

# Prediction of Protein Retention Times in Anion-Exchange Chromatography Systems Using Support Vector Regression

Minghu Song,<sup>†</sup> Curt M. Breneman,<sup>\*,†</sup> Jinbo Bi,<sup>‡</sup> N. Sukumar,<sup>†</sup> Kristin P. Bennett,<sup>‡</sup>  
Steven Cramer,<sup>§</sup> and Nihal Tugcu<sup>§</sup>

Departments of Chemistry, Mathematics, and Chemical Engineering, Rensselaer Polytechnic Institute,  
110 8th Street, Troy, New York, 12180

Received August 14, 2002

Quantitative Structure-Retention Relationship (QSRR) models are developed for the prediction of protein retention times in anion-exchange chromatography systems. Topological, subdivided surface area, and TAE (Transferable Atom Equivalent) electron-density-based descriptors are computed directly for a set of proteins using molecular connectivity patterns and crystal structure geometries. A novel algorithm based on Support Vector Machine (SVM) regression has been employed to obtain predictive QSRR models using a two-step computational strategy. In the first step, a sparse linear SVM was utilized as a feature selection procedure to remove irrelevant or redundant information. Subsequently, the selected features were used to produce an ensemble of nonlinear SVM regression models that were combined using bootstrap aggregation (bagging) techniques, where various combinations of training and validation data sets were selected from the pool of available data. A visualization scheme (star plots) was used to display the relative importance of each selected descriptor in the final set of “bagged” models. Once these predictive models have been validated, they can be used as an automated prediction tool for virtual high-throughput screening (VHTS).

## I. INTRODUCTION

*Ion-Exchange Chromatography* (IEC) is a widely accepted standard bioseparation technique that has been growing in importance during the past decade in keeping with current rapid developments in biotechnology. To date, there are two main kinds of IEC: cation-exchange and anion-exchange chromatography, determined by whether a negative charge (cation-exchange) or a positive charge (anion-exchange) is carried by the functional groups on the surface of the IEC stationary phase. The ionic biopolymers, such as proteins, are separated primarily through the electrostatics interactions between the charged surface of the ion-exchange resin and the ionic solutes bearing the opposite charge. In the case of anion-exchange chromatography, negatively charged proteins bind in a transient fashion to the positively charged stationary phase sites, as long as the salt concentration is kept low. Proteins bound with different degrees of interaction can be separated with the aid of an increasing salt gradient. The selectivity of this technique can be optimized by varying the composition of the stationary phase as well as the pH of the mobile phase. Consequently, one of the major challenges in ion exchange bioseparation is to select appropriate chromatographic materials for a given biological mixture. It has been suggested that virtual screening of separation materials in a manner that parallels current QSAR (*Quantitative Structure-Activity Relationship*) methods in drug design would facilitate the selection of proper chromatographic conditions and speed up development processes.

\* Corresponding author phone: (518)276-2678; fax: (518)276-4887, e-mail: brenecc@rpi.edu.

<sup>†</sup> Department of Chemistry.

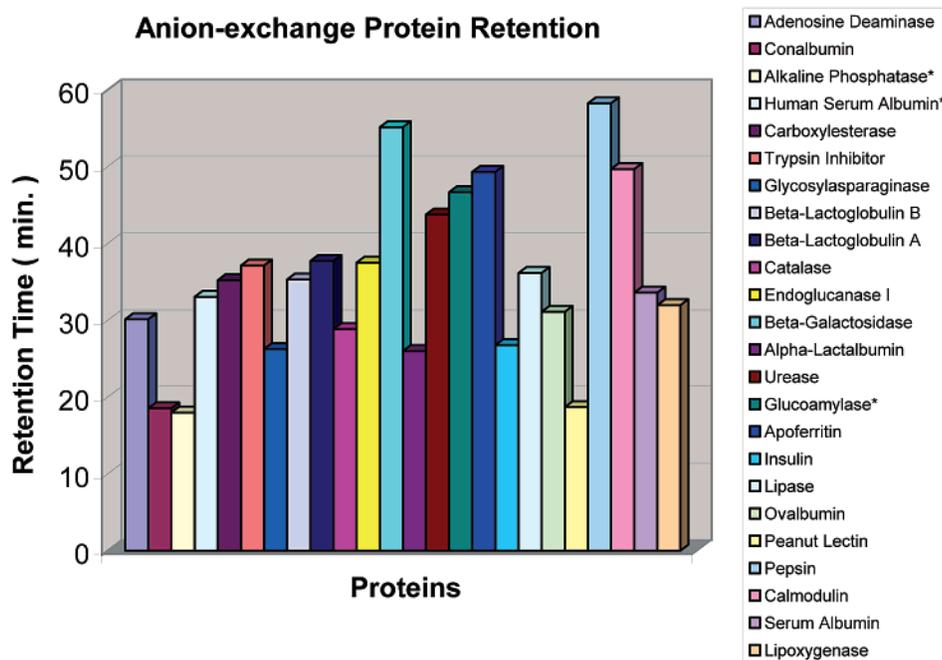
<sup>‡</sup> Department of Mathematics.

<sup>§</sup> Department of Chemical Engineering.

As a result, there is increasing interest within the chromatography community in the development of *Quantitative Structure-Retention Relationship* (QSRR) models<sup>1</sup> based on linear or nonlinear modeling techniques, including *Principal Component Regression* (PCR),<sup>2</sup> *Partial Least Squares* (PLS),<sup>3</sup> and *Artificial Neural Networks* (ANN).<sup>4,5</sup> The major aims of these studies are to construct improved QSRR models to predict the retention behavior of solutes in different stationary phases or salt conditions as well as to build a valuable chromatographic interpretation tool for the solute retention mechanisms. Due to computational bottlenecks in descriptor generation and machine learning algorithms, most current approaches are only applicable for small molecules. Recent research has focused on the adaptation of the TAE (*Transferable Atom Equivalent*) electron density-derived descriptor technique to large molecules such as proteins.<sup>6</sup> In that study, partial least-squares models constructed using subsets of TAE descriptors were found to be capable of predicting protein retention with good accuracy. In the current study, we present a novel modeling approach based on *Support Vector Machine* (SVM) Regression<sup>7</sup> to predict the retention time of proteins in anion exchange systems. A visualization tool, the star plot, is employed to aid in model interpretation. The predictive power of the resulting models is demonstrated by testing them on unseen data that were not used during either descriptor selection or model generation.

## II. DATA SET AND DESCRIPTOR GENERATION

**Protein Retention Data Set.** The crystal structures of 24 structurally diverse proteins with similar isoelectric points (PI) were downloaded from the RSCB Protein Data Bank<sup>8</sup> for analysis. The retention times for these proteins were obtained by carrying out linear gradient chromatography



**Figure 1.** Proteins and their experimental retention times. Entries marked by \* were used as the test set.

81 using the anion exchange stationary phase Source 15Q. The  
 82 names and the experimental retention times of the 24 proteins  
 83 are provided in Figure 1. Three proteins were randomly  
 84 selected as external test cases from this original list.

85 SYBYL v6.5 software<sup>9</sup> was used to preprocess the raw  
 86 macromolecular structures by eliminating the waters of  
 87 crystallization and adding hydrogen atoms to satisfy neutral  
 88 valences on all atoms. A total of 243 descriptors was then  
 89 computed for these proteins using both RECON<sup>10</sup> and MOE  
 90 programs to give a composite set of traditional and electron  
 91 density-derived TAE descriptors.

