List Operations on Chemical Graphs. 5. Implementation of Breadth-First Molecular Path Generation and Application in the Estimation of Retention Index Data and Boiling Points†

R. Gautzsch and P. Zinn*

Lehrstuhl für Analytische Chemie, Ruhr-Universität Bochum, Universitätsstrasse 150, D-44780 Bochum, Germany

Received September 29, 1993®

Breadth-first strategy is described as a basic concept of molecular path generation. Minor modifications on the algorithm also make depth-first strategy available. Breadth-first strategy is advantageous in topological distance type problems. It is applied to generate path counts as topological descriptors. Complete path count sets are compared with atom type, bond type and sphere count descriptor sets by principal components analysis. The tendency of degeneration and the orthogonality of the descriptor sets are investigated. Pure descriptor set models and mixed models are developed to predict boiling points and retention index data of alkanes. It is shown that a mixed model incorporating path count and bond type descriptors covers nearly the whole topological information of the investigated compounds. The estimation error of this model lies near an approximated pure experimental error.

INTRODUCTION

In previous papers[1,2] we have introduced list operations as a tool for deriving information about chemical graphs. In this approach the central data structure is the adjacency list, comparable with the adjacency matrix in the conventional treatment of graph-theoretical problems in chemistry. Most basic list operations on chemical graphs are referring to adjacency lists.

In order to implement list operations, queuing of information is a powerful programming paradigm. It is a basic concept in information sciences[3,4] and applicable to many problems of chemical information handling. Because this technique is of great interest and not restricted to to special programming languages or a special representation of adjacency information, we discuss queuing of information for the implementation of depth-first and breadth-first path generation strategies. Breadth-first strategy is preferable when generating paths and spheres[1,2] of chemical graphs, and the resulting procedure will be described.

These procedures are well suited for application in quantitative structure–property relationship (QSPR) studies. Thereby the structural information is coded in path count and sphere count tables derived from path generation. These tables are compared with formerly described atom-type and bond-type tables[5,6] by principal components analysis. The applicability of the new tables to the estimation of gas chromatographic retention indexes and boiling points of some alkanes will be proved. Alkanes were chosen for analyzing the new descriptor tables, because steric or electronic effects caused by heteroatoms or chemical bonds of higher order in more complex compounds[5,6] often disturb pure topological investigations. In order to examine the estimation models, retention index prediction is compared with boiling point prediction and the lack of fit of the resulting models is discussed.

IMPLEMENTATION OF BREADTH-FIRST STRATEGY FOR MOLECULAR PATH GENERATION

Depth- and breadth-first strategies are basic concepts in different fields of information sciences, e.g. graph theory,[4,7] rule-based expert systems,[8] or theorem proving.[9] When list operations on chemical graphs[1,2] are performed, depth- and breadth-first strategies are useful in order to generate molecular paths and path-based topological indexes. These strategies will be described by referring to adjacency lists[1] as representations of molecular graphs because the adjacency list is the most important data structure with respect to list operations. Nevertheless, minor modifications of the program code allow application to adjacency matrices or connection tables. The resulting implementation of depth- and breadth-first strategies is shown in Figure 1. In order to show how the implementation works, the graph of 1,2-dimethyl-1-ethylcyclopropane of Figure 2 serves as an example.

```
(defun breadth-first-paths (id-number bond-list) ; frame procedure for
  (breadth-first-paths1 nil                      ; breadth-first
                        (list (list id-number))  ; path generation,
                        bond-list))



(defun breadth-first-paths1 (paths      ; recursive implementation of
                             queue      ; breadth-first
                             bond-list) ; path generation,
  (cond ((null queue) (mapcar 'reverse          ; if queue is empty show
                              paths))           ; reverted path list,
        (t (breadth-first-paths1                ; else call recursively,
            (append paths                       ; append first element of
                    (list (car queue)))         ; queue to the path list,
            (append (cdr queue)                 ; append expanded path to
                    (expand-path (car queue)    ; the rest of the queue,
                                 bond-list))
;           (append (expand-path (car queue)    ; modification for
;                                bond-list)     ; depth-first
;                   (cdr queue))                ; strategy,
            bond-list))))



(defun expand-path (path bond-list)             ; path expansion procedure,
  (remove-if                                    ; remove atom if
    #'(lambda (path) (member (car path)         ; it is already member of
                             (cdr path)))       ; path,
    (mapcar #'(lambda (neighbour)               ; construct expanded paths
                      (cons neighbour path))    ; by putting neighbour atoms
            (get-neighbours (car path)          ; in front of path,
                            bond-list
                            nil))))
```

Figure 1. Recursive Implementation of breadth-first and modifications for depth-first molecular path generation strategies in LISP. The queue starts with the ID number of a specified atom and is developed by referring to the molecule's bond list (adjacency list). A list of all molecular paths starting from the specified atom will be generated.

Breadth-first path generation starts from a specified atom of the interesting molecule. In the example this is atom 2, which is the first entry in the queue. The other input parameter of the procedure is the adjacency or bond list of the molecule. As output, a list of all paths starting with the specified atom is generated. At the beginning this list is empty and is filled

† Dedicated to Prof. Dr. Ewald Jackwerth.

® Abstract published in *Advance ACS Abstracts*, May 15, 1994.

0095-2338/94/1634-0791$04.50/0

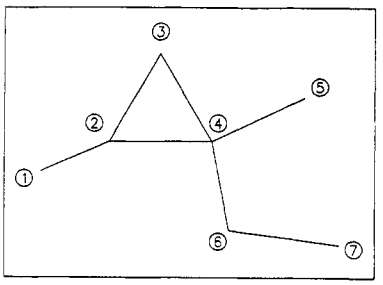

Figure 2. Graph and corresponding ID numbers for the example molecule 1,2-dimethyl-1-ethylcyclopropane.

Breadth-first

PATHS	QUEUE
NIL	((2))
((2))	((1 2) (3 2) (4 2))
((2) (1 2))	((3 2) (4 2))
((2) (1 2) (3 2))	((4 2) (4 3 2))
((2) (1 2) (3 2) (4 2))	((4 3 2) (3 4 2) (5 4 2) (6 4 2))
((2) (1 2) (3 2) (4 2) (4 3 2))	((3 4 2) (5 4 2) (6 4 2) (5 4 3 2) (6 4 3 2))
((2) (1 2) (3 2) (4 2) (4 3 2) (3 4 2))	((5 4 2) (6 4 2) (5 4 3 2) (6 4 3 2))
((2) (1 2) (3 2) (4 2) (4 3 2) (3 4 2) (5 4 2))	((6 4 2) (5 4 3 2) (6 4 3 2))
((2) (1 2) (3 2) (4 2) (4 3 2) (3 4 2) (5 4 2) (6 4 2))	((5 4 3 2) (6 4 3 2) (7 6 4 2))
((2) (1 2) (3 2) (4 2) (4 3 2) (3 4 2) (5 4 2) (6 4 2) (5 4 3 2))	((6 4 3 2) (7 6 4 2))
((2) (1 2) (3 2) (4 2) (4 3 2) (3 4 2) (5 4 2) (6 4 2) (5 4 3 2) (6 4 3 2))	((7 6 4 2) (7 6 4 3 2))
((2) (1 2) (3 2) (4 2) (4 3 2) (3 4 2) (5 4 2) (6 4 2) (5 4 3 2) (6 4 3 2) (7 6 4 2))	((7 6 4 3 2))
((2) (1 2) (3 2) (4 2) (4 3 2) (3 4 2) (5 4 2) (6 4 2) (5 4 3 2) (6 4 3 2) (7 6 4 2) (7 6 4 3 2))	NIL

reverse:
((2) (2 1) (2 3) (2 4) (2 3 4) (2 4 3) (2 4 5) (2 4 6) (2 3 4 5) (2 3 4 6) (2 4 6 7) (2 3 4 6 7))

Figure 3. Development of the queue and path list of the *breadth-first* path building strategy applied to atom 2 of the graph in Figure 2.

up with paths copied from the queue while the procedure is active. The path building is a recursive procedure and incorporates two steps. In the first step the queue is tested and recursion stops if the queue is empty. In the second step the procedure is called again with two changes: The list of paths is appended by the first element of the queue, and the rest of the queue is appended by the result of an expansion of the queue's first element. The expansion of this path actualizes the queue. Therefore the neighbor atoms of the path's first atom were put in front of it to form new expanded paths. To prevent the procedure from going around in the case of cyclic graphs, those neighbors that are already members of the path are removed. The described procedure is illustrated in Figure 3, where the development of the queue and path list is shown. By every call of the procedure, the path list is appended by one element. This element is removed from the queue, and the queue is appended by the expansion of this element. At the exit of the procedure all paths are reverted in order to put the starting atom at the beginning of the paths.

