ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/9032600

Identification of Chemically Selective Displacers Using Parallel Batch Screening Experiments and Quantitative Structure Efficacy Relationship Models

ARTICLE *in* ANALYTICAL CHEMISTRY · DECEMBER 2003

Impact Factor: 5.64 · DOI: 10.1021/ac0341564 · Source: PubMed

CITATIONS
24

READS
21

4 AUTHORS, INCLUDING:

Curt M Breneman
Rensselaer Polytechnic Institute
103 PUBLICATIONS **4,454** CITATIONS
SEE PROFILE

Steven M Cramer
Rensselaer Polytechnic Institute
202 PUBLICATIONS **4,393** CITATIONS
SEE PROFILE

Available from: Curt M Breneman
Retrieved on: 31 January 2016

Anal. Chem. **2003,** *75,* 5806−5816

Identification of Chemically Selective Displacers Using Parallel Batch Screening Experiments and Quantitative Structure Efficacy Relationship Models

Nihal Tugcu,[†,‡] Asif Ladiwala,[†] Curt M. Breneman,[§] and Steven M. Cramer*[,†]

Department of Chemical and Biological Engineering and Department of Chemistry, Rensselaer Polytechnic Institute, Troy, New York 12180

Parallel batch screening experiments were carried out to examine how displacer chemistry and salt counterions affect the selectivity of batch protein displacements in anion exchange chromatographic systems. The results indicate that both salt type and displacer chemistry can have a significant impact on the amount of protein displaced. Importantly, the results indicate that, by changing the displacer, salt counterion, or both, one can induce significant selectivity changes in the relative displacement of two model proteins. This indicates that highly selective separations can be developed in ion exchange systems by the appropriate selection of displacer chemistry and salt counterion. The experimental batch screening data were also used in conjunction with various molecular descriptors to generate quantitative structure efficacy relationship (QSER) models based on a support vector machine feature selection and regression tool. The models resulted in good correlations and successful predictions for an external test set of displacers. A star plot approach was shown to be a powerful tool to aid in the interpretation of the QSER models. These results indicate that this modeling approach can be employed for the a priori prediction of displacer efficacy as well as for providing insight into displacer design and the selection of proper mobile-phase conditions for highly selective separations.

Selectivity in ion exchange systems can be exploited in various ways. A number of studies have addressed the effect of eluting salt type and concentration on protein selectivity for ion exchange systems. The results indicated that both co-ion and counterion had an effect on the protein selectivity.[1−7] Kopaciewicz and coworkers[1] demonstrated that while the cation slightly altered the selectivity, the anion could have a significant effect on the retention time as well as the selectivity in anion exchange systems. In addition, changes in the eluting salt type and gradient mode have been shown to significantly enhance the selectivity of closely related variants.[5]

It has been shown that retention in ion exchange systems is not purely based on electrostatic interactions.[8−10] Further, selectivity changes can occur for proteins due to changes in the stationary-phase backbone chemistry.[11] Homologous sets of molecules have been used to identify the effect of backbone chemistry on the selectivity of displacers on cation exchange stationary phases. Importantly, this work demonstrated that hydrophobic/aromatic interactions had an important effect on the affinity of displacers in cation exchange systems.[12−15] In addition, structural factors affecting displacer efficacy have been studied for cation and anion exchange systems. In that work, various displacers were screened for their efficacy using batch displacements and the results were evaluated using quantitative structure efficacy relationship (QSER) models.[16,17]

Recently, an effective technique for the parallel batch screening of potential low molecular weight displacers has been developed.[16] In contrast to conventional column techniques, this batch screening technique enables the rapid evaluation of potential low molecular weight displacers using parallel displacement experiments. The objective of this study is to explore selectivity changes

* Corresponding author. Tel.: (518) 276-6198. Fax: (518) 276-4030. E-mail: crames@rpi.edu.

† Department of Chemical and Biological Engineering.

‡ Present address: Research Laboratories, RY805S-100, Merck & Co., Inc., 126 E. Lincoln Ave., P. O. Box 2000, Rahway, NJ 07065.

§ Department of Chemistry.

(1) Kopaciewicz, W.; Rounds, M. A.;, Fausnaugh, J.; Regnier F. E. *J. Chromatogr.* **1983**, *266*, 3-21.
(2) Kopaciewicz, W.; Regnier, F. E. *Anal. Chem.* **1983**, *133*, 251−259.
(3) Barron, R. E.; Fritz, J. S. *J. Chromatogr.* **1984**, *284*, 13−25.
(4) Hodder, A. N.; Aguilar, M. I.; Hearn, M. T. W. *J. Chromatogr.* **1989**, *476*, 391−411.
(5) Hodder, A. N.; Aguilar, M. I.; Hearn, M. T. W. *J. Chromatogr.* **1990**, *506*, 17−34.
(6) Rounds, M. A.; Regnier, F. E. *J. Chromatogr.* **1984**, *283*, 37−45.
(7) Malmquist, G.; Lundell, N. *J. Chromatogr.* **1992**, *627*, 107−124.
(8) Stahlberg, J.; Jonsson, B.; Horvath, C. *Anal. Chem.* **1992**, *64*, 3118.
(9) Roth, C. M.; Lenhoff, A. M. *Langmuir* **1993**, *9*, 962.
(10) Roth, C. M.; Unger, K. K.; Lenhoff, A. M. *J. Chromatogr., A* **1996**, *726*, 45.
(11) Mazza, C. B.; Sukumar, N.; Breneman, C. M.; Cramer, S. M. *Anal. Chem.* **2001**, *73*, 5457−5461.
(12) Shukla, A. A.; Barnthouse, K. A.; Bae, S. S.; Moore, J. A.; Cramer, S. M. *J. Chromatogr., A* **1998**, *814*, 83.
(13) Shukla, A. A.; Barnthouse, K. A.; Bae, S. S.; Moore, J. A.; Cramer, S. M. *Ind. Eng. Chem. Res.* **1998**, *37*, 4090.
(14) Shukla, A. A.; Barnthouse, K. A.; Bae, S. S.; Moore, J. A.; Cramer, S. M. *J. Chromatogr., A* **1998**, *827*, 295.
(15) Tugcu, N.; Bae, S. S.; Moore, J. A.; Cramer, S. M. *J. Chromatogr., A* **2002**, *954* (1−2), 127−135.
(16) Mazza, C. B.; Rege, K.; Breneman, C. M.; Dordick, J. S.; Cramer, S. M. *Biotechnol. Bioeng.* **2002**, *80* (1), 60−72.

10.1021/ac0341564 CCC: $25.00 © 2003 American Chemical Society
Published on Web 09/18/2003

that can occur by changing the displacer chemistry, the eluting salt type in displacement chromatography, or both.

In the present work, two similarly retained proteins were used as a model system for screening several displacers using a parallel batch screening technique in the presence of different salts. The data from the batch screening experiments were then used to generate QSER models using support vector machine (SVM) regression analysis.[18] 2D, 3D Molecular Operating Environment (MOE), molecular fragment (FRAG), and transferable atomic equivalent (TAE) descriptors[19,20] were calculated from the energy-minimized structures of the displacers. A *training set* of displacers was selected to derive a predictive QSER model (i.e., learning from the database) using SVM regression and bootstrapping techniques.[18] Using this methodology, the original training set of displacers was further subdivided into a *validation set*, with the remaining displacers used for training. This procedure was repeated 20 times with different training and validation subsets, resulting in the construction of 20 distinct, but similar models. This technique ensures better model generality and is termed "bootstrapping" in the QSPR literature. The predictive capability of the models was initially determined by their performance on the validation sets, but their actual predictive power is only revealed when predictions are made using the "true unknowns" i.e., the displacers held back as the *test set*. When the predictions made for each test molecule by all 20 models are combined, the result is a bootstrap AGGregatED (BAGGED) model. It was shown that the resulting BAGGED QSER models were able to successfully predict the efficacy of the test set displacers in the database. Examination of the descriptor weights within the resulting models enables the importance of various structural and electronic features of the displacers to be elucidated.

