Orthogonalization of Block Variables by Subspace-Projection for Quantitative Structure Property Relationship (QSPR) Research

Yiping Du,[†,§] Yizeng Liang,*[,‡] Boyan Li,[‡] and Chengjian Xu[‡]

Institute of Chemometrics and Chemical Sensing Technology, Hunan University, Changsha 410082, P. R. China, and 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 February 7, 2002

A subspace-projection method is developed to construct orthogonal block variable, which is originally from some kinds of series of topological indices or quantum chemical parameters. With the help of canonical correlation analysis, the orthogonal block variables were used to establish the structure-retention index correlation model. The regression of only few new orthogonal variables obtained by canonical correlation analysis against retention index shows significant improvement both in fitting and prediction ability of the correlation model. Moreover, the quantitative intercorrelation between the different block variables of topological indices can also be evaluated with the help of the subspace-projection technique proposed in this work.

INTRODUCTION

To evaluate quantitatively the degree of similarity or dissimilarity of chemical structures or to find correlations between structure and activities or properties (QSAR or QSPR) one needs to translate structures into numbers. Beginning with Wiener,[1] numerical molecular descriptors, named topological indices (TIs) by Hosoya,[2] have gained gradual acceptance along with other descriptors used in QSAR and QSPR studies. However, the multiplication of TIs, as pointed out by Balaban,[3] caused worry in some parts of the scientific community, due also to the fact that the physical meaning of these descriptors was not clear, and it was also shown that many TIs were intercorrelated; details can be found in several references.[4−7]

There are many descriptors, such as topological indices and quantum chemical parameters, in QSAR/QSPR study. For instance, software CODESSA,[8] designed by Kartritzky et al., may calculate 400 molecular descriptors. How to select proper descriptors from these many candidates becomes the first important task encountered in QSAR/QSPR studies. It is well-known that if too few variables are considered, it will cause underfitting of the model and result in low correlation between structure and activities or properties (QSAR or QSPR). However, if too many variables are included in the regression model it will cause overfitting of the model and result in an unstable model with bad prediction. The task is important and difficult. Selection of descriptors (variables), in our opinion, is to investigate the possible combinations of variables for finding a best model. The exhausting search for all possible combinations of variables is the most reliable method for variable selection. It is unfortunate that the method is not applicable when the number of variables is more than 20, since one needs to calculate 2^n-1 (here n is the number of variables) times for exhausting searches. Thus, several strategies were suggested for variable selection, such as stepwise approaches,[9] leaps-and-bounds regression,[10] genetic algorithm,[11−13] and others.[8,14−18] However, if one has too many variables at hand, it is still a very difficult task for the above methods to deal with.

It is worth noting that many descriptors are similar or even almost the same. They may represent “duplicated” information of the molecular structure. Balaban et al.[5] have investigated the relationship among some commonly used topological indices. They found that some topological indices are quite similar. By accounting for correlation among them, they divided these topological indices into two groups. The Wiener index[1] W, Hosoya index[2] Z, and Randic index[19] $^1\chi$ were considered as one group, while the Balaban’s center indices C and C′ [20] were considered as another group. Topological indices in one group represent high related, i.e., some of them are redundant because they have most of the duplicated information with their similar descriptors. Randic[21] first introduced orthogonalized molecular descriptors in QSAR studies. He illustrated the approach of orthogonalization considering Hosoya’s Z index[2] as a property and connectivity indices,[22] $^1\chi$,$^2\chi$, $^3\chi_p$,$^4\chi_p$ as molecular descriptors, and evaluated the role of the descriptors in the regression, similarities, and differences among molecular descriptors.[21] Later, Trinajstic[23−25] and Xu[26] used orthogonalized molecular descriptors to study descriptor-property correlation and variable selection in QSAR/QSPR. The regression models using orthogonal variables have some interesting features, such as possessing the same correlation coefficient *R*, the standard error *S*, and the *F*-test value as the regression model

* Corresponding author phone: 86-731-8825637; fax: 86-731-8825637; e-mail: yzliang@public.cs.hn.cn. Corresponding 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/ci020283+ CCC: $22.00
Published on Web 08/01/2002

using nonorthogonal variables and the stability of the regression coefficients.[21]

In fact, a series of topological indices (not individual index) with similar calculation strategy was often encountered, such as the molecular connectivity chi indices[22] (${}^0\chi$, ${}^1\chi$, ${}^2\chi$, ${}^3\chi_p$, ${}^3\chi_c$, ${}^4\chi_p$, ...), Kappa indices[27] (${}^0\kappa$, ${}^1\kappa$, ${}^2\kappa$, ${}^3\kappa$, ...). A series of descriptors generally was defined by accounting for more molecular structure information and less redundancy. Thus, a series of descriptors might be considered as an ensemble named block descriptor (variable), which includes all individual descriptors in this series. Being similar to the orthogonalization of individual descriptor mentioned above, orthogonal block descriptors (variables) would also be obtained easily. The advantage of using block descriptor is that one may work with only a few block variables instead of many individual variables.

Canonical correlation analysis (CCA) is an extension of multiple regression.[28] CCA is to establish the maximum correlation among sets of variables. With the help of CCA, one may find out a so-called canonical correlation variable for the block variable, which possesses the maximum correlation with the property (see Theory and Methodology section). This canonical correlation variable captures almost full information of the original block variable. Therefore, the canonical correlation variable that is only one variable (a new variable) can be used to substitute the original block variable that includes more than one individual variables. In this case, the model established between descriptors and property will be simplified with fewer variables but without losing correlation information.

In this work, eight block variables of topological and quantum chemical descriptors, including up to 32 individual descriptors, are selected to investigate the orthogonal variables for every block variable. First, a new subspace-projection procedure is developed to construct orthogonal block variable. The new canonical variables, which could represent the corresponding whole block variables, were extracted with the help of canonical correlation analysis. Then, the structure-retention index correlation model including these new canonical variables is also established. The regression model obtained in this way shows significant improvement both in fitting and prediction ability. Moreover, the intercorrelation between the different block variables of topological indices is also evaluated with the help of the subspace-projection technique proposed in this work.

THEORY AND METHODOLOGY

Orthorgonalization of Block Variables. It seems to be necessary to outline first the procedure of orthogonalization for individual variables developed by Randic[21] before we go into the details of orthogonalization for block variables. Suppose we have n variables (descriptors) $\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_n$, each one of them being a vector containing m elements corresponding to m samples, and a property vector $\mathbf{y}$ with m elements. We will orthogonalize these variables to obtain their orthogonal variables denoted by ω_1, ω_2 ,..., ω_n. A commonly used procedure of orthogonalization[21,26] is described briefly as follows.

First, select a variable, such as, $\mathbf{x}_1$ as the first orthogonal variable ω_1. Then, the second orthogonal variable ω_2 will be calculated by making the second variable, such as, $\mathbf{x}_2$ orthogonalize to $\mathbf{x}_1$, which is the residuals of the regression of $\mathbf{x}_2$ against $\mathbf{x}_1$, i.e., values of calculated $\mathbf{x}_2$ from the regression subtracted from original $\mathbf{x}_2$. In the same way, the orthogonal parts of other variables $\mathbf{x}_3$,..., $\mathbf{x}_n$ to $\mathbf{x}_1$ can be calculated, which are denoted by ω_{31} ,..., $\omega_{n1,}$ respectively. Third step is calculating ω_3, which is made by orthogonalizing ω_{31} to ω_2, i.e., constructing the regression of ω_{31} against ω_2 and considering the residuals of ω_{31} as the third orthogonal variable ω_3. This process continues step by step until obtaining all orthogonal variables ω_1,ω_2 ,..., ω_n. Because of the orthogonality of these new variables, the correlation coefficient R can be easily calculated by the following formula

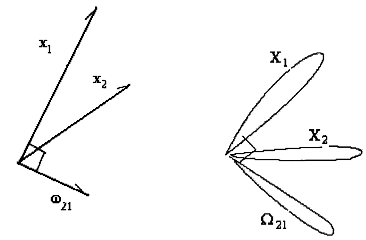

Figure 1. Illustration of the geometric sense of orthogonalization between the individual variables and block variables.

$$R^2 = \sum_{p}^{i=1} R_i^2 \tag{1}$$

where R_i is the correlation coefficient for the regression of the property $\mathbf{y}$ against the ith orthogonal variable ω_i, and R is the correlation coefficient against p orthogonal variables ω_1,ω_2 ,..., ω_p.

Similar to the above procedure, the calculation of orthogonal block variable corresponding to a series of descriptors can be processed. Note that a block variable corresponds to a submatrix of variable matrix $\mathbf{X}$. Suppose that we have n submatrixes, say $\mathbf{X}_1$, ..., $\mathbf{X}_n$. When we want to orthogonalize $\mathbf{X}_i$ to $\mathbf{X}_j$, where $\mathbf{X}_i$ and $\mathbf{X}_j$ are $m\times n_i$ and $m\times n_j$ matrixes, respectively, whose rows correspond variables, the orthogonalized submatrix $\mathbf{X}_{i|j}$ denoted by Ω_i can be simply calculated by the following equation

$$\Omega_i = \mathbf{X}_{i|j} = (\mathbf{I} - \mathbf{X}_j(\mathbf{X}_j^t\,\mathbf{X}_j)^{-1}\mathbf{X}_j^t)\,\mathbf{X}_i \tag{2}$$

Here $\mathbf{X}_{i|j}$ is the orthogonal supplementary part to $\mathbf{X}_j$ but from $\mathbf{X}_i$, meaning that every vector in $\mathbf{X}_{i|j}$ or their linear combinations, say $\mathbf{x}_a = \mathbf{X}_{i|j}\mathbf{a}$, is orthogonal to every vector in $\mathbf{X}_j$ or their linear combinations, say $\mathbf{x}_b = \mathbf{X}_j\mathbf{b}$, that is, $\mathbf{x}_a^t\mathbf{x}_b = 0$. The geometric sense of this orthogonal projection is illustrated in Figure 1.

For n nonorthogonal submatrixes of $\mathbf{X}_1$, $\mathbf{X}_2$, ..., $\mathbf{X}_n$, the procedure to obtain their corresponding orthogonal matrixes Ω_1, Ω_2 ,..., Ω_n is similar to that for individual variable discussed above. The procedure starts by selecting a nonorthogonal matrix, such as $\mathbf{X}_1$ as the first orthogonal matrix Ω_1. Ω_2 can be calculated through eq 2, where $\Omega_2 = \mathbf{X}_{2|1}$ with $\mathbf{X}_j = \mathbf{X}_1$, and $\mathbf{X}_i = \mathbf{X}_2$. When constructing new matrixes of $\mathbf{X}_j$ and $\mathbf{X}_i$ in the way $\mathbf{X}_j = [\Omega_1\ \Omega_2]$, $\mathbf{X}_i = \mathbf{X}_3$, Ω_3 (equals to $\mathbf{X}_{i|j}$) will be calculated by eq 2. Constructing $\mathbf{X}_j$ with Ω_1 and Ω_2 together can ensure the calculated value of $\mathbf{X}_{i|j}$ being orthogonal not only to Ω_1 but also to Ω_2. Other orthogonal matrixes can also be computed by eq 2 through the same procedure. Matrixes such obtained, say Ω_1, Ω_2, ..., Ω_n, are orthogonal with each other. Therefore, the correlation coefficient R for using p orthogonal matrixes can also be obtained from the correlation coefficient R_i for using individual orthogonal matrixes by eq 1. Orthogonalized block variables can be derived from original block variables by this subspace-projection procedure.

Canonical Correlation Analysis (CCA). Notice that either original block variable or orthogonalized block variable is just a matrix that contains more than one variable. Even if we consider a series of descriptors (a block descriptor) as an ensemble, the block still contains several individual descriptors. The dimensionality of regression model is still the same as before, if one simply uses them to establish the regression model. Is it possible for us to reduce the dimension of the model with the help of the obtained orthogonalized block variables? As we know, principal component regression (PCR) is an efficient technique to reduce the data dimensionality. Original variables might be substituted by a few principal components that explain most of the variance of the data matrix **X**. The components corresponding to minimum eigenvalues are removed which are considered as experimental errors. But in QSAR/QSPR study, structural descriptor data, such as topological indices and quantum chemical calculation parameters, are obtained by calculation based on certain procedures. They have no any experimental errors! Every component of PCR for the data without experimental errors should contain some information from the descriptor data matrix and may be helpful to improve the regression against the property. It is dangerous to remove any component even if its corresponding eigenvalue is small! (See Results and Discussion section for more details.) Thus, we use canonical correlation analysis (CCA), instead of PCR, to reduce the dimension for every block variable.