The only difference between breadth- and depth-first path generation is the reverse order of building the actual queue. In depth-first path generation the expanded queue element is put in front of the queue instead of appending it at the end, as can be seen in Figure 1. The resulting different development of the queue and the path list is shown in Figure 4.

Both concepts allow one to derive a series of path-building-based procedures, e.g. finding the smallest paths between two atoms of a molecule, finding all paths between two atoms, generating all paths of a molecule, building path counts, or calculating topological indexes based on path counts or topological distances. In particular breadth-first strategy is very useful in distance-type problems and leads to faster computation time in comparison with depth-first strategies. As typical distance-type problems, searching the smallest paths between two atoms and counting of molecular paths of specified length are to be mentioned. Especially, when counting paths of large molecules, breadth-first is often the only possible strategy, because it allows one to stop counting at a defined

Table 1. Retention Indexes[20] (*I*) and Normal Boiling Points[21] (BP) of Alkanes Fitted with the Models of Table 6

compd no.	name	*I*	BP	compd no.	name	*I*	BP
1	2,2-dimethylpropane	412.3	9.5	36	3,3-dimethylheptane	835.8	137.3
2	2-methylbutane	475.3	27.8	37	2,4-dimethyl-3-ethylpentane	836.5	136.7
3	2,2-dimethylbutane	536.8	49.7	38	2,3,4-trimethylhexane	849.1	139.0
4	2,3-dimethylbutane	567.3	58.0	39	2,3,3,4-tetramethylpentane	858.0	141.5
5	2-methylpentane	569.7	60.3	40	3-methyloctane	870.2	143.3
6	3-methylpentane	584.2	63.3	41	3,3-diethylpentane	877.2	146.2
7	2,2-dimethylpentane	625.6	79.2	42	*n*-butane	400	−0.5
8	2,4-dimethylpentane	629.8	80.5	43	*n*-pentane	500	36.1
9	2,2,3-trimethylbutane	639.7	80.9	44	*n*-hexane	600	69.0
10	3,3-dimethylpentane	658.9	86.1	45	*n*-heptane	700	98.4
11	2-methylhexane	666.6	90.0	46	*n*-octane	800	125.7
12	2,3-dimethylpentane	671.7	89.8	47	*n*-nonane	900	150.8
13	3-ethylpentane	686.0	93.5	48	ethylcyclopropane	510.2	35.9
14	2,2,4-trimethylpentane	689.9	99.2	49	cyclopentane	565.7	49.3
15	2,2-dimethylhexane	719.4	106.8	50	ethylcyclobutane	621.1	70.7
16	2,5-dimethylhexane	728.4	109.0	51	methylcyclopentane	627.9	71.8
17	2,4-dimethylhexane	731.9	109.4	52	cyclohexane	662.7	80.7
18	2,2,3-trimethylpentane	737.1	110.0	53	1,1,2-trimethylcyclopentane	763.2	113.7
19	3,3-dimethylhexane	743.5	112.0	54	1,1,3-trimethylcyclopentane	723.6	105.0
20	2,3,4-trimethylpentane	752.4	113.4	55	methylcyclohexane	725.8	100.9
21	2,3,3-trimethylpentane	759.4	114.7	56	ethylcyclopentane	733.8	103.5
22	2-methylheptane	764.9	117.6	57	1,1-dimethylcyclohexane	787.0	119.5
23	4-methylheptane	767.2	117.7	58	isopropylcyclopentane	812.1	126.4
24	3,4-dimethylhexane	770.6	117.7	59	*n*-propylcyclopentane	830.3	131.0
25	3-methylheptane	772.3	118.0	60	ethylcyclohexane	834.3	131.8
26	2,2,4,4-tetramethylpentane	772.7	122.7	61	1,1,3-trimethylcyclohexane	840.4	136.6
27	3-methyl-3-ethylpentane	774.0	118.2	62	*cis*-1,3-dimethylcyclopentane	682.7	90.8
28	2,2,5-trimethylhexane	776.3	124.0	63	*trans*-1,3-dimethylcyclopentane	686.8	91.7
29	2,2,4-trimethylhexane	789.1	126.5	64	*cis*-1,2-dimethylcyclopentane	720.9	99.6
30	2,4,4-trimethylhexane	807.7	126.5	65	*trans*-1,2-dimethylcyclopentane	689.2	91.9
31	2,3,5-trimethylhexane	812.0	131.3	66	*trans*-1,2-dimethylcyclohexane	801.8	123.4
32	2,2-dimethylheptane	815.4	132.7	67	*cis*-1,2-dimethylcyclohexane	829.3	129.7
33	2,2,3,4-tetramethylpentane	819.6	133.0	68	*trans*-1,3-dimethylcyclohexane	805.6	124.5
34	2,2,3-trimethylhexane	821.6	131.7	69	*cis*-1,3-dimethylcyclohexane	785.0	120.1
35	2,2-dimethyl-3-ethylpentane	822.2	133.8				

Depth first

PATHS	QUEUE
NIL	((2))
((2))	((1 2) (3 2) (4 2))
((2) (1 2))	((3 2) (4 2))
((2) (1 2) (3 2))	((4 3 2) (4 2))
((2) (1 2) (3 2) (4 3 2))	((5 4 3 2) (6 4 3 2) (4 2))
((2) (1 2) (3 2) (4 3 2) (5 4 3 2))	((6 4 3 2) (4 2))
((2) (1 2) (3 2) (4 3 2) (5 4 3 2) (6 4 3 2))	((7 6 4 3 2) (4 2))
((2) (1 2) (3 2) (4 3 2) (5 4 3 2) (6 4 3 2) (7 6 4 3 2))	((4 2))
((2) (1 2) (3 2) (4 3 2) (5 4 3 2) (6 4 3 2) (7 6 4 3 2) (4 2))	((3 4 2) (5 4 2) (6 4 2))
((2) (1 2) (3 2) (4 3 2) (5 4 3 2) (6 4 3 2) (7 6 4 3 2) (4 2) (3 4 2))	((5 4 2) (6 4 2))
((2) (1 2) (3 2) (4 3 2) (5 4 3 2) (6 4 3 2) (7 6 4 3 2) (4 2) (3 4 2) (5 4 2))	((6 4 2))
((2) (1 2) (3 2) (4 3 2) (5 4 3 2) (6 4 3 2) (7 6 4 3 2) (4 2) (3 4 2) (5 4 2) (6 4 2))	((7 6 4 2))
((2) (1 2) (3 2) (4 3 2) (5 4 3 2) (6 4 3 2) (7 6 4 3 2) (4 2) (3 4 2) (5 4 2) (6 4 2) (7 6 4 2))	NIL

reverse:
((2) (2 1) (2 3) (2 3 4) (2 3 4 5) (2 3 4 6) (2 3 4 6 7) (2 4) (2 4 3) (2 4 5) (2 4 6) (2 4 6 7))

Figure 4. Development of the queue and path list of the *depth-first* path building strategy applied to atom 2 of the graph in Figure 2.

path length. On the other hand, depth-first strategy may first find the largest paths and may exceed memory limits. The application of the breadth-first procedure in QSPR studies is discussed in the following.

ORTHOGONALITY AND DEGENERACY OF PATH AND SPHERE COUNTS AS TOPOLOGICAL MOLECULAR DESCRIPTORS

In QSPR studies a wide palette of molecular descriptors has been investigated. Among those a basic concept is the topological description of a molecule. Topological description of a molecule is often presented in the form of topological indexes.[10-14] The general idea of these indexes is to concentrate topological information of a molecule in one or a few parameters. The quality of these indexes with respect to the estimation of a particular molecular property can be determined by looking at the residuals of the estimated property. Using one index for the estimation at a time comparison of the residuals allows a decision for the best index. To decide, about the quality of indexes and the related estimation model is more complicated if two or more indexes are used in the model. The general question in these cases is which combination of topological models gives the best estimation results.

