Journal of Chromatography A, 903 (2000) 145–154

JOURNAL OF CHROMATOGRAPHY A

www.elsevier.com/locate/chroma

# Development of comprehensive descriptors for multiple linear regression and artificial neural network modeling of retention behaviors of a variety of compounds on different stationary phases

M. Jalali-Heravi*, F. Parastar

*Department of Chemistry, Sharif University of Technology, P.O. Box 11365-9516, Tehran, Iran*

Received 23 February 2000; received in revised form 16 August 2000; accepted 16 August 2000

**Abstract**

A new series of six comprehensive descriptors that represent different features of the gas–liquid partition coefficient, $K_L$, for commonly used stationary phases is developed. These descriptors can be considered as counterparts of the parameters in the Abraham solvatochromic model of solution. A separate multiple linear regression (MLR) model was developed by using the six descriptors for each stationary phase of poly(ethylene glycol adipate) (EGAD), *N*,*N*,*N*′,*N*′-tetrakis(2-hydroxypropyl) ethylenediamine (THPED), poly(ethylene glycol) (Ucon 50 HB 660) (U50HB), di(2-ethylhexyl)phosphoric acid (DEHPA) and tetra-*n*-butylammonium *N*,*N*-(bis-2-hydroxylethyl)-2-aminoethanesulfonate (QBES). The results obtained using these models are in good agreement with the experiment and with the results of the empirical model based on the solvatochromic theory. A 6-6-5 neural network was developed using the descriptors appearing in the MLR models as inputs. Comparison of the mean square errors (MSEs) shows the superiority of the artificial neural network (ANN) over that of the MLR. This indicates that the retention behavior of the molecules on different columns show some nonlinear characteristics. The experimental solvatochromic parameters proposed by Abraham can be replaced by the calculated descriptors in this work.
© 2000 Elsevier Science B.V. All rights reserved.

*Keywords:* Stationary phases, LC; Multiple linear regression; Artificial neural network; Retention behavior

## 1. Introduction

For many years chromatographers have sought a method to characterize the solvation properties of stationary phases used in gas chromatography (GC) with the goal of providing a rational approach for selection of an optimum phase for a given separation and to predict retention of solutes on different phases. The most common solvent selectivity scales for gas–liquid chromatography being the system of phase constants proposed by Rohrschnaider [1,2] and later modified by McReynolds [3], Snyder's solvent selectivity triangle [4,5], dispersion selectivity indices [6,7], Hawkes polarity indices [8,9], solubility parameters [10,11], solvatochromic parameters [12,13] and several thermodynamic approaches [14–16].

The principal interactions that affect the solubility of a solvent in a liquid phase, and therefore retention, are dispersion, induction, orientation, and donor–acceptor interactions, including hydrogen bonding [17,18].

Dispersion (or London) forces arise from the electric field generated by rapidly varying dipoles

*Corresponding author.

PII: S0021-9673(00)00871-2

formed between nuclei and electrons at zero-point motion of the molecules. These forces are universal and independent of temperature. Induction (or Debye) forces arise from the interaction of a permanent dipole with a polarizable molecule. Orientation (or Keesom) forces arise from the net attraction between the molecules or portions of the molecules possessing a permanent dipole moment. Induction and orientation forces decrease with increasing temperature and at a sufficient high temperature disappear entirely as all orientations of the dipoles become equally probable. Complementing these physical interactions are donor–acceptor interactions of a chemical nature.

Different features of the gas–liquid partition coefficient, $K_L$, can be considered by using the cavity model of solvation [19]. The model assumes that the transfer of a solute from the ideal gas state to the solvent at the infinite dilution requires: (1) the creating of a cavity in the solvent of suitable size to accommodate the solute; (2) reorganization of the solvent molecules around the cavity (the Gibbs energy change for this process is probably very small compared with the other changes); and (3) interaction of the solute molecule with the surrounding solvent molecules represented by the sum of the individual Gibbs energy contributions to the solvation process [19]. As demonstrated by Abraham and co-workers these changes can be described by the equation:

$$\log K_L = c + rR_2 + s\pi_2^H + a\alpha_2^H + b\beta_2^H + l \log L^{16} \quad (1)$$

where $K_L$ is the solute gas–liquid partition coefficient, $c$ is a constant, $R_2$ is the solute excess molar refraction, $\pi_2^H$ is the effective solute dipolarity/polarizability parameter, $\alpha_2^H$ is the effective hydrogen-bond acidity, $\beta_2^H$ is the effective hydrogen-bond basicity and $L^{16}$ is the gas–liquid partition coefficient of the hexadecane at 25°C. This model is similar to the solvatochromic theory of solution except that the explanatory variables $R_2$, $\pi_2^H$, $\alpha_2^H$, $\beta_2^H$ and log $L^{16}$ are solvation parameters derived from the equilibrium measurements and further refined (and augmented) by multiple linear regression analysis of solvents of assumed characteristic properties. The solvent parameters *r*, *s*, *a*, *b* and *l* are unambiguously defined: the *r* constant refers to the ability of the solvent to interact with the solute π- and n-electron pairs; the *s* constant refers to the ability of the solvent to take part in dipole–dipole and dipole–induced dipole interactions; the *a* constant is a measure of the hydrogen-bond basicity of the solvent; the *b* constant is a measure of the hydrogen-bond acidity of the solvent; and the *l* constant incorporates contributions from solvent cavity formation and dispersion interactions, and more specifically in gas–liquid chromatography indicates how well the solvent will separate members in a homologous series. Experimentally, the solvent specific constants are determined from a number of measurements of log $K_L$ for solutes with known explanatory variables using multiple linear regression analysis [20–22].

Li et al. have proposed a similar model to that of Eq. (1) [23]. However, their model differs in the values taken for the explanatory variables and in the use of an empirical correction term for the influence of the polarizability of the solute on the estimate of the dipole-type interactions. Kollie and co-workers [24,25] have used a general expression, Eq. (2), to represent the various free energy contributions to the solvation process:

$$\Delta G_S^{SOLN}(X) = \Delta G_S^{CAV}(X) + \Delta G_S^{NP}(X) + \Delta G_S^{P}(X) \quad (2)$$

where $\Delta G_S^{SOLN}(X)$ is the partial Gibbs free energy of solution for the transfer of solute $X$ from the gas phase to the stationary phase S; $\Delta G_S^{CAV}(X)$ is the partial Gibbs free energy of cavity formation for solute $X$; $\Delta G_S^{NP}(X)$ and $\Delta G_S^{P}(X)$ are the partial Gibbs free energies of the interactions of the nonpolar and polar contributions of solute $X$ with the surrounding solvent, respectively.