EXPERIMENTAL PROTOCOL

Materials. Bulk strong anion exchange (quaternary ammonium) Source 15Q (15 μm) stationary-phase material was donated by Amersham Biosciences (Uppsala, Sweden). The stationary phase was slurry packed into 50 × 5 mm i.d. columns. TSK-Gel G3000SWXL size exclusion column (300 × 7.8 mm i.d.) and a TSK-Gel SWXL (40 × 6 mm i.d.) guard column were gifts from Tosoh Biosep (Montgomeryville, PA). Amyloglucosidase and apoferritin were purchased from Sigma (St. Louis, MO) and ICN Biomedicals, Inc. (Aurora, OH), respectively. Sodium chloride, sodium bromide, and sodium sulfate were purchased from Fisher Scientific (Pittsburgh, PA). Tris-HCl and Tris base were purchased from Sigma. Benzoic acid, cyclamic acid, 1,2-naphthaquinone-4-sulfonic acid, saccharin (sodium salt, hydrate), pantothenic acid, 1,3,6-naphthalenetrisulfonic acid, sunset yellow FCF, tartrazine, orange G, new coccine, allura red AC, fast green FCF, amaranth, orange I, benzoic acid, 1,5-naphthalenedisulfonic acid, *p*-toluenesulfonic acid, benzenesulfonic acid, 1,2-benzenedisulfonic acid, 2-(4-sulfophenylazo)-1,8-dihydroxy-3,6-naphthalenedisulfonic acid, chromotrope 2*R*, 8-hydroxy-1,3,6-pyrenetrisulfonic acid, caffeine, and calmagite were purchased from Aldrich (Milwaukee, WI). Suramine (hexasodium salt) was obtained from Alexis Corp. (San Diego, CA), and sucrose octasulfate was purchased from Toronto Research Chemicals, Inc. (Ontario, Canada). TA-PSNa_3, IC-PSNa_3, PG(EO1)-PSNa_3, PG-PSNa_3, TA-PhSO_3Na, and PG(EO1)-PhSO_3Na were synthesized in the Department of Chemistry at Rensselaer Polytechnic Institute.[21]

Apparatus. Protein and displacer analysis was carried out using a Waters 600 multisolvent delivery system, a model 712 WISP autoinjector, and a model 996 photodiode array absorbance detector controlled by a Millenium chromatography manager (Waters, Milford, MA).

Procedures. Measurement of Protein Retention. Linear gradient experiments were carried out with a constant slope between buffer A (20 mM Tris, pH 7.5) and buffer B (20 mM Tris + 580 mM NaCl, NaBr, or Na_2SO_4). The linear gradient slope for these experiments was 14.5 mM salt per column volume. Aliquots (20 μL) of 4 mg/mL protein solutions were injected, and the experiments were carried out in duplicate at a flow of 0.5 mL/min. The absorbance was monitored at 280 nm.

Parallel Batch Screening Experiments. Parallel batch screening experiments were carried out separately for each protein. Displacer screening was performed in the presence of either NaCl or Na_2SO_4. The bulk stationary phase (Source 15Q; 2.5 mL) was first washed with deionized water and then the carrier buffer—20 mM Tris + 30 mM NaCl or Na_2SO_4—was added and allowed to equilibrate for 2 h. After gravity settling of the stationary phase, the supernatant was removed and 30 mL of 1.5 mg/mL amyloglucosidase or apoferritin in the carrier buffer was added and then equilibrated in a shaker for 6 h at 23 °C. The supernatant was analyzed by size exclusion chromatography to determine the protein concentration, and the amount adsorbed on the stationary phase was calculated through a mass balance. The supernatant was then removed, and 25-μL aliquots of the stationary phase with adsorbed protein were added to separate vials. Aliquots (300 μL) of 10 mM solutions of each displacer in the carrier buffer were then added to each vial and allowed to equilibrate for 6 h. After equilibration, the stationary phase was allowed to gravity settle, and the supernatants were removed and analyzed to determine the percentage of protein displaced by each displacer. These experiments were carried out in duplicate.

HPLC Analysis of Proteins and Displacers in the Supernatant. Size exclusion chromatography was employed to determine the protein concentrations in all supernatants. In these experiments, the mobile phase was 50 mM phosphate + 100 mM NaCl at pH 6.0. Aliquots (5 μL) of each supernatant solution obtained from the batch screening experiments were injected, and the experiments were carried out in duplicate at a flow rate of 1.0 mL/min. For these experiments, the absorbance was monitored at 280 nm.

QSER Modeling and SVM Regression Models. To construct informative QSER models, electron density-based TAE quantum mechanics descriptors, MOE descriptors and FRAG descriptors were employed. Using this hybrid set of descriptors, a SVM sparse regression algorithm was applied in a feature selection mode to determine a subset of relevant molecular property descriptors for each of the training sets involved in the bootstrapping procedure.

(17) Tugcu, N.; Mazza, C. B.; Breneman, C. M.; Sanghvi, Y.; Cramer, S. M. *Sep. Sci. Technol.* **2002**, *37* (7), 1667−1681.

(18) Vapnik, V. N. *The Nature of Statistical Learning Theory*; Springer: New York, 1995.

(19) Breneman, C. M.; Rhem, M. *J. Comput. Chem.* **1997**, *18*, 182−197.

(20) RECON 5.2, program locally developed by Breneman, C. M. and Sukumar, N., RPI, Troy, NY, 2000.

(21) Tugcu, N.; Park, S. K.; Moore, J. A.; Cramer, S. M. *Ind. Eng. Chem. Res.* **2002**, *41* (25), 6482−6492.

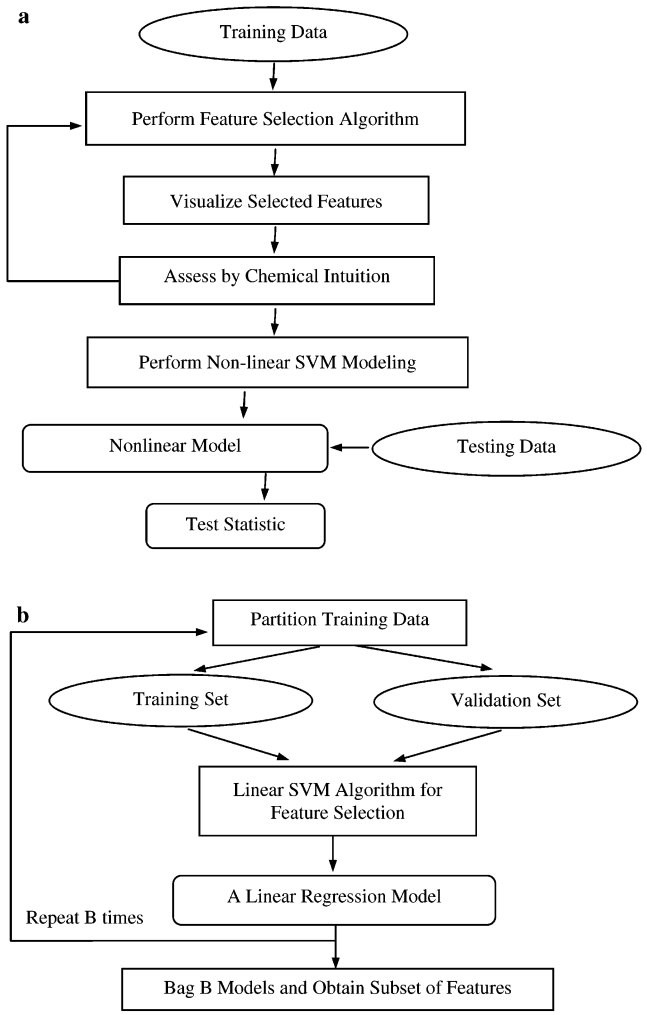

Figure 1. (a) Computational chemical property design and model validation and (b) Feature selection flowchart.

