List Operations on Chemical Graphs. 6. Comparative Study of Combinatorial Topological Indexes of the Hosoya Type

A. Hermann and P. Zinn*

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

Received December 30, 1994[⊗]

The topological index Z introduced by Hosoya can be generated by counting all possible combinations of nonadjacent bonds of a molecule. In this paper a generalization of Hosoya's approach is made by counting nonadjacent paths. The resulting Z_i indexes are incorporating Hosoya's index. The new indexes are compared with well established path count indexes by principal components analysis. The character of degeneracy is pointed out, and the potential for model building is investigated. The Z_i index based models are compared with well established topological models for boiling point estimation.

INTRODUCTION

One of the most famous topological indexes is the combinatorial Z_G index introduced by Hosoya.[1] The Z index results from the summation of nonadjacent numbers $p(G,k)$, where $p(G,k)$ is the number of combinations in which k bonds are so chosen from a graph G that they have no common vertex

$$Z_G = \sum_{k=0}^{m} p(G,k)$$

where $p(G,0)$ has per definition a value of 1, $p(G,1)$ is the number of bonds of G, and $p(G,m)$ is the maximum number of combinations for given k bonds of G. To illustrate Hosoya's Z index Figure 1 shows all possible combinations for $p(G,2) = 11$ and $p(G,3) = 4$ where G is the graph of 2,3,3-trimethylpentane. For this molecule $p(G,1) = 7$ and $p(G,0) = 1$ per definition. Therefore the resulting Z index has a value of 23. The Z index presents several interesting properties. Among these are the narrow relationship to the characteristic polynomial of acyclic molecules which implies the relation to corresponding boiling points and other physical properties. Consequently several papers report the application of the Z index in chemical structure property relationship (QSPR) studies.[2−6] An overview about the application of other indexes is given in ref 7.

As a paradigm summarizing Hosoya's approach one can say that the Z index counts all combinations of nonadjacent bonds within a molecular graph. To generalize this paradigm the approach of this paper is to introduce combinatorial indexes of the Hosoya type that count nonadjacent molecular paths. This approach incorporates the Hosoya index counting nonadjacent bonds which is now called Z_2 characterizing that the bond path consists of two atoms. Higher indexes Z_3, Z_4, Z_5, etc. consider nonadjacent paths including three, four, five, etc. atoms. The higher Z_i indexes are mostly of interest investigating the properties of large molecules. The smallest index of this type is Z_1 counting nonadjacent atoms; Z_1 is also of importance for small molecules.

In this paper we will introduce Z_i indexes and show exemplarily the generation of the Z_1 index. The potential of the Z_i indexes applied in QSPR studies will be discussed using principal components analysis. Models for the estimation of boiling points based on the Z_i indexes will be described and compared with models based on well established topological indexes as molecular descriptors.

METHODOLOGY OF COMBINATORIAL Z_i INDEXES

The definition of the Z_i indexes is completely analogous to Hosoya's definition of Z

$$Z_i = \sum_{k=0}^{m} Z_{ik}$$

where i is the chosen path length (number of included atoms), k is the number of combined paths with length i, and m is the maximal possible number of k. Also in analogy to Hosoya $Z_{i0} = 1$ is determined per definitionem. The other Z_{ik} are the counts of all possible combinations of nonadjacent paths k of length i within a given molecule. To give an illustration of a Z_i index Figure 2 shows all combinations for the Z_1 index of the graph of 2,3,3-trimethylpentane starting from $k = 2$ and ending at $k = 5$. The mathematically combined atoms are signed by a dot. Counting these combinations results in $Z_{12} = 21$, $Z_{13} = 24$, $Z_{14} = 12$, and $Z_{15} = 2$. Because Z_{11} is always equal to the number of atoms in a molecule and under consideration of Z_{10}, the overall value for the Z_1 index of the example molecule is $Z_1 = 68$.

Another illustrated example is the Z_3 index of the same molecule 2,3,3-trimethylpentane. The combinations with $k = 1$ and $k = 2$ are shown in Figure 3. The possible combinations of nonadjacent paths with length 3 are signed by double lines. Here the values of the subindexes are $Z_{31} = 10$ and $Z_{32} = 4$. The resulting index has a value of $Z_3 =$

[⊗] Abstract published in *Advance ACS Abstracts*, April 15, 1995.

0095-2338/95/1635-0551$09.00/0

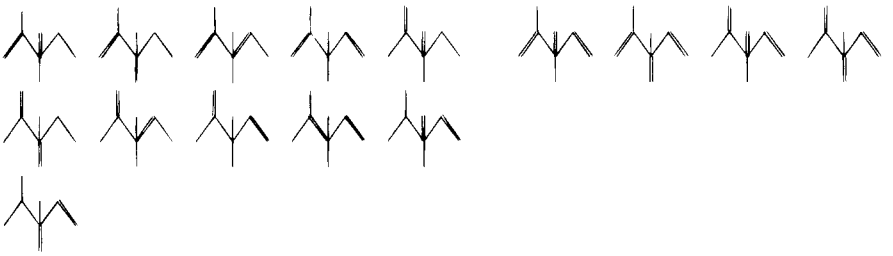

Figure 1. All possible combinations of two and three nonadjacent bonds of 2,3,3-trimethylpentane: $Z_{22} = 11$; $Z_{23} = 4$. Nonadjacent bonds are represented by double lines.

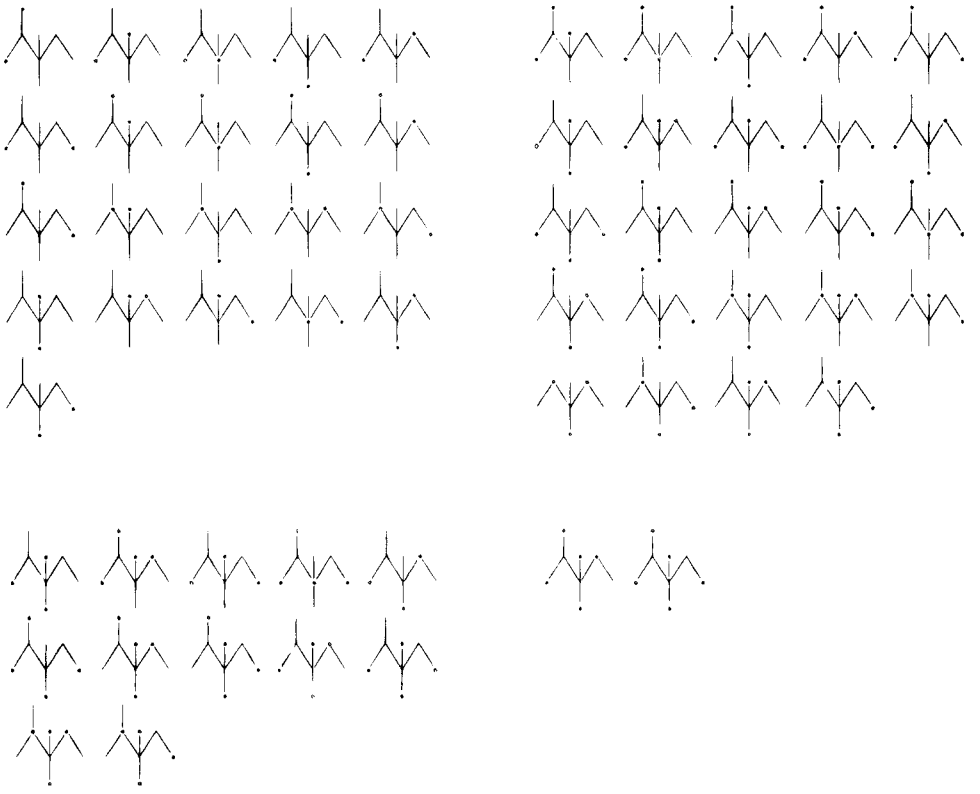

Figure 2. All possible combinations of two to five nonadjacent atoms of 2,3,3-trimethylpentane: $Z_{12} = 21$; $Z_{13} = 24$; $Z_{14} = 12$; $Z_{15} = 2$. Nonadjacent atoms are represented by dots.