The main aim behind modeling is the prediction of different quantities and at the same time to reduce the consumption of solvents and expensive chemicals. Therefore, we believe that the generated models should mainly contain calculated descriptors instead of empirical ones. Keeping this in mind, we have attempted to develop the multiple linear regression (MLR) and artificial neural network (ANN) models to predict the log $K_L$ by using a new series of descriptors that are calculated parameters. These descriptors should represent different interactions that affect the retention phenomena in the chromatographic studies. The results of the present work show that the Abraham's empirical parameters can be replaced by a series of calculated descriptors in modeling of the gas–liquid partition coefficients of a variety of compounds.

## 2. Experimental

This work contains four stages: (1) selection of data set, (2) regression analysis, (3) ANN generation and (4) evaluation of the models.

### 2.1. Data set

The data set was selected from Ref. [25]. This set consists of 54 molecules that were randomly divided into two groups, training set and prediction set. The training set consists of 39 compounds from a variety of organic compounds and the prediction set consists of 15 compounds. The prediction set is a good representative of the training set. The names of the test solutes used in this study are summarized in Table 1.

### 2.2. Regression analysis

Six descriptors were calculated for interpreting solute–solvent interactions in the present work. These descriptors consist of the dipole moment (DIPOL), the highest occupied molecular orbital (HOMO), the partial charge of the most negative atom (PCHNEG), the partial charge of the most positive hydrogen (PCHPOSH), molecular mass ($M_r$) and van der Waals volume (VOLUME) (see Table 3). The quantum–mechanical descriptors of DIPOL, HOMO, PCHNEG and PCHPOSH were obtained using the MOPAC program (version 6) [26]. The van der Waals volume was calculated using a program called BASPRO that was developed in our laboratory [27]. The MLR models were generated by SPSS (for windows 6.0) software [28].

### 2.3. Neural network generation

The detailed theory behind an artificial neural network is adequately described elsewhere [29,30].

Table 1
Chemical names of the molecules studied in this work

| No. | Compound |
|---|---|
| *Training set* | |
| 1 | *n*-Octane |
| 2 | Pentan-2-one |
| 3 | Methyl octanoate |
| 4 | Dimethyl sulfoxide |
| 5 | *n*-Undecane |
| 6 | Methyl nonanoate |
| 7 | Benzodioxane |
| 8 | *cis*-Hydrindane |
| 9 | Butan-1-ol |
| 10 | Dodecane |
| 11 | *n*-Hexadecane |
| 12 | *N*,*N*-Dimethylacetamide |
| 13 | *n*-Butylbenzene |
| 14 | *n*-Tridecane |
| 15 | Methylhexanoate |
| 16 | Hexan-2-one |
| 17 | Decan-2-one |
| 18 | Di-*n*-hexyl ether |
| 19 | Heptan-1-ol |
| 20 | *n*-Pentadecane |
| 21 | *N*,*N*-Dibutylformamide |
| 22 | Benzene |
| 23 | 1-Dodecyne |
| 24 | Methyl tetradecanoate |
| 25 | Dioxane |
| 26 | Nonan-2-one |
| 27 | 2-Methylpentan-2-ol |
| 28 | Nitrobenzene |
| 29 | Nonanal |
| 30 | Heptan-2-one |
| 31 | Nonan-1-ol |
| 32 | Anisole |
| 33 | Dodecan-2-one |
| 34 | Benzonitrile |
| 35 | Methyl dodecanoate |
| 36 | *n*-Decane |
| 37 | Methyl decanoate |
| 38 | Methyl octanoate |
| 39 | Nitropropane |
| *Prediction set* | |
| 40 | 2-Octyne |
| 41 | Phenyl ether |
| 42 | 1-Nitrohexane |
| 43 | Nitrocyclohexane |
| 44 | 4-Phenyl-1,3-dioxane |
| 45 | *n*-Nonane |
| 46 | *n*-Tetradecane |
| 47 | Octan-2-one |
| 48 | Undecan-2-one |
| 49 | Methylheptanoate |
| 50 | Methylundecanoate |
| 51 | Methylhexadecanoate |
| 52 | Octan-1-ol |
| 53 | *N*,*N*-Dimethylformamide |
| 54 | Nitropentane |

Therefore, only the points relevant to this work are described here.The ANN program was written in FORTRAN 90 in our laboratory. A back-propagation strategy was used for the training of the network [31]. Before learning the network, the input vector and output values were normalized between 0.1 and 0.9. The normalizing of the output values between 0.1 and 0.9 allows the network to slightly exceed the minimum and maximum values that were given in the original data file. A sigmoidal function was used as transfer function for the network. The initial weights were selected randomly between −1 and 1. Before training of the ANN, the network parameters were optimized. The optimization strategy was described elsewhere [32,33]. The optimum number of neurons in the hidden layer, momentum and learning rate were 6, 0.4 and 0.2, respectively. Then the network was trained with training set for the optimization of the weights and biases values using back-propagation strategy. The trained network was used for the prediction of log $K_L$ of the compounds included in the prediction set.

## 3. Results and discussion

The main aim of the present work was to define a series of new descriptors that have two properties. First, that they can be used as a general parameters. This means that they can describe the retention behaviors of a variety of organic compounds on commonly used GC stationary phases. Second, that they can be obtained by calculations and in fact can be replaced by empirical parameters. To fulfil the generality of descriptors one needs a very diverse data set. Therefore as can be seen in Table 1 a data set consisting of alkanes, alcohols, ketones, ethers, esters, amides, sulfoxides, nitrile and nitro-containing compounds was chosen to develop the appropriate models. The prediction set also consists of different molecules included in the training set and adequately represents the training set.

The next step was choosing the descriptors. Since it was shown that the cavity model is useful in predicting the gas–liquid coefficient, $K_L$, of different compounds [24,34,35], we have tried to generate a series of calculated parameters that in some way represent different parameters included in Eq. (1), i.e., $R_2$, $\pi_2^H$, $\alpha_2^H$, $\beta_2^H$ and log $L^{16}$. Among different parameters defined, a total of six descriptors; DIPOL, HOMO, PCHNEG, PCHPOSH, $M_r$ and VOLUME show some correlations with Abraham's solvatochromic parameters. Table 2 shows these correlations.