Subsequently, nonlinear SVM models were built based on those relevant descriptors. The overall modeling scheme is shown in Figure 1.

Implementation. The structures of the displacer molecules were obtained from the supplier's website (http://www.sial.com/). These molecules were drawn in SYBYL 6.5 (Tripos Inc., St. Louis, MO) and energy minimized using the MMFF94 force field. Molecular Operating Environment (Chemical Computing Group Inc., Montreal, Canada) software was used to obtain MOE molecular descriptors. The in-house RECON 5.2 package was employed for generating the TAE descriptors for the displacers. An in-house SVM program, developed independently in the Department of Mathematics at Rensselaer Polytechnic Institute, was used in the analysis.[22]

TAE Descriptor Generation. The transferable atomic equivalent/reconstruction (TAE/RECON)[19] method consists of a rapid charge density reconstruction algorithm that utilizes atomic charge density fragments that have been precomputed from ab initio wave functions. In principle, a library of atomic charge density components (TAEs) can be used to construct molecular electron densities in a form that allows for rapid retrieval of the molecular surface properties needed to generate descriptors. For each molecule, the RECON program reads its molecular structure information and then reconstructs the electronic properties of the molecular surface from the atomic fragments. The distributions of several electronic properties on molecular surfaces may then be quantified to give a large variety of numerical QSER descriptors. The CPU and disk resources required for TAE reconstruction are minimal—the electronic property distributions of 31 displacers may be computed in ~30 s on a single-headed 1.7-GHz Linux workstation.

MOE Descriptors. The MOE program provides a combination of several types of traditional molecular property descriptors, including connectivity-based topological 2D descriptors, physicochemical property descriptors, shape-independent 3D molecular features, and some pharmacophoric descriptors. These descriptors were calculated for the displacers via the QuaSAR descriptors module in the MOE package. Prior to MOE descriptor calculation, the displacer structures must be appropriately charged as at the pH of the experiments. For this, the ACD/pK_a DB package (Advanced Chemistry Development Inc., Toronto, ON, Canada) was employed to compute the pK_a's of the charge centers on the displacer molecule. These pK_a values were then used to assign the charges on the displacer molecules at the pH of the experiments.

Support Vector Machine Modeling. The SVM method proposed by Vapnik and co-workers[18] is based on statistical learning theory. This method has proven to be effective for addressing general-purpose classification and regression problems. SVMs have been successfully applied to a wide range of pattern recognition problems, including quality control classifications, "needle in a haystack" classification searches, and robust regression modeling. In most of these cases, the performance of SVM modeling either matches or is significantly better than that of traditional machine learning approaches, including artificial neural networks. The SVM method has a number of interesting properties, including an effective avoidance of overfitting, which improves its ability to build models using large numbers of molecular property descriptors with relatively few experimental results in the training set. Although SVM was originally developed for pattern recognition, it was later extended to solve the regression problem.[18] In the present work, we focus on support vector regression for creating QSER models of displacer efficacy.

To summarize the operation of an SVM modeling procedure, it is important to consider some fundamental principles of SVMs: With a given set of training data, the objective is to find a function $f(x)$, called the ϵ insensitive loss function, that has less than ϵ deviation from the experimental protein retention data for all cases in the training set. In other words, those predicted percent displaced values within the ϵ distance of the actual response are not penalized for being erroneous. Only those prediction points beyond ϵ of the real response values are considered to contain modeling errors and are included in the "loss function". This technique helps to control the complexity of the model and tends to minimize the risk of overfitting. In QSER studies, the magnitude of ϵ will be roughly equivalent to the experimental error in the measurement of percent protein displaced values. In typical QSPR studies, many more descriptors are initially available than the number of molecules in the data set and usually include some

(22) Bennett, K.; Bi, J.; Embrechts, M. J.; Breneman, C. M.; Song, M., in press, *J. Mach. Learn. Res.*.

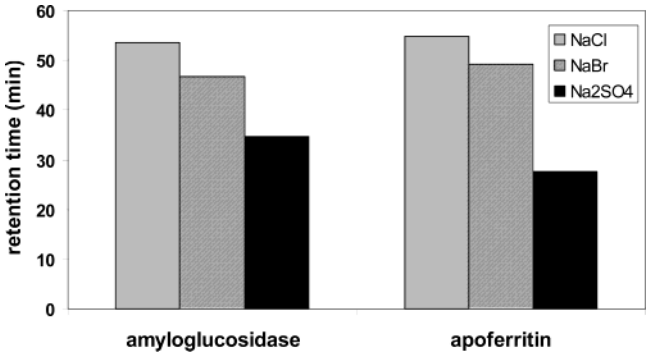

Figure 2. Linear retention of amyloglucosidase and apoferritin in the presence of different salts on Source 15Q material.

redundant or irrelevant variables. To identify the relevant descriptors for a particular problem, variable selection techniques are always employed to choose those informative descriptors and eliminate irrelevant descriptors from consideration. The application of this type of feature selection serves to improve the computational signal-to-noise ratio in the resulting models. In this study, we applied a feature selection approach based on the sparse SVM regression.[22] Within this technique, a series of linear SVMs models (usually 20−40) that exhibit good generalization are constructed. In each linear l_1-norm SVM or bootstrap, the optimal weight vector will have relatively few nonzero weights with the degree of sparsity depending on the SVM model parameters. Those features with nonzero weights then become potential attributes to be used in the nonlinear SVM. The method exploits the fact that linear SVM with l_1-norm regularization inherently performs feature selection as a side effect of minimizing capacity in the SVM model. Finally, the union of all obtained subsets of features produces the final descriptor set that can be used to construct the nonlinear SVM predictive model. To get more robust and general predictive results, multiple QSER models based on the same feature set are built. Thus, instead of using a single model, which is heavily and easily affected by chance correlations, the average of all model predictions is used as our final prediction results. This kind of debiasing technique is referred to as "bagging" in the statistical analysis field.[23]

RESULTS AND DISCUSSION

A relatively large number of anionic proteins were evaluated for their selectivity in the presence of various salt counterions and stationary phase materials.[24] The protein pair amyloglucosidase (AMY) and apoferritin (APO) exhibited significant selectivity changes. As shown in Figure 2, the linear retention data obtained on Source 15Q material indicated that, in the presence of NaCl and NaBr, these two proteins were very closely retained, with a slightly higher retention observed for apoferritin. On the other hand, when the salt was changed to Na_2SO_4, the elution order of these two proteins was reversed and a significant difference in retention was observed.

Since the selectivity of this protein pair is sensitive to the salt counterion, the contributions of various secondary interactions to the retention may be different for these two proteins. To exploit these subtle differences in the displacement mode, a variety of displacers were screened for their efficacy at displacing these proteins using the parallel batch screening technique developed previously in our laboratory.

Parallel batch screening experiments were carried out with amyloglucosidase and apoferritin to evaluate the effect of both salt type and displacer chemistry on the "percent protein displaced". A total of 31 different displacers with various structures, including some anion exchange displacers recently synthesized in our laboratory,[21] were evaluated for each protein in the presence of two different salts, NaCl and Na_2SO_4. The data were then used to generate QSER models.

The percent protein displaced data for each protein by the various displacers in the presence of different salts are shown in Figure 3. Percent amyloglucosidase displaced values varied from ∼3 to ∼68% in NaCl and ∼5 to ∼53% in Na_2SO_4. Percent apoferritin displaced values varied from ∼5 to ∼83% in both NaCl and Na_2SO_4. The actual percent protein displaced values are given in Table A in the Supporting Information.

Effect of Displacing Salt Type on Displacement. As indicated above, the linear retention of amyloglucosidase and apoferritin was very similar in NaCl, with apoferritin exhibiting slightly higher retention. Further, the results indicated that amyloglucosidase had measurably higher retention than apoferritin in the presence of Na_2SO_4 (Figure 2). Based on these data, one would expect that the percent displaced values for amyloglucosidase and apoferritin would be similar in NaCl and that the pecent amyloglucosidase displaced values would be less in Na_2SO_4.