92 **Quantum Theory of Atoms in Molecules (QT-AIM) and**  
 93 **TAE/RECON Descriptors.** Quantum chemical descriptors  
 94 offer an attractive alternative to traditional QSAR/QSPR  
 95 molecular descriptors by expressing a more accurate and  
 96 detailed description of the electronic and geometric molecular  
 97 properties and the interaction between them.<sup>11</sup> However, even  
 98 with the rapid advances in computer architecture and the  
 99 anticipated continued growth in computational power, a direct  
 100 calculation of the properties of large molecules at a high  
 101 level of theory is prohibitive. Bader's quantum theory of  
 102 Atoms in Molecules (AIM)<sup>12,13</sup> provides the framework for  
 103 reconstructing large complicated molecules from a number  
 104 of small electron density fragments while still achieving an  
 105 good approximation to the properties of the intact molecules.  
 106 In AIM theory, the electron density of a molecule can be  
 107 partitioned into distinct electron density basins (the regions  
 108 of space occupied by the corresponding atoms), each  
 109 containing an atomic nucleus. These electron density frag-  
 110 ments are essentially bounded by surfaces of zero net flux  
 111 in the electron density, which correspond to the steepest  
 112 descent pathways from each bond critical point. An atomic  
 113 property ( $A$ ) can then be expressed as the integral of a  
 114 corresponding property density  $\rho_A(r)$  over an atomic basin:

$$A(\Omega) = \int_{\Omega} d\tau \rho_A(r) \text{ where } \rho_A(r) = (N/2) \int d\tau' \{ \psi^* \hat{A} \psi + (\hat{A} \psi)^* \psi \} \quad (1)$$

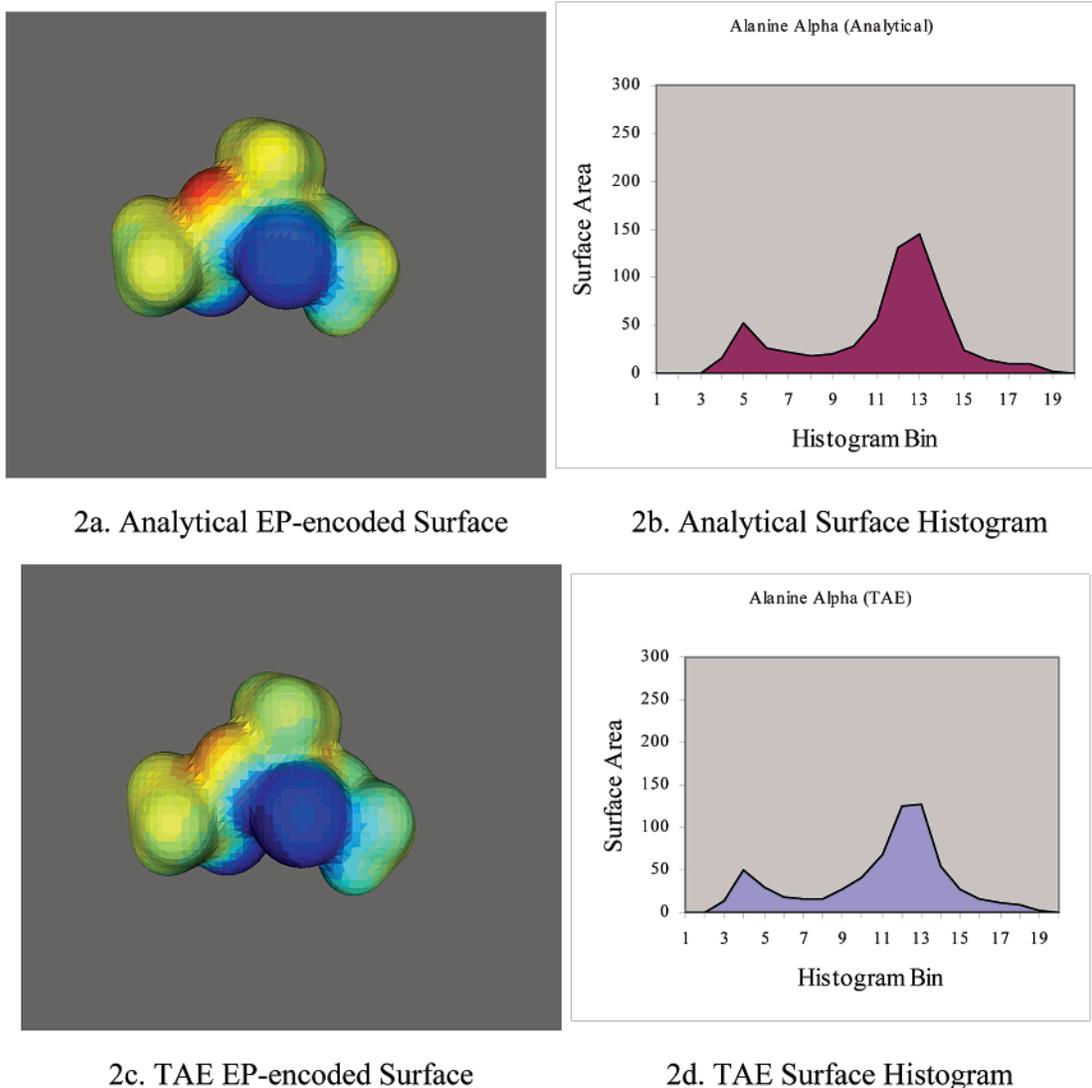
**Table 1.** TAE Atomic Electronic Surface Properties

EP	electrostatic potential
Del(Rho)·N	electron density gradient normal to 0.002 e/au <sup>3</sup> electron density isosurface
G	electronic kinetic energy density $G =$ $(-\hbar/4m) \int \{ \nabla \psi^* \cdot \nabla \psi \} d\tau$
K	electronic kinetic energy density $K =$ $(-\hbar/4m) \int \{ \psi^* \nabla^2 \psi + \psi \nabla^2 \psi^* \} d\tau$
Del(K)·N	gradient of K electronic kinetic energy density normal to surface
Del(G)·N	gradient of B electronic kinetic energy density normal to surface
Fuk	Fukui F <sup>+</sup> function scalar value
Lapl	Laplacian of the electron density $\nabla^2 \rho$
BNP	bare nuclear potential $BNP_{(i)} = \sum_{j=1}^n q_j/r_{ij}$
PIP	local average ionization potential $PIP(r) = \sum_i \rho_i(r)  \epsilon_i  / \rho(r)$

115 These atomic properties possess a high degree of transfer-  
 116 ability from the electronic environment in one molecule to  
 117 another molecule with a similar environment. Consequently,  
 118 the properties of a functional group or whole molecule can  
 119 be obtained by adding these atomic properties together:

$$A_{molecule} = \sum_{\Omega} A(\Omega) \quad (2)$$

120 Based on AIM theory, Breneman introduced the concept  
 121 of "Transferable Atom Equivalents" (TAEs),<sup>10,14</sup> which are  
 122 composed of atomic electron density fragments bounded by  
 123 interatomic zero-flux surfaces ( $\nabla \rho(r) \cdot n(r) = 0$ , for all points  
 124 on the surface) and an extended  $\rho = 0.002$  electron/au<sup>3</sup>  
 125 isodensity surface that approximates the condensed-phase van  
 126 der Waals surface. TAE fragments carry 10 atomic charge  
 127 density-derived properties (listed in Table 1) that were pre-  
 128 computed from small molecules using ab initio wave  
 129 functions at the 6-31+G\* level of theory. As evident from  
 130 the table, TAE electron density reconstructions provide not  
 131 only molecular electron densities but also electronic kinetic  
 132 energy densities and local average ionization potentials as  
 133 well as other first- and second-derivative properties of the



2a. Analytical EP-encoded Surface

2b. Analytical Surface Histogram

2c. TAE EP-encoded Surface

2d. TAE Surface Histogram

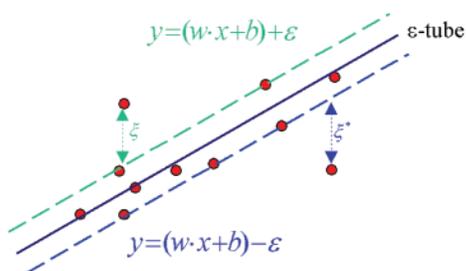
**Figure 2.** Electrostatic potential surface distributions and histograms generated for alanine using both TAE and analytical ab initio methods. The color scheme in parts a and c corresponds to different values of EP on the molecular surface. The distributions of the surface electrostatic potentials are characterized as histogram descriptors using binning techniques as illustrated in parts b and d. Descriptors for larger molecules or proteins can be computed following a similar scheme. TAE reconstruction of the proteins used in this study required approximately 60 s on a single 1.7 GHz processor Linux PC.

134 density. The distributions of these electronic properties  
 135 computed on  $0.002\text{-e}/\text{au}^3$  electronic density isosurfaces may  
 136 be characterized as molecular property descriptors in several  
 137 ways. TAE histogram descriptors can be produced by  
 138 recording the distribution of the properties as surface  
 139 histograms that quantified the molecular surface areas with  
 140 specific ranges of each property value. In addition to these  
 141 histogram descriptors, property extrema, average values and  
 142 standard deviations of the property distributions (in some  
 143 cases with separate  $\sigma$  values for positive and negative  
 144 portions of the range) were also included in the TAE  
 145 descriptor set.