The parameter HOMO is a measure of the ability of a molecule to interact with the π- and n-electron pairs of the other molecules. It can be seen from Table 2 that this parameter shows a correlation coefficient of 0.697 with the solute excess molar refraction, $R_2$. The counterpart of the $\pi_2^H$ parameter in the Abraham's equation is the DIPOL which is the dipole moment of the molecules. Both parameters of the $\pi_2^H$ in the solvatochromic model of Abraham and the DIPOL in our models, represent the ability of a molecule to take part in dipole–dipole and dipole–induced dipole interactions. The parameters of $\alpha_2^H$ and $\beta_2^H$ show some correlation with the PCHPOSH and PCHNEG descriptors, respectively. It is obvious that the partial charges of the most positive hydrogen and the partial charges of the most negative atom can be considered as a measure of acidity and basicity of a molecule, respectively. However, as can be seen from Table 2, correlation between these parameters and their counterparts in Abraham's equation are

Table 2
Correlations between the solvatochromic parameters and different parameters studied in this work

| Descriptor[a] | $R_2$ | $\pi_2^H$ | $\alpha_2^H$ | $\beta_2^H$ | Log $L^{16}$ |
|---|---|---|---|---|---|
| DIPOL | 0.270 | **0.790** | −0.046 | 0.530 | −0.315 |
| HOMO | **0.697** | 0.389 | −0.044 | 0.386 | −0.107 |
| $M_r$ | −0.350 | −0.295 | −0.324 | −0.168 | **0.928** |
| PCHNEG | 0.156 | −0.473 | −0.154 | **−0.542** | 0.247 |
| PCHPOSH | 0.487 | 0.445 | **0.535** | 0.426 | −0.500 |
| VOLUME | −0.523 | −0.488 | −0.261 | −0.315 | **0.914** |

[a] Definition of descriptors is given in the text.

Table 3
The calculated values of different descriptors for all of the molecules studied in this work[a]

| No.[b] | PCHNEG | $M_r$ | DIPOL | HOMO | PCHPOSH | VOLUME |
|---|---|---|---|---|---|---|
| 1 | −0.2104 | 114.23 | 0.00 | −11.07 | 0.0787 | 146.74 |
| 2 | −0.2909 | 86.13 | 2.79 | −10.53 | 0.1069 | 98.72 |
| 3 | −0.3505 | 158.24 | 1.67 | −11.21 | 0.1182 | 174.38 |
| 4 | −0.7780 | 78.13 | 3.95 | −9.53 | 0.1280 | 70.81 |
| 5 | −0.2104 | 156.31 | 0.01 | −11.06 | 0.0787 | 197.21 |
| 6 | −0.3505 | 172.27 | 1.66 | −11.18 | 0.1182 | 191.41 |
| 7 | −0.2000 | 136.15 | 0.91 | −8.94 | 0.1498 | 128.14 |
| 8 | −0.1577 | 124.23 | 0.03 | −10.66 | 0.0874 | 141.65 |
| 9 | −0.3292 | 74.12 | 1.52 | −10.85 | 0.1972 | 87.58 |
| 10 | −0.2104 | 170.34 | 0.00 | −11.06 | 0.0787 | 214.23 |
| 11 | −0.2104 | 226.45 | 0.00 | −10.97 | 0.0788 | 281.48 |
| 12 | −0.3696 | 87.12 | 3.58 | −9.54 | 0.1200 | 93.57 |
| 13 | −0.2107 | 134.22 | 0.34 | −9.30 | 0.1327 | 155.02 |
| 14 | −0.2104 | 184.36 | 0.01 | −11.03 | 0.0788 | 230.95 |
| 15 | −0.3505 | 130.19 | 1.68 | −11.25 | 0.1182 | 140.76 |
| 16 | −0.2905 | 100.16 | 2.77 | −10.53 | 0.1027 | 115.52 |
| 17 | −0.2906 | 156.27 | 2.75 | −10.51 | 0.1069 | 183.01 |
| 18 | −0.2823 | 186.34 | 1.17 | −10.39 | 0.0942 | 230.65 |
| 19 | −0.3292 | 116.20 | 1.52 | −10.85 | 0.1972 | 138.14 |
| 20 | −0.2104 | 212.42 | 0.01 | −10.99 | 0.0788 | 264.70 |
| 21 | −0.3620 | 157.26 | 3.60 | −9.62 | 0.1191 | 178.53 |
| 22 | −0.1301 | 78.11 | 0.00 | −9.65 | 0.1301 | 88.26 |
| 23 | −0.2105 | 138.25 | 0.09 | −10.11 | 0.0961 | 170.85 |
| 24 | −0.3506 | 242.40 | 1.67 | −11.10 | 0.1182 | 275.76 |
| 25 | −0.2694 | 88.11 | 0.00 | −10.21 | 0.1122 | 86.18 |
| 26 | −0.2906 | 142.24 | 2.76 | −10.51 | 0.1069 | 166.12 |
| 27 | −0.3257 | 102.18 | 1.65 | −10.84 | 0.1965 | 121.01 |
| 28 | −0.3586 | 123.11 | 5.24 | −10.56 | 0.1709 | 110.18 |
| 29 | −0.2908 | 142.24 | 2.78 | −10.57 | 0.1155 | 166.31 |
| 30 | −0.2906 | 114.19 | 2.77 | −10.52 | 0.1069 | 132.41 |
| 31 | −0.3292 | 144.26 | 1.52 | −10.85 | 0.1973 | 171.85 |
| 32 | −0.2117 | 108.14 | 1.25 | −9.00 | 0.1481 | 113.48 |
| 33 | −0.2905 | 184.32 | 2.75 | −10.51 | 0.1069 | 216.65 |
| 34 | −0.1349 | 103.12 | 3.34 | −10.02 | 0.1450 | 107.07 |
| 35 | −0.3505 | 214.35 | 1.67 | −11.12 | 0.1182 | 242.02 |
| 36 | −0.2104 | 142.28 | 0.00 | −11.06 | 0.0787 | 180.46 |
| 37 | −0.3505 | 186.29 | 1.67 | −11.15 | 0.1182 | 208.27 |
| 38 | −0.3518 | 298.51 | 1.67 | −11.01 | 0.1176 | 343.09 |
| 39 | −0.3660 | 89.09 | 4.50 | −11.73 | 0.1349 | 84.79 |
| 40 | −0.2107 | 110.20 | 0.08 | −10.11 | 0.0961 | 137.18 |
| 41 | −0.1731 | 170.21 | 1.25 | −8.95 | 0.1502 | 172.32 |
| 42 | −0.3620 | 137.17 | 4.61 | −11.57 | 0.1359 | 135.52 |
| 43 | −0.3661 | 129.16 | 4.51 | −11.35 | 0.1294 | 124.05 |
| 44 | −0.2903 | 164.20 | 2.01 | −9.50 | 0.1474 | 161.66 |
| 45 | −0.2104 | 128.26 | 0.01 | −11.06 | 0.0787 | 163.52 |
| 46 | −0.2104 | 198.39 | 0.00 | −11.01 | 0.0788 | 247.97 |
| 47 | −0.2906 | 128.21 | 2.75 | −10.52 | 0.1069 | 148.93 |
| 48 | −0.2906 | 170.29 | 2.75 | −10.51 | 0.1069 | 199.86 |
| 49 | −0.3505 | 144.21 | 1.67 | −11.24 | 0.1182 | 157.45 |
| 50 | −0.3505 | 200.32 | 1.66 | −11.13 | 0.1182 | 225.06 |
| 51 | −0.3505 | 270.45 | 1.67 | −11.05 | 0.1182 | 309.45 |
| 52 | −0.3292 | 130.23 | 1.51 | −10.85 | 0.1972 | 155.05 |
| 53 | −0.3605 | 73.09 | 3.69 | −9.60 | 0.1215 | 76.90 |
| 54 | −0.3663 | 117.15 | 4.59 | −11.67 | 0.1349 | 118.49 |