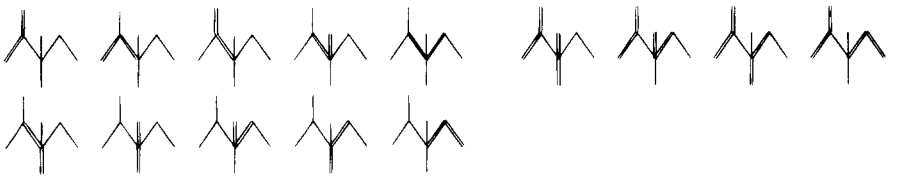

Figure 3. All possible combinations of nonadjacent paths including three atoms of 2,3,3-trimethylpentane: $Z_{31} = 10$; $Z_{32} = 4$. Nonadjacent paths are represented by double lines.

15. It is obvious looking at the Z_i indexes of the example molecule that the number of combinations decreases with increasing length of nonadjacent paths.

The implementation of the Z_i index generation is straightforward and is integrated in the previously described list operating system on chemical graphs.[8,9] The programming technique is a breadth-first search[10] of the combinations of nonadjacent paths. The corresponding LISP code for Z_1 generation is mostly self declaring and consists only of COMMON LISP[11] primitives and basic lisp operations on chemical graphs.[8] It is to be found in Figure 4. A typical interactive application of the developed procedures is shown in Figure 5. This example demonstrates the Z_1 index and subindex generation for the example molecule. In a first step the SMILES[12] like lisp notation of 2,3,3-trimethylpentane is put into the program. Then all combinations of nonadjacent atoms are generated. In the last two steps the combinations are counted to build the subindexes Z_{1k} and the overall Z_1 index. The generation of the higher Z_i indexes and subindexes is to be done analogously.

INVESTIGATED DATA SET

Because the aim of the presented paper is a pure topological discussion of the Z_i indexes only hydrocarbon data are integrated in the investigated data set.[13] The 130 compounds and corresponding boiling points are to be found in Table 1. The first 72 compounds are the complete set of alkane

```
; ** procedure for the calculation of Z1 topological index **
;
(defun z1 (constitution)
  (+ (length (gen-not-adjacent-atoms constitution)) 1))

; ** procedure for the generation of Z1i topological subindexes **
;
(defun gen-z1-count (constitution)
  (do* ((k 0 (+ k 1))
        (not-adjacent-atoms (gen-not-adjacent-atoms constitution))
        (actual-count 1
                      (length
                        (remove-if-not
                          #'(lambda (atoms)
                                    (= (length atoms) k))
                          not-adjacent-atoms)))
        (z1-count '((0 1))
                  (append z1-count
                          (list (list k actual-count)))))
       ((= actual-count 0) (reverse (cdr (reverse z1-count))))))

; ** procedure for the generation of combinations of not-adjacent-atoms
;    and recommended subprocedures **
;
(defun gen-not-adjacent-atoms (constitution)
  (make-not-adjacent-atoms
    nil
    (mapcar 'list (remove-duplicates (make-flat (gen-bond-list constitution))))
    (gen-bond-list constitution)))

(defun make-not-adjacent-atoms (not-adjacent-atoms queue bond-list)
  (cond ((null queue) (reverse not-adjacent-atoms))
        (t (make-not-adjacent-atoms
             (adjoin (car queue) not-adjacent-atoms :test 'subsetp)
             (reverse (adjoin-list-of-sublists
                        (expand-z1-queue (car queue) bond-list)
                        (reverse (cdr queue))))
             bond-list))))

(defun expand-z1-queue (atoms bond-list)
  (let ((linked-atoms
          (remove-duplicates
            (append atoms
                    (mapcan #'(lambda(atom)
                                (get-neighbours atom bond-list nil))
                            atoms))))
        (all-atoms (remove-duplicates (make-flat bond-list))))
    (mapcar #'(lambda (atom) (append atoms (list atom)))
            (set-difference all-atoms linked-atoms))))
```

Figure 4. LISP implementation of procedures and subprocedures for the generation of Z_1 index and Z_{1k} subindexes.

```
; typical output of z1 computing procedures for
; 2,3,3-trimethylpentane

>(setq constitution '(c c (c) c (c) (c) c c))
       ;id-numbers:   1 2  3  4  5   6  7 8

>(gen-not-adjacent-atoms constitution)

((1) (3) (2) (5) (6) (4) (7) (8)
 ;combinations of non-adjacent-atoms determining Z11

 (1 3) (1 5) (1 6) (1 4) (1 7) (1 8) (3 5) (3 6) (3 4) (3 7) (3 8
 (2 5) (2 6) (2 7) (2 8) (5 6) (5 7) (5 8) (6 7) (6 8) (4 8)
 ;combinations of non-adjacent-atoms determining Z12

 (1 3 5) (1 3 6) (1 3 4) (1 3 7) (1 3 8) (1 5 6) (1 5 7) (1 5 8)
 (1 6 7) (1 6 8) (1 4 8) (3 5 6) (3 5 7) (3 5 8) (3 6 7) (3 6 8)
 (3 4 8) (2 5 6) (2 5 7) (2 5 8) (2 6 7) (2 6 8) (5 6 7) (5 6 8)
 ;combinations of non-adjacent-atoms determining Z13

 (1 3 5 6) (1 3 5 7) (1 3 5 8) (1 3 6 7) (1 3 6 8) (1 3 4 8)
 (1 5 6 7) (1 5 6 8) (3 5 6 7) (3 5 6 8) (2 5 6 7) (2 5 6 8)
 ;combinations of non-adjacent-atoms determining Z14

 (1 3 5 6 7) (1 3 5 6 8))
 ;combinations of non-adjacent-atoms determining Z15

>(gen-z1-count constitution)
((0 1) (1 8) (2 21) (3 24) (4 12) (5 2))
  ;Z10   Z11    Z12    Z13    Z14   Z15

>(z1 constitution)

68 ; sum of Z10 to Z15
```

Figure 5. Example of Z_1 index and subindex generation for 2,3,3-trimethylpentane.

isomers from C_4 to C_9. The last 58 compounds include cycloalkanes with one or two rings and almost the same range of carbon numbers. Table 2 incorporates the Z_i indexes and subindexes of the corresponding compounds in Table 1 used as molecular descriptors. Besides the Z_i indexes paths counts P_k are also included in Table 2. The topological properties of path counts within QSPR studies are discussed in a former paper of this series[10] and will be used as a reference in order to compare with Z_i indexes.

PRINCIPAL COMPONENTS ANALYSIS OF Z_i MOLECULAR DESCRIPTORS

Principal components analysis is already applied to investigate path counts P_k and related topological indexes as molecular descriptors and is discussed in detail in ref 10. In general principal components analysis can help to investigate the orthogonality of descriptor sets such as P_k or Z_{ik} and to estimate the character of degeneracy of molecular descriptors with respect to the investigated data set.

A central visualization technique of principal components analysis is the eigenvector projection of each compound's descriptors. If the character of degeneracy is 0 each compound has its own entry in the projection, and the chosen molecular descriptors are completely different. Figures 6 and 7 show eigenvector projections of the Z_{1k}, Z_{2k}, and Z_{3k} subindex sets and as a link to former discussed descriptor sets[10] path count projection. The path count set has the smallest tendency to degenerate. A closer look at Table 2 shows that there is only one pair of compounds nos. 57 and 58 with identical values in the path count set. The compounds are 3-ethyl-3-methyl-hexane and 2,3,4-trimethylhexane. These compounds are easy to differentiate because their atom types as the most simple descriptor set are

