Semi-Empirical Topological Method for Prediction of the Relative Retention Time of Polychlorinated Biphenyl Congeners on 18 Different HR GC Columns

CHROMATOGRAPHIA
2010, *72*, *523–533*

Raouf Ghavami✉, S. Mohammad Sajadi

Department of Chemistry, Faculty of Science, University of Kurdistan, P. O. Box 416, Sanandaj, Iran;
E-Mail: rghavami2000@yahoo.com; rghavami@uok.ac.ir

Received: 17 April 2010 / Revised: 20 June 2010 / Accepted: 14 July 2010
Online publication: 10 August 2010

Abstract

High resolution gas chromatographic relative retention time (HRGC-RRT) models were developed to predict relative retention times of the 209 individual polychlorinated biphenyls (PCBs) congeners. To estimate and predict the HRGC-RRT values of all PCBs on 18 different stationary phases, a multiple linear regression equation of the form RRT = a_o + a_1 (no. *o*-Cl) + a_2 (no. *m*-Cl) + a_3 (no. *p*-Cl) + a_4 (V_M or S_M) was used. Molecular descriptors in the models included the number of *ortho*-, *meta*-, and *para*-chlorine substituents (no. *o*-Cl, *m*-Cl and *p*-Cl, respectively), the semi-empirically calculated molecular volume (V_M), and the molecular surface area (S_M). By means of the final variable selection method, four optimal semi-empirical descriptors were selected to develop a QSRR model for the prediction of RRT in PCBs with a correlation coefficient between 0.9272 and 0.9928 and a leave-one-out cross-validation correlation coefficient between 0.9230 and 0.9924 on each stationary phase. The root mean squares errors over different 18 stationary phases are within the range of 0.0108–0.0335. The accuracy of all the developed models were investigated using cross-validation leave-one-out (LOO), *Y*-randomization, external validation through an odd–even number and division of the entire data set into training and test sets.

Keywords

Gas chromatography
Quantitative structure retention relationship
Environmental analysis
Molecular descriptors and congener identification
Polychlorinated biphenyls

Introduction

Polychlorinated biphenyls (PCBs) are a class of discrete organic compounds with one to ten chlorine atoms attached to a biphenyl nucleus and a general chemical formula of $C_{12}H_{10-n}Cl_n$, where $n = 1–10$ [1]. A general chemical structure of polychlorinated biphenyls is shown in Fig. 1. There are 209 theoretically possible congeners subdivided into ten homologue groups with 1–46 congeners in each. PCBs of a given homolog with different chlorine substitution position are called "isomers". The degree of chlorination varies depending on the reaction conditions, and ranges from 19 to 69% (*w*/*w*). The composition of PCBs is summarized in Table 1 [2]. All congeners have been assigned a systematic number from 1 to 209 corresponding to a specific substitution pattern. The initial scheme was proposed by Ballschmiter and Zell [3] and revised by Guitart et al. [4].

PCBs are hydrophobic compounds with low volatility, and the highly chlorinated ones have poor water solubility. Moreover, they are resistant to acids, bases, and (generally) environmental

Electronic supplementary material The online version of this article (doi:10.1365/s10337-010-1696-5) contains supplementary material, which is available to authorized users.

Original
DOI: 10.1365/s10337-010-1696-5
0009-5893/10/09

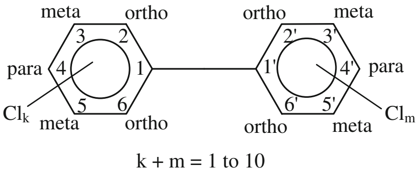

Fig. 1. Generic polychlorinated biphenyl (PCB) chemical structure

degradation processes. They are, therefore, highly persistent in the environment. They have good electrical insulation properties with high thermal conductivity, low flammability and high resistance to thermal degradation [2], and have therefore been widely used as heat transfer fluids and dielectric fluids in electric transformers and large capacitors. They have also been used as organic diluents, plasticizers, additives in pesticides, carbonless copy paper, paint additives, hydraulic fluids, lubricants and many other applications [1, 5–8].

Since PCBs occur as complex mixtures of up to 209 distinct congeners, their properties are dependent on the composition of the mixture. The properties of individual PCB congeners also vary according to the degree of chlorination and location of the chlorine atoms. Generally, the water solubility and vapor pressure decrease as the degree of substitution increases, and the lipid solubility increases with increasing chlorine substitution. PCBs in the environment may be expected to associate with the organic components of soils, sediments, and biological tissues, or with dissolved organic carbon in aquatic systems, rather than being in solution in water [9–11].

The toxicity of a PCB is dependent not only on the number of chlorine atoms present in the biphenyl structures, but also on the positions of the chlorine atoms. PCBs with many chlorines in *ortho*-position (nonco-planar) induce phenobarbital-type responses, while PCBs that lack chlorine atoms in the *ortho*-position but have chlorine atoms in both *para*-positions (4 and 4′) and at least one in the *meta*-position (3, 5, 3′, 5′) (coplanar), can rotate freely around the phenyl–phenyl (1,1′ bond). This means that they can exhibit structural resemblance to the dioxins, i.e. a relatively planar structure (coplanar PCBs), and may hence induce methylcholanthrene-type responses and dioxin-like toxicity [12–14]. Most PCB congeners do not exhibit dioxin-like toxicity but have a different toxicological profile [13, 15]. Co-planar PCBs, like dioxins and furans, bind to the AL-receptor and may exert, thus, dioxin-like effects, in addition to AL-receptor independent effects, which they share with non-coplanar PCBs (e.g. tumor promoter) [13]. Association between elevated exposure to PCB mixtures and alterations in liver enzymes, hepatomegaly, and dermatological effects such as rashes and acne has been reported [13].

Due to PCBs' complex composition, many researchers have also placed emphasis on the identification of the individual PCB congeners [16, 17]. Presently, all 209 PCB congeners have been commercially synthesized and are available for use as standards, and because of advances in high resolution gas chromatography (HRGC), it is possible to determine most of the individual PCB congeners in the environmental samples. However, separation and characterization of all 209 PCB congeners is still an extremely difficult (if not impossible) task that attracts on-going research focus in this field [18].

One of the most successful approaches to the prediction of physicochemical and biological properties of organic molecules, starting only from molecular structure information, is quantitative structure–property/activity relationships modeling (QSPRs/QSARs) [19]. QSPRs/QSARs are mathematical models that attempt to correlate the molecular structure of compounds and their biological, chemical, and physical properties. Among the most extensively studied properties are the chromatographic ones. It is considered that the same basic intermolecular interactions determine the behavior of chemical compounds in both biological and chromatographic environments [20]. Retention in chromatography is the result of a competitive distribution process of the solute between mobile and stationary phases, in which, the partitioning of the solute between these phases is largely determined by the molecular structure [21]. Predicting chromatographic behavior from molecular structure of solutes resulted in the quantitative structure-retention relationships (QSRR) methodology. QSRR are statistically derived relationships between the chromatographic parameters determined for a representative series of analytes in given separation systems and the molecular descriptors accounting for the structural differences among the investigated analytes [22]. Such relationships may provide insight into the molecular mechanism of separation in a given chromatographic system, generate knowledge about the various interactions taking place between the solute and the stationary phase, evaluate physicochemical properties of analytes and identify the most informative structural descriptors. Due to the need to control the PCBs level in the environment, the analytical methods for their analysis are currently based on their separation by GC [23] using on capillary columns with different polarities [24, 25] and specific detectors such as the flame ionization detector (FID) [26], the electron-capture detector (ECD) [27, 28] and mass spectrometry (MS) [29, 30]. The HRGC-RRT on a capillary column with detection by ECD is a unique characteristic of the PCBs and can be used for the identification purpose. Despite the broad range of GC stationary phases available, none can separate all PCBs from each other. Various techniques have been used based on the combination of commercially available GC columns to improve the separation efficiency of PCBs [31–34]. At present, a database of RRTs and coelutions for all 209 congeners on 20 different stationary phases with MS or ECD detection has been reported [17, 35].