It is obvious that if a second index is taken in a model, this index should be orthogonal to the first one. Randic discussed some ideas in this context.[15] In such a model the first index should explain a specific part of the property's variability and the second another part, independent from the first one. The dilemma when using convenient topological indexes is that they are by far not orthogonal and correlations among them are more or less strong.[16] This is understandable because the aim of the construction of a topological index often is the derivation of a parameter that explains the greatest part of the variability of data to establish precisely working estimation models.

To derive topological descriptors with high orthogonality, we have introduced incremental atom- and bond-type models for gas chromatographic retention index prediction.[5,6] The disadvantage of these models is that they are less or more degenerative.[17] That means that two or more topologically different molecules may have the same set of descriptors. For example, if we have a look at ethylcyclopropane and methylcyclobutane, we recognize the same atom types: 1 CH_3; 3 CH_2; 1 CH. The consequence is that prediction of any property of these molecules gives the same results using atom-type models. To overcome this problem, one can refine the estimation model using bond type descriptors. The bond type model has different sets of parameters for both compounds. Ethylcyclopropane has the bond type configuration

$$1\ CH_3\text{–}CH_2;\ \ 1\ CH_2\text{–}CH_2;\ \ 3\ CH_2\text{-}CH$$

and methylcyclobutane

$$1\ CH_3\text{–}CH;\ \ 2\ CH_2\text{–}CH_2;\ \ 2\ CH_2\text{-}CH$$

The degree of degeneracy in the bond-type model is generally much smaller than that in the atom-type model. Nevertheless, degeneracy is not negligable in the bond-type model. As an example, the bond-type sets of 1,3-dimethylcyclohexane and 1,4-dimethylcyclohexane are identical:

$$2\ CH_3\text{–}CH;\ \ 2\ CH_2\text{–}CH_2;\ \ 4\ CH_2\text{-}CH$$

A further decrease in degeneracy can be reached by using path counts as topological descriptors. Randic applies path count bond models for the prediction of C13-NMR shifts.[18] Path counts give the number of how often paths of a particular length are to be found in a molecule. They are calculated by simple counting of molecular paths. A way for the generation of paths is described above in detail.

The low character of degeneracy of the complete path count sets of molecules can be demonstrated by looking at the already mentioned example of the two cyclohexanes. The sets of path counts PL are different for 1,3-dimethylcyclohexane

$$P1 = 8;\ \ P2 = 8;\ \ P3 = 10;\ \ P4 = 10;$$

$$P5 = 11;\ \ P6 = 10;\ \ P7 = 5$$

and for 1,4-dimethylcyclohexane

$$P1 = 8;\ \ P2 = 8;\ \ P3 = 10;\ \ P4 = 10;$$

$$P5 = 10;\ \ P6 = 12;\ \ P7 = 4$$

with L equal to the length of paths.

To use path counts as molecular descriptors in estimation models, one has to ask about the orthogonality of path count sets. Therefore principal component analyses (PCA) of the compounds of Table 1 were performed. In order to compare the results for path counts, atom- and bond-type sets of the same compounds were analyzed too. As additional molecular descriptors, sphere count sets were treated in the same manner. Molecular spheres are related to molecular paths.[1,19] A molecular sphere contains all atoms with a particular topological distance to a central atom. In the case of acyclic molecules, path and sphere counts are identical, whereas path counts of cyclic molecules are always greater than sphere counts. Tables 2–5 summarize the investigated descriptor sets.

PCA analysis was applied to the four feature matrices **D**, including the data of the Tables 2–5. The columns of the matrices are the different types of the four investigated descriptor sets. Each row is associated with one compound. To explain the way in which PCA calculations were performed, the atom-type set of Table 2 serves as an example.

At first, each column was scaled by its mean value to generate the scaled feature matrix $\mathbf{D}^S$. Then $\mathbf{D}^S$ was multiplied by its transpose $(\mathbf{D}^S)^T$ under consideration of the degrees of freedom to form the covariance matrix

$$\mathbf{COV} = (1/(n-1))(\mathbf{D}^S)^T \cdot \mathbf{D}^S$$

example of the covariance matrix for the

atom types of Table 2

$$\mathbf{COV} = \begin{pmatrix} 2.261 & -2.012 & 0.263 & 0.487 \\ -2.012 & 2.949 & -0.563 & -0.310 \\ 0.263 & -0.563 & 0.774 & -0.191 \\ 0.487 & -0.310 & -0.191 & 0.271 \end{pmatrix}$$

followed by the calculation of the correlation matrix with the

Table 2. Atom-Type Descriptor Sets Corresponding to the Compounds of Table 1

	atom types					atom types			
compd no.	CH_3	CH_2	CH	C	compd no.	CH_3	CH_2	CH	C
1	4	0	0	1	36	4	4	0	1
2	3	1	1	0	37	5	1	3	0
3	4	1	0	1	38	5	1	3	0
4	4	0	2	0	39	6	0	2	1
5	3	2	1	0	40	3	5	1	0
6	3	2	1	0	41	4	4	0	1
7	4	2	0	1	42	2	2	0	0
8	4	1	2	0	43	2	3	0	0
9	5	0	1	1	44	2	4	0	0
10	4	2	0	1	45	2	5	0	0
11	3	3	1	0	46	2	6	0	0
12	4	1	2	0	47	2	7	0	0
13	3	3	1	0	48	1	3	1	0
14	5	1	1	1	49	0	5	0	0
15	4	3	0	1	50	1	4	1	0
16	4	2	2	0	51	1	4	1	0
17	4	2	2	0	52	0	6	0	0
18	5	1	1	1	53	3	3	1	1
19	4	3	0	1	54	3	3	1	1
20	5	0	3	0	55	1	5	1	0
21	5	1	1	1	56	1	5	1	0
22	3	4	1	0	57	2	5	0	1
23	3	4	1	0	58	2	4	2	0
24	4	2	2	0	59	1	6	1	0
25	3	4	1	0	60	1	6	1	0
26	6	1	0	2	61	3	4	1	1
27	4	3	0	1	62	2	3	2	0
28	5	2	1	1	63	2	3	2	0
29	5	2	1	1	64	2	3	2	0
30	5	2	1	1	65	2	3	2	0
31	5	1	3	0	66	2	4	2	0
32	4	4	0	1	67	2	4	2	0
33	6	0	2	1	68	2	4	2	0
34	5	2	1	1	69	2	4	2	0
35	5	2	1	1					

elements $cor_{ij} = cov_{ij}/(cov_{ii}cov_{jj})^{1/2}$.

example of the correlation matrix for the atom types of Table 2

$$\mathbf{COR} = \begin{pmatrix} 1 & -0.7792 & 0.1991 & 0.6223 \\ -0.7792 & 1 & -0.3725 & -0.3468 \\ 0.1991 & -0.3725 & 1 & -0.4159 \\ 0.6223 & -0.3468 & -0.4159 & 1 \end{pmatrix}$$

Classical eigenvector analysis is applied on the correlation matrix in order to compute the eigenvectors $\mathbf{E}_1$–$\mathbf{E}_4$.

example of eigenvectors of the correlation matrix of the atom types of Table 2

$$\mathbf{E}_1 = \begin{pmatrix} 0.6463 \\ -0.5937 \\ 0.1317 \\ 0.4610 \end{pmatrix}; \quad \mathbf{E}_2 = \begin{pmatrix} 0.0087 \\ -0.2521 \\ 0.7878 \\ -0.5619 \end{pmatrix};$$

$$\mathbf{E}_3 = \begin{pmatrix} 0.2136 \\ 0.6916 \\ 0.5322 \\ 0.4392 \end{pmatrix}; \quad \mathbf{E}_4 = \begin{pmatrix} 0.7325 \\ 0.3251 \\ -0.2807 \\ -0.5281 \end{pmatrix}$$

To illustrate the results of the eigenvector analysis, two different types of figures were used. The first one shows the plot of the elements (component weights) of the eigenvectors $\mathbf{E}_1$ and $\mathbf{E}_2$ in Figure 5a. Each of these data pairs corresponds with one of the descriptors A_1–A_4 of Table 2 and allows a visual exploration of the descriptor's orthogonality.