Table 1. Boiling Points of 130 Hydrocarbons (72 Alkanes, 58 Cyclic Compounds)

no.	name	bp	no.	name	bp
1	butane	-0.5	66	3-ethyl-2,2-dimethylpentane	133.8
2	2-methylpropane	-11.7	67	3,3,4-trimethylhexane	140.5
3	pentane	36.1	68	3-ethyl-2,3-dimethylpentane	141.6
4	2-methylbutane	27.8	69	2,2,3,4-tetramethylpentane	133
5	2,2-dimethylpropane	9.5	70	2,3,3,4-tetramethylpentane	141.5
6	hexane	69	71	2,2,4,4-tetramethylpentane	122.7
7	2-methylpentane	60.3	72	2,2,3,3-tetramethylpentane	140.3
8	3-methylpentane	63.3	73	cyclopentane	49.3
9	2,3-dimethylbutane	58	74	ethylcyclopropane	35.9
10	2,2-dimethylbutane	49.7	75	methylcyclobutane	36.3
11	heptane	98.4	76	1,1-dimethylcyclopropane	20.7
12	2-methylhexane	90	77	bicyclo[2.1.0]pentane	45.5
13	3-methylhexane	92	78	cyclohexane	80.7
14	3-ethylpentane	93.5	79	methylcyclopentane	71.9
15	2,4-dimethylpentane	80.5	80	ethylcyclobutane	70.7
16	2,3-dimethylpentane	89.8	81	isopropylcyclopropane	58.4
17	2,2-dimethylpentane	79.2	82	1-methyl-1-ethylcyclopropane	56.8
18	3,3-dimethylpentane	86.1	83	1,1,2-trimethylcyclopropane	52.6
19	2,2,3-trimethylbutane	80.9	84	cycloheptane	118.1
20	octane	125.7	85	methylcyclohexane	101.1
21	2-methylheptane	117.6	86	ethylcyclopentane	103.7
22	3-methylheptane	118	87	isopropylcyclobutane	92.7
23	4-methylheptane	117.7	88	1,1-dimethylcyclopentane	87.9
24	3-ethylhexane	118.5	89	bicyclo[3.2.0]heptane	109.5
25	2,5-dimethylhexane	109	90	spiro[2.4]heptane	99
26	2,4-dimethylhexane	109.4	91	cyclooctane	151.1
27	2,3-dimethylhexane	115.6	92	methylcycloheptane	134
28	2-methyl-3-ethylpentane	115.6	93	ethylcyclohexane	131.8
29	2,2-dimethylhexane	106.8	94	propylcyclopentane	130.9
30	3,3-dimethylhexane	112	95	isobutylcyclobutane	119.5
31	3,4-dimethylhexane	117.7	96	isopropylcyclopentane	126.4
32	2,3,4-dimethylpentane	113.4	97	1,1-dimethylcyclohexane	119.8
33	3-ethyl-3-methylpentane	118.2	98	1,1,3-trimethylcyclopentane	104.9
34	2,2,4-trimethylpentane	99.2	99	1,1,2-trimethylcyclopentane	113.7
35	2,2,3-trimethylpentane	110	100	cis-bicyclo[4.2.0]octane	136
36	2,3,3-trimethylpentane	114.7	101	spiro[2.5]octane	125.5
37	2,2,3,3-tetramethylbutane	106.5	102	2-methylbicyclo[2.2.1]heptane	127
38	nonane	150.8	103	ethylcycloheptane	164
39	2-methyloctane	142.8	104	butylcyclopentane	156.8
40	3-methyloctane	143.3	105	propylcyclohexane	156.7
41	4-methyloctane	142.4	106	isobutylcyclopentane	148.3
42	3-ethylheptane	143	107	isopropylcyclohexane	154.6
43	4-ethylheptane	141.2	108	1-methyl-3-propylpentane	148.3
44	2,6-dimethylheptane	135.2	109	1-ethyl-1-butylcyclopropane	140.4
45	2,5-dimethylheptane	136	110	1-methyl-3-isopropylpentane	138
46	2,4-dimethylheptane	133.5	111	tert-butylcyclopentane	145.2
47	2,3-dimethylheptane	140.5	112	dicyclobutylmethane	161
48	3,5-dimethylheptane	136	113	spiro[4.4]nonane	157
49	4-ethyl-2-methylhexane	133.8	114	2-ethylbicyclo[2.2.1]heptane	146.5
50	3,4-dimethylheptane	140.1	115	4-methylspiro[2.5]octane	149
51	3-ethyl-2-methylheptane	138	116	cyclodecane	201
52	3-ethyl-4-methylheptane	140.4	117	ethylcyclooctane	185.5
53	2,2-dimethylheptane	132.7	118	pentylpentane	180
54	3,3-dimethylheptane	137.3	119	butylcyclohexane	181
55	2,3,5-trimethylhexane	131.3	120	isopentylcyclopentane	171.5
56	4,4-dimethylheptane	135.2	121	1-methyl-4-propylcyclohexane	176
57	3-ethyl-3-methylhexane	140.6	122	1-methyl-2-propylcyclohexane	176
58	2,3,4-trimethylhexane	139	123	1-methyl-1-butylcyclopentane	176.4
59	3-ethyl-2,4-dimethylpentane	136.7	124	1-ethyl-2-isopropylcyclopentane	160
60	3,3-diethylpentane	146.2	125	tert-butylcyclohexane	171.6
61	2,2,5-trimethylhexane	124	126	1,2,4,5-tetramethylcyclohexane	173.5
62	2,2,4-trimethylhexane	126.5	127	1,1,2,5-tetramethylcyclohexane	158
63	2,4,4-trimethylhexane	126.5	128	1,1,2,3-tetramethylcyclohexane	167
64	2,2,3-trimethylhexane	131.7	129	spiro[4.5]decane	192
65	2,3,3-trimethylhexane	137.7	130	1-methylhexahydroindane	182.5

different. Also the Z_1, Z_2, and Z_3 indexes of these compounds are different. As a remark it is of interest that the homologous pairs of heptanes or octanes have different path count sets and show no degeneracy. In the opposite to the path count set the three Z_{ik} descriptor sets have a relatively high degree of degeneracy with 10% in the case of the Z_{1k} descriptors and nearly 20% for Z_{2k} and Z_{3k}. With respect to the tendency of degeneration path count descriptors are by far more advantageous for estimation model development than the Z_{ik} descriptor sets.

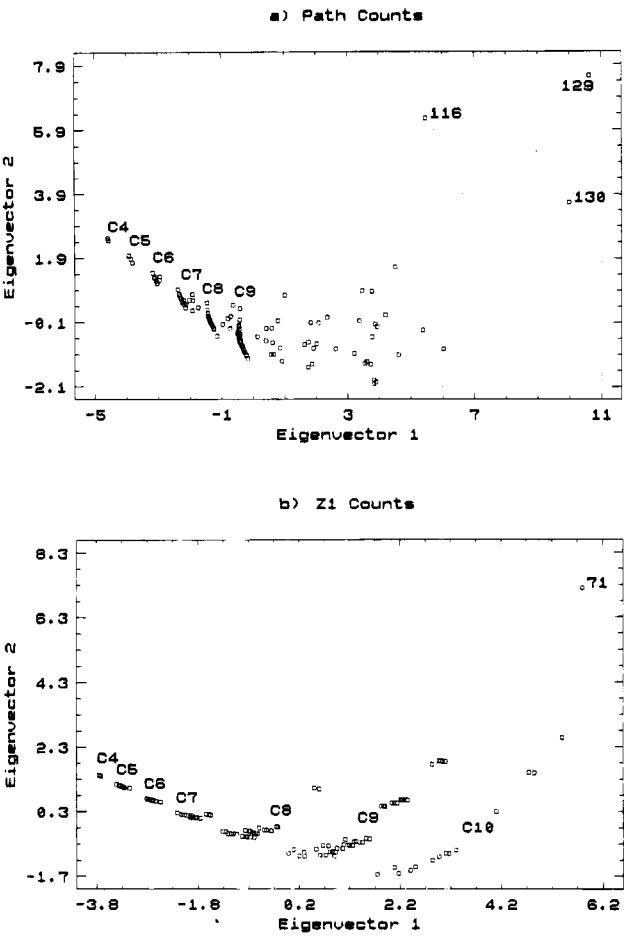

Figure 6. Principal components scatter plots corresponding to (a) pure path count and (b) pure Z_{1k} descriptor sets. Cluster building corresponds to carbon numbers. Extreme compounds are labeled corresponding to Table 1.