Sections a and b of Figure 3 show the relative percent amyloglucosidase and percent apoferritin displaced values in NaCl and Na_2SO_4, respectively. Figure 3a indicates that for most of the displacers, percent amyloglucosidase displaced values were similar or higher than those for apoferritin in the presence of NaCl. This is in good agreement with the linear retention data. However, out of 31 displacers, 3 molecules displaced significantly more apoferritin than amyloglucosidase, in contrast to the linear retention results. Further, several other molecules exhibited significantly higher efficacy at displacing amyloglucosidase relative to apoferritin. Clearly, the presence of the displacer is having a major impact on the effective selectivity of this system. As seen in Figure 3b, when Na_2SO_4 was used as the salt, the majority of displacers resulted in higher percent apoferritin displaced values, as would be expected from the linear retention data. However, as seen in the figure, the selectivity of this system was again profoundly affected by the presence of the various displacers.

Effect of Displacer Chemistry on Displacement. As indicated by the data, the affinity of the displacers tended to increase with increasing number of aromatic rings and charges on the displacers. For the displacement of amyloglucosidase, the highest values were obtained for FCF green (∼68%) and SOS (∼%53) in the presence of NaCl and Na_2SO_4, respectively. Other high-affinity displacers in NaCl were 8-hydroxypyrene, sunset yellow, tartrazine, 2,4-sulfophenylazo, and chromotrope 2R. In Na_2SO_4, displacers such as suramine and PG(EO1)-$PhSO_3Na$ exhibited high percent amyloglucosidase displaced values. All of these displacers had three or more aromatic rings and two or more negative charges.

(23) Breiman, L. *Machine Learn.* **1996**, *24*, 123−140.
(24) Tugcu, N.; Song, M.; Breneman, C. M.; Cramer, S. M. *Anal. Chem.* **2003**, *75*, 3563−3572.

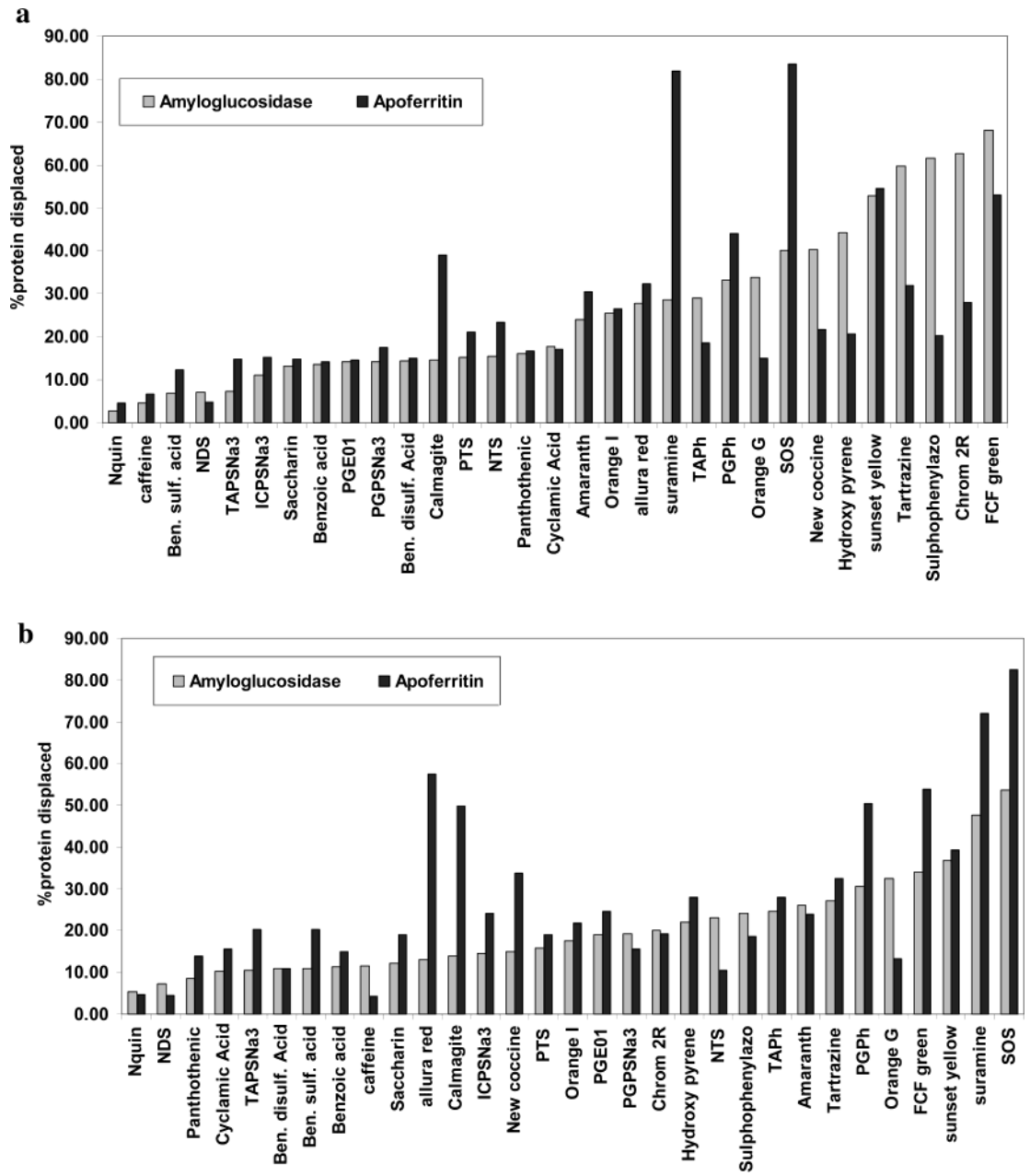

Figure 3. Comparison of percent amyloglucosidase and percent apoferritin displaced values from parallel batch screening experiments in the presence of (a) NaCl and (b) Na_2SO_4.

For the displacement of apoferritin, SOS exhibited the highest percent displaced value (~83%) in the presence of either salt. Other displacers with high percent apoferritin displaced values were suramine and PG(EO1)-$PhSO_3Na$. Again, these results indicate that molecules with more aromatic rings and charges will act as higher affinity displacers in anion exchange systems.

As seen in Figure 3a, the displacers, calmagite, suramine, and SOS resulted in significant increases in the amount of apoferritin displaced relative to amyloglucosidease in NaCl. In addition, *p*-toluenesulfonic acid, naphthalenetrisulfonic acid, amaranth, PG-(EO1)-$PhSO_3Na$, and allura red also resulted in relatively higher percent apoferritin displaced values, albeit to a lesser extent. This result was not anticipated from the protein retention data, where the retention of apoferritin was slightly higher than that of amyloglucosidase. Even though these displacers were all very effective in displacing more apoferritin, a common pattern in the displacer structure could not be ascertained. As shown in Figure 3b, while several displacers exhibited significantly more displacement of apoferritin then amyloglucosidase in the presence of Na_2SO_4, some displacers (e.g., Orange G and naphthalenetrisulfonic acid) resulted in higher displacement of amyloglucosidase. This result was not anticipated from the linear retention data. Again, a common pattern in the displacer structure could not be ascertained. Clearly, these data indicate that both salt type and displacer chemistry have a significant effect on the multicomponent selectivity. To gain more insight into the physicochemical properties that relate to this "selectivity", a QSER modeling approach was employed as described below.