In the second type of figure the autoscaled feature matrix is projected into the coordinate system of eigenvectors $\mathbf{E}_1$ and $\mathbf{E}_2$. Each entry in Figure 6a corresponds to one compound of the feature matrix. In cases of degeneration one entry in Figure 6a may represent two or more compounds.

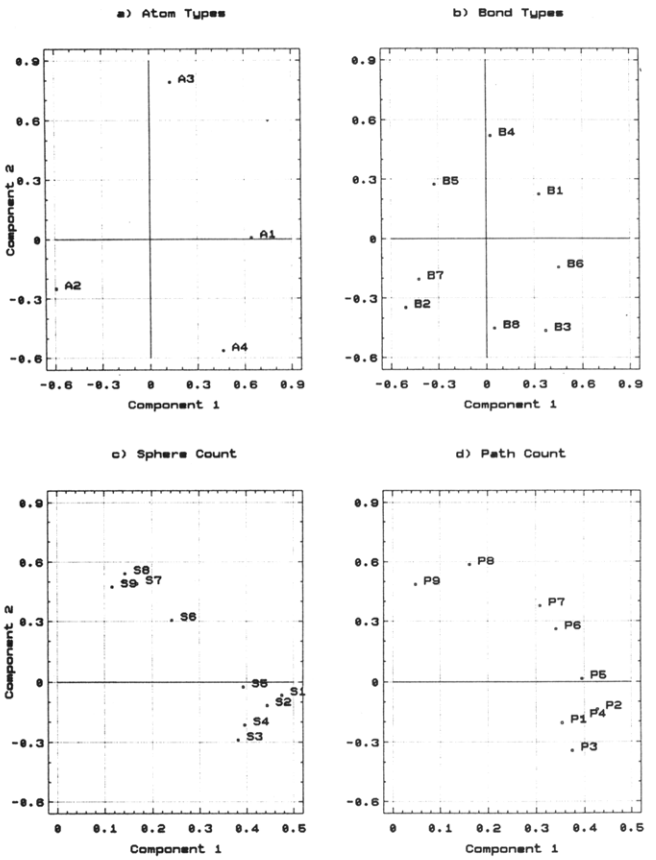

Figure 5. Plots of the first two components of the principle component analysis. The entries in the plots are determined by the first two eigenvector elements of the corresponding descriptor sets.

The results of PCA of all four descriptor sets are illustrated in Figures 5 and 6. In the cases of atom and bond types, the arrangements of entries are rather homogeneously distributed around the origin. Each descriptor of both sets should contribute a specific amount to a corresponding estimation model. On the other hand, the entries of sphere and path count plots in Figure 5c,d show a tendency of cluster building. For example, the sphere counts S1 to S5 and the path counts P1 to P4 are lying relatively close together. The contributions of these descriptors to a property in an estimation model would be very similar. These descriptors correlate strongly with one another, and as a consequence they could be substituted by comprehensive descriptors S15 and P14. The orthogonality of sphere and path count sets is not very strongly pronounced.

Figure 6 shows the PCA scatter plots. In these plots the first and second principal components of each compound determine the entries. If there would be no degeneracy in the descriptor sets, every compound should have its own specific entry. Particularly, the atom type plot in Figure 6a has a strongly decreased number of entries (37) in comparison with the number of compounds in the investigated data set (69). Several compounds have the same sets of atom types because of topological similarity and stereochemical isomerism. Four pairs of cis–trans isomers in the data set have of course the same sets of topological descriptors.

The scatter plots of sphere and path counts in Figure 6c,d show a clear tendency of cluster building. Especially, the sphere counts are building clusters related to the carbon numbers C4–C9 along the component 1 axis. The component

Table 3. Bond-Type Descriptor Sets Corresponding to the Compounds of Table 1

	bond types							
compd no.	CH_3–CH_2	CH_3–CH	CH_3–C	CH_2–CH_2	CH_2–CH	CH_2–C	CH–CH	CH–C
1	0	0	4	0	0	0	0	0
2	1	2	0	0	1	0	0	0
3	1	0	3	0	0	1	0	0
4	0	4	0	0	0	0	1	0
5	1	2	0	1	1	0	0	0
6	2	1	0	0	2	0	0	0
7	1	0	3	1	0	1	0	0
8	0	4	0	0	2	0	0	0
9	0	2	3	0	0	0	0	1
10	2	0	2	0	0	2	0	0
11	1	2	0	2	1	0	0	0
12	1	3	0	0	1	0	1	0
13	3	0	0	0	3	0	0	0
14	0	2	3	0	1	1	0	0
15	1	0	3	2	0	1	0	0
16	0	4	0	1	2	0	0	0
17	1	3	0	0	3	0	0	0
18	1	1	3	0	1	0	0	1
19	2	0	2	1	0	2	0	0
20	0	5	0	0	0	2	0	0
21	1	2	2	0	0	1	0	1
22	1	2	0	3	1	0	0	0
23	2	1	0	2	2	0	0	0
24	2	2	0	0	2	0	1	0
25	2	1	0	2	2	0	0	0
26	0	0	6	0	0	2	0	0
27	3	0	1	0	0	3	0	0
28	0	2	3	1	1	1	0	0
29	1	1	3	0	2	1	0	0
30	1	2	2	0	1	2	0	0
31	0	5	0	0	2	0	1	0
32	1	0	3	3	0	1	0	0
33	0	3	3	0	0	0	1	1
34	1	1	3	1	1	0	0	1
35	2	0	3	0	2	0	0	1
36	2	0	2	2	0	2	0	0
37	1	4	0	0	1	0	2	0
38	1	4	0	0	1	0	2	0
39	0	4	2	0	0	0	0	2
40	2	1	0	3	2	0	0	0
41	4	0	0	0	4	0	0	0
42	2	0	0	1	0	0	0	0
43	2	0	0	2	0	0	0	0
44	2	0	0	3	0	0	0	0
45	2	0	0	4	0	0	0	0
46	2	0	0	5	0	0	0	0
47	2	0	0	6	0	0	0	0
48	1	0	0	1	3	0	0	0
49	0	0	0	5	0	0	0	0
50	1	0	0	2	3	0	0	0
51	0	1	0	3	2	0	0	0
52	0	0	0	6	0	0	0	0
53	0	1	2	2	1	1	0	1
54	0	1	2	1	2	2	0	0
55	0	1	0	4	2	0	0	0
56	1	0	0	3	3	0	0	0
57	0	0	2	4	0	2	0	0
58	0	2	0	3	2	0	1	0
59	1	0	0	4	3	0	0	0
60	1	0	0	4	3	0	0	0
61	0	1	2	2	2	2	0	0
62	0	2	0	1	4	0	0	0
63	0	2	0	1	4	0	0	0
64	0	2	0	2	2	0	1	0
65	0	2	0	2	2	0	1	0
66	2	0	3	2	0	1	0	0
67	0	2	0	3	2	0	1	0
68	0	2	0	2	4	0	0	0
69	0	2	0	2	4	0	0	0

2 axis is useful in investigating the pattern of the descriptor set. The path count plot shows a similar behavior. Herein the *n*-alkanes are labeled by their compound numbers corresponding to Table 1 to give some orientation for compound positions. The scatter plots of sphere and path count sets show no topological degeneracy. Because of the 4 pairs of stereochemical isomers, 65 different descriptor sets are present, corresponding to the entries in the plots. The projection for bond types in Figure 6b has 61 entries. In addition, four pairs of conmpounds (23, 25), (37, 38), (59, 60), and (58, 66) show

