ELSEVIER

Journal of Molecular Structure (Theochem) 424 (1998) 237–247

THEO
CHEM

3D weighting of molecular descriptors for QSPR/QSAR by the method of ideal symmetry (MIS). 1. Application to boiling points of alkanes

Andrey Toropov[a], Alla Toropova[b], Temur Ismailov[b], Danail Bonchev[c,d,*]

[a]Institute of Polymer Chemistry and Physics, Academy of Sciences, Tashkent 700122, Uzbekistan
[b]Tashkent State University, Tashkent 700095, Uzbekistan
[c]Texas A&M University, Galveston, TX 77553-1675, USA
[d]Assen Zlatarov University, Burgas 8010, Bulgaria

Received 4 November 1996; revised 19 March 1997; accepted 4 April 1997

Abstract

The method of ideal symmetry (MIS), developed recently, presents molecules as systems of mutually repulsing atoms connected by covalent bonds of constant length. In this paper we have used MIS optimized geometry to define a vertex 3D weight as a metric analogue of the vertex distance sum in molecular graphs. These 3D weights were used as a substitute for the vertex degrees in several well known topological (2D) indices, thus producing a series of 3D-weighted molecular descriptors. The novel indices were tested in calculating the boiling points of a series of 73 C3–C9 alkanes and showed generally a better performance than the original 2D indices. The best 1-, 2-, and 3-variable linear regression models incorporated 3D zero-order molecular connectivity with correlation coefficients of 0.9892, 0.9961, and 0.9986, and standard deviations of 5.97, 3.64, and 2.17°C, respectively. The approach was further validated by correlations with four other properties of alkanes (heats of formation, heats of vaporization, heats of atomization, and molar volume). The potential of the proposed 3D weighting of topological indices for QSPR/QSAR studies was thus demonstrated. © 1998 Elsevier Science B.V.

Keywords: 3D molecular descriptors; 3D atomic weights; Method of ideal symmetry; QSPR; Alkane properties

1. Introduction

Topological and information-theoretic indices have been widely applied to quantitative structure–property (QSPR) and structure–activity relationships (QSAR) of chemical compounds [1–10]. By describing important details of molecular structure, these indices produce models that fit experimental data fairly well. In attempting to further improve such models, a variety of steric, topographical, and other 3D indices have been added to the numerous topological (2D) molecular descriptors [11–18]. Such methods view the molecules as graphs embedded in certain spatial lattices. Many of the existing graph-theoretic schemes have been modified in this manner. The topographic indices of Randić [14–16], for example, embed molecular graphs in a hexagonal grid, and make use of the metric distances in this idealized lattice. Such an approach does distinguish between stereoisomers,

[*] Corresponding author. Fax: 001 409 740 4429; e-mail: bonchevd @tamug.tamu.edu

0166-1280/98/$19.00 © 1998 Elsevier Science B.V. All rights reserved.
PII S0166-1280(97)00151-6

```
      C             H               H               H
     /             /               /               /
 C - C - C     C - C - C       C - C - C       C - C - H
    /             /               /               /
   C             C               H               H

           R1             R4                    R1  R4
          /              /                     /   /
   R2 - C - C    +    C - C - R5   --->   R2 - C - C - R5
        /              /                     /   /
      R3             R6                    R3  R6
```

which is not possible within the 2-dimensional description. Perhaps the major improvement resulting from 3D molecular descriptors is that they account for 'through-space' interactions whereas the topological indices are based on 'through-bonds' interactions. However, this difference in the physical meaning of the 2D and 3D indices should not be overestimated. The relatively high correlation between the corresponding pairs of such molecular descriptors rather indicates that they only mirror to a different extent the two types of interaction which cannot be clearly separated. An effective 3D descriptor is the 3D Wiener index, developed in Bulgaria and Croatia [17,18], which is calculated from the interatomic distances, taken from experiment or calculated after quantum chemical geometry optimization. This index outperforms its widely used 2D counterpart in correlations with various properties and biological activities (see, for example, [33]).

In this paper, the potential improvement of QSPR (and QSAR) by accounting for the three-dimensional structure of molecules is approached in a different way. A '3D weight', based on the metric distances between atoms in a preferable molecular conformation, can be ascribed to many of the known graph-theoretic and information-theoretic indices. The key feature of our modeling is the presentation of the molecule as a system of mutually repulsing atoms, keeping constant the length of each bond. This concept of molecular simulation was termed the "method of ideal symmetry" (MIS) [19,20]. The MIS method manifests some similarity to the approaches that proceed from the intramolecular repulsion of valence electron pairs [21,22]. The MIS-based 3D-weighting procedure was applied in this study to five topological indices and tested against the boiling points and other properties of alkanes.

2. The method

The method of ideal symmetry (MIS) does not employ any parameters other than bond lengths and valence angles. The MIS model of any compound is an assembly of standard atomic blocks (SAB). In the case of alkanes these are four tetrahedral SABs in which the valence angles are 109.47° and the C–C and C–H bond lengths are 0.154 nm and 0.107 nm, respectively.

The assembly of the SABs proceeds by embedding a bond from a pair of SABs until the molecule is constructed, as represented by the scheme.

