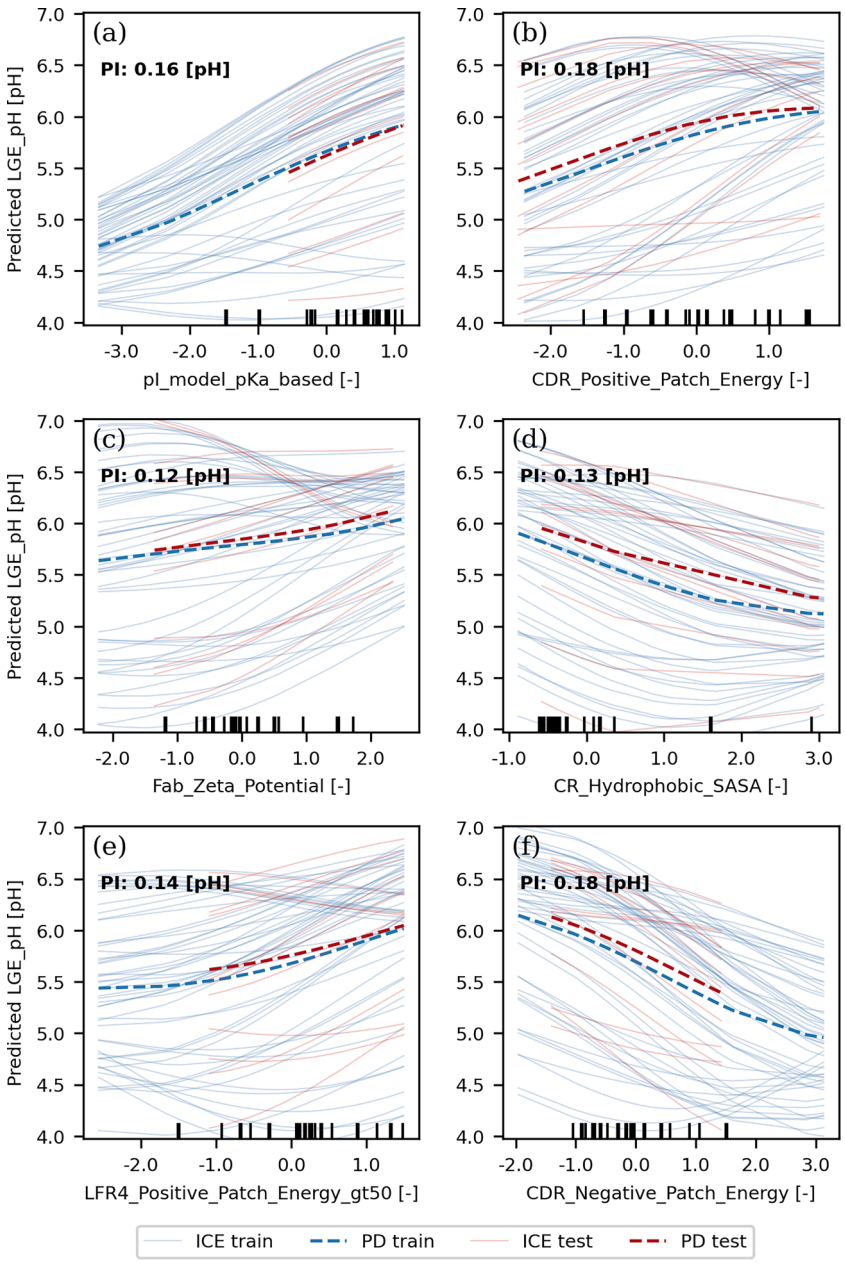

Fig. 6. Partial dependence of pH elution from model features of training and test data. Each subfigure depicts the model response when permutating the features in their normalized feature range. The individual conditional expectation lines represent the target prediction for each molecule, while the partial dependence of the feature is given by their mean values. The feature permutation importance (PI) towards the target prediction is shown in the upper left corner of each subplot. The decile lines at the bottom of each subplot indicate the frequency of the feature values within the data set.

inadequate descriptors to encode for multimodal anionic interaction, as discussed in Section 3.2.