In the past, several attempts have been made to build QSRR models on the prediction of RRT for PCBs on different stationary phases [36–43]. Hasan and Jurs [37] used the five-variable regression equation for prediction of GC-RRT of 209 PCBs with $R^2 = 0.997$ and standard deviation of 0.017. Liu et al. [38] used the five-variable regression equation with $R^2 = 0.9928$ and the root mean square errors of 0.0152 based on molecular electronegativity distance vector (MEDV) descriptors to correlate with the GC-RRTs of 209 PCB congeners on the SE-54 stationary phase. Ren et al. [41] using CODESSA software package and principal component analyzed (PCA) presented a QSRR study for the GC–GC–TOFMS (time-flight mass spectrometry) chromatographic relative retention time of 209 PCB congeners. PCA was used to recognize groups of samples with similar behavior and assist the separation of the data into training and test sets. Jäntschi et al. [42] reported the use of a molecular descriptors family (MDF) in QSRR modeling to predict the chromatographic relative retention times of 209 PCBs on a capillary column of SE-54.

In the preceding paper [43], several estimation models derived from the HRGC-RRT values of all 209 PCB congeners on the 18 stationary phases by a GA-based best multiple linear regression analysis with four optimal descriptors were proposed. The present work is focused on the successful application of a semi-empirical topological method for the prediction of the HRGC-RRTs 209 PCBs congener's values on the same stationary phases. Selecting some semi-empirical chemical descriptors such as molecular volume (V_M) and molecular surface area (S_M) as well as the number of substituted chlorine atoms as descriptors, the QSRR models correlating to the RRTs of 209 PCB congeners are developed using the elimination selection stepwise best multiple linear regression (BMLR) analysis. It has been found that the QSRR models have not only high estimation qualities and high stabilities but also good predictive potentials.

Experimental

Experimental Database Set

The observed HRGC-RRTs of 209 PCBs on 18 different stationary phases, 30m DB1, 30m SPB-Octyl, 60m SPB-Octyl, 100m CP-Sil5-C18, 30m DB5-MS, 60m RTX-5, 50m CP-Sil-13, 30m SPB-20, 30m HP-35, 60m RTX-35, 30m DB-17, 60m HP-1301, 30m DP-XLB, 30m DB-35-MS, 50m HT-8, 30m Apiezon L, 30m CNBP#2, 48m 007-23, reported by Frame [17, 35] served as experimental data in this study and the HRGC-RRTs designed as dependent variables. The structures of the PCBs together with their relative retention time values are listed in Table S1 (see Supplementary Material).

Table 1. Composition of PCBs by homologs

Homolog	Molecular formula	Chlorine (% by weight)	Number of isomers
Monochlorobiphenyl	$C_{12}H_9Cl$	19	3
Dichlorobiphenyl	$C_{12}H_8Cl_2$	32	12
Trichlorobiphenyl	$C_{12}H_7Cl_3$	41	24
Tetrachlorobiphenyl	$C_{12}H_6Cl_4$	49	42
Pentachlorobiphenyl	$C_{12}H_5Cl_5$	54	46
Hexachlorobiphenyl	$C_{12}H_4Cl_6$	59	42
Heptachlorobiphenyl	$C_{12}H_3Cl_7$	63	24
Octachlorobiphenyl	$C_{12}H_2Cl_8$	66	12
Nonachlorobiphenyl	$C_{12}HCl_9$	69	3
Decachlorobiphenyl	$C_{12}Cl_{10}$	71	1
Total congeners	–	–	209

Computer Hardware and Software

All calculations were run on a Pentium IV personal computer (CPU at 2.6 MB) under Windows XP operating system. The ISIS/Draw version 2.3 software was used for drawing the molecular structures [44]. Molecular modeling and geometry optimization were employed by Hyperchem (version 7.1, HyperCube) [45]. Dragon software [46] was employed for calculation of theoretical molecular descriptors. SPSS software (version 13.0, SPSS) (http://www.spss.com/) was used for stepwise MLR analysis and other calculations were performed in the MATLAB (version 7.0, Math Works) environment.

Descriptor Generation

To obtain QSRR models, PCB congeners must be represented using molecular descriptors. Descriptors are generated solely from the molecular structures and aimed to numerically encode meaningful features of each molecule. The calculation process of the molecular descriptors is described as below: all the two-dimensional structures of the molecules were drawn using ISIS/Draw 2.3 program [44]. Then the 3D geometry structures of the molecules were pre-optimized using MM+ molecular mechanics force filed and precisely optimized with semi-empirical AM1 method implemented in HyperChem software package (Hypercube, version 7.0) [45]. All calculations were carried out at restricted Hartree–Fock level without configuration interaction. The molecular structures were optimized using the Polak–Ribiere algorithm until the root mean square gradient was 0.01 kcal Å mol^{-1} (200 K, gas phase). In order to prevent the structures locating at local minima, geometry optimization was run many times with different stating points for each molecule.

During this investigation, 22 molecular descriptors were studied to characterize all 209 PCB congeners. Among these, six quantum chemical descriptors, like the dipole moment of the molecules at X, Y and Z directions (μ), the standard heat of formation (ΔH_f°), the energy of the highest occupied molecular orbital (E_{HOMO}), the energy of the lowest unoccupied molecular orbital (E_{LUMO}), E_{HOMO}^2, E_{LUMO}^2; 11 theoretical descriptors, molar mass (M_w), the natural logarithm of molar mass ($\ln M_w$), molecular volume (V_M), molecular surface area (S_M), molecular polarizability (α), hydration energy (H_e), total energy of the molecule (E_{tot}), bending energy (E_{ben}), electronic energy (E_{ele}), nuclear energy (E_{nuc}), molar refractivity (R_f); and the other five descriptors, the number of *ortho*-substituted chlorines (o-Cl), *meta*-substituted chlorines (m-Cl) and *para*-substituted chlorines (p-Cl), the root square of the number of chlorine substituents ($Cl^{1/2}$) and the square number of chlorine substituents (Cl^2), belonged to topological-type ones; were calculated by the HyperChem and Dragon programs. The number of these descriptors presented the electronic

Table 2. Correlation matrix for the inter-correlation of various semi-imperial descriptors of 209 PCBs congeners

	No. *o*-Cl	No. *m*-Cl	No. *p*-Cl	S_M	V_M
No. *o*-Cl	1.0000				
No. *m*-Cl	−0.0038	1.0000			
No. *p*-Cl	−0.0182	−0.0116	1.0000		
S_M	0.4062	0.7467	0.4878	1.0000	
V_M	0.5207	0.7041	0.4556	0.9891	1.0000

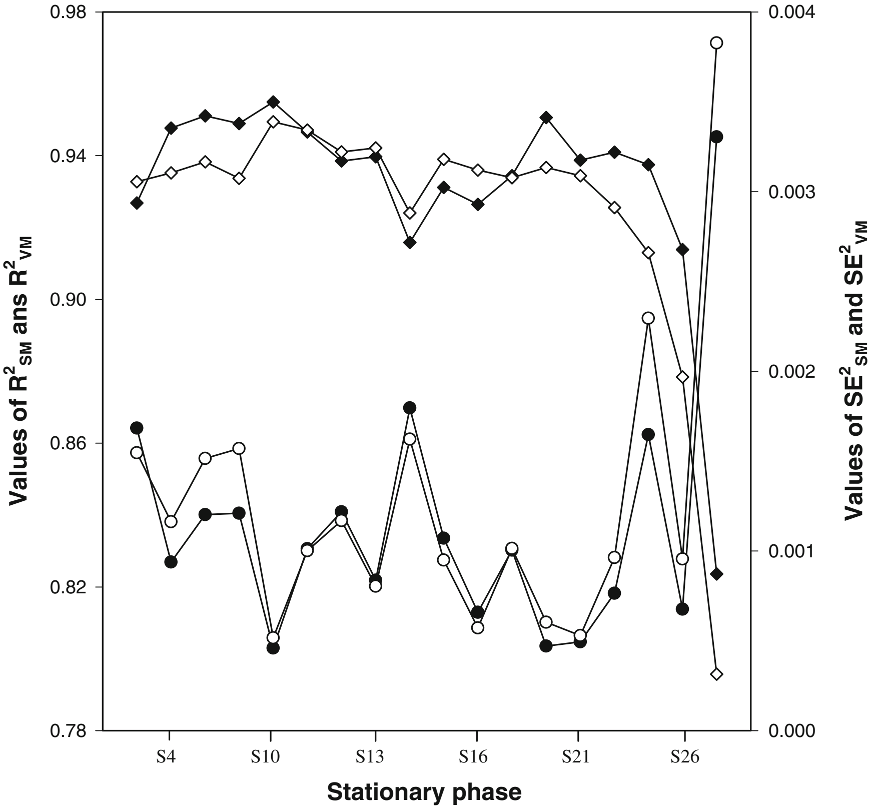

Fig. 2. The influences of the molecular surface area (S_M) and molecular volume (V_M) descriptors on the correlation coefficient (R^2) and the standard deviation (SE^2) for the resulted 18 univariate models versus stationary phases

structure-related information. For instance, E_{HOMO} imply the electron-donating ability, E_{LUMO} denote electron-acceptance ability, μ refers to molecular dipolar property, ΔH_f° is the heat released or absorbed (enthalpy change) during the formation of a pure substance from its elements at constant pressure. Some other descriptors revealed the molecular size and topological information such as M_w, *o*-Cl, *m*-Cl, and *p*-Cl. In addition, the logarithm transferation and square along with root square operations are to calculate a number of nonlinear components in linear model.