Then a geometry optimization follows, with rotations about each C–C bond. The MIS model selected is the conformer that minimizes the energy-like E_0 function

$$E_0 = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d(i,j)^{-6} \qquad (1)$$

in which $d(i,j)$ is the metric distance between the ith and jth atoms, and n is the total number of atoms. Indeed, geometry optimization can be performed by using quantum chemical methods or the distance geometry methods of Crippen [23] and Havel [24]. However, for the series of congeneric compounds used in QSPR/QSAR the optimization based on Eq. (1) is much faster and generally provides results in agreement with experimental geometries [19,20].

The MIS models thus obtained might be regarded as molecular graphs whose vertices and edges possess some out-of-plane space. Correspondingly, the MIS geometries may be used for a 3D weighting of graph vertices, thus defining an entire class of *3D-weighted topological indices*. The weighting factor 3DW was introduced as a function of the interatomic

Table 1
Atomic coordinates in 2,3-dimethyl butane (Å) calculated by the MIS method

Atom	x	y	z
C1	1.54	2.43	0.66
H2	2.09	3.34	0.56
H3	0.50	2.62	0.48
H4	1.67	2.03	1.65
C5	2.05	1.40	−0.37
H6	1.69	1.67	−1.35
C7	0.00	0.00	0.00
H8	−0.36	−1.01	0.00
H9	−0.36	0.51	−0.87
H10	−0.36	0.50	0.88
C11	1.54	0.00	0.00
H12	1.90	−0.26	0.97
C13	2.05	−1.02	−1.03
H14	1.50	−1.93	−0.93
H15	3.09	−1.22	−0.85
H16	1.92	−0.63	−2.02
C17	3.59	1.40	−0.37
H18	3.95	0.90	0.50
H19	3.95	2.41	−0.37
H20	3.95	0.90	−1.25

distances $d(i,k)$ in the MIS model:

$$3DW_i = \sum_{\substack{k \neq i \\ k,i \text{ nonadjacent}}}^{n} \exp(d(i,k)^{-2}) \tag{2}$$

The summation in Eq. (2) is taken over all n atoms, including the hydrogens. 3DW may then be regarded as a metric analogue of the so-called distance degrees or distance sums introduced earlier for the integer graph distances [4,25] (though distance degrees are normally of use in hydrogen depleted graphs). The exponent $d(i,k)^{-2}$ was selected from a series of terms approximating the attracting interatomic potentials.

To better illustrate the method, we present in Table 1 the optimized MIS coordinates of 2,3-dimethyl butane:

```
                       19
                      /
       4   18 - C17 - 20   12        14
      /        /          /         /
3 - C1    -   C5    -   C11    -   C13  - 15
   /         /          /         /
  2         6    8 - C7 - 10    16
                     /
                    9
```

The $3DW_i$ weights of the carbon atoms in this molecule, calculated from the atomic MIS coordinates and Eq. (2), are 16.78 for C1, C7, C13, and C17, and 17.79 for C5 and C11, respectively. By substituting these values into Eqs. (3–7) (see below), one obtains the values of the five MIS indices for 2,3-dimethyl butane given in Table 2.

The atomic 3D weights were used instead of the vertex degrees a_i to redefine those of the topological indices that are based on a_i. These weights may be regarded as a further 3D generalization of the 2D graph-theoretic indices following the earlier development of the 3D Wiener index [17,18]. We performed a 3D modification of the two Zagreb indices M1 and M2 [26], the kth order molecular connectivity indices of Randić [27], Kier and Hall $^{k}\chi$ [9,28], and the weighted self-returning walks index SRW2 [29,30]. The MIS-based 3D analogues of these topological indices will be generally denoted by 3DTI. Their formulas are shown below (zero-order and first-order molecular connectivities were used in this paper):

$$3DM1 = \sum_{i=1}^{n} (3DW_i)^2 \tag{3}$$

$$3DM2 = \sum_{\text{all edges}} (3DW_i\, 3DW_j) \tag{4}$$

$$3D^0\chi = \sum_{i=1}^{n} (3DW_i)^{-1/2} \tag{5}$$

$$3D^1\chi = \sum_{\text{all edges}} (3DW_i\, 3DW_j)^{-1/2} \tag{6}$$

$$3DSRW2 = 2 \sum_{\text{all edges}} (3DW_i\, 3DW_j)^{1/2} \tag{7}$$

The values of the five MIS indices for the 73 C3–C9 alkanes in the test series are shown in Table 2.

