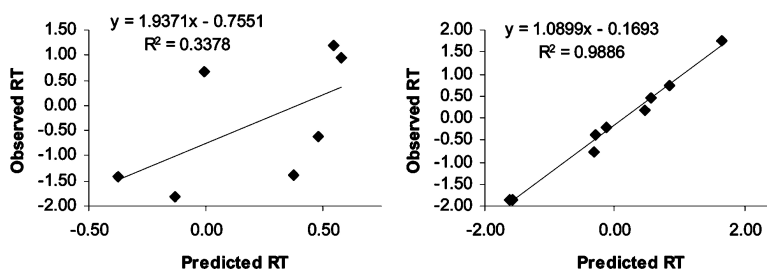


## Prediction of HPLC Conditions Using QSPR Techniques: an Effective Tool to Improve Combinatorial Library Design

Sabine Schefzick, Chris Kibbey, and Mary P. Bradley

*J. Comb. Chem.*, **2004**, 6 (6), 916-927 • DOI: 10.1021/cc049914y • Publication Date (Web): 07 October 2004

Downloaded from <http://pubs.acs.org> on March 20, 2009



### More About This Article

Additional resources and features associated with this article are available within the HTML version:

- Supporting Information
- Links to the 1 articles that cite this article, as of the time of this article download
- Access to high resolution figures
- Links to articles and content related to this article
- Copyright permission to reproduce figures and/or text from this article

[View the Full Text HTML](#)

# Prediction of HPLC Conditions Using QSPR Techniques: an Effective Tool to Improve Combinatorial Library Design

Sabine Schefzick,\* Chris Kibbey, and Mary P. Bradley

*Pfizer Global Research and Development, Discovery Technologies, Ann Arbor Laboratories,  
2800 Plymouth Road, Ann Arbor, Michigan 48105*

*Received May 3, 2004*

The purification and characterization of compounds resulting from parallel synthesis or combinatorial chemistry has not yet been optimized to operate as a completely automated high-throughput process. Liquid chromatography/mass spectroscopy (LC/MS) is most commonly employed to carry out the characterization and identification of combinatorial compounds. This desired level of automation can only be accomplished if the separation conditions for every compound in the combinatorial array are known prior to the analysis. This study presents a quantitative structure retention relationship (QSRR) approach to predict the retention time of structurally diverse solutes under 75 different LC/MS conditions. Sixty-two compounds were analyzed using 15 commonly used HPLC columns under 5 different gradient conditions. The solute retention time was used as the dependent variable, and more than 1000 molecular descriptors were calculated for this compound set to generate QSRR models. After the elimination of highly correlated variables and those with zero variance, two different genetic algorithms were applied to identify the most significant descriptors. Following the variable selection, the identified descriptors were used to create QSRR models for each separation condition. The calculated stepwise multiple linear regression models have been proven to be statistically significant and highly predictive, with an average coefficient of determination ( $R^2$ ) of 0.86, an average cross-validated  $r^2$  of 0.62,  $r^2 = 0.76$ , and an average  $F$  value of 27.29. The QSRR models can be used to design “analysis-friendly” library purification plates, in which compounds are arranged on the basis of their predicted separation condition and can also be used during the library design phase to flag compounds not amenable to the separation methods in use.

## Introduction

With the advent of combinatorial chemistry, high-throughput synthesis methods have made it possible to synthesize multiple compounds in parallel. Hence, follow-up analysis methods, such as purification and characterization, have been challenged to increase throughput to meet the demands of combinatorial chemistry. To achieve the desired degree of automation in the purification and characterization of libraries, the liquid chromatography/mass spectroscopy (LC/MS) conditions must be known prior to the analysis. This involves significant effort in method development to ensure the optimal conditions for each compound. This requirement is impractical to achieve, because there are typically a limited number of methods in routine use at any given time. A reasonable substitute would be to estimate which of the available methods would give the best result for each compound in the library. Moreover, this prediction could be performed on virtual compounds, thereby facilitating the plating of compounds for purification and analysis.

We applied a quantitative structure–retention relationship (QSRR) analysis to relate the retention time of each of the 62 different compounds under 75 specific LC/MS separation conditions to structural features of the compound using 1381

calculated molecular descriptors. Because the pool of available descriptors was large, genetic algorithms (GA) were used to select the most relevant descriptors for further analysis.

## Background

The developments of combinatorial chemistry and high-throughput screening have enabled the synthesis and screening of a greater number of new chemical entities (NCEs) than would have been possible by traditional techniques. However, these technologies are not sufficiently mature to allow the synthesis and screening of the googol ( $10^{100}$ ) of compounds<sup>1</sup> that is estimated to exist in the virtual chemistry space. Consequently, it is imperative to limit the chemical space through efficient library design. Lipinski’s “Rule of five”<sup>2</sup> has been used to identify compounds with a high risk of poor bioavailability on the basis of molecular mass, lipophilicity, the number of hydrogen bond donors, and acceptors. Lipinski’s “Rule of five” in combination with ADME/T (adsorption, distribution, metabolism, excretion and toxicity) filters can increase the chance to identify “drug-like” bioactive compounds. Hence, a lot of time is invested to select the most appropriate compound library for a particular therapeutic area project. These compounds are then enumerated and synthesized as combinatorial array(s). The synthesized products are submitted for purification by high performance liquid chromatography (HPLC), and the identi-

\* To whom correspondence should be addressed. Phone: (734) 622-2357. Fax: (734) 622-2782. E-mail: sabine.schefzick@pfizer.com.

ties of the final products are confirmed by analytical LC/MS. Unfortunately, not all of the carefully designed compounds will end up in the corporate compound database. In practice, it is common to find attrition rates of 50–70% within a combinatorial array due to a combination of failed synthesis and loss of material during the purification and characterization steps.

RP-HPLC, or reversed-phase high performance liquid chromatography, is the most widely used purification and separation technique in pharmaceutical companies. One major advantage of HPLC is the possibility to analyze solutes with a wide range of polarity in an automated process. However, the process can only be completely automated if the correct separation conditions are known prior to the separation. The method development, including the selection of HPLC stationary phase, mobile phase, gradient, etc., is time-consuming and would be impractical to perform for all compounds in a library. Hence, the traditional HPLC paradigm needs to be adapted such that (a) many samples can be processed in the shortest possible time, (b) there is no interim method development, and (c) reequilibrations of the HPLC system are kept to a minimum. The current approach for purification and characterization to identify separation conditions is driven by the experience of the experimentalist. A tool that could be used to rapidly predict the HPLC method on the basis of structural information would serve to guide the compound flow during the purification process and allow for greater optimization, leading to higher throughput and success rates. Furthermore, it is possible to use this information during the library design process so that compounds that would be more likely to experience difficulties during purification would obtain a lower priority.

Several approaches have been published to predict the solute retention behavior on a selected column. Every effort is based on the linear solvation energy relationship (LSER),<sup>3–5</sup> also known as solvatochromic equation, which relates the reactivity parameter with solvent–solute interactions on the basis of physicochemical properties.

$$\log k' = c + rR_2 + s\pi_2^H + \alpha \sum \alpha_2^H + \beta \sum \beta_2 + \nu V_x \quad (1)$$

Equation 1 defines the relationship between the capacity factor,  $k'$ , and solute descriptors, where  $R_2$  is excess molar refraction,  $\pi_2^H$  is the solute polarizability/dipolarity,  $\sum \alpha_2^H$  and  $\sum \beta_2$  are the solute hydrogen-bond acidity and basicity, and  $V_x$  is the solute volume. The constants  $c$ ,  $r$ ,  $s$ ,  $a$ ,  $b$ , and  $\nu$  are specific for the system condition employed. This approach describes contributions of individual intermolecular interactions that are responsible for the partition behavior of neutral molecules in octanol–water or reversed-phase separation systems. The solute properties are empirical descriptors, which are only available for about 4000 compounds. Hence, this approach is limited to several thousand compounds and, therefore, not feasible for our study.

