Distance-Related Indexes in the Quantitative Structure−Property Relationship Modeling†

Bono Lučić,‡ István Lukovits,§ Sonja Nikolić,*,‡ and Nenad Trinajstić‡

The Rugjer Bošković Institute, P.O. Box 180, HR-10002 Zagreb, Croatia, and Chemical Research Center, Hungarian Academy of Sciences, P.O. Box 17, H-1525 Budapest, Hungary

Received June 30, 2000

A comparative study of structure−boiling point modeling for a set of 180 acyclic and cyclic hydrocarbons (DS-180) and two of its subsets (one containing a selection of 76 acyclic and cyclic alkanes (DS-76), and the other containing 104 (DS-104) mono- and polycyclic butanes through octanes) using several known and novel distance-related indices is reported. The distance-related indices used were as follows: Wiener index, hyper-Wiener index, detour index, hyper-detour index, Harary index, Pasaréti index, Vérhalom index, Wiener-sum index, inverse Wiener-sum index and the product-form version of the Wiener index. Additional indices used were the total number of paths, the Hosoya *Z* index, the total walk count index, the number of carbon atoms, and the number of rings in the hydrocarbon. The best models for predicting the boiling points of 76, 104, and 180 acyclic and cyclic alkanes contain the natural logarithm of the cross-products of the Hosoya and detour index and of the Pasaréti index and the number of rings. This result extends earlier work by us[2,7] and Rücker and Rücker[8] on the use of the Wiener, detour, and Hosoya indices in modeling boiling points of alkanes and cycloalkanes. It also supports later work by Rücker and Rücker[9] on the use of the descriptor combination for the same purpose.

INTRODUCTION

In recent years a number of distance-related indices have been proposed in the literature.[1−3] There is also a vast literature on the oldest distance-related index, the Wiener index.[4−6] Therefore, it is of interest to investigate how these novel distance-related indices compare with the Wiener index and some other topological indices used in the quantitative structure−property relationship (QSPR) modeling. We decided to carry out this analysis using boiling points of a set of 180 lower acyclic and cyclic hydrocarbons since for them a careful analysis using a combination of the Wiener and detour indices, the walk-based topological index, and the Hosoya index was already carried out.[2,7−9]

To simplify our analysis we will use the graph theoretical terminology[10] in referring to studied hydrocarbons and their structural characteristics because the topological indices that we will consider here were obtained using (chemical) graph theoretical concepts.[11]

DEFINITIONS OF CONSIDERED DISTANCE-RELATED INDICES

Wiener Index. The Wiener index[12] $W = W(G)$ of a molecular graph G is defined[13] as the half-sum of the off-diagonal elements of the molecular distance matrix $\mathbf{D} = \mathbf{D}(G)$:[14]

$$W = (1/2)\sum_{i=1}^{N}\sum_{j=1}^{N}(D)_{ij} \qquad (1)$$

where $(D)_{ij}$ represents the length of a shortest path between vertexes i and j of G. There are also other definitions of the Wiener number available in the literature[4,5] and several proposals for computing the distance matrix.[15]

Hyper-Wiener Index. This index, denoted as WW = WW(G), was introduced by Randić.[16] However, Randić's algorithm for computing the hyper-Wiener index could be applied only to acyclic structures. It was later shown that WW can be computed for all structures as follows:[1,17,18]

$$\mathrm{WW} = (1/4)\sum_{i=1}^{N}\sum_{j=1}^{N}[(D)_{ij} + (D^2)_{ij}] \qquad (2)$$

where the summation goes over all pairs of vertexes i and j.

Detour Index. The detour index[7] $\omega = \omega(G)$ of a molecular graph G is defined in a way similar to the Wiener index; that is, the detour index is equal to the half-sum of the off-diagonal elements of the molecular detour matrix $\Delta = \Delta(G)$:[2,7,19−21]

$$\omega = (1/2)\sum_{i=1}^{N}\sum_{j=1}^{N}(\Delta)_{ij} \qquad (3)$$

where $(\Delta)_{ij}$ represents the length of a longest path between vertexes i and j of G. Several methods were proposed for computing the detour matrix in the literature[2,8,19,21−24]

Hyper-Detour Index. Replacing $(D)_{ij}$ by $(\Delta)_{ij}$ in eq 2, one immediately obtains the hyper-detour index $\omega\omega = \omega\omega(G)$:[7,25]

$$\omega\omega = (1/4)\sum_{i=1}^{N}\sum_{j=1}^{N}[(\Delta)_{ij} + (\Delta^2)_{ij}]/2 \qquad (4)$$

Harary Index. The Harary index[26,27] $H = H(G)$ of a molecular graph G is defined as the half-sum of the off-

* To whom correspondence should be addressed. E-mail: sonja@rudjer.irb.hr.