3. MIS models for the boiling points of alkanes

Alkanes are convenient objects for testing new structural descriptors, due to the lack of electronic effects caused by heteroatoms or/and conjugation. The alkane series selected for testing our MIS models contained all 73 C3–C9 compounds from the paper of Basak et al. [8] with six of the boiling points corrected by Herndon [31] who verified all alkane boiling points

Table 2
Values of the five MIS indices defined by Eqs. (3)–(7)

No.	Compound	$3D^0\chi$	$3D^1\chi$	3DM1	3DM2	3DSRW2
1	C3	1.1210	0.2763	154	104	28.9
2	2M–C3	1.2320	0.2791	445	346	64.4
3	n–C4	1.2370	0.2841	438	334	63.3
4	2,2–MMC3	1.3420	0.2806	966	813	114.1
5	2M–C4	1.3460	0.2858	953	783	112.0
6	n–C5	1.3520	0.2903	936	760	110.3
7	2,2–MMC4	1.4480	0.2859	1771	1530	174.9
8	2,3–MMC4	1.4510	0.2877	1759	1510	173.8
9	2M–C5	1.4570	0.2916	1728	1471	171.5
10	3M–C5	1.4550	0.2905	1740	1483	172.2
11	n–C6	1.4620	0.2950	1704	1437	169.5
12	2,2,3–MMMC4	1.5480	0.2882	2933	2602	249.9
13	2,2–MMC5	1.5540	0.2912	2888	2548	247.3
14	3,3–MMC5	1.5500	0.2898	2918	2575	248.5
15	2,3–MMC5	1.5540	0.2920	2886	2534	246.6
16	2,4–MMC5	1.5570	0.2932	2864	2514	245.6
17	2M–C6	1.5620	0.2962	2827	2464	243.1
18	3M–C6	1.5590	0.2950	2846	2484	244.1
19	3E–C5	1.5570	0.2939	2865	2503	245.1
20	n–C7	1.5660	0.2989	2796	2420	241.0
21	2,2,3,3–MMMMC4	1.6420	0.2892	4518	4101	338.9
22	2,2,3–MMMC5	1.6470	0.2923	4461	4018	335.4
23	2,3,3–MMMC5	1.6450	0.2917	4482	4034	336.0
24	2,2,4–MMMC5	1.6490	0.2931	4438	3995	334.4
25	2,2–MMC6	1.6540	0.2958	4383	3924	331.4
26	3,3–MMC6	1.6500	0.2941	4429	3970	333.4
27	3,3–MEC5	1.6460	0.2927	4467	4007	334.9
28	2,3,4–MMMC5	1.6480	0.2934	4446	3986	334.1
29	2,3–MMC6	1.6540	0.2962	4384	3913	331.0
30	2,3–MEC5	1.6500	0.2946	4424	3954	332.7
31	2,4–MMC6	1.6540	0.2963	4379	3908	330.8
32	2,5–MMC6	1.6580	0.2977	4344	3872	329.2
33	2–MC7	1.6610	0.2999	4304	3816	326.9
34	3–MC7	1.6590	0.2988	4330	3845	328.1
35	4–MC7	1.6560	0.2985	4336	3853	328.4
36	3–EC6	1.6560	0.2975	4360	3878	329.5
37	n–C8	1.6650	0.3020	4266	3763	324.6
38	2,2,3,3–MMMMC5	1.7350	0.2925	6532	5989	437.7
39	2,2,3,4–MMMMC5	1.7360	0.2937	6502	5940	435.9
40	2,2,3–MMMC6	1.7420	0.2962	6414	5838	432.2
41	2,2,3–MMEC5	1.7390	0.2950	6459	5887	434.0
42	2,3,3,4–MMMMC5	1.7350	0.2931	6522	5961	436.7
43	2,3,3–MMMC6	1.7400	0.2956	6443	5864	433.1
44	2,3,3–MMEC5	1.7360	0.2939	6507	5932	435.6
45	2,2,4,4–MMMMC5	1.7320	0.2918	6577	6013	438.6
46	2,2,4–MMMC6	1.7430	0.2963	6410	5834	432.0
47	2,4,4–MMMC6	1.7410	0.2959	6430	5850	432.6
48	2,2,5–MMMC6	1.7460	0.2977	6360	5780	430.0
49	4,4–MMC7	1.7450	0.2975	6379	5790	430.4
50	3,3–EEC5	1.7370	0.2946	6499	5904	434.6
51	2,3,4–MEMC5	1.7400	0.2957	6447	5858	432.9
52	2,3,5–MMMC6	1.7450	0.2977	6372	5780	430.0
53	2,3–MEC6	1.7420	0.2964	6427	5833	432.0

Table 2 (*continued*)

No.	Compound	$3D^0\chi$	$3D^1\chi$	3DM1	3DM2	3DSRW2
54	3,4-EMC6	1.7440	0.2974	6392	5793	430.5
55	2,4-MEC6	1.7470	0.2987	6348	5743	428.7
56	3,4-MMC6	1.6520	0.2954	4406	3935	331.9
57	n-C9	1.7590	0.3046	6169	5521	420.3
58	2-MC8	1.7560	0.3029	6214	5583	422.7
59	3-MC8	1.7540	0.3019	6248	5620	424.0
60	4-MC8	1.7530	0.3016	6258	5634	424.6
61	3-EC7	1.7510	0.3008	6287	5664	425.7
62	4-EC7	1.7510	0.3005	6295	5676	426.1
63	2,2-MMC7	1.7500	0.2995	6308	5711	427.4
64	2,3-MMC7	1.7490	0.2997	6315	5704	427.2
65	2,4-MMC7	1.7490	0.2995	6313	5709	427.4
66	2,5-MMC7	1.7510	0.3002	6291	5682	426.4
67	2,6-MMC7	1.7530	0.3011	6261	5648	425.1
68	3,3-MMC7	1.7450	0.2979	6371	5774	429.8
69	3,4-MMC7	1.7470	0.2988	6346	5740	428.5
70	3,5-MMC7	1.7480	0.2990	6337	5730	428.2
71	3,3-MEC6	1.7410	0.2961	6437	5846	432.4
72	3,3,4-MMMC6	1.7390	0.2951	6462	5885	433.9
73	2,3,4-MMMC6	1.7420	0.2964	6427	5833	432.0