The general mathematical problem of CCA is to establish the maximum correlation among sets of variables. For two matrixes $m \times p$ **X** and $m \times q$ **Y**, CCA is to find two linear combinations for **X** and **Y**, i.e., $\mathbf{Xa}_1$ and $\mathbf{Yb}_1$, respectively. CCA will give the maximum correlation between $\mathbf{Xa}_1$ and $\mathbf{Yb}_1$. $\mathbf{Xa}_1$ and $\mathbf{Yb}_1$ are called canonical correlation variables. This can be indicated by

$$R(\mathbf{Xa}_1, \mathbf{Yb}_1) = \max R(\mathbf{Xa}, \mathbf{Yb}) \tag{3}$$

Subject to the constrains

$$\mathrm{v}(\mathbf{Xa}) = 1 \text{ and } \mathrm{v}(\mathbf{Yb}) = 1 \tag{4}$$

where $R(\cdot)$ denotes correlation coefficient, and $\mathrm{v}(\cdot)$ denotes variance. $\mathbf{Xa}_1$ and $\mathbf{Yb}_1$ is the first pair of canonical correlation variable. Similar calculations can be conducted to find the second pair, the third pair, and so forth. The standard statistical method[28] to solve this problem may result from the calculation of eigenvalues and eigenvectors of the matrix **K**

$$\mathbf{K} = (\mathbf{V}_{\mathbf{XX}}^{-1/2})\mathbf{V}_{\mathbf{XY}}(\mathbf{V}_{\mathbf{YY}}^{-1/2}) \tag{5}$$

where **V** denotes the covariance matrix

$$\mathbf{V}_{\mathbf{XX}} = \mathrm{E}(\mathbf{X} - \bar{\mathbf{X}})(\mathbf{X} - \bar{\mathbf{X}})^{\mathrm{t}} \tag{6}$$

$$\mathbf{V}_{\mathbf{XY}} = \mathrm{E}(\mathbf{X} - \bar{\mathbf{X}})(\mathbf{Y} - \bar{\mathbf{Y}})^{\mathrm{t}} \tag{7}$$

$$\mathbf{V}_{\mathbf{YY}} = \mathrm{E}(\mathbf{Y} - \bar{\mathbf{Y}})(\mathbf{Y} - \bar{\mathbf{Y}})^{\mathrm{t}} \tag{8}$$

where

$$\bar{\mathbf{X}} = \begin{bmatrix} \bar{x}_1 & \bar{x}_2 & \cdots & \cdots & \bar{x}_p \\ \bar{x}_1 & \bar{x}_2 & \cdots & \cdots & \bar{x}_p \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \bar{x}_1 & \bar{x}_2 & \cdots & \cdots & \bar{x}_p \end{bmatrix}, \bar{\mathbf{Y}} = \begin{bmatrix} \bar{y}_1 & \bar{y}_2 & \cdots & \cdots & \bar{y}_q \\ \bar{y}_1 & \bar{y}_2 & \cdots & \cdots & \bar{y}_q \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \bar{y}_1 & \bar{y}_2 & \cdots & \cdots & \bar{y}_q \end{bmatrix}$$

with $\bar{x}_i = 1/m \sum_{j=1}^{m} x_{ji}$, ($i$ = 1, 2, ..., p), and $\bar{y}_i = 1/m \sum_{j=1}^{m} y_{ji}$, ($i$ = 1, 2, ..., q), and t indicates transpose of the matrix. One could obtain the scores $\mathbf{u}_{\mathrm{i}}$ and loadings $\mathbf{v}_{\mathrm{i}}$ with the help of the singular value decomposition of the matrix **K**, such that

$$\mathbf{K} = [\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_{\mathrm{r}}]\ \mathbf{S}\ [\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_{\mathrm{r}}]^{\mathrm{t}} \tag{9}$$

Here **S** is the diagonal matrix of singular values of the matrix **K** which is degressive, then the canonical correlation variables will be calculated by the formulas

$$\mathbf{b_i} = \mathbf{V}_{\mathbf{XX}}^{-1/2}\mathbf{u}_{\mathrm{i}} \tag{10}$$

$$\mathbf{a_i} = \mathbf{V}_{\mathbf{YY}}^{-1/2}\mathbf{v}_{\mathrm{i}}\ (i{=}1, 2, ...\mathrm{r}) \tag{11}$$

$\mathbf{Xa}_{\mathrm{i}}$ and $\mathbf{Yb}_{\mathrm{i}}$ are the ith pair of canonical correlation variables. In the case of **Y** being a vector **y**, such as the property of retention index encountered in this work, r equals to 1, i.e., there will exist only one pair of canonical correlation variable $\mathbf{Xa}_1$ and $\mathbf{Yb}_1$ because the matrix **K** has rank one. Note that $\mathbf{a}_1$ is a vector of the same number of elements as the number of variables in **X**, while $\mathbf{b}_1$ is a scalar. The regression of $\mathbf{Xa}_1$ against **Y** will give almost the same correlation coefficient as the regression of **X** against **y**. Thus, the canonical correlation variable, which is a vector instead of a matrix, may be used to replace the original variable set (a matrix) without losing information of the original variable set. In this way, the canonical correlation analysis might provide a new way to reduce the dimensionality of QSAR/QSPR model. The study of descriptor-property correlation will be therefore simplified by introducing only a few canonical correlation variables.

Outline of the Procedure of Calculation. (1) Split all given descriptors into a few subsets based on a certain criterion, such as considering a series of descriptors proposed by same authors as a subset, which was adopted in this paper, to get block variables $\mathbf{X}_1$, $\mathbf{X}_2$, ..., $\mathbf{X}_n$.

(2) The block variables $\mathbf{X}_1$, $\mathbf{X}_2$, ..., $\mathbf{X}_n$ were centered by subtracting their means.

(3) Orthogonalize block variables by eq 2. The order of variables strongly impacts on the orthogonalization result.[26] Here we use the "based on R_i"[26] approach to orthogonalize variables. First pick up a block variable in the set of $\mathbf{X}_1$, $\mathbf{X}_2$, ..., $\mathbf{X}_n$ with maximum correlation coefficient R against the property **y** as the first orthogonal block variable Ω_1. Then, calculate their orthogonal block variables to Ω_1 by eq 2 and select the orthogonal block variables with maximum R in the left ones as the second orthogonal block variables Ω_2 for the remaining block variables. The third orthogonal block variable Ω_3 is such orthogonal one to Ω_1 and Ω_2 that have

Table 1. Topological Indices and Experimental Retention Times of Anthocyanins (Data Set 1, from Ref 23)

compd no.	RT (min)	p_1	p_2	p_3	$^1\chi$	$^2\chi$	$^3\chi$	H	H′	H″
1	8.62	36	54	62	11.6201	9.9599	7.6915	223.927	242.345	653.763
2	11.29	35	52	59	11.5286	9.8109	7.6362	212.376	229.643	618.050
3	13.48	37	55	64	12.0059	10.1667	7.9717	232.493	246.244	656.946
4	14.47	34	50	56	11.4370	9.6717	7.5559	201.179	217.065	582.375
5	16.21	36	53	61	11.9143	10.0177	7.9147	220.664	233.465	621.205
6	18.53	38	56	66	12.3916	10.3735	8.2520	241.291	250.197	660.134
7	17.15	42	62	69	13.3744	11.2305	8.3520	275.474	295.748	798.776
8	20.73	41	60	66	13.2828	11.0816	8.2967	263.338	282.988	763.050
9	22.74	43	63	71	13.8292	11.3689	8.6830	284.587	299.697	801.957
10	24.14	40	58	63	13.1913	10.9424	8.2164	251.554	270.353	727.364
11	25.97	42	61	68	13.6686	11.2884	8.5752	272.171	286.860	766.210
12	27.08	44	64	73	14.1459	11.6442	8.9125	293.931	303.700	805.150

maximum R in the remaining ones. Other orthogonal block variables have the same calculation procedure.

(4) For variables **X** and **Y**, their canonical correlation variables, say $\mathbf{Xa}_i$ and $\mathbf{Yb}_i$, can be calculated by using eqs 5−11. In this work, **Y** is actually the property vector **y**, so $\mathbf{b}_1$ is a scalar, and there is only one pair of canonical correlation variable, say $\mathbf{Xa}_1$ and $\mathbf{Yb}_1$. Here we calculate canonical correlation variables for orthogonal block variables and get new variables ω_1, ω_2, ..., ω_n, corresponding to the orthogonal block variables.

(5) Establish regression models of the property **y** against every orthogonal canonical correlation variable derived from block variables. Select a few variables with maximum correlation coefficient R_i to establish the descriptor-property correlation model.

DATA SETS

Data Set 1. This data set is collected from Amic et al.[23] who calculated the HPLC retention times (RT) of 12 anthocyanidin malonylglucosides by using several structure−property models based on three different types of orthogonalized topological indices, say path numbers (path counts) p_1, p_2, and p_3, molecular connectivity indices $^1\chi$, $^2\chi$, and $^3\chi$, and Harary indices H, H′, and H″. They found the best agreement between the experimental and calculated values with path numbers. Here we consider the three types of topological indices as three block variable (three matrixes), say **P**, χ, and **H**, and then study the relationship between their orthogonal variable matrixes and retention time. The retention time and original topological indices of the 12 compounds are listed in Table 1.

Data Set 2. This data set contains a GC retention index, 26 topological indices, and 6 quantum chemical descriptors of 149 alkane molecules. The retention index data are collected from a GC retention index database established in our laboratary.[29] Topological indices used are molecular connectivity series indices[22] $^1\chi$,$^2\chi$,$^2\chi_p$, $^2\chi_c$; kappa series indices,[27] say $^0\kappa$, $^1\kappa$, $^2\kappa$, $^3\kappa$; path counts series indices[30] p_1, p_2, p_3, p_4; walk counts series indices[30] w_1, w_2, w_3, w_4; path/walk counts series indices[30] pw_1, pw_2, pw_3, pw_4; the indices proposed and used by Schultz,[31,32] say molecular topological index (MTI), the principal eigenvalue of the distance matrix (PED), the principal eigenvalue of the adjacency-plus-distance matrix (PEAD), the logarithm of determinant of the adjacency-plus-distance matrix (DET); the indices Yx[33] and EAID[34] proposed by Xu and co-workers. The quantum chemical descriptors are heat of formation, electronic energy, core−core repulsion energy, dipole moment, ionization potential, and LUMO energy. Eight capital bold characters $\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{X}_3$, $\mathbf{X}_4$, $\mathbf{X}_5$, $\mathbf{X}_6$, $\mathbf{X}_7$, and $\mathbf{X}_8$ are used to denote these eight series of descriptors (block variables), respectively. The quantum chemical descriptors were calculated by using the MOPAC method in Chem3D software. The routines for calculating the topological indices were programmed in our laboratory using MATLAB language of version 5.3. This data set is given in Table 2.

RESULTS AND DISCUSSION

Improvement of Correlation by using CCA for Data Set 1. There are three types of topological indices in data set 1. Every one of them has three individual topological indices. Table 3 gives the regression results of the model between retention times and orthogonal variables of every type of topological indices. They reveal some features of orthogonalization procedure. The standard error S and the correlation coefficient R do not change with the orthogonalization, and the regression coefficients are constant (see the columns 4−7 in Table 3). The model with three path numbers p_1, p_2, and p_3 is superior to others, which was considered as being quite satisfactory for experimental chemists[23] with $R = 0.9989$ and $S = 0.3260$.

Is it possible to further improve the model without including more variables into the model? Here we consider the three types of descriptors as three block variables, say **P**, χ, and **H**. Following the calculation procedure stated in the Theory and Methodology section, three orthogonal canonical correlation variables corresponding to the three block variables are obtained. The correlation model including these three variables is then established with significantly improved results, say $R = 0.9997$ and $S = 0.1680$, respectively (see Table 4). The comparison of the residuals between the model including p_1, p_2, and p_3 and the model including the three canonical correlation variables of orthogonal block variables **P**, χ, and **H** is shown in Table 5. From this table one can easily see that residuals corresponding to the latter correlation model are much less than those of the former model for almost all compounds. The reason for this, in our opinion, is that these three new canonical orthogonal variables collect all the information from the original nine variables in the data. They are something like the principal components extracted from nine variables, but not being based only on their variation of themselves, as done by classical principal component analysis. This point will be further explained in the following paragraphs.

