Journal of Chromatography A, 1223 (2012) 118–125

Contents lists available at SciVerse ScienceDirect

Journal of Chromatography A

journal homepage: www.elsevier.com/locate/chroma

# Comparison of quantitative structure–retention relationship models on four stationary phases with different polarity for a diverse set of flavor compounds

Jun Yan[a,1], Dong-Sheng Cao[a,1], Fang-Qiu Guo[a], Liang-Xiao Zhang[b], Min He[a], Jian-Hua Huang[a], Qing-Song Xu[c], Yi-Zeng Liang[a,*]

[a] Research Center of Modernization of Traditional Chinese Medicine, Central South University, Changsha 410083, China
[b] Key Laboratory of Separation Science for Analytical Chemistry, Dalian Institute of Chemical Physics, Chinese Academy of Sciences, Dalian 116023, China
[c] School of Mathematical Sciences and Computing Technology, Central South University, Changsha 410083, China

ARTICLE INFO

*Article history:*
Received 5 October 2011
Received in revised form 3 December 2011
Accepted 5 December 2011
Available online 13 December 2011

*Keywords:*
Quantitative structure–retention relationship (QSRR)
Retention index
Flavor compounds
Polarity

ABSTRACT

A quantitative structure–retention relationship study was performed for 656 flavor compounds with highly structural diversity on four stationary phases of different polarities, using topological, constitutional, quantum chemical and geometrical descriptors. Statistical methods were employed to find an informative subset that can accurately predict the gas chromatographic retention indices (RIs). Multivariable linear regression (MLR) was used to map the descriptors to the RIs. The stability and validity of models have been tested by internal and external validation, and good stability and predictive ability were obtained. The resulting QSRR models were well-correlated, with the square of correlation coefficients for cross validation, $Q^2$, values of 0.9595, 0.9528, 0.9595 and 0.9223 on stationary phase OV101, DB5, OV17 and C20M, respectively. The molecular properties known to be relevant for GC retention index, such as molecular size, branching, electron density distribution and hydrogen bond effect were well covered by generated descriptors. The descriptors used in models on four stationary phases were compared, and some reasonable explanations about gas chromatographic retention mechanism were obtained. The model may be useful for the prediction of flavor compounds while experimental data is unavailable.

## 1. Introduction

Flavor compounds are these chemicals can bind to proteins on the olfactory receptor neurons (ORNs) at the surface of the olfactory epithelium, excitation of ORNs generates a topographic map of sensory information in the brain that is a representation of the stimulating chemical features of the external world. In fact, a seemingly infinite number of perceptions are invoked by less than 1000 flavor compounds that make up human odor space, i.e. flavor compounds are the material foundation of olfaction [1]. Therefore, the identification of flavor compounds is the basis of olfaction research. In addition, it is very important in flavor industry for essence preparation, imitation and new essence exploration [2], and it is helpful for quality control of food products [3].

To determine compounds that are responsible for the flavor of a product, the crucial step is the identification of the odor-active compound. One method is gas chromatograph-olfactometry (GCO) which is the collection of techniques that combine olfactometry or the use of human detectors to access odor activity in defined air streams with the gas chromatographic separation of volatiles, this approach has been widely used in the field of flavor compounds analysis [4], the other is gas chromatography–mass spectrometry (GC–MS), the most popular and important tool in analytical chemistry, many works about this have been published in recent years [5,6]. However, both of the two approaches have drawbacks. For GCO, the result is easily affected by the subjectivities of observers and the changes of environment conditions. For GC–MS, ambiguous identification can be obtained when structurally related compounds that give similar mass spectra, for instance, in the case of isomeric compounds. Except for that, standard spectra of some compounds cannot be found in mass spectra database or commercially unavailable for standard sample is also a problem. Consequently, RI as a useful tool for identification purpose has been applied by many analysts. RI is firstly proposed by Kováts in 1958 and further developed by van den Dool and Kratz to LTPRI for linear temperature programming [7,8], which is independent from the operation conditions, except for the experimental temperature and the polarity of stationary phases. Hence, it is very suitable for interlaboratory comparison and provides a feasible way

[*] Corresponding author. Tel.: +86 731 8830831; fax: +86 731 8830831.
*E-mail addresses:* yizeng_liang@263.net, yanjun03@gmail.com (Y.-Z. Liang).
[1] These authors contributed equally to this paper.

0021-9673/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.chroma.2011.12.020

to study possible mechanism of retention behavior. The identification results from GCO and GC–MS combined with RI will be more accurate [9].

QSRR is a technique for relating the variations in one response variable (*Y*-retention index) to the variations of several descriptors (*X*-molecular descriptors), with predictive or explanatory purpose. Since the pioneering work of Kaliszan [10], numerous investigators have reported very good correlation between experimental chromatographic RI and various molecular descriptors in the past few years [11–15]. For instance, a QSRR study performed for RIs of biologically and environmentally important organic compounds on capillary columns with low-polar by Isidorov and Szczepaniak [16], Lu et al. studied QSRR of the RIs of 90 saturated esters on different stationary phases using novel topological indices [17], Görgényi studied the relationship between RIs and structure for 35 aliphatic ketones and aldehydes by principal component analysis [18] and partial least squares regression [19], and a set of 846 compounds with diverse chemical structures has been used to investigate the applicability of QSRRs approaches for the prediction of Kovats retention indices by Garkani-Nejad et al. [20]. Besides that, a lot of work about QSRR can be found in the comprehensive review given by Héberger in 2007, in which some important conclusions were reviewed and suggestions for future works were proposed [21].