against those of the original measurements. The new MIS indices were tested jointly with other topological indices and molecular descriptors included in the OASIS software pack [32,33]. Besides the five 2D indices used to generate Eqs. (3)–(7), these were the topological indices of Wiener [34] W and Hosoya [35] Z, the 3D Wiener [17,18] WG, their information-theoretic analogues [17,36] I_W, I_z, and I_{WG}, electropy [37] ϵ, the Balaban indices [38,39] J and D2, the maximum metric distance [17] D_{max}, molecular weight MW and the number of carbon atoms N_C. The models were obtained by a standard multivariate regression analysis; their significance was evaluated by the overall and partial Fischer ratios. Model verification was performed by the leave-one-out procedure. The best models obtained are presented below.

3.1. 1-variable models

The best two models are shown below by Eqs. (8) and (9). A comparison of the models obtained with all molecular descriptors used is made in Table 3.

Table 3
Statistics of the 1-variable regressions of the 22 molecular descriptors with the boiling point of the C3–C9 alkanes

Descriptor	*r*	*s*	*F*	Descriptor	*r*	*s*	*F*
$3D^0\chi$	0.9892	5.97	3238	$^1\chi$	0.9857	6.88	2424
MW, N_C	0.9844	7.17	2229	ϵ	0.9840	7.26	2168
I_{WG}	0.9807	7.96	1790	I_W	0.9786	8.40	1603
3DSRW2	0.9608	11.30	854	$^0\chi$	0.9569	11.84	771
SRW2	0.9445	13.40	586	WG	0.9410	13.80	549
3DM1	0.9351	14.45	494	W	0.9338	14.58	484
3DM2	0.9304	14.94	458	Z	0.8832	19.12	252
M1	0.8818	19.23	248	M2	0.8693	20.15	220
$3D^1\chi$	0.8541	21.20	191	D2	0.8050	24.19	131
I_Z	0.7950	24.73	122	L_{max}	0.6415	31.27	50
J	0.6392	31.35	49	I_χ	0.6137	32.19	43

Table 4
Experimental vs calculated (by Eq. (12)) boiling points

Structure	Bp	Bp (calc.)	Bp - Bp (calc.)
propane	-42.07	-42.14	0.07
n-butane	-0.50	-3.65	3.15
2-methylpropane	-11.73	-13.70	1.97
n-pentane	36.07	33.91	2.16
2-methylbutane	27.85	26.07	1.78
2,2-dimethylpropane	9.50	13.08	-3.58
n-hexane	68.74	68.04	0.70
2-methylpentane	60.27	60.92	-0.65
3-methylpentane	63.28	62.40	0.88
2,2-dimethylbutane	49.74	50.72	-0.98
2,3-dimethylbutane	57.99	57.17	0.82
n-heptane	98.43	97.80	0.63
2-methylhexane	90.05	91.41	-1.36
3-methylhexane	91.85	92.16	-0.31
3-ethylpentane	93.48	93.63	-0.15
2,2-dimethylpentane	79.20	81.94	-2.74
2,3-dimethylpentane	89.78	89.14	0.64
2,4-dimethylpentane	80.50	84.49	-3.99
3,3-dimethylpentane	86.06	85.67	0.39
2,2,3-trimethylbutane	80.88	82.87	-1.99
n-octane	125.66	123.92	1.74
2-methylheptane	117.65	117.51	0.14
3-methylheptane	118.93	119.01	-0.08
4-methylheptane	117.71	119.01	-1.30
3-ethylhexane	118.53	119.75	-1.22
2,2-dimethylhexane	106.84	108.79	-1.95
2,3-dimethylhexane	115.61	115.99	-0.38
2,4-dimethylhexane	109.43	111.89	-2.46
2,5-dimethylhexane	109.10	112.06	-2.96
3,3-dimethylhexane	111.97	112.52	-0.55
3,4-dimethylhexane	117.73	117.46	0.27
3-ethyl-2-methylpentane	115.65	116.01	-0.36
3-ethyl-3-methylpentane	118.26	116.05	2.21
2,2,3-trimethylpentane	109.84	111.93	-2.09
2,2,4-trimethylpentane	99.24	101.87	-2.63
2,3,3-trimethylpentane	114.76	114.18	0.58
2,3,4-trimethylpentane	113.47	112.24	1.23
2,2,3,3-tetramethylbutane	106.47	111.65	-5.18
n-nonane	150.80	146.41	4.39
2-methyloctane	143.26	140.75	2.51
3-methyloctane	144.18	142.22	1.96
4-methyloctane	142.48	141.49	0.99
3-ethylheptane	143.00	142.97	0.03
4-ethylheptane	142.10	142.97	-0.87
2,2-dimethylheptane	132.69	132.73	-0.04
2,3-dimethylheptane	140.50	139.20	1.30
2,4-dimethylheptane	133.50	135.11	-1.61
2,5-dimethylheptane	136.00	136.56	-0.56
2,6-dimethylheptane	135.21	135.28	-0.07
3,3-dimethylheptane	137.30	135.73	1.57
3,4-dimethylheptane	140.10	140.68	-0.58
3,5-dimethylheptane	136.00	137.31	-1.31
4,4-dimethylheptane	135.20	135.73	-0.53