Limitation of PCR for Data Set 2. The data set 2 contains 32 individual variables (descriptors). Classical PCR was

Table 2. Block Descriptors and Retention Index of Alkanes (Data Set 2)

	$\mathbf{X}_1$				$\mathbf{X}_2$				$\mathbf{X}_3$				$\mathbf{X}_4$				$\mathbf{X}_5$			
compd	$^1\chi$	$^2\chi$	$^3\chi_p$	$^3\chi_c$	$^0\kappa$	$^1\kappa$	$^2\kappa$	$^3\kappa$	p_1	p_2	p_3	p_4	w_1	w_2	w_3	w_4	pw_1	pw_2	pw_3	pw_4
22334m5C5	4.1934	5.1264	3.3764	2.366	7.3645	10	2.56	1.9911	9	15	15	6	9	24	54	132	10	6.5675	3.6921	0.6857
2233m4C4	3.25	4.5	2.25	2.5	1.9538	8	1.75	2.2222	7	12	9	0	7	19	40	97	8	5.3571	2.5714	0
2233m4C6	4.3107	4.8839	2.9053	2.2071	7.9666	10	2.9388	2.6509	9	14	13	6	9	23	50	121	10	6.1786	3.3644	0.8583
2233m4C5	3.8107	4.4874	2.9142	2.2071	6.5548	9	2.3195	2	8	13	12	3	8	21	46	111	9	5.7786	3.2357	0.45
22344m5C5	4.1547	5.4537	2.5981	2.8764	4.729	10	2.56	3.1111	9	15	12	9	9	24	51	129	10	6.5	2.9524	0.8571
2234m4C5	3.8541	4.3987	2.366	1.866	6.5548	9	2.7222	2.88	8	12	10	6	8	20	42	100	9	5.6083	2.7393	0.7286
2235m4C6	4.3372	4.8966	2.3034	1.9784	7.9666	10	3.4083	4.48	9	13	10	7	9	22	45	106	10	6.0714	2.6968	0.8431
223m3-3eC5	4.3713	4.5178	3.3713	1.9786	7.3645	10	2.9388	1.9911	9	14	15	7	9	23	52	126	10	6.1841	3.8638	1.013
223m3C4	2.9434	3.5207	1.7321	1.6547	3.8823	7	1.8519	2.6667	6	9	6	0	6	15	30	69	7	4.4167	2	0
223m3C7	4.4814	4.4093	2.4691	1.5701	8.5686	10	4	4.48	9	12	10	6	9	21	43	98	10	5.7548	2.8091	0.8126
223m3C6	3.9814	4.0557	2.2001	1.5701	7.1568	9	3.2397	3.5556	8	11	9	5	8	19	39	90	9	5.2548	2.6145	0.7771
223m3C5	3.4814	3.6753	2.0908	1.5701	5.7934	8	2.52	2.8125	7	10	8	3	7	17	35	81	8	4.8214	2.4574	0.4912
2244m4C6	4.2678	5.2552	1.966	2.7678	7.9666	10	2.9388	5.5309	9	14	9	10	9	23	46	117	10	6.1333	2.3754	1.0146
2244m4C5	3.7071	5.2981	1.0607	3.1213	3.3172	9	2.3195	8	8	13	6	9	8	21	40	105	9	5.65	1.5	0.9
2245m4C6	4.3272	4.9861	2.0724	2.1297	7.9666	10	3.4083	5.5309	9	13	9	8	9	22	44	105	10	6.0643	2.4704	0.8794
224m3-3eC5	4.3921	4.6248	2.3569	1.8172	7.9666	10	3.4083	3.1111	9	13	12	11	9	22	47	114	10	5.9833	3.1641	1.4165
224m3C7	4.4545	4.6586	1.7423	1.8493	8.5686	10	4	7	9	12	8	9	9	21	41	97	10	5.6643	2.216	1.1649
224m3C6	3.9545	4.2782	1.6578	1.8493	7.1568	9	3.2397	5.8776	8	11	7	7	8	19	37	88	9	5.231	2.0624	0.8796
224m3C5	3.4165	4.1586	1.0206	1.9689	5.1913	8	2.52	7.2	7	10	5	6	7	17	32	78	8	4.7476	1.436	0.7773
2255m4C6	4.2071	5.6213	1.625	3.1213	4.127	10	2.9388	9.1429	9	14	7	6	9	23	44	109	10	6.2333	1.8566	0.5641
225m3C7	4.4545	4.6128	2.0841	1.8493	8.5686	10	4	7	9	12	8	6	9	21	41	94	10	5.7833	2.3454	0.7003
225m3C6	3.9165	4.4932	1.4717	1.9689	6.5548	9	3.2397	8	8	11	6	5	8	19	36	84	9	5.3	1.7465	0.5952
226m3C7	4.4165	4.8467	1.7083	1.9689	7.9666	10	4	9.1429	9	12	7	6	9	21	40	92	10	5.8	1.9371	0.658
22m2-3eC6	4.5194	4.2599	2.3547	1.5104	8.5686	10	4	3.7025	9	12	11	9	9	21	44	103	10	5.6417	3.0494	1.3544
22m2-3eC5	4.0194	3.8794	2.2103	1.5104	5.9527	9	3.2397	2.88	8	11	10	7	8	19	40	94	9	5.2083	2.9	1.1
22m2-4eC6	4.4925	4.4473	2.0557	1.7648	7.3645	10	4	5.5309	9	12	9	9	9	21	42	99	10	5.6643	2.5871	1.2085
22m2C3	2	3	0	2	1.0866	5	1	0	4	6	0	0	4	10	16	40	5	3	0	0
22m2C4	2.5607	2.9142	1.0607	1.5607	3.2375	6	1.6327	5.3333	5	7	3	0	5	12	22	51	6	3.55	1.2	0
22m2C7	4.0607	4.0178	1.5303	1.5607	7.1568	9	3.92	8	8	10	6	5	8	18	34	77	9	4.95	1.8095	0.6982
22m2C6	3.5607	3.6642	1.2803	1.5607	5.7934	8	3.1111	7.2	7	9	5	4	7	16	30	69	8	4.45	1.6	0.625
22m2C5	3.0607	3.3107	1	1.5607	4.4843	7	2.3438	6	6	8	4	3	6	14	26	61	7	3.95	1.3639	0.5417
22m2C8	4.5607	4.3713	1.7803	1.5607	8.5686	10	4.7603	9.1429	9	11	7	6	9	20	38	85	10	5.45	2.0595	0.8174
2334m4C6	4.4248	4.2854	3.3705	1.4035	8.7959	10	3.4083	2.2857	9	13	14	7	9	22	49	114	10	6.0714	3.7286	0.9832
2334m4C5	3.8868	4.1308	2.9761	1.488	4.9758	9	2.7222	2	8	12	12	4	8	20	44	102	9	5.6667	3.2857	0.5714
2335m4C6	4.3599	4.7413	2.4973	1.7474	8.1938	10	3.4083	3.7025	9	13	11	8	9	22	46	109	10	6.0595	2.9419	0.9717
233m3C7	4.504	4.2468	2.7376	1.3392	8.7959	10	4	3.7025	9	12	11	6	9	21	44	100	10	5.7619	3.0901	0.8457
233m3C6	4.004	3.8933	2.4573	1.3392	7.3841	9	3.2397	2.88	8	11	10	5	8	19	40	92	9	5.2619	2.8924	0.8139
233m3C5	3.504	3.4968	2.4742	1.3392	6.0206	8	2.52	2.2222	7	10	9	2	7	17	36	82	8	4.8619	2.7563	0.3761
2344m4C6	4.4147	4.3748	3.1439	1.5505	8.7959	10	3.4083	2.6509	9	13	13	8	9	22	48	113	10	6.0536	3.4806	1.0249
234m3-3eC5	4.4474	4.1688	3.3997	1.2701	6.9897	10	3.4083	1.9911	9	13	15	8	9	22	50	117	10	6.0722	3.9192	1.1636
234m3C6	4.0914	3.4887	2.5931	0.7615	7.9861	9	3.92	2.88	8	10	10	6	8	18	38	83	9	5.1381	2.9896	0.9248
234m3C5	3.5534	3.3472	2.1031	0.8591	4.2144	8	3.1111	2.8125	7	9	8	4	7	16	33	72	8	4.7048	2.4791	0.6154
235m3C6	4.0366	3.8508	1.9813	0.9773	7.3841	9	3.92	4.5	8	10	8	6	8	18	36	79	9	5.15	2.3929	0.8373
236m3C7	4.5366	4.1925	2.3374	0.9773	8.7959	10	4.7603	5.5309	9	11	9	6	9	20	40	86	10	5.6833	2.637	0.7464
23m2-3eC6	4.5647	3.9123	3.0226	1.0913	9.3979	10	4	2.6509	9	12	13	8	9	21	46	106	10	5.6833	3.5715	1.2914
23m2-3eC5	4.0647	3.5158	3.0092	1.0913	6.782	9	3.2397	2	8	11	12	5	8	19	42	96	9	5.2833	3.4429	0.8857
23m2-4eC6	4.6294	3.6797	2.8439	0.6925	8.1938	10	4.7603	3.1111	9	11	12	9	9	20	43	95	10	5.5429	3.4586	1.3982
23m2C4	2.6427	2.488	1.3333	0.6667	1.6586	6	2.2222	3	5	6	4	0	5	11	21	43	6	3.4667	1.6	0
23m2C7	4.1807	3.3635	2.1511	0.569	7.9861	9	4.8395	4.5	8	9	8	5	8	17	34	71	9	4.8333	2.5103	0.7963
23m2C6	3.6807	3.01	1.8821	0.569	6.6227	8	3.9375	3.6735	7	8	7	4	7	15	30	63	8	4.3333	2.3015	0.7394
23m2C5	3.1807	2.6295	1.782	0.569	5.3136	7	3.0612	2.6667	6	7	6	2	6	13	26	54	7	3.9	2.1333	0.4167
23m2C8	4.6807	3.7171	2.4011	0.569	9.3979	10	5.76	5.5309	9	10	9	6	9	19	38	79	10	5.3333	2.7364	0.8859
244m3C7	4.4772	4.5122	1.8876	1.6154	8.7959	10	4	5.5309	9	12	9	10	9	21	42	100	10	5.631	2.4597	1.3595
244m3C6	3.9772	4.1157	1.9179	1.6154	7.3841	9	3.2397	4.5	8	11	8	7	8	19	38	90	9	5.231	2.3224	0.9221
246m3C7	4.5197	4.3643	1.7475	1.1052	6.3876	10	4.7603	7	9	11	8	9	9	20	39	87	10	5.5667	2.2222	1.1991
24m2-3eC5	4.0914	3.5601	2.184	0.8027	5.5779	9	3.92	2.88	8	10	10	8	8	18	38	85	9	5.0917	2.9286	1.2571
24m2-3ipC5	4.4641	4.3094	2.3094	1.1111	3.8997	10	4	3.1111	9	12	12	12	9	21	45	105	10	5.8667	3.2	1.6
24m2-4eC6	4.5378	4.1157	2.6082	1.335	8.1938	10	4	3.7025	9	12	11	9	9	21	44	103	10	5.6762	3.0808	1.2603
24m2C7	4.1639	3.5234	1.6552	0.6969	7.9861	9	4.8395	5.8776	8	9	7	7	8	17	33	71	9	4.75	2.1361	1.1168
24m2C6	3.6639	3.143	1.5707	0.6969	6.6227	8	3.9375	5	7	8	6	5	7	15	29	62	8	4.3167	1.9745	0.8065
24m2C5	3.1259	3.0234	0.9428	0.8165	2.9054	7	3.0612	6	6	7	4	4	6	13	24	52	7	3.8333	1.3333	0.6667
24m2C8	4.6639	3.8769	1.9243	0.6969	9.3979	10	5.76	7	9	10	8	8	9	19	37	79	10	5.25	2.3643	1.1944
255m3C7	4.4772	4.4503	2.3556	1.6154	8.7959	10	4	5.5309	9	12	9	6	9	21	42	96	10	5.7833	2.6079	0.7304
25m2-3eC6	4.5746	4.0419	2.2413	0.9082	8.7959	10	4.7603	4.48	9	11	10	9	9	20	41	91	10	5.5548	2.8599	1.3175
25m2C7	4.1639	3.4846	1.9337	0.6969	7.9861	9	4.8395	5.8776	8	9	7	5	8	17	33	69	9	4.85	2.2393	0.7267
25m2C6	3.6259	3.365	1.3214	0.8165	3.6124	8	3.9375	7.2	7	8	5	4	7	15	28	59	8	4.3667	1.6263	0.5989
25m2C8	4.6639	3.865	2.0183	0.6969	9.3979	10	5.76	7	9	10	8	7	9	19	37	78	10	5.2833	2.411	1.06
26m2C7	4.1259	3.7186	1.5629	0.8165	4.9758	9	4.8395	8	8	9	6	5	8	17	32	67	9	4.8667	1.8318	0.6791
26m2C8	4.6639	3.8382	2.1753	0.6969	9.3979	10	5.76	7	9	10	8	6	9	19	37	77	10	5.35	2.4643	0.8256
27m2C8	4.6259	4.0722	1.8129	0.8165	5.7856	10	5.76	9.1429	9	10	7	6	9	19	36	75	10	5.3667	2.0985	0.8239
2m-33e2C5	4.6254	3.5777	3.3371	0.8985	6.5352	10	4	1.9911	9	12	15	9	9	21	48	111	10	5.6889	4.0909	1.4909
2m-3eC7	4.7187	3.5546	2.3957	0.5	9.3979	10	5.76	4.48	9	10	10	8	9	19	39	83	10	5.2381	2.9722	1.2935
2m-3eC6	4.2187	3.201	2.1267	0.5	7.9861	9	4.8395	3.5556	8	9	9	7	8	17	35	75	9	4.7381	2.7777	1.2547
2m-3eC5	3.7187	2.8206	1.9916	0.5	5.4185	8	3.9375	2.8125	7	8	8	5	7	15	31	66	8	4.3048	2.6205	0.9744
2m-3ipC6	4.5914	3.9406	2.3346	0.8027	6.9897	10	4.7603	3.7025	9	11	11	10	9	20	42	94	10	5.525	3.0779	1.5086