A lot of organic compounds have been studied for RI prediction, however, typically works in this field deal with 20–300 samples that often belonging to a strictly defined class of substances [22–24]. So far, the RI of flavor compounds have not been studied systematically, hence, the aim of this study is to develop correlative models between gas chromatographic RI of four sets of flavor compounds with highly structural diversity on different stationary phases (OV101, DB5, OV17 and C20M) and four subsets of meaningful and straightforward molecular descriptors, respectively. Furthermore, we also expect to demonstrate the various effects of different molecular descriptors on a given stationary phase and a given molecular descriptor on different stationary phases. Compared with the previous work, highly structural diversity of molecules and four stationary phases of different polarity were more helpful to investigate the interaction between the solute and the stationary phase on a more general level.

## 2. Experiments

### 2.1. Data set

The data used in this study were collected from Acree's Flavornet for stationary phases OV101, DB5, OV17 and C20M [1]. For these columns, OV101 and DB5 are non-polar, OV17 is mid-polar and C20M is strongly polar (738 compounds were downloaded from http://www.flavornet.org, among them, some compounds lack of experimental retention indices on corresponding columns and some molecular structures cannot be optimized for descriptor calculation, so these compounds were excluded). Finally, a set of 297, 405, 205 and 434 molecules has been selected for this investigation on four columns, respectively. More than 20 kinds of odors including fruit, herb, baked, sulfur, rose and grape, etc., can be represented by these flavor compounds. It indicated that global diversity and local similarity are two main features of this data and also a big challenge for model development. Statistical results showed that the number of C is 1–23, the number of H is 3–48, the number of ring is 0–5, and the molecular mass is 40.02–276.25 for all molecules. A complete list of the compounds names, structures and retention indices is shown in supporting information.

### 2.2. Software

Six types of molecular descriptors were calculated with ChemoPy descriptor calculation program, developed by our group, including constitutional descriptors, topological descriptors, electronic state indices, quantum chemical descriptors, the descriptors related to molecular properties and geometrical descriptors.

Statistical evaluation of the data and multivariate data analysis has been performed mainly in Matlab 7.0. All work has been performed on personal computers running under operating system Microsoft Windows.

### 2.3. Molecular modeling

Some descriptors such as quantum chemical and geometrical descriptors are conformation-dependent. For all molecules used in our study, the energy-lowest structures were considered. A famous semi-empirical molecular orbital MOPAC 2007 program was used for optimizing the geometrical conformers of these aliphatic alcohols using the AM1 method. Thus, all geometrical and quantum chemical descriptors are then calculated based on the optimized structural features.

### 2.4. Descriptors generation

In QSRR study, molecular descriptors of chemical structures are important factors affecting the quality of the models. Various structural attributes of the molecule are used as descriptors. As to retention index, plenty of studies show that molecular size, molecular mass, shape, branching and electron density distribution, etc., are the main factors [25].

Descriptor generation contain the following steps: (1) 195 molecular descriptors, which can represent structural information more or less related to gas chromatographic retention, were calculated using the Chemopy of our group, including constitutional descriptors, topological descriptors, geometrical descriptors, and quantum chemical descriptors. (2) All descriptors were preselected by eliminating: (i) those descriptors are not available for each compounds; (ii) descriptors having a small variation in magnitude for all structures and (iii) the value of descriptors equal to zero for more than 80% compounds, which in order to avoid matrix calculation error. And then, 127 molecular descriptors remained. (3) Finally, stepwise method was used to select the most important descriptors affecting the quality of models.

Based on the results of mathematical selection and chemical research experiences, four small subsets of molecular descriptors used in four QSRRs models, respectively, were selected from the origin pool. The all molecular descriptors used in this paper were listed in Table 1.

### 2.5. Sample splitting

The molecular descriptors and the chromatographic retention indices were correlated by MLR. Origin data includes four sets, 297 samples on non-polar stationary phase OV101, 405 samples on non-polar stationary phase DB5, 205 samples on mid-polar stationary phase OV17 and 434 samples on strongly polar stationary phase C20M, respectively. Four sets of flavor compounds were divided into two groups randomly: training set and test set. The training set, representing about 3/4 of the total number of compounds, was used to build the QSRR model; the remaining 1/4 was assigned to the test set and used to validate the model. In this paper, four sets of compounds were split into 230 and 67, 305 and 100, 165 and 40, 330 and 104 for training and test set, respectively.

**Table 1**
Molecular descriptors introduction.

| No. | Name | Meaning | Category |
|---|---|---|---|
| 1 | *weight*$^{a,b}$ | Molecular weight | Constitutional |
| 2 | *ndonr*$^{a,b,c,d}$ | Number of H-bond donors | Constitutional |
| 3 | *ipc*$^{c,d}$ | Information content from adjacent matrix | Topological structure |
| 4 | $^{4}\chi^{d}$ | Simple molecular connectivity<br>Chi indices for four cluster | Topological structure |
| 5 | $^{1}\chi^{a,b}$ | Simple molecular connectivity<br>Chi indices for path order 1 | Topological structure |
| 6 | $^{0}\chi^{a,b}$ | Simple molecular connectivity<br>Chi indices for path order 0 | Topological structure |
| 7 | *qhmax*$^{c,d}$ | Most positive charge on *H* | Quantum chemical |
| 8 | $\mu^{c,d}$ | Dipole moment | Quantum chemical |
| 9 | *DPSA1*$^{d}$ | Difference between partial positively and negatively charged surface areas | CPSA descriptors |
| 10 | *FPSA1*$^{c,d}$ | Fractional partial positive surface area | CPSA descriptors |

*Note*: superscript a, b, c, d indicates the corresponding molecular descriptor is included in model 1, 2, 3 and 4, respectively.