† Reported in part at the Second Indo-U.S. Workshop on Mathematical Chemistry held during May 30−June 3, 2000 at the University of Minnesota−Duluth, Duluth, MN.

‡ The Rugjer Bošković Institute.

§ Hungarian Academy of Sciences.

10.1021/ci0000777 CCC: $20.00
Published on Web 01/09/2001

diagonal elements of the reciprocal molecular distance matrix $\mathbf{D}^r = \mathbf{D}^r(G)$:

$$H = (1/2)\sum_{i=1}^{N}\sum_{j=1}^{N}(D^r)_{ij} \quad (5)$$

The reciprocal distance matrix $\mathbf{D}^r$ can be obtained by replacing all off-diagonal elements of the distance matrix $(D)_{ij}$ by their reciprocals:

$$(D^r)_{ij} = 1/(D)_{ij}; \quad i \neq j \quad (6)$$

while diagonal elements $(D^r)_{ii}$ are equal to zero by definition.

Pasaréti Index. The Pasaréti index $P = P(G)$ is an all-path version of the Wiener index.[28] It is defined as follows:

$$P = \sum_{i<j}\sum_{p_{ij}}|p_{ij}| \quad (7)$$

where p_{ij} denotes a path between vertexes i and j and $|p_{ij}|$ denotes the length of this path. The summation is between all pairs of vertexes i and j and for all paths between i and j. This index is called the Pasaréti index because it was derived in the home of one of us (I.L.) that is located in the part of Budapest called Pasarét.

Vérhalom Index. The Pasaréti index P is an exponential function of graph size in terms of the number of its vertexes (N). This property makes it unwieldy to use in the QSPR/QSAR modeling because it becomes soon too large a number in the homologous series of molecules. Therefore, the Pasaréti index was transformed into a new variant $V = V(G)$ that was called the Vérhalom index:[28]

$$V = P/k \quad (8)$$

where k is the total number of paths in G divided by $N(N-1)/2$. The name Vérhalom index is given to this distance-related index because it was originated in the Chemical Research Center of the Hungarian Academy of Sciences located in the district Vérhalom in Budapest. Both the Pasaréti index and the Vérhalom index prior to the present study have not been used in QSPR modeling.

Wiener Sum of the D/Δ Matrix. We call the Wiener sum of the $\mathbf{D}/\Delta$ matrix the Wiener-sum index $WS = WS(G)$ of a molecular graph G.[29] It is defined as the half-sum of the off-diagonal elements of the molecular quotient matrix $\mathbf{D}/\Delta$:[30]

$$WS = (1/2)\sum_{i=1}^{N}\sum_{j=1}^{N}(D)_{ij}/(\Delta)_{ij} \quad (9)$$

Randić used this index for a characterization of cyclic structures.[30]

Wiener Sum of the Δ/D Matrix. We call the Wiener sum of the $\Delta/\mathbf{D}$ matrix the inverse Wiener-sum index $ws = ws(G)$ of a molecular graph G. It is defined in the same way as the Wiener-sum index except the off-diagonal elements in the molecular quotient matrix are inverted:[29]

$$ws = (1/2)\sum_{i=1}^{N}\sum_{j=1}^{N}(\Delta)_{ij}/(\mathbf{D})_{ij} \quad (10)$$

This index and the Wiener-sum index have so far been used only in the structure−boiling point modeling for condensed benzenoid hydrocarbons.[29]

Product−Form Version of the Wiener Index. The multiplicative version of the Wiener index $\pi = \pi(G)$ of a molecular graph G is equal to the product of shortest distances between all pairs of vertexes in G:[31]

$$\pi = \prod_{i<j}(D)_{ij} \quad (11)$$

The π index is, even for small graphs, a big number; e.g., $\pi = 34\,560$ for *n*-hexane. Therefore, it is recommended that $\ln \pi$ be used instead of π in the QSPR modeling. The use of $\ln \pi$ in QSPR was so far limited to the structure−octane number study.[31]

