# Data Mining for Seeking an Accurate Quantitative Relationship between Molecular Structure and GC Retention Indices of Alkenes by Projection Pursuit

Yiping Du,[†,§] Yizeng Liang,*[,‡] and Dong Yun[‡]

Institute of Chemometrics and Chemical Sensing Technology, Hunan University, Changsha 410082, P. R. China, College of Chemistry and Chemical Engineering, Central South University, Changsha, 410083, P. R. China, and College of Chemical Engineering, Shandong University of Science and Technology, ZiBo, 255012, P. R. China

Received March 6, 2002

Primary data mining on alkenes for seeking an accurate quantitative relationship between the molecular structure and retention indices of gas chromatography is developed in this paper. Based on the results obtained from projection pursuit, all alkenes investigated show an interesting classification. Thus, a new variable named class distance variable of alkenes, which essentially describes information about the branch, position of the double bonds, the number of double bonds, and so on for alkenes, is proposed. With the help of the new variable, both fitting and prediction accuracy of the regression model can be dramatically improved. The results obtained in this work show that the technique of projection pursuit developed in statistics is a quite promising tool for seeking an accurate quantitative structure-retention relationship (QSRR).

## INTRODUCTION

Many papers on the relationship between chemical structure of alkenes and their GC retention indices have been published since 1972.[1−16] Sojak and co-workers[1−12] have reported a series of work, in which retention behavior of the isomers of *n*-alkene and *n*-alkadiene was shown being dependent on the number of carbon atoms in the molecule, the position of double bond(s), geometric isomerism, and column temperature. Retention indices of homologous series show certain regularity with the increase of number of carbon atoms in the molecules. They also found that retention anomalies are expected for compounds with molecular structures that make it possible to form a ring conformation of the propyl group with the $\pi$-electron system of the remainder of the molecule, such as 1-pentene and 1,*trans*-3-heptadiene. That phenomenon was called propyl effect.[13]

Researching for structure-retention correlation is helpful in identifying compounds with similar retention indices and predicting retention indices. Since the mid 1980s, the study of quantitative structure retention relationship (QSRR) between structural descriptors, such as topological descriptors and quantum chemical descriptors, and retention has become a hot field.[13−16] By building a structure-retention relationship model, one can predict retention indices and identify compounds. Rohrbaugh and Jurs[14] used 4 descriptors selected from 33 to build some models for expressing descriptor-retention index correlation of alkenes for 3 column types. "Excellent fit of the experimental values" has been obtained with the number of compounds $n = 86$, the correlation coefficient $R = 0.997$ or 0.996, standard error $S = 7.19$ to 7.78, and $F = 3233$ to 3509. In the paper, authors introduced the boiling point of alkenes as a variable (descriptor) into the regression model. It is well-known that the retention index has a high correlation with the boiling point. The authors also stated that "the boiling point was found to be the major contributor to the high correlation". However, the boiling point is a measuring parameter, not structural descriptor! If only the structural descriptors, such as topological indices and quantum chemical descriptors, were included in the model without using the boiling point, the regression model would not be so good.

Twenty topological and quantum chemical descriptors were used to establish a regression model for the retention index of alkenes in our laboratory, the results are quite unsatisfactory (see details later). Furthermore, even with a correlation coefficient $R$ higher than 0.996, the model may not be said to be good enough. Buja et al. have pointed out that careful and creative regression modeling yielded fits with good global properties ($R^2 = 0.995$), but there were still unacceptably large residuals and poor performance on cross-validation tests.[17] As in the work[14] of Rohrbaugh and Jurs, many residuals are more than 10 i.u., the maximum residual reaches to 24 i.u., which is 5 times bigger than measurement errors in the GC retention index. Thus, the measurement errors should be taken into account when we establish some statistic models in order to make their usage in practice. How to find a regression model whose residuals can be reduced to the level of the measurement error is the major concern of this work. Furthermore, is it possible for us to only use the information merely from the chemical structure to predict the retention behavior of alkenes? Is it possible for us to give a simple picture to explain the main factors influencing the retention behaviors of alkenes? The main purpose of this work lies in these aspects.

A new technique, named knowledge discovery in databases (KDD) or data mining (DM) has become more pervasive in chemical research,[18−25] with the rapid increase of the

* Corresponding author phone: 86-731-8825637; fax: 86-731-8825637; e-mail: yzliang@public.cs.hn.cn. Corresponding author address: Institute of Chemometrics and Intelligent Analytical Instruments, College of Chemistry and Chemical Engineering, Central South University, Changsha 410083, P. R. China.

[†] Hunan University.

[‡] Central South University.

[§] Shandong University of Science and Technology.

10.1021/ci020285u CCC: $22.00 
Published on Web 10/03/2002

information on chemical compounds, such as the spectral databases, chromatographic databases, or a large amount of data on molecular structures and their properties. The DM aims to discover something new from the facts recorded in the databases or huge amount of data collected.[22−25] The major idea of data mining lies in that it combines the techniques both from classification and regression. Recently, with the help of projection pursuit technique we found an interesting classification for the alkane compounds and then proposed a new class distance variable based on the classification to establish a regression model.[18] With the help of information from the class distance variable, the regression model was improved dramatically. An accurate quantitative relationship between molecular structure and gas chromatographic retention index for alkanes was obtained.[18]

In this work, projection pursuit technique developed in statistics is used to establish the quantitative relationship between molecular structure and retention index for alkenes. A similar class distance variable for alkenes is proposed. The accuracy for both fitting and prediction of the regression model can be dramatically improved by introducing such a regression variable.

## THEORY AND METHODOLOGY

**Projection Pursuit.** Projection pursuit (PP) is a useful tool developed in statistics for seeking "interesting" projections to try to find the intrinsic structure hidden in the high dimensional data.[26] Once one could fortunately find some interesting projection direction that gives the clear intrinsic data structure on low dimension space, it is possible for one to further obtain the good fitting results for regression[27] and/or good classification picture in lower dimension for visual pattern recognition.[28] The aim of this paper is to find some "interesting" projections to reveal the data structure of both the descriptor matrix **X** and the retention index vector **y**, thus, to seek an efficient method to build an accurate quantitative relationship between molecular structure and retention indices of the alkenes collected. The matrix **X** containing the descriptors of compound structures, which are obtained from some kind of rational calculation without any measurement noise, is usually used to do the projections. However, this data structure can hardly provide some useful information about relationship between **X** and **y**, if one uses simply matrix **X** to do so. If we simply include the retention index vector **y** into the matrix **X** and then do projections on this augment matrix, the data structure is just similar to the one obtained merely from **X**. Thus, the key step is to find out a good projection direction and a reasonable projection index.

We propose the following projection index of entropy function form

$$\text{minimize } \Sigma\xi = -\,p_i \log(p_i) \tag{1}$$

where $p_i = m_i/m$, $m_i$ is the number of compounds whose $\alpha_j$ ($j = 1, ...m$) is falling into the interval [$i$*0.05, ($i$+1)*0.05] ($i = 0, 1, ..., 39$) and $m$ is the number of compounds, and

$$\alpha_j = [(\mathbf{x}_j^{\mathrm{T}}\mathbf{a} - y_j) - \min(\mathbf{x}_j^{\mathrm{T}}\mathbf{a} - y_j)]/ [\max(\mathbf{x}_j^{\mathrm{T}}\mathbf{a} - y_j) - \min(\mathbf{x}_j^{\mathrm{T}}\mathbf{a} - y_j)],\ (j = 1, ..., m) \tag{2}$$

with

$$\mathbf{a} = (\mathbf{X}^{(\alpha)\mathrm{T}}\mathbf{X}^{(\alpha)})^{-1}\,\mathbf{X}^{(\alpha)\,\mathrm{T}}\mathbf{y}^{(\alpha)} \tag{3}$$

subject to correlation coefficient $R \geq 0.999$. Here **a** is a vector representing the projection direction. Instead of taking all the samples into regression calculation in eq 3, we here only include a very small part of the samples, say 2 or 3 times the number of variables, to construct $\mathbf{X}^{(\alpha)}$ and $\mathbf{y}^{(\alpha)}$. The reason we to select a small subset of samples to calculate the projection direction vector **a** lies in that there may be diverse structures in alkenes, which intimates that there may be several different classes in alkenes. If the projection direction vector **a** is good enough, then the index $\xi$ defined in eq 1 will reach a small value, which hints that the distribution of $\alpha_j$ ($j = 1, ..., m$) should be of some structure instead of just uniformly scattering in the interval [0, 1]. Notice that if the values of $\alpha_j$ for all compounds show some different classes, the differences between $\alpha_j$ will represent some kind of distance between classes in alkenes. Thus, we can further use the informative projection plot of $\alpha_j$ to do regression by the following formula

$$\hat{\mathbf{y}} = \Sigma f_i(\mathbf{x}_j, \alpha_j) \tag{4}$$

where $f_i$ are some kind function to be found.

**Calculation Procedure.** The projection pursuit procedure can be completed in the following steps:

(1) Select randomly k rows from data matrix **X** and **y**, that is, randomly select k samples from the alkenes to construct $\mathbf{X}^{(\alpha)}$ and $\mathbf{y}^{(\alpha)}$;

(2) Obtain one possible projection direction vector **a** using eq 3;

(3) If the correlation coefficient between $\mathbf{X}^{(\alpha)}$ and $\mathbf{y}^{(\alpha)}$ is greater than 0.999, calculate $\alpha_j$ ($j = 1, ..., m$) using eq 2, otherwise, go back to step (1);

(4) Take count of $\alpha_j$ ($j = 1, ...m$) which falls into the interval [$i$*0.05, ($i$+1)*0.05] ($i = 0, 1, ..., 39$), respectively;

(5) Calculate the projection index $\xi$ using eq 1;

(6) Sort the indices obtained with the increasing order and keep 100 smallest indices in computer.