4. Conclusion

In this study, we developed a QSPR model to predict pH gradient elution in multimodal chromatography using a diverse set of therapeutic antibodies. In total 64 IgG-derivates were included, categorized into five therapeutic antibody formats, which exhibited distinctive chromatographic behavior. Throughout the QSPR workflow, physiochemical characteristics tailored for antibody description were derived from homology modeling and regressed to the pH retention. A rigorous feature selection was conducted, reducing the initial descriptor count from 1312 to six in course of a two-staged feature selection process. The utilization of GPR as a Bayesian modeling approach proved advantageous due to its strong model performance, heteroscedastic uncertainty estimation, and non-linear feature identification.

Our experimental results demonstrate that the IgG backbone significantly impact chromatographic retention, as indicated by the comparable elution behavior of the IgG4 molecules. However, the main driver of multimodal interaction is presumed to be in the antibody Fv domain, as homologous IgG1 derivatives showed diverging elution trends. The feature dependencies of our QSPR model support these findings and shed light on a complex adsorption mechanism in multimodal chromatography. The proposed mechanism originates from the CDR region but involves the entire antibody structure due to a combination of electrostatic and hydrophobic contributions.

The overall model performance and its mechanistic interpretability allow for its application in an accelerated process development of IgG-like purification based solely on sequence information. Our model can

replace the experimental screening of initial process pH in multimodal chromatography, which is particularly beneficial for the material and time-constrained early process development. Furthermore, our model can serve as an *in silico* screening approach to identify candidates suitable for purification by multimodal chromatography. Lastly, the identified feature dependencies could aid in the development of improved mechanistic chromatography models by considering a molecular level of detail. In this context, multiscale modeling through the correlation of mechanistic isotherm parameters to molecular-level descriptors could be considered as an intermediate step. To enhance current QSPR models and address the structural diversity of engineered biologics, future research should target the development of global descriptors to encompass multimodal surface interactions, as well as protein topology, independent from molecule structure.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Rudger Hess: Conceptualization, Methodology, Software, Validation, Data curation, Formal analysis, Investigation, Visualization, Writing – original draft, Writing – review & editing, Visualization. **Jan Faessler:** Data curation, Investigation. **Doil Yun:** Data curation, Investigation. **David Saleh:** Conceptualization, Writing – review & editing. **Jan-Hendrik Grosch:** Conceptualization, Writing – review & editing, Project administration, Funding acquisition. **Thomas Schwab:** Resources, Supervision, Project administration, Funding acquisition. **Jürgen Hubbuch:** Conceptualization, Supervision, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The authors would like to thank the department of Development Biologicals (Boehringer Ingelheim, Biberach, Germany) and the Biotherapeutics Discovery department Biologics Research CMC (Boehringer Ingelheim, Ridgefield, USA) and especially Michael Marlow, Helen Wu, and James Hu, for provision of therapeutic antibodies. Furthermore, we would like to thank the BI modeling team for their constant support, as well as the long and fruitful discussions.

Appendix

Fig. A1

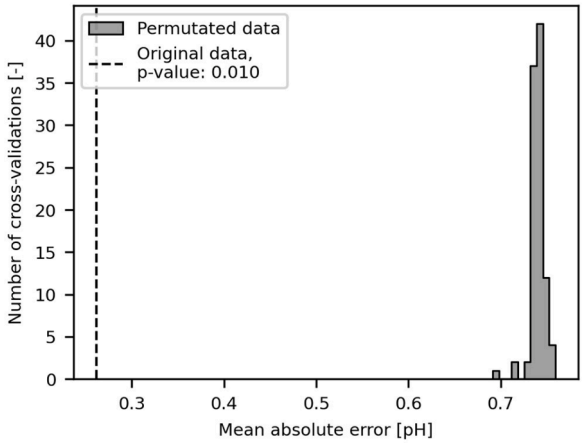

Fig. A1. Significance of y-scrambled cross-validation for pH elution prediction. The cross-validated error of the original data set is plotted against the cross-validated error during 100 permutations of the target variable. The p-value represents the probability of obtaining the original cross-validation score by chance, serving as an indicator of the true dependency between the target variable and the model features.

References

[1] A.A. Shukla, B. Hubbard, T. Tressel, S. Guhan, D. Low, Downstream processing of monoclonal antibodies—application of platform approaches, J. Chromatogr. B 848 (2007) 28–39, https://doi.org/10.1016/j.jchromb.2006.09.026.