146 The TAE library consists of a set of precalculated atomic  
 147 fragments structured in a form that allows the atomic  
 148 fragments involved in the new molecule to be rapidly  
 149 retrieved. The RECON (*RECON*struction) program reads the  
 150 atomic connectivity information of the protein and assigns  
 151 the closest fragment match from the TAE library to each  
 152 atom based on atom type, hybridization and structural  
 153 environment. By summing up the corresponding atomic

properties of the constituent fragments, we can obtain a large  
 set of electron density-based TAE descriptors for macro-  
 molecules. These descriptors provide information about  
 basicity, hydrophobicity, hydrogen-bonding capacity and  
 polarity as well as molecular polarizability. For example,  
 surface property histograms such as the electrostatic potential  
 distribution of alanine histogram shown in Figure 2 may be  
 computed using TAE/RECON program. As shown in the  
 figure, the TAE electrostatic potential distribution represents  
 the analytical ab initio result quite effectively.

154  
 155  
 156  
 157  
 158  
 159  
 160  
 161  
 162  
 163  
 164 The TAE/RECON approach has been shown to be effective  
 165 in QSPR studies.<sup>15</sup> It is a resource-efficient alternative  
 166 to HF/SCF or DFT ab initio calculations, which can be  
 167 prohibitive even for molecules of modest size. The CPU and  
 168 disk resources required for TAE reconstruction are compa-  
 169 rable to those utilized by molecular mechanics energy  
 170 computations. The TAE QSPR descriptors for individual  
 171 proteins or large databases can be computed within seconds  
 172 on modest workstations.



**Figure 3.** Graphical depiction of an  $\epsilon$ -insensitive loss function and an  $\epsilon$ -tube. Only the deviations of data points outside the  $\epsilon$ -tube, such as  $\xi$  and  $\xi^*$ , will be considered as the errors and thus be penalized—in this case two points are used as examples.

173 **MOE Descriptors.** The MOE program provides a widely  
 174 applicable set of classical molecular descriptors, including  
 175 traditional physicochemical properties, connectivity-based  
 176 topological 2D and shape-dependent 3D molecular features.  
 177 These descriptors have been applied to the construction of  
 178 QSAR/QSPR models for boiling point, vapor pressure, and  
 179 the free energy of solvation in water as well as water  
 180 solubility and blood-brain barrier penetration.<sup>16</sup>

### 181 III. MODELING METHODOLOGY

182 **Support Vector Regression (SVR) Overview.** In recent  
 183 years, there has been a lot of interest in studying support  
 184 vector machines (SVMs) in the field of machine learning.  
 185 SVMs are a class of supervised learning algorithms initially  
 186 proposed by Vapnik.<sup>17,18</sup> To date, SVMs have been applied  
 187 successfully to a wide range of pattern recognition problems,  
 188 such as image recognition,<sup>19</sup> microarray gene expression  
 189 classification,<sup>20</sup> protein folding recognition,<sup>21</sup> protein struc-  
 190 tural class prediction,<sup>22</sup> identification of protein cleavage  
 191 sites,<sup>23</sup> QSAR and other pharmaceutical data analysis.<sup>20,24</sup>  
 192 Although SVMs were originally developed for classifi-  
 193 cation, Vapnik enabled them to solve regression problems  
 194 by choosing a suitable cost function ( $\epsilon$ -insensitive loss  
 195 function) that enables a sparse set of support vectors to be  
 196 obtained.<sup>17</sup>

197 Normal regression procedures are often stated as the  
 198 processes deriving a function  $f(x)$  that has the least deviation  
 199 between predicted and experimentally observed responses  
 200 for all training examples. One of the main characteristics of  
 201 SVR is that instead of minimizing the observed training error,  
 202 SVR attempts to minimize the generalization error bound  
 203 so as to achieve higher generalization performance. This  
 204 generalization error bound is the combination of the training  
 205 error and a regularization term that controls the complexity  
 206 of hypothesis space. The first term is calculated by the  
 207  $\epsilon$ -insensitive losses<sup>17</sup>

$$L_\epsilon(y - f(x)) = |y - f(x)|_\epsilon = \min(0, |y - f(x)| - \epsilon) \quad (3)$$

208 in which  $\epsilon$  is the tolerance to error and we only consider  
 209 those deviations larger than  $\epsilon$  as errors. The  $l_2$ -norm  $1/2\|\omega\|^2$   
 210 of normal vector is typically adopted as a regularization factor  
 211 and  $\omega$  is the weight vector to be determined in the function  
 212  $f$ . This algorithm, called  $\epsilon$ -SVR, seeks to find a function  $f^*$   
 213  $\in F = \{f: R^N \rightarrow R\}$  based on a training set of  $M$  examples

$(x_i, y_i)$  with  $x_i \in R^N$  by minimizing the overall regularized  
 214 risk functional<sup>25</sup> 215

$$\frac{1}{2}\|\omega\|^2 + C \sum_{i=1}^M |y_i - f(x_i)|_\epsilon \quad (4)$$

216 where  $C$  is a fixed regularization constant determining the  
 217 tradeoff between training error (empirical loss) and model  
 218 complexity. Figure 3 illustrates what the  $\epsilon$ -insensitive loss  
 219 function looks like.

220 If the hypothesis space  $F$  consists of a linear function in  
 221 the form  $\langle w \cdot x \rangle + b$ , then the SVR problem can be posed as  
 222 a convex optimization problem as follows:

$$\left[ \begin{array}{l} \text{minimize} \quad \frac{1}{2}\|\omega\|^2 + C \sum_{i=1}^M (\xi_i + \xi_i^*) \\ \text{subject to} \quad y_i - \langle w \cdot x_i \rangle - b \leq \epsilon + \xi_i, \quad \xi_i \geq 0, \\ \quad \quad \quad \langle w \cdot x_i \rangle + b - y_i \leq \epsilon + \xi_i^*, \quad \xi_i^* \geq 0, \\ \quad \quad \quad i = 1, 2, \dots, M \end{array} \right] \quad (5)$$

223 A favorable property of the above formulation is that its  
 224 solution is robust with respect to small changes in the training  
 225 set.

226 Another major characteristic of Support Vector methods  
 227 is that it implicitly maps the original input space to a high  
 228 dimensional feature space  $x \mapsto \Phi(x)$  by means of so-called  
 229 kernel functions based on Mercer's theorem, whereupon a  
 230 linear regression function  $f(x) = \langle w \cdot \Phi(x) \rangle + b$  is constructed  
 231 upon the feature space to achieve a nonlinear model in the  
 232 original input space. Thus Support Vector generalization  
 233 error, unlike those of other machine learning methods, is not  
 234 directly related to the original input dimensionality of the  
 235 problem. By the optimality conditions of the quadratic  
 236 programming formulation of SVMs, the normal vector  $w$  can  
 237 be expressed as  $w = \sum_{i=1}^M \alpha_i \Phi(x_i)$  and the function  $f$  can be  
 238 written in the form of a kernel expansion as

$$f(x) = \sum_{i=1}^M \alpha_i k(x_i, x) + b \quad \text{where } k(x_i, x) = \langle \Phi(x_i) \cdot \Phi(x) \rangle \quad (6)$$

239 In classical support vector regression, the proper value for  
 240 the parameter  $\epsilon$  is difficult to determine beforehand. Fortu-  
 241 nately, this problem is partially resolved in a new algorithm,  
 242  $\nu$  support vector regression ( $\nu$ -SVR),<sup>26,27</sup> in which  $\epsilon$  itself is  
 243 a variable in the optimization process and is controlled by  
 244 another new parameter  $\nu \in (0, 1]$ .  $\nu$  is the upper bound on  
 245 the fraction of error points or the lower bound on the fraction  
 246 of points inside the  $\epsilon$ -insensitive tube. Thus a good  $\epsilon$  can be  
 247 automatically found by choosing  $\nu$ , which adjusts the  
 248 accuracy level to the data at hand. This makes  $\nu$  a more  
 249 convenient parameter than the one used in  $\epsilon$ -SVR.

250 Since solving quadratic programming problems is usually  
 251 more computationally expensive than solving linear pro-  
 252 gramming problems, efforts have been made to derive a  
 253 linear programming formulation for SVR. Instead of using  
 254 the Euclidean norm i.e.,  $l_2$ -norm regularization of  $w$ , the  
 255 sparse  $\nu$ -SVR always regularizes through applying  $l_1$ -norm,  
 256 a sparse favoring norm, directly to coefficients  $\alpha_j, j = 1, \dots, M$   
 257 in the kernel expansion of  $f$ . The  $l_1$ -norm of the vector  $\alpha$  is  
 258  $\sum_{j=1}^M |\alpha_j|$ , which can be rewritten as  $\sum_{j=1}^M (\alpha_j + \alpha_j^*)$  if we  
 259 define  $\alpha_j = \alpha_j - \alpha_j^*$ , where  $\alpha_j \geq 0$  and  $\alpha_j^* \geq 0$ .