Results and Discussions

The calculated semi-empirical topological descriptors were collected in a data matrix (***D***) whose number of rows and columns were the number of molecules and descriptors. At the beginning, in order to minimize the information overlapping in descriptors and to reduce the number of descriptors required in regression equation, the concept of non-redundant descriptors (NRD) [47] was used in our study. That is, when two descriptors are correlated by a linear correlation coefficient value greater than 0.85, both descriptors are correlated with the dependent variables and the better correlation is used for the actual analysis, leaving out the descriptors showing a lower correlation. This objective-based feature selection left reduced and predictive descriptors for the studied compounds. By using these criteria, 17 out of 22 original descriptors were eliminated. These descriptors can give some information on the affecting degree for RRTs of different descriptors and well understanding the correlation between the experimental and calculated values. Therefore, *o*-Cl, *m*-Cl, *p*-Cl, and one of S_M or V_M set of descriptors has been used in the QSRR models for RRTs prediction on all capillary stationary phases.

Linear Regression of RRTs with S_M and V_M

To reduce redundancy in the descriptor data matrix, correlation of descriptors with each other and with the RRT values of the PCB congeners was examined to detect collinear descriptors (i.e. $r > 0.9$). The correlation coefficient matrix for the descriptors used in this study, is listed in Table 2. Among these descriptors, molecular volume (V_M) and molecular surface area (S_M) are presenting the molecular size. They are highly correlated with each other. The correlation coefficient amounts to as high as 0.9891 for the present set of 209 PCBs congeners. Consequently, at the start, the correlation of the semi-empirical topological descriptors S_M and V_M along with the RRT of all PCB congeners of each stationary phase has been studied. The influences of the S_M and V_M descriptors on the correlation coefficient (R^2) and the standard deviation (SE^2) for the resulted 18 univariate models versus stationary phases are plotted in Fig. 2. Obviously, low to very high correlation coefficients were obtained for both semi-empirical topological descriptors. As can be shown, for

Table 3. QSRR models and statistical parameters of GC-RRT values versus (52 + 180) for the total sets ($n = 209$) of polychlorinated biphenyl (PCB) congeners on 18 GC capillary columns

System no.	Models of the training sets	R^2	RMS	REP	F	R^2_{cv}	$\mathrm{RMS}_{\mathrm{CV}}$	$R^2_{\max}$
S1	RRT = 1.3633(±0.0987) + 0.0930(±0.0031) no. *o*-Cl + 0.1734(±0.0054) no. *m*-Cl + 0.1825(±0.0056) no. *p*-Cl − 0.0037(±0.0003) S_{M}	0.9885	0.0134	2.9204	4404.6	0.9879	0.0137	0.0012
S4	RRT = 0.6151(±0.0753) + 0.0577(±0.0024) no. *o*-Cl + 0.1194(±0.0041) no. *m*-Cl + 0.1296(±0.0043) no. *p*-Cl − 0.0014(±0.0002) S_{M}	0.9915	0.0108	2.3345	5923.3	0.9910	0.0111	0.0017
S6	RRT = 0.5652(±0.0809) + 0.0659(±0.0025) no. *o*-Cl + 0.1363(±0.0044) no. *m*-Cl + 0.1480(±0.0046) no. *p*-Cl − 0.0015(±0.0002) S_{M}	0.9928	0.0113	2.4726	7032.9	0.9924	0.0116	0.0020
S8	RRT = 0.5610(±0.0841) + 0.0632(±0.0027) no. *o*-Cl + 0.1342(±0.0046) no. *m*-Cl + 0.1447(±0.0048) no. *p*-Cl − 0.0015(±0.0002) S_{M}	0.9919	0.0119	2.5939	6260.4	0.9914	0.0122	0.0014
S10	RRT = 0.3320(±0.0826) + 0.0387(±0.0026) no. *o*-Cl + 0.0737(±0.0045) no. *m*-Cl + 0.0792(±0.0047) no. *p*-Cl − 0.0003(±0.0002) S_{M}	0.9819	0.0115	2.2614	2773.7	0.9809	0.0118	0.0015
S11	RRT = 0.7229(±0.0915) + 0.0680(±0.0029) no. *o*-Cl + 0.1273(±0.0050) no. *m*-Cl + 0.1354(±0.0052) no. *p*-Cl − 0.0018(±0.0003) S_{M}	0.9881	0.0135	2.8985	4221.0	0.9874	0.0138	0.0016
S12	RRT = 0.8629(±0.1070) + 0.0737(±0.0034) no. *o*-Cl + 0.1367(±0.0058) no. *m*-Cl + 0.1460(±0.0061) no. *p*-Cl − 0.0022(±0.0003) S_{M}	0.9844	0.0158	3.4007	3213.7	0.9835	0.0162	0.0015
S13	RRT = 0.6793(±0.1005) + 0.0585(±0.0032) no. *o*-Cl + 0.1074(±0.0055) no. *m*-Cl + 0.1153(±0.0057) no. *p*-Cl − 0.0015(±0.0003) S_{M}	0.9803	0.0147	3.1238	2543.4	0.9793	0.0151	0.0022
S14	RRT = 1.4366(±0.1217) + 0.0929(±0.0038) no. *o*-Cl + 0.1712(±0.0066) no. *m*-Cl + 0.1818(±0.0069) no. *p*-Cl − 0.0038(±0.0003) S_{M}	0.9813	0.0179	3.8538	2670.8	0.9802	0.0183	0.0023
S15*	RRT = 0.8083(±0.1217) + 0.0669(±0.0038) no. *o*-Cl + 0.1184(±0.0066) no. *m*-Cl + 0.1277(±0.0070) no. *p*-Cl − 0.0019(±0.0003) S_{M}	0.9743	0.0179	3.7843	1926.2	0.9729	0.0183	0.0023
S16	RRT = 0.6064(±0.1075) + 0.0480(±0.0034) no. *o*-Cl + 0.0822(±0.0059) no. *m*-Cl + 0.0897(±0.0061) no. *p*-Cl − 0.0011(±0.0003) S_{M}	0.9651	0.0151	3.1515	1407.8	0.9630	0.0155	0.0019
S17	RRT = 0.9159(±0.0925) + 0.0658(±0.0029) no. *o*-Cl + 0.1263(±0.0050) no. *m*-Cl + 0.1337(±0.0053) no. *p*-Cl − 0.0022(±0.0003) S_{M}	0.9849	0.0136	2.9117	3333.5	0.9841	0.0140	0.0020
S20	RRT = 0.2871(±0.0865) + 0.0332(±0.0027) no. *o*-Cl + 0.0689(±0.0047) no. *m*-Cl + 0.0750(±0.0049) no. *p*-Cl − 0.0001(±0.0002) S_{M}	0.9788	0.0121	2.3845	2351.6	0.9775	0.0125	0.0019
S21	RRT = 0.4342(±0.0934) + 0.0373(±0.0029) no. *o*-Cl + 0.0709(±0.0051) no. *m*-Cl + 0.0771(±0.0053) no. *p*-Cl − 0.0006(±0.0003) S_{M}	0.9708	0.0131	2.7321	1692.4	0.9690	0.0134	0.0016
S22	RRT = 0.4503(±0.0992) + 0.0438(±0.0031) no. *o*-Cl + 0.0935(±0.0054) no. *m*-Cl + 0.1009(±0.0057) no. *p*-Cl − 0.0008(±0.0003) S_{M}	0.9795	0.0143	3.0452	2430.7	0.9782	0.0147	0.0018
S24	RRT = 0.6231(±0.0936) + 0.0625(±0.0030) no. *o*-Cl + 0.1454(±0.0051) no. *m*-Cl + 0.1587(±0.0053) no. *p*-Cl − 0.0017(±0.0003) S_{M}	0.9910	0.0123	2.7188	5624.4	0.9905	0.0127	0.0012
S26	RRT = 0.3447(±0.0999) + 0.0241(±0.0031) no. *o*-Cl + 0.0672(±0.0054) no. *m*-Cl + 0.0729(±0.0057) no. *p*-Cl − 0.0003(±0.0001) S_{M}	0.9657	0.0141	2.9458	1436.5	0.9636	0.0145	0.0017
S27	RRT = 1.5251(±0.2246) + 0.0692(±0.0071) no. *o*-Cl + 0.1696(±0.0123) no. *m*-Cl + 0.1787(±0.0128) no. *p*-Cl − 0.0039(±0.0006) S_{M}	0.9272	0.0335	7.2682	649.9	0.9230	0.0345	0.0030