Baczek and Kaliszan<sup>6–16</sup> have demonstrated the prediction of solute retention under a given set of linear, reversed-phase gradient HPLC conditions by generating a quantitative structure–retention relationship (QSRR) model. Specifically,

they have shown that the retention time of a solute under gradient reversed-phase HPLC can be determined from the following equation

$$t_R = k_1 + k_2\mu + k_3\delta_{\min} + k_4A_{\text{was}} \quad (2)$$

where  $\mu$  is the total dipole moment,  $\delta_{\min}$  is the electron excess charge of the most negatively charged atom, and  $A_{\text{was}}$  is the water-accessible molecular surface area of the solute. The constants  $k_1$ ,  $k_2$ ,  $k_3$ , and  $k_4$  are related to the specific stationary phase and mobile phase gradient employed in the separation.

Another simple quantitative structure retention time model correlates the retention behavior with the logarithm of *n*-octanol/water partition coefficient  $\log P$ , which can be calculated with a variety of computer programs with various error margins.

$$\text{retention parameter} = k_1 + k_2 \log P \quad (3)$$

Recently, Kaliszan presented a study that compared the latter two approaches for predicting gradient retention.<sup>16</sup> In this study, eqs 1 and 2 revealed statistically significant QSRR models; however, Kaliszan also showed that the predictive power of these models is rather limited. Thus, Kaliszan concludes “... a suitable translation which would reveal the properties encoded into the structure in a reliable manner is still lacking”.

Another recent approach uses the response (retention factor  $\log k_w$ ) to build a decision tree based on 266 molecular descriptors (0D, 1D, and 2D) to predict the retention time under isocratic conditions.<sup>17</sup> Despite the fact that a statistically relevant model with good predictive power is achieved, the author feels that a “... more diverse set of substances with more diverse retention times ...” might be needed to predict the chromatographic behavior. Therefore, the goal of our study is to identify a standard set of structurally diverse compounds along with the most suitable molecular descriptors to predict the retention times of these solutes under 75 different HPLC conditions.

Several commercially available software programs<sup>17–23</sup> are available either to optimize the HPLC separation conditions or to predict HPLC retention time; however, none of these programs was developed for rapid analysis of combinatorial libraries. Therefore, the predicted retention times are typically > 5 min.

We employed a QSRR approach to generate statistical models that are used to predict a set of reversed-phase gradient HPLC conditions best suited for the characterization of combinatorial compounds. A database containing all 75 models is used to determine the predicted retention time of every compound (using the solute’s QSRR descriptors) in the library under each of the chromatographic conditions.

The calculated retentions of the solute under the analytical chromatographic conditions are assigned to one of three bins:

$$t_R < 1.5 \quad \text{not retained}$$

$$1.5 < t_R < 4.5 \quad \text{moderately retained}$$

$$t_R > 4.5 \quad \text{highly retained}$$

**Table 1.** List of HPLC Columns Frequently Used to Characterize Compounds

	dimension	particle size	vendor
YMC Pro C <sub>18</sub>	4.6 × 50 nm	3	Waters Corp. <sup>a</sup>
YMC Pack Phenyl	4.6 × 50 nm	3	Waters Corp. <sup>a</sup>
Aquasil C <sub>18</sub>	4.6 × 50 nm	3	Thermo Electron Corp. <sup>b</sup>
YMC Pack ODS AQ C <sub>18</sub>	4.6 × 50 nm	3	Waters Corp. <sup>a</sup>
Res.Sys. Hydropore C <sub>18</sub>	4.6 × 50 nm	3	Resolution Systems <sup>c</sup>
MetaChem Polaris C <sub>18</sub>	4.6 × 50 nm	3	MetaChem <sup>d</sup>
Xterra MS C <sub>8</sub>	4.6 × 50 nm	3.5	Waters Corp. <sup>a</sup>
Xterra MS C <sub>18</sub>	4.6 × 50 nm	3.5	Waters Corp. <sup>a</sup>
Waters Symmetry C <sub>18</sub>	4.6 × 50 nm	3.5	Waters Corp. <sup>a</sup>
Alltima C <sub>18</sub>	4.6 × 50 nm	3	Alltech Associates, Inc. <sup>e</sup>
Polymer Laboratories PLRP S	4.6 × 50 nm	5	Polymer Laboratories Ltd. <sup>f</sup>
Phenomenex Prodigy Phenyl	4.6 × 50 nm	3	Phenomenex Inc. <sup>g</sup>
Phenomenex SYNERGI MAX RP C <sub>12</sub>	4.6 × 50 nm	4	Phenomenex Inc. <sup>g</sup>
Phenomenex SYNERGI Polar RP	4.6 × 50 nm	4	Phenomenex Inc. <sup>g</sup>
Phenomenex Luna C <sub>8</sub> (2)	4.6 × 50 nm	3	Phenomenex Inc. <sup>g</sup>

<sup>a</sup> See ref 24. <sup>b</sup> See ref 25. <sup>c</sup> See ref 26. <sup>d</sup> See ref 27. <sup>e</sup> See ref 28. <sup>f</sup> See ref 29. <sup>g</sup> See ref 30.

**Table 2.** Physicochemical Properties of the HPLC Columns Used in This Study

	surface area (m <sup>2</sup> /g)	pore size (Å)	pore volume (mL/g)	C content (%)	bonded phase coverage (mmol/m <sup>2</sup> )
YMC Pro C <sub>18</sub>	335	120	1.06	16	2.5
YMC Pack Phenyl	300	120	1.0	9	3.2
Aquasil C <sub>18</sub>	310	100	0.9	12	1.8
YMC Pack ODS AQ C <sub>18</sub>	300	120	1.0	14	2.2
Res.Sys. Hydropore C <sub>18</sub>	300	120	1.0	15	NA
MetaChem Polaris C <sub>18</sub>	200	200	1.0	NA	NA
Xterra MS C <sub>8</sub>	175	125	0.7	12	2.3
Xterra MS C <sub>18</sub>	175	125	0.7	15.5	2.2
Waters Symmetry C <sub>18</sub>	340	100	0.9	19.1	3.2
Alltima C <sub>18</sub>	340	100	NA	16	NA
Polymer Laboratories PLRP S	NA	NA	NA	NA	NA
Phenomenex Prodigy Phenyl	450	100	1.06	10	NA
Phenomenex SYNERGI MAX RP C <sub>12</sub>	475	80	1.05	15	NA
Phenomenex SYNERGI Polar RP	475	80	1.05	11	NA
Phenomenex Luna C <sub>8</sub> (2)	400	100	NA	13.5	5.5

A chromatographic method that yields moderate retention for a solute receives a score of 1 for that solute. Otherwise, the chromatographic method receives a score of 0 for that solute. The chromatographic method with the highest score for all solutes in the virtual library is selected as the recommended analytical method for the library. A similar approach is applied to recommend a preparative chromatographic method for library purification. In this case, analytical HPLC columns with the same stationary phase and lengths as the preparative HPLC columns are used to obtain the retention time information. The retention time of the solute under preparative HPLC condition is proportional to the retention time of the solute under analytical HPLC conditions and the flow rates of the mobile phase and indirectly proportional to the square of the column diameters.

However, in this study we will focus our efforts on predicting the solute's retention behavior exclusively for analytical HPLC conditions, which are implemented in the characterization process of combinatorial compounds.

### Experimental Section

Because the selection of the best HPLC conditions for a given compound can be subjective, on the basis of the experience and knowledge of the analytical chemist, we declined to use preexisting analytical data to generate our

QSRR models. Therefore, we selected a standard solute dataset to be analyzed under specific HPLC conditions that would likely be used for high-throughput characterization in our labs. These conditions were chosen in collaboration with the experimentalist.