260 Due to these features of linear  $\nu$  support vector regression,  
 261 we adopted it for our numerical experiments on the QSRR  
 262 problem. A two-step computational strategy was adopted:  
 263 First, a sparse linear SVM was utilized as a variable selection  
 264 method to identify relevant molecular descriptors; and then  
 265 in the next step, a set of nonlinear SVM models derived by  
 266 kernel mapping were constructed using the selected features.  
 267 In addition, a statistical technique called “bagging” (Bootstrap  
 268 Aggregation) was employed to improve model generalization  
 269 performance.

270  **$l_1$ -Norm SVR Linear Feature Selection.** In ion-exchange  
 271 chromatography systems, the solutes interact with the  
 272 stationary phase in the column through a combination of  
 273 intermolecular interactions as the mobile phase flows down  
 274 through the column. Since it is not possible to know a priori  
 275 which molecular descriptors are most relevant for describing  
 276 these interactions, a comprehensive set of descriptors is  
 277 employed in the initial steps of QSRR model generation.  
 278 This results in a situation where there are far fewer  
 279 observations than the number of molecular descriptors. As  
 280 is well-known in both the chemical and statistical communi-  
 281 ties, the accuracy of prediction is not monotonic with respect  
 282 to the number of features employed in the model, because  
 283 some descriptors may be found to be unnecessary or  
 284 irrelevant, while inclusion of too many descriptors may  
 285 produce fortuitous correlations and over-trained models.  
 286 Therefore, in this extreme of very few observations with very  
 287 many descriptors, it is essential to utilize efficient feature  
 288 selection and regularization methods. Even though SVMs  
 289 are claimed to be insensitive to the problem of dimensionality  
 290 with kernels implemented as discussed above, reduction of  
 291 the input space can still help to speed up the learning process  
 292 by removing irrelevant features and emphasizing only a few  
 293 relevant features to make the interpretation more convenient.  
 294 That is why feature selection methods have received much  
 295 attention recently in QSAR or QSPR studies. Several  
 296 algorithms, such as forward selection,<sup>28</sup> simulated anneal-  
 297 ing,<sup>29</sup> genetic algorithms,<sup>30,31</sup> K-nearest neighbor,<sup>32</sup> evolution-  
 298 ary programming<sup>33,34</sup> artificial ants<sup>35,36</sup> and binary particle  
 299 swarms,<sup>37</sup> have been implemented for feature selection in  
 300 the scientific literature.

301 The feature selection method used in this work exploits  
 302 the fact that sparse SVM modeling using a linear hypotheses  
 303 with  $l_1$ -norm regularization inherently performs feature  
 304 selection as a side effect of minimizing function capacity  
 305 during the modeling process.<sup>38</sup> In a linear regression model  
 306 of the form  $y = \langle a \cdot x \rangle + b$ , each component of  $\alpha$  provides a  
 307 weight for the corresponding feature, thus providing a  
 308 measure of its significance in the model. Moreover, the sign  
 309 of each component  $\alpha_i$  indicates the effect of the  $i^{\text{th}}$  feature  
 310 on the hypothesis. If  $\alpha_i > 0$ , the feature contributes positively  
 311 to the observed response  $y$ , and when negative it diminishes  
 312  $y$ . In linear Support Vector regression, the pertinent process  
 313 involves maximization of the “margin”, a term that is  
 314 inversely proportional to the norm of the weights  $\|w\|$ . The  
 315 margin is defined as the geometric size of the  $\epsilon$ -tube. In the  
 316 case of linear SVMs, this size of the margin provides a  
 317 measure of model complexity. An effect of maximizing the  
 318 margin (or minimizing the norm of the weights) is to make  
 319 the optimal weight vector more sparse. Sparsity is defined  
 320 here as the average number of nonzero components (descrip-  
 321 tor weights) in the optimal weight vector. This method of

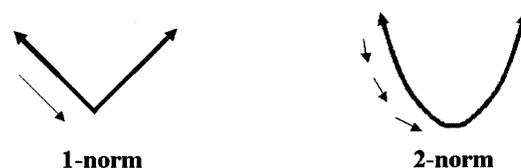


Figure 4. The weights of irrelevant descriptors will converge to zero much faster when using the  $l_1$ -norm compared to the  $l_2$ -norm.

feature selection is formulated as a sparse  $\nu$ -SVR without kernel mapping, which can be stated in the following manner:

$$\left[ \begin{array}{l} \text{minimize} \quad \frac{1}{2} \sum_{j=1}^N (\alpha_j + \alpha_j^*) + C \frac{1}{M} \sum_{i=1}^M (\xi_i + \xi_i^*) + C\nu\epsilon \\ \text{subject to} \quad y_i - \sum_{j=1}^N (\alpha_j - \alpha_j^*) x_{ij} - b \leq \epsilon + \xi_i, \quad i = 1, 2, \dots, M \\ \sum_{j=1}^N (\alpha_j - \alpha_j^*) x_{ij} + b - y_i \leq \epsilon + \xi_i, \quad i = 1, 2, \dots, M \\ \alpha_j, \alpha_j^*, \xi_i, \xi_i^*, \epsilon \geq 0, \quad j = 1, 2, \dots, N, \quad i = 1, 2, \dots, M \end{array} \right] \quad (7)$$

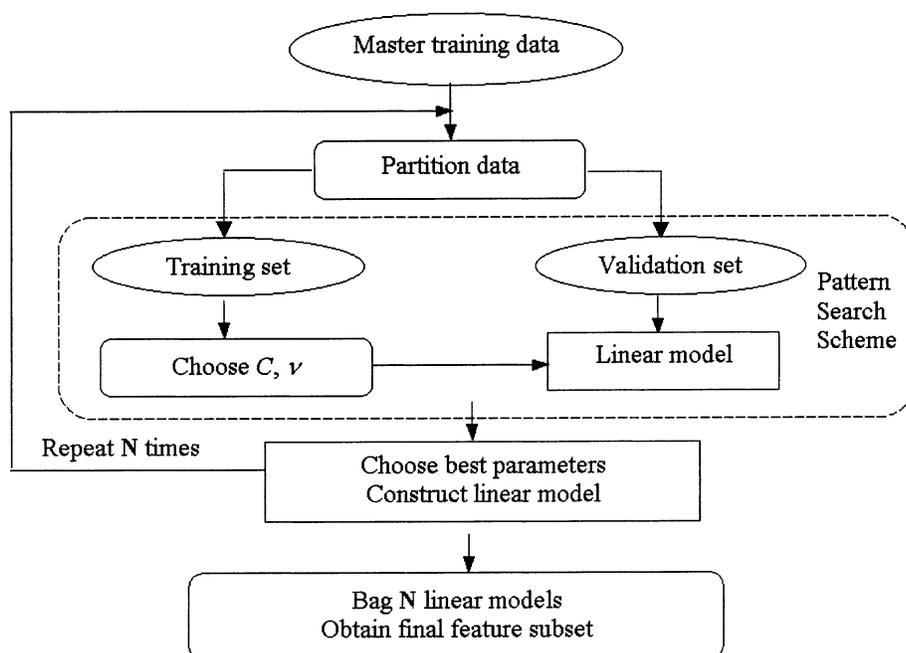
One-norm sparse SVR optimization can enhance the sparsity of the  $l_1$ -norm of  $\alpha$  as shown in Figure 4, because it is easier to drive the weights of irrelevant descriptors to zero. Those descriptors with nonzero weights then become potentially relevant features to be selected and used to build a subsequent nonlinear model.

Since QSRR data are sometimes comprised of relatively few examples represented by many correlated descriptors, even small perturbations of the training set may lead to large variations in the learning process. This eventuality results in the generation of different linear models and different sets of nonzero-weighted descriptors for related training sets. Recent research reported in the literature has shown that if used with care, ensemble modeling can improve the generalization performance particularly for unstable nonlinear models, such as those involving neural networks.<sup>39</sup> Thus to stabilize the learning process and ensure that a robust set of features are selected in the present work, the technique of bootstrap aggregation (or “bagging”) was used in the form originally proposed by Breiman.<sup>40,41</sup> The idea is to construct a series of individual sparse SVR predictors (models) using a bootstrap resampling technique,<sup>42</sup> record the selected descriptors for each individual bootstrap and then take a union of all descriptors into a single final feature set.