[2] A. Voitl, T. Müller-Späth, M. Morbidelli, Application of mixed mode resins for the purification of antibodies, J. Chromatogr. A 1217 (2010) 5753–5760, https://doi.org/10.1016/j.chroma.2010.06.047.

[3] K. Kallberg, H. Johansson, L. Bulow, Multimodal chromatography: an efficient tool in downstream processing of proteins, Biotechnol. J. 7 (2012) 1485–1495, https://doi.org/10.1002/biot.201200074.

[4] M.A. Holstein, A.A.M. Nikfetrat, M. Gage, A.G. Hirsh, S.M. Cramer, Improving selectivity in multimodal chromatography using controlled pH gradient elution, J. Chromatogr. A 1233 (2012) 152–155, https://doi.org/10.1016/j.chroma.2012.01.074.

[5] I.F. Pinto, M.R. Aires-Barros, A.M. Azevedo, Multimodal chromatography: debottlenecking the downstream processing of monoclonal antibodies, Pharm. Bioprocess. 3 (2015) 263–279, https://doi.org/10.4155/pbp.15.7.

[6] K. Zhang, X. Liu, Mixed-mode chromatography in pharmaceutical and biopharmaceutical applications, J. Pharm. Biomed. Anal. 128 (2016) 73–88, https://doi.org/10.1016/j.jpba.2016.05.007.

[7] G. Zhao, X.Y. Dong, Y. Sun, Ligands for mixed-mode protein chromatography: principles, characteristics and design, J. Biotechnol. 144 (2009) 3–11, https://doi.org/10.1016/j.jbiotec.2009.04.009.

[8] S.M. Cramer, M.A. Holstein, Downstream bioprocessing: recent advances and future promise, Curr. Opin. Chem. Eng. 1 (2011) 27–37, https://doi.org/10.1016/j.coche.2011.08.008.

[9] Cytiva, Multimodal Chromatography, (2021). https://www.cytivalifesciences.com/en/us/support/handbooks (accessed April 27, 2022).