Table 2 (Continued)

	X$_1$				**X**$_2$				**X**$_3$				**X**$_4$				**X**$_5$			
compd	$^1\chi$	$^2\chi$	$^3\chi_p$	$^3\chi_c$	$^0\kappa$	$^1\kappa$	$^2\kappa$	$^3\kappa$	p_1	p_2	p_3	p_4	w_1	w_2	w_3	w_4	pw_1	pw_2	pw_3	pw_4
2m-4eC7	4.7019	3.6925	2.079	0.6124	9.3979	10	5.76	5.5309	9	10	9	9	9	19	38	82	10	5.1833	2.6579	1.4327
2m-4eC6	4.2019	3.3121	1.9594	0.6124	6.782	9	4.8395	4.5	8	9	8	7	8	17	34	73	9	4.75	2.5095	1.1667
2m-5eC7	4.7019	3.6537	2.307	0.6124	8.1938	10	5.76	5.5309	9	10	9	7	9	19	38	80	10	5.2833	2.7537	1.0821
2mC3	1.7321	1.7321	0	0.5774	.9769	4	1.3333	0	3	3	0	0	3	6	9	18	4	2	0	0
2mC4	2.2701	1.8021	.8165	0.4082	2.8928	5	2.25	4	4	4	2	0	4	8	14	27	5	2.5833	1	0
2mC7	3.7701	2.8896	1.385	0.4082	6.6227	8	5.1429	7.2	7	7	5	4	7	14	26	52	8	4.0167	1.6921	0.6943
2mC10	5.2701	3.9503	2.135	0.4082	10.8533	11	8.1	10	10	10	8	7	10	20	38	76	11	5.5167	2.4421	1.0631
2mC6	3.2701	2.5361	1.135	0.4082	5.3136	7	4.1667	6	6	6	4	3	6	12	22	44	7	3.5167	1.4659	0.5928
2mC9	4.7701	3.5967	1.885	0.4082	9.3979	10	7.1111	9.1429	9	9	7	6	9	18	34	68	10	5.0167	2.1921	0.9381
2mC12	6.2701	4.6574	2.635	0.4082	13.8792	13	10.0833	12	12	12	10	9	12	24	46	92	13	6.5167	2.9421	1.3131
2mC5	2.7701	2.1825	.866	0.4082	4.0668	6	3.2	5.3333	5	5	3	2	5	10	18	36	6	3.0167	1.21	0.4675
2mC8	4.2701	3.2432	1.635	0.4082	7.9861	9	6.125	8	8	8	6	5	8	16	30	60	9	4.5167	1.9421	0.8199
3344m4C6	4.3713	4.4749	3.5999	1.9142	6.9897	10	2.9388	1.9911	9	14	15	6	9	23	52	125	10	6.2	3.8952	0.8674
334m3C7	4.542	4.0319	2.967	1.2546	9.3979	10	4	3.1111	9	12	12	7	9	21	45	103	10	5.7	3.3532	1.0838
334m3C6	4.042	3.6514	2.8577	1.2546	7.9861	9	3.2397	2.3802	8	11	11	5	8	19	41	94	9	5.2667	3.2056	0.8225
335m3C7	4.5152	4.2353	2.5551	1.4958	9.3979	10	4	4.48	9	12	10	8	9	21	43	100	10	5.7143	2.9306	1.0045
33m2-4eC6	4.58	3.8556	3.002	1.1948	8.1938	10	4	2.6509	9	12	13	9	9	21	46	107	10	5.6536	3.6413	1.3949
33m2C7	4.1213	3.6213	2.1642	1.2071	7.9861	9	3.92	4.5	8	10	8	5	8	18	36	81	9	4.9333	2.4776	0.7844
33m2C6	3.6213	3.2678	1.8839	1.2071	6.6227	8	3.1111	3.6735	7	9	7	4	7	16	32	73	8	4.4333	2.2637	0.7303
33m2C5	3.1213	2.8713	1.9142	1.2071	4.1095	7	2.3438	2.6667	6	8	6	1	6	14	28	63	7	4.0333	2.1167	0.25
33m2C8	4.6213	3.9749	2.4142	1.2071	9.3979	10	4.7603	5.5309	9	11	9	6	9	20	40	89	10	5.4333	2.6872	0.8527
33e2C6	4.7426	3.3107	3.0303	0.7071	7.1373	10	4.7603	2.6509	9	11	13	9	9	20	44	100	10	5.3	3.7287	1.6047
33e2C5	4.2426	2.9142	3	0.7071	3.7717	9	3.92	2	8	10	12	6	8	18	40	90	9	4.9	3.6	1.2
344m3C7	4.542	4.0479	2.8408	1.2546	9.3979	10	4	3.1111	9	12	12	8	9	21	45	104	10	5.6667	3.3369	1.2411
34m2-3eC6	4.6027	3.6704	3.4175	1.0067	8.7959	10	4	2.2857	9	12	14	8	9	21	47	108	10	5.6881	3.8857	1.2965
34m2C7	4.2187	3.1515	2.3594	0.4714	8.5882	9	4.8395	3.5556	8	9	9	6	8	17	35	74	9	4.7667	2.8111	1.0643
34m2C6	3.7187	2.7711	2.2593	0.4714	4.8165	8	3.9375	2.8125	7	8	8	4	7	15	31	65	8	4.3333	2.6538	0.7715
34e2C6	4.7948	3.1532	2.7642	0.3333	4.5815	10	5.76	3.1111	9	10	12	10	9	19	41	89	10	5.1429	3.6	1.7544
35m2C7	4.2019	3.2625	2.1986	0.5774	6.1799	9	4.8395	4.5	8	9	8	6	8	17	34	72	9	4.8	2.5934	0.9196
3m-3eC7	4.682	3.6213	2.841	0.9268	8.7959	10	4.7603	3.7025	9	11	11	7	9	20	42	94	10	5.3786	3.2291	1.1365
3m-3eC6	4.182	3.2678	2.5607	0.9268	7.3841	9	3.92	2.88	8	10	10	6	8	18	38	86	9	4.8786	3.0313	1.1033
3m-3eC5	3.682	2.8713	2.5607	0.9268	4.362	8	3.1111	2.2222	7	9	9	3	7	16	34	76	8	4.4786	2.8952	0.6667
3m-4eC6	4.2567	2.9621	2.4974	0.4024	7.3841	9	4.8395	2.88	8	9	10	7	8	17	36	77	9	4.7381	3.131	1.2817
3m-5eC7	4.7399	3.4316	2.5873	0.4928	8.7959	10	5.76	4.48	9	10	10	8	9	19	39	83	10	5.2333	3.1123	1.2606
3mC7	3.8081	2.6556	1.7474	0.2887	7.2247	8	5.1429	5	7	7	6	4	7	14	27	54	8	4	2.0984	0.7515
3mC6	3.3081	2.3021	1.4784	0.2887	5.9157	7	4.1667	3.84	6	6	5	3	6	12	23	46	7	3.5	1.8702	0.6653
3mC9	4.8081	3.3628	2.2474	0.2887	10	10	7.1111	7	9	9	8	6	9	18	35	70	10	5	2.5746	0.9647
3mC12	6.3081	4.4234	2.9974	0.2887	14.4813	13	10.0833	9.9174	12	12	11	9	12	24	47	94	13	6.5	3.3246	1.3328
3mC5	2.8081	1.9217	1.3938	0.2887	3.4648	6	3.2	3	5	5	4	1	5	10	19	37	6	3.0667	1.6857	0.2857
3mC8	4.3081	3.0092	1.9974	0.2887	8.5882	9	6.125	5.8776	8	8	7	5	8	16	31	62	9	4.5	2.3246	0.8464
3eC7	4.3461	2.8248	2.1206	0.2041	7.3841	9	6.125	4.5	8	8	8	6	8	16	32	65	9	4.4333	2.6269	1.1296
3eC6	3.8461	2.4712	1.8516	0.2041	6.0206	8	5.1429	3.6735	7	7	7	5	7	14	28	57	8	3.9333	2.4182	1.07
3eC5	3.3461	2.0908	1.7321	0.2041	3.053	7	4.1667	2.6667	6	6	6	3	6	12	24	48	7	3.5	2.25	0.75
3eC8	4.8461	3.1783	2.3706	0.2041	8.7959	10	7.1111	5.5309	9	9	9	7	9	18	36	73	10	4.9333	2.8531	1.2192
44m2C7	4.1213	3.6642	1.8536	1.2071	6.1799	9	3.92	4.5	8	10	8	7	8	18	36	83	9	4.8333	2.4056	1.1881
44m2C8	4.6213	4.0178	2.1339	1.2071	9.3979	10	4.7603	5.5309	9	11	9	8	9	20	40	91	10	5.3333	2.6194	1.24
4pC7	4.8461	3.2321	2.0908	0.2041	5.7059	10	7.1111	5.5309	9	9	9	9	9	18	36	75	10	4.8	2.7238	1.6299
4m-3eC7	4.7567	3.3425	2.5975	0.4024	8.7959	10	5.76	3.7025	9	10	11	9	9	19	40	86	10	5.1714	3.2794	1.5514
4m-4eC7	4.682	3.6642	2.5607	0.9268	10	10	4.7603	3.7025	9	11	11	9	9	20	42	96	10	5.2786	3.1627	1.521
4mC7	3.8081	2.6825	1.5629	0.2887	5.4185	8	5.1429	5	7	7	6	5	7	14	27	55	8	3.9333	2.0419	1.0071
4mC9	4.8081	3.3896	2.082	0.2887	10	10	7.1111	7	9	9	8	7	9	18	35	71	10	4.9333	2.4963	1.1793
4mC12	6.3081	4.4503	2.832	0.2887	14.4813	13	10.0833	9.9174	12	12	11	10	12	24	47	95	13	6.4333	3.2463	1.5407
4mC8	4.3081	3.0361	1.832	0.2887	8.5882	9	6.125	5.8776	8	8	7	6	8	16	31	63	9	4.4333	2.2701	1.0887
4eC7	4.3461	2.8516	1.9712	0.2041	6.782	9	6.125	4.5	8	8	8	7	8	16	32	66	9	4.3667	2.5754	1.3611
4eC8	4.8461	3.2052	2.2402	0.2041	10	10	7.1111	5.5309	9	9	9	8	9	18	36	74	10	4.8667	2.7841	1.4171
4ipC7	4.7187	3.5814	2.2617	0.5	7.5918	10	5.76	4.48	9	10	10	9	9	19	39	84	10	5.1714	2.9253	1.5116
5mC9	4.8081	3.3896	2.101	0.2887	7.5918	10	7.1111	7	9	9	8	7	9	18	35	71	10	4.9333	2.4983	1.1663
5mC12	6.3081	4.4503	2.851	0.2887	14.4813	13	10.0833	9.9174	12	12	11	10	12	24	47	95	13	6.4333	3.2245	1.4934
6mC12	6.3081	4.4503	2.851	0.2887	14.4813	13	10.0833	9.9174	12	12	11	10	12	24	47	95	13	6.4333	3.2007	1.4659
C3	1.4142	0.7071	0	0	.8293	3	2	0	2	1	0	0	2	3	4	6	3	1	0	0
C4	1.9142	1	0.5	0	1.2041	4	3	4	3	2	1	0	3	5	8	13	4	1.6667	.6667	0
C7	3.4142	2.0607	1.2071	0	4.1095	7	6	6	6	5	4	3	6	11	20	37	7	3.1667	1.5357	0.6762
C10	4.9142	3.1213	1.9571	0	6.9897	10	9	9.1429	9	8	7	6	9	17	32	61	10	4.6667	2.2857	1.0595
C6	2.9142	1.7071	0.9571	0	2.8627	6	5	5.3333	5	4	3	2	5	9	16	29	6	2.6667	1.2857	0.5333
C9	4.4142	2.7678	1.7071	0	6.1799	9	8	8	8	7	6	5	8	15	28	53	9	4.1667	2.0357	0.9345
C12	5.9142	3.8284	2.4571	0	9.3378	12	11	11.1111	11	10	9	8	11	21	40	77	12	5.6667	2.7857	1.3095
C16	7.9142	5.2426	3.4571	0	14.4494	16	15	15.0769	15	14	13	12	15	29	56	109	16	7.6667	3.7857	1.8095
C13	6.4142	4.182	2.7071	0	10.8689	13	12	12	12	11	10	9	12	23	44	85	13	6.1667	3.0357	1.4345
C14	6.9142	4.5355	2.9571	0	11.8314	14	13	13.0909	13	12	11	10	13	25	48	93	14	6.6667	3.2857	1.5595
C15	7.4142	4.8891	3.2071	0	13.4269	15	14	14	14	13	12	11	14	27	52	101	15	7.1667	3.5357	1.6845
C11	5.4142	3.4749	2.2071	0	8.445	11	10	10	10	9	8	7	10	19	36	69	11	5.1667	2.5357	1.1845
C5	2.4142	1.3536	.7071	0	2.2907	5	4	4	4	3	2	1	4	7	12	21	5	2.1667	1	0.3333
C8	3.9142	2.4142	1.4571	0	4.8165	8	7	7.2	7	6	5	4	7	13	24	45	8	3.6667	1.7857	0.8095
C2	1	0	0	0	0	2	0	0	1	0	1	0	1	1	1	1	2	0	2	0