The overall feature selection scheme is illustrated in Figure 5. The following process was carried out in this work:

- Multiple training and validation sets were developed from a master training data set using a bootstrapping protocol;
- A series of sparse linear SVMs was created that exhibit good generalization following the common accession using  $n$ -fold cross-validation and quantified by cross-validated correlation coefficients;
- Subsets of features having nonzero weights in the linear models were selected;
- Finally, all features obtained in the last step were aggregated to produce the final candidate set of descriptors.

**Nonlinear Regression Bagging Models.** Once a set of features is selected, a nonlinear  $\nu$ -SVR with a kernel



**Figure 5.** General framework of feature selection scheme.

362 formulation such as shown in eq 9 is used to construct the  
363 QSRR models. The Radial Basis Function (RBF)

$$k(x, x') = \exp(-||x - x'||^2/2\sigma^2) \quad (8)$$

364 was chosen as the kernel in our computational studies.

365 This allows us to obtain the regression function  $f$  as a linear  
366 combination of only a few kernel functions. The sparse  
367  $\nu$ -SVR is formulated as follows:

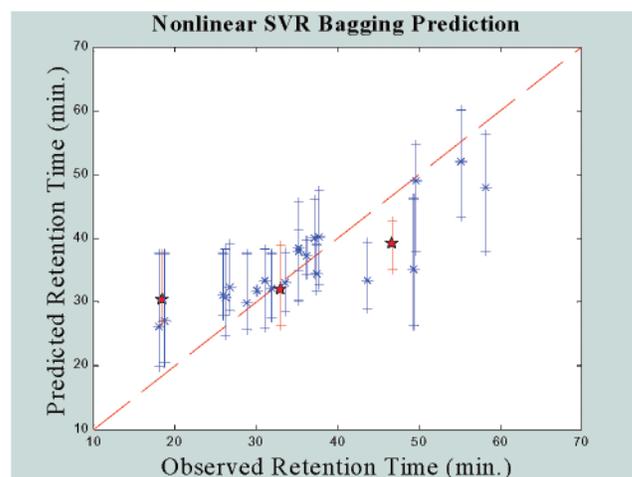
$$\left[ \begin{array}{l} \text{minimize } \frac{1}{2} \sum_{j=1}^M (\alpha_j + \alpha_j^*) + C \frac{1}{M} \sum_{i=1}^M (\xi_i + \xi_i^*) + C\nu\epsilon \\ \text{subject to } y_i - \sum_{j=1}^M (\alpha_j - \alpha_j^*) k(x_i, x_j) - b \leq \epsilon + \xi_i, i = 1, 2, \dots, M \\ \sum_{j=1}^M (\alpha_j - \alpha_j^*) k(x_i, x_j) + b - y_i \leq \epsilon + \xi_i, i = 1, 2, \dots, M \\ \alpha_j, \alpha_j^*, \xi_i, \xi_i^*, \epsilon \geq 0, i, j = 1, 2, \dots, M \end{array} \right] \quad (9)$$

368 A simple grid search<sup>43</sup> was employed to choose appropriate  
369 values for the kernel parameter  $\sigma$  as well as the capacity  
370 factor  $C$  and the parameter  $\nu$ . More details of how the  
371 parameters  $C, \nu$  are selected using a pattern search technique  
372 can be found in Bennett's recent publication.<sup>38</sup> To again  
373 reduce the variance of the predicted values, the same  
374 "bagging" technique was utilized in training the final  
375 regression model over the selected features based on the  
376 nonlinear SVR predictors  $\phi_n(x)$ .

$$\phi_{bag}(x) = \frac{1}{N} \sum_{n=1}^N \phi_n(x),$$

where  $N$  is the cardinality of the ensemble (10)

377 The same cross-validation procedure as described earlier  
378 was used to quantify the predictive capabilities of individual  
379 predictors and that of the whole predictor ensemble.



**Figure 6.** The prediction scatter plot using all descriptors before feature selection.

**Implementation.** The SVR feature selection and modeling  
program was implemented using the CPLEX optimization  
toolbox<sup>44</sup> and the C programming language as available in  
the Department of Mathematics at RPI and installed on an  
IBM-AIX Unix platform. Star plot visualization graphics  
were generated using the S-PLUS 2000 software package.<sup>45</sup>

#### IV. RESULTS AND DISCUSSION

**SVR Feature Selection and Bagging Prediction Results.**  
The aim of this work was to generate predictive models for  
protein ion-exchange chromatographic retention times with  
high accuracy as well as to characterize the main interaction  
mechanisms that account for the retention behavior in anion  
exchange systems.

Figure 6 shows retention time modeling results obtained  
before any feature selection using all topological and  
quantum mechanical descriptors mentioned in Section II. In  
this figure, the observed retention times (horizontal axis) are

**Table 2.** Definition of the Relevant Descriptors Obtained from Sparse SVR Feature Selection

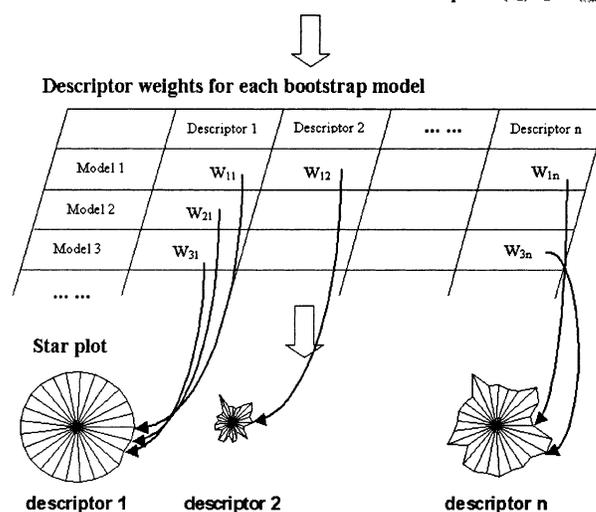
descriptor name	chemical information encoded in these descriptors
PEOE.VSA.FPPOS (MOE)	Fraction of positive polar van der Waals surface area. The Partial Equalization of Orbital Electronegativities (PEOE) method of calculating the atomic charges was developed by Gasteiger <sup>46</sup>
FCHARGE (MOE)	Total charge of molecule (sum of formal charges)
PIP2 (TAE)	The second histogram bin of PIP property. Local average ionization potential in the low range
PIP20 (TAE)	The last histogram bin of PIP property. Local average ionization potential in the high range
SIKIA(TAE)	K electronic kinetic energy density, which correlates with the presence and strength of Bronsted basic sites. (integral average)
SIGIA (TAE)	Derived from the G electronic kinetic energy density on the molecular surface. Similar in interpretation to SIKIA, but provide supplemental information.
VSA.POL	Sum of van der Waals surface of "polar" atoms

397 plotted against the corresponding predicted values for each  
 398 protein obtained using nonlinear SVR models. The blind test  
 399 data, as indicated in red, were held out and were not involved  
 400 in model generation or validation. The asterisk on each  
 401 vertical bar shows the bagged result of 12 bootstraps for each  
 402 protein, and the length of the bar represents the full prediction  
 403 range of retention time for each protein generated by the 12  
 404 bagged models. The cross-validation step produced an  $R_{CV}^2$   
 405 = 0.851 and the blind test set an  $R_{bag}^2 = 0.926$ .

406 As discussed above, sparse  $\nu$ -SVR approaches were  
 407 adapted to select only those features relevant to anion-  
 408 exchange protein retention under the experimental conditions  
 409 used to develop the data set. In this feature selection  
 410 procedure, 20 sparse linear SVM models were constructed  
 411 based on 20 different random partitions of the training data.  
 412 In the final aggregate SVR model, there were only seven  
 413 descriptors remaining with nonzero weights. These seven  
 414 descriptors and their primary definitions are shown in Table  
 415 2. Although some of the descriptors are not directly associ-  
 416 ated with specific physicochemical effects, they have been  
 417 found to contain chemical information relevant to the  
 418 interaction mechanisms involved in the anion exchange  
 419 system. As explained below, this can facilitate the under-  
 420 standing of QSRR modeling for protein retention.