**HPLC Columns.** Cluster analysis of the chromatographic conditions used in the analytical characterization of combinatorial libraries over the past several years revealed that the majority of the characterization is performed on approximately 15 columns (listed below). These 15 HPLC columns are listed in Table 1 together with some of their physicochemical properties (Table 2). All HPLC columns used in this study were newly purchased.

**HPLC Gradients.** For the purpose of this analysis, we limited the analytical HPLC conditions to five linear gradient programs listed below. Formic acid (1%) was added as a modifier to the aqueous as well as the organic mobile phase. Acetonitrile was chosen as organic mobile phase, since this is the preferred organic phase in-house. Each of these 5 gradients was used in each of the 15 HPLC columns for a total of 75 experiments for every compound in this study.

- 10% CH<sub>3</sub>CN → 100% CH<sub>3</sub>CN, 5 min
- 10% CH<sub>3</sub>CN → 50% CH<sub>3</sub>CN, 5 min
- 50% CH<sub>3</sub>CN → 100% CH<sub>3</sub>CN, 5 min
- 30% CH<sub>3</sub>CN → 70% CH<sub>3</sub>CN, 5 min
- 20% CH<sub>3</sub>CN → 80% CH<sub>3</sub>CN, 5 min

**Equipment.** The chromatographic data were obtained using an Alliance HT Waters 2795 LC/MS apparatus, which is equipped with a pump, a variable-wavelength UV-vis detector (Waters 996 Photodiodes Array Detector), autosampler and thermostat. The mass spectrometer (Micromass ZMD 2000) used for this study was a z-spray mass detector equipped with a single quadrupole mass analyzer and API interface (electrospray & APCI). Data were collected and processed using MassLynx 3.5, which is distributed by Micromass, Ltd.

**Chemicals and Solutes.** ChromAR water; acetonitrile; and formic acid, 88%, AR (ACS) were purchased from Mallinckrodt Laboratory Chemicals.

Of the 62 compounds used in this study, 48 compounds (structures not shown) were obtained from Pfizer's compound database, and 14 compounds (Table 3) were purchased from Sigma-Aldrich. The following test solutes were obtained from Sigma Aldrich: [2*S*-(2*a*,3*b*,8*ab*)]-(+)-hexahydro-3-(hydroxymethyl)-8*a*-methyl-2-phenyl-5*H*-oxazolo[3,2-*a*]pyridin-5-*on*, 4,5-diphenylimidazole, 4-hydroxy-2,5-diphenyl-3-thiophenone 1,1-dioxide, 4-hydroxy-3-( $\alpha$ -iminobenzyl)-1-methyl-6-phenylpyridin-2(1*H*)-*one*, 4-isobutyl- $\alpha$ -methylphenylacetic acid, 6-hydroxy-1,3-benzoxathiol-2-*on*, 7-hydroxy-4-coumarinylacetic acid, 8-methoxypsoralen, furoin, gramine, hydrocortisone, prednisolone, reserpine, and sulfadiazine. The concentrations of the solutes in the standard solution varied from 1.5 to 2.0 mg/mL. The sample solutions contained three or four compounds. The solutes were dissolved in 1:1 v/v % CH<sub>3</sub>CN/H<sub>2</sub>O solution. Since it is well-known that DMSO (dimethyl sulfoxide) influences the peak shape, addition of DMSO was avoided as much as possible.

It was our intent to work with a solute set that is representative of druglike compounds that shows histogram plots of Lipinski's Rule of 5 descriptors as well as the number of rotatable bonds and polar surface area descriptors. All compounds used in this study fell within "druglike" boundaries (MW < 500, ClogP < 5, hydrogen bond acceptors < 10, hydrogen bond donors < 5) with a distribution covering a wide range of chemical space within the Lipinski guidelines. In addition, to further characterize that the diversity of our solute set was representative of chemical space likely to be encountered with a typical combinatorial library, pairwise tanimoto coefficients ( $T_c$ ) were used to analyze the relative molecular diversity of the data set. Tanimoto coefficients and the average  $T_c$  value were calculated using the daylight fingerprints. The calculated averaged tanimoto coefficient for 64 compounds used in the standard solute set used in this study was  $0.19 \pm 0.20$ , indicating a high level of diversity within the data set. For comparison, the MDDR (MDL drug data report) database was used as a reference database. For 81 796 MDDR compounds with a molecular weight between 100 and 600 and rotatable bonds  $\leq 15$ , the pairwise tanimoto coefficient was  $0.29 \pm 0.27$ .

**Determination of the Retention Parameter for the QSRR Studies.** Each solute in the standard set was chromatographed three times on each of the 15 columns under the five gradient conditions. All chromatographic measurements were performed at 20 °C with a mobile phase flow rate of 1 mL/min. The injected sample volume was 20  $\mu$ L.

The average retention times, obtained from these experiments, were used as the dependent variable to generate the QSRR models. The retention times were manually determined from the chromatograms. The peak's shape and its influences (e.g., DMSO) were not considered in this study.

**Molecular Descriptors of Solutes.** Initially, 2419 descriptors were calculated using the geometry-optimized 3-D molecular structures.

MMFF94<sup>31</sup> force field, embedded in SYBYL 6.9, was used to generate the optimized 3D conformation. The same force field was used to compute partial charges. Table 4 summarizes the molecular modeling software and the descriptors generated for this analysis.

## Result and Discussion

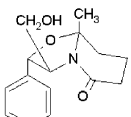
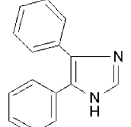
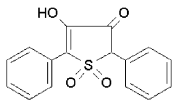
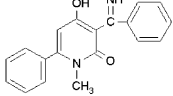
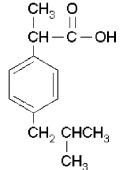
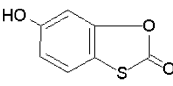
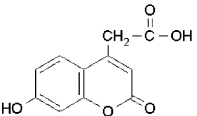
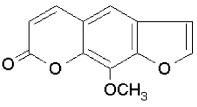
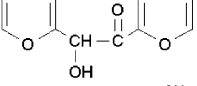
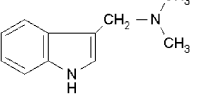
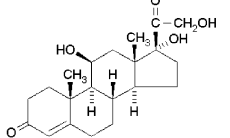
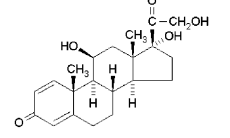
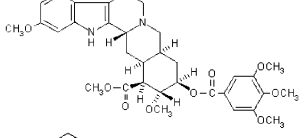
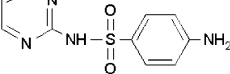
**Genetic Algorithm and Data Analysis.** For each HPLC method, the average retention time of every successfully separated solute was combined with the available set of descriptors. It is important to point out that it was not possible to retrieve retention times for all 62 solutes under all 75 HPLC conditions. Because of variability in early and late elution of solutes for each chromatographic analysis, our datasets include  $\sim 43$  observations for each method used. With that in mind, it was important to perform the collinearity and zero variance checking at this point because the datasets for each method no longer include the identical number (or composition) of observations. After the removal of descriptors with a collinearity >90% and zero variance variables,  $\sim 650$  descriptors remained in the data set. Prior to the application of genetic algorithm as variable selection tool, the data sets were normalized using the BoxCoxAuto technique in Partek,<sup>41</sup> which is a power transformation tool that automatically determines the most feasible normalization algorithm.

Two different genetic algorithms (GAs) were used for further variable selection. The first genetic algorithm utilized a feature selection tool available through Partek. The other genetic algorithm applied was the genetic function approximation available through Cerius2.<sup>42,43</sup> In both cases, 10 000 generations were calculated with a population of 100. The mutation probability was set to 0.05. All evolved linear equations were of a fixed length of 10 variables. For the GFA in Cerius2, the initial equation length was set to 10. Because of the random nature of these algorithms, each GA was applied three times to each dataset. The three best equations per GA were selected, and each descriptor was ranked on the basis of the frequency of selection by the GA (maximum score = 3).