Table 2 (Continued)

	X_6				X_7		X_8						
compd	MTI	PED	PEAD	DET	Y_x	EAID	heat form.	elec. energy	core energy	ion. energy	dipole	HUMO	retention index
22334m5C5	390	22.2797	23.7206	4.5666	3.8116	27.9672	−43.50514	−10359.75	8775.176	10.6757	.01136	3.566	953.4
2233m4C4	214	14.9373	16.3459	3.5465	3.3946	19.6514	−40.89208	−7334.599	6061.211	10.9787	.0003	3.619	728.69
2233m4C6	416	23.8806	25.3631	4.6612	3.8763	26.4029	−49.85479	−10184.86	8600.009	10.81639	.02406	3.597	928.8
2233m4C5	298	18.844	20.2756	4.105	3.6301	23.0138	−45.31129	−8744.597	7315.479	10.85189	.00424	3.61	855.13
22344m5C5	402	22.8477	24.3074	4.5733	3.8389	27.1709	−44.50718	−10298.45	8713.832	10.68827	.02712	3.532	921.7
2234m4C5	312	19.7257	21.1798	4.1545	3.6737	20.9282	−47.21812	−8660.957	7231.757	10.799	.01352	3.6	822.07
2235m4C6	446	25.4087	26.9248	4.7082	3.9396	23.6434	−56.17962	−9956.045	8370.919	10.8477	.0083	3.557	873.3
223m3-3eC5	396	22.8011	24.2476	4.6631	3.8261	26.1074	−47.96047	−10288.09	8703.325	10.60393	.02938	3.593	965.7
223m3C4	156	12.3945	13.8023	3.1284	3.1917	14.5406	−40.31712	−5901.524	4783.7	10.96893	.01124	3.651	641.46
223m3C7	472	27.013	28.5623	4.775	3.9974	22.6971	−59.41519	−9726.869	8141.603	10.91321	.01203	3.564	914.4
223m3C6	334	21.225	22.7219	4.2475	3.7425	19.914	−52.5564	−8440.196	7010.765	10.91405	.01278	3.586	823.18
223m3C5	230	16.3152	17.7574	3.6913	3.4691	17.0939	−45.7546	−7168.514	5894.916	10.94434	.00713	3.628	738.98
2244m4C6	432	24.5739	26.0746	4.6685	3.9075	24.884	−50.73219	−10054.49	8469.602	10.73394	.03091	3.545	888.6
2244m4C5	322	20.1263	21.608	4.1242	3.701	22.0027	−46.37195	−8617.877	7188.713	10.84266	.03158	3.616	774.77
2245m4C6	450	25.572	27.0918	4.7098	3.9465	23.2751	−55.2219	−9924.192	8339.107	10.85359	.02394	3.529	872.1
224m3-3eC5	414	23.8035	25.2717	4.7163	3.8658	23.5956	−49.32553	−10167.72	8582.889	10.74407	.02811	3.546	903.9
224m3C7	476	27.1627	28.7142	4.7807	4.0018	22.0956	−59.38599	−9710.798	8125.532	10.86166	.02109	3.552	875.7
224m3C6	342	21.6063	23.1144	4.2522	3.7619	19.4968	−52.5686	−8412.023	6982.592	10.91496	.02438	3.588	790.60
224m3C5	242	17.0338	18.5101	3.7038	3.516	16.847	−46.77102	−7108.715	5835.073	11.00219	.01496	3.65	691.55
2255m4C6	464	26.116	27.6594	4.672	3.975	25.2168	−57.7112	−9866.361	8281.169	10.9892	.00112	3.603	820.2
225m3C7	488	27.6823	29.2455	4.7778	4.0245	22.1878	−61.44294	−9642.028	8056.674	10.95924	.00715	3.573	878.1
225m3C6	358	22.4662	24.0028	4.2545	3.8056	19.6122	−55.55449	−8299.761	6870.199	11.03712	.01606	3.608	777.07
226m3C7	508	28.667	30.253	4.772	4.063	22.2744	−62.44179	−9549.044	7963.646	10.97658	.01417	3.584	873
22m2-3eC6	440	25.3826	26.8878	4.8009	3.9254	22.4631	−57.65448	−9931.48	8346.29	10.80742	.01895	3.572	902.1
22m2-3eC5	318	20.2981	21.7593	4.2533	3.6915	19.7664	−50.82289	−8571.786	7142.43	10.8455	.01823	3.61	824.28
22m2-4eC6	456	26.101	27.6241	4.8005	3.9572	22.0183	−58.77768	−9914.725	8329.485	10.84355	.01602	3.574	881.3
22m2C3	64	6.6056	8	2.1072	2.5809	8.8515	−32.83527	−3528.125	2721.7	11.53374	.00009	3.905	412.57
22m2C4	106	9.6702	11.0769	2.6665	2.937	11.1138	−37.71373	−4631.849	3669.675	11.18348	.00502	3.752	537.77
22m2C7	380	23.9634	25.5329	4.3224	3.866	18.7515	−58.2038	−8111.55	6681.874	11.02231	.00948	3.621	816.16
22m2C6	260	18.4133	19.9327	3.7938	3.5898	16.1431	−51.34291	−6916.208	5642.368	11.03788	.01187	3.657	720.17
22m2C5	170	13.6353	15.097	3.2375	3.2799	13.6669	−44.48559	−5751.119	4633.113	11.09381	.00789	3.713	626.55
22m2C8	534	30.2636	31.8749	4.8351	4.1126	21.2942	−65.06525	−9344.763	7759.251	11.01967	.01297	3.59	914.9
2334m4C6	414	23.8455	25.3161	4.6988	3.8705	23.8665	−52.34581	−10131.5	8546.545	10.70593	.01201	3.549	949.1
2334m4C5	304	19.3005	20.7375	4.1486	3.6498	21.0577	−46.5384	−8698.136	7268.965	10.77915	.01342	3.593	861.15
2335m4C6	434	24.8365	26.3362	4.7065	3.9139	23.4154	−54.76602	−10004.06	8418.996	10.80911	.03516	3.518	903.3
233m3C7	460	26.4376	27.9722	4.7761	3.9732	22.4914	−58.46761	−9791.058	8205.832	10.86492	.02361	3.556	931.7
233m3C6	326	20.7945	22.2755	4.2451	3.7192	19.7529	−51.60917	−8488.587	7059.197	10.85973	.01938	3.579	841.89
233m3C5	226	16.0683	17.4976	3.687	3.4522	17.0151	−44.85519	−7204.111	5930.551	10.90266	.01942	3.623	761.71
2344m4C6	418	24.0207	25.4963	4.7005	3.8786	23.556	−51.48589	−10100	8515.08	10.69922	.0295	3.504	935
234m3-3eC5	402	23.2165	24.6664	4.7065	3.8402	23.6579	−49.67133	−10244.79	8659.945	10.58441	.02111	3.584	969.4
234m3C6	332	21.197	22.6797	4.2858	3.7354	18.093	−53.75174	−8436.201	7006.718	10.77514	.01169	3.536	850.88
234m3C5	236	16.8079	18.2572	3.7356	3.4943	15.7844	−47.86289	−7112.205	5838.515	10.82493	.00176	3.59	754.14
235m3C6	348	22.0627	23.5748	4.2919	3.7791	18.0256	−55.06606	−8365.301	6935.76	10.89763	.01732	3.556	813.05
236m3C7	494	28.117	29.6824	4.813	4.0367	20.3567	−62.51926	−9585.047	7999.646	10.91229	.01066	3.53	919
23m2-3eC6	428	24.7922	26.2807	4.7945	3.901	22.1567	−54.72075	−10040.06	8455.002	10.66324	.02532	3.575	949.4
23m2-3eC5	310	19.8563	21.3004	4.2451	3.6687	19.5688	−47.99347	−8644.609	7215.376	10.70035	.02189	3.612	875.00
23m2-4eC6	442	25.5399	27.0397	4.836	3.9295	20.289	−58.02816	−9937.877	8352.671	10.74124	.01698	3.554	930.6
23m2C4	108	10	11.4031	2.7093	2.9605	10.5556	−38.8557	−4607.479	3645.256	11.05099	.0167	3.723	568.32
23m2C7	370	23.5541	25.1011	4.3621	3.8408	17.1759	−58.44774	−8147.541	6717.854	10.93728	.02129	3.579	855.34
23m2C6	254	18.1815	19.6756	3.8305	3.5678	14.9316	−51.58796	−6938.178	5664.326	10.95772	.01924	3.611	760.79
23m2C5	168	13.6346	15.0731	3.2723	3.2702	12.761	−44.76133	−5761.819	4643.802	11.05977	.01857	3.663	672.28
23m2C8	520	29.7093	31.3017	4.8737	4.0878	19.3839	−65.30891	−9387.703	7802.181	10.91738	.01932	3.553	952.1
244m3C7	460	26.4127	27.9454	4.7826	3.9695	21.8502	−59.20395	−9780.46	8195.202	10.87702	.01596	3.539	889.4
244m3C6	334	21.1839	22.6773	4.2498	3.7397	19.3591	−52.44712	−8450.309	7020.881	10.89239	.0256	3.574	809.56
246m3C7	490	27.9451	29.5065	4.8178	4.0278	20.1022	−63.25566	−9624.39	8038.957	10.91867	.00972	3.524	870.1
24m2-3eC5	324	20.7438	22.2092	4.2975	3.7087	18.0267	−51.25299	−8531.783	7102.408	10.65053	.03613	3.58	838.17
24m2-3ipC5	420	24.219	25.6902	4.7604	3.8786	21.434	−48.93039	−10124.45	8539.641	10.62554	.04698	3.536	915.1
24m2-4eC6	440	25.3399	26.8429	4.7959	3.9254	21.6949	−56.09631	−9979.855	8394.732	10.70683	.01501	3.586	920.7
24m2C7	370	23.5441	25.0901	4.3664	3.8385	16.9909	−58.48207	−8137.65	6707.962	10.95256	.01966	3.551	821.28
24m2C6	258	18.3964	19.8985	3.8335	3.5808	14.8294	−51.66034	−6914.026	5640.171	11.0251	.0195	3.587	732.69
24m2C5	176	14.176	15.6472	3.2833	3.3137	12.7196	−45.75026	−5703.545	4585.485	11.12265	.01742	3.631	630.25
24m2C8	516	29.529	31.1174	4.8813	4.0785	19.1423	−65.34251	−9392.599	7807.075	10.92806	.01948	3.527	915.8
255m3C7	476	27.1219	28.6717	4.7788	4.0018	22.0142	−60.18868	−9710.834	8125.534	10.96445	.01845	3.567	891.7
25m2-3eC6	458	26.3381	27.8591	4.842	3.961	20.1704	−59.64742	−9863.368	8278.092	10.77954	.01585	3.527	891.4
25m2C7	378	23.9292	25.4857	4.3639	3.8585	17.0502	−59.24589	−8097.021	6667.3	10.98061	.01081	3.568	833.21
25m2C6	270	19.1115	20.6428	3.8366	3.6242	14.9164	−53.37681	−6830.822	5556.894	11.05755	.00024	3.604	728.82
25m2C8	520	29.6877	31.2789	4.8805	4.085	19.1634	−66.06484	−9373.651	7788.096	10.93027	.00467	3.536	921.8
26m2C7	394	24.7896	26.3702	4.3595	3.8978	17.1264	−60.25541	−8014.733	6584.968	11.00163	.00672	3.589	827.46
26m2C8	532	30.2142	31.8156	4.8729	4.1066	19.2187	−66.13414	−9328.573	7743.015	10.93965	.00636	3.55	931.5
27m2C8	552	31.1984	32.8195	4.8691	4.1412	19.2864	−67.11765	−9234.642	7649.041	11.02139	.02196	3.565	928.5
2m-33e2C5	408	23.7103	25.1647	4.8028	3.8538	21.9335	−49.1577	−10216.2	8631.379	10.60048	.06617	3.582	984
2m-3eC7	484	27.9218	29.4739	4.8966	4.0162	19.2208	−63.36866	−9641.531	8056.093	10.87932	.01557	3.595	941
2m-3eC6	346	22.2198	23.7242	4.384	3.7725	17.034	−56.51208	−8341.148	6911.546	10.90291	.01216	3.63	844.75
2m-3eC5	242	17.4187	18.8774	3.8345	3.5181	14.8446	−49.68597	−7067.82	5794.051	10.87628	.00826	3.677	762.57
2m-3ipC6	446	25.7896	27.2967	4.8436	3.9366	20.3033	−54.33541	−9923.896	8338.85	10.62574	.03415	3.578	915.5