QSER Models. QSER models were generated to predict the displacer efficacy for amyloglucosidase and apoferritin in the presence of different salts, as well as to aid in identifying the structural components that contribute to the efficacy and selectivity

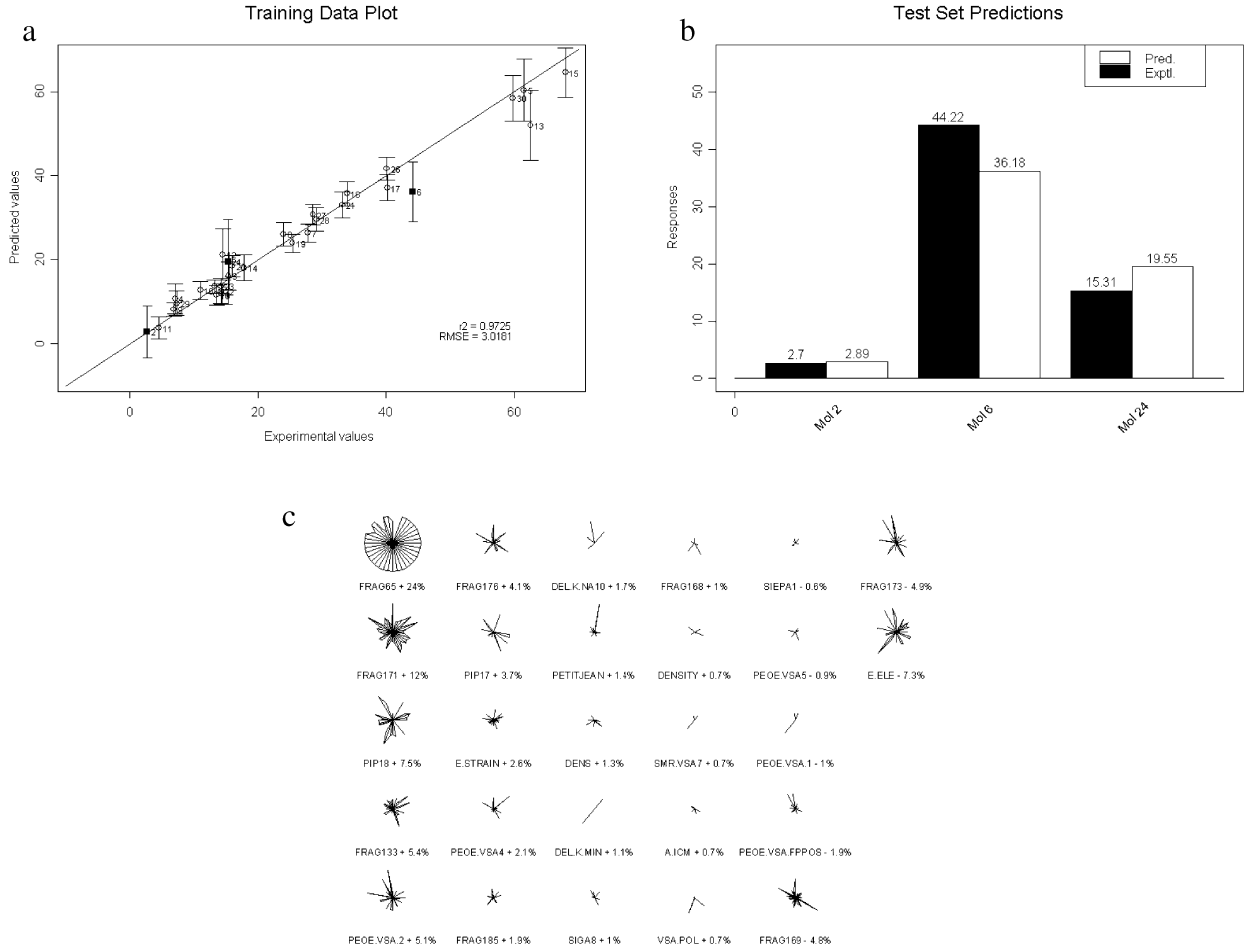

Figure 4. QSER model for percent AMY displaced in the presence of NaCl: (a) predicted values vs observed values for the entire data set, (b) comparison of the predicted and experimental values for the test set, and (c) star plots with weight percentages for the descriptors in the model.

of the displacers in anion exchange systems. The responses employed in these models were the four sets of parallel batch screening data as well as the ratio of the percent protein displaced of the two proteins in the two salts. The motivation behind this was to see whether the QSER models could successfully explain or capture the observed selectivity reversals. As described in the theory section, a wide variety of MOE, FRAG, and TAE descriptors were used to generate these models. The initial data set consisted of 6 experimental responses, 189 MOE descriptors, 208 FRAG descriptors, and 147 TAE descriptors. Descriptors having the same values for all 31 molecules were eliminated from the data set, since these provide no additional information in the model. Furthermore, descriptors showing a variance of greater than 4 times a standard deviation were removed from the data set to reduce outliers and to enable interpretation of the models. The final data set consisted of 6 responses and 233 descriptors and was used to generate six independent QSER models. This data set was subjected to SVM feature selection to give six independent feature sets, one corresponding to each response. Finally, the data set was divided into training and test sets for the purpose of model building, with 10% of the molecules (i.e., 3 molecules) in the test set and the remaining 28 molecules in the training set. 1,2-Naphthoquinone-4-sulfonic acid (Mol 2), 8-hydroxy-1,3,6-pyrenetrisulfonic acid (Mol 6), and *p*-toluenesulfonic acid (Mol 24) were arbitrarily chosen to be in the external test set.

Once developed, the QSER models were tested for their predictive ability for the external test set of displacers. The key descriptors in the final QSER models were then examined to determine the physicochemical phenomena that influence displacer efficacy in the presence of different salts. As a part of the modeling process, graphic visualization plots (star plots) were generated to give a better understanding of relative descriptor importance for the different models.

The relevant QSER descriptors selected in these models include shape, size, surface property, molecular fragment, and electron density derived descriptors. The definitions of the important descriptors are given in Table B (Supporting Information). As seen in the table, many descriptors have direct physical/chemical significance and are relatively easy to interpret. For descriptors that are more difficult to interpret, tools such as correlation matrices/plots and molecular surface visualization were employed to gain insight into the physicochemical information provided by these descriptors in the model.

Figures 4a, 5a, and Ia−IVa (Supporting Information) show the correlation between the experimental and predicted results. The open circles represent the "bagged" predictions for the training set molecules when left out in the validation set during the bootstrapping procedure and the dark squares represent "bagged" predictions for the test set molecules. The error bars in these figures represent the standard deviation error range in the