(7) Go back to step (1) until the number of loops reaches $10^8$.

## DATA COLLECTION AND DESCRIPTORS OF THE ALKENES

**Data Collection.** More than 2000 retention indices of alkenes on the column of squalane and at different temperatures collected from different sources[2−4,29−46] were picked out from the GC retention index database.[47] Then, a very careful check of possible mistakes both in data transformation and different sources was conducted.[47] Notice that the retention indices of alkenes collected are run at different temperatures, and the calibration of temperature for some compounds is necessary. We used the simple regression method to calibrate the values measured at some temperatures to a fixed temperature 60 °C. With the data of the same compound measured from different laboratories at hand, we could easily estimate the measurement errors of the retention indices of gas chromatography. In general, the varying level of the measurement errors of the GC retention indices for alkenes is within ±4 index units (i.u.). Only a few of the n-monoenes have errors up to 6 i.u. because they have only one retention index at a fixed temperature (130 °C) and are

**Table 1.** Value of Retention Index and Class Distance Variable of Alkanes

| compound[a] | I | $C_D$ | compound[a] | I | $C_D$ | compound[a] | I | $C_D$ | compound[a] | I | $C_D$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 10m-1-C12 | 1253.10 | 0.3637 | 3m-3e-1-C5 | 743.60 | 0.6456 | 1-C14 | 1383.50 | 0.1173 | c-3-C14 | 1382.60 | 0.2334 |
| 11m-1-C12 | 1247.20 | 0.3506 | 3m-t-2-C7 | 798.40 | 0.4845 | 1-C15 | 1484.70 | 0.1173 | c-3-C15 | 1485.30 | 0.2334 |
| 22m2-t-3-C6 | 692.60 | 0.7919 | 3m-t-2-C6 | 694.30 | 0.4845 | 1-C11 | 1081.10 | 0.1173 | c-3-C11 | 1083.30 | 0.2334 |
| 22m2-c-3-C6 | 717.80 | 0.7458 | 3m-t-2-C5 | 613.10 | 0.4845 | 1-C5 | 481.80 | 0.1173 | c-3-C8 | 788.40 | 0.2334 |
| 233m3-1-C4 | 629.90 | 0.8799 | 3m-t-3-C7 | 784.20 | 0.4885 | 1-C8 | 781.40 | 0.1173 | c-4-C10 | 980.90 | 0.2379 |
| 233m3-1-C5 | 736.20 | 0.8799 | 3m-t-3-C6 | 685.60 | 0.4885 | 1-C2 | 177.70 | 0.1173 | c-4-C9 | 883.50 | 0.2379 |
| 234m3-2-C5 | 766.30 | 0.9693 | 3m-c-2-C7 | 789.00 | 0.4925 | t-2-C4 | 406.50 | 0.2406 | c-4-C12 | 1176.60 | 0.2379 |
| 23m2-1-C4 | 559.40 | 0.6065 | 3m-c-2-C6 | 700.00 | 0.4925 | t-2-C7 | 698.60 | 0.2406 | c-4-C16 | 1578.80 | 0.2379 |
| 23m2-1-C6 | 739.90 | 0.6065 | 3m-c-2-C5 | 603.50 | 0.4925 | t-2-C10 | 996.80 | 0.2406 | c-4-C13 | 1275.30 | 0.2379 |
| 23m2-1-C5 | 651.60 | 0.6065 | 3m-c-3-C7 | 777.70 | 0.4957 | t-2-C6 | 596.90 | 0.2406 | c-4-C14 | 1375.50 | 0.2379 |
| 23m2-2-C4 | 625.60 | 0.7172 | 3m-c-3-C6 | 684.20 | 0.4957 | t-2-C9 | 896.60 | 0.2406 | c-4-C15 | 1479.30 | 0.2379 |
| 23m2-2-C6 | 789.50 | 0.7172 | 3e-1-C6 | 738.30 | 0.3787 | t-2-C12 | 1196.96 | 0.2406 | c-4-C11 | 1077.90 | 0.2379 |
| 23m2-2-C5 | 704.00 | 0.7172 | 3e-1-C5 | 648.70 | 0.3787 | t-2-C16 | 1597.00 | 0.2406 | c-4-C8 | 787.00 | 0.2379 |
| 23m2-t-3-C6 | 751.30 | 0.7325 | 3e-2-C5 | 697.70 | 0.4853 | t-2-C13 | 1297.30 | 0.2406 | c-5-C10 | 979.90 | 0.2483 |
| 23m2-c-3-C6 | 749.20 | 0.7364 | 3e-3-C6 | 772.70 | 0.4981 | t-2-C14 | 1397.60 | 0.2406 | c-5-C12 | 1172.52 | 0.2483 |
| 244m3-1-C5 | 705.50 | 0.8659 | 3e-t-2-C6 | 781.80 | 0.5006 | t-2-C15 | 1497.00 | 0.2406 | c-5-C16 | 1573.00 | 0.2483 |
| 244m3-2-C5 | 715.90 | 0.9898 | 3e-c-2-C6 | 783.60 | 0.4973 | t-2-C11 | 1096.50 | 0.2406 | c-5-C13 | 1271.20 | 0.2483 |
| 24m2-1-C7 | 829.80 | 0.6080 | 44m2-1-C6 | 724.60 | 0.6516 | t-2-C5 | 500.20 | 0.2406 | c-5-C14 | 1369.30 | 0.2483 |
| 24m2-1-C6 | 739.00 | 0.6080 | 44m2-1-C5 | 606.00 | 0.6516 | t-2-C8 | 797.90 | 0.2406 | c-5-C15 | 1473.50 | 0.2483 |
| 24m2-1-C5 | 638.50 | 0.6080 | 44m2-t-2-C6 | 747.60 | 0.7673 | t-3-C7 | 687.50 | 0.2325 | c-5-C11 | 1075.30 | 0.2483 |
| 24m2-2-C6 | 730.80 | 0.7507 | 44m2-t-2-C5 | 614.80 | 0.7673 | t-3-C10 | 985.30 | 0.2325 | c-6-C12 | 1171.63 | 0.2558 |
| 24m2-2-C5 | 640.50 | 0.7507 | 44m2-c-2-C6 | 743.80 | 0.7504 | t-3-C6 | 592.20 | 0.2325 | c-6-C16 | 1568.10 | 0.2558 |
| 24m2-t-3-C6 | 728.00 | 0.7557 | 44m2-c-2-C5 | 637.10 | 0.7504 | t-3-C9 | 886.70 | 0.2325 | c-6-C13 | 1268.40 | 0.2558 |
| 24m2-c-3-C6 | 725.40 | 0.7604 | 45m2-1-C6 | 735.00 | 0.5897 | t-3-C12 | 1184.79 | 0.2325 | c-6-C14 | 1365.80 | 0.2558 |
| 25m2-1-C6 | 741.50 | 0.5937 | 45m2-t-2-C6 | 735.30 | 0.7307 | t-3-C16 | 1584.90 | 0.2325 | c-6-C15 | 1470.00 | 0.2558 |
| 25m2-2-C6 | 750.20 | 0.7188 | 45m2-c-2-C6 | 736.80 | 0.7280 | t-3-C13 | 1284.70 | 0.2325 | c-7-C16 | 1565.00 | 0.2611 |
| 25m2-t-3-C6 | 694.80 | 0.7369 | 4m-1-C7 | 746.90 | 0.3604 | t-3-C14 | 1384.10 | 0.2325 | c-7-C14 | 1363.80 | 0.2611 |
| 25m2-c-3-C6 | 701.00 | 0.7256 | 4m-1-C6 | 658.90 | 0.3604 | t-3-C15 | 1485.30 | 0.2325 | c-7-C15 | 1467.00 | 0.2611 |
| 2m-1-C3 | 383.00 | 0.3520 | 4m-1-C12 | 1242.20 | 0.3604 | t-3-C11 | 1085.70 | 0.2325 | c-8-C16 | 1564.30 | 0.2616 |
| 2m-1-C4 | 488.00 | 0.3520 | 4m-1-C5 | 550.20 | 0.3604 | t-3-C8 | 788.80 | 0.2325 | 1,3-C4 | 386.79 | 0.1244 |
| 2m-1-C7 | 776.40 | 0.3520 | 4m-2p-1-C5 | 822.90 | 0.5990 | t-4-C10 | 981.70 | 0.2356 | 1,4-C5 | 463.76 | 0.1244 |
| 2m-1-C6 | 678.90 | 0.3520 | 4m-2e-1-C5 | 737.00 | 0.5876 | t-4-C9 | 884.00 | 0.2356 | 1,5-C6 | 563.86 | 0.1244 |