Table 2 (Continued)

	$\mathbf{X}_6$				$\mathbf{X}_7$		$\mathbf{X}_8$						
compd	MTI	PED	PEAD	DET	Y_x	EAID	heat form.	elec. energy	core energy	ion. energy	dipole	HUMO	retention index
2m-4eC7	484	27.8955	29.4461	4.9117	4.0129	19.0307	−64.46569	−9631.298	8045.813	10.92379	.01548	3.585	907.4
2m-4eC6	354	22.6204	24.1368	4.3837	3.7932	16.9243	−57.62773	−8306.392	6876.741	10.91731	.01834	3.619	824.88
2m-5eC7	500	28.6303	30.1974	4.8906	4.0454	19.1006	−64.29136	−9529.33	7943.852	10.87701	.0179	3.576	924
2mC3	36	4.6458	6	1.6812	2.2279	5.9437	−29.42095	−2501.393	1850.654	11.29216	.00968	3.834	365.61
2mC4	68	7.4593	8.8535	2.2455	2.6637	8.0168	−35.45876	−3459.122	2652.583	11.19864	.00959	3.747	475.34
2mC7	288	20.4792	22.045	3.9081	3.6885	14.1687	−56.00821	−6668.959	5394.916	11.04465	.01298	3.622	764.95
2mC10	778	39.9915	41.6629	5.4415	4.3971	20.3513	−76.59119	−10295.41	8553.864	11.03033	.01108	3.549	1062.3
2mC6	190	15.4048	16.9205	3.3751	3.3918	12.1052	−49.14733	−5546.744	4428.536	11.06544	.01107	3.656	666.89
2mC9	578	32.7777	34.4198	4.9349	4.1853	18.3046	−69.73028	−9047.42	7461.706	11.03547	.01301	3.569	963.9
2mC12	1308	56.5338	58.2516	6.4435	4.7672	24.3879	−90.3129	−12893.23	10840.01	11.00102	.01105	3.518	1264.1
2mC5	118	11.0588	12.5154	2.8169	3.0527	10.0078	−42.28732	−4472.773	3510.4	11.11751	.01295	3.7	570.07
2mC8	416	26.2722	27.8796	4.4247	3.9508	16.2497	−62.86934	−7837.876	6407.998	11.03852	.01113	3.593	864.86
3344m4C6	400	23.0224	24.4764	4.6545	3.8377	26.1282	−49.66267	−10230.82	8645.975	10.71371	.00565	3.577	983.7
334m3C7	444	25.6108	27.1231	4.779	3.9386	22.3821	−57.44845	−9880.618	8295.437	10.81671	.01221	3.557	936.6
334m3C6	318	20.3172	21.7797	4.2407	3.6947	19.6926	−50.63471	−8543.778	7114.431	10.84068	.00802	3.6	855.25
335m3C7	456	26.1375	27.662	4.7819	3.961	21.8998	−57.18019	−9831.447	8246.277	10.8112	.02293	3.543	907.7
33m2-4eC6	424	24.5445	26.0248	4.7943	3.891	22.2137	−52.34786	−10056.6	8471.642	10.66368	.03666	3.596	937.8
33m2C7	356	22.6772	24.2083	4.3253	3.8043	18.5345	−55.98761	−8252.36	6822.78	10.98146	.01333	3.607	837.09
33m2C6	244	17.4426	18.9193	3.787	3.5285	15.9681	−49.12979	−7023.271	5749.526	10.98769	.00914	3.645	744.81
33m2C5	162	13.0698	14.4942	3.2253	3.2327	13.5841	−42.37236	−5827.91	4709.996	11.10235	.00755	3.697	660.39
33m2C8	502	28.7178	30.2966	4.84	4.0554	21.0528	−62.8485	−9503.495	7918.08	10.95726	.01023	3.577	932
33e2C6	434	25.2716	26.7631	4.8858	3.9128	20.4791	−53.97946	−10035.38	8450.349	10.69107	.06001	3.622	954.1
33e2C5	316	20.3923	21.8425	4.3415	3.6867	18.1419	−47.31306	−8636.945	7207.742	10.68934	.05316	3.665	880.34
344m3C7	440	25.4248	26.9318	4.7799	3.9295	22.3057	−57.37407	−9898.987	8313.81	10.8048	.01354	3.564	932.2
34m2-3eC6	420	24.3503	25.8251	4.7902	3.8832	22.0958	−53.07305	−10107.53	8522.538	10.68526	.0147	3.594	964.6
34m2C7	354	22.6789	24.199	4.3653	3.7982	17.0667	−57.60669	−8236.623	6806.973	10.87213	.01526	3.579	859.56
34m2C6	246	17.6759	19.1471	3.8261	3.5367	14.8677	−50.77153	−6993.059	5719.243	10.92834	.01135	3.623	771.84
34e2C6	448	26.0476	27.5509	4.9247	3.9406	19.0066	−60.60211	−9889.468	8304.149	10.8173	.02543	3.601	945.8
35m2C7	362	23.0687	24.5995	4.3665	3.8176	16.9605	−63.85483	−8597.404	7140.171	10.96583	.02732	3.575	834.26
3m-3eC7	466	26.9281	28.4646	4.8564	3.9832	20.7674	−58.69095	−9929.194	8343.959	10.65178	.02638	3.653	953
3m-3eC6	332	21.3349	22.8207	4.3365	3.7354	18.299	−52.46083	−8499.878	7070.451	10.69321	.01361	3.677	855.42
3m-3eC5	232	16.6705	18.1096	3.7836	3.4784	15.8462	−45.74374	−7192.715	5919.117	10.74486	.00829	3.69	776.13
3m-4eC6	338	21.7527	23.2408	4.3796	3.7509	16.9702	−54.8406	−8375.911	6946.381	10.85269	.01559	3.58	856.16
3m-5eC7	480	27.6466	29.1905	4.8993	4.0069	19.0093	−61.30532	−9773.897	8188.549	10.71181	.01249	3.647	924
3mC7	276	19.7628	21.3026	3.9099	3.6483	14.0988	−54.39288	−6804.25	5530.277	10.91392	.00561	3.644	772.67
3mC6	182	14.8636	16.3497	3.3707	3.3504	12.0447	−48.26163	−5614.455	4496.286	10.98817	.00928	3.667	676.60
3mC9	558	31.7954	33.4197	4.9364	4.1529	18.2377	−68.10435	−9251.791	7666.147	10.85035	.00955	3.62	969.62
3mC12	1276	55.2712	56.9787	6.4439	4.745	24.3416	−88.68764	−13160.83	11107.68	10.84349	.00889	3.577	1270.1
3mC5	114	10.7424	12.1719	2.8089	3.0191	9.9767	−41.43589	−4520.207	3557.871	11.07902	.00815	3.699	584.70
3mC8	400	25.4119	26.9977	4.4279	3.9143	16.18	−61.24434	−8003.184	6573.375	10.85159	.00857	3.632	870.35
3eC7	376	24.0988	25.6486	4.4445	3.853	16.0738	−60.33007	−8157.519	6727.75	10.84026	.01168	3.621	867.45
3eC6	260	18.7788	20.2794	3.9243	3.588	14.0048	−54.20422	−6874.462	5600.497	10.91421	.00197	3.643	773.10
3eC5	174	14.2969	15.7492	3.3714	3.3047	11.9902	−47.37917	−5689.736	4571.604	11.00171	.00611	3.681	686.80
3eC8	526	30.2108	31.8041	4.9463	4.0956	18.1274	−67.18213	−9428.95	7843.347	10.80058	.00824	3.613	964
44m2C7	348	22.2705	23.7886	4.328	3.7817	18.4386	−55.88533	−8283.897	6854.322	10.9544	.00123	3.602	828.71
44m2C8	486	27.9952	29.558	4.8491	4.0237	20.9201	−62.74371	−9568.881	7983.47	10.93572	.00839	3.569	918
4pC7	498	28.8746	30.4392	4.9823	4.0383	18.032	−66.48675	−9611.935	8026.361	10.88716	.01408	3.585	906
4m-3eC7	468	27.1117	28.644	4.8999	3.985	19.1099	−61.60451	−9742.1	8156.738	10.78757	.01358	3.554	940.5
4m-4eC7	454	26.3477	27.8691	4.8716	3.9562	20.6551	−59.32462	−9809.879	8224.616	10.77674	.01162	3.608	937.6
4mC7	272	19.542	21.0738	3.9117	3.634	14.0661	−54.98279	−6755.896	5481.897	11.00483	.01211	3.598	767.48
4mC9	546	31.2641	32.879	4.9421	4.1323	18.1891	−68.70417	−9179.29	7593.621	10.9493	.01153	3.543	960
4mC12	1252	54.4216	56.1224	6.4447	4.7276	24.287	−89.287	−13061.63	11008.45	10.90795	.01206	3.496	1258.3
4mC8	392	25.0207	26.5968	4.4337	3.8948	16.1365	−61.84335	−7951.1	6521.266	10.97348	.01262	3.567	863.16
4eC7	368	23.6799	25.2175	4.4594	3.8305	16.034	−60.04227	−8136.145	6706.389	10.8722	.00392	3.586	857.82
4eC8	510	29.4624	31.0402	4.9657	4.0637	18.0758	−66.89917	−9423.204	7837.612	10.84146	.0103	3.562	951.5
4ipC7	472	27.3408	28.8787	4.9168	3.9895	19.1699	−62.29878	−9688.523	8103.132	10.85306	.02841	3.515	925
5mC9	542	31.0969	32.7089	4.9468	4.1252	18.1821	−68.70337	−9196.188	7610.518	10.94881	.01683	3.541	957.4
5mC12	1236	53.9017	55.5987	6.4467	4.7156	24.2764	−89.28639	−13103.44	11050.26	10.90097	.01268	3.491	1252.4
6mC12	1228	53.6556	55.351	6.4466	4.7094	24.275	−89.28254	−13122.95	11069.78	10.89464	.01135	3.489	1249.9
C3	16	2.7321	4	1.2041	1.7404	4.0019	−24.3021	−1635.942	1140.963	11.32464	.00428	3.92	300
C4	38	5.1623	6.5311	1.7782	2.3032	5.7397	−31.17807	−2454.979	1804.164	11.17192	.0003	3.829	400
C7	204	16.6254	18.1783	3.4533	3.4622	11.4735	−51.75943	−5404.033	4285.712	11.07227	.00643	3.682	700
C10	604	34.3429	36.0037	5.0004	4.2264	17.3227	−72.34198	−8866.062	7280.234	11.06314	.00051	3.577	1000
C6	128	12.1093	13.6116	2.9122	3.1364	9.5262	−44.89873	−4353.036	3390.55	11.08452	.00034	3.736	600
C9	438	27.7422	29.3729	4.4915	3.9996	15.3724	−65.48109	−7666.323	6236.332	11.06405	.00645	3.605	900
C12	1054	49.6287	51.3375	6.0061	4.62	21.2144	−86.06287	−11381.7	9484.198	11.05911	.00079	3.537	1200
C16	2538	88.5384	90.3119	7.9801	5.2431	28.9215	−113.5067	−16796.89	14276.06	10.9727	.00109	3.488	1600
C13	1346	58.3139	60.042	6.5035	4.7932	23.1511	−92.924	−12691.11	10637.77	11.03233	.0065	3.521	1300
C14	1688	67.6939	69.439	6.998	4.9537	25.0807	−99.78485	−14031.5	11822.33	11.00937	.00095	3.509	1400
C15	2084	77.7687	79.5288	7.4901	5.1032	27.0042	−106.6458	−15400.72	13035.72	10.98969	.0064	3.498	1500
C11	808	41.6384	43.3249	5.5052	4.432	19.2722	−79.20287	−10105.7	8364.034	11.06302	.0064	3.555	1100
C5	74	8.2882	9.7297	2.3502	2.7566	7.6364	−38.03782	−3366.513	2559.863	11.11116	.00619	3.775	500
C8	306	21.8364	23.4316	3.9771	3.747	13.4156	−58.62032	−6510.521	5236.364	11.06658	.00032	3.639	800
C2	4	1	2	.6021	.9542	2.2927	−17.44032	−930.5031	591.3596	11.766	.00008	4.117	200