### 2.6. Model validation

In the present study, 10-fold cross validation was performed to evaluate the robustness and validity of models. The sample is partitioned into 10 mutuality exclusive subsets of similar size. Then each of the subsets is sequentially used as the validation group, while being excluded from the calibration. The statistical parameter for the LMO cross validation has been used to indicate the predictive ability of a model. Generally, many authors consider a high $Q^2$ value as an indicator or even as the ultimate proof of the high predictive power of a QSRR model. Furthermore, a test set with no information used in QSRR model development was introduced for external validation. Then, the goodness of fit of the models was evaluated using the following statistical parameters: squared correlation coefficient for model fitting, $R^2$, squared correlation coefficient for cross validation, $Q^2$, squared correlation coefficient for test set, $R_{\text{test}}{}^2$, Fisher ratio value, $F$, and root mean square error, RMSE.

## 3. Results and discussion

### 3.1. Analysis by MLR

After descriptor selection, four small subsets that contained the maximum retention mapping information were extracted from an origin pool of 127 descriptors. These descriptors have been used to represent the relationships between molecular structures and retention indices by MLR, and then four best MLR models were built on OV101, DB5, OV17 and C20M using 4, 4, 6 and 6 molecular descriptors, respectively. All models were analyzed based on the criteria proposed by Golbraikh and Tropsha: (a) high value of cross-validated $Q^2$ value; (b) correlation coefficient $R$ between the predicted and the observed activities of compounds from an external test set close to 1; (c) at least one (but better both) of the correlation coefficients for regressions through the origin (predicted versus observed activities, or observed versus predicted activities) should be close to $R^2$ and (d) at least one slope of regression lines through the origin should be close to 1 [26].

#### 3.1.1. Model 1

Equation:

$$I = 1166.2 - 267.0(\pm 24.2)^{0}\chi + 347.9(\pm 25.4)^{1}\chi + 36.2(\pm 3.7)ndonr + 199.7(\pm 17.7)MW \quad (1)$$

Statistics and validation:

$$R^2 = 0.9605,\quad F = 1844,\quad \text{RMSEF} = 59.61,\quad Q^2 = 0.9595,$$
$$\text{RMSE}_{\text{cv}} = 60.30,\quad n = 230$$

Based on the criteria above, the high values of $R^2$, $F$ and $Q^2$ suggested that the generated model is robust and significant. Two topological descriptors ($^{0}\chi$ and $^{1}\chi$) and two constitutional descriptors (*ndonr* and *MW*) were included in model 1 on OV101 which is an apolar stationary phase packed with 100% dimethyl polysiloxane. We know that the main interaction between apolar stationary phase and apolar molecules, or apolar stationary phase and polar molecules, mainly depends on dispersion force and induction force, and dispersion force often play a leading role for majority of molecules while induction force is usually very small. As well-known, molecular mass and molecular deformability are responsible for dispersion force, the bigger molecular mass and the higher molecular deformability are, the stronger dispersion force is. In Eq. (1), the constitutional descriptor *MW* represents molecular mass which perform a positive effect on RI has been reported in several publications [18,27,28]. $^{0}\chi$ and $^{1}\chi$ proposed by Kier and Hall, usually known as Kier–Hall connectivity index, are calculated from the vertex degree of hydrogen-suppressed graph for zero-order path and one-order path. They are viewed as a measure of molecular branching, so the two descriptors encode the information about molecular shape and molecular deformability which can obviously influence dispersion force. Moreover, steric hindrance effect depending on molecular shape in gas chromatographic retention behavior also can be represented by $^{0}\chi$ and $^{1}\chi$. Besides that, hydrogen bond effect is another important factor on RI sometimes, especially when the presence of heteroatom, such as O, N and S with lone pair electrons, which can form hydrogen bond easily. Consequently, the constitutional descriptor *ndonr*, i.e. the number of hydrogen bond donor, was included in model 1. As a word, the molecular descriptors used in model 1 can well represent the structural information which related to the interaction between the solute and the apolar stationary phase OV101.

#### 3.1.2. Model 2

Equation:

$$I = 1150.6 - 269.2(\pm 19.1)^{0}\chi + 420.5(\pm 20.7)^{1}\chi + 48.0(\pm 3.2)ndonr + 179.0(\pm 14.0)MW \quad (2)$$

Statistics and validation:

$$R^2 = 0.9532,\quad F = 2062,\quad \text{RMSEF} = 61.05,\quad Q^2 = 0.9528,$$
$$\text{RMSE}_{\text{cv}} = 61.33,\quad n = 305$$

Model 2 is for 305 flavor compounds on column DB5, which is an apolar stationary phase packed with 5% phenyl 95% dimethyl poly siloxane. It is observed that the descriptors used in model 2 are the same as in model 1, perhaps because of that both OV101 and DB5 are apolar stationary phases, the interaction between the solute and the stationary phase is similar, so it can be reflected by the same molecular descriptors. One can find that the trend of effect on RI by each descriptor is consistent in Eqs. (1) and (2). $^1\chi$, *MW* and *ndonr* perform positive effect (with coefficient value of 347.93, 199.72, 36.91 and 420.49, 178.98, 48.01 in Eqs. (1) and (2), respectively) to a certain degree. This result is reasonable and understandable from the chemical meaning of these descriptors. As mentioned in model 1, *MW* and *ndonr* can represent information about molecular mass and the number of hydrogen bond donor which is positively correlated with dispersion force and hydrogen bond intensity. Hence, it is easy to understand their positive effects on RI. As to $^1\chi$, a measure of molecular branching, the value of $^1\chi$ is negative correlated with molecular branching, namely that the bigger $^1\chi$ value is, the less branched of corresponding molecule. As we know that with the increase of molecular branching, the molecular become more compacted, so the intermolecular contact area reduces and the molecular deformability decreases [29]. Based on this, the value of $^1\chi$ is also positively correlated with RI. However, in spite of reflecting similar structural information as $^1\chi$, $^0\chi$ performed a negative effect (with coefficient value of −297.0 and −29.91 in Eqs. (1) and (2), respectively). This is due to that each index defines a specific branching measure according to different algorithms, the connectivity index $^0\chi$ increase with the increase of branching, while $^1\chi$ decreases when the molecule becomes more compact (more branched) [30], so $^0\chi$ is negatively correlated with RI. The consideration from real experiences and chemical meaning may be very helpful to reliably reflect the role of descriptors on the predicted response.