| 2m-1-C12 | 1274.90 | 0.3520 | 4m-3e-t-2-C5 | 756.90 | 0.7310 | t-4-C12 | 1179.03 | 0.2356 | 1,6-C7 | 664.20 | 0.1244 |
| 2m-1-C5 | 580.50 | 0.3520 | 4m-3e-c-2-C5 | 768.20 | 0.7104 | t-4-C16 | 1579.70 | 0.2356 | 1,7-C8 | 764.17 | 0.1244 |
| 2m-2-C4 | 514.40 | 0.4729 | 4m-t-2-C7 | 750.70 | 0.4956 | t-4-C13 | 1279.20 | 0.2356 | 1,8-C9 | 862.70 | 0.1244 |
| 2m-2-C7 | 789.50 | 0.4729 | 4m-t-2-C6 | 657.30 | 0.4956 | t-4-C14 | 1378.40 | 0.2356 | 1,9-C10 | 964.00 | 0.1244 |
| 2m-2-C6 | 691.30 | 0.4729 | 4m-t-2-C5 | 561.40 | 0.4956 | t-4-C15 | 1480.20 | 0.2356 | 1,t-3-C7 | 713.10 | 0.1464 |
| 2m-2-C5 | 597.90 | 0.4729 | 4m-t-3-C7 | 778.90 | 0.4877 | t-4-C11 | 1079.90 | 0.2356 | 1,t-3-C10 | 1008.50 | 0.1464 |
| 2m-3e-1-C5 | 736.20 | 0.6325 | 4m-c-2-C7 | 746.20 | 0.5021 | t-4-C8 | 783.70 | 0.2356 | 1,t-3-C6 | 614.16 | 0.1464 |
| 2m-3e-2-C5 | 779.10 | 0.7471 | 4m-c-2-C6 | 656.00 | 0.5021 | t-5-C10 | 983.30 | 0.2379 | 1,t-3-C9 | 909.70 | 0.1464 |
| 2m-t-3-C7 | 741.20 | 0.4802 | 4m-c-2-C5 | 556.50 | 0.5021 | t-5-C12 | 1178.61 | 0.2379 | 1,t-3-C5 | 516.58 | 0.1464 |
| 2m-t-3-C6 | 646.80 | 0.4802 | 4m-c-3-C7 | 773.90 | 0.4968 | t-5-C16 | 1578.80 | 0.2379 | 1,t-3-C8 | 811.18 | 0.1464 |
| 2m-c-3-C7 | 736.00 | 0.4871 | 4e-1-C6 | 757.40 | 0.3647 | t-5-C13 | 1278.10 | 0.2379 | 1,t-4-C7 | 677.20 | 0.2335 |
| 2m-c-3-C6 | 644.40 | 0.4871 | 4e-t-2-C6 | 742.70 | 0.5307 | t-5-C14 | 1376.90 | 0.2379 | 1,t-4-C10 | 965.50 | 0.2335 |
| 2e-1-C4 | 592.30 | 0.3390 | 4e_c-2-C6 | 748.80 | 0.5196 | t-5-C15 | 1479.30 | 0.2379 | 1,t-4-C6 | 582.52 | 0.2335 |
| 2e-1-C6 | 779.30 | 0.3390 | 55m2-1-C6 | 706.30 | 0.6520 | t-5-C11 | 1079.80 | 0.2379 | 1,t-4-C9 | 869.90 | 0.2335 |
| 2e-1-C5 | 682.60 | 0.3390 | 55m2-t-2-C6 | 707.50 | 0.7941 | t-6-C12 | 1177.44 | 0.2431 | 1,t-4-C8 | 769.00 | 0.2335 |
| 2ip-1-C5 | 750.70 | 0.5723 | 55m2-c-2-C6 | 724.40 | 0.7632 | t-6-C16 | 1575.30 | 0.2431 | 1,t-5-C7 | 681.90 | 0.2417 |
| 334m3-1-C5 | 724.20 | 0.8575 | 5m-1-C7 | 755.50 | 0.3661 | t-6-C13 | 1275.30 | 0.2431 | 1,t-5-C10 | 964.00 | 0.2417 |
| 33m2-1-C4 | 507.80 | 0.6526 | 5m-1-C6 | 650.80 | 0.3661 | t-6-C14 | 1373.90 | 0.2431 | 1,t-5-C9 | 864.50 | 0.2417 |
| 33m2-1-C6 | 714.20 | 0.6526 | 5m-1-C12 | 1237.90 | 0.3661 | t-6-C15 | 1476.70 | 0.2431 | 1,t-5-C8 | 768.30 | 0.2417 |
| 33m2-1-C5 | 627.90 | 0.6526 | 5m-t-2-C7 | 767.60 | 0.4862 | t-7-C16 | 1573.60 | 0.2455 | 1,t-6-C10 | 962.30 | 0.2426 |
| 33m2-2e-1-C4 | 731.40 | 0.8453 | 5m-t-2-C6 | 659.80 | 0.4862 | t-7-C14 | 1372.50 | 0.2455 | 1,t-6-C9 | 869.10 | 0.2426 |
| 344m3-1-C5 | 700.70 | 0.8791 | 5m-t-3-C7 | 755.90 | 0.4799 | t-7-C15 | 1475.30 | 0.2455 | 1,t-6-C8 | 779.30 | 0.2426 |
| 344m3-t-2-C5 | 747.10 | 1.0000 | 5m-c-2-C7 | 776.80 | 0.4664 | t-8-C16 | 1571.70 | 0.2481 | 1,t-7-C10 | 967.80 | 0.2424 |
| 344m3-c-2-C5 | 748.00 | 0.9984 | 5m-c-2-C6 | 672.30 | 0.4664 | c-2-C4 | 417.20 | 0.2316 | 1,t-7-C9 | 879.10 | 0.2424 |
| 34m2-1-C6 | 756.00 | 0.5709 | 5m-c-3-C7 | 760.10 | 0.4723 | c-2-C7 | 703.50 | 0.2316 | 1,t-8-C10 | 977.10 | 0.2482 |
| 34m2-1-C5 | 638.60 | 0.5709 | 6m-1-C7 | 748.20 | 0.3714 | c-2-C10 | 1000.20 | 0.2316 | 1,c-3-C7 | 716.50 | 0.1406 |
| 34m2-t-2-C6 | 761.40 | 0.7374 | 6m-1-C12 | 1233.90 | 0.3714 | c-2-C6 | 603.90 | 0.2316 | 1,c-3-C10 | 1007.20 | 0.1406 |
| 34m2-t-2-C5 | 678.70 | 0.7374 | 6m-t-2-C7 | 768.50 | 0.4688 | c-2-C9 | 900.40 | 0.2316 | 1,c-3-C6 | 622.50 | 0.1406 |
| 34m2-t-3-C6 | 778.90 | 0.7640 | 6m-t-3-C7 | 748.60 | 0.4749 | c-2-C12 | 1199.88 | 0.2316 | 1,c-3-C9 | 909.50 | 0.1406 |
| 34m2-c-2-C6 | 757.20 | 0.7479 | 6m-c-2-C7 | 771.90 | 0.4625 | c-2-C16 | 1603.20 | 0.2316 | 1,c-3-C5 | 525.99 | 0.1406 |
| 34m2-c-2-C5 | 671.40 | 0.7479 | 6m-c-3-C7 | 750.00 | 0.4723 | c-2-C13 | 1300.70 | 0.2316 | 1,c-3-C8 | 810.50 | 0.1406 |
| 34m2-c-3-C6 | 784.40 | 0.7540 | 7m-1-C12 | 1232.50 | 0.3908 | c-2-C14 | 1400.90 | 0.2316 | 1,c-4-C7 | 678.60 | 0.2309 |
| 35m2-1-C6 | 699.50 | 0.6207 | 8m-1-C12 | 1235.90 | 0.3845 | c-2-C15 | 1503.60 | 0.2316 | 1,c-4-C10 | 964.50 | 0.2309 |
| 35m2-t-2-C6 | 751.70 | 0.7371 | 9m-1-C12 | 1241.40 | 0.3737 | c-2-C11 | 1099.90 | 0.2316 | 1,c-4-C6 | 588.34 | 0.2309 |
| 35m2-c-2-C6 | 752.30 | 0.7360 | 1-C3 | 287.30 | 0.1173 | c-2-C5 | 505.00 | 0.2316 | 1,c-4-C9 | 868.10 | 0.2309 |
| 3m-1-C4 | 450.40 | 0.3686 | 1-C4 | 384.80 | 0.1173 | c-2-C8 | 801.60 | 0.2316 | 1,c-4-C8 | 771.60 | 0.2309 |
| 3m-1-C7 | 740.40 | 0.3686 | 1-C7 | 682.30 | 0.1173 | c-3-C7 | 691.00 | 0.2334 | 1,c-5-C7 | 688.10 | 0.2391 |
| 3m-1-C6 | 645.80 | 0.3686 | 1-C10 | 981.60 | 0.1173 | c-3-C10 | 984.40 | 0.2334 | 1,c-5-C10 | 961.40 | 0.2391 |
| 3m-1-C12 | 1239.30 | 0.3686 | 1-C6 | 582.80 | 0.1173 | c-3-C6 | 592.90 | 0.2334 | 1,c-5-C9 | 865.60 | 0.2391 |
| 3m-1-C5 | 552.50 | 0.3686 | 1-C9 | 881.90 | 0.1173 | c-3-C9 | 885.90 | 0.2334 | 1,c-5-C8 | 769.20 | 0.2391 |
| 3m-2e-1-C4 | 659.90 | 0.5835 | 1-C12 | 1181.60 | 0.1173 | c-3-C12 | 1182.72 | 0.2334 | 1,c-6-C10 | 962.50 | 0.2406 |
| 3m-2e-1-C5 | 750.90 | 0.5835 | 1-C16 | 1584.40 | 0.1173 | c-3-C16 | 1585.50 | 0.2334 | 1,c-6-C9 | 868.00 | 0.2406 |
| 3m-2ip-1-C4 | 712.40 | 0.8276 | 1-C13 | 1282.60 | 0.1173 | c-3-C13 | 1282.60 | 0.2334 | 1,c-6-C8 | 783.50 | 0.2406 |