[a] Definition of the descriptors is given in the text.

[b] Numbers refer to the molecules given in Table 1.

relatively low compared with the other parameters. This is due to the fact that the $\alpha_2^H$ and $\beta_2^H$ parameters have similar values for a large number of molecules [20], while the values of the calculated descriptors of PCHNEG and PCHPOSH are different for each molecule of the data set. Another important parameter in cavity model is log $L^{16}$ that incorporates contributions from solvent cavity formation and dispersion interactions. It is obvious that as the molecular mass and the volume of a molecule increase, the cavitation energy and dispersion interactions increase. Therefore, the parameter log $L^{16}$ in the solvatochromic model can be replaced by the parameters $M_r$ and VOLUME in the models given in this work (see Table 2).

Table 3 demonstrates the calculated values of different descriptors for all of the molecules included in the training and the prediction sets. The dipole moment of molecules varies from 0.00 to 5.24 Debye indicating that the data set consists of polar and nonpolar molecules. All of the other descriptors also show a large variation that is due to the diversity of molecules studied in this work.

### 3.1. Analysis of MLR models

A separate MLR model was developed for each stationary phase using the above mentioned descriptors for which the specifications are summarized in Table 4. The linear equations where obtained using the ENTER strategy in the SPSS for Windows software. The statistics for each model are also given in this table. As can be seen the correlation coefficients range from 0.944 to 0.966 with $F$ values of 44 to 75. In addition, the standard errors for different models are low indicating the suitability and generality of the descriptors. It is noteworthy that except for the parameter VOLUME, the signs of the other coefficients are the same for different stationary phases. The signs for the coefficients of the parameters DIPOL, HOMO, $M_r$ and PCHPOSH are positive and the sign of the coefficient of PCHNEG is negative. Consideration of these signs indicates that as these parameters increase the solute gas–liquid partition coefficient increases. This is in agreement with the experiment and with the empirical solvatochromic parameters of the cavity model.

Table 4
Specification of MLR models for different stationary phases

| Column[a] | Variable | | | | | | |
|---|---|---|---|---|---|---|---|
| | DIPOL | HOMO | $M_r$ | PCHNEG | PCHPOSH | VOLUME | (Constant) |
| EGAD | 0.1477 | 0.4276 | 0.0235 | −1.6863 | 4.6202 | −0.0093 | 3.7620 |
| | (±0.0361) | (±0.0619) | (±0.0048) | (±0.4245) | (±1.2465) | (±0.0041) | (±0.6341) |
| | | | $n=39$, $r=0.958$, $F=59$, SE=0.221 | | | | |
| THPED | 0.1434 | 0.3613 | 0.0129 | −1.5868 | 5.5741 | 0.0014 | 2.9767 |
| | (±0.0391) | (±0.0672) | (±0.0052) | (±0.4605) | (±1.3522) | (0.0044) | (±0.6878) |
| | | | $n=39$, $r=0.952$, $F=52$, SE=0.240 | | | | |
| U50HB | 0.0952 | 0.2914 | 0.0193 | −0.7858 | 3.6905 | −0.0036 | 2.7733 |
| | (±0.0376) | (±0.0646) | (±0.0050) | (±0.4427) | (±1.3000) | (±0.0043) | (±0.6613) |
| | | | $n=39$, $r=0.955$, $F=55$, SE=0.231 | | | | |
| DEHPA | 0.0667 | 0.2955 | 0.0090 | −1.1971 | 4.3135 | 0.0069 | 2.5414 |
| | (±0.0369) | (±0.0634) | (±0.0049) | (±0.4349) | (±1.2770) | (±0.0042) | (±0.6496) |
| | | | $n=39$, $r=0.966$, $F=74$, SE=0.227 | | | | |
| QBES | 0.1828 | 0.3668 | 0.0169 | −1.0050 | 10.5564 | −0.0049 | 2.6027 |
| | (±0.0419) | (±0.0720) | (±0.0056) | (±0.4936) | (±1.4493) | (±0.0047) | (±0.7372) |
| | | | $n=39$, $r=0.944$, $F=44$, SE=0.257 | | | | |

[a] EGAD, Poly(ethylene glycol adipate); THPED, *N*,*N*,*N*′,*N*′-tetrakis(2-hydroxypropyl)ethylenediamine; U50HB, poly(ethylene glycol) (Ucon 50 HB 660); DEHPA, di(2-ethylhexyl)phosphoric acid; QBES, tetra-*n*-butylammonium *N*,*N*-(bis-2-hydroxylethyl)-2-aminoethanesulfonate.