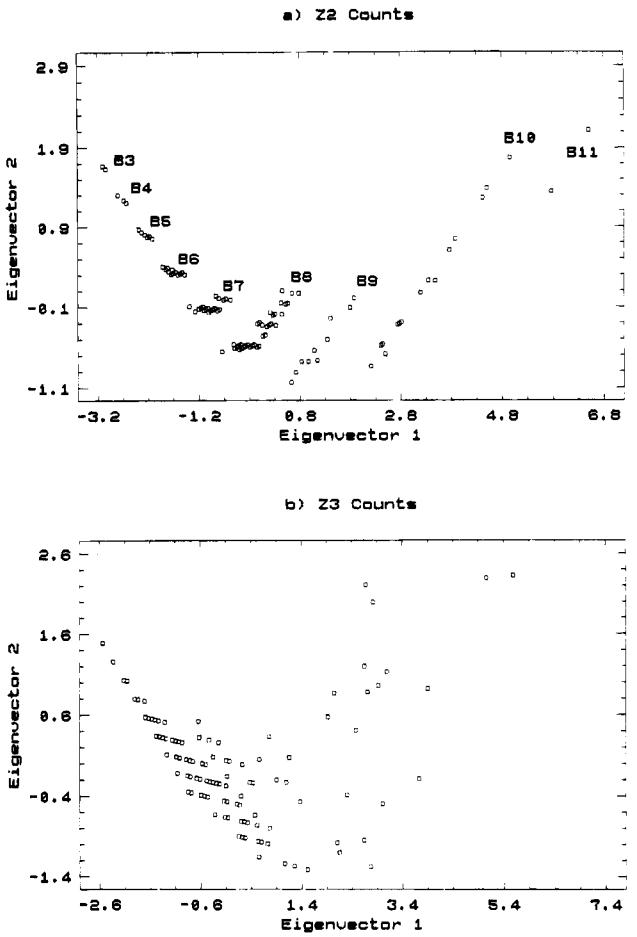

Figure 7. Principal components scatter plots corresponding to (a) pure Z_{2k} and (b) pure Z_{3k} descriptor sets. Z_{2k} cluster building corresponds to the number of bonds. Z_{3k} projection shows less clustering tendency but strongly degenerative characteristic.

Another interesting point of view looking at the eigenvector projection is the cluster building of the compounds. The acyclic isomers are clustering relatively close together in the path count projection of Figure 6a what is in accordance with their boiling points. The cyclic compounds show a less pronounced cluster building. Especially the compounds 116, 129, and 130 are extreme outliers with respect to the other compounds. This is also in good accordance with the high boiling points of these compounds. The Z_{1k} eigenvector projection in Figure 6b shows a more homogeneous ordering in cluster building. Here the cyclic compounds as well as the acyclic ones are lying on traces corresponding to their carbon numbers. As an extraordinary compound 2,2,4,4-tetramethylpentane no. 71 has an extreme cut off to the rest of the C_9 compounds. This results from the high symmetry of the associated graph with the consequence of a high number of possible combinations of nonadjacent atoms. The cluster building in the Z_{2k} eigenvector projection in Figure 7a has a similar principle of order as Z_{1k}. In this case the clustering traces are determined by the numbers of bonds of the compounds. The Z_{3k} descriptors show in Figure 7b a less pronounced tendency of cluster building. The impression of a degenerative descriptor set is predominant.

Besides the investigated complete descriptor sets of the Z_{ik} subindexes, the principal components analysis was further applied to descriptor sets consisting of Z_i indexes and combinations of Z_i indexes with the numbers of atoms (P_1) and bonds (P_2) of the compounds. The associated eigenvector projections are to be seen in Figures 8 and 9. The first remarkable result is that the set of the indexes Z_1, Z_2, and Z_3 in Figure 8a shows completely no tendency to degenerate. Also combinations of Z_1 and Z_2 with P_1 in Figure 8b respectively Z_1 and Z_3 with P_2 in Figure 9a have only small degenerative character. If only one of the Z_i indexes remains to be seen in Figure 9b, the degree of degeneracy increases strongly. The exact percentage values of degeneracy of all investigated descriptor sets are tabulated in Table 3. The cluster building in Figures 8 and 9 is to be interpreted in a similar way. The eigenvector-1-axis in the four eigenvector projections represents the carbon number of the compounds. It characterizes the magnitude of the molecules while the eigenvector-2-axis represents the molecular shape. Particularly in the cases of the descriptor sets Z_1, Z_2, Z_3 and Z_1, Z_3, P_2 in Figures 8a and 9a the compounds are linearly separable into three compound classes: without rings, with one ring, and with two rings. Overall the principal components analysis and eigenvector projection of the Z indexes and subindexes are helpful for investigating the tendency of degeneration and the potential of compound classification. The results are of interest in developing models for the prediction of boiling points based on these indexes.

Table 2. Molecular Descriptors of the Compounds of Table 1[a]