* For system 15 $n = 208$

nine stationary phases (S4, S6, S8, S10, S20, S22, S24, S26, and S27), the S_{M} produced a better model as the V_{M}, while for four stationary phases (S1, S14, S15, and S16) the V_{M} produced a better model as the S_{M} but for other stationary phases (S11, S12, S13, S17, and S21), the correlation coefficient (R^2) and the standard deviation (SE^2) are almost equally. Being colinear of both descriptors (S_{M} and V_{M}), to derive the best QSRR models for all stationary phases, caused to eliminate one of them from the selected set of descriptors. According to the obtained results, the molecular volume semi-empirical descriptor should be removed. Therefore, the ability of the resulting QSRR regression models to enable accurate prediction of the relative retention time is not related to colinearity between the variables.

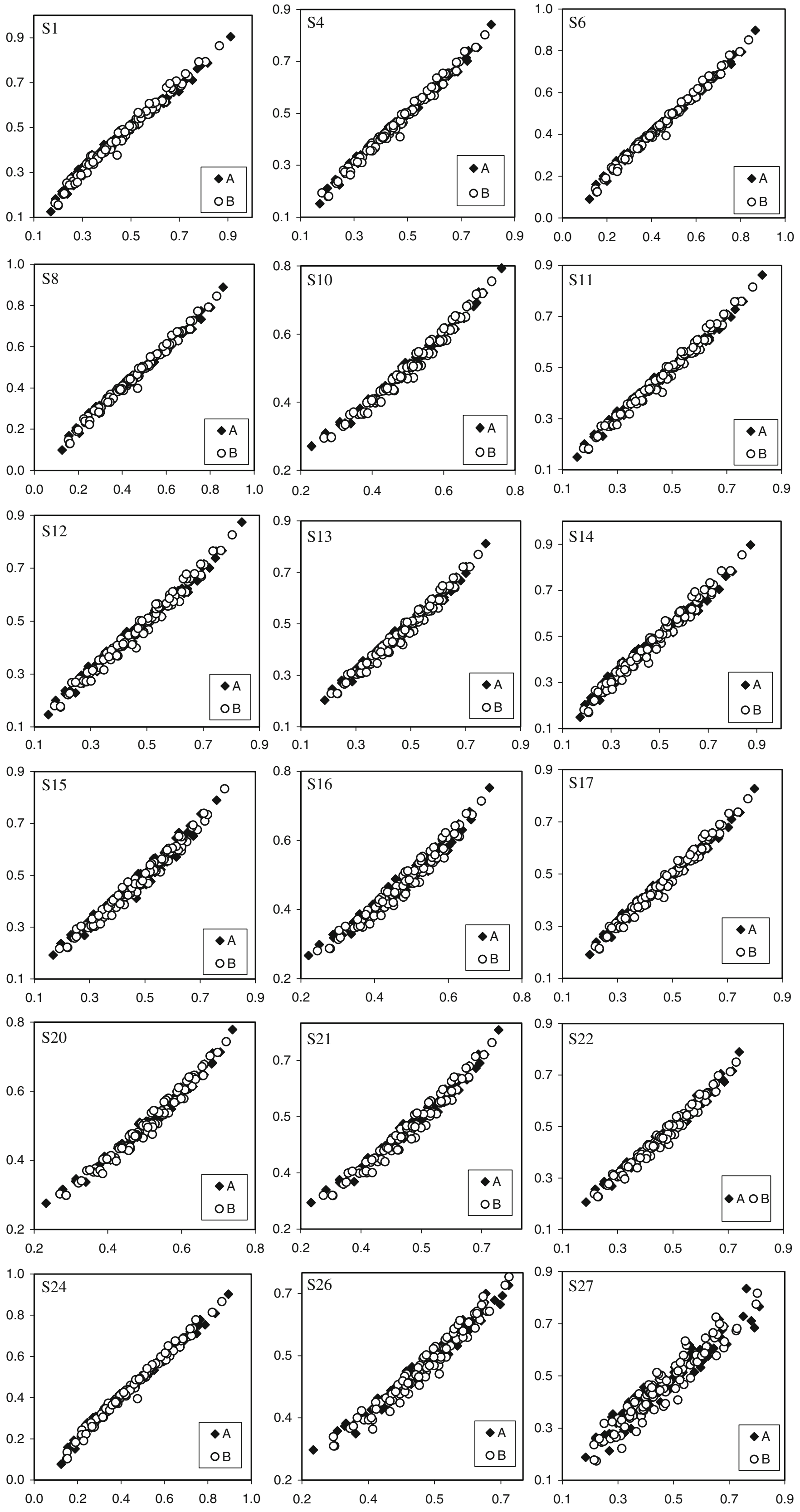

Fig. 3. Plots of the RRTs estimated for the odd set (◆, **A**) and even set (○, **B**) samples by holdout model versus that observed RRTs experimentally for all stationary phases of PCBs

Multilinear Regressions with Descriptors

The general purpose of multiple regressions is to quantify the relationship between several independent or predictor and dependent variables. A set of coefficients defines the single linear combination of independent variables (molecular descriptors) that best describes RRT values. The RRT values for each PCB on the stationary phase would then be calculated as a composite of each semi-empirical topological descriptor weighted by the respective coefficients. A multiple linear model can be represented as:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k + \varepsilon \qquad (1)$$

where $\{X_1, \ldots, X_k\}$ are molecular descriptors, β_0 is the regression model constant, β_1– β_k are the coefficients corresponding to the descriptors X_1–X_k and y is dependent variable. The values for $\beta_0 - \beta_k$ are chosen by minimizing the sum of squares of the vertical distances of the points from the hyperplane so as to give the best prediction of y from X. Regression coefficients represent the independent contributions of each calculated molecular descriptor. In matrix notation, we will write the MLR model is defined in Eq. 2 as:

$$y = \mathbf{X}b + e \qquad (2)$$

When **X** is of full rank the least squares solution is: $\hat{b} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$ where $\hat{b}$ is the estimator for the regression coefficients in $\hat{b}$. The advantages of MLR are that it is simple to use and the derived models are easy to interpret. The sign of the coefficients $\beta_0 - \beta_k$ shows whether the molecular descriptors contribute positively or negatively to the target property and their magnitudes indicates the relative importance of the descriptors to the target property. Table 3 summarizes the best correlation between the computed semi-empirical molecular structure descriptors including *o*-Cl, *m*-Cl, *p*-Cl, and S_M and RRT through MLR analysis for the total sets of PCB congeners on 18 GC capillary columns, where R^2 is the calibration squared correlation coefficients, RMS is the root mean square error, REP is the relative error of prediction, *F* is the Fisher's criterion at the 95% level probability of the equations. The value after the symbol "±" in the parenthesis is the standard deviation related to the regression coefficient. As can be seen in this Table, the R^2 values above 0.9651 with the exception of system 27 ($R^2 = 0.9272$) and the RMS and REP below 0.0179, 3.8538; except for system 27 (RMS = 0.0335, REP = 7.2682), indicated that the MLR models have good statistical qualities with low prediction error and demonstrated an excellent predictive power of the obtained QSRR models for all stationary phases. In addition, taking into account the signs of the correlation coefficients, the following explanation of RRT-semi-empirical descriptors relationships can be given. The terms of *o*-Cl, *m*-Cl and *p*-Cl are positively while the S_M term is negatively correlated with the RRT in all QSRR models, which indicate increasing the number of chlorine atoms substituted in *ortho*, *meta* and *para* positions on biphenyl rings caused to increase further interaction with the stationary phases possessing different polarities, while increasing the molecular size caused to increase its surface area and stronger intermolecular dispersion force will be achieved, therefore the PCB molecule with larger S_M value does not tend to be adsorbed onto the stationary phase (the RRT become less).