**Table 1** (Continued)

| compound[a] | I | $C_D$ | compound[a] | I | $C_D$ | compound[a] | I | $C_D$ | compound[a] | I | $C_D$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1,c-7-C10 | 967.10 | 0.2389 | c-2,c-4-C8 | 835.90 | 0.2432 | t-3,c-6-C9 | 881.40 | 0.3428 | c-2,c-5-C7 | 716.40 | 0.3394 |
| 1,c-7-C9 | 883.60 | 0.2389 | c-2,c-5-C7 | 712.30 | 0.346 | t-3,c-7-C10 | 970.10 | 0.3599 | c-2,c-5-C10 | 988.10 | 0.3394 |
| 1,c-8-C10 | 981.10 | 0.2409 | c-2,c-5-C10 | 985.30 | 0.346 | t-4,t-6-C10 | 1030.20 | 0.2207 | c-2,c-5-C9 | 891.60 | 0.3394 |
| t-2,t-4-C7 | 746.50 | 0.24 | c-2,c-5-C9 | 887.90 | 0.346 | t-4,c-6-C10 | 1019.40 | 0.2404 | c-2,c-5-C8 | 796.50 | 0.3394 |
| t-2,t-4-C10 | 1040.50 | 0.24 | c-2,c-5-C8 | 792.70 | 0.346 | c-2,t-4-C7 | 751.10 | 0.2331 | c-2,c-6-C10 | 983.30 | 0.3545 |
| t-2,t-4-C6 | 645.92 | 0.24 | c-2,c-6-C10 | 978.50 | 0.3636 | c-2,t-4-C10 | 1041.70 | 0.2331 | c-2,c-6-C9 | 887.90 | 0.3545 |
| t-2,t-4-C9 | 940.50 | 0.24 | c-2,c-6-C9 | 883.60 | 0.3636 | c-2,t-4-C6 | 654.09 | 0.2331 | c-2,c-6-C8 | 804.00 | 0.3545 |
| t-2,t-4-C8 | 837.90 | 0.24 | c-2,c-6-C8 | 798.00 | 0.3636 | c-2,t-4-C9 | 942.60 | 0.2331 | c-2,c-7-C10 | 985.30 | 0.3575 |
| t-2,t-5-C7 | 708.30 | 0.3439 | c-2,c-7-C10 | 980.10 | 0.3656 | c-2,t-4-C8 | 840.60 | 0.2331 | c-2,c-7-C9 | 900.60 | 0.3575 |
| t-2,t-5-C10 | 990.00 | 0.3439 | c-2,c-7-C9 | 897.00 | 0.3656 | c-2,t-5-C10 | 992.90 | 0.3383 | c-2,c-8-C10 | 1001.40 | 0.3545 |
| t-2,t-5-C9 | 889.20 | 0.3439 | c-2,c-8-C10 | 998.70 | 0.3595 | c-2,t-5-C9 | 892.40 | 0.3383 | c-3,t-5-C10 | 1026.60 | 0.2354 |
| t-2,t-5-C8 | 795.20 | 0.3439 | t-3,t-5-C10 | 1037.40 | 0.218 | c-2,t-5-C8 | 797.80 | 0.3383 | c-3,t-5-C9 | 925.00 | 0.2354 |
| t-2,t-6-C10 | 976.40 | 0.3682 | t-3,t-5-C9 | 934.50 | 0.218 | c-2,t-6-C10 | 982.30 | 0.358 | c-3,t-6-C10 | 974.00 | 0.3461 |
| t-2,t-6-C9 | 881.60 | 0.3682 | t-3,t-5-C8 | 838.10 | 0.218 | c-2,t-6-C9 | 887.20 | 0.358 | c-3,c-5-C10 | 1031.40 | 0.2243 |
| t-2,t-6-C8 | 794.59 | 0.3682 | t-3,t-6-C10 | 977.90 | 0.3369 | c-2,t-6-C8 | 799.90 | 0.358 | c-3,c-5-C9 | 932.30 | 0.2243 |
| t-2,t-7-C10 | 984.60 | 0.3633 | t-3,t-6-C9 | 884.00 | 0.3369 | c-2,t-7-C10 | 986.80 | 0.3551 | c-3,c-5-C8 | 835.90 | 0.2243 |
| t-2,t-7-C9 | 895.00 | 0.3633 | t-3,t-7-C10 | 972.30 | 0.3559 | c-2,c-4-C7 | 754.60 | 0.225 | c-3,c-6-C10 | 976.40 | 0.3401 |
| t-2,t-8-C10 | 995.40 | 0.3655 | t-3,c-5-C10 | 1022.80 | 0.2372 | c-2,c-4-C10 | 1043.80 | 0.225 | c-3,c-6-C9 | 882.00 | 0.3401 |
| c-2,c-4-C7 | 745.40 | 0.2432 | t-3,c-5-C9 | 925.00 | 0.2372 | c-2,c-4-C6 | 661.95 | 0.225 | c-3,c-7-C10 | 970.90 | 0.3585 |
| c-2,c-4-C10 | 1032.00 | 0.2432 | t-3,c-5-C8 | 830.60 | 0.2372 | c-2,c-4-C9 | 945.80 | 0.225 | c-4,c-6-C10 | 1028.30 | 0.2242 |
| c-2,c-4-C9 | 934.50 | 0.2432 | t-3,c-6-C10 | 974.00 | 0.3428 | c-2,c-4-C8 | 846.30 | 0.225 | | | |

[a] The digits following C show the number of the carbons in the straight chain; m, e, p, and ip show methyl, ethyl, propyl, and isopropyl, respectively; digits in front of these characters denote the position of the substituents, and the ones behind them denote the number of these substituents; t and c denote trans and cis; I denotes retention index; $C_d$ shows the class distance variable.

of the big deviation between measure temperature and calibration temperature (60 °C). Retention indices for 383 alkenes including n-monoenes, n-dienes, and branched monoenes were obtained. They are listed in Table 1.

**Topological Descriptors of the Alkenes.** Investigation of chromatographic retention is one of the most active areas for QSPR studies using molecular connectivity indices.[14,48] Molecular connectivity indices are defined as the summation of subgraph terms over all the graph edges:

$$^0\chi = \sum_{edges} (\delta_i)^{-0.5}$$

$$^1\chi = \sum_{edges} (\delta_i\delta_j)^{-0.5}$$

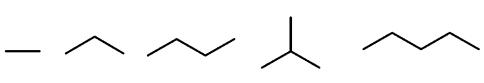
$$^2\chi = \sum_{edges} (\delta_i\delta_j\delta_k)^{-0.5}$$

$$^3\chi_p = \sum_{edges} (\delta_i\delta_j\delta_k\delta_l)^{-0.5}$$

$$^3\chi_c = \sum_{edges} (\delta_i\delta_j\delta_k\delta_l)^{-0.5}$$

$$^4\chi_p = \sum_{edges} (\delta_i\delta_j\delta_k\delta_l\delta_o)^{-0.5}$$

where $\delta_i$,$\delta_j$,$\delta_k$, $\delta_l$, and $\delta_o$ are the valences of the atoms of *i*, *j*, *k*, *l*, and *o*. The zero-order subgraph term is based on each individual graph vertex. The other 5 subgraph terms are based on one until four edges illustrated as follows,



respectively. The summation is then to be taken again over all subgraph terms. For the molecules of alkene, each valence of two adjacent atoms of a double bond increases 1 comparing to that of single bond. However, when considering the subgraph terms over all graph edges, a double bond is regarded as a single bond, that is to say the multiple edges are taken only once.

Other 9 topological indices and 5 quantum chemical parameters were also used in this paper. They are the kappa series, say $^1\kappa$, $^2\kappa$, $^3\kappa$;[49] molecular topological index (MTI); the principal eigenvalue of the distance matrix (MED); the determinant of the adjacency-plus-distance matrix (DET) and geometrical isomerism index S index (MTI′) proposed by Schultz et al.;[50−53] and the indices Yx[54] and EAID[55] proposed by Xu and co-workers, respectively; and quantum chemical descriptors, like heat of formation, electronic energy, dipole moment, ionization potential, and LUMO energy.

All the routines for calculating the topological indices were programmed in our laboratory using MATLAB language of version 5.3. The quantum chemical descriptors were calculated using the MOPAC method in Chem3D software. The values of all descriptors mentioned above were provided in a supplementary table.

## RESULTS AND DISCUSSION

**Limitation of Common Regression Model.** Figure 1 shows the fitting results of the regression model including 20 molecular descriptors. Their corresponding statistic parameters are as follows, that is, correlation coefficient $R = 0.9982$, standard error $S = 17.48$ i.u., and F-test value $F = 5148.83$, respectively. The statistic parameters of the regression seem to be acceptable. However, the model residuals are too large with the maximum one reaching up to 50 i.u.! What is wrong with the regression model? It seems to us that there must be several classes in the studied 383 alkene molecules which could not be described well by the used 20 molecular descriptors. Moreover, to our best knowledge, we could not even find out better results by simply using common regression for modeling alkenes merely using molecular descriptors so far. As shown by previous works on retention behavior of alkenes, we know

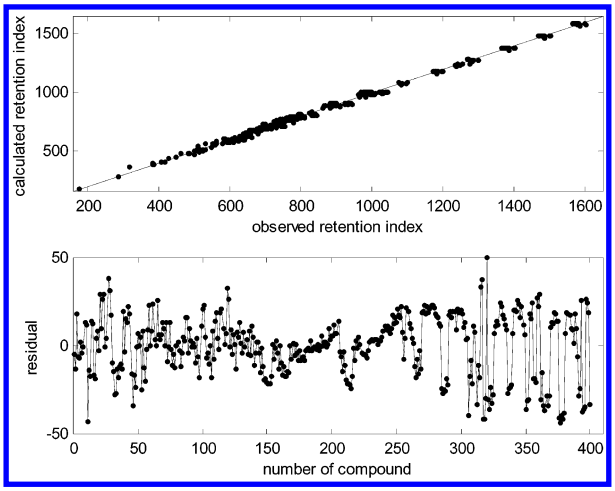



**Figure 1.** The fitting and residual plots of regression model for whole alkenes including 20 variables. Top part: The fitting plot of the regression model. Lower part: The residual plots of the regression model.

that there may be many factors influencing retention behavior of the alkenes, such as number of carbon atoms, number and position of the double bonds, alkyl branches and their position connected to the backbone, and cis- and trans-configuration. However, some of the factors mentioned above were not well taken into account in present molecule descriptors. Thus, there must be several subclasses hidden in alkenes. One may at least split alkenes into some classes such as monoenes or dienes; straight chain alkenes or branched alkenes; 1 branch alkenes, 2 branches alkenes, and etc. Moreover, it was also found that retention anomalies are expected for compounds with molecular structures that make it possible to form a ring conformation of the propyl group. Thus, simply using one model to describe several kinds of samples certainly results in poor result. To seek an accurate quantitative retention-structure relationship, the first thing, in our opinion, may be to find out reasonably the subclasses in alkenes and the relationship among them.