no.	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	Z_{11}	Z_{12}	Z_{13}	Z_{14}	Z_{15}	Z_{16}	Z_1	Z_{21}	Z_{22}	Z_{23}	Z_{24}	Z_{25}	Z_2	Z_{31}	Z_{32}	Z_{33}	Z_3
1	4	3	2	1	0	0	0	0	4	3	0	0	0	0	8	3	1	0	0	0	5	2	0	0	3
2	4	3	3	0	0	0	0	0	4	3	1	0	0	0	9	3	0	0	0	0	4	3	0	0	4
3	5	4	3	2	1	0	0	0	5	6	1	0	0	0	13	4	3	0	0	0	8	3	0	0	4
4	5	4	4	2	0	0	0	0	5	6	2	0	0	0	14	4	2	0	0	0	7	4	0	0	5
5	5	4	6	0	0	0	0	0	5	6	4	1	0	0	17	4	0	0	0	0	5	6	0	0	7
6	6	5	4	3	2	1	0	0	6	10	4	0	0	0	21	5	6	1	0	0	13	4	1	0	6
7	6	5	5	3	2	0	0	0	6	10	5	1	0	0	23	5	5	0	0	0	11	5	1	0	7
8	6	5	5	4	1	0	0	0	6	10	5	0	0	0	22	5	5	1	0	0	12	5	0	0	6
9	6	5	6	4	0	0	0	0	6	10	6	1	0	0	24	5	4	0	0	0	10	6	1	0	8
10	6	5	7	3	0	0	0	0	6	10	7	2	0	0	26	5	3	0	0	0	9	7	0	0	8
11	7	6	5	4	3	2	1	0	7	15	10	1	0	0	34	6	10	4	0	0	21	5	3	0	9
12	7	6	6	4	3	2	0	0	7	15	11	3	0	0	37	6	9	2	0	0	18	6	4	0	11
13	7	6	6	5	3	1	0	0	7	15	11	2	0	0	36	6	9	3	0	0	19	6	2	0	9
14	7	6	6	6	3	0	0	0	7	15	11	1	0	0	35	6	9	4	0	0	20	6	0	0	7
15	7	6	7	4	4	0	0	0	7	15	12	5	1	0	41	6	8	0	0	0	15	7	5	0	13
16	7	6	7	6	2	0	0	0	7	15	12	3	0	0	38	6	8	2	0	0	17	7	2	0	10
17	7	6	8	4	3	0	0	0	7	15	13	6	1	0	43	6	7	0	0	0	14	8	3	0	12
18	7	6	8	6	1	0	0	0	7	15	13	4	0	0	40	6	7	2	0	0	16	8	0	0	9
19	7	6	9	6	0	0	0	0	7	15	14	6	1	0	44	6	6	0	0	0	13	9	3	0	13
20	8	7	6	5	4	3	2	1	8	21	20	5	0	0	55	7	15	10	1	0	34	6	6	0	13
21	8	7	7	5	4	3	2	0	8	21	21	8	1	0	60	7	14	7	0	0	29	7	8	0	16
22	8	7	7	6	4	3	1	0	8	21	21	7	0	0	58	7	14	8	1	0	31	7	6	0	14
23	8	7	7	6	5	2	1	0	8	21	21	7	1	0	59	7	14	8	0	0	30	7	5	0	13
24	8	7	7	7	5	2	0	0	8	21	21	6	0	0	57	7	14	9	1	0	32	7	3	0	11
25	8	7	8	5	4	4	0	0	8	21	22	11	2	0	65	7	13	4	0	0	25	8	11	0	20
26	8	7	8	6	5	2	0	0	8	21	22	10	2	0	64	7	13	5	0	0	26	8	8	0	17
27	8	7	8	7	4	2	0	0	8	21	22	9	1	0	62	7	13	6	0	0	27	8	6	0	15
28	8	7	8	8	4	1	0	0	8	21	22	8	0	0	60	7	13	7	1	0	29	8	4	0	13
29	8	7	8	8	5	0	0	0	8	21	22	8	1	0	61	7	13	7	0	0	28	8	3	0	12
30	8	7	9	5	4	3	0	0	8	21	23	13	3	0	69	7	12	3	0	0	23	9	9	0	19
31	8	7	9	7	4	1	0	0	8	21	23	11	2	0	66	7	12	5	0	0	25	9	4	0	14
32	8	7	9	8	4	0	0	0	8	21	23	11	2	0	66	7	12	4	0	0	24	9	7	0	17
33	8	7	9	9	3	0	0	0	8	21	23	9	0	0	62	7	12	7	1	0	28	9	0	0	10
34	8	7	10	5	6	0	0	0	8	21	24	16	6	1	77	7	11	0	0	0	19	10	12	0	23
35	8	7	10	8	3	0	0	0	8	21	24	13	3	0	70	7	11	3	0	0	22	10	6	0	17
36	8	7	10	9	2	0	0	0	8	21	24	12	2	0	68	7	11	4	0	0	23	10	4	0	15
37	8	7	12	9	0	0	0	0	8	21	26	17	6	1	80	7	9	0	0	0	17	12	9	0	22
38	9	8	7	6	5	4	3	2	9	28	35	15	1	0	89	8	21	20	5	0	55	7	10	1	19
39	9	8	8	6	5	4	3	2	9	28	36	19	4	0	97	8	20	16	2	0	47	8	13	1	23
40	9	8	8	7	5	4	3	1	9	28	36	18	2	0	94	8	20	17	4	0	50	8	11	0	20
41	9	8	8	7	6	4	2	1	9	28	36	18	3	0	95	8	20	17	3	0	49	8	10	1	20
42	9	8	8	8	6	4	2	0	9	28	36	17	1	0	92	8	20	18	5	0	52	8	8	0	17
43	9	8	8	8	7	4	1	0	9	28	36	17	2	0	93	8	20	18	4	0	51	8	7	1	17
44	9	8	9	6	5	4	4	0	9	28	37	23	7	1	106	8	19	12	0	0	40	9	17	1	28
45	9	8	9	7	5	5	2	0	9	28	37	22	5	0	102	8	19	13	2	0	43	9	15	0	25
46	9	8	9	7	7	3	2	0	9	28	37	22	7	1	105	8	19	13	0	0	41	9	13	1	24
47	9	8	9	8	5	4	2	0	9	28	37	21	4	0	100	8	19	14	2	0	44	9	12	1	23
48	9	8	9	8	6	4	1	0	9	28	37	21	4	0	100	8	19	14	3	0	45	9	12	0	22
49	9	8	9	8	7	4	0	0	9	28	37	21	5	0	101	8	19	14	2	0	44	9	11	0	21
50	9	8	9	9	6	3	1	0	9	28	37	20	3	0	98	8	19	15	3	0	46	9	9	0	19
51	9	8	9	9	7	3	0	0	9	28	37	20	4	0	99	8	19	15	2	0	45	9	8	1	19
52	9	8	9	10	7	2	0	0	9	28	37	19	2	0	96	8	19	16	4	0	48	9	6	0	16
53	9	8	10	6	5	4	3	0	9	28	38	26	9	1	112	8	18	10	0	0	37	10	16	0	27
54	9	8	10	8	5	4	1	0	9	28	38	24	6	0	106	8	18	12	2	0	41	10	11	0	22
55	9	8	10	8	6	4	0	0	9	28	38	25	8	1	110	8	18	10	0	0	37	10	16	1	28
56	9	8	10	8	7	2	1	0	9	28	38	24	8	1	109	8	18	12	0	0	39	10	9	1	21
57	9	8	10	10	6	2	0	0	9	28	38	22	4	0	102	8	18	14	3	0	44	10	5	0	16
58	9	8	10	10	6	2	0	0	9	28	38	23	5	0	104	8	18	12	2	0	41	10	11	0	22
59	9	8	10	10	8	0	0	0	9	28	38	23	7	1	107	8	18	12	0	0	39	10	9	1	21
60	9	8	10	12	6	0	0	0	9	28	38	20	1	0	97	8	18	16	5	0	48	10	0	0	11
61	9	8	11	6	5	6	0	0	9	28	39	30	12	2	121	8	17	6	0	0	32	11	22	0	34
62	9	8	11	7	7	3	0	0	9	28	39	29	12	2	120	8	17	7	0	0	33	11	18	0	30
63	9	8	11	8	7	12	0	0	9	28	39	28	11	2	118	8	17	8	0	0	34	11	15	0	27
64	9	8	11	9	5	3	0	0	9	28	39	27	9	1	114	8	17	9	0	0	35	11	13	0	25
65	9	8	11	10	5	2	0	0	9	28	39	26	8	1	112	8	17	10	0	0	36	11	10	1	23
66	9	8	11	10	7	0	0	0	9	28	39	26	9	1	113	8	17	10	0	0	36	11	9	0	21
67	9	8	11	11	5	1	0	0	9	28	39	25	6	0	108	8	17	11	2	0	39	11	8	0	20
68	9	8	11	12	5	0	0	0	9	28	39	24	5	0	106	8	17	12	2	0	40	11	5	0	17
69	9	8	12	10	6	0	0	0	9	28	40	30	12	2	122	8	16	6	0	0	31	12	16	0	29
70	9	8	12	12	4	0	0	0	9	28	40	28	9	1	116	8	16	8	0	0	33	12	11	1	25
71	9	8	13	6	9	0	0	0	9	28	41	37	21	7	145	8	15	0	0	0	24	13	27	0	41
72	9	8	13	12	3	0	0	0	9	28	41	31	12	2	124	8	15	6	0	0	30	13	12	0	26
73	5	5	5	5	5	0	0	0	5	5	0	0	0	0	11	5	5	0	0	0	11	5	0	0	6
74	5	5	6	4	2	0	0	0	5	5	0	0	0	0	11	5	4	0	0	0	10	6	0	0	7

Table 2 (Continued)