Table 4. Statistical parameters of the over-fitting and predictive ability of the models

System no.	Odd samples					Even samples		
	$RMSE_{RS}$	R^2_{RS}	$RMSE_{HO}$	R^2_{HO}	$RMSE_{RS}$	R^2_{RS}	$RMSE_{HO}$	R^2_{HO}
1	0.0150	0.9905	0.0158	0.9903	0.0169	0.9870	0.0176	0.9867
4	0.0110	0.9934	0.0115	0.9932	0.0134	0.9897	0.0137	0.9895
6	0.0120	0.9943	0.0126	0.9941	0.0141	0.9915	0.0146	0.9913
8	0.0124	0.9936	0.0132	0.9933	0.0147	0.9905	0.0153	0.9902
10	0.0124	0.9853	0.0127	0.9850	0.0144	0.9788	0.0148	0.9784
11	0.0133	0.9908	0.0138	0.9906	0.0163	0.9855	0.0166	0.9853
12	0.0157	0.9878	0.0162	0.9876	0.0189	0.9811	0.0193	0.9810
13	0.0149	0.9844	0.0153	0.9841	0.0177	0.9764	0.0181	0.9762
14	0.0179	0.9853	0.0186	0.9851	0.0214	0.9775	0.0220	0.9774
15	0.0207	0.9720	0.0210	0.9715	0.0189	0.9769	0.0192	0.9765
16	0.0162	0.9713	0.0167	0.9706	0.0187	0.9593	0.0191	0.9585
17	0.0136	0.9881	0.0141	0.9880	0.0163	0.9819	0.0167	0.9817
20	0.0129	0.9827	0.0133	0.9822	0.0151	0.9751	0.0155	0.9745
21	0.0141	0.9758	0.0145	0.9751	0.0162	0.9661	0.0166	0.9653
22	0.0149	0.9832	0.0154	0.9827	0.0173	0.9762	0.0178	0.9755
24	0.0138	0.9929	0.0147	0.9927	0.0164	0.9894	0.0170	0.9892
26	0.0151	0.9716	0.0156	0.9706	0.0173	0.9605	0.0178	0.9593
27	0.0335	0.9413	0.0345	0.9406	0.0393	0.9139	0.0402	0.9130

Model Prediction-Validation

Model validation is a critical component of QSRR development. A number of procedures have been established to determine the quality of QSRR models. Therefore, a leave-one-out cross-validation (LOO-CV), *Y*-randomization, and external validation (EV) procedures through an odd–even number and division of the entire data set into training and test sets are used to validate the predictive ability and check the statistical significance of the developed 18 QSRR models.

Cross-Validation

The most popular validation method is cross-validation (CV), known as jackknifing or leave-one-out (LOO). This method systematically removes one data point at a time from the training set, and constructs a model with the reduced data set. Subsequently, the model is used to predict the data point that has been left out. By repeating the procedure for the entire data set, a complete set of predicted properties and cross-validated statistics can be obtained. It has been argued that the LOO procedure often overestimates the predictivity of the model and that, subsequently, the QSRR models are overoptimistic [48]. For cross-validated statistics, it has been suggested that prediction residual error sum of squares (PRESS), cross-validated square correlation coefficient (R^2_{cv}) and root mean square error in cross-validation (RMS_{cv}) are good estimates of the real prediction error of a model:

$$\mathrm{PRESS} = \sum_{i=1}^{N} (y_{\mathrm{pred},i} - y_{\mathrm{obs},i})^2 \tag{3}$$

$$R^2_{cv} = 1 - \frac{\sum_{i=1}^{N} (y_{\mathrm{pred},i} - y_{\mathrm{obs},i})^2}{\sum_{i=1}^{N} (y_{\mathrm{obs},i} - \bar{y}_{\mathrm{obs}})^2} = 1 - \frac{\mathrm{PRESS}}{\sum_{i=1}^{N} (y_{\mathrm{obs},i} - \bar{y}_{\mathrm{obs}})^2} \tag{4}$$

$$\mathrm{RMS_{cv}} = \sqrt{\frac{\sum_{i=1}^{N} (y_{\mathrm{pred},i} - y_{\mathrm{obs},i})^2}{N}} \tag{5}$$

where N is the number of training patterns, $y_{\mathrm{obs},i}$ and $y_{\mathrm{pred},i}$ are the experimental, and predicted RRTs of the left-out compound i, respectively and $\bar{y}_{\mathrm{obs}}$ is the average experimental RRT of left-in compounds different from i. Values of R^2_{cv} can range from one to less than zero. A value of one indicates a perfect prediction, and a value of 0 means that the QSRR derived has no modeling power. Negative values arise from a situation where the derived QSRR is a poorer description of data than no model at all. The R^2_{cv} values can be considered as a measure of the predictive power of a model: whereas R^2 can always be increased artificially by adding more parameters, R^2_{cv} decreases if a model is over parameterized [49], and is therefore a more meaningful summary statistic for predictive models. The correlation coefficients (R^2_{cv}) and RMS_{cv} for each subset are presented in Table 3 and the resulted values are plotted in Fig. 3. The cross-validation results show that the R^2_{cv} are higher than 0.9230 and RMS_{CV} lower than 0.0345 for all GC stationary phases. Furthermore, in all cases, the cross-validated R^2_{cv} values are very close to the corresponding R^2 values and the cross-validated RMS_{cv} values are only slightly larger than the corresponding RMS values. Clearly, the cross-validation demonstrates the final models to be statistically significant.

This method is not a very rigorous model predictivity test and suffers from two other major deficiencies: the

Table 5. QSRR models of GC-RRT values versus (52 + 180) for the training sets ($n = 160$) of polychlorinated biphenyl (PCB) congeners on 18 GC capillary columns