**Projection Pursuit and Classification of Alkenes.** Molecular connectivity indices developed by Randić, Kier, and Hall have been proven to be a very efficient series of topological descriptors in QSRR researches.[14,48] We first tried to use only ${}^0\chi$, ${}^1\chi$, ${}^2\chi$, ${}^3\chi_p$ as descriptors in this work to do the projection pursuit, since we want to include as few as possible variables at the first step. The projection pursuit research was conducted followed the calculation procedure as stated in the theory and methodology section. In this procedure, the key step is how to construct reasonably $\mathbf{X}^{(\alpha)}$ and $\mathbf{y}^{(\alpha)}$ to obtain the best projection direction vector $\mathbf{a}$. We select 9 samples to construct $\mathbf{X}^{(\alpha)}$ and corresponding $\mathbf{y}^{(\alpha)}$, because one needs samples of at least 2 times of the number of variables to get a stable statistical estimate. To avoid doing too heavy a calculation, we restricted the correlation coefficient between $\mathbf{X}^{(\alpha)}$ and $\mathbf{y}^{(\alpha)}$ to be greater than 0.999 and calculated $10^8$ candidate sets of 9 samples (taking about 16 h) selected randomly in this work.

We found four typical projection patterns with different projection index values from all projections obtained in the $10^8$ calculations. They are shown in Figures 2−5. The projection pattern with the maximum value of projection index $\xi$ (in Figure 2) shows only a scatter diagram and has no clear structure, while others (Figures 3−5) are quite

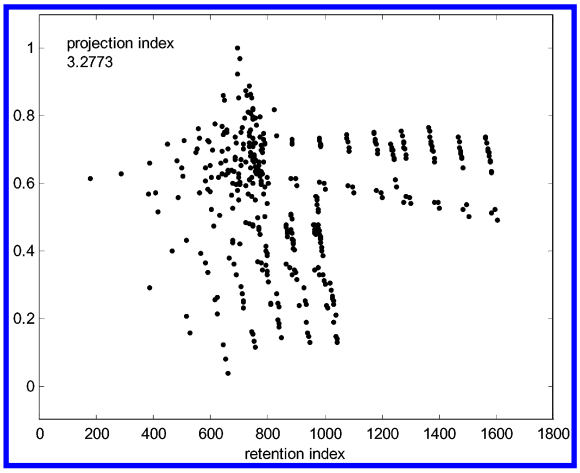



**Figure 2.** The projection plot of pattern 1.

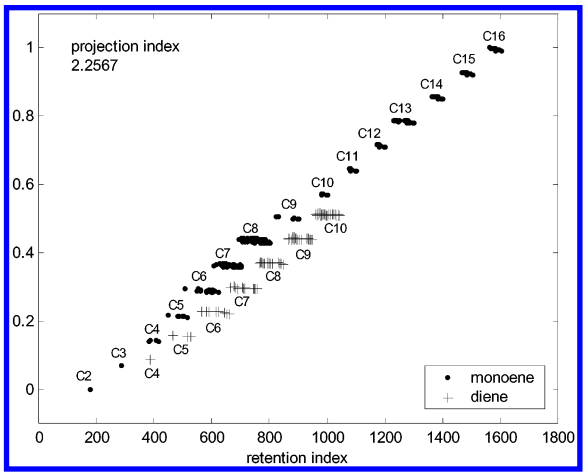



**Figure 3.** The projection plot of pattern 2. The number behind the capital C denotes the number of carbon atoms of the molecule.

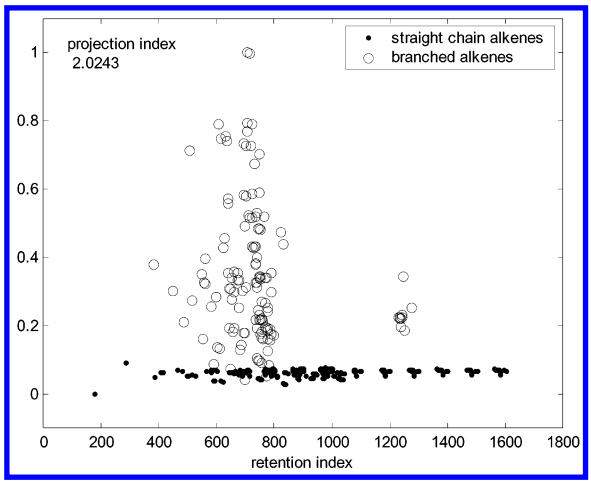



**Figure 4.** The projection plot of pattern 3.

informative with some interesting structures. A close look of the latter 3 figures reveals that there is plentiful chemical structure information in them.

In Figure 3, all alkenes are classified into some classes based on the numbers of carbon atoms in their molecules, in which monoenes and dienes are marked by "•" and "+", respectively. The digits following the character "C" in the plot show the numbers of carbon atoms. The values of projection $\alpha_j$ increase regularly with the increasing number of carbon atoms and are very close to each other for

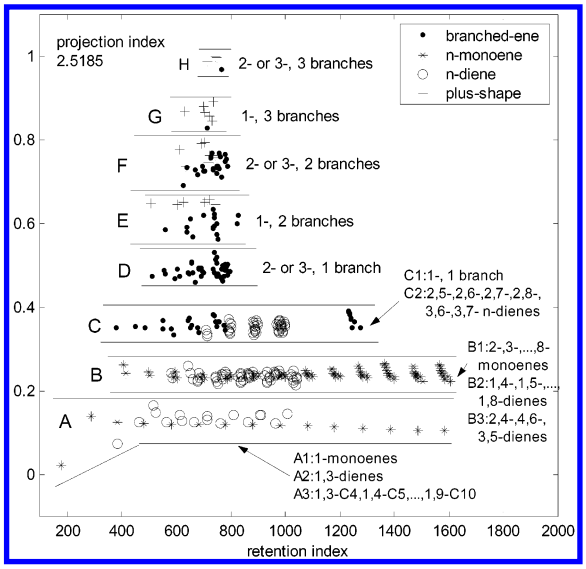
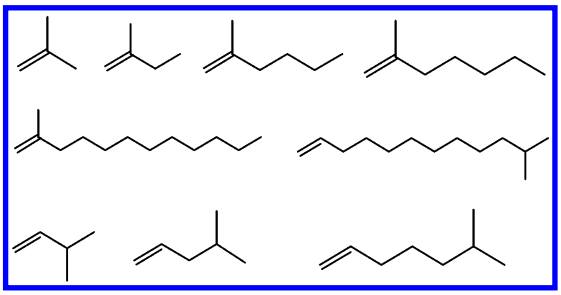



**Figure 5.** The projection plot of pattern 4.

compounds with same number of carbon atoms. It is interesting to see that the 9 samples selected to construct $\mathbf{X}^{(\alpha)}$ and $\mathbf{y}^{(\alpha)}$ are all molecules containing 8 carbon atoms. They are 2,3-dimethyl-*trans*-3-hexene, 2,4-dimethyl-*trans*-3-hexene, 2,5-dimethyl-*cis*-3-hexene, 3,3-dimethyl-1-hexene, 3,4-dimethyl-1-hexene, 3,4-dimethyl-*cis*-3-hexene, 4,4-dimethyl-1-hexene, 4,4-dimethyl-*cis*-2-hexene, and 4-methyl-3-ethyl-*trans*-2-pentene, respectively. It seems to hint that the structure feature of the group of molecules selected to construct $\mathbf{X}^{(\alpha)}$ and $\mathbf{y}^{(\alpha)}$ decides the structure feature of the projection pattern.

In Figure 4, the projection pattern that has the minimum projection index value shows two classes, in which one embraces all the straight chain alkenes including n-monoenes and n-dienes (marked by "•") and the other is all branched monoenes (marked by "o"). From the plot, one can see that the straight chain alkenes including monoenes and dienes concentrate on a line, while the branched monoenes scatter very much without forming a close class. However, the classification of this pattern is not so informative. Samples included for obtaining this projection direction are as follows: 4-ethyl-*cis*-2-hexene, 4-ethyl-1-hexene, *trans*-4-octene, *cis*-4-dodecene, *cis*-5-tetradecene, *cis*-8-hexadecene, 1,*trans*-4-hexadiene, *trans*-2,*trans*-4-octadiene, and *trans*-2,*cis*-8-decadiene, respectively.

The most interesting projection pattern is the fourth one shown in Figure 5. The 9 compounds used to obtain the projection direction **a** are as follows:



They can be regarded as a group of analogue compounds, since they all have one methyl substitute at the first position in the backbone and all have a first position double band. To include more compounds of the same structure feature into the class to obtain better projection direction **a**, two other compounds, 2-methyl-1-pentene and 5-methyl-1-hexene, were added into this set as a new sample set for projecting. The projective result (see Figure 5) obtained by using the new set is almost the same as the one obtained by using only 9 compounds shown above.