DATA SETS AND COMPUTATIONAL METHODS

Data Sets. We carried out a comparative study of structure−boiling point modeling for a set of 180 lower acyclic and cyclic hydrocarbons. The set of 180 hydrocarbons (DS-180) was split into two subsets: (1) a subset of 76 lower alkanes and cycloalkanes (DS-76) and (2) a subset of 104 mono- and polycyclic butanes through octanes (DS-104). The composition of the set DS-180 was dictated by previous studies. It was Lukovits who first studied the set DS-76,[7] without giving a selection criterion for the compounds in his sample. Later the Zagreb group[21] and Rücker and Rücker[8] used the same sample, not because it was a nice sample, but, though it was a rather ill-defined sample, simply because it had been studied before. Rücker and Rücker,[8] in order to have a more consistent sample, added the set DS-104.

Structures of studied acyclic and cyclic hydrocarbons can be found in Figure 1 of ref 9 (pp 790−792). Therefore, we are not repeating these structures in our paper. Their distance-related indices computed by us and experimental boiling points, taken from Rücker and Rücker,[9] are given in Table 1.

In Table 1 we also give the total number of paths (TNP), the Hosoya Z index,[13] the total walk count index (twc),[9] the number of atoms n, and the number of rings n_r in the hydrocarbon. The values of Z for DS-76 appear in ref 2, while the values of Z for DS-104 were taken from ref 8. Some values in the cited sources were erroneous. They are corrected in this paper. The twc indices were taken from Rücker and Rücker.[9]

In addition to these 15 initial descriptors we have also used the natural logarithmic transformation of initial descriptors (except of descriptor $\ln \pi$, which was not logarithmically transformed) giving an additional 14 descriptors. Because some descriptors had zero values for certain molecules, we used logarithmic transformation of initial descriptors of the form $\ln(d_i + 1)$, where d_i represents the ith descriptor value. Moreover, by introducing the logarithms of the cross-products of 14 initial descriptors, we extended the descriptor set from 15 to 120 descriptors. Logarithmic transformation was chosen after we performed scatter plots BP vs each of 14 descriptors ($\ln \pi$ was not considered) and found that most plots indicated nonlinear logarithmic relations.

Computational Methods. The first set of models presented in this paper are "the best multivariate regression

Table 1. Calculated Distance Indexes for 180 Acyclic and Cyclic Hydrocarbons and Their Boiling Points