Table 4. Sphere Count Descriptor Sets Corresponding to the Compounds of Table 1

	sphere counts										sphere counts								
compd no.	S1	S2	S3	S4	S5	S6	S7	S8	S9	compd no.	S1	S2	S3	S4	S5	S6	S7	S8	S9
1	5	8	12	0	0	0	0	0	0	36	9	16	20	16	10	8	2	0	0
2	5	8	8	4	0	0	0	0	0	37	9	16	20	20	16	0	0	0	0
3	6	10	14	6	0	0	0	0	0	38	9	16	20	20	12	4	0	0	0
4	6	10	12	8	0	0	0	0	0	39	9	16	24	24	8	0	0	0	0
5	6	10	10	6	4	0	0	0	0	40	9	16	16	14	10	8	6	2	0
6	6	10	10	8	2	0	0	0	0	41	9	16	20	24	12	0	0	0	0
7	7	12	16	8	6	0	0	0	0	42	4	6	4	2	0	0	0	0	0
8	7	12	14	8	8	0	0	0	0	43	5	8	6	4	2	0	0	0	0
9	7	12	18	12	0	0	0	0	0	44	6	10	8	6	4	2	0	0	0
10	7	12	16	12	2	0	0	0	0	45	7	12	10	8	6	4	2	0	0
11	7	12	12	8	6	4	0	0	0	46	8	14	12	10	8	6	4	2	0
12	7	12	14	12	4	0	0	0	0	47	9	16	14	12	10	8	6	4	2
13	7	12	12	12	6	0	0	0	0	48	5	10	6	4	0	0	0	0	0
14	8	14	20	10	12	0	0	0	0	49	5	10	10	0	0	0	0	0	0
15	8	14	18	10	8	6	0	0	0	50	6	12	10	6	2	0	0	0	0
16	8	14	16	10	8	8	0	0	0	51	6	12	14	4	0	0	0	0	0
17	8	14	16	12	10	4	0	0	0	52	6	12	12	6	0	0	0	0	0
18	8	14	20	16	6	0	0	0	0	53	8	16	24	16	0	0	0	0	0
19	8	14	18	14	8	2	0	0	0	54	8	16	24	12	4	0	0	0	0
20	8	14	18	16	8	0	0	0	0	55	7	14	16	10	2	0	0	0	0
21	8	14	20	18	4	0	0	0	0	56	7	14	16	8	4	0	0	0	0
22	8	14	14	10	8	6	4	0	0	57	8	16	22	14	4	0	0	0	0
23	8	14	14	12	10	4	2	0	0	58	8	16	20	12	8	0	0	0	0
24	8	14	16	16	8	2	0	0	0	59	8	16	18	10	8	4	0	0	0
25	8	14	14	12	8	6	2	0	0	60	8	16	18	14	6	2	0	0	0
26	9	16	26	12	18	0	0	0	0	61	9	18	26	18	10	0	0	0	0
27	8	14	18	18	6	0	0	0	0	62	7	14	18	8	2	0	0	0	0
28	9	16	22	12	10	12	0	0	0	63	7	14	18	8	2	0	0	0	0
29	9	16	22	14	14	6	0	0	0	64	7	14	18	10	0	0	0	0	0
30	9	16	22	16	14	4	0	0	0	65	7	14	18	10	0	0	0	0	0
31	9	16	20	16	12	8	0	0	0	66	8	16	20	16	4	0	0	0	0
32	9	16	20	12	10	8	6	0	0	67	8	16	20	16	4	0	0	0	0
33	9	16	24	20	12	0	0	0	0	68	8	16	20	14	6	0	0	0	0
34	9	16	22	18	10	6	0	0	0	69	8	16	20	14	6	0	0	0	0
35	9	16	22	20	14	0	0	0	0										

Table 5. Path Count Descriptor Sets Corresponding to the Compounds of Table 1

	path counts										path counts								
compd no.	P1	P2	P3	P4	P5	P6	P7	P8	P9	compd no.	P1	P2	P3	P4	P5	P6	P7	P8	P9
1	5	4	6	0	0	0	0	0	0	36	9	8	10	8	5	4	1	0	0
2	5	4	4	2	0	0	0	0	0	37	9	8	10	10	8	0	0	0	0
3	6	5	7	3	0	0	0	0	0	38	9	8	10	10	6	2	0	0	0
4	6	5	6	4	0	0	0	0	0	39	9	8	12	12	4	0	0	0	0
5	6	5	5	3	2	0	0	0	0	40	9	8	8	7	5	4	3	1	0
6	6	5	5	4	1	0	0	0	0	41	9	8	10	12	6	0	0	0	0
7	7	6	8	4	3	0	0	0	0	42	4	3	2	1	0	0	0	0	0
8	7	6	7	4	4	0	0	0	0	43	5	4	3	2	1	0	0	0	0
9	7	6	9	6	0	0	0	0	0	44	6	5	4	3	2	1	0	0	0
10	7	6	8	6	1	0	0	0	0	45	7	6	5	4	3	2	1	0	0
11	7	6	6	4	3	2	0	0	0	46	8	7	6	5	4	3	2	1	0
12	7	6	7	6	2	0	0	0	0	47	9	8	7	6	5	4	3	2	1
13	7	6	6	6	3	0	0	0	0	48	5	5	6	4	2	0	0	0	0
14	8	7	10	5	6	0	0	0	0	49	5	5	5	5	5	0	0	0	0
15	8	7	9	5	4	3	0	0	0	50	6	6	7	8	4	2	0	0	0
16	8	7	8	5	4	4	0	0	0	51	6	6	7	7	7	2	0	0	0
17	8	7	8	6	5	2	0	0	0	52	6	6	6	6	6	6	0	0	0
18	8	7	10	8	3	0	0	0	0	53	8	8	12	13	11	6	2	0	0
19	8	7	9	7	4	1	0	0	0	54	8	8	12	11	13	8	0	0	0
20	8	7	9	8	4	0	0	0	0	55	7	7	8	8	8	8	2	0	0
21	8	7	10	9	2	0	0	0	0	56	7	7	8	9	9	4	2	0	0
22	8	7	7	5	4	3	2	0	0	57	8	8	11	10	10	10	4	0	0
23	8	7	7	6	5	2	1	0	0	58	8	8	10	11	11	6	4	0	0
24	8	7	8	8	4	1	0	0	0	59	8	8	9	10	11	6	4	2	0
25	8	7	7	6	4	3	1	0	0	60	8	8	9	10	10	10	4	2	0
26	9	8	13	6	9	0	0	0	0	61	9	9	13	12	14	12	8	0	0
27	8	7	9	9	3	0	0	0	0	62	7	7	9	9	10	5	0	0	0
28	9	8	11	6	5	6	0	0	0	63	7	7	9	9	10	5	0	0	0
29	9	8	11	7	7	3	0	0	0	64	7	7	9	10	9	4	1	0	0
30	9	8	11	8	7	2	0	0	0	65	7	7	9	10	9	4	1	0	0
31	9	8	10	8	6	4	0	0	0	66	8	8	10	11	10	10	4	1	0
32	9	8	10	6	5	4	3	0	0	67	8	8	10	11	10	10	4	1	0
33	9	8	12	10	6	0	0	0	0	68	8	8	10	10	11	10	5	0	0
34	9	8	11	9	5	3	0	0	0	69	8	8	10	10	11	10	5	0	0
35	9	8	11	10	7	0	0	0	0										

topological equivalence with respect to bond-type sets.

To summarize the results of the principal component analysis of atom-type, bond-type, sphere count, and path count molecular descriptor sets, two facts are of importance. At first, the degree of orthogonality of the descriptor sets prefers the atom- and bond-type sets as molecular descriptors in

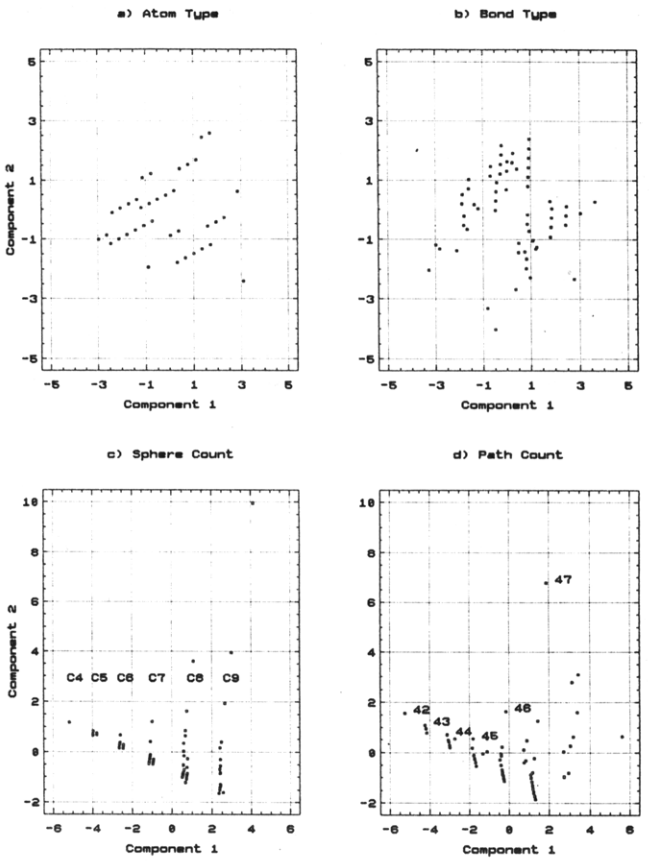

Figure 6. Principle component analysis scatter plots corresponding to the four investigated descriptor sets. The entries in the plots are determined by the first two principle components of each compound included in Table 1. C4–C9 characterize the number of carbon atoms of the entries in plot c; 42–47 are the positions of the *n*-alkanes in plot d.