Using JMP,<sup>44</sup> all 150 GA results (two different GAs applied to 75 data tables) were combined using the sum-rank fusion method<sup>45-47</sup> (eq 4). Here,  $R_i(x)$  symbolizes the rank position of the descriptor  $x$  for a specific HPLC method  $i$ , and  $N$  is the number of different HPLC methods. The descriptors with the highest  $\text{SUM}_x$  value were considered the most significant descriptors for predicting the solutes retention.

$$\text{SUM}_x = \sum_{i=1}^N R_i(x) \quad (4)$$

**Table 3.** Structures Obtained from Sigma-Aldrich

Structure	Name	Sigma Aldrich Catalogue No.
	[2S-(2a,3b,8ab)]-(+)-Hexahydro-3-(hydroxymethyl)-8a-methyl-2-phenyl-5H-oxazolo[3,2-a]pyridin-5-one	38,811-4
	4,5-Diphenylimidazole	D20,860-4
	4-Hydroxy-2,5-diphenyl-3-thiophenone 1,1-dioxide	34,357-9
	4-Hydroxy-3-(a-aminobenzyl)-1-methyl-6-phenylpyridin-2(1H)-one	25,033-3
	4-Isobutyl-a-methylphenylacetic acid	28,474-2
	6-Hydroxy-1,3-benzoxathiol-2-on	21,707-7
	7-Hydroxy-4-coumarinylacetic acid	33,566-5
	8-Methoxypsoralen	23,272-6
	Furoin	19,265-1
	Gramine	G1,080-6
	Hydrocortisone	28,609-5
	Prednisolone	28,698-2
	Reserpine	R0875 (Sigma)
	Sulfadiazine	28,719-9

**Table 4.** List of Calculated Descriptors

no. of descriptors	software	description of descriptors
1496	Dragon <sup>32</sup>	2D autocorrelations descriptors, 3D-MoRSE descriptors, BCUT descriptors, GETAWAY descriptors, Galvez topological charge indices, RDF descriptors, Randic molecular profiles, WHIM descriptors, aromaticity indices, atom-centered fragments, charge descriptors, constitutional descriptors, empirical descriptors, functional group counts, geometrical descriptors, molecular walk counts, properties, topological descriptors
199	MOE <sup>33</sup>	physical properties, subdivided surface areas atom counts and bond counts, Kier & Hall connectivity and kappa shape indices, adjacency and distance matrix descriptors, pharmacophore feature descriptors, partial charge descriptors, potential energy descriptors, surface area, volume and shape descriptors, conformation dependent charge descriptors
36	Qikprop <sup>34</sup>	2D, 3D descriptors and pharmaceutically relevant properties of organic molecules, e.g., aqueous solubility (log S), brain/blood partition coefficient (log BB), CNS activity
50	HiVol <sup>35</sup> /Sybyl <sup>36</sup>	2D and 3D descriptors, topology descriptors, dipole moment
401	MolconnZ <sup>37</sup>	molecular connectivity, shape, and information indices
88	Volsurf <sup>38</sup>	2D molecular descriptors derived from 3D molecular interaction energy grid maps
142	TSAR <sup>39</sup>	molecular attributes, number of connectivity, shape, topology, and electrotopology indices, counts of atoms, rings, groups, and H-bond donors and acceptors, electrostatic calculations
17	webpk <sup>40</sup>	in-house program
2419	total	

**Table 5.** List of the 20 Most Frequently Selected Descriptors by Genetic Algorithm

descriptors	description	frequency count	% selected
#amine	number of nonconjugated amine (QikProp)	150	33.33
CLogP	Biobyte's log P (Sybyl6.9)	127	28.22
nNHRPh	number of secondary amines (aromatic) (Dragon)	100	22.22
vsa_other	approximation to the sum of VDW surface areas of atoms typed as "other" (MOE)	52	11.56
Atype_N_68	AlogP N in:Al3N (Cerius2)	49	10.89
DCASA	absolute value of the difference between CASA+ (positive charge weighted surface area, ASA+ times max {qi > 0}) and CASA- (MOE)	47	10.44
C-027	response to R-CH-X, Ghose-Crippen atom centered fragment (Dragon)	45	10.00
donors	HIVol donor (Sybyl6.9)	36	8.00
DASA	absolute value of the difference between ASA+ (water-accessible surface area of all atoms with positive partial charge (strictly >0)) and ASA- (MOE)	35	7.78
QplogKp	predicted skin permeability (QikProp)	35	7.78
FASA+	fractional ASA+ calculated as ASA+/ASA (MOE)	34	7.56
PDsol (mcg/mL)	aqueous solubility (webPK)	34	7.56
Group_count_for_chain_c=n	group_count_for_chain_c=n (TSAR)	33	7.33
nCOOHPh	number of carboxylic acids (aromatic) (Dragon)	33	7.33
H8m	H autocorrelation of lag 8/weighted by atomic masses GETAWAY (dragon)	31	6.89
HATS3u	leverage-weighted autocorrelation of lag3/unweighted GETAWAY (Dragon)	31	6.89
estate_sCH3	estate for CH <sub>3</sub> group (Sybyl 6.9)	30	6.67
Group_count_for_Phenyl	group_count_for_Phenyl (TSAR)	29	6.44
CLogP_error	biobyte's log P (Sybyl6.9)	29	6.44
SlogP_VSA9	sum of v <sub>i</sub> such that L <sub>i</sub> > 0.40 (MOE)	28	6.22

**Significant Descriptors.** Table 5 lists the 20 most frequently selected descriptors. Many of the selected descriptors represent molecular parameters that are known to influence the separation in RP HPLC. For example, it is well-known that the ionization state (neutral or charged) of a compound will affect the retention behavior of the solute. Moreover, under the HPLC conditions used in this study, amines will be charged, whereas the carboxylic group will

be neutral. Therefore, it makes sense that the number of amines (#amine) and aromatic amines (nNHRPh) impact the prediction behavior more than the number of carboxylic acids (nCOOHPh).

Likewise, it is not surprising that ClogP was selected as a significant variable since the partitioning of a compound between liquid aqueous and organic phases is related to the solute's partition equilibrium between mobile and stationary

phase. Particular, the nitrogen in aliphatic substructures is of importance, because the AlogP of nitrogen (:Al<sub>3</sub>N)<sup>48</sup> is ranked as the fourth most important descriptor identified by the GAs. In addition, we observed a relevant contribution of different water-accessible surface descriptors (DCASA, DASA, FASA+), which is in agreement with the results of Baczek et al.,<sup>16</sup> who found that a water-accessible molecular surface area descriptor, calculated in Hyper-Cube, also showed a significant contribution in their QSRR studies.

The retention behavior of solutes was originally described by eq 1. Two of the four terms describe the relationship of the retention time of a solute and the hydrogen bonding capability of the solute. Hydrogen bond properties can also be described by the number of donor atoms in the compound, another descriptor identified by the GAs. The last term of Abraham's equation is the cavity term, which is related to the energy necessary to form a cavity for the solute in the solvent. This term is dependent on the volume of the solute. This dependency is described using GETAWAY (geometry, topology, and atom-weights assembly)<sup>49,50</sup> descriptors, in which two of these descriptors (HATS3u, H8m) significantly contribute to the retention time. Both descriptors reveal information about size and shape of the molecules. (HATS3u and H8m belong to the H-GETAWAY descriptors, in which the molecular influence matrix and especially the diagonal elements of this matrix are used to determine specific size and shape properties.) HATS3u and H8m are spatial auto-correlation descriptors, which summarize contributions of a specific path length (lag) in the molecular graph. Overall, all descriptors selected by the genetic algorithm can be physically explained, which is somewhat surprising, considering the large number of descriptors available for the analysis and the considerable variable reduction afforded by removing the correlated and invariant descriptors.