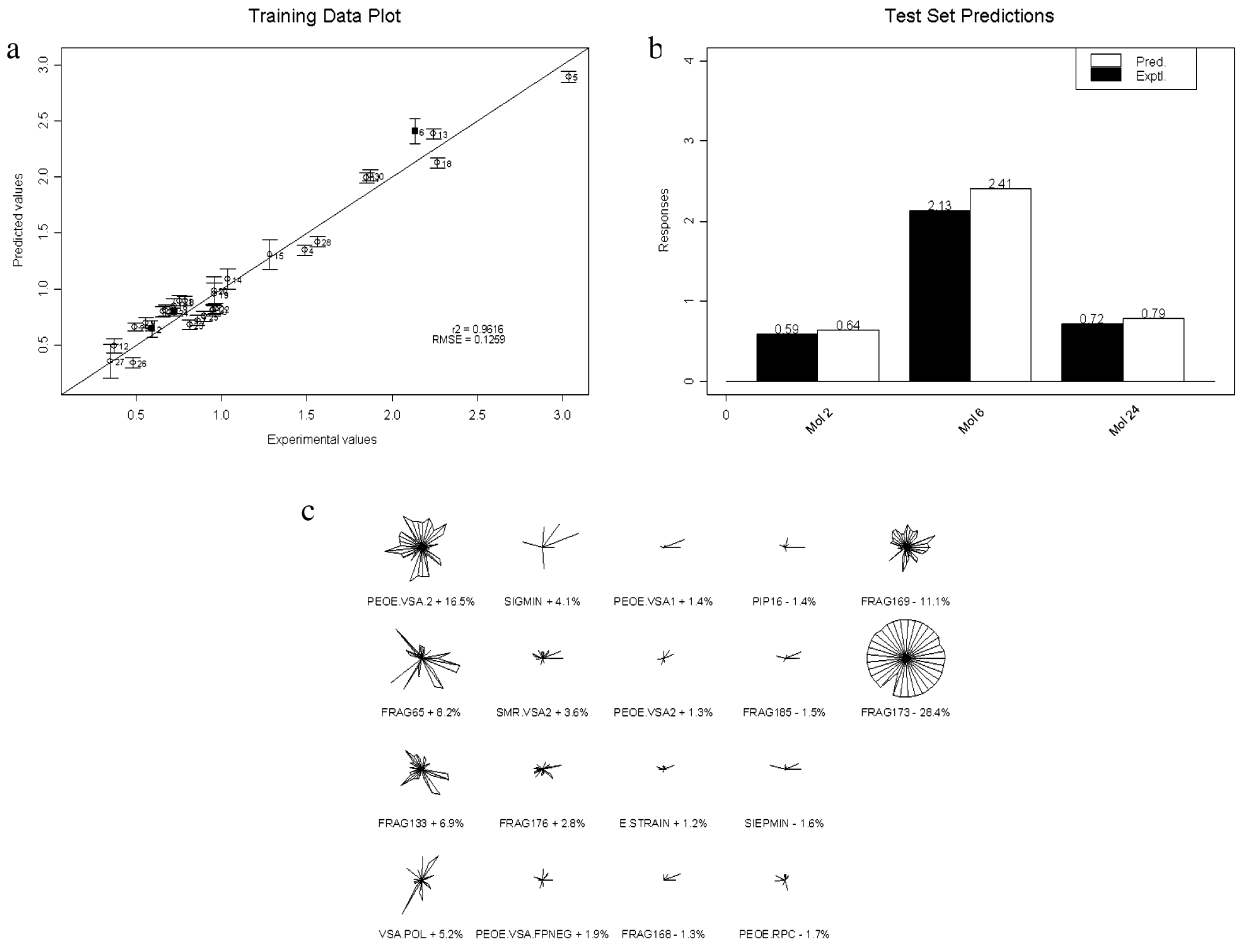

Figure 5. QSER model for percent AMY to percent APO displaced in the presence of NaCl: (a) predicted values vs observed values for the entire data set, (b) comparison of the predicted and experimental values for the test set, and (c) star plots with weight percentages for the descriptors in the model.

predicted percent protein displaced values. The predictions for the test set molecules are presented in Figures 4b, 5b, and Ib−IVb (Supporting Information).

Figure 4 and Figure I show the results for the percent amyloglucosidase displaced in the presence of NaCl and Na_2SO_4, respectively. The cross-validated r^2 for these models were 0.9725 and 0.9442, which indicate that the predicted values for the percent amyloglucosidase displaced were in good agreement with the experimental data. In fact, the percent amyloglucosidase displaced values that were predicted for the test set successfully verified the predictive power of these models (Figures 4b and Ib). Similarly, QSER models generated for apoferritin exhibited good correlations between the experimental and predicted percent apoferritin displaced values with cross-validated r^2 of 0.9523 and 0.9807 in the presence of NaCl and Na_2SO_4, respectively (Figures IIa and IIIa, Supporting Information). Furthermore, the percent apoferritin displaced values for the three displacers in the test set were successfully predicted by the QSER models, as shown in Figures IIb and IIIb (Supporting Information). It is also important to be able to predict the selectivity reversals produced by the displacers in these batch screening experiments. QSER models developed for the ratio of the percent amyloglucosidase displaced to the percent apoferritin displaced for NaCl (Figure 5a) and Na_2SO_4 (Figure IVa. Supporting Information) showed cross-validated r^2 values of 0.9616 and 0.8689, respectively. The predictions for the test set of displacers were also quite good for five out of the six cases (Figures 5b and IVb). These results indicate that the QSER models can indeed capture the subtle effects of different displacer chemistry on the displacement of these two proteins.

While these QSER models are very well correlated with the experimental data and are capable of making good predictions, they also offer the opportunity to provide insight into the displacement phenomenon by the interpretation of the descriptors used for these models. To facilitate interpretation, it was necessary to determine which descriptors were consistently important when different combinations of training and validation molecules were used. Each different set of training molecules is called a “fold”, and the model created using this set is used to make predictions on the validation molecules left out of the training set for that particular “fold”. Star plots were then created to evaluate the relative importance of each of these descriptors selected throughout each of the 40 bootstrap “folds” used for creating the composite model set. In these plots, each star corresponds to a specific descriptor and the length of each ray represents the weight or importance of this descriptor in one of the 40 bootstrap iterations. For each star plot, the selected descriptors are ranked according to their sum of the ray radii for all bootstraps, so that the most significant descriptor with positive weight appears in the upper left-hand corner while the most significant negative

Table 1. Categorization of Descriptors for All the Models on the Basis of the Physicochemical Effects That They Represent[a,b]

model	type of effect			
	electrostatic	hydrophobic	H-bonding	size/shape/VDW
AMY-Na_2SO_4	*A.ACID (A.NS ~0.993)* *PIP11 (A.POL ~0.972)* *PIP17 (A.ACID ~0.872)* PIPMAX PEOE.VSA.3	*E (Q.VSA.HYD ~ 0.863)* SLOGP.VSA8	*VSA.ACC* *PIP17 (A.POL ~0.965)*	*PETITJEAN* *PIP11 (VOL ~0.972)* *E (WEIGHT ~0.878)*
AMY-NaCl	*PIP18 (A.ACID ~0.872)* *PEOE.VSA.2* *PIP17 (A.ACID ~ 0.872)* E.ELE PEOE.VSA.FPPOS	*FRAG133*	*PIP18 (A.POL ~ 0.97)* *PIP17 (A.POL ~0.965)*	*PETITJEAN* *DENS*
APO-Na_2SO_4	*PIP17 (A.ACID ~0.872)* PEOE.VSA.FPPOS	*E (Q.VSA.HYD ~0.863)* *SLOGP.VSA9* *PEOE.VSA.FHYD* FRAG133	*PIP17 (A.POL ~0.965)*	*E (WEIGHT ~0.878)* *SMR.VSA7* PETITJEAN
APO-NaCl	*PIP16 (PEOE.PC− ~0.87)* *PIP17 (A.POL ~0.965)* EP2	*SLOGP.VSA9* *E (Q.VSA.HYD ~0.863)*	*PIP17 (A.POL ~0.965)* EP2	*E (WEIGHT ~0.878)* *PETITJEAN*
AMY/APO-Na_2SO_4	*SIGMIN* *SIKA1* PEOE.RPC SIKMIN PEOE.VSA.3 VSA.POL	E (Q.VSA.HYD ~0.863)	*SIGMIN* *SIKA1* PEOE.RPC SIKMIN VSA.POL	*BALABANJ* E (WEIGHT ~0.878) STD.DIM3 GLOB
AMY/APO-NaCl	*PEOE.VSA.2* *VSA.POL* *SIGMIN* *PEOE.VSA.FPNEG* PEOE.RPC	*FRAG133*	*SIGMIN* VSA.POL	*SMR.VSA2*

[a] Italicized descriptors contribute positively to the models. [b] Text in parentheses indicates the best correlations for the difficult to interpret descriptors.

contributor appears on the lower right-hand side. The order proceeds from left to right in a columnar fashion. The contribution of each descriptor to the model is quantified as the percentage of the total weight of the descriptor in terms of the total weights of all descriptors in the model. The star plots for the resultant models are shown in Figures 4c, 5c, and Ic−IVc (Supporting Information). The descriptors found to be of significant importance in the models can also be classified on the basis of the physical/chemical interactions that they might represent, as shown in Table 1. As seen in the table, the descriptors can be classified as being primarily electrostatic, hydrophobic, H-bonding, or size/shape/VDW-type interactions. However, for several of the descriptors that contain complex chemical information, we have employed correlation matrices/plots to determine their behavior, often resulting in their being associated with several types of interactions (e.g., PIPs and E) (Figure 6). The contributions of the important descriptors as identified by the star plots for each model are given in Table 2.