Table 4 (*continued*)

Structure	Bp	Bp (calc.)	Bp – Bp (calc.)
3-ethyl-2-methylhexane	138.00	139.95	-1.95
4-ethyl-2-methylhexane	133.80	136.58	-2.78
3-methyl-3-ethylhexane	140.60	139.27	1.33
3-ethyl-4-methylhexane	140.40	141.23	-0.83
2,2,3-trimethylhexane	133.60	135.14	-1.54
2,2,4-trimethylhexane	126.54	127.09	-0.55
2,2,5-trimethylhexane	124.08	126.54	-2.46
2,3,3-trimethylhexane	137.68	137.39	0.29
2,3,4-trimethylhexane	139.00	137.66	1.34
2,3,5-trimethylhexane	131.34	132.81	-1.47
2,4,4-trimethylhexane	130.65	129.34	1.31
3,3,4-trimethylhexane	140.46	139.60	0.86
3,3-diethylpentane	146.17	143.00	3.17
2,2-dimethyl-3-ethylpentane	133.83	135.89	-2.06
2,3-dimethyl-3-ethylpentane	142.00	140.93	1.07
2,4-dimethyl-3-ethylpentane	136.72	136.20	0.52
2,2,3,3-tetramethylpentane	140.27	139.86	0.41
2,2,3,4-tetramethylpentane	133.01	131.39	1.62
2,2,4,4-tetramethylpentane	122.28	112.70	9.58
2,3,3,4-tetramethylpentane	141.55	139.06	2.49

$$\mathrm{Bp}\ (^\circ\mathrm{C}) = 279.25(\pm 4.91)3\mathrm{D}^0\chi - 348.94(\pm 8.08) \tag{8}$$

$$n = 73,\ r = 0.9892,\ s = 5.97,\ s' = 6.18,\ F = 3238$$

$$\mathrm{Bp}\ (^\circ\mathrm{C}) = 59.11(\pm 1.20)^1\chi - 104.51(\pm 4.41) \tag{9}$$

$$n = 73,\ r = 0.9857,\ s = 6.88,\ s' = 7.23,\ F = 2424$$

Here r is the correlation coefficient, s the standard deviation, s' the averaged standard deviation of the leave-one-out procedure, and F the Fischer ratio.

As seen, the $3\mathrm{D}^0\chi$ MIS index outperforms all other indices tested. Four of the five MIS indices show better statistics than their 2D analogues, the exception being Randić's connectivity index $^1\chi$, the champion of the topological indices.

3.2. 2-variable models

$$\mathrm{Bp}\ (^\circ\mathrm{C}) = 244.87(\pm 4.32)3\mathrm{D}^0\chi + 27.02(\pm 2.45)\mathrm{I}_z - 336.47(\pm 5.05) \tag{10}$$

$$n = 73,\ r = 0.9961,\ s = 3.63,\ s' = 3.80,\ F = 4428$$

$$\mathrm{Bp}\ (^\circ\mathrm{C}) = 34.35(\pm 2.18)^1\chi + 25.43(\pm 2.12)\mathrm{I_W} - 130.44(\pm 3.34) \tag{11}$$

$$n = 73,\ r = 0.9953,\ s = 3.96,\ s' = 4.18,\ F = 3722$$

The next three best models included the Randić molecular connectivity $^1\chi$ with 3D Wiener index WG (r = 0.9951, s = 4.07, s' = 4.28, F = 3520), the Zagreb M2 index in combination with the $3\mathrm{D}^1\chi$ index (r = 0.9947, s = 4.21, s' = 4.36, F = 3289), and molecular weight MW combined with the information-theoretic analogue of the Hosoya index I_z (r = 0.9946, s = 4.27, s' = 4.54, F = 3199).

Once again, the best model included the $3\mathrm{D}^0\chi$ MIS index. The leave-one-out procedure produced exactly the same averaged correlation coefficients and averaged standard deviations s' that are only slightly larger than those of the basic models.