**Stepwise Multiple Linear Regression.** After identifying a suitable subset of variables, stepwise multiple linear regression was chosen to generate QSRR equations for all HPLC conditions tested. Stepwise multiple linear regression produces a multiple-term linear equation; however, not all independent variables are used. Step-by-step variables are added to the equation, and a new regression is performed. If the new variable contributes significantly to the regression equation, the variable is retained; otherwise, the variable is excluded, hence preventing overfitting. Stepwise multiple linear regressions were performed in the QSAR module of Cerius2. All 20 variables were used to generate stepwise regression equations (parameters: forward search, 100 max. steps, *F* value 2.00) for all HPLC conditions. A Cerius2 script was used to generate and export all the regression equations.

During the following discussion of the generated QSRR models, we will refer to  $r^2$  as the square of the correlation coefficient obtained from the stepwise multiple linear regression in the training set and to  $q^2$  as the leave-one-out  $r^2$ . Alternatively, we will refer to  $R^2$  as the square of the correlation coefficient in the test set, estimating the predictive ability of the generate QSRR models.

Twelve different training and test sets of each data table were used to evaluate the predictive power of the QSRR models. All 75 datasets were randomly divided into 12

**Table 6.** Averaged Stepwise Multiple Linear Regression Results Obtained from the Most Predictive Test and Training Set per HPLC Condition

	test sets	training sets
$R^2/r^2$	0.863	0.762
$q^2$		0.621
$F$		27.29
$N_{\text{obs}}$	8.00	35.00
$N_{\text{var}}$	2.00	7.42

different training sets, each including 80% of the original dataset. The remaining 20% of the original datasets were used to estimate the predictability of the generated QSRR models. For each training set, a stepwise regression model was constructed using the procedure described above.

Afterward, these QSRR equations were used to predict the retention time for observations included in test set. Golbraikh et al.<sup>51,52</sup> suggested that QSRR models are only acceptable if the  $q^2$  is  $>0.5$  and the predictive  $R^2$  is  $>0.6$ . Moreover, the square of the correlation coefficient  $R^2$  must be close to  $R_0^2$ , the square of the correlation coefficient for a regression with 0 intercept. The slope of the regression models is also identified as a critical factor and should take values close to 1.0.

Table 6 lists the average results for all 75 datasets. The model with the best  $R^2$  (best predictably) found for each dataset was used to identify the best predictive QSRR model for a specific HPLC condition. The average predictability of all 75 QSRR models is quite good, with a  $R^2 = 0.86$ . A pictorial presentation of the distribution of the statistical results is depicted in Figure 2. Figure 2 shows the distribution of the statistical parameters for QSRR models of the training sets, whereas Figure 3 represents the distribution of QSRR statistics for the test sets. The average number of descriptors for all models was 7.4. It can be seen in Figure 3 that  $R^2$  is  $>0.6$  for the majority of QSRR models (71 out of 75), and two models have  $R^2$  between 0.5 and 0.6.

Two additional QSRR models, both from the anticipated Aquasil HPLC column, are not predictive for the test set compounds. It is not surprising that the Aquasil C18 experiments were poorly predicted. A frequency table (Table 9) indicates that only 6 of 20 variables chosen by the GAs are important for this particular column. This implies that this stationary phase must have physicochemical properties different from all the other columns. Moreover, it can be observed from Figure 3 that the linear regression between observed and predicted retention time shows an average slope,  $k$ , of 1.07 and an average intercept of 0.04. The square of the correlation coefficient for a linear regression through the origin ( $R_0^2 = 0.76$ ) is close to the square of the correlation coefficient  $R^2 = 0.86$ , which indicates statistically stable models. The predictability of the QSRR models is evaluated by identifying the square of the correlation coefficient,  $R^2$ , of the observed versus the predicted retention time for a set of compounds that were not used to generate the QSRR models.

Figure 4 shows the actual versus predicted plots of the test set compounds for the models with the best and poorest predictive ability. The corresponding predicted and actual retention time values for these compounds are listed in



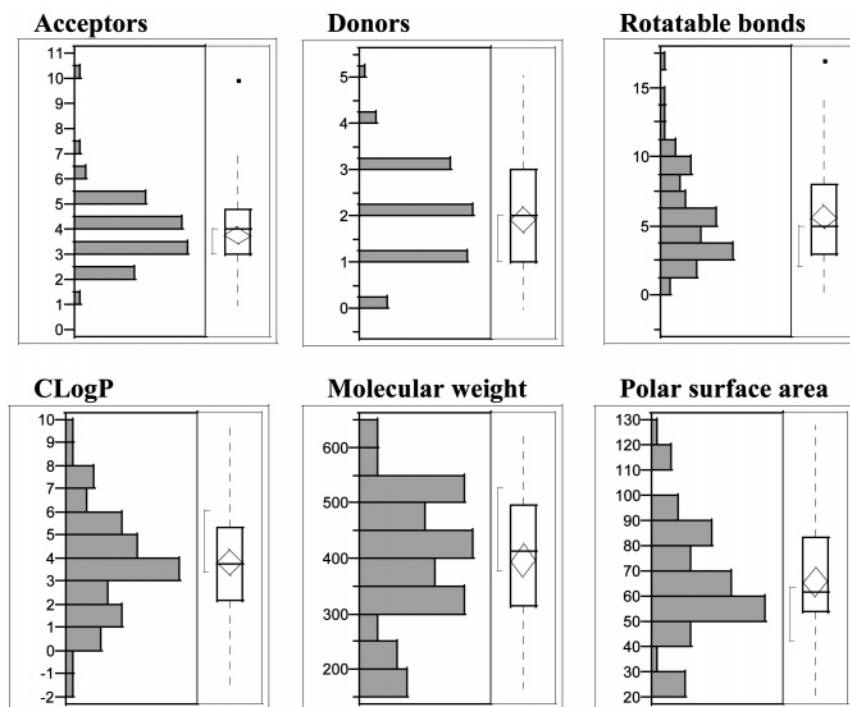


Figure 1. Illustration of the structural diversity in the standard solute set.