Table 5
The experimental and calculated values of log $K_L$ using the MLR and ANN models for different stationary phases[a]

| No. | EGAD | | | THPED | | | U50HB | | | DEHPA | | | QBES | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Experimental | MLR | ANN | Experimental | MLR | ANN | Experimental | MLR | ANN | Experimental | MLR | ANN | Experimental | MLR | ANN |
| *Training set* | | | | | | | | | | | | | | | |
| 1 | 1.09 | 1.07 | 1.05 | 1.38 | 1.42 | 1.44 | 1.68 | 1.68 | 1.69 | 1.91 | 1.91 | 1.86 | 0.70 | 0.79 | 0.75 |
| 2 | 1.56 | 1.76 | 1.52 | 1.67 | 1.87 | 1.66 | 1.67 | 1.90 | 1.70 | 1.63 | 1.89 | 1.72 | 1.44 | 1.64 | 1.34 |
| 3 | 2.34 | 2.45 | 2.23 | 2.52 | 2.66 | 2.44 | 2.64 | 2.80 | 2.58 | 2.74 | 2.91 | 2.70 | 2.04 | 2.20 | 2.02 |
| 4 | 3.45 | 3.35 | 3.33 | 3.30 | 3.15 | 3.15 | 2.98 | 2.71 | 3.05 | 2.91 | 2.67 | 2.78 | 3.14 | 2.93 | 3.12 |
| 5 | 1.61 | 1.59 | 1.55 | 2.07 | 2.04 | 2.03 | 2.42 | 2.31 | 2.40 | 2.73 | 2.64 | 2.70 | 1.26 | 1.25 | 1.22 |
| 6 | 2.55 | 2.63 | 2.41 | 2.76 | 2.87 | 2.66 | 2.88 | 3.02 | 2.81 | 3.01 | 3.16 | 2.96 | 2.23 | 2.37 | 2.21 |
| 7 | 3.19 | 3.11 | 3.07 | 3.07 | 2.96 | 3.06 | 3.30 | 3.13 | 3.17 | 3.08 | 2.96 | 3.11 | 3.12 | 2.94 | 2.89 |
| 8 | 1.69 | 1.48 | 1.69 | 1.99 | 1.66 | 1.95 | 2.07 | 2.00 | 2.21 | 2.39 | 2.06 | 2.32 | 1.49 | 1.17 | 1.45 |
| 9 | 1.81 | 1.74 | 1.72 | 2.08 | 1.97 | 1.88 | 1.93 | 1.86 | 1.80 | 2.02 | 1.96 | 1.82 | 2.27 | 2.13 | 2.08 |
| 10 | 1.79 | 1.76 | 1.70 | 2.30 | 2.24 | 2.22 | 2.66 | 2.52 | 2.61 | 3.00 | 2.89 | 2.96 | 1.42 | 1.40 | 1.37 |
| 11 | 2.49 | 2.49 | 2.57 | 3.22 | 3.09 | 3.22 | 3.64 | 3.39 | 3.56 | 4.09 | 3.89 | 4.06 | 2.19 | 2.05 | 2.12 |
| 12 | 2.93 | 2.56 | 2.84 | 2.90 | 2.55 | 2.79 | 2.66 | 2.41 | 2.69 | 2.66 | 2.35 | 2.58 | 2.64 | 2.40 | 2.74 |
| 13 | 2.14 | 2.51 | 2.11 | 2.29 | 2.68 | 2.28 | 2.48 | 2.78 | 2.46 | 2.64 | 2.93 | 2.61 | 1.92 | 2.37 | 1.88 |
| 14 | 1.96 | 1.94 | 1.90 | 2.53 | 2.45 | 2.45 | 2.91 | 2.74 | 2.84 | 3.27 | 3.14 | 3.23 | 1.63 | 1.57 | 1.55 |
| 15 | 1.93 | 2.08 | 1.80 | 2.05 | 2.23 | 1.97 | 2.17 | 2.37 | 2.09 | 2.20 | 2.41 | 2.14 | 1.66 | 1.88 | 1.58 |
| 16 | 1.77 | 1.91 | 1.66 | 1.90 | 2.05 | 1.81 | 1.91 | 2.09 | 1.85 | 1.90 | 2.11 | 1.89 | 1.63 | 1.74 | 1.44 |
| 17 | 2.62 | 2.62 | 2.52 | 2.86 | 2.89 | 2.75 | 2.89 | 2.95 | 2.75 | 3.01 | 3.11 | 2.89 | 2.42 | 2.40 | 2.28 |
| 18 | 2.20 | 2.63 | 2.19 | 2.55 | 3.08 | 2.55 | 2.66 | 3.19 | 2.69 | 3.10 | 3.57 | 3.05 | 1.79 | 2.29 | 1.77 |
| 19 | 2.47 | 2.26 | 2.18 | 2.83 | 2.58 | 2.54 | 2.70 | 2.49 | 2.47 | 2.88 | 2.69 | 2.68 | 2.89 | 2.59 | 2.56 |
| 20 | 2.31 | 2.31 | 2.32 | 2.99 | 2.88 | 2.94 | 3.40 | 3.18 | 3.30 | 3.82 | 3.64 | 3.77 | 2.01 | 1.89 | 1.92 |
| 21 | 3.54 | 3.37 | 3.39 | 3.72 | 3.52 | 3.64 | 3.39 | 3.43 | 3.54 | 3.61 | 3.54 | 3.67 | 3.21 | 3.12 | 3.14 |
| 22 | 1.35 | 1.47 | 1.37 | 1.39 | 1.55 | 1.49 | 1.57 | 1.73 | 1.60 | 1.58 | 1.72 | 1.62 | 1.29 | 1.45 | 1.24 |
| 23 | 2.27 | 1.91 | 2.13 | 2.54 | 2.22 | 2.37 | 2.73 | 2.41 | 2.61 | 2.99 | 2.66 | 2.83 | 2.03 | 1.62 | 1.89 |
| 24 | 3.58 | 3.53 | 3.50 | 3.93 | 3.92 | 3.92 | 4.06 | 4.10 | 4.03 | 4.37 | 4.40 | 4.37 | 3.16 | 3.16 | 3.20 |
| 25 | 1.85 | 1.64 | 1.75 | 1.77 | 1.59 | 1.66 | 1.86 | 1.81 | 1.78 | 1.69 | 1.72 | 1.69 | 1.56 | 1.37 | 1.54 |
| 26 | 2.40 | 2.45 | 2.36 | 2.62 | 2.69 | 2.54 | 2.64 | 2.74 | 2.55 | 2.73 | 2.86 | 2.66 | 2.22 | 2.25 | 2.13 |