		distance index[b]														
hydrocarbon[a]	BP[a]	W	WW	ω	$\omega\omega$	H	P	V	WS	ws	TNP	ln π	Z	twc	n_r	n
n1	161.5	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
n2	88. 6	1	1	1	1	1.000	1	1.000	1.00 0	1.000	1	0.000	2	1	0	2
n3	42.1	4	5	4	5	2.500	4	4.00 0	3.00 0	3.000	3	0.693	3	5	0	3
n4	0.5	10	15	10	15	4.33 3	10	10.000	6.00 0	6.000	6	2.485	5	16	0	4
2mn3	11.7	9	12	9	12	4.500	9	9.00 0	6.00 0	6.000	6	2.079	4	18	0	4
n5	36.0	20	35	20	35	6.41 7	20	20.0 00	10.0 00	10.000	10	5.663	8	44	0	5
2mn4	27.8	18	28	18	28	6.66 7	18	18.0 00	10.0 00	10.000	10	4.970	7	53	0	5
22mn3	9.5	16	22	16	22	7.00 0	16	16.0 00	10.0 00	10.000	10	4.159	5	70	0	5
n6	68.7	35	70	35	70	8.70 0	35	35.0 00	15.0 00	15.000	15	10.450	13	111	0	6
3mn5	63.3	31	54	31	54	9.08 3	31	31.0 00	15.0 00	15.000	15	9.246	12	142	0	6
2mn5	60.3	32	58	32	58	9.00 0	32	32.0 00	15.0 00	15.000	15	9.534	11	134	0	6
23mn4	58.0	29	47	29	47	9.33 3	29	29.0 00	15.0 00	15.000	15	8.553	10	165	0	6
22mn4	49.7	28	44	28	44	9.50 0	28	28.0 00	15.0 00	15.000	15	8.148	9	185	0	6
n7	98.5	56	126	56	126	11.1 50	56	56.0 00	21.0 00	21. 00	21	17.030	21	268	0	7
3en5	93.5	48	90	48	90	11.7 50	48	48.0 00	21.0 00	21.000	21	14.909	20	378	0	7
3mn6	92.0	50	99	50	99	11.6 17	50	50.0 00	21.0 00	21.000	21	15.420	19	354	0	7
2mn6	90	52	108	52	108	11.4 83	52	52.0 00	21.0 00	21.000	21	15.931	18	329	0	7
22mn5	79.2	46	84	46	84	12.0 83	46	46.0 00	21.0 00	21.000	21	14.099	14	489	0	7
33mn5	86.1	44	76	44	76	12.2 50	44	44.0 00	21.0 00	21.000	21	13.523	16	526	0	7
24mn5	80.5	48	91	48	91	11.8 33	48	48.0 00	21.0 00	21.000	21	14.792	15	399	0	7
23mn5	89.8	46	83	46	83	12.0 00	46	46.0 00	21.0 00	21.000	21	14.216	17	436	0	7
223mn4	80.9	42	69	42	69	12.5 00	42	42.0 00	21.0 00	21.000	21	12.830	13	588	0	7
c4	12.6	8	10	16	30	5.00 0	24	12.0 00	3.33 3	14.000	12	1.386	7	28	1	4
1ec4	70.7	29	49	45	101	9.75 0	70	38.8 89	10.5 33	26. 333	27	8.148	17	240	1	6
1mc4	36.3	16	23	28	57	7.33 3	43	22.6 32	6.33 3	20.000	19	3.871	10	89	1	5
c7	118.4	42	70	105	322	12.8 33	147	73.5 00	9.217	68.833	42	12.542	29	441	1	7
c6	80.7	27	42	63	168	10.0 00	90	45.0 00	7.20 0	45.000	30	7.455	18	186	1	6
1bc6	180.9	133	323	217	737	21.6 43	330	185.625	30.2 97	89.067	80	41.198	114	7954	1	10
1ipc6	154.8	88	176	160	488	19.1 50	239	130. 364	22.4 48	78.667	66	27.662	62	4020	1	9
1sbc6	179.3	121	263	205	653	22.3 83	308	173. 250	29.8 76	90.467	80	38.336	106	1024 0	1	10
1tbc6	171.5	114	233	198	611	23.0 17	296	166. 500	29.6 38	91.333	80	36.321	80	1333 3	1	10
1m4ec6	150.8	90	187	162	487	19.0 67	247	130. 765	22.1 24	80.333	68	28.068	64	4020	1	9
1pc6	156.7	94	203	166	527	18.6 83	250	136. 364	22.6 86	77.800	66	29.272	70	3396	1	9
113mc6	136.6	82	152	158	462	19.7 50	235	124. 412	21.1 33	83. 000	68	25. 830	48	5396	1	9