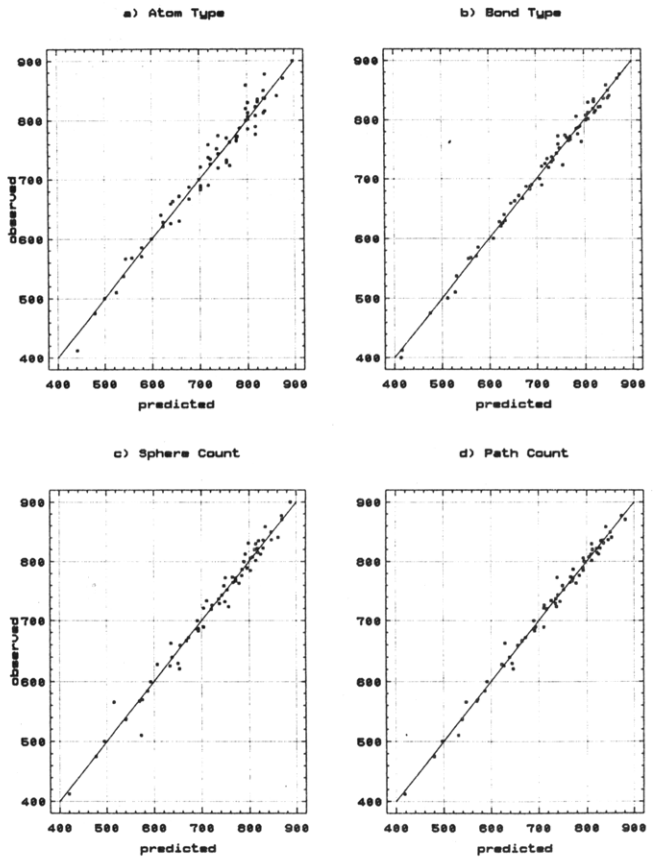

Figure 7. Comparison of *retention indexes* predicted by pure descriptor set estimation models (predicted) summarized in *Table 6, nos. 1–4*, with data from the literature[20] (observed).

estimation models. The second important factor is that the tendency of degeneracy is strong with respect to atom-type sets. A closer look at cyclic compounds shows that degeneracy is also relevant for sphere count sets. The investigated compounds show small degenerative tendencies for bond-type sets and none for path count sets (with the exception of cis–trans isomers). With an investigation of other data sets, bond types will generally show degeneracy for compounds with very similar topological constitutions. In these cases path counts may be helpful, showing the smallest tendency of degeneration. The consequence for the development of prediction models is that one has to make a compromise between orthogonality of descriptor sets on the one hand and degeneracy on the other.

TOPOLOGICAL DESCRIPTOR SET MODELS FOR THE PREDICTION OF BOILING POINTS AND RETENTION INDEX DATA

On the basis of the discussed descriptor sets, estimation models have been developed in order to predict gas chromatographic retention index data and boiling points of the alkanes given in Table 1. As a model building technique we used multilinear regression with descriptor sets as independent variables. Table 6 summarizes the models and corresponding results, and Figures 7–9 illustrate observed versus predicted

Table 6. Results of Pure Descriptor Set Models and Mixed Descriptor Models Fitting Retention Index Data and Boiling Points of the Compounds of Table 1[a]

model no.	model type	variable: dependent	variable: independent	no. of independent variables	MAE	SE	r	corresponding figure
1	atom type	*I*	*C*, A1–A4	5	14.70 RIU	19.89 RIU	0.9843	7a
2	bond type	*I*	*C*, B1–B8	9	8.12 RIU	11.47 RIU	0.9948	7b
3	sphere count	*I*	S1–S6	6	11.12 RIU	16.73 RIU	0.9997	7c
4	path count	*I*	*C*, P1–P8	9	7.82 RIU	11.54 RIU	0.9947	7d
5	atom type	BP	*C*, A1–A4	5	3.84 K	5.04 K	0.9886	8a
6	bond type	BP	*C*, B1–B8	9	2.67 K	3.60 K	0.9942	8b
7	sphere count	BP	*C*, S1–S3, S5, S6	6	2.94 K	4.23 K	0.9920	8c
8	path count	BP	*C*, P1–P7	8	2.60 K	3.43 K	0.9947	8d
9	mixed	*I*	*C*, B1–B7, P4–P7	12	6.05 RIU	9.98 RIU	0.9960	9a
10	mixed	*I*	B1–B8, S5–S7	11	7.36 RIU	10.48 RIU	0.9999	9b
11	mixed	*I*	A1–A4, S4–S6	7	8.89 RIU	12.90 RIU	0.9998	9c
12	mixed	*I*	A1–A4, S4–S6, B3,[c] B4,[c] P6,[c] P7[c]	11	5.77 RIU	8.12 RIU	0.9999	9d

[a] Parameters: *I* = retention index; BP = boiling point; *C* = regression constant; Ai^b = atom type; Bi^b = bond type; Si^b = sphere count; Pi^b = path count; MAE = mean absolute error; SE = standard error of estimate; r = correlation coefficient. Units: K = Kelvin; RIU = retention index units. [b] The numbering of the descriptor set components corresponds to Tables 2–5. [c] Additional descriptors used for cyclic compounds only.

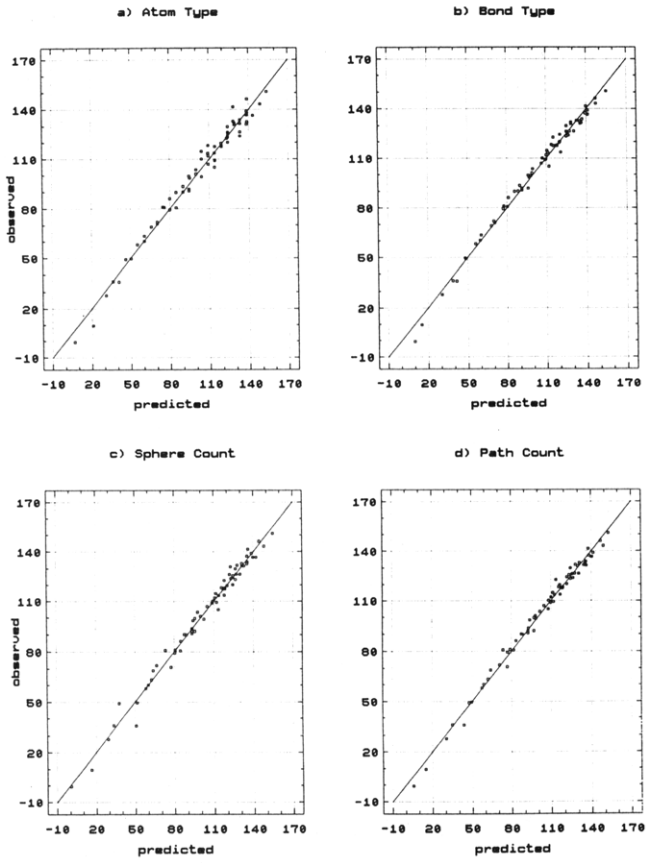

Figure 8. Comparison of *normal boiling points* predicted by pure descriptor set estimation models (predicted) summarized in *Table 6, nos. 5–8*, with data from the literature[21] (observed).