no.	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	Z_{11}	Z_{12}	Z_{13}	Z_{14}	Z_{15}	Z_{16}	Z_1	Z_{21}	Z_{22}	Z_{23}	Z_{24}	Z_{25}	Z_2	Z_{31}	Z_{32}	Z_{33}	Z_3
75	5	5	6	6	2	0	0	0	5	5	1	0	0	0	12	5	4	0	0	0	10	6	0	0	7
76	5	5	8	4	0	0	0	0	5	5	2	0	0	0	13	5	2	0	0	0	8	8	0	0	9
77	5	6	9	10	7	0	0	0	5	4	0	0	0	0	10	6	6	0	0	0	13	9	0	0	10
78	6	6	6	6	6	6	0	0	6	9	2	0	0	0	18	6	9	2	0	0	18	6	3	0	10
79	6	6	7	7	7	2	0	0	6	9	3	0	0	0	19	6	8	1	0	0	16	7	2	0	10
80	6	6	7	8	4	2	0	0	6	9	3	0	0	0	19	6	8	2	0	0	17	7	1	0	9
81	6	6	8	6	4	0	0	0	6	9	3	0	0	0	19	6	7	0	0	0	14	8	3	0	12
82	6	6	9	7	2	0	0	0	6	9	4	0	0	0	20	6	6	1	0	0	14	9	0	0	10
83	6	6	10	8	2	0	0	0	6	9	5	1	0	0	22	6	5	0	0	0	12	10	1	0	12
84	7	7	7	7	7	7	7	0	7	14	7	0	0	0	29	7	14	7	0	0	29	7	7	0	15
85	7	7	8	8	8	8	2	0	7	14	8	1	0	0	31	7	13	5	0	0	26	8	7	0	16
86	7	7	8	9	9	4	2	0	7	14	8	0	0	0	30	7	13	6	0	0	27	8	4	0	13
87	7	7	9	10	6	4	0	0	7	14	9	2	0	0	33	7	12	4	0	0	24	9	6	0	16
88	7	7	10	9	9	4	0	0	7	14	10	3	0	0	35	7	11	2	0	0	21	10	6	0	17
89	7	8	11	15	15	11	9	0	7	13	6	0	0	0	27	8	17	9	0	0	35	11	9	0	21
90	7	8	12	13	13	8	4	0	7	13	6	0	0	0	27	8	16	7	0	0	32	12	10	0	23
91	8	8	8	8	8	8	8	8	8	20	16	2	0	0	47	8	20	16	2	0	47	8	12	0	21
92	8	8	9	9	9	9	9	2	8	20	17	4	0	0	50	8	19	13	1	0	42	9	13	0	23
93	8	8	9	10	10	10	4	2	8	20	17	3	0	0	49	8	19	14	2	0	44	9	10	0	20
94	8	8	9	10	11	6	4	2	8	20	17	3	0	0	49	8	19	14	1	0	43	9	9	0	19
95	8	8	10	10	8	6	4	0	8	20	18	7	1	0	55	8	18	10	0	0	37	10	15	0	26
96	8	8	10	11	11	6	4	0	8	20	18	5	0	0	52	8	18	11	0	0	38	10	11	0	22
97	8	8	11	10	10	10	4	0	8	20	19	8	1	0	57	8	17	8	0	0	34	11	14	0	26
98	8	8	12	11	13	8	0	0	8	20	20	10	2	0	61	8	16	5	0	0	30	12	16	0	29
99	8	8	12	13	11	6	2	0	8	20	20	8	1	0	58	8	16	7	0	0	32	12	11	0	24
100	8	9	12	16	16	17	12	10	8	19	14	2	0	0	44	9	24	20	3	0	57	12	18	0	31
101	8	9	13	14	14	14	8	4	8	19	14	2	0	0	44	9	23	17	2	0	52	13	20	0	34
102	8	9	13	17	22	19	11	2	8	19	15	3	0	0	46	9	23	16	1	0	50	13	20	0	34
103	9	9	10	11	11	11	11	4	9	27	31	11	0	0	79	9	26	27	8	0	71	10	17	1	29
104	9	9	10	11	12	8	6	4	9	27	31	11	0	0	79	9	26	27	7	0	70	10	16	2	29
105	9	9	10	11	12	12	6	4	9	27	31	11	1	0	80	9	26	27	7	0	70	10	16	3	30
106	9	9	11	11	13	8	6	4	9	27	32	15	3	0	87	9	25	22	2	0	59	11	21	2	35
107	9	9	11	12	12	12	6	4	9	27	32	14	2	0	85	9	25	23	4	0	62	11	19	3	34
108	9	9	11	12	14	10	6	3	9	27	32	14	2	0	85	9	25	23	4	0	62	11	18	2	32
109	9	9	12	12	9	6	3	0	9	27	32	13	0	0	82	9	24	22	6	0	62	12	15	0	28
110	9	9	12	13	14	11	6	0	9	27	33	17	3	0	90	9	24	19	2	0	55	12	22	2	37
111	9	9	13	13	13	8	6	0	9	27	34	20	5	0	96	9	23	16	0	0	49	13	23	0	37
112	9	10	13	16	12	12	12	8	9	26	27	10	1	0	74	10	32	36	12	0	91	13	34	1	49
113	9	10	14	18	22	16	12	8	9	26	28	9	0	0	73	10	31	34	9	0	85	14	24	4	43
114	9	10	14	19	25	23	16	7	9	26	28	9	0	0	73	10	31	33	9	0	84	14	27	2	44
115	9	10	15	18	18	16	10	6	9	26	28	9	0	0	73	10	30	30	7	0	78	15	29	3	48
116	10	10	10	10	10	10	10	10	10	35	50	25	2	0	123	10	35	50	25	2	123	10	25	10	46
117	10	10	11	12	12	12	12	12	10	35	51	28	3	0	128	10	34	46	22	2	115	11	25	5	42
118	10	10	11	12	13	9	8	6	10	35	51	28	3	0	128	10	34	46	21	1	113	11	24	6	42
119	10	10	11	12	13	14	8	6	10	35	51	28	4	0	129	10	34	46	21	2	114	11	24	10	46
120	10	10	12	12	13	10	8	6	10	35	52	33	8	0	139	10	33	40	13	0	97	12	31	8	52
121	10	10	12	13	14	16	10	6	10	35	52	32	7	0	137	10	33	41	16	1	102	12	28	7	48
122	10	10	12	14	15	15	8	5	10	35	52	31	6	0	135	10	33	42	17	1	104	12	24	8	45
123	10	10	13	14	15	11	6	4	10	35	53	35	9	0	143	10	32	38	13	1	95	13	25	8	47
124	10	10	13	16	18	12	7	3	10	35	53	34	8	0	141	10	32	38	13	0	94	13	24	6	44
125	10	10	14	14	14	14	8	6	10	35	54	40	15	2	157	10	31	32	6	0	80	14	34	9	58
126	10	10	14	16	16	18	10	2	10	35	54	39	12	1	152	10	31	32	9	0	83	14	35	4	54
127	10	10	15	16	16	16	10	2	10	35	55	43	17	3	164	10	30	28	5	0	74	15	37	6	59
128	10	10	15	17	16	14	10	3	10	35	55	42	15	2	160	10	30	29	7	0	77	15	34	2	52
129	10	11	14	18	22	21	22	16	10	34	46	21	1	0	113	11	41	61	31	2	147	14	37	14	66
130	10	11	15	20	25	25	19	17	10	34	47	24	3	0	119	11	40	56	25	2	135	15	39	15	70

[a] P_i = path counts; Z_i = combinatorial indexes, Z_{ik} = combinatorial subindexes.

APPLICATION OF Z_i INDEXES AS TOPOLOGICAL DESCRIPTORS IN BOILING POINT ESTIMATION MODELS

On the basis of principal components analysis the suitability of the Z_i indexes as topological molecular descriptors was studied and multilinear regression models were developed in order to estimate the boiling points of the compounds in Table 1. The investigated models are summarized in Table 3, and some examples of the plots of the resulting estimated versus observed boiling points are shown in Figures 10−12.

As a reference the models nos. 1−4 of Table 3 consisting of the previously discussed descriptors,[10] atom types A_i, bond types B_i, and paths counts P_i, were chosen. Model 1 is based on a pure path count descriptor set with a standard estimation error of about 5 °C. The corresponding plot is shown in Figure 10a. To improve this model the path counts are combined with atom and bond type descriptors as summarized in models nos. 2 and 3 of Table 3. Applying these

Table 3. Selected Z_i Index and Z_{ik} Subindex Based Models for Boiling Point Estimation of the Compounds of Table 1[a]

no.	regression parameter	no. of parameters	MAE, °C	SE, °C	r^2	% degeneracy of descriptor set	remark
1	C, P1-P7	8	3.80	5.04	0.9852	0.8	
2	C, A1-A4, P4-P7	9	3.53	4.69	0.9872		reference
3	C, B1-B7, P4-P7	12	2.94	4.02	0.9906		models
4	C, P1, P2, SP	4	6.54	8.63	0.9566		
5	C, A1-A4, P4, P5, Z12, P6·(1-R), P7·R	10	2.58	3.65	0.9922		best
6	C, A3, B1-B6, P4, P5, Z12 P6·(1-R), P7·R	12	2.03	3.03	0.9946		models
7	Z11, Z12, Z13	3	6.94	8.93		10.0	
8	A1-A4, Z12, Z13	7	4.78	6.35			Z subindex
9	Z21, Z22, Z23	3	10.85	13.29		18.5	models
10	C, Z31, Z32	3	21.14	26.12	0.6029	19.2	
11	C, Z1, Z2, Z3	4	12.45	17.41	0.8326	0.0	
12	C, Z1, Z2, Z3	11	4.36	6.28	0.9770	0.0	
	$Z1^2$, $Z2^2$, $Z1^3$, $Z2^3$, $Z3^3$	11	4.36	6.28	0.9770	0.0	
	Z1·Z2, Z1·Z2·Z3						Z index
13	C, Z1, Z2, Z3, P1, P2	6	4.42	6.32	0.9767	0.0	models
14	C, Z1, Z2, P1	4	4.84	6.77	0.9733	3.8	
15	C, Z1, Z3, P2	4	7.20	9.50	0.9474	1.5	
16	C, Z2, P1, P2	4	6.20	8.37	0.9592	16.2	

[a] MAE = mean absolute error; SE = standard error of estimate; r^2 = squared correlation coefficient.