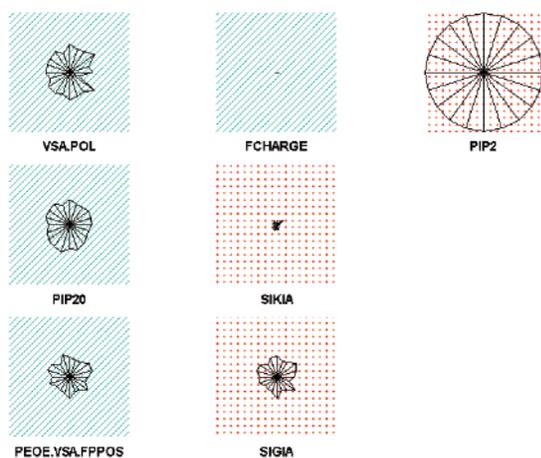
421 During the model construction, one of main tasks is to  
 422 determine the significance of the selected QSRR descriptors  
 423 for later model interpretation. In earlier work, traditional  
 424 QSRR equations made up of linear combinations of physi-  
 425 cally interpretable structural descriptors were employed to  
 426 elucidate the relative importance of several molecular mech-  
 427 anisms involved in chromatographic processes.<sup>47</sup>

428 In contrast to earlier techniques that often used descriptor  
 429 weights within single models for chemical interpretation, a  
 430 graphic visualization tool known as "star plots"<sup>48</sup> was used  
 431 in the current work to characterize the relative importance  
 432 of the seven selected descriptors across the multiple models  
 433 present in the bootstrap aggregate. In most multivariate  
 434 visualization applications, star plots are generated in a multi-  
 435 plot format where each plot represents one case, and each  
 436 radial line represents the magnitude of a particular variable  
 437 (or column) in the data matrix. When the endpoints of the  
 438 rays are connected together with a line, the resulting figure  
 439 resembles a "star". In the current work, each star corresponds  
 440 to a single selected relevant descriptor, where the radius of  
 441 each spoke is the weight of that descriptor in one of the  
 442 sparse SVR models used in the bootstrap (normalized by  
 443 the maximum magnitude of the weights of all descriptors in

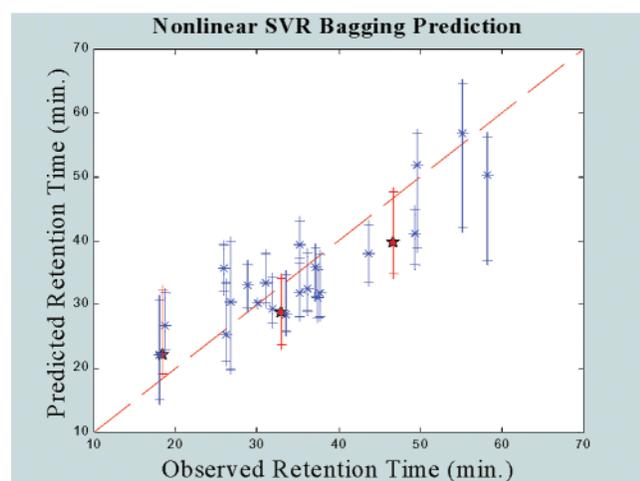
**Linear SVR model ensemble for relevant descriptors ( $x_1, x_2 \dots x_n$ )****Figure 7.** Star plot generation process.

the same bootstrap). This technique visually represents the  
 444 relative importance of each descriptor in each of the predictor  
 445 models used in the bootstrap aggregate and provides a  
 446 measure of the consistent importance of the descriptor over  
 447 all of the bootstrap models. For each descriptor, the sum or  
 448 average of all 20 radii (or the surface area of the star) can  
 449 be used to represent the overall relative importance of the  
 450 descriptor over all 20 bootstraps. The descriptor weights from  
 451 all 20 of the linear SVR models used in the bootstrap  
 452 aggregation procedure are mapped onto the star plots in the  
 453 manner shown in Figure 7. In the example shown in that  
 454 figure, descriptor 1 is consistently important to all models,  
 455 while descriptor 2 has less uniform significance. 456

457 Since descriptor contributions may be either positive or  
 458 negative, background color is used indicate the consistent  
 459 sign of the weight across all bootstraps. Finally, the descrip-  
 460 tors may be ranked over all 20 bootstrap iterations, such that  
 461 the most significant negatively weighted descriptor appears  
 462 in the upper left of the graphic, while the most positively  
 463 weighed descriptor appears in the lower right-hand side of  
 464 the figure. The ordering is performed in a column-wise  
 465 fashion. For instance, in Figure 8, the star plot graph shows  
 466 seven stars representing the weights of the seven selected  
 467 descriptors over 20 bootstraps. As shown in the figure,  
 468 PEOE.VSA.FPPOS has the largest negative effect on reten-  
 469 tion time and PIP2 has the largest positive effect on retention  
 470 time. This kind of graphical approach offers a direct way to 470



**Figure 8.** Star plots for the seven descriptors selected by the feature selection algorithm. Descriptor starplots with a cyan background have negative contributions to the retention time, while those with a red dot background have a positive effect on retention.



**Figure 9.** The prediction scatter plot using the nonlinear SVR model with seven selected descriptors.

471 examine the relative significance of molecular property  
472 descriptors in a semiquantitative manner.

473 The scatter plot for nonlinear SVR prediction based on  
474 these seven descriptors with 12 bootstraps is shown in Figure  
475 9. In this case, the cross-validated  $R_{CV}^2 = 0.882$  and the test  
476 set  $R_{bag}^2 = 0.988$ . It may be observed that the final nonlinear  
477 model performs better with only seven features than with  
478 the original 243 descriptors. The reduction in features also  
479 simplifies the model and allows for better interpretation.  
480 While the predictive accuracy of the model is subject to  
481 improvement, the technique is clearly capable of providing  
482 useful estimates of retention time that should prove useful  
483 in chromatography planning.

484 **Model Interpretation.** Besides the development of direct  
485 prediction models, one of main challenges in QSRR lies in  
486 extracting chemical meaning from the descriptor patterns  
487 found in the models. It is hypothesized that the application  
488 of a fundamental physicochemical modeling approach such  
489 as that used in this investigation will aid in the understanding  
490 of the interaction mechanisms of ion-exchange systems. The  
491 development of predictive models and a greater level of  
492 understanding of the underlying processes of protein chro-  
493 matography will be valuable for future experimental design.

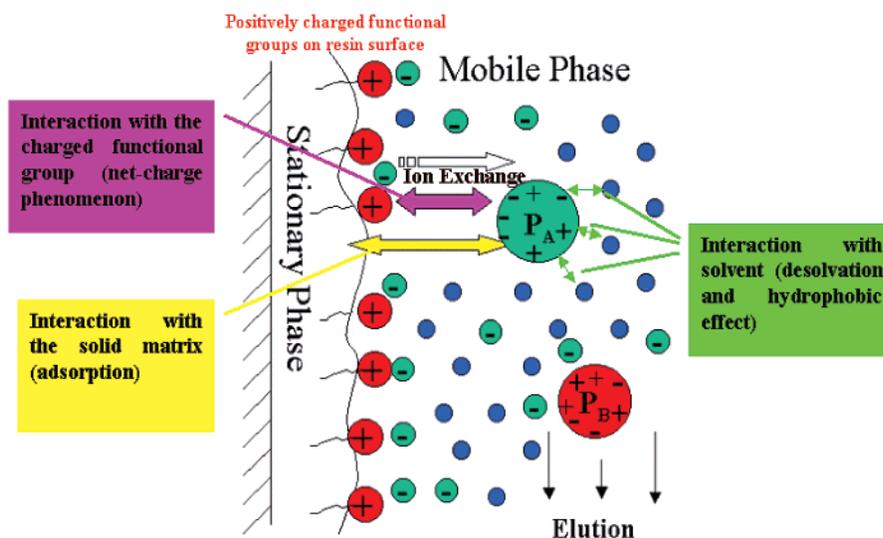
494 Despite its widespread application, the exact mechanism  
495 of protein retention in ion-exchange chromatography is still  
496 controversial. Protein retention on an anion-exchange resin  
497 is controlled by the balance of interactions between the  
498 protein and a set of charged functional groups as well as by  
499 the characteristics of the surrounding medium and the  
500 stationary phase solid matrix. It is known that this kind of  
501 mixed-mode separation mechanism within ion-exchange  
502 columns can offer unique selectivity for the separation of  
503 proteins. One such scenario is depicted in Figure 10.