System no.	Models of the training sets
S1	$\mathrm{RRT} = 1.2899(\pm 0.1128) + 0.0905(\pm 0.0036)$ no. o-Cl $+ 0.1680(\pm 0.0063)$ no. m-Cl $+ 0.1769(\pm 0.0066)$ no. p-Cl $- 0.0035(\pm 0.0003)\ S_M$
S4	$\mathrm{RRT} = 0.6307(\pm 0.0851) + 0.0858(\pm 0.0027)$ no. o-Cl $+ 0.1203(\pm 0.0047)$ no. m-Cl $+ 0.1307(\pm 0.0050)$ no. p-Cl $- 0.0015(\pm 0.0002)\ S_M$
S6	$\mathrm{RRT} = 0.5916(\pm 0.0925) + 0.0669(\pm 0.0030)$ no. o-Cl $+ 0.1375(\pm 0.0052)$ no. m-Cl $+ 0.1494(\pm 0.0054)$ no. p-Cl $- 0.0016(\pm 0.0003)\ S_M$
S8	$\mathrm{RRT} = 0.5849(\pm 0.0970) + 0.0642(\pm 0.0031)$ no. o-Cl $+ 0.1352(\pm 0.0054)$ no. m-Cl $+ 0.1460(\pm 0.0057)$ no. p-Cl $- 0.0015(\pm 0.0003)\ S_M$
S10	$\mathrm{RRT} = 0.4113(\pm 0.0933) + 0.0417(\pm 0.0030)$ no. o-Cl $+ 0.0788(\pm 0.0052)$ no. m-Cl $+ 0.0849(\pm 0.0055)$ no. p-Cl $- 0.0005(\pm 0.0003)\ S_M$
S11	$\mathrm{RRT} = 0.7831(\pm 0.1022) + 0.0703(\pm 0.0033)$ no. o-Cl $+ 0.1308(\pm 0.0057)$ no. m-Cl $+ 0.1394(\pm 0.0060)$ no. p-Cl $- 0.0020(\pm 0.0003)\ S_M$
S12	$\mathrm{RRT} = 0.9207(\pm 0.1185) + 0.0759(\pm 0.0038)$ no. o-Cl $+ 0.1400(\pm 0.0066)$ no. m-Cl $+ 0.1498(\pm 0.0070)$ no. p-Cl $- 0.0024(\pm 0.0003)\ S_M$
S13	$\mathrm{RRT} = 0.7553(\pm 0.1112) + 0.0614(\pm 0.0036)$ no. o-Cl $+ 0.1121(\pm 0.0062)$ no. m-Cl $+ 0.1205(\pm 0.0065)$ no. p-Cl $- 0.0017(\pm 0.0003)\ S_M$
S14	$\mathrm{RRT} = 1.4395(\pm 0.1355) + 0.0930(\pm 0.0044)$ no. o-Cl $+ 0.1706(\pm 0.0076)$ no. m-Cl $+ 0.1815(\pm 0.0080)$ no. p-Cl $- 0.0039(\pm 0.0004)\ S_M$
S15[a]	$\mathrm{RRT} = 0.8821(\pm 0.1323) + 0.0698(\pm 0.0042)$ no. o-Cl $+ 0.1233(\pm 0.0074)$ no. m-Cl $+ 0.1329(\pm 0.0078)$ no. p-Cl $- 0.0021(\pm 0.0004)\ S_M$
S16	$\mathrm{RRT} = 0.6943(\pm 0.1202) + 0.0514(\pm 0.0039)$ no. o-Cl $+ 0.0880(\pm 0.0067)$ no. m-Cl $+ 0.0959(\pm 0.0071)$ no. p-Cl $- 0.0014(\pm 0.0003)\ S_M$
S17	$\mathrm{RRT} = 0.9449(\pm 0.1037) + 0.0670(\pm 0.0033)$ no. o-Cl $+ 0.1277(\pm 0.0058)$ no. m-Cl $+ 0.1354(\pm 0.0061)$ no. p-Cl $- 0.0023(\pm 0.0003)\ S_M$
S20	$\mathrm{RRT} = 0.3808(\pm 0.0967) + 0.0368(\pm 0.0031)$ no. o-Cl $+ 0.0751(\pm 0.0054)$ no. m-Cl $+ 0.0817(\pm 0.0057)$ no. p-Cl $- 0.0004(\pm 0.0002)\ S_M$
S21	$\mathrm{RRT} = 0.5130(\pm 0.1050) + 0.0404(\pm 0.0034)$ no. o-Cl $+ 0.0761(\pm 0.0059)$ no. m-Cl $+ 0.0829(\pm 0.0062)$ no. p-Cl $- 0.0008(\pm 0.0003)\ S_M$
S22	$\mathrm{RRT} = 0.5440(\pm 0.1100) + 0.0476(\pm 0.0035)$ no. o-Cl $+ 0.0994(\pm 0.0061)$ no. m-Cl $+ 0.1072(\pm 0.0065)$ no. p-Cl $- 0.0011(\pm 0.0003)\ S_M$
S24	$\mathrm{RRT} = 0.5887(\pm 0.1106) + 0.0613(\pm 0.0036)$ no. o-Cl $+ 0.1424(\pm 0.0062)$ no. m-Cl $+ 0.1559(\pm 0.0065)$ no. p-Cl $- 0.0016(\pm 0.0003)\ S_M$
S26	$\mathrm{RRT} = 0.4424(\pm 0.1125) + 0.0278(\pm 0.0036)$ no. o-Cl $+ 0.0735(\pm 0.0063)$ no. m-Cl $+ 0.0800(\pm 0.0066)$ no. p-Cl $- 0.0006(\pm 0.0003)\ S_M$
S27	$\mathrm{RRT} = 1.5647(\pm 0.2523) + 0.0707(\pm 0.0081)$ no. o-Cl $+ 0.1705(\pm 0.0141)$ no. m-Cl $+ 0.1807(\pm 0.0148)$ no. p-Cl $- 0.0040(\pm 0.0007)\ S_M$

[a] For system 15 $n = 159$

time to carry out the cross-validation increases as the square of the size of training set; the method produces n final models (each corresponding to one of the training set molecules being left out) and it is not clear which is the 'best' model. To further check the prediction ability of the resulting QSRR models two better methods are applied here, one by Hawkins and co-workers in 2004 [50] namely the odd–even external validation and the other better method is to remove a percentage of the training set into a prediction set [49, 51].

Odd–Even External Validation

To validate and develop a credible QSRR model, it is not enough to build a model for the whole data set. So, the 209 (208 for System 15) data set PCBs for all stationary phases were sorted in the ascending order of RRT values and then divided into two sets namely "odd set" and "even set" RRTs [49, 51]. This way of splitting ensures that the distribution of RRT values of the two subsets were very similar. The QSRR models were fitted to the odd set and even set samples separately and the resulted fitness were assessed by applying QSRR models to both samples. To compare the estimation abilities of the models, two statistical parameters namely root mean squares error (RMSE) and R^2, were calculated. The same data set (i.e. 'calibration set') that was already used to fit the models was employed to determine resubstitution parameters, i.e. $\mathrm{RMSE}_{\mathrm{RS}}$ and R^2_{RS}, also to determine holdout parameters, i.e. $\mathrm{RMSE}_{\mathrm{HO}}$ and R^2_{HO} for the other data set, which was not involved in the fitting. The resubstitution statistical parameters of the samples base their predictions on the regression fitted to those samples and this is while the holdout statistical parameters base their predictions on the regression fitted to the other samples. The plots of RRTs estimated by odd- and even-set QSRR models (holdout prediction) versus the RRTs observed experimentally are given in Fig. 3, also Table 4 summarizes these statistical parameters achieved by this approach. As can be seen, in the odd- and even-set samples, the resubstitution and holdout RMSE are very similar, indicating that the same sample and other sample predictions are equally precise for all stationary phases.

Y-Randomization Test

Another procedure that is easy to perform is a randomization test called *Y*-randomization (randomization of response, i.e. in our case RRT). In this method for each column stationary phase, the output RRTs values of the compounds are shuffled randomly, and the resulting data set is examined by the QSRR method against real (unscrambled) input descriptors to determine the correlation and predictivity of the resulting "model" [43, 52–54]. The whole procedure is repeated on many different scrambled data sets. The rationale behind this test is that the significance of the real QSRR model would be suspect if there is a strong correlation between the selected descriptors and the randomized response variables. The randomization was repeated ten times. If the statistical qualities of these models are much lower than the original model, it can be considered that the model is reasonable and had not been obtained by chance. The results are shown in last column of Table 2. Very low level of $R^2_{\max}$ (in the interval of 0.0012 for S1 and 0.0030 for S27) indicates good results in our original models and is not due to a chance correlation or structural dependency of the training set for each stationary phase of the GC column.

Calibration and Prediction Sets

In this investigation, for further testing the predictive ability of the models for the external compounds without the models, part of the congeners are picked up from 209 (208 for System 15) PCBs to construct a training set which is used to develop a prediction model and then predict the values of RRTs in the remaining congeners. How to pick up the compounds in the training set is very important for developing of the predictive QSRR models. In this case, before each training run, all data sets were split randomly into two separate sub-matrices: the training set matrix and external testing set matrix. Out of 160 congeners (159 for System 15) (76.6%) were used for the training set and 49 congeners (23.4%) were used as external validation. The PCBs constituting the training and testing sets are clearly presented in Table S1. Moreover, the same divisions were repeated with corresponding RRTs values. The test examples are marked as bold font and training set was also used to obtain the best fit equation of MLR with four semi-empirical descriptors.