3.3. 3-variable models

$$\mathrm{Bp}\ (^\circ\mathrm{C}) = 727.26(\pm 20.76)3\mathrm{D}^0\chi - 19.46(\pm 0.91)3\mathrm{DSRW2} + 7.99(\pm 0.39)\mathrm{M2} - 779.42(\pm 20.08) \tag{12}$$

$$n = 73,\ r = 0.9986,\ s = 2.17,\ s' = 2.50,\ F = 8340$$

Table 5
Statistics for the best 4-variable models of alkane boiling point

Variables	r	s	s'	F
I_W, M1, M2, D2	0.9994	1.41	1.69	14800
I_{WG}, M1, M2, D2	0.9994	1.46	1.66	13800
$^1\chi$, 3DSRW2, M2, D2	0.9992	1.67	1.82	10500
MW, ϵ, SRW2, M2	0.9991	1.81	2.02	9034
ϵ, SRW2, M1, D2	0.9990	1.84	2.13	8651

$$\begin{aligned}\mathrm{Bp}\ (^\circ\mathrm{C}) = {} & 650.01(\pm 18.58)3\mathrm{D}^0\chi \\ & -73.39(\pm 3.73)^{o}\chi + 3.50(\pm 0.20)\mathrm{M2} \\ & -583.83(\pm 12.26) \qquad (13)\end{aligned}$$

$n=73$, $r=0.9984$, $s=2.32$, $s'=2.51$, $F=7304$

$$\begin{aligned}\mathrm{Bp}\ (^\circ\mathrm{C}) = {} & 61.60(\pm 1.52)^1\chi \\ & -0.0081(\pm 0.0008)3\mathrm{DM2} \\ & +1.690(\pm 0.099)\mathrm{M2} - 130.46(\pm 3.84) \\ & \qquad (14)\end{aligned}$$

$n=73$, $r=0.9974$, $s=3.01$, $s'=3.27$, $F=4329$

The next best combinations of three variables are the Randić molecular connectivity $^1\chi$ with the MIS 3DM1 index and the Zagreb M2 index with r = 0.9973, s = 3.04, s' = 3.31, and F = 4276, and the MIS index 3D$^0\chi$ in combination with $^0\chi$ and SRW2, with r=0.9973, s = 3.05, s' = 3.24, and F = 4204.

As seen, all five best regressions include some of the MIS indices. The regressions do not change their correlation coefficients in the leave-one-out procedure, while the averaged standard deviations are larger than those of the respective initial models by only 0.2–0.3°C.

In Table 4, the boiling points calculated according to Eq. (12) are compared to the experimental values.

3.4. 4-variable models

The five models with the best statistics are compared in Table 5. They are highly significant, with correlation coefficients within the 0.9990–0.9994 range and standard deviations of 1.38°–1.83°C. The MIS index 3DSRW2 is contained in the third best model.

4. Discussion

As demonstrated in the foregoing and summarized in Table 6, the 3D weighting of the five vertex-degree-based topological indices offers opportunities for better quantitative structure–property (and, possibly, structure–activity) correlations.

Our results compare favorably with other QSPR studies that make use of multiple linear regressions for calculating alkane boiling points. (The use of non-linear models can additionally improve the correlation, as convincingly shown by Mihalić and Trinajstić [40].)

The same series of 73 alkanes plus ethane has been modeled by Basak et al. [8] using principal component analysis. The best three-parameter model obtained has r = 0.993, s = 5.7, and F = 1608. In the recent work of Gautzsch and Zinn [41], based on the specific group contribution method, the best model derived for 69 C4–C9 alkanes includes eight parameters with r = 0.9947 and s = 3.43. Our best 3-variable model with r = 0.9986, s = 2.17, and F = 8340 for the 73 C3–C9 alkane series compares favorably with both the above-mentioned models. This comparison indicates the advantage of using well selected overall topological and 3D parameters for structure-based molecular property calculations.

Commenting on the comparison with the Gautzsch and Zinn method, one should mention an essential difference from our approach. The first method deals with descriptors with high orthogonality whereas the topological and 3D indices are known to be (sometimes strongly) interdependent. Different authors approach this problem in different ways, e.g. by assuming a different threshold of ‘strong’ interrelation such as r = 0.8, 0.9, or 0.95 [42]. Some of our best models also contain parameters with pairwise correlation higher than 0.95. These refer to one of the

Table 6
Comparison of the best linear regressions with and without MIS indices

Variables	r	s	F
1 variable			
(a) $3D^0\chi$	0.9892	5.97	3238
(b) $^1\chi$	0.9857	6.88	2424
2 variables			
(a) $3D^0\chi$, I_z	0.9961	3.63	4428
(b) $^1\chi$, I_w	0.9953	3.97	3722
3 variables			
(a) $3D^0\chi$, 3DSRW2, M2	0.9986	2.17	8340
(b) D2, I_z, J	0.9967	3.36	3464
4 variables			
(a) $^1\chi$, 3DSRW2, M2, D2	0.9992	1.67	10 500
(b) I_w, M1, M2, D2	0.9994	1.41	14 800

five 2-variable models ($^1\chi$, WG), three of the five 3-variable models (SRW2, $3D^0\chi$; $3D^0\chi$, $^0\chi$), and all five best 4-variable models (M1, M2; ϵ, SRW2; ϵ, MW; $^1\chi$, 3DSRW2). The increased involvement of interdependent parameters in this sequence seems in parallel with Herndon's conjecture [43] that one cannot explain the entire variance in the experimental data when dealing only with independent parameters.