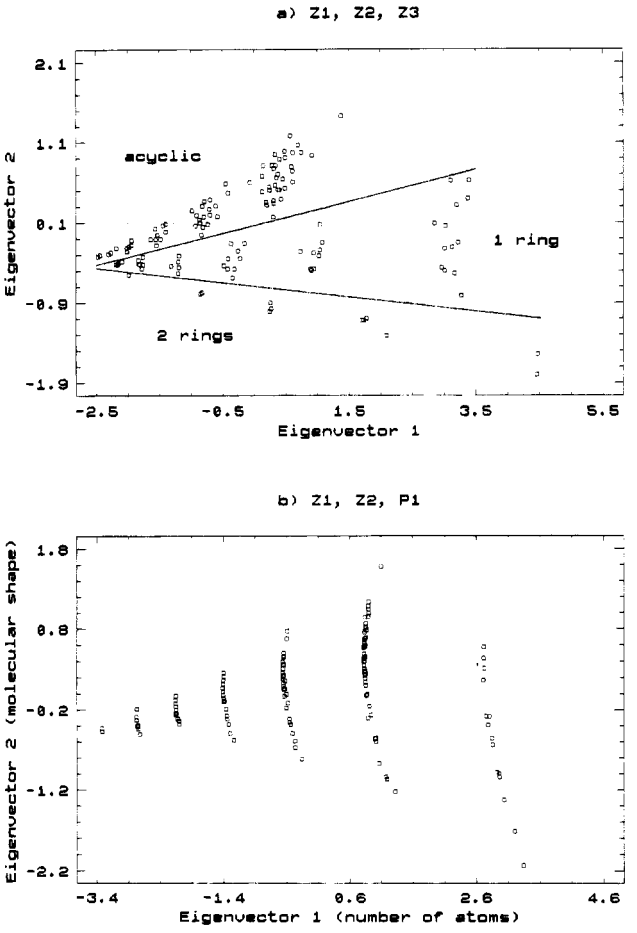

Figure 8. Principal components scatter plots corresponding to (a) Z_1, Z_2, Z_3 and (b) Z_1, Z_2, P_1 descriptor sets. With respect to descriptor set (a) a linear separation into acyclic and cyclic compounds with one or two rings is possible. The compounds in projection (b) are lying on iso carbon number curves.

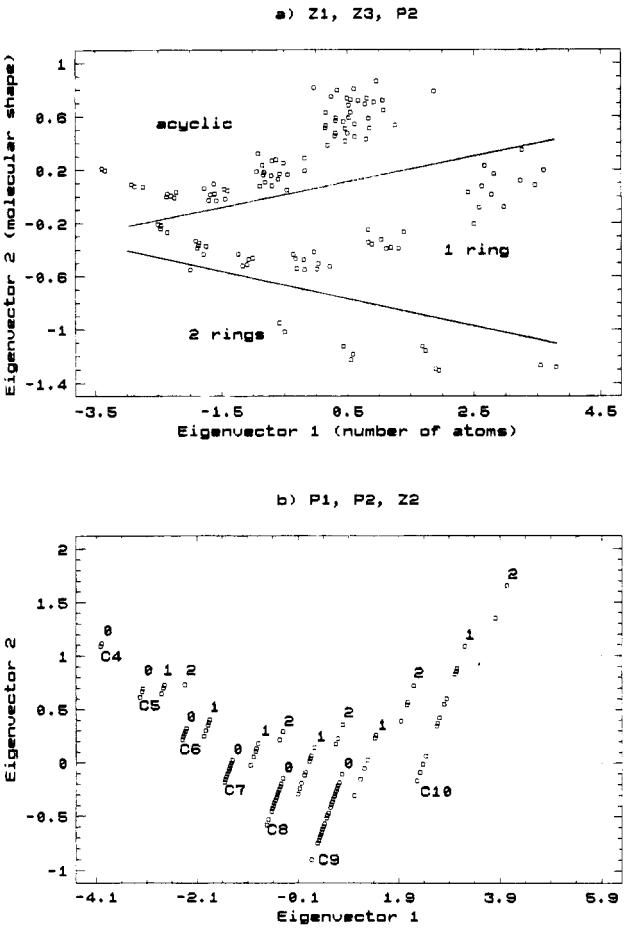

Figure 9. Principal components scatter plots corresponding to (a) Z_1, Z_3, P_2 and (b) P_1, P_2, Z_2 descriptor sets. With respect to descriptor set (a) a linear separation into acyclic and cyclic compounds with one or two rings is possible. The clusters in projection (b) are corresponding to carbon numbers and the subclusters to the molecular shape (0, 1, 2 rings).

additional descriptors the standard errors of estimate decrease about 0.25 °C per descriptor. The atom and bond type descriptors of the improved models substitute the smaller paths counts of model no. 1, e.g., the number of atoms P_1 and the number of bonds P_2. A further reference model no. 4 consisting of a strongly reduced descriptor set including P_1, P_2, and the number of all paths SP is not sufficient to predict boiling points.

The models nos. 7, 9, and 10 are using pure Z_i count sets. Comparing these models with the pure path count

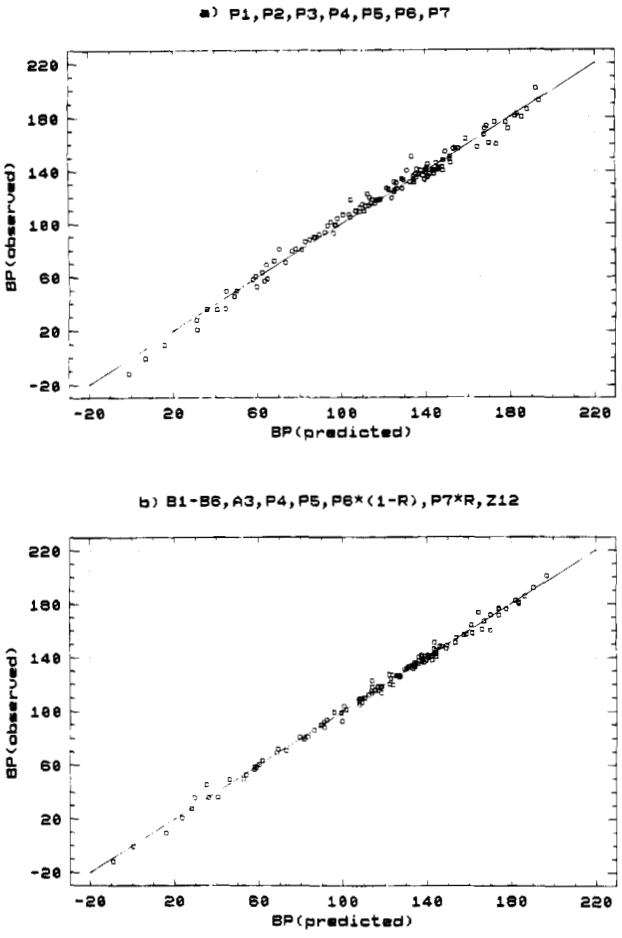

Figure 10. Predicted versus observed boiling points of the compounds in Table 1: (a) pure path count model; no. 1 of Table 3 and (b) best model; no. 6 of Table 3.