### 3.1.3. Model 3

Equation:

$$I = 1230.0 + 49.0(\pm 7.2)DPSA1 - 75.2(\pm 4.5)FPSA1 + 57.7(\pm 12.6)qh\max + 67.3(\pm 9.6)\mu + 289.2(\pm 8.2)ipc \quad (3)$$

Statistics and validation:

$$R^2 = 0.9607,\quad F = 863,\quad \text{RMSEF} = 57.63,\quad Q^2 = 0.9595,\quad \text{RMSE}_{\text{cv}} = 58.55,\quad n = 165$$

Model 3 is built for 165 flavor compounds on OV17, a mid-polar stationary phase packed with 50% phenyl 50% methyl poly siloxane, using 6 molecular descriptors. Through analysis of descriptors, one can find that 5 new descriptors (*DPFA1*, *FPSA1*, *qhmax*, $\mu$ and *ipc*) arise and only the descriptor *ndonr* remained compared with model 1 and model 2. This change of descriptors in the model is mainly caused by the variation of stationary phase polarity. We know that the dipole–dipole interaction increases with the molecular polarity increasing, so the influence of directional force must be taken into account when we study the retention relationship between the solute and a polar stationary phase.

*FPSA1* and *DPSA1* belong to CPSA (charge partial surface area) descriptors proposed by Stanton and Jurs [31]. In fact, CPSA descriptors are of thirty different descriptors, which combine shape and electronic information to characterize molecules and therefore encode features responsible for polar interactions between molecules. In order to calculate CPSA descriptors, all atoms are viewed as hard sphere defined by the Van der Waals radius and the solvent-accessible surface area is used as molecular surface area. As an example, Fig. 1 shows the molecular shape and electronic density distribution after structure optimization for 2-methyl-3-furanthiol.

Two fundamental descriptors can then be defined as PPSA1 (partial positive surface area) and PNSA 1(partial negative surface area) based on the figure of molecular electronic density distribution such obtained. They are calculated as follows:

$$PPSA1 = \sum_{a+} SA_{a+}, \quad (4)$$

Here $SA_{a+}$ denotes the solvent-accessible surface area of positively charged atom. And

$$PNSA1 = \sum_{a-} SA_{a-}, \quad (5)$$

where $SA_{a-}$ represents the solvent-accessible surface area of negatively charged atom. In this study, two variables, named *DPSA1* and *FPSA1*, are used. They can be simply calculated as follows, respectively,

$$DPSA1 = PPSA1 - PNSA1 \quad (6)$$

and

$$FPSA1 = \frac{PPSA1}{SASA}. \quad (7)$$

Another descriptor, say *qhmax*, represents most positive charge on *H*, which is also a variable reflecting electron density distribution information. $\mu$ is molecular dipole moment. The bigger the value of $\mu$ is, the stronger the directional force is. Since dispersion force always working among any molecules, information about molecular mass, deformability is still included here. *Ipc*, proposed by Bonchev, is a comprehensive descriptor, representing the total information on distances in a molecular graph, which can discriminate molecules from molecular size and branching [32]. From the aspect of chemical structural information, here, *ipc* could be viewed as the combination of $^0\chi$, $^1\chi$ and *MW* in model 1 and model 2.

### 3.1.4. Model 4

Equation:

$$I = 1558.6 - 154.6(\pm 8.2)FPSA1 + 126.8(\pm 12.7)qh\max + 71.2(\pm 5.5)\mu - 53.2(\pm 5.4)^4\chi_c + 57.2(\pm 12.9)ndonr + 393.7(\pm 5.7)ipc \quad (8)$$

Statistics and validation:

$$R^2 = 0.9228,\quad F = 852,\quad \text{RMSEF} = 104.24,\quad Q^2 = 0.9223,\quad \text{RMSE}_{\text{cv}} = 104.58,\quad n = 330$$

Model 4 is developed for 330 compounds on C20M, a strong polar stationary phase packed with 100% poly ethylene glycol. Jurs once correlated molecular structure and gas chromatographic retention indices of 107 pyrazines on C20M employing CPSA descriptors [29]. However, CPSA descriptors are not enough if the samples observed with highly structural diversity. Compared with date set 3, there were more molecules in data set 4 and the polarity of stationary phase was also stronger. From Eqs. (3) and (8), one can find that most descriptors used in model 3 and 4 were the same, such as *FPSA1*, *qhmax*, $\mu$, *ndonr* and *ipc*, and the trend of influence by these five descriptors are consistent. It elucidated that the interaction between solute molecules and stationary phase OV17 is similar with the interaction between solute molecules and stationary phase C20M. Thus, we can describe it with the same descriptors. Besides that, the presence of $^4\chi_c$, this is also a Kier-Hall connectivity index for the type of 4 clusters of molecular graph that can reflect the complexity of molecules, may be caused by the presence of many structural complicated compounds in data set 4. There were 205 and 434 flavor compounds in data set 3 and data set 4, respectively, and more complicated compounds in data 4. Hence, $^4\chi_c$ was introduced to discriminate these compounds.