504 Due to the nature of the ion-exchange processes, it is  
505 expected that electrostatic effects will play a dominant role  
506 in protein retention. This dominant electrostatic effect arises  
507 because the acidic amino acid side chains, i.e., aspartate and  
508 glutamate, are partially deprotonated under the experimental  
509 conditions in which the mobile phase is buffered at pH =  
510 7.4 and produce negative charges on the periphery of the  
511 protein. Since anion exchange sites (quaternary ammonium  
512 functional groups  $N(CH_3)_3^+$ ) on the resin surface are  
513 completely ionized under these conditions, proteins with high  
514 negative charge densities on their surfaces will show greater  
515 affinity for these sites and will elute later. Proteins with low  
516 negative charge densities will interact more weakly with the  
517 resin and will elute first. As described later, additional effects  
518 are also present that can influence elution selectivity,  
519 including overall charge asymmetry and other factors.

520 According to the SVM modeling results from this work,  
521 a dominant set of electrostatic interactions may be proposed  
522 to explain the protein retention behavior using three main  
523 factors: net charge, polarity/polarizability (charge asym-  
524 metry) and a desolvation penalty.

525 The first of the three electrostatic factors is represented  
526 by a fractional surface area descriptor and atomic formal  
527 charges. The MOE descriptor PEOE.VSA.FPPOS represents  
528 the fraction of the molecular surface area bearing a positive  
529 partial charge, as calculated by the PEOE (Partial Equaliza-  
530 tion of Orbital Electronegativities) approach. As shown in  
531 Figure 8, this descriptor bears a negative weight, meaning  
532 that greater fractional positive surface area decreases the  
533 protein retention time. This result is consistent with the net  
534 charge hypothesis, which suggests that fixed positively  
535 charged sites in the resin will exhibit a favorable affinity  
536 for negatively charged amino acids and repel positively  
537 charged regions of the protein surface. The descriptor  
538 FCHARGE represents the total formal charge of the protein,  
539 which is negative in cases where the solution pH is higher  
540 than their isoelectric point (PI). The small negative value of  
541 its weight in the sparse SVR models is consistent with the  
542 explanation that proteins with more negative charge have a  
543 tendency to interact more strongly with the positively charged  
544 function groups present on the surface of the resin. The  
545 apparent insignificance of this seemingly important descriptor  
546 is due to the fact that the electrostatic effect is better  
547 represented more by other selected electrostatic-related  
548 descriptors. The importance of this descriptor may prove to  
549 be more significant in data sets involving proteins with  
550 diverse PI.

551 Although the above net charge model has been frequently  
552 used to explain the phenomenon, retention mapping studies  
553 on the strong ion-exchange columns showed it to be  
554 inadequate. The influence of intramolecular charge asym-  
555 metry in the proteins has been successfully employed as an



**Figure 10.** A simple cartoon illustration of multi-mode interaction involved in protein retention. The symbols  $P_A$  and  $P_B$  represent two proteins with different binding affinities to the stationary phase.

556 alternative explanation for deviations from the net charge  
 557 model together with the fact that protein tertiary structure is  
 558 know to affect retention.<sup>49</sup> Recent studies suggested that  
 559 protein local dipolarity should also be taken into consider-  
 560 ation, since it appears that only a fraction of locally charged  
 561 protein surfaces interact with the stationary phase.<sup>50</sup> These  
 562 regions of localized charge are postulated to orient the protein  
 563 with respect to the oppositely charged ion-exchange support.  
 564 It is clear that in large, complex macromolecules, the  
 565 distribution of charged groups may not be uniform through-  
 566 out the structure. As a result of this inhomogeneity, even  
 567 proteins with zero net charge may exhibit significant  
 568 electrostatic fields. Consequently, the retention behavior will  
 569 depend not only on the net charge itself but also on the spatial  
 570 distribution of charge throughout the protein structure. Other  
 571 effects include the potential for reorganizing these dipoles  
 572 (and higher multipoles) in response to an applied electronic  
 573 field originating from the neighboring medium. Descriptors  
 574 associated with dipolarity and polarizability effects are  
 575 expected to account for differences among the solutes as to  
 576 their propensity to participate in dipole–dipole, dipole–  
 577 induced dipole and charge-transfer interactions.

578 Several TAE electron density-based descriptors listed in  
 579 Table 2 were found to be significant to retention, e.g. SIKIA,  
 580 SIGIA and PIP. These descriptors have also been found to  
 581 correlate with molecular properties such as acid/base strength  
 582 and polarity as well as polarizability.<sup>10</sup> The PIP descriptor  
 583 family can be associated with regions of donor and acceptor  
 584 capabilities that relate to the tendency of analytes to take  
 585 part in charge-transfer interactions. Prior to feature selection,  
 586 there were twenty PIP descriptors present in the descriptor  
 587 set, where PIP1 and PIP2 represent regions of the molecular  
 588 surface where electron density is easily ionizable, while  
 589 PIP20 is associated with regions of tightly held electron  
 590 density, such as on exchangeable protons. SIGIA and SIKIA  
 591 describe the integrals of G and K electronic kinetic energy  
 592 densities found on the molecular van der Waals surface.  
 593 These descriptors are related to the Laplacian of the density  
 594 and are associated with the presence and the strength of  
 595 Lewis basic sites. It has been shown in this work that these  
 596 dipolarity/polarizability-related descriptors (PIP2, SIKIA and

SIGIA) correlate with increased retention time. This may  
 597 be due to their representation of increased dipole/induced-  
 598 dipole or charge/induced-dipole forces between the protein  
 599 and the strong ion-exchanger groups as well as induced-  
 600 dipole/induced-dipole interactions between the polarizable  
 601 aromatic groups of the stationary phase and polarizable  
 602 regions of the protein. The PIP20 descriptor was found to  
 603 be anticorrelated with retention time, indicating that the  
 604 presence of nonacidic hydrogen bond donors (serine, etc.)  
 605 increases solute/mobile-phase interactions at the expense of  
 606 solute/stationary phase interactions.  
 607

608 In addition to the charge characteristics of the protein and  
 609 resin surface as well as the underlying matrix, the nature of  
 610 the solvent, e.g. polarity, is also known to be an important  
 611 contributor to protein retention. In the current model, this  
 612 effect is described by the MOE descriptor VSA.POL, which  
 613 approximates the VDW surface area of polar atoms (both  
 614 hydrogen bond donors and acceptors). The importance of  
 615 this descriptor implies that the hydrogen bonding capacity  
 616 of proteins may also be involved in the intermolecular  
 617 interactions responsible for column retention behavior. To a  
 618 first approximation, most charged and polar groups on the  
 619 solvated protein can interact favorably with the surrounding  
 620 water before attaching to the support surface. Thus, even a  
 621 protein with a moderate polarity has to pay an energetic  
 622 penalty which increases in proportion to the overall polar  
 623 surface of the protein. In this way a protein with more polar  
 624 atoms on the exposed van der Waals surface will have a  
 625 stronger hydrogen-bonding capacity with the mobile phase  
 626 and will elute out of the column first, accounting for the  
 627 negative effect of VSA.POL for retention shown in the star  
 628 plot. Fortunately, this kind of desolvation penalty can be  
 629 offset, although never completely overcome, by more favor-  
 630 able electrostatic interactions between the resin and the  
 631 protein that lead to increased protein retention.

## V. CONCLUSIONS

633 In this study, Support Vector Machine (SVM) regression  
 634 was introduced as a method for generating predictive QSRR  
 635 models of protein retention times in anion exchange chro-

matographic systems. It was demonstrated that models developed using this technique encompass a wide range of proteins including a variety of sizes, shapes, functionalities and selectivity for these resins. In the future, this method should prove useful for performing comparative QSRR studies under different chromatographic conditions and be important for determining appropriate protein purification conditions.

In summary, the behavior of protein solutes in anion-exchange chromatography conditions may be quantified through the use of traditional and electron density-based molecular property descriptors. Models may be constructed using SVR methods for both feature selection and property prediction. Extensive cross-validation of the modeling results was accomplished using multiple sets of training and validation cases in a bootstrap scheme, the results of which are visualized and interpreted using star plots.

## ACKNOWLEDGMENT

This work was supported in part by NSF grants IIS-9979860 and BES-0079436.