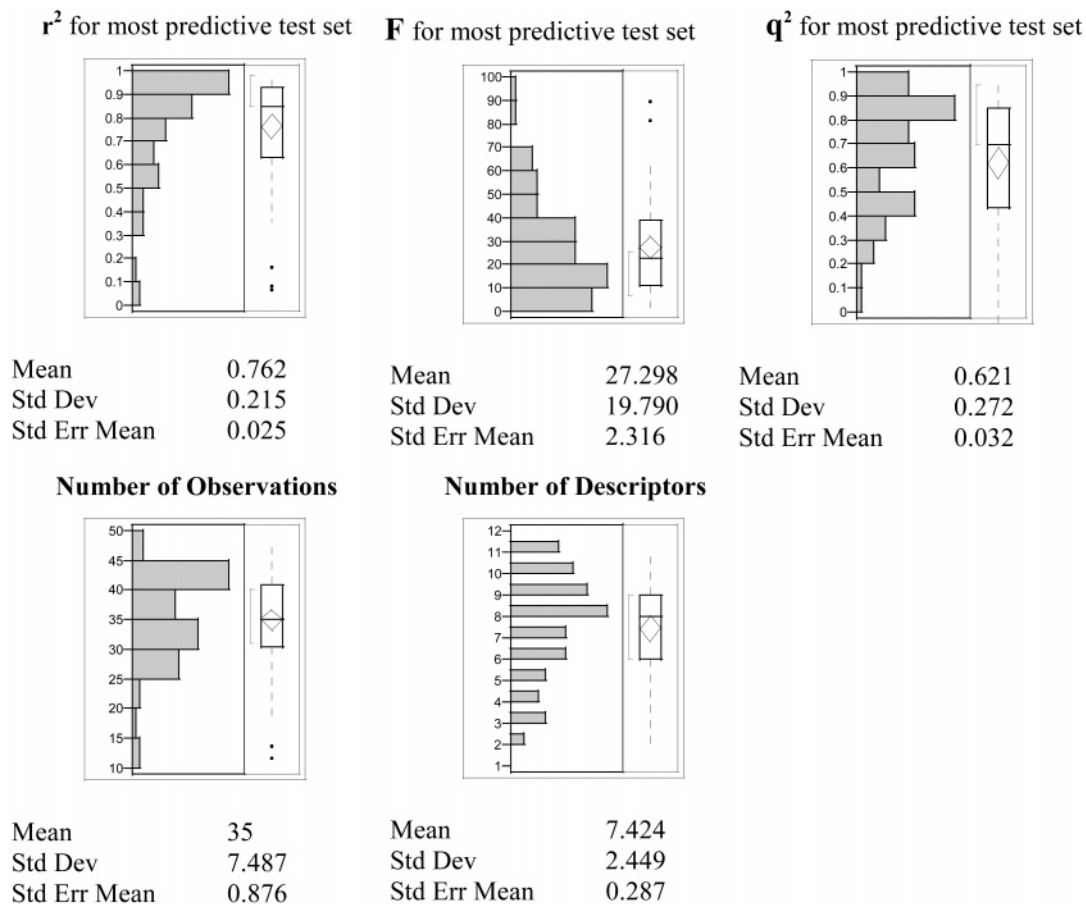
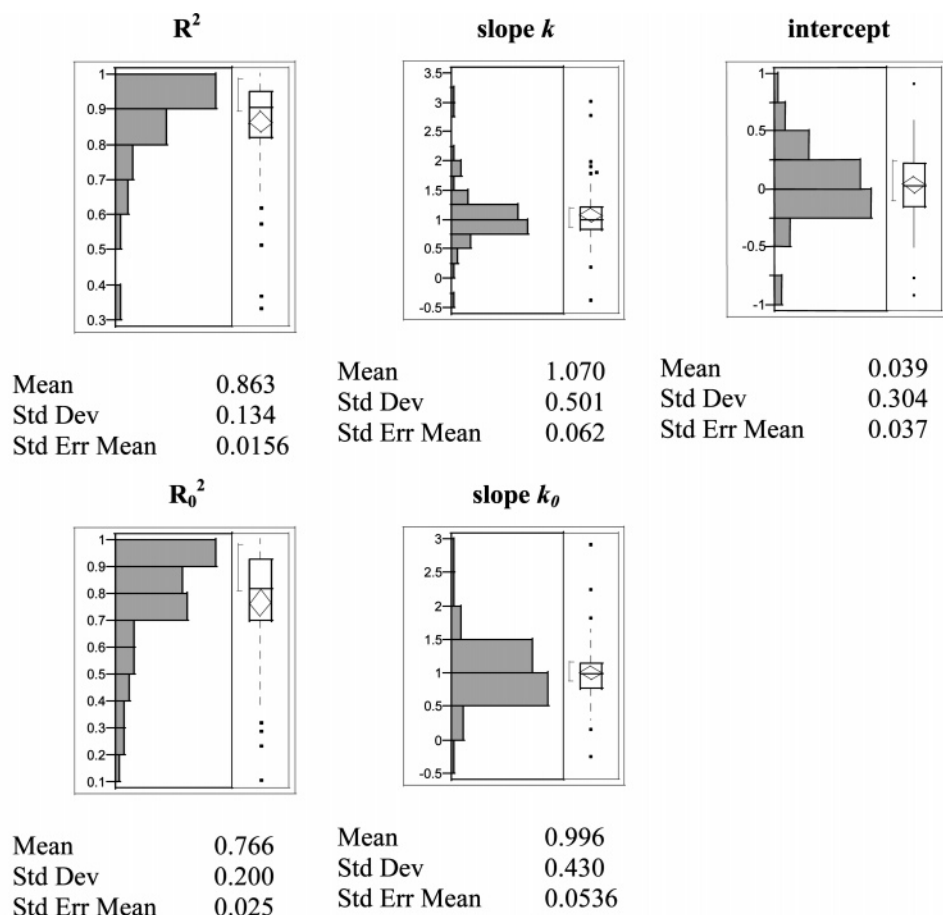


Figure 2. Statistical result from the stepwise multiple linear regressions for the most predictive QSRR model for each HPLC condition experiments ( $r^2$  symbolizes the square of the correlation coefficient,  $F$  value indicates a “signal-to-noise” ratio,  $q^2$  stands for a LOO- $r^2$ , NObs is an abbreviation for the number of observation, and NVars the number of variables).

Table 7. Since the structures of some compounds are not disclosed, we looked for any differences in the descriptors for test set compounds in the models with good predictability

versus test set compounds in the model with bad predictability. In general, the values for the descriptors are in the same range. The only differences we are able to point out is



**Figure 3.** Statistical results of observed versus predicted linear regression for the most predictive QSRR model used as a measure for the predictability.

**Table 7.** Normalized Observed and Predicted Retention Time Values for the Best and Worst Predictive QSRR Model

compd	best predictive model		worse predictive model	
	observed RT	predicted RT	observed RT	predicted RT
Pfizer 1			0.95	0.58
Pfizer 2			1.18	0.54
Pfizer 3			-0.62	0.48
Pfizer 4			-1.38	0.38
Pfizer 5			-1.82	-0.13
Pfizer 6	-1.87	-1.61	-1.43	-0.38
Reserpine	-0.20	-0.13	0.67	-0.01
Pfizer 7	-1.85	-1.57		
7-Hydroxy-4-coumarinylacetic acid	-0.76	-0.31		
Pfizer 8	-0.40	-0.30		
Pfizer 9	0.19	0.47		
Pfizer 10	0.47	0.57		
Pfizer 11	0.75	0.85		
Pfizer 12	1.74	1.65		

that the randomly selected test compounds in the data set with good predictability have twice the number of hydrogen bond donor atoms and only half of the average value of descriptors DCASA and DASA. These findings might indicate that the ratio of positive versus negative charged surface areas is slightly smaller for compounds in the good predictive model.

The worst predictive model is the QSRR model for experiment 14, in which the Aquasil HPLC column was combined with a 30–70% CH<sub>3</sub>CN gradient. The QSRR model with the best predictability can be used to predict the retention time of new solutes separated on the Waters

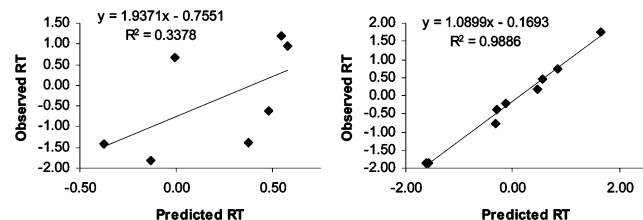
Symmetry C<sub>18</sub> with a 10–100% CH<sub>3</sub>CN gradient. As expected, the type of packing material in the HPLC column influences the quality of the QSRR. Conversely, the mobile phase gradient seems to have no significant influence on the QSRR model (Figure 5). However, the average predictability for QSRR models using a mobile phase gradient of 10–50% CH<sub>3</sub>CN and 30–70% CH<sub>3</sub>CN are below the grand mean of the average square of the correlation coefficient ( $R^2$ ).

To verify that all of the variables were making a significant contribution to the model (i.e., no single descriptor was overwhelming the model), we regenerated each model, leaving out each descriptor in turn. The averaged  $q^2$  value

**Table 8.** Averaged  $r^2$  and Cross-Validated  $q^2$  Values for Five Different Randomized Datasets<sup>a</sup>

	$r^2$	$q^2$
original dataset	0.762	0.621
randomized dataset 1	0.225	0.077
randomized dataset 2	0.222	0.074
randomized dataset 3	0.242	0.088
randomized dataset 4	0.242	0.069
randomized dataset 5	0.215	0.065

<sup>a</sup> For the sake of comparison, the same statistical measures are listed for the original data sets.