[10] D. Gao, L.L. Wang, D.Q. Lin, S.J. Yao, Evaluating antibody monomer separation from associated aggregates using mixed-mode chromatography, J. Chromatogr. A 1294 (2013) 70–75, https://doi.org/10.1016/j.chroma.2013.04.018.
[11] L.S. Wolfe, C.P. Barringer, S.S. Mostafa, A.A. Shukla, Multimodal chromatography: characterization of protein binding and selectivity enhancement through mobile phase modulators, J. Chromatogr. A 1340 (2014) 151–156, https://doi.org/10.1016/j.chroma.2014.02.086.
[12] B.K. Nfor, M. Noverraz, S. Chilamkurthi, P.D.E.M. Verhaert, L.A.M. van der Wielen, M. Ottens, High-throughput isotherm determination and thermodynamic modeling of protein adsorption on mixed mode adsorbents, J. Chromatogr. A 1217 (2010) 6829–6850, https://doi.org/10.1016/j.chroma.2010.07.069.
[13] Y.F. Lee, H. Graalfs, C. Frech, Thermodynamic modeling of protein retention in mixed-mode chromatography: an extended model for isocratic and dual gradient elution chromatography, J. Chromatogr. A 1464 (2016) 87–101, https://doi.org/10.1016/j.chroma.2016.08.026.
[14] R.B. Gudhka, D.J. Roush, S.M. Cramer, A thermodynamic evaluation of antibody-surface interactions in multimodal cation exchange chromatography, J. Chromatogr. A 1628 (2020), 461479, https://doi.org/10.1016/j.chroma.2020.461479.
[15] W.K. Chung, Y. Hou, M. Holstein, A. Freed, G.I. Makhatadze, S.M. Cramer, Investigation of protein binding affinity in multimodal chromatographic systems using a homologous protein library, J. Chromatogr. A 1217 (2010) 191–198, https://doi.org/10.1016/j.chroma.2009.08.005.
[16] H.S. Karkov, B.O. Krogh, J. Woo, S. Parimal, H. Ahmadian, S.M. Cramer, Investigation of protein selectivity in multimodal chromatography using in silico designed Fab fragment variants, Biotechnol. Bioeng. 112 (2015) 2305–2315, https://doi.org/10.1002/bit.25642.
[17] J. Robinson, D. Roush, S. Cramer, Domain contributions to antibody retention in multimodal chromatography systems, J. Chromatogr. A 1563 (2018) 89–98, https://doi.org/10.1016/j.chroma.2018.05.058.
[18] S.S. Parasnavis, B. Niu, M. Aspelund, W.K. Chung, M. Snyder, S.M. Cramer, Systematic workflow for studying domain contributions of bispecific antibodies to selectivity in multimodal chromatography, Biotechnol. Bioeng. 119 (2022) 211–225, https://doi.org/10.1002/bit.27967.
[19] K. Srinivasan, S. Banerjee, S. Parimal, L. Sejergaard, R. Berkovich, B. Barquera, S. Garde, S.M. Cramer, Single molecule force spectroscopy and molecular dynamics simulations as a combined platform for probing protein face-specific binding, Langmuir 33 (2017) 10851–10860, https://doi.org/10.1021/acs.langmuir.7b03011.
[20] R.B. Gudhka, C.L. Bilodeau, S.A. McCallum, M.A. McCoy, D.J. Roush, M.A. Snyder, S.M. Cramer, Identification of preferred multimodal ligand-binding regions on IgG1 FC using nuclear magnetic resonance and molecular dynamics simulations, Biotechnol. Bioeng. 118 (2021) 809–822, https://doi.org/10.1002/bit.27611.
[21] E. O'Connor, M. Aspelund, F. Bartnik, M. Berge, K. Coughlin, M. Kambarami, D. Spencer, H. Yan, W. Wang, Monoclonal antibody fragment removal mediated by mixed mode resins, J. Chromatogr. A 1499 (2017) 65–77, https://doi.org/10.1016/j.chroma.2017.03.063.
[22] K. Dhingra, R.B. Gudhka, S.M. Cramer, Evaluation of preferred binding regions on ubiquitin and IgG1-FC for interacting with multimodal cation exchange resins using DEPC labeling/mass spectrometry, Biotechnol. Bioeng. (2023), https://doi.org/10.1002/bit.28361.
[23] H.F. Tong, C. Cavallotti, S.J. Yao, D.Q. Lin, Molecular insight into protein binding orientations and interaction modes on hydrophobic charge-induction resin, J. Chromatogr. A 1512 (2017) 34–42, https://doi.org/10.1016/j.chroma.2017.06.071.
[24] J. Robinson, D. Roush, S.M. Cramer, The effect of pH on antibody retention in multimodal cation exchange chromatographic systems, J. Chromatogr. A 1617 (2020), 460838, https://doi.org/10.1016/j.chroma.2019.460838.