**Relationship between Classes and Molecular Structure.** In Figure 5, one can see that all the alkenes can be roughly classed in 8 groups, marked by A, B, C, D, E, F, G, and H, respectively. All branched alkenes, n-monoenes, and n-dienes are classified into several classes based on their structure features. In this plot, the branched monoenes are marked by "•" and "+" in order to distinguish two different topological shapes of branching. For the alkenes of two or three branches may be considered as two kinds of molecular topological shapes, say plus-shape and star-shape as shown in Figure 6.

Furthermore, to distinguish n-monoenes and n-dienes, we use the mark "*" to represent the former and mark "o" to indicate the last.

After careful checking of the structure features of the molecules in these 8 classes, we can find some interesting relationships between classes and molecular structure, which will be discussed in the following paragraphs.

First, let us have a close look at class A. In this class, we can see that there are three subclasses, that is, A1: linear 1-monoenes, A2: 1,c-3-dienes, and 1,t-3-dienes and A3 whose molecules are with two double bonds on the two edges of the molecule (1,3-$C_4$, 1,4-$C_5$, 1,5-$C_6$, 1,6-$C_7$, 1,7-$C_8$, 1,8-$C_9$, and 1,9-$C_{10}$). Notice that A1 representing linear 1-monoenes is a kind of monoene having its double band at the first position in the molecules, and A2 representing linear conjugated dienes having also their conjugated double band at edged position in the molecules. A3 represents nonconjugated dienes. The structure feature is that their double bands are all at edged positions. As found by Katritzky et al., the position of the substitute group in alkanes is a very important factor to influence the retention behaviors of alkanes.[16] The same conclusion can be reached for the alkenes studied in this work. It is very interesting to compare the structure features between class A and class B. In class B, there are also 3 subclasses, say B1 including linear 2-, 3-,..., 8-monoenes, B2 including linear 1,4-, 1,5-,...,1,8-dienes, and B3 including conjugated linear dienes say 2,4-, 4,6-, 3,5-dienes. All these indicate that when the double bands, no matter whether they are single double bands or conjugated double bands, move to the midsection of the carbon backbone their corresponding project indexes move from class A to B. This fact reveals clearly that the retention behavior of alkenes is strongly dependent on the position of the double band in the backbone. This conclusion is quite similar to the one drawn by Katritzky and his colleagues for methyl-branched hydrocarbons[16] for the substitute position of methyl group.

Class C contains two subclasses, in which one is a group containing n-dienes (marked C2 in the plot) and the other is a subclass containing 1-monoenes with 1 branch (marked C1 in the plot). If one has a close look at the subclass C2, one can find that it embraces linear 2,5-, 2,6-, 2,7-, 2,8-, 3,6-, 3,7-dienes. Comparing with the corresponding B2 (including linear 1,4-, 1,5-,...,1,8-dienes) in class B, the only difference between these two subclasses lies in that the edged double band move from the first position to the second or

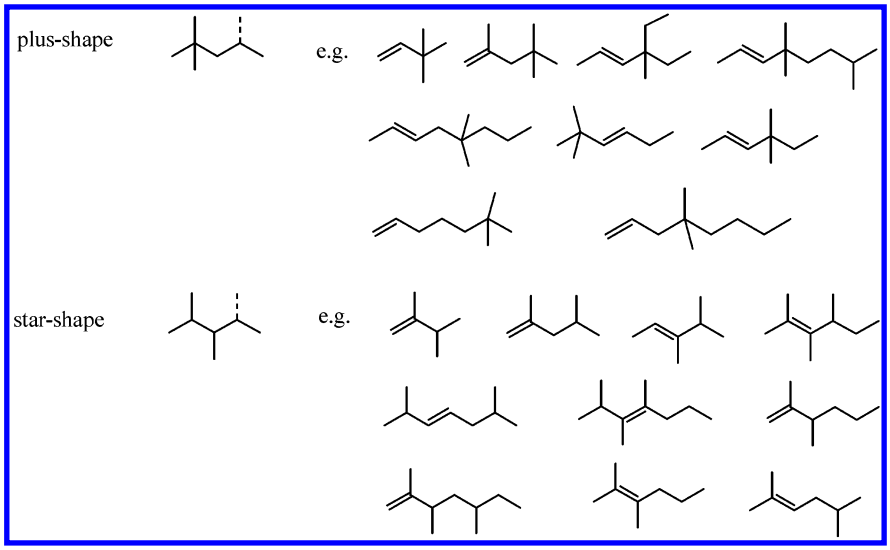



**Figure 6.** Plus-shape and star-shape molecules.

**Table 2.** Quantitative Relationship between Classification Shown in Figure 5 and the Chemical Structure

| no. | class | subclass | compound | structure feature[a] |
|---|---|---|---|---|
| 1 | A | A1 | 1-monoenes | 1 double, 1 edge |
| | | A2 | 1,3-dienes | 2 double, 1 edge, 1 conjugate |
| | | A3 | 1,3-C4, 1,4-C5, ..., 1,9-C10 | 2 double, 2 edge |
| 2 | B | B1 | 2-,3-,...,8-monoenes | 1 double |
| | | B2 | 1,4-, 1,5-,..., 1,8-dienes | 2 double, 1 edge |
| | | B3 | 2,4-, 4,6-, 3,5-dienes | 2 double, 1 conjugate |
| 3 | C | C1 | 1-, 1 branch | 1 double, 1 branch, 1 edge |
| | | C2 | 2,5-, 2,6-, 2,7-, 2,8-, 3,6-, 3,7-dienes | 2 double |
| 4 | D | D | 2- or 3-, 1 branch | 1 double, 1 branch |
| 5 | E | E | 1-, 2 branches | 1 double, 2 branch, 1 edge |
| 6 | F | F | 2- or 3-, 2 branches | 1 double, 2 branch |
| 7 | G | G | 1-, 3 branches | 1 double, 3 branch, 1 edge |
| 8 | H | H | 2- or 3-, 3 branches | 1 double, 3 branch |

[a] Double, edge, and conjugate mean double bonds, double bonds in edge, and conjugated double bonds, respectively. The Arabic numerals in front of these items denote the number of these structural items.

the third position in the backbone. It is why their class position moves from B to C. As for C1, we know it belongs to the branched monoenes. The structure feature of branched alkenes is the topic in the following paragraph.

In our previous work on alkanes, we have found that the branching topological shape of the molecules is the main factor influencing retention behavior of alkanes.[18] Katritzky and his colleagues (see Figure 2 in their paper[16]) also reached the same conclusion. Here for alkenes, we also find that the branching topological shape is really a very important factor influencing the retention behavior of the branched alkenes. From the projection plot (Figure 5), one can easily see that the branched alkenes (strictly speaking, all being branched monoenes involved in this work) are classed by mono- (subclasses C and D), di- (subclasses E and F), and trisubstituted monoenes (subclasses G and H). If the other two factors influencing the retention behavior, say position of the double band (see above paragraphs in results and discussion section) and topological branched shapes (plus shape and star shape shown in Figure 6) are also taken into account, how and why these branched monoenes can be classed into these six subclasses can be explained perfectly.

It can be summarized from the above discussions that the position of the double band (edged and nonedged), the conjugated double bands, the branching degree, and their corresponding topological shapes all are very important factors influencing the retention behavior of the alkenes. Thus, the classification shown in Figure 5 does describe the retention behavior of every kind of alkene with different chemical structure features. With all these in the mind, the following can be concluded from the classification pattern shown in Figure 5:

(1) When the number of branches increases by one, the position for classification in Figure 5 moves up two, say from C to E or D to F, for instance.

(2) When the number of double bonds increases by one, the class position increases by one correspondingly.

(3) When the edged double bond moves to the midsection of the backbone of the molecule, say from edged to nonedged, the class position will move up one, say from A to B or B to C, for instance.

(4) When the conjugated double bands become nonconjugated double bands, the class position will move up one, say from A to B, for instance.

After carefully checking all the compounds involved in this work, the above 4 rules will work without one exception. They are summarized in Table 2.

**Behavior of Homologues and Class Distance Variable of Alkenes.** As we all know, there is a good linear relationship between the retention value and the number of carbon atoms for a homologous series molecules. In our early work,[18] every homologous series of alkanes shows a very

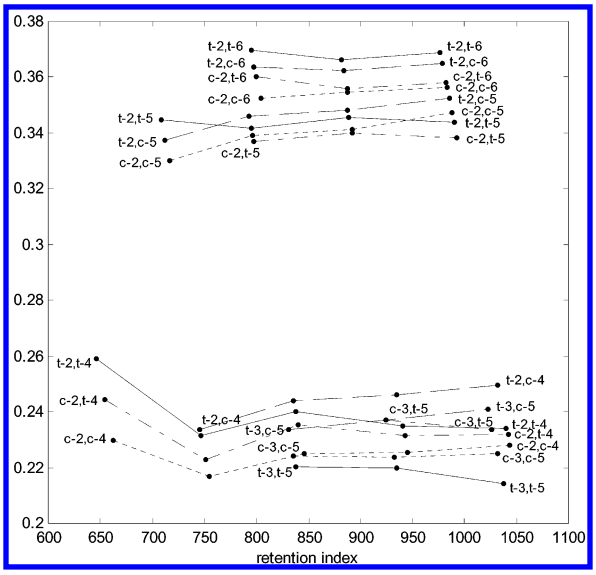



**Figure 7.** The projection plot of 16 homologous series of alkenes embraced in class B3 and class C2.

closed value of projection $\alpha_j$ for all compounds. What is the behavior of alkenes? From the enlarged view of Figure 5, we found that almost all samples for every homologue investigated in this work have closed projection value $\alpha_j$. Figure 7 shows projection results for 16 homologous series in subclasses B3 and C2 in Figure 5. The results obtained hint that every homologue can be described by a number, which essentially expresses a kind of class distance or difference between it and the class that is used to calculate the projection direction **a**. Thus, the distance of every homologous series is denoted by the mean of all values of $\alpha_j$ for this series, and then a new variable $C_d$ called class distance variable of alkene is proposed which represents the distance for the homologous series. The calculation procedure of $C_d$ is described as follows:

(1) Calculate projection values $\alpha_j$ of compounds in every homologous series using eq 2.