| 27 | 1.78 | 2.10 | 1.88 | 1.99 | 2.39 | 2.23 | 1.88 | 2.29 | 2.14 | 2.09 | 2.44 | 2.33 | 1.95 | 2.45 | 2.28 |
| 28 | 3.02 | 3.28 | 3.11 | 3.01 | 3.17 | 3.11 | 3.14 | 3.09 | 3.03 | 2.93 | 2.81 | 2.91 | 3.14 | 3.38 | 3.05 |
| 29 | 2.36 | 2.47 | 2.55 | 2.61 | 2.72 | 2.72 | 2.66 | 2.76 | 2.72 | 2.76 | 2.89 | 2.80 | 2.25 | 2.32 | 2.40 |
| 30 | 1.98 | 2.11 | 1.98 | 2.14 | 2.28 | 2.11 | 2.16 | 2.32 | 2.14 | 2.18 | 2.37 | 2.19 | 1.83 | 1.94 | 1.78 |
| 31 | 2.88 | 2.60 | 2.95 | 3.31 | 2.99 | 3.24 | 3.19 | 2.91 | 3.23 | 3.45 | 3.17 | 3.37 | 3.27 | 2.90 | 3.13 |
| 32 | 2.25 | 2.62 | 2.23 | 2.22 | 2.61 | 2.25 | 2.42 | 2.66 | 2.29 | 2.33 | 2.62 | 2.29 | 2.16 | 2.57 | 2.09 |
| 33 | 3.04 | 2.97 | 2.92 | 3.34 | 3.30 | 3.22 | 3.37 | 3.37 | 3.19 | 3.57 | 3.59 | 3.41 | 2.81 | 2.71 | 2.62 |
| 34 | 2.77 | 2.29 | 2.67 | 2.78 | 2.33 | 2.72 | 2.90 | 2.42 | 2.62 | 2.65 | 2.26 | 2.57 | 2.90 | 2.41 | 2.76 |
| 35 | 3.17 | 3.17 | 3.02 | 3.46 | 3.51 | 3.39 | 3.59 | 3.67 | 3.52 | 3.83 | 3.91 | 3.79 | 2.79 | 2.85 | 2.80 |
| 36 | 1.44 | 1.41 | 1.38 | 1.84 | 1.83 | 1.83 | 2.17 | 2.10 | 2.17 | 2.46 | 2.40 | 2.43 | 1.08 | 1.10 | 1.07 |
| 37 | 2.75 | 2.82 | 2.61 | 2.99 | 3.09 | 2.89 | 3.11 | 3.24 | 3.04 | 3.28 | 3.41 | 3.23 | 2.41 | 2.53 | 2.41 |
| 38 | 4.41 | 4.25 | 4.34 | 4.86 | 4.76 | 4.83 | 5.00 | 4.96 | 4.92 | 5.45 | 5.40 | 5.36 | 3.91 | 3.80 | 3.84 |
| 39 | 1.98 | 1.96 | 1.93 | 2.00 | 1.98 | 2.04 | 2.10 | 1.98 | 2.00 | 1.89 | 1.79 | 1.94 | 2.09 | 2.00 | 2.02 |
| *Prediction set* | | | | | | | | | | | | | | | |
| 40 | | 1.56 | 1.67 | 1.73 | 1.81 | 1.89 | 1.85 | 1.99 | 2.07 | 2.02 | 2.17 | 2.21 | 1.22 | 1.32 | 1.48 |
| 41 | 3.61 | 3.50 | 3.43 | 3.58 | 3.46 | 3.65 | 3.89 | 3.64 | 3.75 | 3.43 | 3.56 | 3.87 | | 3.32 | 3.28 |
| 42 | 3.08 | 2.55 | 2.77 | 2.81 | 2.66 | 2.90 | 2.81 | 2.67 | 2.85 | 2.81 | 2.57 | 2.81 | | 2.54 | 2.74 |
| 43 | 2.95 | 2.67 | 2.99 | 3.03 | 2.66 | 3.03 | 3.09 | 2.71 | 2.97 | 2.97 | 2.51 | 2.86 | | 2.56 | 2.90 |
| 44 | 3.72 | 3.52 | 3.47 | 3.71 | 3.45 | 3.63 | 3.71 | 3.56 | 3.65 | 3.65 | 3.45 | 3.68 | | 3.30 | 3.32 |
| 45 | 1.26 | 1.24 | 1.21 | 1.61 | 1.63 | 1.63 | 1.93 | 1.89 | 1.92 | 2.18 | 2.15 | 2.14 | 0.89 | 0.94 | 0.91 |
| 46 | 2.14 | 2.12 | 2.09 | 2.76 | 2.67 | 2.68 | 3.15 | 2.96 | 3.07 | 3.55 | 3.39 | 3.50 | 1.82 | 1.73 | 1.72 |
| 47 | 2.20 | 2.28 | 2.17 | 2.38 | 2.48 | 2.33 | 2.40 | 2.53 | 2.35 | 2.46 | 2.62 | 2.42 | 2.03 | 2.10 | 1.96 |
| 48 | 2.83 | 2.80 | 2.71 | 3.10 | 3.10 | 2.97 | 3.13 | 3.16 | 2.96 | 3.29 | 3.35 | 3.14 | 2.62 | 2.56 | 2.44 |
| 49 | 2.13 | 2.26 | 2.02 | 2.29 | 2.44 | 2.21 | 2.40 | 2.58 | 2.34 | 2.47 | 2.65 | 2.42 | 1.85 | 2.04 | 1.81 |
| 50 | 2.96 | 3.00 | 2.80 | 3.22 | 3.30 | 3.13 | 3.35 | 3.46 | 3.27 | 3.55 | 3.66 | 3.51 | 2.60 | 2.69 | 2.60 |
| 51 | 3.99 | 3.89 | 3.97 | 4.39 | 4.34 | 4.44 | 4.53 | 4.53 | 4.54 | 4.91 | 4.90 | 4.93 | 3.54 | 3.49 | 3.56 |
| 52 | 2.68 | 2.43 | 2.57 | 3.07 | 2.78 | 2.91 | 2.95 | 2.70 | 2.86 | 3.17 | 2.93 | 3.04 | 3.07 | 2.74 | 2.87 |
| 53 | 2.67 | 2.37 | 2.66 | 2.63 | 2.33 | 2.57 | 2.48 | 2.19 | 2.45 | 2.29 | 2.10 | 2.32 | 2.49 | 2.25 | 2.63 |
| 54 | 2.38 | 2.34 | 2.51 | 2.45 | 2.42 | 2.63 | 2.55 | 2.43 | 2.58 | 2.39 | 2.30 | 2.53 | | 2.34 | 2.52 |

[a] Definition of the stationary phases is given in Table 4.