[25] C.L. Bilodeau, E.Y. Lau, D.J. Roush, M.A. Snyder, S.M. Cramer, Behavior of water near multimodal chromatography ligands and its consequences for modulating protein–ligand interactions, J. Phys. Chem. B 125 (2021) 6112–6120, https://doi.org/10.1021/acs.jpcb.1c01549.
[26] R.B. Gudhka, M. Vats, C.L. Bilodeau, S.A. McCallum, M.A. McCoy, D.J. Roush, M. A. Snyder, S.M. Cramer, Probing IgG1 FC–multimodal nanoparticle interactions: a combined nuclear magnetic resonance and molecular dynamics simulations approach, Langmuir 37 (2021) 12188–12203, https://doi.org/10.1021/acs.langmuir.1c02114.
[27] Y. Hou, S.M. Cramer, Evaluation of selectivity in multimodal anion exchange systems: a priori prediction of protein retention and examination of mobile phase modifier effects, J. Chromatogr. A 1218 (2011) 7813–7820, https://doi.org/10.1016/j.chroma.2011.08.080.
[28] J.F. Buyel, J.A. Woo, S.M. Cramer, R. Fischer, The use of quantitative structure–activity relationship models to develop optimized processes for the removal of tobacco host cell proteins during biopharmaceutical production, J. Chromatogr. A 1322 (2013) 18–28, https://doi.org/10.1016/j.chroma.2013.10.076.
[29] J.A. Woo, H. Chen, M.A. Snyder, Y. Chai, R.G. Frost, S.M. Cramer, Defining the property space for chromatographic ligands from a homologous series of mixed-mode ligands, J. Chromatogr. A 1407 (2015) 58–68, https://doi.org/10.1016/j.chroma.2015.06.017.
[30] J. Woo, S. Parimal, M.R. Brown, R. Heden, S.M. Cramer, The effect of geometrical presentation of multimodal cation-exchange ligands on selective recognition of hydrophobic regions on protein surfaces, J. Chromatogr. A 1412 (2015) 33–42, https://doi.org/10.1016/j.chroma.2015.07.072.
[31] J.R. Robinson, H.S. Karkov, J.A. Woo, B.O. Krogh, S.M. Cramer, QSAR models for prediction of chromatographic behavior of homologous Fab variants, Biotechnol. Bioeng. 114 (2017) 1231–1240, https://doi.org/10.1002/bit.26236.
[32] J. Robinson, M.A. Snyder, C. Belisle, J. Liao, H. Chen, X. He, Y. Xu, S.M. Cramer, Investigating the impact of aromatic ring substitutions on selectivity for a multimodal anion exchange prototype library, J. Chromatogr. A 1569 (2018) 101–109, https://doi.org/10.1016/j.chroma.2018.07.049.
[33] L.E. Crowell, C. Goodwine, C.S. Holt, L. Rocha, C. Vega, S.A. Rodriguez, N. C. Dalvie, M.K. Tracey, M. Puntel, A. Wigdorovitz, V. Parreño, K.R. Love, S. M. Cramer, J.C. Love, Development of a platform process for the production and purification of single-domain antibodies, Biotechnol. Bioeng. 118 (2021) 3348–3358, https://doi.org/10.1002/bit.27724.
[34] J. Pezzini, G. Joucla, R. Gantier, M. Toueille, A.M. Lomenech, C.L. Sénéchal, B. Garbay, X. Santarelli, C. Cabanne, Antibody capture by mixed-mode chromatography: a comprehensive study from determination of optimal purification conditions to identification of contaminating host cell proteins, J. Chromatogr. A 1218 (2011) 8197–8208, https://doi.org/10.1016/j.chroma.2011.09.036.
[35] S.M. Timmick, N. Vecchiarello, C. Goodwine, L.E. Crowell, K.R. Love, J.C. Love, S. M. Cramer, An impurity characterization based approach for the rapid development of integrated downstream purification processes, Biotechnol. Bioeng. 115 (2018) 2048–2060, https://doi.org/10.1002/bit.26718.
[36] A.T. Hanke, M. Ottens, Purifying biopharmaceuticals: knowledge-based chromatographic process development, Trends Biotechnol. 32 (2014) 210–220, https://doi.org/10.1016/j.tibtech.2014.02.001.
[37] D.K. Babi, J. Griesbach, S. Hunt, F. Insaidoo, D. Roush, R. Todd, A. Staby, J. Welsh, F. Wittkopp, Opportunities and challenges for model utilization in the biopharmaceutical industry: current versus future state, Curr. Opin. Chem. Eng. 36 (2022), 100813, https://doi.org/10.1016/j.coche.2022.100813.