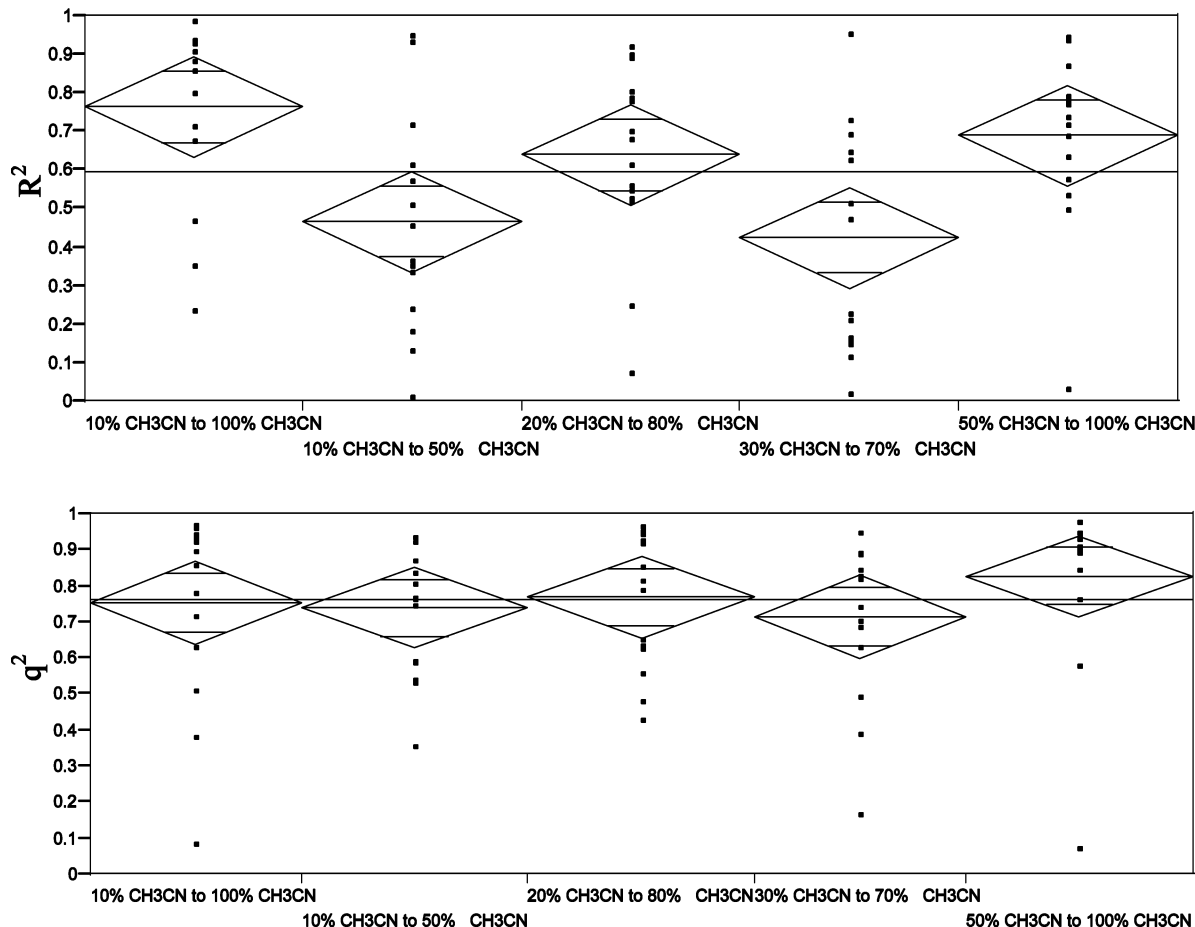
**Figure 4.** Plot 1 illustrates the observed versus the predicted retention time for the worse data set. Plot 2 depicts the observed versus the predicted retention time for the best data set.

obtained from these models indicates that the models are extremely stable ( $q^2 = 0.623 \pm 0.01$ ); no single variable is dominant. To further validate that the models generated were not statistical anomalies, we randomized each data set by scrambling the dependent variable for each observation. This randomization procedure should yield models with little or no statistical significance. To measure the randomization,

we took the ratio of the number of standard deviations of the mean value of the correlation coefficient of all random trials to the nonrandom correlation coefficient value. The larger this number, the greater the likelihood that the model generated with nonrandom data represents a true relationship between the data variables and activity. The mean value for this ratio (automatically generated in the Cerius2 software) was 3.44, considering all 75 QSRR models. The averaged  $r^2$  and  $q^2$  values for the randomized data sets were  $0.229 \pm 0.02$  and  $0.074 \pm 0.01$ , whereas the  $r^2$  and  $q^2$  values for the original data set were  $r^2 = 0.784$  and  $q^2 = 0.678$  (Table 8).

## Conclusion

In this study, we successfully provided QSRR models for 75 different HPLC experiments (5 gradients for each of 15 different columns) that predict the retention time for a given standard solute set under all the different HPLC conditions tested. Over 2000 descriptors were calculated for each of the solutes, and two different genetic algorithms were deployed to identify statistically significant variables in a huge pool of descriptors. Using a fusion method, the results of the GAs were combined, and the 20 best overall variables were identified for all HPLC methods. Afterward, these 20 variables were used as starting points to generate QSRR models for all HPLC conditions using stepwise linear regression. Overall, it was possible to generate 71 of 75 statistically significant and predictive QSRR models, with an average  $r^2 = 0.76$  and  $q^2 = 0.62$ . Work is currently



**Figure 5.** Correlation between gradient and predictability of the QSRR models.

Table 9. Frequency Scores of Descriptors for 15 HPLC Columns

descriptors	YMC_ Pro_C18	YMC_ ODS_ AQ	YMC_ Phenyl	MetaChem_ Polaris	Xterra_ MS_ C8	Xterra_ MS_ C18	Waters_ Symmetry_ C18	Alltima_ C18	PolymLab PLRPS	Prodigy PHE	MAXRP C12	PolarRP	LUNAC8	total count for columns	% of total count
#amine	17	11	9	10	12	10	17	5	17	15	10	3	5	150	33.33
CLogP	5	6	11	11	8	7	10	13	13	9	6	17	5	127	28.22
nNHRPh	6	3	2	8	7	8	13	5	9	15	13	2	0	100	22.22
vs_a_other	1	1	1	2	4	4	2	4	10	3	5	9	3	52	11.56
Atype_N_68	5	3	0	3	2	1	9	6	2	8	3	0	1	49	10.89
DCASA	1	4	0	7	6	5	5	1	5	5	3	1	0	47	10.44
C-027	0	4	0	2	7	3	2	2	3	3	3	1	4	45	10
Donors	6	0	0	3	1	2	3	4	2	1	4	2	3	36	8
DASA	1	0	0	7	3	4	1	4	1	3	0	3	2	35	7.78
QlogKp	3	0	1	3	5	5	5	0	0	3	1	0	0	35	7.78
FASA+	6	1	0	0	1	0	0	1	2	2	2	12	7	34	7.56
PDsol(mcg/mL)	1	0	0	0	5	0	1	2	5	3	6	8	1	34	7.56
Group_count_ for_chain_c=n	2	3	1	3	5	5	4	0	1	3	2	0	0	33	7.33
nCOOHPh	1	4	0	2	3	0	2	2	2	6	2	0	0	33	7.33
H8m	4	0	0	5	0	5	5	1	0	0	1	0	4	31	6.89
HATS3u	1	4	11	1	0	4	1	1	1	1	1	0	0	31	6.89
estate_sCH3	0	1	0	1	2	5	1	1	4	1	7	0	0	30	6.67
Group_count_ for_Phenyl	1	2	0	5	8	2	2	1	0	0	0	1	3	29	6.44
CLogP_error	0	0	7	4	1	4	1	0	1	0	0	2	0	29	6.44
SlogP_VSA9	3	1	0	4	3	3	2	1	1	1	0	0	6	28	6.22
variables_select	17	16	13	18	18	17	19	17	17	17	16	12	12	20	100

underway to utilize these models in the design of combinatorial libraries as well as to select an analytical method for purification and characterization of singleton compounds in an open access analytical lab, where expert analysis may not be readily available.