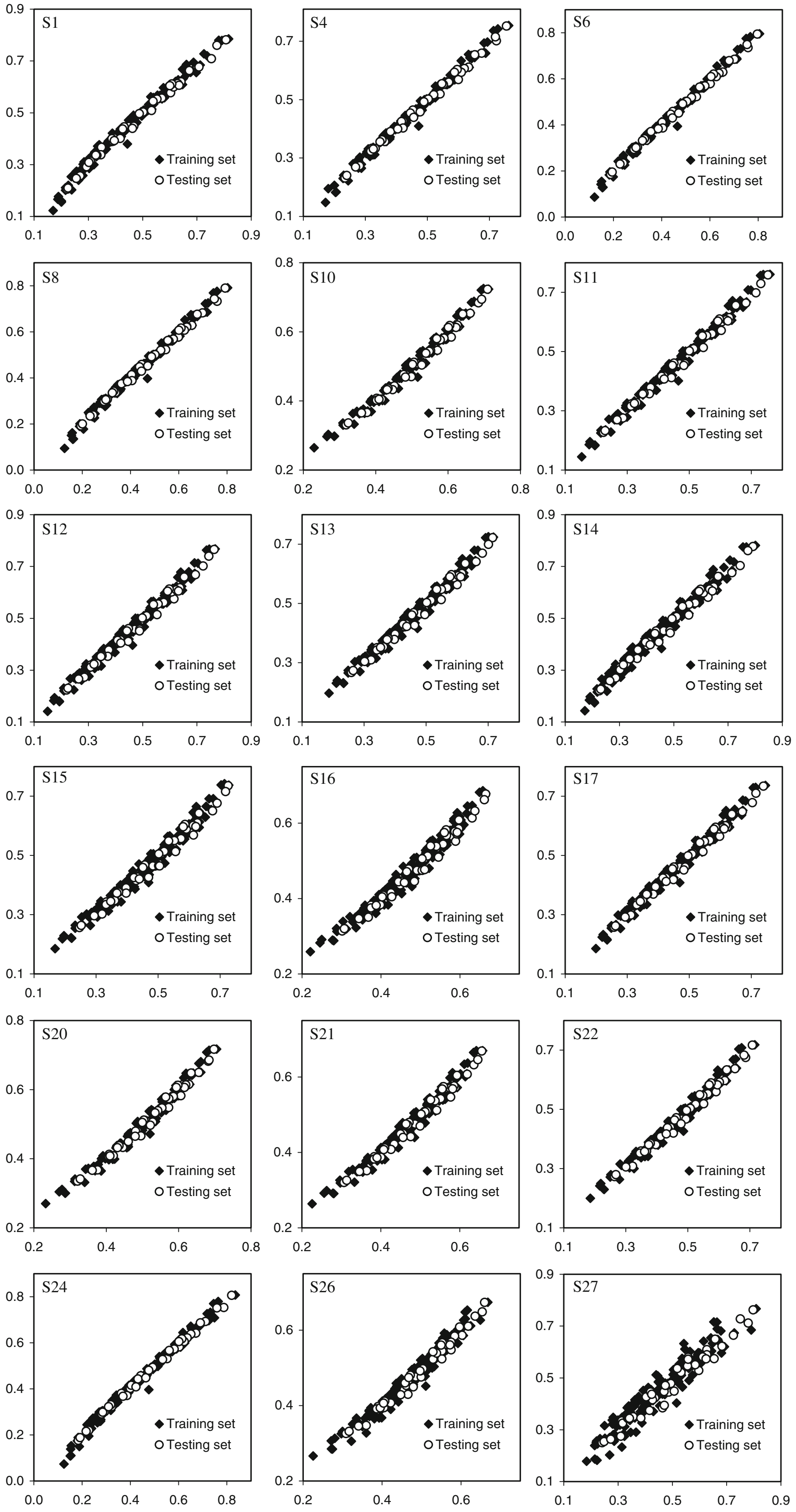

Fig. 4. Plots of the RRTs estimated by the QSPR models in Table 5 vs. that observed for 160 (system 15 159) training set PCBs (◆) and 49 testing set ones (○) for all stationary phases

Table 6. Statistical parameters of the QSRR models obtained using different molecular descriptors

System no.	Training set						Testing set		
	RMS	REP	*F*	R^2	R^2_{cv}	RMS_{cv}	RMS	REP	R^2
S1	0.0165	3.7285	2743.5	0.9861	0.9850	0.0171	0.0160	3.1327	0.9941
S4	0.0124	2.7646	3905.7	0.9902	0.9894	0.0129	0.0120	2.3654	0.9936
S6	0.0135	3.0657	4528.8	0.9915	0.9909	0.0140	0.0123	2.4304	0.9950
S8	0.0142	3.2060	3967.4	0.9903	0.9896	0.0147	0.0124	2.4454	0.9948
S10	0.0136	2.7298	1890.4	0.9799	0.9784	0.0142	0.0135	2.4849	0.9869
S11	0.0149	3.3100	2865.7	0.9867	0.9857	0.0155	0.0152	2.9895	0.9906
S12	0.0173	3.8469	2208.1	0.9828	0.9816	0.0179	0.0183	3.5743	0.9874
S13	0.0163	3.5377	1773.6	0.9786	0.9771	0.0168	0.0173	3.3901	0.9837
S14	0.0198	4.4217	1768.5	0.9786	0.9770	0.0205	0.0207	4.0290	0.9869
S15	0.0193	4.2104	1392.7	0.9731	0.9712	0.0200	0.0209	4.0603	0.9790
S16	0.0176	3.7411	977.4	0.9619	0.9590	0.0182	0.0181	3.5409	0.9726
S17	0.0152	3.3311	2216.0	0.9828	0.9816	0.0157	0.0152	2.9866	0.9891
S20	0.0142	2.8370	1649.6	0.9770	0.9753	0.0147	0.0147	2.7339	0.9831
S21	0.0154	3.2720	1167.8	0.9679	0.9654	0.0159	0.0155	3.0485	0.9775
S22	0.0161	3.5004	1707.7	0.9778	0.9762	0.0167	0.0171	3.3706	0.9824
S24	0.0162	3.6989	3352.6	0.9886	0.9876	0.0168	0.0125	2.4660	0.9959
S26	0.0165	3.5057	993.4	0.9625	0.9595	0.0171	0.0164	3.2443	0.9728
S27	0.0369	8.2946	421.0	0.9157	0.9093	0.0383	0.0370	7.2078	0.9497

Furthermore, the testing set was used to monitor overfitting the MLR models. The resulted MLR models for training set congeners were the same as those obtained for the entire set of all PCBs in each subset subject to use descriptors of all congener's models supporting sufficient ability for the prediction set of 49 PCBs. The resulting regression equations of the training set for individual HRGC column stationary phases with the optimal four descriptors are indexed in Table 5, and results obtained are plotted in Fig. 4. Statistical parameters for the best-fitted models are also presented in Table 6. The correlation coefficients (R^2) of the obtained models are >0.96 for all the stationary phases except for system 27 (0.9157), and the highest one is 0.9915 for System 6. The root mean square errors (RMS) and relative error prediction (REP) of estimation ranged from 0.0124, 2.7646 of System 4 to 0.0198, 4.4217 of System 14 (except for System 27), respectively, also the F statistic values are >977.4 (except System 27). The LOO-CV method was used to examine the stability of QSRR models, and the values of R^2_{cv} and RMS_{cv} for the models were above 0.9590 and in the range of 0.0129 and 0.0205 (except for System 27). The predicted RRTs versus the observed RRTs of the 160 (159 for System 15) PCB training sets are plotted in Fig. 4 (diamond). As shown in Table 6 and Fig. 4, the QSRR statistical results exhibit good estimation capacity and stability for internal training set PCB samples to individual stationary phases. High predictive ability of QSRR models for external examples is another criterion of a good QSRR model. The predicted RRTs of 49 PCBs in the external testing set by the models in Table 5 are also demonstrated in Fig. 4 (circle) versus the observed RRTs of 18 GC stationary phases. For all 18 HRGC stationary phases, the regression of the observed and predicted RRTs had a high agreement with the diagonal of each chart. The predicted correlation coefficients (R^2) over 0.9726 with the exception of System 27 ($R^2 = 0.9497$), the root mean square errors (RMS) and relative error prediction (REP) below 0.0209 and 4.603, respectively, except for System 27 ($RMS = 0.0370$, $REP = 7.2079$) demonstrated an excellent predictive power of the obtained QSRR models.

Conclusion

The HRGC-RRT values of PCBs on 18 capillary stationary phases (S1, S4, S6, S8, S10, S11, S12, S13, S14, S15, S16, S17, S20, S21, S22, S24, S26, S27), depend on four semi-empirical molecular descriptors, *o*-Cl, *m*-Cl, *p*-Cl and S_M. MLR with non-redundant descriptors (NRD) produced more predictive, informative and significantly improved QSRR models. The validation and predictive ability of the models were examined by three methods of leave-one-out cross-validation, *Y*-randomization, and external validation. The methods indicated that the resulted multiparametric QSRR models have high prediction ability and low overfitting. All QSRR models but one related to the System 27 column stationary phase provide a reasonably well calibrated correlation coefficient ($R^2 = 0.9619$–0.9915) and the LOO cross-validation correlation coefficient ($R^2_{CV} = 0.9590$–0.9909).