[38] M. Bailly, C. Mieczkowski, V. Juan, E. Metwally, D. Tomazela, J. Baker, M. Uchida, E. Kofman, F. Raoufi, S. Motlagh, Y. Yu, J. Park, S. Raghava, J. Welsh, M. Rauscher, G. Raghunathan, M. Hsieh, Y.L. Chen, H.T. Nguyen, N. Nguyen, D. Cipriano, L. Fayadat-Dilman, Predicting antibody developability profiles through early stage discovery screening, MAbs 12 (2020), 1743053, https://doi.org/10.1080/19420862.2020.1743053.
[39] J.T. Heads, S. Kelm, K. Tyson, A.D.G. Lawson, A computational method for predicting the aggregation propensity of IgG1 and IgG4(P) mAbs in common storage buffers, MAbs 14 (2022), 2138092, https://doi.org/10.1080/19420862.2022.2138092.
[40] R. Hess, D. Yun, D. Saleh, T. Briskot, J.H. Grosch, G. Wang, T. Schwab, J. Hubbuch, Standardized method for mechanistic modeling of multimodal anion exchange chromatography in flow through operation, J. Chromatogr. A (2023), 463789, https://doi.org/10.1016/j.chroma.2023.463789.
[41] F. Kröner, J. Hubbuch, Systematic generation of buffer systems for pH gradient ion exchange chromatography and their application, J. Chromatogr. A 1285 (2013) 78–87, https://doi.org/10.1016/j.chroma.2013.02.017.
[42] R. Wälchli, M. Ressurreição, S. Vogg, F. Feidl, J. Angelo, X. Xu, S. Ghose, Z.J. Li, X. L. Saoût, J. Souquet, H. Broly, M. Morbidelli, Understanding mAb aggregation during low pH viral inactivation and subsequent neutralization, Biotechnol. Bioeng. 117 (2020) 687–700, https://doi.org/10.1002/bit.27237.
[43] D. Saleh, R. Hess, M. Ahlers-Hesse, F. Rischawy, G. Wang, J. Grosch, T. Schwab, S. Kluters, J. Studts, J. Hubbuch, A multiscale modeling method for therapeutic antibodies in ion exchange chromatography, Biotechnol. Bioeng. 120 (2023) 125–138, https://doi.org/10.1002/bit.28258.
[44] K. Zhu, T. Day, D. Warshaviak, C. Murrett, R. Friesner, D. Pearlman, Antibody structure determination using a combination of homology modeling, energy-based refinement, and loop prediction, Protein Struct. Funct. Bioinform. 82 (2014) 1646–1655, https://doi.org/10.1002/prot.24551.
[45] K.R. Abhinandan, A.C.R. Martin, Analysis and improvements to Kabat and structurally correct numbering of antibody variable domains, Mol. Immunol. 45 (2008) 3832–3839, https://doi.org/10.1016/j.molimm.2008.05.022.
[46] E.O. Saphire, P.W.H.I. Parren, R. Pantophlet, M.B. Zwick, G.M. Morris, P.M. Rudd, R.A. Dwek, R.L. Stanfield, D.R. Burton, I.A. Wilson, Crystal structure of a neutralizing human IgG Against HIV-1: a template for vaccine design, Science 293 (2001) 1155–1159, https://doi.org/10.1126/science.1061692.
[47] G. Scapin, X. Yang, W.W. Prosise, M. McCoy, P. Reichert, J.M. Johnston, R.S. Kashi, C. Strickland, Structure of full-length human anti-PD1 therapeutic IgG4 antibody pembrolizumab, Nat. Struct. Mol. Biol. 22 (2015) 953–958, https://doi.org/10.1038/nsmb.3129.
[48] K. Zhu, T. Day, Ab initio structure prediction of the antibody hypervariable H3 loop, Protein Struct. Funct. Bioinform. 81 (2013) 1081–1089, https://doi.org/10.1002/prot.24240.
[49] G.M. Sastry, M. Adzhigirey, T. Day, R. Annabhimoju, W. Sherman, Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments, J. Comput. Aid Mol. Des. 27 (2013) 221–234, https://doi.org/10.1007/s10822-013-9644-8.
[50] M.H.M. Olsson, C.R. Søndergaard, M. Rostkowski, J.H. Jensen, PROPKA3: consistent treatment of internal and surface residues in empirical pKa predictions, J. Chem. Theory Comput. 7 (2011) 525–537, https://doi.org/10.1021/ct100578z.
[51] C. Lu, C. Wu, D. Ghoreishi, W. Chen, L. Wang, W. Damm, G.A. Ross, M.K. Dahlgren, E. Russell, C.D.V. Bargen, R. Abel, R.A. Friesner, E.D. Harder, OPLS4: improving force field accuracy on challenging regimes of chemical space, J. Chem. Theory Comput. 17 (2021) 4291–4300, https://doi.org/10.1021/acs.jctc.1c00302.
[52] K. Sankar, K. Trainor, L.L. Blazer, J.J. Adams, S.S. Sidhu, T. Day, E. Meiering, J.K. X. Maier, A descriptor set for quantitative structure-property relationship