The calculated values of log $K_L$ using the generated MLR models are given in Table 5 for all of the molecules included in the training and the prediction sets. These values are given for different stationary phases in this table. Comparison of the calculated and the experimental values reveals that a good agreement exists between them.

### 3.2. Analysis of the artificial neural network

The ANN was generated by using the descriptors appearing in the MLR models as inputs. A 6-6-5 neural network was developed with the optimum momentum and learning rate of 0.4 and 0.2, respectively. In order to prevent the overfitting, the mean square errors (MSEs) for the training and the prediction sets were plotted against the number of iterations (Fig. 1). The overfitting will start after 35500 training of the network. The ANN calculated values of log $K_L$ for the training and the prediction sets on different columns are included in Table 5.

To evaluate the neural network, the MSEs of its results for the training and the prediction sets are compared with the MSEs of the regression models for different stationary phases in Table 6. Comparison of the MSEs shows the superiority of the ANN model over that of the MLRs. This indicates that some of the descriptors appearing in the MLR models interacts with each other and on the whole the retention behavior of the molecules on different columns show some nonlinear characteristics.

Table 6
Comparison of the MSEs for the results obtained using the ANN and the regression models

| Column[a] | MSE | | | |
|---|---|---|---|---|
| | Training | | Prediction | |
| | MLR | ANN | MLR | ANN |
| EGAD | 0.020 | 0.005 | 0.021 | 0.010 |
| THPED | 0.024 | 0.005 | 0.016 | 0.005 |
| U50HB | 0.022 | 0.006 | 0.018 | 0.005 |
| DEHPA | 0.021 | 0.004 | 0.019 | 0.010 |
| QBES | 0.027 | 0.007 | 0.012 | 0.009 |

[a] Definition of the columns is given in Table 4.

The calculated ANN values of log $K_L$ of the prediction set are plotted against the experimental values for different columns in Fig. 2. As shown in this figure, all values fit the regression lines indicating the ability of the ANN in predicting of the retention behavior of organic compounds on the commonly used stationary phases. Fig. 3 shows the propagation of residuals. Since the residuals are propagated on both sides of the zero line, there is no systematic error in developing of the ANN model. In order to compare the ANN results with the results obtained by using the solvatochromic model and the

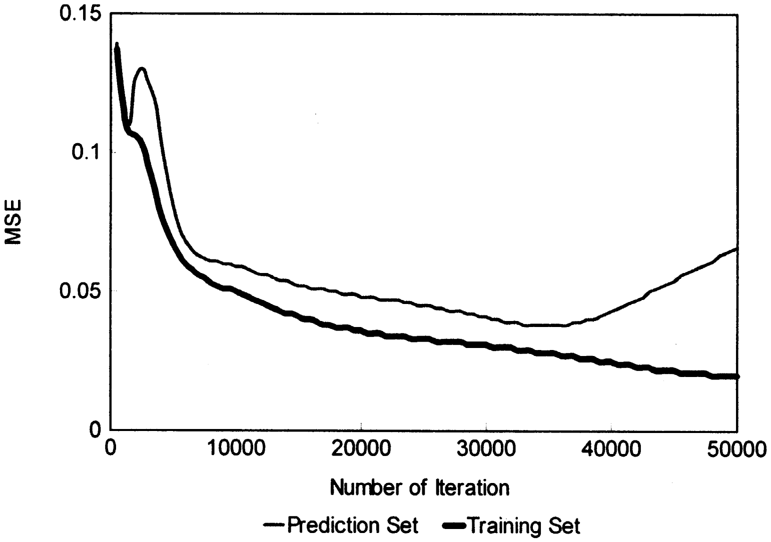

Fig. 1. Variations of MSE vs. the number of iterations for the training and the prediction sets.

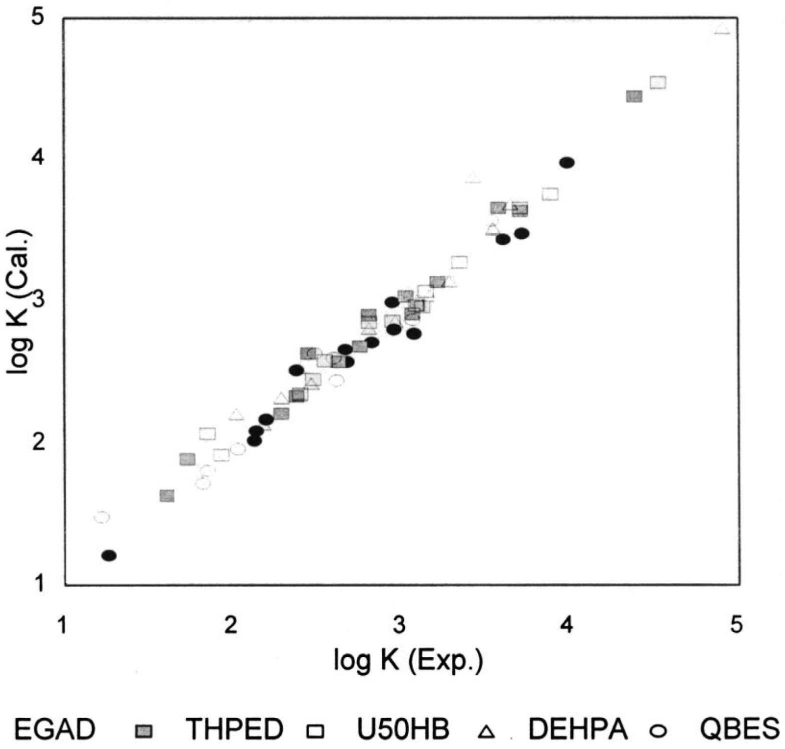

Fig. 2. Plot of the calculated values of the log $K_L$ for the prediction set against the experimental values.

MLR models generated in this work, the correlation coefficients between the calculated and the experimental values are given in Table 7. Inspection of these results indicate the superiority of the ANN model over that of the MLR model.