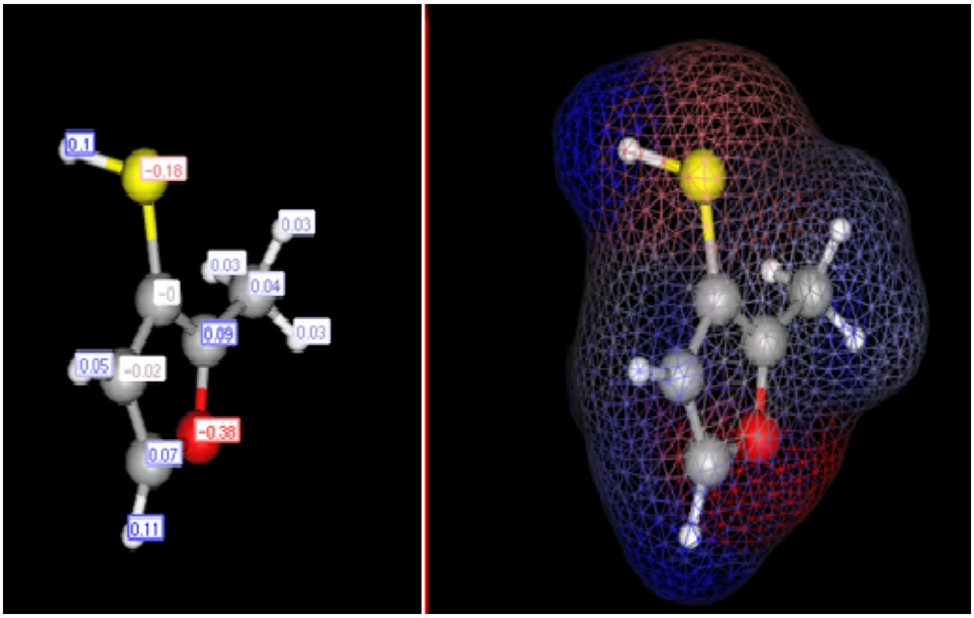

**Fig. 1.** Electronic density distribution for 2-methyl-3-furanthiol.

**Table 2**
Statistical parameters for four QSRR models.

| Column | Descriptors | $R^2$ | RMSEF | $Q^2$ | $\mathrm{RMSE}_{\mathrm{cv}}$ | ${R_{\mathrm{test}}}^2$ | $\mathrm{RMSE}_{\mathrm{test}}$ |
|---|---|---|---|---|---|---|---|
| OV101 | *Weight*, $^1\chi$, $^0\chi$, *ndonr* | 0.9605 | 59.61 | 0.9595 | 60.30 | 0.9585 | 58.02 |
| DB5 | *Weight*, $^1\chi$, $^0\chi$, *ndonr* | 0.9532 | 61.05 | 0.9528 | 61.33 | 0.9501 | 65.68 |
| OV17 | *Ipc*, $\mu$, *DPSA1*, *FPSA1*, *ndonr*, *qhmax* | 0.9607 | 57.63 | 0.9595 | 58.55 | 0.9581 | 54.33 |
| C20M | *Ipc*, *ndonr*, *FPSA1*, $\mu$, $^4\chi c$, *qhmax* | 0.9228 | 104.24 | 0.9223 | 104.58 | 0.9255 | 105.48 |

### 3.2. External validation

For validation purposes, four external validation sets randomly selected from the four sets of original sample were constructed, respectively. Fig. 2 shows the experimental RIs versus the prediction of RIs for all compounds, the training sets and the test sets. In all cases, the $R^2$, $Q^2$, ${R_{\mathrm{test}}}^2$ and RMSE to assess the quality of the models were calculated, the results can be found in Table 2. The values of cross-validated $Q^2$ bigger than 0.9500 (except for strong polar column C20M, $Q^2 = 0.9223$), besides that, all the $Q^2$ and the ${R_{\mathrm{test}}}^2$ are close to the corresponding $R^2$ on four columns. Based on the criteria proposed by Golbraikh and Tropsha, these models are robust and significant.

### 3.3. Contribution analysis of descriptors

One of the aims of this paper is to investigate the contributions of different descriptors in different conditions. As we know, the diversity of molecular structure can affect descriptor selection. Hence, the results will be more provable if molecular structure diversity factor be excluded. Based on this, 107 common compounds from four columns were extracted, and then MLR models were performed for these compounds with the 10 descriptors used in 4 models. Model parameters can be found in Table 3.

**Table 3**
Statistical parameters for four QSRR models of 107 common molecules.

| Column | $R^2$ | RMSEF | $Q^2$ | $\mathrm{RMSE}_{\mathrm{cv}}$ |
|---|---|---|---|---|
| OV101 | 0.9718 | 56.2714 | 0.9587 | 68.1360 |
| DB5 | 0.9741 | 54.0304 | 0.9619 | 65.5337 |
| OV17 | 0.9597 | 67.6811 | 0.9464 | 78.0032 |
| C20M | 0.9309 | 105.7966 | 0.9084 | 121.8397 |

In order to find the structural features that would be important to the RIs based on different stationary phases, the relative and fraction contribution of each index is estimated. The relative contribution ($\Psi_r$) and fraction contribution ($\Psi_f$) of the corresponding descriptor to RI are defined as follows [33]:

$$\psi_r(i) = a_i X_i \tag{9}$$

$$\psi_f(i) = \frac{r^2 |\psi_r(i)|}{\sum_i |\psi_r(i)|} \times 100\% \tag{10}$$

where $a_i$ and $X_i$ are the coefficient and the average value of the $i$th descriptor in the model and $r_2$ is the coefficient of the determination of the model. The sum is over all indices in the model. The results for above four models are listed in Table 4.