prediction in biologics, Mol. Inform. (2022), 2100240, https://doi.org/10.1002/minf.202100240.
[53] S.A. Wildman, G.M. Crippen, Prediction of physicochemical parameters by atomic contributions, J. Chem. Inf. Comp. Sci. 39 (1999) 868–873, https://doi.org/10.1021/ci990307l.
[54] K. Sankar, S.R. Krystek, S.M. Carl, T. Day, J.K.X. Maier, AggScore: prediction of aggregation-prone regions in proteins based on the distribution of surface patches, Proteins Struct. Funct. Bioinform. 86 (2018) 1147–1156, https://doi.org/10.1002/prot.25594.
[55] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, G. Louppe, P. Prettenhofer, R. Weiss, R.J. Weiss, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[56] O. Obrezanova, G. Csányi, J.M.R. Gola, M.D. Segall, Gaussian processes: a method for automatic QSAR modeling of ADME properties, J. Chem. Inf. Model. 47 (2007) 1847–1857, https://doi.org/10.1021/ci7000633.
[57] C.E. Rasmussen, C.K.I. Williams, Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning), The MIT Press, 2005.
[58] P. Zhou, F. Tian, X. Chen, Z. Shang, Modeling and prediction of binding affinities between the human amphiphysin SH3 domain and its peptide ligands using genetic algorithm-Gaussian processes, Peptide Sci. 90 (2008) 792–802, https://doi.org/10.1002/bip.21091.
[59] C. Zhu, R.H. Byrd, P. Lu, J. Nocedal, Algorithm 778: ʟ-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization, ACM Trans. Math. Softw. 23 (1997) 550–560, https://doi.org/10.1145/279232.279236. Toms.
[60] V. Kumar, Feature selection: a literature review, Smart Comput. Rev. 4 (2014) 211–229, https://doi.org/10.6029/smartcr.2014.03.007.
[61] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32, https://doi.org/10.1023/a:1010933404324.
[62] A. Tropsha, Best practices for QSAR model development, validation, and exploitation, Mol. Inform. 29 (2010) 476–488, https://doi.org/10.1002/minf.201000061.
[63] M. Ojala, G.C. Garriga, Permutation tests for studying classifier performance, in: Proceedings of the Ninth IEEE International Conference on Data Mining, 2009, pp. 908–913, https://doi.org/10.1109/icdm.2009.108.
[64] T. Hastie, The Elements of Statistical learning : Data mining, inference, and Prediction, 2nd ed., N.Y. Springer, New York, 2009.
[65] A. Goldstein, A. Kapelner, J. Bleich, E. Pitkin, Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation, J. Comput. Graph. Stat. 24 (2015) 44–65, https://doi.org/10.1080/10618600.2014.907095.
[66] J.T. Heads, R. Lamb, S. Kelm, R. Adams, P. Elliott, K. Tyson, S. Topia, S. West, R. Nan, A. Turner, A.D.G. Lawson, Electrostatic interactions modulate the differential aggregation propensities of IgG1 and IgG4P antibodies and inform charged residue substitutions for improved developability, Protein Eng. Des. Sel. 32 (2019) 277–288, https://doi.org/10.1093/protein/gzz046.
[67] S. Ghose, B. Hubbard, S.M. Cramer, Protein interactions in hydrophobic charge induction chromatography (HCIC), Biotechnol. Progr. 21 (2005) 498–508, https://doi.org/10.1021/bp049712+.
[68] P. Baumann, K. Baumgartner, J. Hubbuch, Influence of binding pH and protein solubility on the dynamic binding capacity in hydrophobic interaction chromatography, J. Chromatogr. A 1396 (2015) 77–85, https://doi.org/10.1016/j.chroma.2015.04.001.
[69] S. Koley, S.H. Altern, M. Vats, X. Han, D. Jang, M.A. Snyder, C. Belisle, S. M. Cramer, Evaluation of guanidine-based multimodal anion exchangers for protein selectivity and orthogonality, J. Chromatogr. A 1653 (2021), 462398, https://doi.org/10.1016/j.chroma.2021.462398.
[70] L.K. Kimerer, T.M. Pabst, A.K. Hunter, G. Carta, Role of configurational flexibility on the adsorption kinetics of bivalent bispecific antibodies on porous cation exchange resins, J. Chromatogr. A 1655 (2021), 462479, https://doi.org/10.1016/j.chroma.2021.462479.
[71] C. Molnar. Interpretable Machine Learning. A Guide for Making Black Box Models Explainable, 2019. https://christophm.github.io/interpretable-ml-book.