From the results obtained in this paper one may

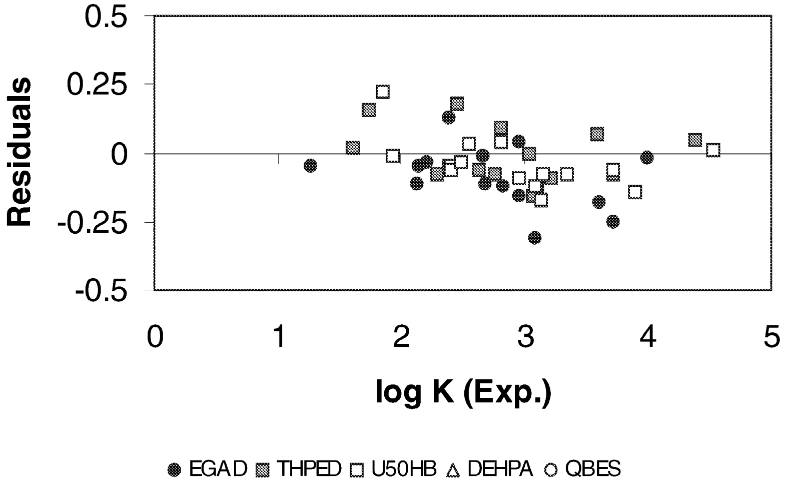

Fig. 3. Plot of residuals vs. experimental values of log $K_L$.

Table 7
Correlation coefficients between the experimental and the calculated values of log $K_L$ for different columns

| Column | Training | | Prediction | |
|---|---|---|---|---|
| | *r* | SE | *r* | SE |
| EGAD | 0.993 | 0.087 | 0.988 | 0.118 |
| THPED | 0.992 | 0.094 | 0.990 | 0.105 |
| U50HB | 0.990 | 0.102 | 0.993 | 0.091 |
| DEHPA | 0.995 | 0.081 | 0.981 | 0.153 |
| QBES | 0.988 | 0.110 | 0.986 | 0.144 |

conclude that the parameters of DIPOL, HOMO, PCHNEG, PCHPOSH, VOLUME and $M_r$ can be considered as comprehensive descriptors for predicting of the partition coefficient of a variety of molecules on different columns. Also, the experimental solvatochromic parameters proposed by Abraham can be replaced by the calculated descriptors developed in this work [20].

## Acknowledgements

The authors wish to acknowledge Professor C.F. Poole for allowing use of his experimental results in this work. We also acknowledge M.H. Fatemi for his helpful discussion.

## References

[1] L. Rohrschneider, J. Chromatogr. 22 (1966) 6.

[2] L. Rohrschneider, J. Chromatogr. 39 (1969) 383.

[3] W.O. McReynolds, J. Chromatogr. Sci. 8 (1970) 685.

[4] L.R. Snyder, J. Chromatogr. 92 (1974) 223.

[5] L.R. Snyder, J. Chromatogr. Sci. 16 (1978) 223.

[6] M.B. Evans, J.K. Haken, T. Toth, J. Chromatogr. 351 (1986) 155.

[7] M.B. Evans, J.K. Haken, J. Chromatogr. 406 (1987) 105.

[8] W. Burns, S.J. Hawkes, J. Chromatogr. Sci. 15 (1997) 185.

[9] E. Chong, B. de Bricero, G. Miller, S.J. Hawkes, Chromatographia 20 (1985) 293.

[10] R.A. Keller, J. Chromatogr. Sci. 11 (1973) 49.

[11] E. Fernandez-Sanchez, A. Fernandez-Torres, J.A. Garcia-Dominguez, J.M. Santiuste, E. Pertierra-Rimda, J. Chromatogr. 457 (1988) 55.

[12] J.E. Brady, D. Bjorkman, C.D. Herter, P.W. Carr, Anal. Chem. 56 (1984) 278.

[13] S.K. Polle, P.H. Shetty, C.F. Poole, Anal. Chim. Acta 218 (1989) 241.

[14] C.F. Poole, R.M. Pomaville, T.A. Dean, Anal. Chim. Acta 225 (1989) 193.

[15] J.A. Garcia-Dominguez, J.M. Santiuste, Q. Dai, J. Chromatogr. A 787 (1997) 145.

[16] W.J. Cheong, J.D. Choi, Anal. Chim. Acta 342 (1997) 51.

[17] K.H. Lamparczy, A. Radeck, Chromatographia 18 (1984) 615.

[18] K.H. Lamparczy, Chromatographia 20 (1985) 223.

[19] S.K. Poole, C.F. Poole, Analyst 120 (1995) 289.

[20] M.H. Abraham, Chem. Soc. Rev. 22 (1993) 73.

[21] M.H. Abraham, J. Phys. Org. Chem. 6 (1993) 660.

[22] G. Park, C.F. Poole, J. Chromatogr. A 726 (1996) 141.

[23] J. Li, A.J. Dollas, P.W. Carr, J. Chromatogr. 517 (1990) 103.

[24] T.O. Kolie, C.F. Polle, J. Chromatogr. 556 (1991) 547.

[25] T.O. Kollie, C.F. Poole, M.H. Abraham, G.S. Witing, Anal. Chim. Acta 259 (1992) 1.

[26] MOPAC Package, Version 6, US Air Force Academy, Colorado Spring, CO, 1990.

[27] F. Parastar, Computer Modeling of Biological Activity for Some Benzoic Acid Derivatives, M.Sc. Thesis, Shahid Bahonar University of Kerman, Kerman, 1995.

[28] SPSS for Windows, Statistical Package for IBM PC, SPSS Inc., 1993 (http://www.spss.com).

[29] J. Zupan, J. Gasteiger, Neural Networks for Chemist – An Introduction, VCH, Weinheim, 1993.

[30] M.T. Hagan, H.B. Demuth, M. Beal, Neural Network Design, PWS, Boston, MA, 1996.

[31] N.K. Bose, P. Liang, Neural Network, Fundamentals, McGraw-Hill, New York, 1996.

[32] M. Jalali-Heravi, F. Parastar, J. Chem. Inf. Comput. Sci. 40 (2000) 147.

[33] M. Jalali-Heravi, M.H. Fatemi, J. Chromatogr. A 825 (1998) 161.

[34] T.O. Kollie, C.F. Poole, J. Chromatogr. 550 (1990) 213.

[35] C.F. Poole, T.O. Kollie, S.K. Poole, Chromatographia 34 (1992) 281.