From Table 4 we can find that the contributions of individual molecular descriptors to the four stationary phases cover a wide range of $\Psi_f$ values which are depending on the polarity of the columns. For all columns, $^0\chi$, $^1\chi$ and *MW* make a major contribution to RIs, and the average of $\Psi_{\mathrm{f}}$ values is 0.2230, 0.3651 and 0.1553, respectively, which is far bigger than other descriptors. The results elucidate that the size, shape and deformability of a molecule, which reflect dispersion force, induction force and steric hindrance effect, always play a dominant role in determining RIs on all stationary phases with different polarities, because $^0\chi$, $^1\chi$ and *MW* characterize the information of these three structural features with a positive correlation. As to descriptor *ipc*, its fraction contribution is not so significant due to several descriptors contain similar information exist in the models. On the other hand, *FPSA1*, *qhmax*, $\mu$ and *DPSA1* have smaller contributions to RIs relying on the polarity of columns, and the average of $\Psi_f$ values is 0.0205, 0.0519, 0.0526 and 0.0157, respectively. Interaction between solute and stationary phase depend not only on dispersion force, induction force and steric effect but also on dipole–dipole interaction namely

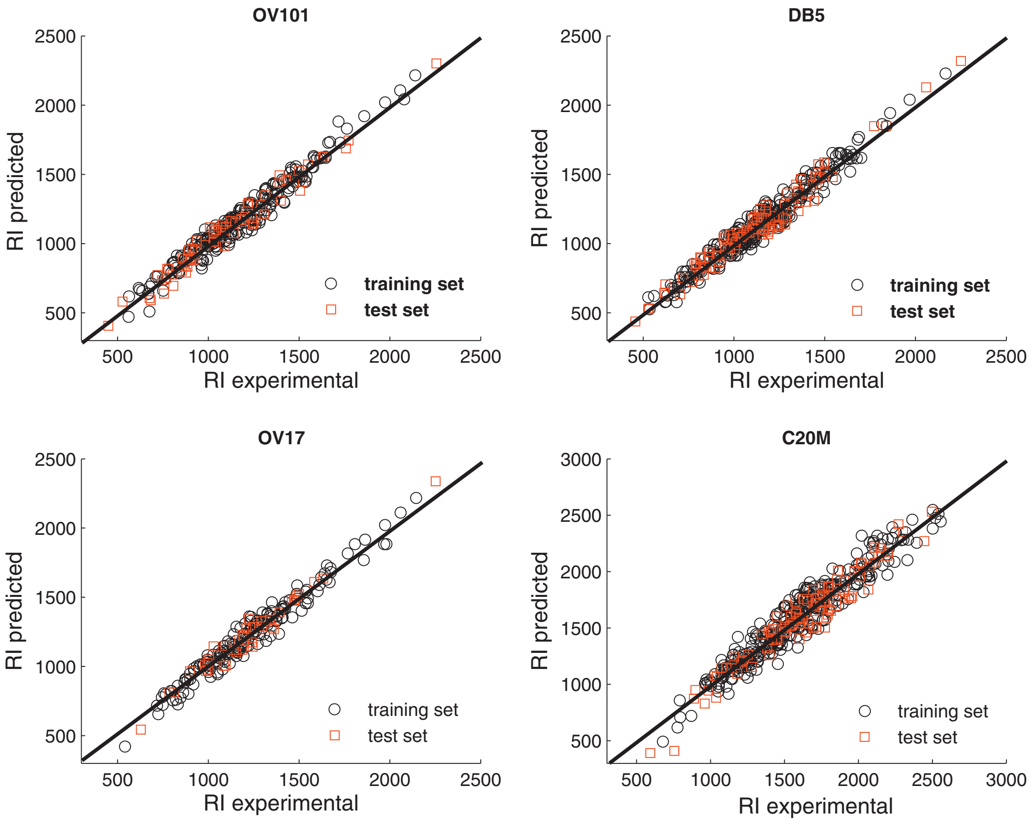

**Fig. 2.** Plot of RI experimental vs. predicted for training set and test set.

directional force, especially on polar stationary phases. From the variation of $\Psi_f$ values of *FPSA1*, *qhmax* $\mu$ *and DPSA1*, one can observe that the $\Psi_f$ values from these descriptors increase with the column polarity increasing, 0.0175–0.0363 for *FPSA1*, and 0.0091–0.1122 for *qhmax*, 0.0357–0.0576 for $\mu$ and 0.0045–0.0056 for *DPSA1*. The reason may be that polar interaction between the solute and the stationary phases become stronger with the increasing polarity of the columns, whereas the relative importance of electron density distribution or molecular polarizability encoded by *FPSA1*, *qhmax*,$\mu$ *and DPSA1* will more and more strong. On the contrary, the $\Psi_f$ values of $^0\chi$, $^1\chi$, *MW* and *ipc*, on the whole, steadily decrease with the increasing polarity of the columns. However, they always keep a dominant status compared with other descriptors. It should be noted that the decrease of the $\Psi_f$ values of $^0\chi$, $^1\chi$, *MW* and *ipc* is not due to their weak effects in polar stationary phases but a relative decrease caused by electrical related descriptors' influence. Finally, there are two special descriptors *ndonr* and $^4\chi_c$ which have very small contributions to RIs with average $\Psi_f$ values 0.0100 and 0.0035, respectively, and the variety of $\Psi_f$ values is not obvious as others. This may be due to that the two descriptors are not valid for all molecular structures but a fraction of them, for example, the compounds including more hetero atoms or the compounds with highly complicated structure, so they can be viewed as local variables, making big contribution to RIs for some compounds but nothing for others. Hence, the values of $\Psi_f$ is very small in models but cannot absolutely be excluded. Figs. 3–5 depict scatter plots of the $\Psi_f$ values of different descriptors against the stationary phase with different polarities.

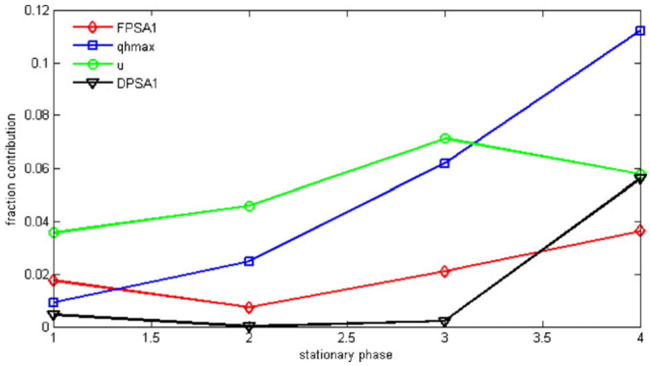

**Fig. 3.** Plot of the $\Psi_f$ values of electrical related descriptors against the stationary phase with different polarity.