(2) All compounds in an individual homologous series have the same value of $C_d$, which equals the mean of $\alpha_j$ of all molecules in this series.

For example, the homologous series of t-2, t-5-diene contains four compounds, i.e., *trans*-2,*trans*-5-heptadiene; *trans*-2,*trans*-5-octadiene; *trans*-2,*trans*-5-nonadiene, and *trans*-2,*trans*-5-decadiene, respectively. Their values of $\alpha_j$ are 0.3448, 0.3415, 0.3456, and 0.3437, respectively. The average of them equals 0.3439, which is just the value of the class distance variable $C_d$ for these compounds. Values of class distance variable for all alkenes are listed in Table 1.

The regression model including 20 molecular descriptors mentioned above did not provide satisfactory fitting (see Figure 1). The regression model including only 4 molecular connectivity indices ${}^0\chi$, ${}^1\chi$, ${}^2\chi$, ${}^3\chi_p$ gives even worse results with $R = 0.99453$, $S = 28.94$ i.u., and the maximum residual

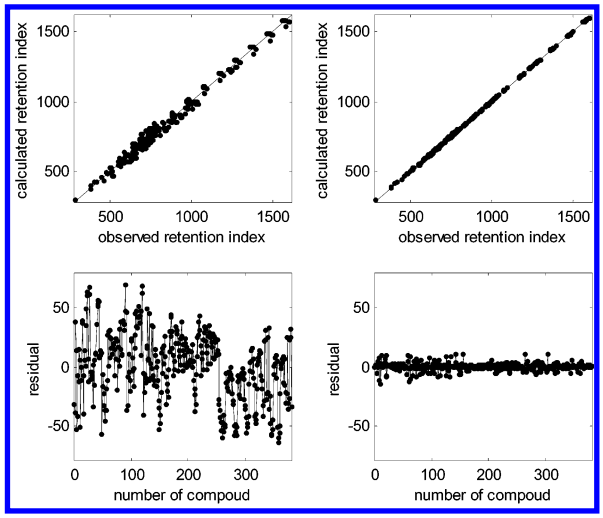



**Figure 8.** The fitting and residual plots of regression model for alkenes including 4 topological indices without and with class distance variable proposed. Left top part: The fitting plot of the regression model including only 4 molecule connectivity variables. The correlation coefficient of the model is 0.99453. Right top part: The fitting plot of the regression model including 4 variables together with the class distance variable. The correlation coefficient of the model is 0.99993. Left low part: The residual plots of regression model including only 4 variables. Right low part: The residual plots of regression model including 4 variables together with class distance variable.

up to 70.2 i.u.! The class distance variable $C_d$ denotes the distance among all classes in samples. It essentially describes the deficient part of the above 4 descriptors. Thus, a new regression model including both the 4 molecular connectivity indices and the class distance variable of alkenes was established, and both models without and with the variable $C_d$ were shown as follows:

$$I = -76.1003 + 146.3080\,{}^0\chi + 86.5863\,{}^1\chi - 114.2411\,{}^2\chi - 48.0294\,{}^3\chi_p$$

$$I = -340.5916 + 480.0324\,{}^0\chi - 427.6214\,{}^1\chi - 88.5717\,{}^2\chi + 20.3675\,{}^3\chi_p - 484.7760\,C_d$$

To our surprise, the regression model with the variable $C_d$ is very good. The fitting and residual plots are shown in Figure 8. The statistics parameters are listed in Table 3. From this figure, one can easily see that all the residuals are small and quite close to measurement errors. Furthermore, we also check its prediction ability by cross-validation technique. The results are also shown in Table 3. The results are quite satisfactory. The correlation coefficient, say $R$, reaches 0.9999. The corresponding $F$ test value and standard error are 512946 and 3.35 i.u., respectively. The model also shows a good prediction ability with almost the same prediction standard error ($S_{cv} = 3.39$ i.u.) as the estimate standard error ($S = 3.35$ i.u.).

**Prediction of Retention Indices Measured on the Column of Ucon LB 550x.** The new class distance variable

**Table 3.** Statistical Parameters of Models with and without Class Distance Variable

| | fitting (squalane) | | | | | prediction (Ucon LB 550x) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| descriptors | n | R | S (i.u.) | F | $S_{cv}$ (i.u.) | n | R | S (i.u.) | F | $S_{cv}$ (i.u.) |
| ${}^0\chi,{}^1\chi,{}^2\chi,{}^3\chi_p$ | 383 | 0.9945 | 28.94 | 8524 | 29.13 | 152 | 0.9809 | 18.62 | 936 | 18.91 |
| ${}^0\chi,{}^1\chi,{}^2\chi,{}^3\chi_p$, $C_d$ | 383 | 0.9999 | 3.35 | 512946 | 3.39 | 152 | 0.9982 | 5.78 | 8054 | 5.93 |

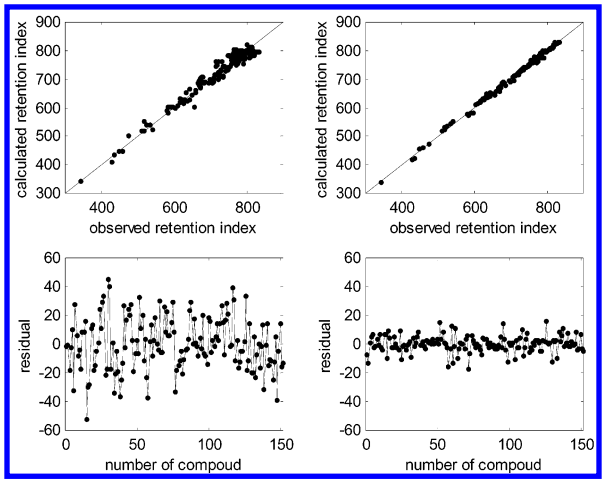



**Figure 9.** The fitting and residual plots of regression model for 152 alkenes measured on Ucon LB 550x at column temperature 40 °C including 4 variables without and with class distance variable. Left top part: The fitting plot of the regression model including only 4 molecule connectivity variables. The correlation coefficient of the model is 0.9809. Right top part: The fitting plot of the regression model including 4 variables together with the class distance variable. The correlation coefficient of the model is 0.9982. Left low part: The residual plots of regression model including only 4 variables. Right low part: The residual plots of regression model including 4 variables together with class distance variable.

of alkenes is just obtained from the structure feature of molecules; it should be independent of concrete experiment conditions, such as the chromatographic columns, column temperature, and so on. To further confirm this assumption, we collected another data set[33] measured on the stationary phase of Ucon LB 550x and at the column temperature of 40 °C, in which there are 152 monoenes. Together with the 4 molecular connectivity indices, say ${}^0\chi,{}^1\chi,{}^2\chi,{}^3\chi_p$, and the corresponding class distance variable (equal $C_d$ value of the same compound in the Table 1), the regression model is established (see Figure 9). The fitting and residual results with only 4 molecular connectivity indices are also shown in Figure 9. From this figure, one can easily see that the improvement from the class distance variable is quite significant. The regression parameters are listed in Table 3 too. Their corresponding correlation coefficient *R*, *F* test values, standard errors (*S*), and prediction standard errors ($S_{cv}$) are 0.9982, 8054, 5.78, and 5.93 (with class distance variable) and 0.9809, 936, 18.692, and 18.91 (without class distance variable), respectively. The fact shows that the class distance variable of alkenes proposed in this work performs fairly well.

## ACKNOWLEDGMENT

The project is financially supported by National Nature Foundation Committee of P. R. China (No. 20175036).

**Supporting Information Available:** Table of descriptors and retention indices of alkenes. This material is available free of charge via the Internet at http://pubs.acs.org.