## REFERENCES AND NOTES

- Kaliszan, R. Correlation between Retention Indexes and Connectivity Indexes of Alcohols and Methyl-Esters with Complex Cyclic Structure. *Chromatographia* **1977**, *10*, 529–531.
- Katritzky, A. R.; Petrukhin, R.; Tatham, D.; Basak, S.; Benfenati, E. et al. Interpretation of quantitative structure–property and -activity relationships. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 679–685.
- Montana, M. P.; Pappano, N. B.; Debattista, N. B.; Raba, J.; Luco, J. M. High-performance liquid chromatography of chalcones: Quantitative structure-retention relationships using partial least-squares (PLS) modeling. *Chromatographia* **2000**, *51*, 727–735.
- Sutter, J. M.; Peterson, T. A.; Jurs, P. C. Prediction of gas chromatographic retention indices of alkylbenzenes. *Anal. Chim. Acta* **1997**, *342*, 113–122.
- Loukas, Y. L. Artificial neural networks in liquid chromatography: efficient and improved quantitative structure-retention relationship models. *J. Chromatogr. A* **2000**, *904*, 119–129.
- Mazza, C. B.; Sukumar, N.; Breneman, C. M.; Cramer, S. M. Prediction of protein retention in ion-exchange systems using molecular descriptors obtained from crystal structure. *Anal. Chem.* **2001**, *73*, 5457–5461.
- Vapnik, V. N. An overview of statistical learning theory. *IEEE Trans. Neural Networks* **1999**, *10*, 988–999.
- Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
- SYBYL 6.5. Tripos Associate Inc. 1699 S. Hanley Rd., Suite 303, St. Louis, MO 63144-2913.
- Breneman, C. M.; Thompson, T. R.; Rhem, M.; Dung, M. Electron-Density Modeling of Large Systems Using the Transferable Atom Equivalent Method. *Comput. Chem.* **1995**, *19*, 161–179.
- Karelson, M.; Lobanov, V. S.; Katritzky, A. R. Quantum-chemical descriptors in QSAR/QSPR studies. *Chem. Rev.* **1996**, *96*, 1027–1043.
- Bader, R. F. W. *Atoms in Molecules: A Quantum Theory*; Oxford University Press: Oxford, UK, 1994.
- Matta, C. F. Theoretical reconstruction of the electron density of large molecules from fragments determined as proper open quantum systems: The properties of the oripavine PEO, enkephalins, and morphine. *J. Phys. Chem. A* **2001**, *105*, 11088–11101.
- Breneman, C. M. Transferable Atom Equivalents. Molecular Electrostatic Potentials from the Electric Multipoles of PROAIMS Atomic Basins. *The Application of Charge Density Research to Chemistry and Drug Design*; Plenum Press: 1991; pp 357–358.
- Breneman, C. M.; Rhem, M. QSPR analysis of HPLC column capacity factors for a set of high-energy materials using electronic van der Waals surface property descriptors computed by transferable atom equivalent method. *J. Comput. Chem.* **1997**, *18*, 182–197.
- Labute, P. A widely applicable set of descriptors. *J. Mol. Graph. Model.* **2000**, *18*, 464–477.
- Vapnik, V. N. *The Nature of Statistical Learning Theory*; Springer, Berlin, 1995.
- Cortes, C.; Vapnik, V. Support-Vector Networks. *Machine Learning* **1995**, *20*, 273–297.
- Zhang, L.; Zhou, W. D.; Jiao, L. C. Support vector machine for 1-D image recognition. *J. Infrared Millimeter Waves* **2002**, *21*, 119–123.
- Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S. Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Comput. Chem.* **2001**, *26*, 5–14.
- Ding, C. H. Q.; Dubchak, I. Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics* **2001**, *17*, 349–358.
- Karchin, R.; Karplus, K.; Haussler, D. Classifying G-protein coupled receptors with support vector machines. *Bioinformatics* **2002**, *18*, 147–159.
- Cai, Y. D.; Liu, X. J.; Xu, X. B.; Chou, K. C. Support vector machines for predicting HIV protease cleavage sites in protein. *J. Comput. Chem.* **2002**, *23*, 267–274.
- Czerminski, R.; Yasri, A.; Hartsough, D. Use of Support Vector Machine in pattern classification: Application to QSAR studies. *Quant. Struct.-Act. Relat.* **2001**, *20*, 227–240.
- Vapnik, V. N. *Estimation of Dependences Based on Empirical Data*; Springer-Verlag: Berlin, 1982.
- Alex, J.; Smola, B. S. Linear Programs for Automatic Accuracy Control in Regression. *Proceedings ICANN'99, Int. Conf. on Artificial Neural Networks*; Springer: Berlin, 1999.
- Scholkopf, B.; Smola, A. J.; Williamson, R. C.; Bartlett, P. L. New support vector algorithms. *Neural Comput.* **2000**, *12*, 1207–1245.
- Whitley, D. C.; Ford, M. G.; Livingstone, D. J. Unsupervised forward selection: A method for eliminating redundant variables. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1160–1168.
- Sutter, J. M.; Dixon, S. L.; Jurs, P. C. Automated Descriptor Selection for Quantitative Structure–Activity-Relationships Using Generalized Simulated Annealing. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 77–84.
- Rogers, D.; Hopfinger, A. J. Application of Genetic Function Approximation to Quantitative Structure–Activity-Relationships and Quantitative Structure–Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854–866.
- So, S. S.; Karplus, M. Genetic neural networks for quantitative structure–activity relationships: Improvements and application of benzodiazepine affinity for benzodiazepine/GABA(A) receptors. *J. Med. Chem.* **1996**, *39*, 5246–5256.
- Zheng, W. F.; Tropsha, A. Novel variable selection quantitative structure–property relationship approach based on the k-nearest-neighbor principle. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 185–194.
- Luke, B. T. Evolutionary Programming Applied to the Development of Quantitative Structure–Activity-Relationships and Quantitative Structure–Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1279–1287.
- Kubinyi, H. Variable Selection in Qsar Studies .1. An Evolutionary Algorithm. *Quant. Struct.-Act. Relat.* **1994**, *13*, 285–294.
- Izrailev, S.; Agrafiotis, D. A novel method for building regression tree models for QSAR based on artificial ant colony systems. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 176–180.
- Izrailev, S.; Agrafiotis, D. K. Variable selection for QSAR by artificial ant colony systems. *SAR QSAR Environ. Res.* **2002**, *13*, 417–423.
- Agrafiotis, D. K.; Cedeno, W. Feature selection for structure–activity correlation using binary particle swarms. *J. Med. Chem.* **2002**, *45*, 1098–1107.
- Bennett, K.; Bi, J.; Embrechts M.; Breneman, C.; Song, M. Dimensionality Reduction via Sparse Support Vector Machines. *J. Machine Learning Research 2002 (Special Issue on Feature Selection)* (In press).
- Dimitris K. Agrafiotis, W. C.; Victor S. On the use of Neural Network Ensembles in QSAR and QSPR. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 903–911.
- Breiman, L. Bagging predictors. *Machine Learning* **1996**, *24*, 123–140.
- Breiman, L. Using iterated bagging to debias regressions. *Machine Learning* **2001**, *45*, 261–277.
- Efron, B.; Tibshirani, R. J. *An introduction to the bootstrap*; Chapman and Hall: New York, 1993.
- Demiriz A. B. K.; Breneman C.; Embrechts, M. Support Vector Machine Regression in Chemometrics. *33<sup>rd</sup> Symposium on Computing Science and Statistics: Proceedings of Interface*, June, 2001.
- Using the CPLEX(TM) Linear Optimizer and CPLEX(TM) Mixed Integer Optimizer (version 2.0). CPLEX optimization Inc., Incline Village, Nevada.
- S-PLUS 2000; Data Analysis Products Division, Mathsoft Inc., Seattle, Washington, 98109.

## ANION-EXCHANGE CHROMATOGRAPHY SYSTEMS

PAGE EST: 10.1 J. Chem. Inf. Comput. Sci. K

- 786 (46) Gasteiger, J.; Marsili, M. Iterative Partial Equalization of Orbital  
787 Electronegativity – a Rapid Access to Atomic Charges. *Tetrahedron*  
788 **1980**, *36*, 3219–3228. 794
- 789 (47) Kaliszan, R. Quantitative Structure-Retention Relationships Applied  
790 to Reversed-Phase High-Performance Liquid-Chromatography. *J.*  
791 *Chromatogr. A* **1993**, *656*, 417–435. 795
- 792 (48) Chambers, J.; Cleveland, W.; Kleiner, B.; Tukey, P. *Graphical Methods*  
793 *for Data-Analysis*; Wadsworth, 1983. 796
- (49) Kopaciewicz, W.; Rounds, M. A.; Fausnaugh, J.; Regnier, F. E.  
Retention Model for High-Performance Ion-Exchange Chromatogra-  
phy. *J. Chromatogr.* **1983**, *266*, 3-21. 797
- (50) Cohen, B. E.; McAnaney, T. B.; Park, E. S.; Jan, Y. N.; Boxer, S. G.  
et al. Probing protein electrostatics with a synthetic fluorescent amino  
acid. *Science* **2002**, *296*, 1700–1703. 798  
799
- CI025580T 800