From these figures, we can observe the different influences of electrical related descriptors (including *FPSA1*, *qhmax*, $\mu$ and *DPSA1*) and topological descriptors (including $^0\chi$, $^1\chi$, *MW* and *ipc*,

**Table 4**
The fraction contribution of individual descriptor to RI in four models.

| Column | *FPSA1* | *qhmax* | $\mu$ | *ndonr* | $^0\chi$ | $^1\chi$ | *Weight* | *Ipc* | $^4\chi_c$ | *DPSA1* |
|---|---|---|---|---|---|---|---|---|---|---|
| OV101 | 0.0175 | 0.0091 | 0.0357 | 0.0248 | 0.2035 | 0.4363 | 0.1405 | 0.0928 | 0.0070 | 0.0045 |
| DB5 | 0.0074 | 0.0248 | 0.0459 | 0.0120 | 0.2317 | 0.4025 | 0.1862 | 0.0626 | 0.0008 | 0.0001 |
| OV17 | 0.0210 | 0.0617 | 0.0712 | 0.0006 | 0.2422 | 0.3720 | 0.1685 | 0.0183 | 0.0021 | 0.0021 |
| C20M | 0.0363 | 0.1122 | 0.0576 | 0.0027 | 0.2145 | 0.2498 | 0.1258 | 0.0718 | 0.0042 | 0.0560 |
| Average | 0.0205 | 0.0519 | 0.0526 | 0.0100 | 0.2230 | 0.3651 | 0.1553 | 0.0614 | 0.0035 | 0.0157 |

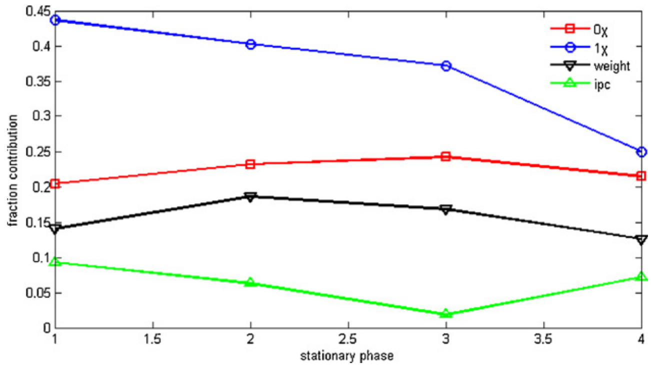

**Fig. 4.** Plot of the $\Psi_f$ values of topological descriptors against the stationary phase with different polarity.

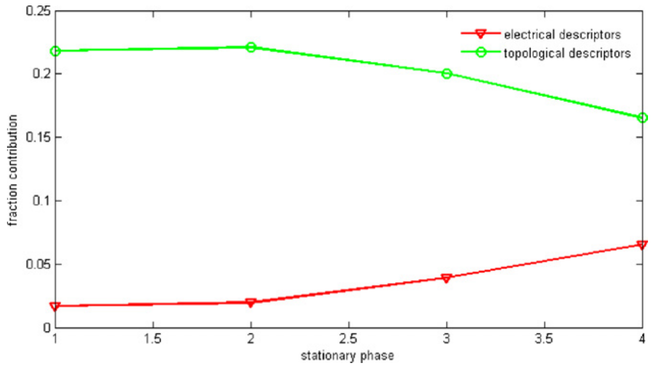

**Fig. 5.** Plot of the average $\Psi_f$ values of electrical related descriptors and topological descriptors against the stationary phase with different polarity.

it should be noted that *MW* often be seen as constitutional descriptor, it is viewed as topological descriptor for convenience here), and also find the variations of fraction contributions for each descriptor in QSRR models with the increase of stationary phase polarity. The variations for some descriptors are not consistent as expected, for example, the $\Psi_f$ value of $\mu$ in model 4 is smaller than in model3, due to the interaction among different descriptors with similar information about molecular structure. However, the main variation for electrical related descriptors and topological descriptors is quite clear as reflected in Fig. 5. In order to find more obviously changing trends for different descriptors, other four models were built with fewer descriptors to avoid strong interaction among descriptors, i.e. two electrical related descriptors *FPSA1* and $\mu$, a topological descriptor *ipc* and a descriptor about hydrogen bond *ndonr*. The $R^2$s of four models are 0.9431, 0.9428, 0.9176 and 0.9000, respectively. Then, the same analysis as previous was performed and the results were showed in Fig. 6. One can observe that *FPSA1* (the $\Psi_f$ value 0.0665, 0.0777, 0.1034 and 0.2136) and $\mu$ increase obviously (the $\Psi_f$ value 0.0613, 0.0773, 0.1295 and 0.1308) with the increase of polarity while *ipc* (the $\Psi_{\text{f}}$ value 0.8018, 0.7840, 0.6684 and 0.4686) is on the contrary.

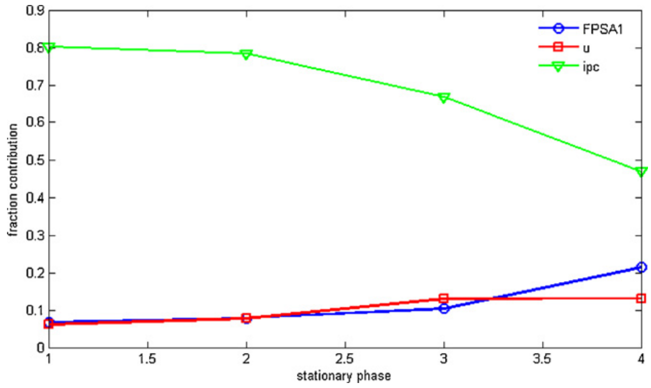

**Fig. 6.** Plot of the $\Psi_f$ values of three different descriptors against the stationary phase with different polarity.