## REFERENCES AND NOTES

(1) Sojak, L.; Ostrovsky I.; Kubinec, R.; Kraus, G.; Kraus, A. High-resolution gas chromatography with liquid crystal glass capillaries xi. separation of isomeric $C_8$ and $C_9$ hydrocarbons. *J. Chromatogr.* **1990**, *509*, 93−99.
(2) Sojak, L.; Kral'Ovicova, E.; Ostrovsky, I.; Leclercq, P. A. Retention behaviour of cinjugated and isolated n-alka-dienes. identification of n-nona- and n-decadienes by capillary gas chromatography using structure-retention correlations and mass spectrometry. *J. Chromatogr.* **1984**, *292*, 241−261.
(3) Sojak, L.; Ostrovsky, I.; Leclercq, P. A.; Rijks. J. A. Identification of n-hepta-and n-octadienes by high-resolution gas chromatography using structure-retention corelations and mass spectrometry. *J. Chromatogr.* **1980**, *191*, 187−198.
(4) Sojak, L.; Hrivnak, J.; Majer. P. Capillary Gas chromatography of linear alkenes on squalane. *Anal. Chem.* **1973**, *45*, 293−302.
(5) Sojak, L.; Ostrovsky, I.; Janak, J. Propyl Effect and retention-structure correlation as a means of gas chromatographic identification. *J. Chromatogr.* **1987**, *406*, 43−49.
(6) Sojak, L.; Ostrovsky, I.; Kubinec, R.; Kraus, G.; Kraus. A. Separation and identification of all isomeric n-nonadecenes by capillary gas chromatography on a mesogenic stationary phase with Fourier transform infrared and mass spectrometric detection. *J. Chromatogr.* **1991**, *609*, 283−288.
(7) Sojak, L.; Krupcik, J.; Janak, J. Gas chromatography of all $C_{15}$−$C_{18}$ linear alkenes on capillary columns with very high-resolution power. *J. Chromatogr.* **1980**, *195*, 43−64.
(8) Sojak, L.; Ostrovsky, I.; Farkas, P.; Janak, J. High-resolution gas chromatography with liquid crystal glass capillaries. ix. separation of isomeric $C_9$−$C_{11}$ n-alkenes and n-alkanes. *J. Chromatogr.* **1986**, *356*, 105−114.
(9) Sojak, L.; Ostrovsky, I.; Kubinec, R.; Kraus, G.; Kraus, A. High-resolution gas chromatography with liquid crystal glass capillaries. xii. separation of isomeric $C_{17}$−$C_{18}$ n-alkenes. *J. Chromatogr.* **1990**, *520*, 75−83.
(10) Sojak, L.; Hrivnak, J.; Ostrovsky, I.; Janak, J. Capillary gas chromatography of linear alkenes on squalane separation and identification of n-pentadecenes and n-hexadecenes. *J. Chromatogr.* **1974**, *91*, 613−622.
(11) Sojak, L.; Xrupcik, J.; Tesarik, L.; Janak, J. Correlation of the boiling points of nonbranched $C_6$ and $C_{10}$ Olefins with the gas chromatographic retention indices. *J. Chromatogr.* **1972**, *65*, 93−102.
(12) Sojak, L.; Majer, P.; Skalak, P.; Janak, J. Identification of straight-chain undecenes by capillary gas chromatography on squalane. *J. Chromatogr.* **1972**, *65*, 137−142.
(13) Boneva, S.; Dimov, N. Gas chromatographic retention indices for alkenes on ov-101 and squalane capillary columns. *Chromatographia* **1986**, *21*, 149−150.
(14) Rohrbaugh, R. H.; Jurs, P. C. Prediction of gas chromatographic retention indexes of selected olefins. *Anal. Chem.* **1985**, *57*, 2770−3.
(15) Kaliszan, R. *Quantitative structure chromatographic retention relationships*; Wiley: New York, 1987.
(16) Katritzky, A. R.; Chen, K.; Maran, U.; Carlson, D. A. QSPR correlation and prediction of GC retention indexes for methyl-branched hydrocarbons produced by insects. *Anal. Chem.* **2000**, *72*, 101−109.
(17) Buja, A.; Duffy, D.; Hastie, T.; Tibshirani, R. Discussion of "Multivariate adaptive regression splines". *Ann. Statist.* **1991**, *19*, 93−98.
(18) Du, Y. P.; Liang, Y. Z. Data mining for seeking accurate quantitative relationship between molecular structure and GC retention indices of alkenes by projection pursuit. *Comput. Chem.* Submitted for publication.
(19) Liang,Y. Z.; Gan, F. Chemical knowledge discovery from mass spectral database. I. Isotope distribution and Beynon table. *Anal. Chim. Acta* **2001**, *446*, 107−114.
(20) Cundari, T. R.; Russo, M. Database mining using soft computing techniques. An integrated neural network-fuzzy logic-genetic algorithm approach. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 281−287.
(21) Debska, B. J.; Guzowska-Swider, B. Knowledge discovery in an Infrared Database. *Computers Chem.* **1997**, *21*, 51−59.
(22) Fayyad, M. U.; Uthurusamy R. Data mining and knowledge discovery in databases (introduction to the special section). *Commun. ACM* **1996**, *39*(11), 24−26.
(23) Glymour, C.; Madigan, D.; Pregibon, D.; Smyth, P. Statistical inference and data mining. *Commun. ACM* **1996**, *39*(11), 35−41.
(24) Inmon, W. H. The data warehouse and data mining. *Commun. ACM* **1996**, *39*(11), 49−50.
(25) Fayyad, M, U.; Haussler, D.; Stolorz, P. Mining scientific data. *Commun. ACM* **1996**, *39*(11), 51−57.
(26) Huber, P. J. Projection pursuit. *Ann. Statist.* **1985**, *13*, 435−475.
(27) Friedman, J. H.; Stuetzle, W. Projection pursuit regression. *J. Am. Statist. Assoc.* **1981**, *76*, 817−823.
(28) Friedman, J. H. Stanford University, unpublished manuscript.
(29) Zulaica, J.; Guiochon, G. Analysis of high polymers and their pyrolysis products by gas chromatography. I. method of study. *Bull Soc. Chim. (Fr.)* **1966**, 1343−1351.
(30) Tourres, D. A. Structural analysis of industrial butene dimers by gas chromatography. *J. Gas Chromatogr.* **1967**, *5*, 35−40.

(31) Hively, R. A.; Hinton, R. E. Variation of the retention index with temperature on squalane substrates. *J. Gas Chromatogr*. **1968**, *6*, 203−17.
(32) Loewenguth, J. C.; Tourres, D. A. Etude des variations des indices de retention en fonction de la temperature. *Z. Anal. Chem*. **1968**, *236*, 170.
(33) Matukuma, A. Retention indeces of alkanes through $C_{10}$ and alkenes through $C_8$ and relation between boiling points and retention data. *Gas Chromatography 1968*; Inst. of Petroleum: London, 1969; p 55.
(34) Schomburg, G.; Henneberg, D. Analysis of olefin mixtures by combination of capillary gas chromatograph and mass spectrometer. *Gas Chromatography 1968*; Inst. of Petroleum: London, 1969; p 45.
(35) Sojak, L.; Bucinska. Open tubular column gas chromatography of dehydrogenation products of $C_6-C_{10}$ n-alkanes. Separation and identification of mixtures of $C_6-C_{10}$ straight-chain alkanes, alkenes, and aromatics. *J. Chromatogr*. **1970**, *51*, 75−82.
(36) Eisen, O.; Orav, A.; Rang, S. Identification of normal alkenes, cyclopentenes, and cyclohexenes by capillary gas chromatography. *Chromatographia* **1972**, *5*, 229−39.
(37) Rijks, J. A.; Cramers, C. A. High precision capillary gas chromatography of hydrocarbons. *Chromatographia* **1974**, *7*, 99.
(38) Vaneertum, R. On the retention index calculation according to Takacs. *J. Chromatogr. Sci.* **1975**, *13*, 150.
(39) Chretiem, J. R.; Dubois, J. E. Topological analysis of gas−liquid chromatographic behavior of alkenes. *Anal. Chem.* **1977**, *49*, 747−56.
(40) Pacakova, V.; Kozlik, V. Capillary reaction gas chromatography. 1. Catalytic decomposition of hydrocarbons. *Chromatographia* **1978**, *11*, 266.
(41) Welsch, T.; Engewald, W. Molecular structure and retention behaviour. IX. Retention behaviour of isomeric octynes and octadiynes. *Chromatographia* **1978**, *11*, 5.
(42) Dubois, J. E.; Chretien, J. R.; Sojak, L.; Rijks, J. A. Topological analysis of the behavior of linear alkenes up to tetradecenes in gas−liquid chromatography on squalane. *J. Chromatogr*. **1980**, *194*, 121−34.
(43) Sojak, L.; Krupcik, J.; Janak, J. Comparison of capillary columns coated with $C_{87}$ hydrocarbon and squalane in the analysis of *n*-pentadecene isomers. *J. Chromatogr*. **1980**, *191*, 199−206.
(44) Sojak, L.; Kraus, G.; Farkas, P.; et al. High-performance gas chromatography in liquid-crystal glass capillaries. V. Separation of isomeric n-tridecenes and n-tetradecenes. *J. Chromatogr*. **1982**, *238*, 51−7.
(45) Schroder, H. *HRC&CC* **1980**, 3, 38, 95, 119, 362.
(46) Papazova, D.; Milina, R.; Dimov, N. Comparative evaluation of retention of hydrocarbons present in the C5- petroleum fraction on methylsilicone and squalane phases. *Chromatographia* **1988**, *25*, 177−180.
(47) Du, Y. P.; Liang, Y. Z.; Wu, C. J. Database construction of GC retention index and correction of mistakes in it. Chinese 8th computers and applied chemistry conference; 2001; Huangshan, pp 147−149.
(48) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.
(49) Kier, L. B.; Hall, L. H. The kappa indices for modeling molecular shape and flexibility. *Topological Indices and Related Descriptors in QSAR and QSPR;* Devillers, J., Balaban, A. T., Eds.; Gordon and Breach Science Publishers: The Netherlands, 1999; pp 455−489.
(50) Schultz, H. P. Topological organic chemistry. 1. Graph theory and topological indices of alkanes. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 227−228.
(51) Schultz, H. P.; Schultz, E. B.; Schultz, T. P. Topological organic chemistry. 2. Graph theory, matrix determinants and eigenvalues, and topological indices of alkanes. *J. Chem. Inf. Comput. Sci*. **1990**, *30*, 27−29.
(52) Schultz, H. P.; Schultz, E. B.; Schultz, T. P. Topological organic chemistry. 7.1 Graph theory and molecular topological indices of unsaturated and aromatic hydrocarbons. *J. Chem. Inf. Comput. Sci*. **1990**, *33*, 863−867.
(53) Schultz, H. P.; Schultz, E. B.; Schultz, T. P. Topological organic chemistry. 9. Graph theory and molecular topological indices of stereoisomeric organic compounds. *J. Chem. Inf. Comput. Sci*. **1995**, *35*, 864−870.
(54) Yao, Y. Y.; Xu, L.; Yuan, X. S. A new topological index for research on structure−property relationship of alkanes. *Chinese ACTA Chimica Sinica* **1993**, *51*, 463−469.
(55) Hu, C. Y.; Xu, L. On Highly Discriminating Molecular Topological Index. *J. Chem. Inf. Comput. Sci*. **1996**, *36*, 82−90.

CI020285U