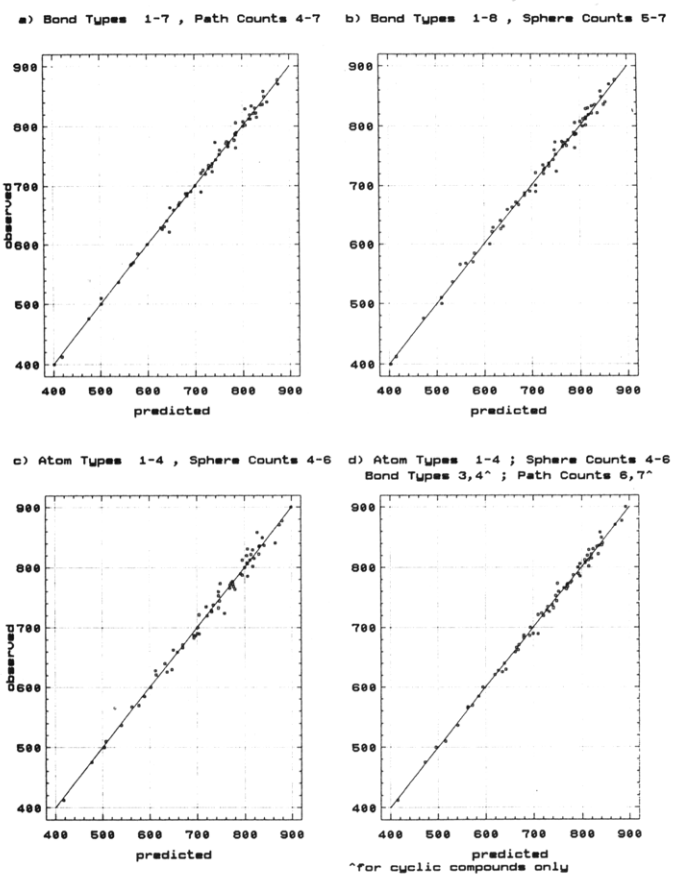

Figure 9. Comparison of *retention indexes* predicted by mixed descriptor estimation models (predicted) summarized in *Table 6, nos. 9–12*, with data from the literature[20] (observed).

properties with respect to the different estimation models.

The models can be classified in three groups. The first four models are based on the four pure descriptor sets and estimate retention index data. Models 5–8 use nearly the same descriptors for the estimation of boiling points. Using exactly the same descriptors produced some regression coefficients with low significance. With an investigation of retention indexes and boiling points quasi simultaneously, an estimation of the lack of fit of the models is to be expected. Because both properties are strongly correlated, the influence of the descriptors on the estimation results should be studied with better reliability than investigation of only one of the properties. The last group of the estimation models 9–12 surrounds mixed models with the intention to improve prediction errors. The latter models were developed with an interactive regression technique.

With respect to mean absolute errors (MAE) and standard errors of estimate (SE), the best pure descriptor set models are the bond-type and path count based models. They show the smallest prediction errors for retention indexes as well as for boiling points. The quality of these models is almost the same for estimating retention indexes or boiling points.

An improvement of the estimation errors can be reached by application of mixed descriptor sets. The last four models in Table 6 give some examples of retention index estimation using a refined palette of topological descriptors. It should be mentioned that all regression coefficients derived from the presented mixed models are of high significance. The possibility of improving the pure descriptor set models by mixing topological descriptors indicates that the topology of the investigated compounds is not exhaustively described by pure descriptor sets as introduced above. Combining topological information will cover a greater part of the variability of the data, so that the error of an ideal model should only incorporate the pure experimental error of the observed data.

DISCUSSION OF TOPOLOGICALLY BASED ESTIMATION MODELS

In order to validate the quality of the developed models, it is useful to get information about the residuals containing the remaining variability of the data. The residuals can be broken down into two parts, the pure experimental variability and the variability associated with the lack of fit of the developed models. The lack of fit is a measure of how well the models describe the observed data, while the experimental variability is immanent in the data and of course independent of the chosen model. To determine the experimental error, at least two observations of each compound's property must be known. For both investigated properties, boiling points as well as retention index data, only one observation was available. On the other hand, the properties are strongly correlated, so that a rough approximation of the pure experimental error can be made by calculating each of both properties with the other property as the regressor.

The relationship between boiling points and retention indexes is illustrated in Figure 10. Herein a strong correlation is recognized with a small tendency toward nonlinearity. It is a well-known fact that boiling points of *n*-alkanes are not directly proportional to their carbon numbers. Therefore the Kovats' retention index definition is based on a logarithmic transformation that results in a direct proportionality between the retention index and the carbon number of the *n*-alkanes. Regression analysis of boiling points versus retention index data and vice versa verifies the nonlinear relationship. In

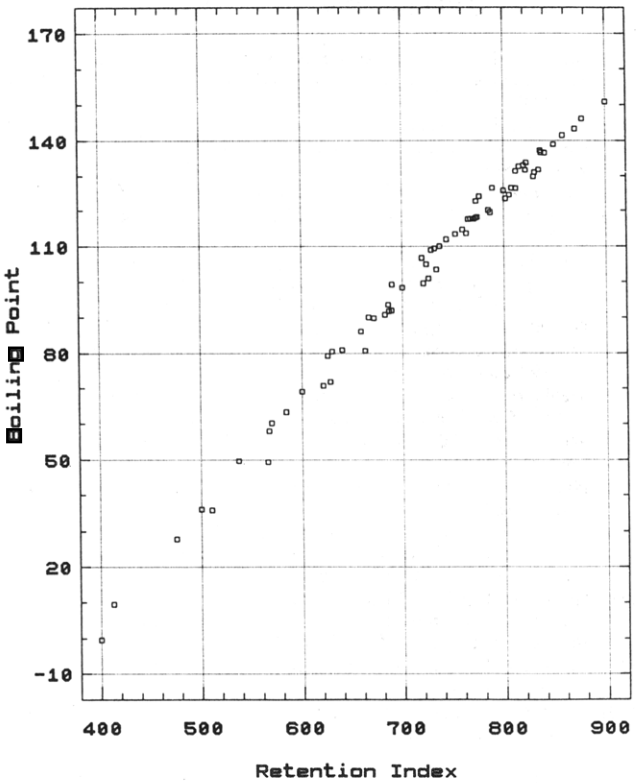

Figure 10. Normal boiling points[21] versus retention indexes[20] for the compounds of Table 1.

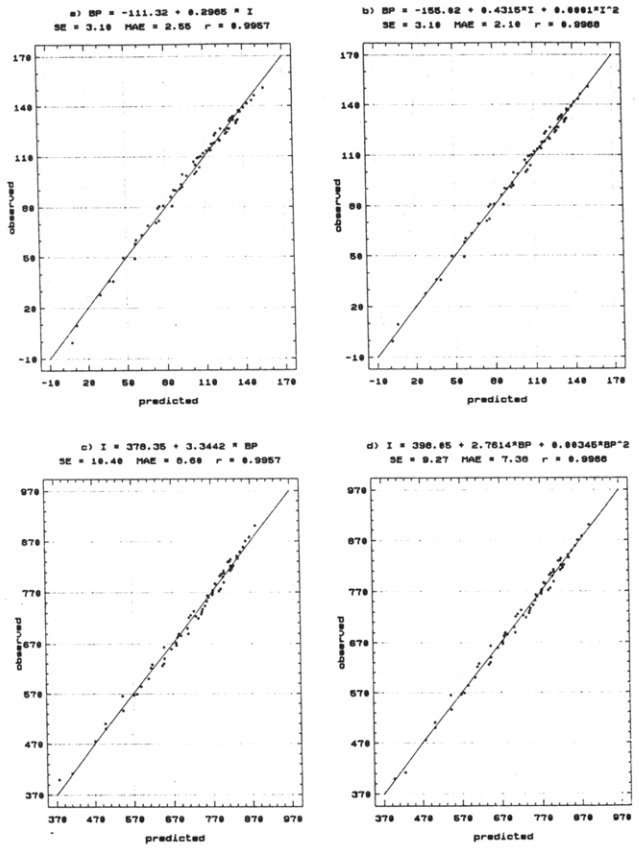

Figure 11. Comparison of predicted normal boiling points (BP) with data from the literature[20,21] (observed) by linear regression models of observed retention indexes (*I*) (a, b) and vice versa (c, d).