## References and Notes

- (1) Kasner, E.; Newman, J. *Mathematics and Imagination*; Penguin Books: London, 1940.
- (2) Lipinski, C. A. In *Tools for Oral Absorption. Part Two. Predicting Human Absorption. BIOTEC, PDD symposium*, AAPS: Miami, 1995.
- (3) Abraham, M. H.; Rosés, M.; Poole, C. F.; Poole, S. K. *J. Phys. Org. Chem.* **1997**, *10*, 358–368.
- (4) Platts, J. A.; Abraham, M. H.; Butina, D.; Hersey, A. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 835–845.
- (5) Platts, J. A.; Abraham, M. H.; Butina, D.; Hersey, A. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 71–80.
- (6) Kaliszan, R. *Quantitative Structure–Chromatographic Retention Relationships*; John Wiley & Sons: New York, 1987.
- (7) Kaliszan, R. *Computer aids to chemistry*; Vernin, G., Chanon, M., Eds.; Ellis Horwood, 1986; ISBN 0-8531-2774-3; *Trends Anal. Chem.* **1987**, *6*, XXIII–XXIV.
- (8) Kaliszan, R. Multivariate chemometrics in QSAR (quantitative structure–activity relationships): A dialogue, by P. P. Mage; Research Studies Press, Letchworth; Wiley: New York, 1988; ISBN 0-471-91570-X; *Chemom. Intell. Lab. Syst.* **1989**, *5*, 100–101.
- (9) Kaliszan, R. *J. Chromatogr., A* **1993**, *656*, 417–435.
- (10) Kaliszan, R. *J. Chromatogr., B* **1998**, *715*, 229–244.
- (11) Kaliszan, R. *Trends Anal. Chem.* **1999**, *18*, 400–410.
- (12) Kaliszan, R.; Haber, P.; Baczek, T.; Siluk, D.; Valko, K. *J. Chromatogr., A* **2002**, *965*, 117–127.
- (13) Kaliszan, R.; van Straten, M. A.; Markuszewski, M.; Cramers, C. A.; Claessens, H. A. *J. Chromatogr., A* **1999**, *855*, 455–486.
- (14) Baczek, T.; Kaliszan, R. *J. Biochem. Biophys. Methods* **2001**, *49*, 83–98.
- (15) Baczek, T.; Kaliszan, R. *J. Chromatogr., A* **2002**, *962*, 41–55.
- (16) Baczek, T.; Kaliszan, R. *Pre J. Chromatogr., A* **2003**, *987*, 29–37.
- (17) Galushko, S. V. *GIT Spezial Chromatogr.* **1996**, *2*, 88–93.
- (18) Snyder, L. R.; Quarry, M. A. *J. Liq. Chromatogr.* **1987**, *10*, 1789–1820.
- (19) Dolan, J. W.; Snyder, L. R.; Djordjevic, N. M.; Hill, D. W.; Waeghe, T. J. *J. Chromatogr., A* **1999**, *857*, 1–20.
- (20) Dolan, J. W.; Snyder, L. R.; Djordjevic, N. M.; Hill, D. W.; Waeghe, T. J. *J. Chromatogr., A* **1999**, *857*, 21–39.
- (21) Dolan, J. W.; Snyder, L. R.; Wolcott, R. G.; Haber, P.; Baczek, T.; Kaliszan, R.; Sander, L. C. *J. Chromatogr., A* **1999**, *857*, 41–68.
- (22) Williams, A.; Kolovanov, E.; Hofmann, H. *GIT Labor-Fachzeitschrift* **2000**, *44*, 154–157.
- (23) Delaurent, C.; Brenier-Maurel, C.; Galushko, S. V.; Tanchuk, V.; Shishkina, I.; Pylypchenko, O. *Spectra Analyse* **2002**, *31*, 34–37.
- (24) Waters Cooperation; <http://www.waters.com/WatersDivision/ContentD.asp?ref=JDRS-5KJKSP> (accessed 2003).
- (25) Thermo Electron Corporation; <http://www.keystonescientific.com/aquasil.htm> (accessed 2003).
- (26) Resolution Systems; <http://www.resolutionssystem.biz/portal/html/index.php> (accessed 2003).
- (27) Ansys Technologies MetaChem; <http://www.metachem.com/index.htm> (accessed 2003).
- (28) Alltech Associates Inc.; <http://www.alltechweb.com/US/Home.asp> (accessed 2003).
- (29) Polymer Laboratories Ltd; <http://www.polymerlabs.com/index.htm> (accessed 2003).

- (30) Phenomenex Inc.; <http://www.Phenomenex.com/Phen/Products/Brand.asp> (accessed 2003).
- (31) Halgren, T. J. *J. Comput. Chem.* **1996**, *17*, 490–519.
- (32) Todeschini, R. *Dragon*; <http://www.disat.unimib.it/chm/Dragon.htm>; Web ed.: Milan, Italy (accessed 2003).
- (33) Chemical Computing Group Inc. *MOE*, 2003.02 ed.; Montreal, Quebec, Canada.
- (34) Jorgensen, W. L. *QikProp*, 1.6 ed.; Yale University: New HavenCT.
- (35) Tripos Inc. *HiVol*; St. Louis, MO 63144.
- (36) Tripos Inc. *Sybyl*, 6.9 ed.; St. Louis, MO 63144.
- (37) Tripos Inc. *Molconn-Z*; St. Louis, Missouri, 63144.
- (38) Molecular Discovery Ltd. *Volsurf*, 3.0 ed.; Pinner, Middlesex, U.K.
- (39) Accelrys Inc. *TSAR 3D*, 3.3 ed.; San Diego, CA.
- (40) Pfizer Inc.; WebPK; PGRD; Ann Arbor, Michigan.
- (41) Partek Inc. *Partek Pro*, 5.1 ed.; St. Charles, MO.
- (42) Accelrys Inc. *Cerius2*, 4.8 ed.; San Diego, CA.
- (43) Rogers, D.; Hopfinger, A. J. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854–866.
- (44) SAS Institute Inc. *JMP*, 5.0 ed.
- (45) Raymond, J.; Willett, P. J. *Comput. Aided Mol. Des.* **2002**, *16*, 59–71.
- (46) Ginn, C. M. R.; Willett, P.; Bradshaw, J. *Perspect. Drug Discovery* **2000**, *20*, 1–16.
- (47) Belkin, N. J.; Cool, C.; Croft, W. B.; Callan, R. K. *Proc. SIGIR* **1993**, 339–346.
- (48) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. *J. Phys. Chem. A* **1998**, *102*, 3762–3772.
- (49) Consonni, V.; Todeschini, R.; Pavan, M. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 682–692.
- (50) Consonni, V.; Todeschini, R.; Pavan, M.; Gramatica, P. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 693–705.
- (51) Golbraikh, A.; Tropsha, A. *J. Mol. Graphics Modell.* **2002**, *20*, 269–276.
- (52) Golbraikh, A.; Shen, M.; Xiao, Z.; Xiao, Y.-D.; Lee, K.-H.; Tropsha, A. *J. Comput. Aided Mol. Des.* **2002**, *17*, 241–253.

CC049914Y