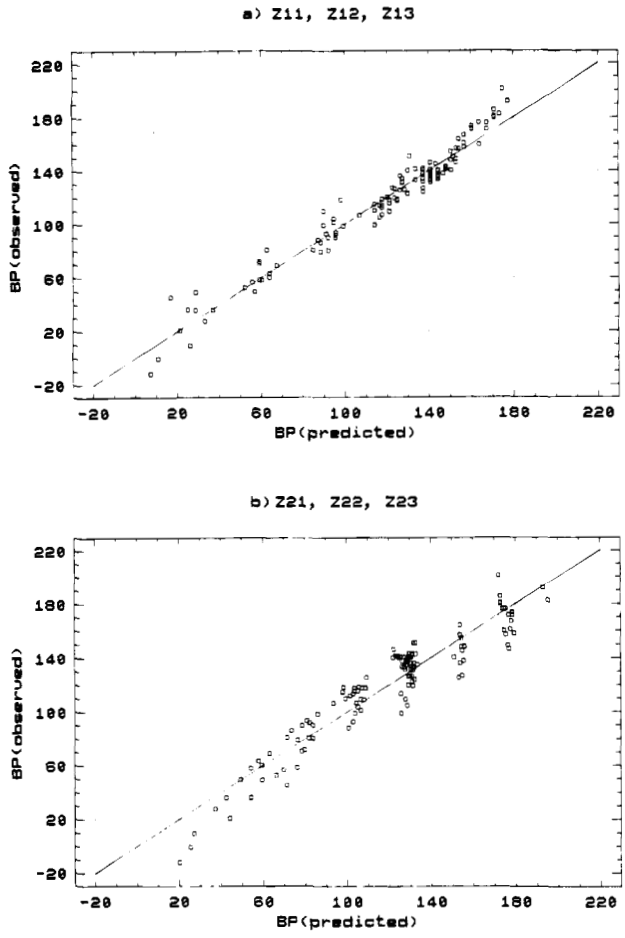

Figure 11. Predicted versus observed boiling points of the compounds in Table 1: (a) Pure Z_{1i} subindex model; no. 7 of Table 3 and (b) Pure Z_{2i} subindex model; no. 9 of Table 3.

model no. 1 several disadvantages are obvious. Especially the increasing prediction error and the unacceptably high character of degeneracy are to be mentioned. In order to give a graphical comparison with the pure path count model of Figure 10a the corresponding plots for the pure Z_1 and Z_2 count sets are shown in Figure 11. A substitution of the atom count Z_{11} by the corresponding atom types in model no. 8 reduces the prediction error more strongly than the analogous substitution from model no. 1 to no. 2, respectively, on a higher error level. As a result it must be stated that boiling point prediction using pure Z_i subindex sets is less suitable than using pure path count sets. The generation of the pure Z_i subindex sets starting from the adjacency information of a molecule is accompanied with a loss of topological information.

To get a more complete transformation of topological information, Z_i index based models nos. 11−16 were investigated. Model no. 11 makes a linear approach using the indexes Z_1, Z_2, and Z_3. Although the tendency of degeneracy is 0 the standard estimation error is extremely high with 17.41 °C. The result is a nonlinear relationship between observed and estimated boiling points as to be seen in Figure 12a. A linearization of this relationship can be reached introducing nonlinear and interaction terms into the model. This is considered in model no. 12 and a mainly linearized relationship shows Figure 12b. The estimation error of this model is decreased about 10 °C comparing with the linear approach. Generating Z_1 to Z_3 indexes the loss of topological information is much smaller than generating the former discussed pure Z_i subindex sets. It is to be recognized that the resulting linear relationship between observed and estimated boiling points is based on a nonlinear model. A linearization can also be reached using additional descriptors, e.g., the number of atoms P_1 and the number of bonds P_2 in model no. 13. Using a smaller set of descriptors by combining two of the Z_i indexes with the important descriptors P_1 or P_2 is not successful. The corresponding models nos. 14−16 have increasing estimation errors.

The most complete transformation of topological information into molecular descriptors is presented in models nos. 5 and 6 of Table 3. A combination of different descriptor types results in small estimation errors down to 2 °C and a highly correlated relationship as to be seen for model no. 6 in Figure 10b. We found that the Z_{12} subindex is of significance in such models. A closer look shows that Z_{12} is a combination of a molecules atom and bond number P_1 and P_2.

Z_{12} classifies molecular graphs into groups with identical numbers of atoms and rings

$$Z_{12} = \frac{1}{2}P_1 (P_1 - 1) - P_2$$

CONCLUSIONS

Overall our investigations have shown that the combinatorial indexes generated by counting nonadjacent molecular paths are supplementary with respect to the well established Hosoya index. The indexes are of more theoretical interest

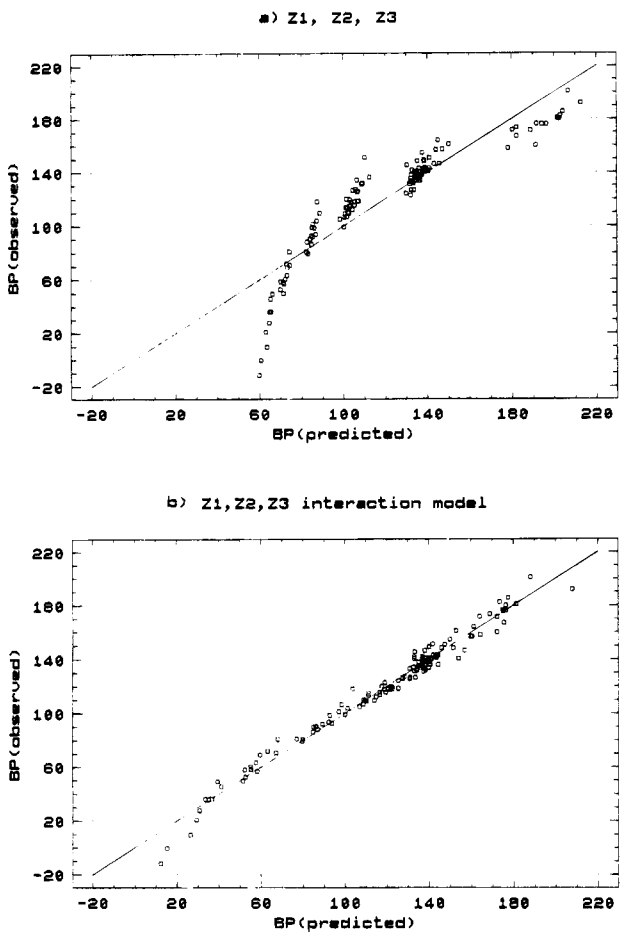

Figure 12. Predicted versus observed boiling points of the compounds in Table 1: (a) Z_i index model; no. 11 of Table 3 and (b) linearized Z_i model; no. 12 of Table 3.

than of practical use within estimation models. Especially developing topological indexes or index sets which have unique values for each molecular graph the Z_i indexes are of importance. Combinations of the Z_i indexes have a very low degree of degeneracy. Despite favorable degenerative characteristics, the Z_i indexes are less suitable as descriptors of estimation models. The tendency to build linear relationships between observed and estimated properties such as boiling points is not strongly pronounced and is overcome by other topological descriptors, e.g., path counts.

REFERENCES AND NOTES

(1) Hosoya, H. Topological index. A newly proposed quantity characterizing the topological nature of structure isomeres of saturated hydrocarbons. *Bull. Chem. Soc. Jpn.* **1971**, *44*, 2332−2339.
(2) Gao, Y.; Hosoya H. Topological index and thermodynamic properties. IV. Size dependency of the structure-activity correlation of alkanes. *Bull. Chem. Soc. Jpn.* **1988**, *61*, 3093−3102.
(3) Hosoya, H. Introduction to graph theory. *Understanding Chem. React.* **1994**, *9*, 1−36.
(4) Buydens, L.; Massart, D. L. Prediction of gaschromatographic retention indices from linear free energy and topological parameters. *Anal. Chem.* **1981**, *53*, 1990.
(5) Buydens, L.; Coomans, D.; Vanbelle, M.; Massart, D. L.; Van den Driesche, R. Comparative study of topological and linear free energy-related parameters for the prediction of GC retention indices. *J. Pharm. Sci.* **1983**, *72*, 1327.
(6) Randić, M. Representation of molecular graphs by basic graphs. *J. Chem. Inf. Comput. Sci.* **1992**, 32, 57−69.
(7) Balaban, A. Applications of graph theory in chemistry. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 334−343.
(8) Gautzsch, R.; Zinn, P. List operations on chemical graphs. 1. Basic list structures and operations. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 541−550.
(9) Gautzsch, R.; Zinn, P. List operations on chemical graphs. 2. Combining basic list operations. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 551−555.
(10) Gautzsch, R.; Zinn, P. 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. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 791−800.
(11) Steele, G. L. *COMMON LISP THE LANGUAGE*. Digital Press: USA, 1990.
(12) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.
(13) *Beilstein's Handbuch der Organischen Chemie*, Band 1 und 5; Springer-Verlag: Berlin, 1925.

CI940142A