124mc6	144.8	84	160	162	481	19.5 33	243	126. 783	20.8 95	83.833	69	26.458	56	4779	1	9
c8	149	64	120	160	552	15.6 67	224	112. 000	12.6 10	97.333	56	19.879	47	1016	1	8
c5	49.3	15	20	35	80	7.50 0	50	25.0 00	4.58 3	27.500	20	3.466	11	75	1	5
12ec5	150.5	87	171	151	447	19.2 00	230	125. 455	22.6 51	68.933	66	27.439	67	4814	1	9
11mc5	87.9	39	61	75	185	13.3 33	110	59.2 31	12.1 83	45.833	39	11.326	21	766	1	7
12mc5	95.6	40	64	79	203	13.1 67	117	61.4 25	11.6 83	46.833	40	11.731	24	690	1	7
1ec5	103.5	43	75	79	207	12.8 33	118	63.5 38	12.4 83	44.667	39	12.712	27	590	1	7
1mc5	71.8	26	39	54	131	10.1 67	79	40.8 62	7.88 3	36.167	29	7.049	16	225	1	6
1m2pc5	149.5	90	185	151	451	19.0 17	230	127. 385	23.2 89	67.583	65	28.173	64	4525	1	9
c3	32. 8	3	3	6	9	3.00 0	9	4.50 0	1.50 0	6.000	6	0.000	4	9	1	3
11mc3	20.6	15	20	22	38	7.50 0	33	19.4 12	7.16 7	15.000	17	3.466	8	116	1	5
12mc3	32.6	16	23	24	45	7.33 3	38	21.1 11	6.91 7	15.333	18	3.871	9	107	1	5
1ec3	35.9	17	26	24	46	7.16 7	37	21.7 65	7.33 3	14.667	17	4.277	10	93	1	5
1mc3	0.7	8	10	13	22	5.00 0	20	10.9 09	3.83 3	10.000	11	1.386	6	32	1	4
112mc3	52.6	26	39	37	71	10.1 67	58	33.4 62	11.0 00	21.667	26	7.049	12	377	1	6
n8	125.7	84	210	84	210	13.743	84	84.0 00	28.0 00	28.000	28	25.555	34	627	0	8
3mn7	118.9	76	170	76	170	14.2 67	76	76.0 00	28.0 00	28.000	28	23.609	31	838	0	8
3en6	118.5	72	150	72	150	14.4 83	72	72.0 00	28.0 00	28.000	28	22.693	32	928	0	8
34mn6	117.7	68	134	68	134	14.8 67	68	68.0 00	28.0 00	28.000	28	21.489	29	1136	0	8
3e3mn5	118.2	64	118	64	118	15.2 50	64	64.0 00	28.0 00	28.000	28	20.285	28	1441	0	8
4mn7	117.7	75	165	75	165	14.3 17	75	75.0 00	28.0 00	28.000	28	23.386	30	856	0	8
2mn7	117.6	79	185	79	185	14.1 00	79	79.0 00	28.0 00	28.000	28	24.302	29	764	0	8
3e2mn5	115.6	67	129	67	129	14.9 17	67	67.0 00	28.0 00	28. 000	28	21.266	28	1152	0	8
23mn6	115.6	70	143	70	143	14.7 33	70	70.0 00	28.0 00	28. 000	28	22.000	27	1068	0	8
24mn6	109.4	71	147	71	147	14.6 50	71	71.0 00	28.0 00	28.000	28	22.287	26	997	0	8
33mn6	112.0	67	131	67	131	15.0 33	67	67.0 00	28.0 00	28. 000	28	21.083	25	1301	0	8
25mn6	109.1	74	161	74	161	14.4 67	74	74.0 00	28.0 00	28. 000	28	23.021	25	911	0	8
234mn5	113.5	65	122	65	122	15.1 67	65	65.0 00	28.0 00	28. 000	28	20.572	24	1296	0	8
233mn5	114.8	62	111	62	111	15.5 00	62	62.0 00	28.0 00	28. 000	28	19.592	23	1609	0	8
22mn6	106.8	71	149	71	149	14.7 67	71	71.0 00	28.0 00	28.000	28	22. 105	23	1142	0	8
223mn5	109.8	63	115	63	115	15.4 17	63	63.0 00	28.0 00	28. 000	28	19.879	22	1536	0	8
224mn5	99.2	66	127	66	127	15.1 67	66	66.0 00	28.0 00	28. 000	28	20. 742	19	1317	0	8
2233 mn4	106.5	58	97	58	97	16.0 00	58	58.0 00	28.0 00	28. 000	28	18.205	17	2047	0	8