We suppose that the potential difficulties related to the use of highly intercorrelating parameters are not in the physical interpretability of models, but rather in the unstable predictive patterns of such models. This quality is, however, easy to control by applying the leave-one-out verification procedure (or any cross-validation technique), and our 3D-weighted models passed this test successfully. On the other hand, there should be no strong objection against the use of intercorrelating parameters because such pairs of parameters can be orthogonalized (Randić recently advocated the use of orthogonalized parameters and proposed an original orthogonalizing procedure) [44,45]. It is essentially the same either to use a pair of dependent variables that explain a certain variance in the experimental data, or to separate the variance explained by the first parameter from the small additional part explained only by the second orthogonalized parameter.

5. Validation of the approach with other alkane properties

Boiling points seem to be the property most frequently used for testing QSPR techniques. Thus, five papers have been published on this topic in the *Journal of Chemical Information and Computer Sciences* during 1994 only, while in 1992 Horvath [46] listed 36 (!) different methods for calculating boiling points. Indeed, our approach needs further testing on other properties and series of compounds. This is neither possible nor needed to be done in a single paper. Yet, for a further validation, we present several illustrative examples of our MIS modeling of other alkane properties. The data used for the heats of formation and heats of atomization are those used by Kier and Hall [47] for correlations with their molecular connectivity indices, whereas the molar volumes and heats of vaporization are taken from Edward [48]. These are basically smaller series, comprising all 38 C3–C8 or slightly more alkanes.

5.1. Heats of atomization

The best 4-variables model contains the MIS analog of the Zagreb index M2:

$$\begin{aligned}\Delta H_{\mathrm{at}} = {} & 267.76(\pm 0.82)N_{\mathrm{C}} + 5.03(\pm 0.29)\mathrm{SRW2} \\ & -0.0024(\pm 0.0007)\mathrm{3DM2} \\ & +0.1157(\pm 0.0405)W \\ & -1.51(\pm 0.14)\mathrm{M2} + 133.68(\pm 1.62)\end{aligned} \qquad (15)$$

$n=38,\ r=1.0000,\ s=0.40,\ s'=0.48,\ F=8.64\times 10^6$

5.2. Heats of vaporization

Three of the five MIS indices are included in the best model with six variables:

$$\begin{aligned}\Delta H_{\mathrm{v}}\ (\mathrm{kkal\ mol^{-1}}) = {} & -0.0565(\pm 0.0069)\mathrm{W} \\ & -0.0501(\pm 0.0050)\mathrm{3DSRW2} \\ & +0.0340(\pm 0.0023)\mathrm{3DM1} \\ & -0.0330(\pm 0.0021)\mathrm{3DM2} \\ & +1.552(\pm 0.036)\mathrm{D2} \\ & +0.1397(\pm 0.0141)\mathrm{M2} + 1.0874(\pm 0.0250)\end{aligned} \qquad (16)$$

$n=47,\ r=0.9985,\ s=0.088,\ s'=0.10,\ F=2251$

5.3. Heats of formation

The 2-variable model with the best statistics incorporates the Randić first-order connectivity index and its zero-order MIS counterpart:

$$\Delta H_{\mathrm{v}}\ (\mathrm{kkal\ mol}^{-1}) = 98.77(\pm 3.46)3\mathrm{D}^{0}\chi - 11.23(\pm 0.76)^{1}\chi - 70.58(\pm 3.02) \qquad (17)$$

$n=38$, $r=0.9964$, $s=0.62$, $s'=0.70$, $F=2422$

5.4. Molar volumes

The same zero-order MIS connectivity index was found to be the parameter that correlated best with molar volumes:

$$V_{\mathrm{M}}\ (\mathrm{ml\ mol}^{-1}) = 155.29(\pm 3.38)3\mathrm{D}^{0}\chi - 95.13(\pm 5.52) \qquad (18)$$

$n=46$, $r=0.9897$, $s=2.50$, $s'=2.62$, $F=2104$

Models (15–18), as well as many others not shown here, confirm our finding for boiling points that the studied 3D analogues of some of the most frequently used 2D indices are always among the parameters included in the best regression equations. In the light of the above we may conclude that our MIS models (based on intramolecular atom–atom repulsion) could compete well, in the simulation and modeling of molecular properties, with other QSPR/ QSAR methods.

Acknowledgements

The comments made by W.C. Herndon (El Paso) are highly appreciated. D. Bonchev was supported by the Robert Welch Foundation, Texas. A. Toropov gratefully acknowledges the support of the Vostok Innovation Company (D.V. Rikov, President).