## 4. Conclusion

Although great diversity the experimental data shows in flavor compounds structures, four models for the prediction of Kovàts retention index were developed on stationary phase OV101, DB5, OV17 and C20M with different polarities. The models have good predictive capacity and statistical parameters by internal and external validation. Study on the prediction of retention index for flavor compounds are rarely reported before, the four models proposed are useful for further investigation about the gas chromatographic retention behavior of flavor compounds in future. Furthermore, the contributions of different molecular descriptors in different conditions were analyzed. The molecular descriptors used in this paper can well reflect the interactions between the solute and the stationary phases, i.e. dispersion force, directional force, induction force, ability of H bond donation and steric effect. Results showed that molecular descriptors that encode information of molecular size, shape and deformability make a major contribution in all cases, and the influence of molecular descriptors which are responsible for polarizability and electron density distribution become more and more important with the column polarity increasing. Besides that, some local variables such as *ndonr* and ${}^4\chi_c$ also make some contributions to retention index depending on the structural features of compounds. Compared with other works, highly structural diversity of samples and four stationary phases of different polarity were more helpful to investigate the interaction between the solute and stationary phase on a more general level.

## Acknowledgements

This work is financially supported by the National Nature Foundation Committee of PR China (Grant Nos. 20875104 and 21075138) and the international cooperation project on Traditional Chinese Medicines of Ministry of Science and Technology of China (Grant No. 2007DFA40680).

## Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.chroma.2011.12.020.

## References

[1] http://www.flavornet.org/.
[2] M. Ávila, M. Zougagh, A. Escarpa, Á. Ríos, J. Chromatogr. A 1216 (2009) 7179.
[3] D. Saison, D.P. De Schutter, B. Uyttenhove, F. Delvaux, F.R. Delvaux, Food Chem. 114 (2009) 1206.
[4] S. Rochat, J. Egger, A. Chaintreau, J. Chromatogr. A 1216 (2009) 6424.
[5] D. Ryan, R. Shellie, P. Tranchida, A. Casilli, L. Mondello, P. Marriott, J. Chromatogr. A 1054 (2004) 57.
[6] S.M. Rocha, E. Coelho, J. Zrostlíková, I. Delgadillo, M.A. Coimbra, J. Chromatogr. A 1161 (2007) 292.
[7] E.Sz. Kovàts, Helv. Chim. Acta 41 (1958) 1915.
[8] H. Van Den Dool, P.D. Kratz, J. Chromatogr. 11 (1963) 463.
[9] V.I. Babushok, I.G. Zenkevich, Chromatographia 69 (2009) 257.
[10] R. Kaliszan, Quantitative Structure–Chromatographic Retention Relationships, Wiley, New York, 1987.
[11] F. Luan, C.X. Xue, R.S. Zhang, C.Y. Zhao, M.C. Liu, Z.D. Hu, B.T. Fan, Anal. Chim. Acta 537 (2005) 101.
[12] H.X. Zhao, X.Y. Xue, Q. Xu, F.F. Zhang, X.M. Liang, J. Chromatogr. A 1107 (2006) 248.
[13] R.J. Hu, H.X. Liu, R.S. Zhang, C.X. Xue, Talanta 68 (2005) 31.

[14] R.D. de Mello Castanho Amboni, B. da Silva Junkes, V.E.F. Heinzen, R.A. Yunes, Theochem 579 (2002) 53.
[15] Svein A. MjØs, O. Grahl-Nielsen, J. Chromatogr. A 1110 (2006) 171.
[16] V.A. Isidorov, L. Szczepaniak, J. Chromatogr. A 1216 (2009) 8998.
[17] C.H. Lu, W.M. Guo, C.S. Yin, Anal. Chim. Acta 561 (2006) 96.
[18] K. Héberger, M. Görgényi, J. Chromatogr. A 845 (1999) 21.
[19] K. Héberger, M. Görgényi, M. Sjöström, Chromatographia 51 (2000) 595.
[20] Z. Garkani-Nejad, M. Karlovits, W. Demuth, J. Chromatogr. A 1028 (2004) 287.
[21] K. Héberger, J. Chromatogr. A 1158 (2007) 273.
[22] O. Farkas, I.G. Zenkevich, F. Stout, J.H. Kalivas, K. Héberger, J. Chromatogr. A 1198–1199 (2008) 188.
[23] J.M. Sutter, T.A. Peterson, P.C. Jurs, Anal. Chim. Acta 342 (1997) 113.
[24] B. Ren, Chemom. Intell. Lab. Syst. 66 (2003) 29.
[25] F. Liu, Y.Z. Liang, C.Z. Cao, N. Zhou, Anal. Chim. Acta 594 (2007) 279.
[26] A. Golbraikh, A. Tropsha, J. Mol. Graph. Modell. 20 (2002) 269.
[27] M. Jalali-Heravi, M.H. Fatemi, J. Chromatogr. A 897 (2000) 227.
[28] J. Ruther, J. Chromatogr. A 890 (2000) 313.
[29] X.H. Du, C.J. Feng, Chin. J. Anal. Chem. 31 (2003) 486.
[30] J. Devillers, A.T. Balaban, Topological Indices and Related Descriptors in QSAR and QSPR, Gordon and Breach Science Publishers, The Netherlands, 1999, p. 112.
[31] C.G. Georgakopoulos, J.C. Kiburis, P.C. Jurs, Anal. Chem. 63 (1991) 2012.
[32] D. Bonchev, N. Trinajstic, J. Chem. Phys. 67 (1977) 4517.
[33] D.E. Needham, I.-C. Wei, P.G. Seybold, J. Am. Chem. Soc. 110 (1988) 4186.