Open Access

References

1. Schantz SL (1996) Neurotoxicol Teratol 18:217–227. doi:10.1016/S0892-0362(96)90001-X
2. Erickson MD (1997) Analytical chemistry of PCBs, 2nd edn. CRC Press LLC, Boca Raton, FL, USA
3. Ballschmiter K, Zell M (1980) Fresenius Z Anal Chem 302:20–31. doi:10.1007/BF00469758
4. Guitart R, Puig P, Gomez-Catalan J (1993) Chemosphere 27:1451–1459. doi:10.1016/0045-6535(93)90239-2
5. Safe S (1990) Crit Rev Toxicol 21:51–88. doi:10.3109/10408449009089873
6. Jones KC, Burnett V, Daurte-Davidson R, Waterhouse KS (1991) Chemistry in Britain
7. Newman CN, Unger MA (2003) Fundamentals of ecotoxicology, 2nd edn. CRC Press LLC, Boca Raton, FL, USA
8. Mills SA III, Thal DI, Barney J (2007) Chemosphere 68:1603–1612. doi:10.1016/j.chemosphere.2007.03.052
9. Tanabe S, Kannan N, Subramanian A, Watanabe S, Tatsukawa R (1987) Environ Pollut 47:147–163. doi:10.1016/0269-7491(87)90044-3
10. Iwata H, Tanabe S, Sakal N, Tatsukawa R (1993) Environ Sci Technol 27:1080–1098. doi:10.1021/es00043a007
11. Harrad SJ, Sewart AP, Alcock R, Boumphrey R, Burnett V, Duartedavidson R, Halsall C, Sanders G, Waterhouse

K, Wild SR, Jones KC (1994) Environ Pollut 85:131–146. doi:10.1016/0269-7491(94)90079-5
12. Safe SH (1994) Crit Rev Toxicol 24:87–149. doi:10.3109/10408449409049308
13. WHO (1993) Polychlorinated biphenyls and terphenyls. Environmental health criteria, vol 140, 2nd edn. World Health Organization, Geneva
14. EPA (1998) PCB ID-Table of PCB congeners and other species. http://www.epa.gov/ toxteam/pcbid/table.htm. Accessed 5 May 2007
15. Van den Berg M, Birnbaum LS, Denison M, De Vito M, Farland W, Feeley M, Fiedler H, Hakansson H, Hanberg A, Haws L, Rose M, Safe S, Schrenk D, Tohyama C, Tritscher A, Tuomisto J, Tysklind M, Walker N, Peterson RE (2006) Toxicol Sci 93:223–241. doi:10.1093/toxsci/kfl055
16. Vetter W, Luckas B, Buijten J (1998) J Chromatogr A 799:249–258. doi:10.1016/S0021-9673(97)01101-1
17. Frame GM (1997) Fresenius J Anal Chem 357:714–722. doi:10.1007/s002160050238
18. Grob RL, Barry EF (2004) Modern practice of gas chromatography, 4th edn. Wiley, New York
19. Kaliszan R (1987) Quantitative structure–chromatographic retention relationships. Wiley, New York
20. McFarland JW, Berger CM, Froshauer SA, Hayashi SF, Hecker SJ, Jaynes BH, Jefson MR, Kamicker BJ, Lipinski CA, Lundy KM, Reese CP, Vu CB (1997) J Med Chem 40:1340–1346. doi:10.1021/jm960436i
21. Gramatica P, Navas N, Todeschini R (1998) Chemom Intell Lab Syst 40:53–63. doi:10.1016/S0169-7439(97)00079-8
22. Kaliszan R, van Straten MA, Markuszewski M, Cramers CA, Claessens HA (1999) J Chromatogr A 855:455–486. doi:10.1016/S0021-9673(99)00742-6
23. Hyötyläinen T, Kallio M, Hartonen K, Jussila M, Palonen S, Riekkola M-L (2002) Anal Chem 74:4441–4446. doi:10.1021/ac0201528
24. Larsen B, Bowadt S, Tilio R (1993) In: Albaigés J (ed) Environmental analytical chemistry of PCBs, vol 16. Gordon & Breach, Singapore, pp 3–24
25. Storr-Hansen E (1993) In: Albaigés J (ed) Environmental analytical chemistry of PCBs, vol 16. Gordon & Breach, Singapore, pp 24–38
26. Eganhouse R, Gould B, Olaguer D, Phinney C, Sherblom P (1993) In: Albaigés J (ed) Environmental analytical chemistry of PCBs, vol 16. Gordon & Breach, Singapore, pp 111–134
27. Pavoni B, Sfriso A, Raccanelli S (1993) In: Albaigés J (ed) Environmental analytical chemistry of PCBs, vol 16. Gordon & Breach, Singapore, pp 101–110
28. Bedard D, May R (1996) Environ Sci Technol 30:237–245. doi:10.1021/es950262e
29. Plomley J, Laušević M, March R (2000) Mass Spectrom Rev 19:305–365. doi:10.1002/1098-2787
30. March R (1997) J Mass Spectrom 32:351–369. doi:10.1002/(SICI)1096-9888
31. Liu Z, Philips JB (1991) J Chromatogr Sci 29:227–231
32. Phillips JB, Xu J (1995) J Chromatogr A 703:327–334. doi:10.1016/0021-9673(95)00297-Z
33. Haglund P, Harju M, Ong R, Marriott P (2001) J Micro Sep 13:306–311
34. Focant JF, Sjödin A, Patterson DG (2004) J Chromatogr A 1040:227–238. doi:10.1016/j.chroma.2004.04.003
35. Frame GM (1997) Fresenius J Anal Chem 357:701–713. doi:10.1007/s002160050237
36. Makino M (1999) Chemosphere 39:893–903. doi:10.1016/S0045-6535(99)00032-6
37. Hasan MN, Jurs PC (1988) Anal Chem 60:978–982. doi:10.1021/ac00161a007
38. Liu SS, Liu Y, Yin DQ, Wang XD, Wang LS (2006) J Sep Sci 29:296–301. doi:10.1002/jssc.200301592
39. Gramatica P, Navas N, Todeschini R (1998) Chemom Intell Lab Syst 40:53–63. doi:10.1016/S0169-7439(97)00079-8
40. Krawczuk A, Voelkel A, Lulek J, Urbaniak R, Szyrwinska K (2003) J Chromatogr A 1018:63–71. doi:10.1016/j.chroma.2003.08.037
41. Ren Y, Liu H, Yao X, Liu M (2007) Anal Bioanal Chem 388:165–172. doi:10.1007/s00216-007-1188-0
42. Jäntschi L, Bolboacă SD, Diudea MV (2007) Int J Mol Sci 8:1125–1157. doi:10.3390/i8111125
43. Ghavami R, Sadeghi F (2009) Chromatographia 70:851–868. doi:10.1365/s10337-009-1233-6
44. ISIS Draw 2.3. MDL information systems, Inc., 1990–2000
45. HyperChem Release 7.1 for windows molecular modeling system program package, HyperCube, 2002
46. Todeschini R, Consonni V, Dragon software version 2.3. (http://www.disat.unimib.it/chm/dragon.htm)
47. Olivero J, Garcia T, Payares P, Vivas R, Diaz D, Daza E, Geerliger P (1997) J Pharm Sci 86:625–630. doi:10.1021/js960196u
48. Shao J (1993) J Am Stat Assoc 88:486–494
49. Hawkins DM, Basak SC, Mills D (2003) J Chem Inf Comput Sci 43:579–586. doi:10.1021/ci025626i
50. Hawkins DM (2004) J Chem Inf Comput Sci 44:1–12. doi:10.1021/ci0342472
51. Tetko IV, Livingstone DJ, Luik AI (1995) J Chem Inf Comput Sci 35:826–833. doi:10.1021/ci00027a006
52. Golbraikh A, Tropsha A (2000) Mol Divers 5:231–243. doi:10.1023/A:1021372108686
53. Melagraki G, Afantitis A, Sarimveis H, Koutentis PA, Markopoulose J, Igglessi-Markopouloua O (2007) Bioorg Med Chem 15:7237–7247. doi:10.1016/j.bmc.2007.08.036
54. Hemmateeneja B, Sanchooli M (2007) J Chemom 21:96–107. doi:10.1002/cem.1039