References

[1] A.T. Balaban, I. Motoc, D. Bonchev, O. Mekenyan, Topics Curr. Chem. 114 (1983) 21.
[2] N. Trinajstić, Chemical Graph Theory, 2nd edn, CRC Press, Boca Raton, FL, 1992.
[3] M. Randić, N. Trinajstić, J. Mol. Struct. 300 (1993) 551.
[4] D. Bonchev, N. Trinajstić, Int. J. Quantum Chem. Symp. 16 (1982) 463.
[5] D. Bonchev, Information-Theoretic Indices for Characterization of Chemical Structures, Research Studies Press, Chichester, UK, 1983.
[6] P.G. Seybold, M. May, U.A. Bagal, J. Chem. Educ. 64 (1987) 575.
[7] S.C. Basak, G.J. Niemi, G.D. Veith, J. Math. Chem. 4 (1990) 185.
[8] S.C. Basak, G.J. Niemi, G.D. Veith, J. Math. Chem. 7 (1991) 243.
[9] L.B. Kier, L.H. Hall, Molecular Connectivity in Chemistry and Drug Research, Academic Press, New York, 1976; Molecular Connectivity in Structure–Activity Analysis, Research Studies Press, Chichester, UK, 1986.
[10] O. Mekenyan, S.C. Basak, in D. Bonchev, O. Mekenyan (Eds.) Graph Theoretical Approaches to Chemical Reactivity, Kluwer Academic, Dordrecht, The Netherlands, 1994, p. 221.
[11] A.T. Balaban, A. Chiriac, I. Motoc, Z. Simon, Steric Fit in QSAR, Lecture Notes in Chemistry, No. 15, Springer, Berlin, 1980.
[12] M. Randić, J. Math. Chem. 9 (1992) 97.
[13] A.R. Katritzky, E.V. Gordeeva, J. Chem. Inf. Comput. Sci. 33 (1993) 835.
[14] M. Randić, Int. J. Quantum Chem. Symp. 15 (1988) 201.
[15] M. Randić, J. Chem. Inf. Comput. Sci. 34 (1994) 277.
[16] M. Randić, J. Chem. Inf. Comput. Sci. (in press).
[17] O. Mekenyan, D. Peitchev, D. Bonchev, N. Trinajstić, I. Bangov, Drug Design 36 (1986) 176.
[18] B. Bogdanov, S. Nikolić, N. Trinajstić, J. Math. Chem. 3 (1989) 299.
[19] A.A. Toropov, B.G. Ishakov, R.A. Muftahov, T. Ismailov, A.T. Mamadalimov, Russ. J. Phys. Chem. 66 (1992) 1074.
[20] A.A. Toropov, A.F. Toropova, R.A. Muftahov, T. Ismailov, A.G. Muftahov, Russ. J. Phys. Chem. 68 (1994) 577.
[21] R.J. Gillespie, Molecular Geometry, Addison-Wesley, New York, 1972.
[22] D.L. Keppert, Inorganic Stereochemistry, Springer, New York, 1972.
[23] G.M. Crippen, A.S. Smellie, J.W. Peng, J. Chem. Inf. Comput. Sci. 28 (1988) 125.
[24] T.F. Havel, in: Encyclopedia of NMR, Wiley, New York, 1995.
[25] D. Bonchev, A.T. Balaban, O. Mekenyan, J. Chem. Inf. Comput. Sci. 20 (1980) 106.
[26] I. Gutman, B. Ruščic, N. Trinajstić, C.W. Wilcox Jr, J. Chem. Phys. 69 (1975) 3399.
[27] M. Randić, J. Am. Chem. Soc. 97 (1975) 6609.
[28] L.B. Kier, L.H. Hall, W.J. Murray, M. Randić, J. Pharm. Sci. 64 (1975) 1971.
[29] D. Bonchev, L.B. Kier, J. Math. Chem. 9 (1992) 75.
[30] D. Bonchev, X. Liu, D.J. Klein, Croat. Chem. Acta 66 (1993) 141.
[31] W.C. Herndon (private communication).
[32] O. Mekenyan, S. Karabunarliev, D. Bonchev, Comput. Chem. 14 (1990) 193.

[33] D. Bonchev, C.F. Mountain, W.A. Seitz, A.T. Balaban, J. Med. Chem. 36 (1993) 1562.

[34] H. Wiener, J. Am. Chem. Soc., 69 (1947) 17; 69 (1947) 2636.

[35] H. Hosoya, Bull. Chem. Soc. Jpn 44 (1971) 2332.

[36] D. Bonchev, N. Trinajstić, J. Chem. Phys. 67 (1977) 4517.

[37] W.Y. Yee, K. Sakamoto, Y.J. I'Haya, Rep. Univ. Electro-comm., 27 (1976) 53; K. Sakamoto, W.Y. Yee, Y.J. I'Haya, Rep. Univ. Electro-comm., 27 (1977) 227.

[38] A.T. Balaban, Chem. Phys. Lett. 89 (1982) 399.

[39] A.T. Balaban, Theor. Chim. Acta 53 (1979) 355.

[40] Z. Mihalić, N. Trinajstić, J. Chem. Educ. 69 (1992) 701.

[41] R. Gautzsch, P. Zinn, J. Chem. Inf. Comput. Sci. 34 (1994) 791.

[42] L.M. Egolf, M.D. Wessel, P.C. Jurs, J. Chem. Inf. Comput. Sci. 34 (1994) 941.

[43] W.C. Herndon (private communication).

[44] M. Randić, J. Chem. Inf. Comput. Sci. 31 (1991) 311.

[45] M. Randić, New J. Chem. 15 (1991) 517.

[46] A.L. Horvath, Molecular Design: Chemical Structure Generation from the Properties of Pure Organic Compounds, Elsevier, Amsterdam, 1992.

[47] L.B. Kier, L.H. Hall, Molecular Connectivity in Structure–Activity Analysis. Research Studies Press, Chichester, UK, 1986.

[48] J.T. Edward, Can. J. Chem. 60 (1982) 480.