Table 1. (continued)

	distance index[b]															
hydrocarbon[a]	BP[a]	W	WW	ω	$\omega\omega$	H	P	V	WS	ws	TNP	ln π	Z	twc	n_r	n
11mc6	119.5	59	103	119	337	16.3 33	174	91.9 25	15.9 33	68. 667	53	18.087	34	1829	1	8
12mc6	126.6	60	106	124	362	16.1 67	182	94.3 70	15.3 62	70.000	54	18.493	39	1657	1	8
13mc6	122.3	61	110	123	355	16.0 83	182	94.3 70	15.6 00	69.167	54	18.781	37	1583	1	8
14mc6	121.8	62	115	122	349	16.0 33	182	94.3 70	15.9 33	68.667	54	19.004	38	1572	1	8
1ec6	131.8	64	122	124	368	15.7 83	184	97.2 08	16.2 57	67.000	53	19.697	44	1401	1	8
1pc5	131	67	135	111	315	15.5 67	168	94.0 80	18.2 93	53.567	50	20.495	43	1449	1	8
1ipc5	126.4	62	114	106	286	16.0 00	159	89.0 40	18.0 83	54.167	50	19.068	38	1701	1	8
112mc5	114	56	92	106	278	16.6 67	157	84.5 38	16.4 83	58.500	52	17.107	32	2368	1	8
113mc5	104.9	58	100	104	266	16.5 00	157	84.5 38	17.0 83	57.000	52	17.682	30	2211	1	8
bc110b	8	7	8	17	33	5.50 0	39	12.3 16	2.50 0	15.500	19	0.693	8	51	2	4
s22p	39	14	18	28	58	8.00 0	66	23.5 71	5.00 0	20.000	28	2.773	12	166	2	5
mbc110b	33.5	14	18	29	60	8.00 0	66	22.7 59	5.16 7	22.000	29	2.773	11	188	2	5
bc210p	46	14	18	37	88	8.00 0	82	25.6 25	3.91 7	30.000	32	2.773	13	150	2	5
bc11 1p	36	14	18	32	69	8.00 0	84	25.4 55	4.50 0	25.000	33	2.773	13	147	2	5
1e1mc3	57	27	42	36	68	10.0 00	53	33.125	11.6 67	20.667	24	7.455	14	331	1	6
1ipc3	58.3	28	45	37	73	9.83 3	56	35.0 00	11.8 33	20.333	24	7.860	14	282	1	6
123mc3	63	27	42	39	78	10.0 00	63	35.0 00	10.7 50	22.000	27	7.455	14	354	1	6
1e2mc3	63	29	49	40	84	9.75 0	64	36.9 23	11.2 17	21.250	26	8.148	15	301	1	6
1pc3	69	31	56	40	86	9.50 0	61	38.1 25	11.9 33	20.167	24	8.841	16	245	1	6
12mc4	62	27	42	45	99	10.0 00	70	37.5 00	9.93 3	27.667	28	7.455	15	278	1	6
13mc4	59	28	46	44	94	9.91 7	70	37.5 00	10.3 33	27.000	28	7.742	14	270	1	6
bcpr	76	27	43	45	103	10.3 33	103	41.7 57	9.06 7	25.667	37	7.1 67	20	403	2	6
s23hx	69.5	25	37	50	119	10.6 67	117	41.7 86	7.93 3	32.667	42	6.356	20	457	2	6
13mbcb	55	24	34	45	97	10.8 33	104	37.1 43	8.58 3	29.833	42	5.951	15	686	2	6
bc31 0hx	81	24	34	66	182	10.8 33	148	46.2 50	5.83 3	49.333	48	5.951	21	408	2	6
bc21 1hx	71	23	31	62	163	11.0 00	152	45.6 00	5.75 0	45.000	50	5.545	21	390	2	6
bc22 0hx	83	25	37	67	186	10.6 67	149	45.6 12	5.73 3	48.333	49	6.356	22	403	2	6
mbc210p	60.5	24	34	56	139	10.8 33	123	41.0 00	6.96 7	39.333	45	5.951	18	527	2	6
1sbc3	90.3	45	81	56	119	12.5 00	83	54.4 69	17.4 33	26.833	32	13.523	24	756	1	7
1122 mc3	76	39	61	54	107	13.3 33	85	49.5 83	15.8 33	29.333	36	11.326	16	1355	1	7
1tbc3	80.5	41	67	52	103	13.0 00	77	50.5 31	17.3 33	27. 000	32	12. 137	18	940	1	7
1123 mc3	78	40	64	56	114	13.1 67	90	51.0 81	15.5 83	29.667	37	11.731	19	1275	1	7
11ec3	88.6	43	74	54	108	12.7 50	77	50.5 31	17.1 67	27.333	32	12.830	24	956	1	7
12ec3	90	47	90	62	144	12.3 67	101	58.9 17	16.3 50	28.367	36	14.034	25	836	1	7