Table 3. Regression Results with Orthogonal Variables (Vectors) of the Three Types of Topological Indices

statistical data			regression coefficients			
F	*R*	*S*	p_1	p_2	p_3	constant
22.3620	0.8313	3.4360	1.4362			−37.6432
211.6729	0.9895	0.9401	1.4362	−7.1808		−37.6432
1195.7	0.9989	0.3260	1.4362	−7.1808	0.6295	−37.6432

statistical data			regression coefficients			
F	*R*	*S*	$^1\chi$	$^2\chi$	$^3\chi$	constant
39.7598	0.8939	2.7710	5.3497	−49.5685		
72.4536	0.9703	1.5756	5.3497	−36.4782	−49.5685	
44.9390	0.9716	1.6356	5.3497	−36.4782	−2.7314	−49.5685

statistical data			regression coefficients			
F	*R*	*S*	H	H′	H″	constant
17.6527	0.7990	3.7171	0.1553	−20.1174		
9.5747	0.8248	3.6842	0.1553	−0.3276	−20.1174	
14.3894	0.9188	2.7326	0.1553	−0.3276	−4.4981	−20.1174

Table 4. Regression Results with Orthogonal Variable Matrix of the Three Types of Topological Indices

statistical data			regression coefficients			
F	*R*	*S*	**P**	χ	**H**	constant
4484.05	0.9989	0.2916	5.8870			18.3675
2207.27	0.9990	0.2939	5.8870	0.08136		18.3675
4511.24	0.9997	0.1680	5.8870	0.08136	0.2239	18.3675

Table 5. Residuals of the Regressions of Retention Time against Descriptors p_1, p_2, and p_3 and against Three Canonical Correlation Variables of Orthogonal Block Variables **P**, χ, and **H**

	residual			residual	
compd no.	p_1, p_2, p_3	**P**, χ, **H**	compd no.	p_1, p_2, p_3	**P**, χ, **H**
1	0.2853	0.1026	7	−0.3863	−0.0787
2	−0.2087	−0.0558	8	0.0297	0.0509
3	0.1613	−0.0431	9	0.2197	0.0001
4	−0.1927	0.1317	10	0.2757	−0.1317
5	−0.2727	−0.2615	11	0.2857	0.2665
6	0.2273	0.1261	12	−0.4243	−0.1071

Table 6. Correlation Coefficient and Standard Error of the Retention-Component in PCR

pc	*R*	*S*	pc	*R*	*S*	pc	*R*	*S*
1	0.9689	47.8761	12	0.0202	193.5115	23	0.0065	193.5470
2	0.2043	189.4675	13	0.0009	193.5510	24	0.0024	193.5505
3	0.0667	193.1196	14	0.0130	193.5348	25	0.0081	193.5448
4	0.0723	193.0450	15	0.0160	193.5264	26	0.0050	193.5487
5	0.0218	193.5050	16	0.0150	193.5294	27	0.0052	193.5485
6	0.0591	193.2128	17	0.0009	193.5510	28	0.0019	193.5507
7	0.0347	193.4345	18	0.0320	193.4521	29	0.0000	193.5511
8	0.0082	193.5446	19	0.0044	193.5492	30	0.0000	193.5511
9	0.0121	193.5369	20	0.0036	193.5498	31	0.0000	193.5511
10	0.0113	193.5387	21	0.0062	193.5474	32	0.0000	193.5511
11	0.0375	193.4149	22	0.0024	193.5505			

made for the data set of 32 descriptors. Table 6 shows the correlation coefficients *R* and the standard errors *S* of the regression model between retention index (property) and all the individual principal components, respectively. From this table one can see that the first two principal components describe most of systematic variation (variance) of the data set and also explain most of the property retention index with $R = 0.9689$ and 0.2043. Others contribute small with very closed correlation coefficients *R* and having almost the same standard errors *S*. In this case, can we say the first two

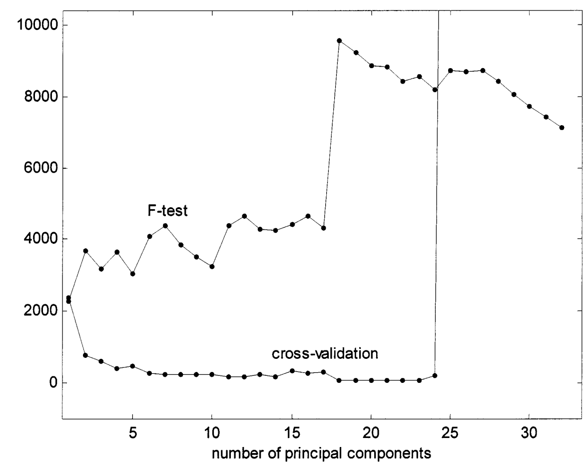

Figure 2. Relationships of cross-validation (leave-one-out) and *F* test vs the number of principal components.

principal components have explained "full" of the property? According to eq 1, the first two principal components can only reach 0.9902 correlation (i.e., $\sqrt{0.9689^2+0.2043^2}$) (for orthogonal variables, $R^2_{(x,y)} = \sum_{i=1}^{n} R^2_{(xi,y)}$, where $R_{(xi,y)}$ is the correlation coefficient of **y** against a variable $\mathbf{x}_i$, $R_{(x,y)}$ against all variables, say $\mathbf{x}_1, \mathbf{x}_2,...,\mathbf{x}_n$), while all 32 variables can reach 0.9997 correlation. So, additional principal components should be introduced into the regression model in order to improve the correlation furthermore. But the principal components of third to 32th (Table 6) show closed *R* and *S*, and the order of the values of *R* does not match the order of the principal components by their corresponding eigenvalues, e.g. the fourth $R >$ the third *R*, the sixth $R >$ the fifth *R*, and so forth. Thus, it is hard to select reasonably the number of principal components. If one selects the principal component according to the order of their corresponding eigenvalues (variances), one needs to select 22 principal components such obtained and the regression correlation coefficient will reach 0.9996 at this time. The cross-validation (leave-one-out) and *F*-test were also chosen to select number of components in this work. The results are shown in Figure 2. From this plot, the reasonable conclusion can hardly be made. From the results obtained by cross-validation, one can see that if one includes more than 25 principal components into regression model, the prediction will be unaccepted. However, there are several minima in the PRESS curve, where which one is the best point is really difficult to decide. The same situation is met for *F*-tests. Is it possible for one to include only a few orthogonal variables to get a better correlation model for this data set?

Selection of Orthogonal Block Variables for the Data Set 2. In the data set 2, there are eight block variables $\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{X}_3$, $\mathbf{X}_4$, $\mathbf{X}_5$, $\mathbf{X}_6$, $\mathbf{X}_7$, and $\mathbf{X}_8$. Their orthogonal matrixes Ω_1, Ω_2, Ω_3, Ω_4, Ω_5, Ω_6, Ω_7, and Ω_8 can be calculated following the calculation procedure (3) in the section of Theory and Methodology. The order of variables to orthogonalize impacts on results strongly. We use the greedy procedure "based on R_i"[26] to ascertain the order of the block variables, which is 3, 6, 8, 1, 5, 4, 7, 2 (first column in Table 7). Based on this order, we obtain the orthogonal block variables. The second column in Table 7 shows the correlation coefficients of the retention index against these orthogonal block variables. To improve the correlation, we investigate the combination of the orthogonal block variables by

Table 7. Correlation Coefficients of Regressions for Various Combinations of Eight Orthogonal Block Variables against Retention Index

block var.[a]	R_i	orth. var. comb.[a]	R	F-test	CCA var. comb.[a]	R	F-test
$\mathbf{X}_3$	0.99653	Ω_3	0.99653	5156.7	ω_3	0.99653	21056.6
$\mathbf{X}_6$	0.06468	$\Omega_3\Omega_6$	0.99863	6348.0	$\omega_3\omega_6$	0.99862	26480.2
$\mathbf{X}_8$	0.04153	$\Omega_3\Omega_6\Omega_8$	0.99949	9332.1	$\omega_3\omega_6\omega_8$	0.99949	47124.8
$\mathbf{X}_1$	0.01944	$\Omega_3\Omega_6\Omega_8\Omega_1$	0.99968	11161.4	$\omega_3\omega_6\omega_8\omega_1$	0.99968	55635.2
$\mathbf{X}_5$	0.00630	$\Omega_3\Omega_6\Omega_8\Omega_1\Omega_7$	0.99970	9429.3	$\omega_3\omega_6\omega_8\omega_1\omega_7$	0.99968	44214.0
$\mathbf{X}_4$	0.00576	$\Omega_3\Omega_6\Omega_8\Omega_1\Omega_7\Omega_5$	0.99972	8172.4	$\omega_3\omega_6\omega_8\omega_1\omega_7\omega_5$	0.99968	36587.4
$\mathbf{X}_7$	0.00685	$\Omega_3\Omega_6\Omega_8\Omega_1\Omega_7\Omega_5\Omega_4$	0.99974	8129.6	$\omega_3\omega_6\omega_8\omega_1\omega_7\omega_5\omega_4$	0.99970	33583.3
$\mathbf{X}_2$	0.00434	$\Omega_3\Omega_6\Omega_8\Omega_1\Omega_7\Omega_5\Omega_4\Omega_2$	0.99975	7131.7	$\omega_3\omega_6\omega_8\omega_1\omega_7\omega_5\omega_4\omega_2$	0.99970	29209.2

[a] Term of block var., orth. var. comb. and CCA var. comb. mean block variables, orthogonal block variable combination, and canonical correlation variable combination.