As seen in Table 2, descriptors associated with negative charges on the displacer molecules, such as A.ACID, PIP17, PIP18, and PEOE.VSA.2, were found to exist in all models. A.ACID, which was found to be the most important positive contributor in the model for AMY-Na_2SO_4, is the number of acidic atoms on the displacer molecule. This descriptor has a correlation of 0.993 with A.NS, the number of sulfur atoms in the molecule. For the molecules included in our study, most of the negative charges are found on the sulfonic groups (given by A.NS), and thus A.ACID is essentially a count of the number of negative charges on the molecule. PIP17 and PIP18 are surface area histogram bin-type descriptors associated with regions of the surface of the displacer molecule having high ionization potential, i.e., regions of tightly held electron density. This in turn corresponds to atoms having high partial negative charges. The PEOE.VSA.2 descriptor is the sum of the van der Waals (VDW) surface area with a specific range of negative partial charges. As seen in Figure 7, this range of PEOE partial charge values is mostly found on sulfur atoms in sulfonic groups in the molecule. These descriptors have high positive contributions to the models, which suggest that the higher the negative charge on the displacer molecule the higher is its efficacy. This, of course, agrees very well with the aspects of retention in anion exchange systems where an increase in the number of negative charges is know to increase retention and hence displacer efficacy.

At the same time, descriptors associated with partial positive charges on the displacers, e.g., PEOE.VSA.FPPOS, EP2, were found to be among the important negative contributors to the models. PEOE.VSA.FPPOS defines the fractional VDW surface area associated with positive partial charge. The EP2 descriptor represents the surface area of the molecule having a low electrostatic potential, such as regions having loosely held electrons (i.e., regions with high positive partial charge). The negative contributions of these descriptors suggest that as the fraction of polar positive surface area increases, the displacer efficacy decreases. Again, this is consistent with the aspects of anion exchange chromatography where positively charged solutes are less retained.

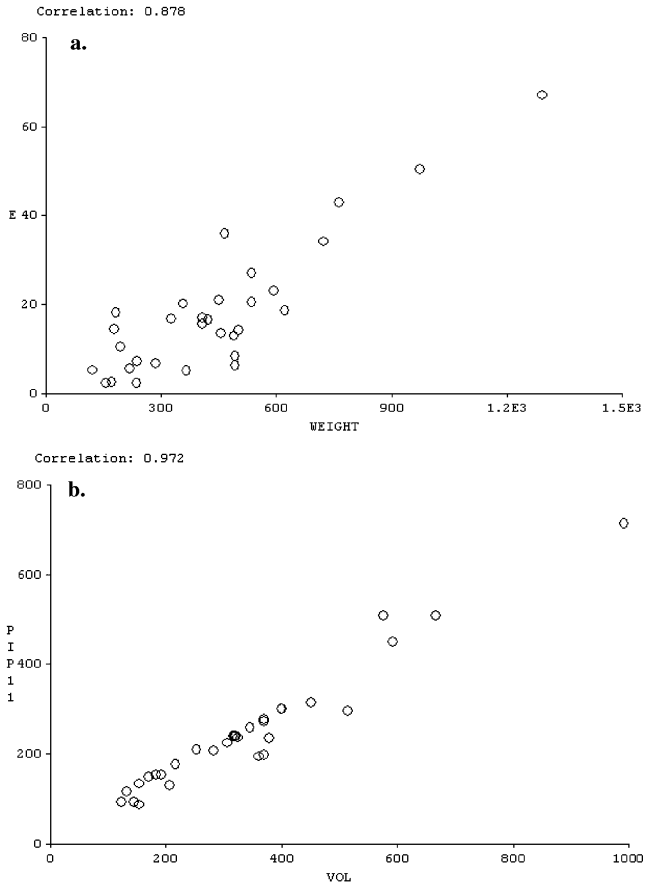

Figure 6. Correlation plots of (a) *E* vs weight (i.e., molecular weight) and (b) PIP11 vs volume (i.e., van der Waals volume of molecule).

E.ELE and PEOE.RPC are related to the charge distribution on the displacer molecules. E.ELE, which is the electrostatic component of potential energy, is associated with the proximity of charged groups on the molecule. This descriptor tends to have higher values when like charges reside in close proximity. The negative contribution of E.ELE in the model for AMY-NaCl suggests that greater charge distribution on the molecules favors displacer efficacy. The relative positive partial charge, PEOE.RPC, is defined as the ratio of the largest positive partial charge to the total positive partial charge on the molecule. PEOE.RPC is shown to have a positive contribution for the APO-Na_2SO_4 model. Although, at first this might appear to be counterintuitive for anion exchange systems (since PEOE.RPC refers to positive charges), a closer review of the mathematical definition of this descriptor hints at the fact that more concentrated positive partial charges will result in higher PEOE.RPC values for the molecules. Thus, this descriptor suggests that greater distribution of positive partial charge on the molecule surface is unfavorable for high displacer efficacy in anion exchange systems.

Shape and size descriptors were also found to be present in all of the models. For example, PETITJEAN, which refers to the aspect ratio of the molecule,[25] is a shape parameter found to have a positive contribution in most of these models. Based on the definition of PETITJEAN, the positive contribution in the models suggests that "flatter" displacer molecules will exhibit greater efficacy in displacing the proteins. Intuitively, this makes sense because a planar conformation of the displacer molecule will enable easy access of the charged groups on the displacer to the resin surface. DENS/DENSITY, which is a purely size-based descriptor, was also seen in one of the models. This descriptor had a small positive contribution to the models, suggesting that increasing the size increases displacer efficacy in anion exchange systems. Apart from the above descriptors that provide direct shape/size information, these models also contained descriptors where size information is implicit. For example, the potential energy *E* of the displacer molecule was found to have moderate to high positive contributions in the models. It is common knowledge that *E* is highly correlated with the molecular size (Figure 6a), and hence, this suggests that increasing the size of the displacer molecule increases its efficacy. The PIP-bin descrip-

Table 2. Contributions of the Key Descriptors As Identified by the Star Plots for Each Model

% AMY Displaced

NaCl		Na_2SO_4	
descriptor	% contribn	descriptor	% contribn
Descriptors with POSITIVE Contributions			
FRAG65	24.0	A.ACID	20.1
FRAG171	12.0	PETITJEAN	10.7
PIP18	7.5	PIP11	9.9
FRAG133	5.4	FRAG168	5.9
PEOE.VSA.2	5.1	E	5.0
FRAG176	4.1	FRAG171	5.0
		VSA.ACC	4.7
		PIP17	4.6
Descriptors with NEGATIVE Contributions			
E.ELE	7.3	SLOGP.VSA8	2.9
FRAG173	4.9	PIPMAX	1.6
FRAG169	4.8	PEOE.VSA.3	1.6

% APO Displaced

NaCl		Na_2SO_4	
descriptor	% contribn	descriptor	% contribn
Descriptors with POSITIVE Contributions			
E	21.9	PIP17	17.9
PIP16	19.7	E	17.0
PIP17	13.0	FRAG173	4.8
SLOGP.VSA9	11.2	SLOGP.VSA9	3.8
FRAG173	9.9	FRAG177	3.7
PETITJEAN	9.4	PEOE.VSA.FHYD	3.6
FRAG185	7.5	SMR.VSA7	3.3
		DEL. G.NMAX	3.3
Descriptors with NEGATIVE Contributions			
EP2	7.4	SIGMIN	4.6
		FRAG133	2.5

% AMY to % APO Displaced Ratios

NaCl		Na_2SO_4	
descriptor	% contribn	descriptor	% contribn
Descriptors with POSITIVE Contributions			
PEOE.VSA.2	16.5	SIGMIN	14.6
FRAG65	8.2	SIKA1	5.4
FRAG133	6.9	BALABANJ	5.4
VSA.POL	5.2	PETITJEAN	4.8
SIGMIN	4.1		
Descriptors with NEGATIVE Contributions			
FRAG173	28.4	E	11.0
FRAG169	11.1	PEOE.RPC	22.0
		SIKMIN	7.8

(25) Petitjean, M.; *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 331−337.