Figure 11 observed boiling points and retention indexes are plotted against model derived values. Besides linear models in Figure 11a,c, quadratic models in Figure 11b,d are presented to approximate the logarithmic relationship. For both

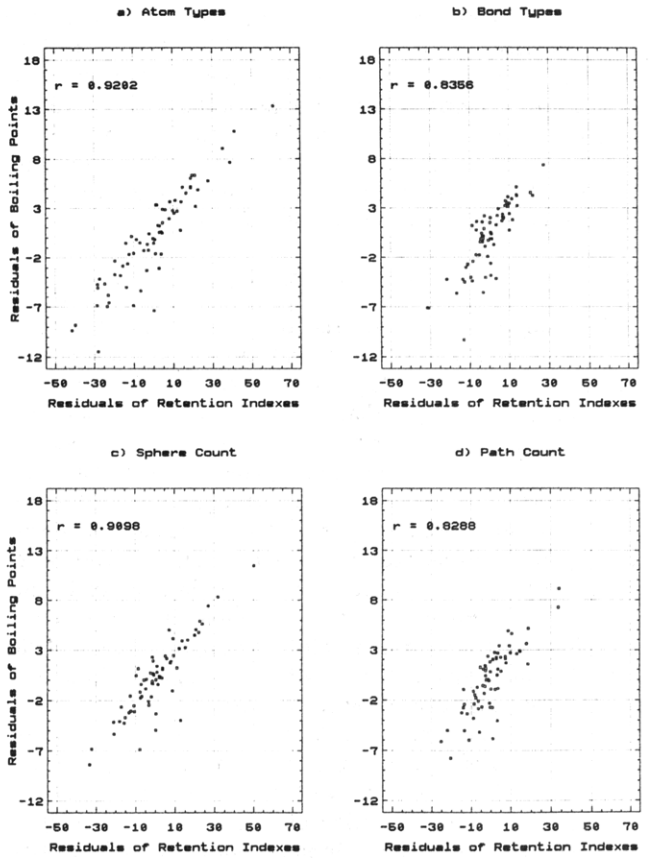

Figure 12. Residuals of boiling points versus residuals of retention indexes. Both properties are predicted with pure descriptor set models. Residuals of atom-type and sphere count set models show high correlation coefficients. The estimation errors are dominated by a lack of topological information in the models. The estimation errors of bond-type and path count set models are dominated by the pure experimental error. Correlation coefficients become smaller.

properties the coefficients of the quadratic terms are of high significance and improve the estimation errors significantly. As approximations for the pure experimental errors, the standard errors of estimate of the quadratic models are received with SE = 3.10 K corresponding to boiling points and SE = 9.27 corresponding to retention index data.

Comparing the approximation of the pure experimental errors with the standard errors of the topologically based models of Table 6 allows an estimation of the lack of fit. In an ideal model, the lack of fit is zero and the model's standard error is equal to the pure experimental error.

The pure descriptor set models 1–4, predicting retention indexes, and models 5–8, predicting boiling points, show higher standard errors in comparison with the approximated experimental errors. As expected, the lack of fit is much greater for atom-type and sphere count models than for bond-type and path count models. This is also illustrated in Figure 12, in which the residuals of the predicted boiling points are plotted versus the residuals of the predicted retention indexes. With respect to atom-type and sphere count models, the residuals still show a relatively strong correlation. Obviously the residuals incorporate further topological information not covered by the used descriptor sets. On the other hand, bond-type and path count models cover a wider range of topological information. The residuals are dominated by the pure experimental errors, and the correlation of the residuals becomes smaller. Nevertheless the lack of fit is also not negligible for pure bond-type and path count descriptor sets.

To minimize the lack of fit, mixed descriptor models were investigated and are examplarily shown in Table 6 (models

9–12). The error of model 9 with a value of 9.98 lies near the approximated experimental error with 9.27. Nearly all of the topological information of the investigated compounds is covered by model 9. The error of model 12 with 8.12 causes suspicion because it is even smaller than the experimental error. In order to fit the few cyclic compounds of the data set, model 12 applies four additional independent variables. The result is obviously an overfitting of the data.

CONCLUSIONS

In a study of quantitative structure–property relations, atom-type, bond-type, sphere count, and path count sets are basic descriptors of the topology of a molecule. The presented investigation has shown that PCA is an effective tool for the comparison of molecular descriptor sets. The PCA scatter plots give an impression of the degeneracy, while the PCA weight plots correspond to the orthogonality within the descriptor sets. Multilinear regression applied to complete descriptor sets as independent variables clearly favors bond-type and path count models. For an estimation of boiling points and retention index data, a combination of the latter descriptors gives the best results. An approximation of the lack of fit shows that nearly the whole topological information of a molecule is covered by such models.

Our investigation on molecular topology based on a software system that performs list operations on chemical graphs.[1,2] All the four descriptor sets were derived by list operations. Especially for the generation of path counts, queueing of topological information combined with the breadth-first strategy is the preferable technique for computing these descriptors. Also combinational topological information like the Hosoya index and other related indexes can effectively be derived by those basic techniques. We will report on this topic in a later paper.

ACKNOWLEDGMENT

On the occasion of his retirement from his profession, the authors thank Prof. Dr. E. Jackwerth, Lehrstuhl für Analytische Chemie, Ruhr-Universität Bochum, for his generous financial and ideal support of our work on computer applications in chemistry during the past years.

REFERENCES AND NOTES

(1) Gautzsch, R.; Zinn, P. List operations on chemical graphs. 1. Basic list structures and operations. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 541–550.
(2) Gautzsch, R.; Zinn, P. List operations on chemical grafts. 2. Combining basic list operations. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 551–555.
(3) Abelson, H.; Susman, G. J.; Susman, I. *Structure and interpretation of computer programs*; MIT Press: Cambridge, MA, 1986.
(4) Sedgewick, R. *Algorithmen in C*; Addison-Wesley (Deutschland): Bonn, 1992.
(5) Duvenbeck, Ch.; Zinn, P. List operations on chemical graphs. 3. Development of vertex and edge models for fitting retention index data. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 211–219.
(6) Duvenbeck, Ch.; Zinn, P. List operations on chemical graphs. 4. Using edge models for prediction of retention index data. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 220–230.
(7) Mehlhorn, K. *Data structures and algorithms. 2: Graph algorithms and NP-Completeness*; Springer-Verlag: Berlin, 1984.
(8) Winston, P. H. *Künstliche Intelligenz*; Addison-Wesley (Deutschland): Bonn, 1987.
(9) Nilsson, N. N. *Principles of artificial intelligence*; Springer-Verlag: Berlin, 1982.
(10) Wiener, H. J. Correlation of heats of isomerization and differences in heat of vaporization of isomers, among the paraffin hydrocarbons. *J. Am. Chem. Soc.* **1947**, *69*, 2636–2638.
(11) Hosoya, H. Topological index. A newly proposed quantity characterizing the topological nature of structure isomers of saturated hydrocarbons. *Bull. Chem. Soc. Jpn.* **1971**, *44*, 2332–2339.
(12) Randic, M. On characterization of molecular branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609–6615.
(13) Balaban, A. T. Chemical graphs XXXIV. Five new topological indices for the branching of tree-like graphs [1]. *Theor. Chim. Acta* **1979**, *53*, 355–375.
(14) Hall, L. H.; Kier, L. B. The molecular connectivity X indexes and κ shape indexes in structure-property modeling. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers: New York, 1991; p 367.
(15) Randic, M. Resolution of ambiguities in structure–property studies by use of orthogonal descriptors. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 311–320.
(16) Hermann, A.; Zinn, P. Unpublished investigations.
(17) Balaban, A. T. Applications of graph theory in chemistry. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 334–343.
(18) Miyashita, Y.; Ohsako, H.; Okuyama, T.; Sasaki, S.; Randic, M. Computer-assisted studies of structure-property relationships using graph invariants. *Magn. Reson. Chem.* **1991**, *29*, 362–365.
(19) Diudea, M. V.; Minailiuc, O.; Balaban, A. T. Molecular topology. IV. Regressive vertex degrees (new graph invariants) and derived topological indices. *J. Comput. Chem.* **1991**, *12*, 527–535.
(20) Rijks, J.; Cramers, C. High precision capillary gas chromatography of hydrocarbons. *Chromatographia* **1974**, *77*, 99–106.
(21) *Beilsteins Handbuch der Organischen Chemie*, Band 1 und 5; Springer-Verlag: Berlin, 1925.