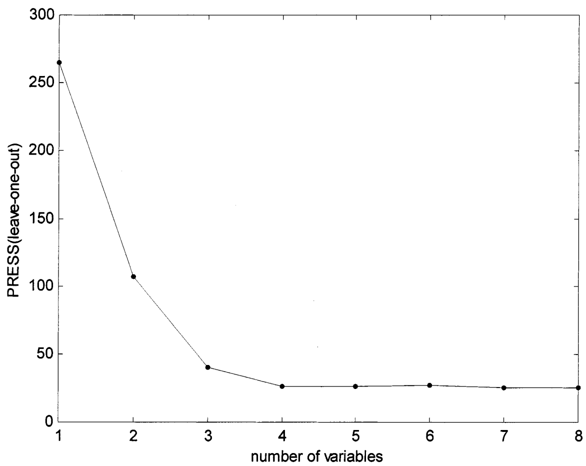

Figure 3. Relationships of cross validation (leave-one-out) vs the number of eight canonical orthogonal components with the same order as listed in Table 7.

selecting those with maximum values of R_i. The third and fourth columns in Table 7 show the combinations and correlation coefficients R. From this table, one can see that an increase of R slows down when selecting more than four variables. So the four block variables Ω_3, Ω_6, Ω_8, and Ω_1 might be used to establish the regression model, which can explain the retention index with $R = 0.99968$. To reduce the number of the variables, we calculate further the canonical correlation variables for the eight orthogonal block variables. With the help of canonical correlation analysis, eight new orthogonal variables (vectors), say ω_1, ω_2, ..., ω_8, are obtained. The combinations of these eight new variables and their contributions to explain retention index with R are listed in the fifth and sixth columns in Table 7. From the results listed in Table 7, it is clearly seen that the eight new orthogonal variables (vectors), say ω_1, ω_2, ..., ω_8, can provide almost the same information as eight orthogonal block variables do. To check the prediction ability of the correlation model with the eight new orthogonal variables, cross-validation of leave-one-out and F-test are also made, which are shown in Figure 3 and Table 7, respectively. These two results are quite agreement with each other. Both cross-validation (minimum point at 4, see Figure 3) and F-test value (with maximum value of 55635.2) seem to suggest that the correlation model of four orthogonal variables, say ω_1, ω_3 ω_6, and ω_8, will show the best behavior both in fitting and prediction.

Intercorrelation between the Different Descriptor Variables. The reason selection of variables will be so difficult, in our opinion, lies in that many descriptors are similar and/or represent "duplicated" information of the molecular structure. Thus, how to reasonably evaluate the intercorrelation between the different descriptor variables or even between the different block variables of topological indices will become very helpful for QSAR/QSPR studies. With the help of the subspace-projection technique proposed in this work, this task seems to be possible to fulfill. Table 8 gives correlation coefficients of retention index against block variables (variable matrix) and their corresponding orthogonal block variables. The numbers in the first row and the first column in table denote the number of the block variables. Values in the position of row i and column j in Table 8 are correlation coefficients of the regression between the retention index and the block variable obtained after subspace-projection orthogonalizing the jth block variable to the ith block variable. While, the diagonal values give the correlation coefficients of regression of the retention index against the ith or jth block variable ($i=j$) without orthogonalization.

Table 8. Correlation Coefficients of Retention Index against Block Variables and Their Corresponding Orthogonal Block Variables

	1	2	3	4	5	6	7	8
1	0.9952	0.0794	0.0764	0.0806	0.0689	0.0738	0.0121	0.0633
2	0.0877	0.9945	0.0892	0.0846	0.0810	0.0408	0.0436	0.0621
3	0.0560	0.0622	0.9965	0.0412	0.0413	0.0647	0.0315	0.0380
4	0.0874	0.0831	0.0744	0.9946	0.0808	0.0895	0.0413	0.0591
5	0.0674	0.0705	0.0648	0.0719	0.9953	0.0808	0.0178	0.0540
6	0.1179	0.0934	0.1238	0.1238	0.1233	0.9909	0.0413	0.0925
7	0.2733	0.2739	0.2797	0.2740	0.2740	0.2604	0.9570	0.2675
8	0.1170	0.1102	0.1176	0.1097	0.1131	0.0989	0.0820	0.9903

From the table, we can see that regression of retention index against every individual block variable gives the following values of correlation coefficients R (see the diagonals in Table 8): 0.9952, 0.9945, 0.9965, 0.9946, 0.9953, 0.9909, 0.9570, 0.9903, respectively. All these values except the seventh one are more than 0.99, which explain most of the property retention index. For instance, if we select the third block variable (path variables, say p_1, p_2, ..., p_4) to establish the regression model against the retention index (see the row of $i = 3$ in Table 8) with $R = 0.9965$, additional descriptors should be introduced in order to improve further the correlation. How much can the candidates of the other seven block variables improve the regression? Here we consider orthogonalized block variables to the third one, to avoid the duplication between the orthogonal block variables and the third block variable. The contributions of the orthogonal block variables give the values for $R = 0.0560$ (χ variables, say ${}^1\chi$, ${}^2\chi_p$, ${}^3\chi_p$, and ${}^3\chi_c$), 0.0622 (Kappa variables, say ${}^0\kappa$, ${}^1\kappa$, ${}^2\kappa$, and ${}^3\kappa$), 0.0412 (walk account variables, say w_1, w_2, ..., w_4), 0.0413 (path/walk account variables, say pw_1, pw_2, ,.., pw_4), 0.0647 (the indices proposed and used by Schultz, say MTI, PED, PEAD, DET), 0.0315 (the indices proposed by Xu and co-workers, say Yx and EAID), and 0.0380 (quantum chemical descriptors, say heat form., elec. energy, core energy, ion. energy, dipole, HUMO), respectively. They are much smaller than that of the third block variable. This fact illustrates that the third one explains most of the retention index, while other orthogonal block variables after subspace-projection by it explain only a little. It is worth noting that the orthogonal block variables after subspace-projection by the third block variable (path variables, say p_1, p_2, ..., p_4) are essentially the remaining portions after removing the "overlapping part" with the path variables. This indicates that the overlapping between the block variables is very serious. The information for explaining the retention index from different block variables is almost duplicated. From the discussion above, one can see that the duplicated parts of all the seven other block variables with the third block variable for explaining retention of alkanes investigated in this work seem to be the following, that is, (0.9952−0.0560) = 0.9292; (0.9945−0.0622) = 0.9323; (0.9946−0.0412) = 0.9534; (0.9953−0.0413) = 0.9540; (0.9909−0.0647) = 0.9252; (0.9570−0.0315) = 0.9255; (0.9903−0.0380) = 0.9523, respectively. These numbers suggest that the overlapping between the path account variables and the other seven block variables can be roughly estimated by them. They may be further classified into two groups, in which one embraces walk account variables (0.9534), path/walk account variables (0.9540), and quantum chemical descriptors (0.9523) and the other may include χ variables (0.9292), Kappa variables (0.9323), the indices proposed and used by Schultz (0.9252), and the indices proposed by Xu and co-workers (0.9255). However, all these eight block variables are quite similar in providing molecular description for alkanes. How to evaluate reasonably and quantitatively the similarity between the block variables is just conducted in our laboratory.

ACKNOWLEDGMENT

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

REFERENCES AND NOTES

(1) Wiener, H. Structural determination of paraffin boiling points. *J. Am. Chem. Soc.* **1947**, *69*, 17−20.
(2) Hosoya, H. Topological index. A newly proposed quantity characterizing topological nature of structural isomers of saturated hydrocarbons. *Bull. Chem. Soc. Jpn.* **1971**, *44*, 2332−2339.
(3) Balaban, A. T.; Ivanciuc, O. Historical development of topological indices. *Topological indices and related descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach Science Publishers: The Netherlands, 1999; pp 21−57.
(4) Motoc, I.; Balaban A. T. Toplogical indices: Intercorrelations, physical meaning, correlational indices ability. *Rev. Roum. Chim.* **1981**, *26*, 593−600.
(5) Motoc, I.; Balaban, A. T.; Mekenyan, O.; Bonchev, D. Topological indices: inter-relations and composition. *MATCH (Commun. Math. Chem.)* **1982**, *13*, 369−404.
(6) Basak, S. C.; Niemi, G. J.; Regal, R. R.; Veith, G. D. Topological indices: Their nature relatedness, and applications. *Math Modelling Sci. Technol.* **1987**, *8*, 300−305.
(7) Morales, D. A.; Araujo, O. On the search for best correlation between graph theoretical invariants and physicochemical properties. *J. Math. Chem.* **1993**, *13*, 95−106.
(8) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. CODESSA Version 2.0 Reference Manual; University of Florida: Gainesville, FL, 1994.
(9) Draper, N. R.; Smith, H. *Applied Regression Analysis;* John Wiley and Sons: New York, 1981.
(10) Furnival, G. M.; Wilson, R. W. Regressions by leaps and bounds. *Technometrics* **1974**, *16*, 499−511.
(11) Goldberg, D. E. *Genetic Algorithms in Search, Optimization & Machine Learning*; Addison-Wesley: New York, 1989.
(12) Hibbert, D. B. Genetic algorithm in chemistry. *Chemom. Intell. Lab. Syst.* **1993**, *19*, 277−93.
(13) Hasegawa, K.; Funatsu, K. GA strategy for variable selection in QSAR studies: GAPLS and D-optimal designs for predictive QSAR model. *J. Mol. Struct. (THEOCHEM)* **1998**, *425*, 255−262.
(14) Wikel, J. H.; Dow, E. R. The use of neural networks for variable selection in QSAR. *Bioorg. Med. Chem. Lett.* **1993**, *3*, 645−51.
(15) Sutter, J. M.; Dixon, S. L.; Jurs, P. C. Automated descriptor selection for quantitative-structure−activity relationships using generalized simulated annealing. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 77−84.
(16) Lucic, B.; Trinajstic, N. A. New Efficient Approach for Variable Selection Based on Multiregression: Prediction of Gas Chromatographic Retention Times and Response Factors. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 610−621.
(17) Taraviras, S. L.; Ivanciuc, O.; Bass, D. C. Identification of Groupings of Graph Theoretical Molecular Descriptors Using a Hybrid Cluster Analysis Approach. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1128−1146.
(18) Ivanciuc, O.; Taraviras, S. L.; Cabrol-Bass, D.Quasi-orthogonal Basis Sets of Molecular Graph Descriptors as a Chemical Diversity Measure. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 126−134.
(19) Randic, M. On characterization of molecular branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609−6615.
(20) Balaban, A. T. Chemical graphs. XXXIV. Five new topological indexes for the branching of tree-like graphs. *Theor. Chim. Acta* **1979**, *53*, 355−75.
(21) Randic, M. Resolution of ambiguities in structure−property studies by use of orthogonal descriptors. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 311−320.
(22) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.
(23) Amic, D.; Davidovic-Amic, D.; Trinajstic, N. Calculation of retention times of anthocyanins with orthogonalized topological indices *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 136−139.
(24) Lucic, B.; Nikolic, S.; Trinajstic, N.; Juretic, D. The structure−property models can be improved using the orthogonalized descriptors *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 532−538.
(25) Soskic, M.; Plavsic, D.; Trinajstic, N. 2-Difluoromethylthio-4,6-bis-(monoalkylamino) -1,3,5-triazines as inhibitors of Hill reaction: A QSAR study with orthogonal descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 829−832.
(26) Xu, L.; Zhang, W. J. Comparison of different methods for variable selection *Anal. Chim. Acta* **2001**, *446*, 477−483.
(27) 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.
(28) Mardia, K. V.; Kent, J.; Bibby, J. *Multivariate Analysis;* Academic Press: New York: 1979.
(29) 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, Huangshan, 2001; pp 147−149.
(30) Randic, M.; Zupan, J. On Interpretation of well-known topological indices *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 550−560.
(31) Schultz, H. P. Topological organic chemistry. 1. Graph theory and topological indices of alkanes. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 227−228.
(32) 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.
(33) 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.
(34) Hu, C. Y.; Xu, L. On highly discriminating molecular topological index. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 82−90.

CI020283+