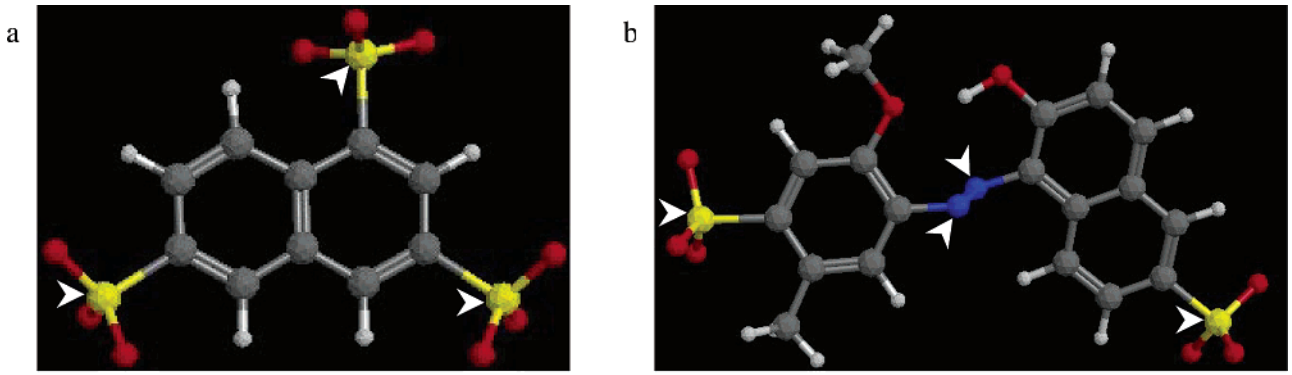

Figure 7. Molecular surface visualization tp enable identification of atoms/groups associated with the PEOE.VSA.2 descriptor having PEOE partial charges in the range [−0.15,−0.10). Atoms that contribute to PEOE.VSA.2 are indicated with arrows.

tors (i.e., PIP11, PIP17, and PIP18) were also found to have fairly good correlations with several size-related descriptors (Figure 6b). As such, they capture the charge−size relation for the displacer molecules. The positive contributions of the PIP-bin descriptors in the models coupled with their positive correlation with molecular size suggests that, for the molecules in our study, as size increases the charge is seen to increase, and hence, the efficacy also tends to increase.

An interesting observation that can be made in the star plots is the fact that hydrophobicity-related descriptors, such as PEOE.VSA.FHYD and SLOGP.VSA9, appear to be important only in the QSER models for apoferritin. PEOE.VSA.FHYD gives the fractional hydrophobic van der Waals surface area of the displacer molecule. SLOGP.VSA9 is a histogram surface area bin descriptor associated with high log P(o/w) values, i.e., regions of high hydrophobicity on the molecular surface. These descriptors exhibited positive contributions in the two apoferritin models, suggesting that increasing the hydrophobicity of the displacer increases its ability to displace apoferritin. Based on the fraction of hydrophobic amino acids in each protein, apoferritin is more hydrophobic than amyloglucosidase. Thus, one possible explanation for the above observation is that the hydrophobic displacer interactions with the stationary-phase surface are similar to those of apoferritin, resulting in enhanced displacer efficacy.

All the models were found to contain several molecular fragment-based descriptors (e.g., FRAG65, FRAG133, FRAG168, FRAG169, FRAG171, FRAG176, and FRAG185). Almost all of these FRAG-type descriptors are associated with substituted aromatic groups, which is largely a result of the structures of the molecules used in this study. Hence, they provide us with limited information about the factors responsible for displacer efficacy or selectivity. Thus, while the apparent importance of FRAG descriptors is indicative of the latent structural information contained within the training set, this must be tempered by the possibility of these descriptors serving as surrogates for other less well represented effects.

The ratios of percent displaced values for amyloglucosidase and apoferritin were modeled in an attempt to interpret the selectivity changes observed in different salts for different displacers. As seen in Table 2, the localized positive charges (PEOE.RPC) on a displacer have a greater effect on displacing apoferritin than amyloglucosidase in the presence of both salts. While SIGMIN had a positive value for the amyloglucosidase in Na_2SO_4, it had a negative value for apoferritin in Na_2SO_4. This produced a very strongly positive value for the ratio of amyloglucosidase to apoferritin in Na_2SO_4. This variable was observed to act as a class variable, in that it only exhibited three value ranges within the chemotypes used in the training data. As mentioned above, the more hydrophobic the displacer, the more efficacious it is at displacing apoferritin in the presence of either salt. This is reflected in the negative contribution of E and PEOE.VSA.FHYD on the percent displaced ratio of amyloglucosidase to apoferritin in Na_2SO_4.

It is well known that Na_2SO_4 can enhance the hydrophobic interactions between a solute and the stationary phase.[26] In our analysis, more descriptors that sample hydrophobic interactions were selected and had higher contributions in models for Na_2SO_4 as compared to those for NaCl. Further, for the displacement of apoferritin, which is more hydrophobic, we observed an increase in the number and contribution of hydrophobic descriptors as compared to the QSER models for amyloglucosidase. These results indicate that the star plots for the QSER models can provide us with useful insights into the factors responsible for displacer selectivity in our batch screening system.

CONCLUSIONS

In this study, batch displacement experiments have indicated that salt type and displacer chemistry have a significant impact on the percent protein displaced values and the selectivity of displacers for different proteins. While the percent protein displaced data obtained in the presence of different salts were qualitatively similar to the protein linear retention data, there were several exceptions for both NaCl and Na_2SO_4. For some of the displacers, significant selectivity changes were observed in both cases. These displacers may, in fact, be the first examples of "chemically selective" displacers. If these findings are applicable to column experiments, these results indicate that one may be able to create highly selective elution methodologies by changing the salt type, displacer chemistry, or both.

The experimental batch screening data were used in conjunction with various molecular descriptors for generating QSER models based on SVM feature selection and regression. The

(26) Lin, F.; Chen, W.; Hearn, M. T. W. *Anal. Chem.* **2001**, *73*, 3875−3883.

models resulted in good correlations and successful predictions for the external test set of displacers. The star plot approach has been shown to be a useful tool to aid in the interpretation of the QSER models. In addition to capturing the general aspects of anion exchange chromatography based on the selected molecular descriptors, specific trends regarding displacer design were determined such as the enhancement of displacer efficacy with a "flat" or planar conformation. Further, descriptors for hydrophobicity and charge distribution were found to be useful in describing selectivity changes that can occur based on displacer chemistry. The enhancement of hydrophobic interactions in the presence Na_2SO_4 was also captured by these models. In conclusion, this modeling approach has been shown to be quite useful not only for the a priori prediction of displacer efficacy but also for interpreting some specific effects related to displacer design and the selection of proper mobile-phase conditions. Future work will extend this analysis to a more diverse library of displacer molecules to enable us to explore the behavior of a wider structural space. In addition, we will examine the potential of chemically selective displacers for a wide variety of experimental systems.

ACKNOWLEDGMENT

The authors acknowledge NSF Grant BES-0079436, NIH Grant GM 47372-04A2, and Amersham Biosciences for funding this research. We thank Professor Kristin Bennett and Jinbo Bi from the Department of Mathematics at RPI for allowing us to use the SVM code. We also thank Yelena Shaipshaikes for her assistance in the laboratory.

SUPPORTING INFORMATION AVAILABLE

Additional information as noted in the text. This material is available free of charge via the Internet at http://pubs.acs.org.

Received for review February 17, 2003. Accepted July 17, 2003.

AC0341564