1bc3	98	51	106	62	148	11.9 83	93	61.0 31	17.6 00	26.567	32	15.238	26	615	1	7
1m1I pc3	81.5	41	67	52	101	13.0 00	75	49.2 19	17.1 67	27.333	32	12.137	20	1022	1	7
1e23 mc3	91	44	78	60	132	12.6 67	98	55.6 22	15.8 50	29.167	37	13.118	23	1001	1	7
1m2pc3	93	48	94	62	145	12.2 83	99	59.4 00	16.6 50	27.950	35	14.322	24	791	1	7
1ipc4	92.7	44	78	64	148	12.6 67	99	57.7 50	15.7 33	33.667	36	13.118	24	705	1	7
1e3mc4	89.5	46	87	66	153	12.5 33	107	59.1 32	15.5 33	34.333	38	13.629	24	732	1	7
1e2mc4	94	44	78	68	164	12.6 67	107	59.1 32	14.8 00	35.500	38	13.118	25	751	1	7
13mc5	91.3	41	68	78	197	13.0 83	117	61.4 25	11.9 83	46.083	40	12.019	23	670	1	7
1mc6	101	42	71	90	251	12.9 17	131	67.0 98	11.0 67	56.333	41	12.425	26	547	1	7
dcprm	102	46	87	68	171	12.8 33	150	67.0 21	14.3 33	32.333	47	13.405	32	987	2	7
11ms22p	78	38	59	64	147	13.8 33	154	58.8 00	13.0 67	35.667	55	10.633	24	1913	2	7
s24h	98.5	38	59	85	233	13.8 33	187	67.7 07	9.81 7	52.167	58	10.633	32	1216	2	7
s33h	96.5	40	66	80	214	13.5 83	192	67.2 00	11.4 00	47.667	60	11.326	33	1204	2	7
122mbcb	84	38	59	67	155	13.8 33	158	57.2 07	12.8 67	39.000	58	10.633	23	2330	2	7
14mbc210p	74	37	56	80	205	14.0 00	176	60.5 90	10.6 17	50.333	61	10.227	25	1894	2	7
13mbc111p	71.5	38	60	68	153	13.9 17	176	59.6 13	12.5 00	42.000	62	10.515	22	1367	2	7
mbc310hx	92	37	56	93	265	14.0 00	206	67.5 94	9.25 0	61.833	64	10.227	29	1433	2	7
6mbc310 hx	103	39	63	96	282	13.7 50	225	70.5 22	9.30 0	61.583	67	10.920	32	1320	2	7
2mbc310hx	100	38	59	94	271	13.8 33	216	69.7 85	9.35 0	61.333	65	10.633	31	1253	2	7
bc32 0h	110. 5	38	59	111	354	13.8 33	244	74.2 61	7.41 7	74.833	69	10.633	35	1037	2	7
mbc211hx	81.5	36	53	86	229	14.1 67	208	66.1 82	9.45 0	56.167	66	9.822	28	1327	2	7
bc221h	105. 5	36	53	103	312	14.1 67	250	73.9 44	7.50 0	68.500	71	9.822	34	987	2	7
bc410h	116	39	62	109	344	13.6 67	242	75.8 51	8.08 3	74.333	67	11.038	34	1063	2	7
bc311h	110	38	59	102	307	13.8 33	248	74.4 00	8.46 7	68.333	70	10.633	34	1003	2	7
tc311024h	107	35	51	106	329	14.6 67	400	77.7 78	7.15 0	73.000	108	9.129	41	1868	3	7
tc410024h	105	36	54	113	367	14.5 00	391	78.9 52	7.00 0	79.833	104	9.534	40	1867	3	7
tc410013h	107. 5	35	51	112	360	14.6 67	385	77.0 00	6.86 7	80.333	105	9.129	39	2234	3	7
tc221026h	106	33	45	111	354	15.0 00	444	79.6 92	6.30 0	78.000	117	8.318	41	1890	3	7
tc410027h	110	37	57	115	378	14.3 33	429	81.1 62	7.23 3	81.500	111	9.940	42	2113	3	7
tec410h	104	34	49	119	402	15.1 67	681	84.6 21	6.20 0	87.000	169	8.436	50	3138	4	7
tec320h	108. 5	32	43	125	435	15.5 00	762	85.5 72	5.40 0	92.500	187	7.625	51	3131	4	7
1122 3mc3	100. 5	56	92	77	160	16.6 67	124	70.8 57	21.1 67	38.667	49	17.107	26	4654	1	8
112m2ec3	104. 5	59	103	78	167	16.3 33	123	73.2 77	21.9 33	37.500	47	18.087	28	3967	1	8
11m2ipc3	94.4	62	114	81	182	16.0 00	130	77.4 47	22.2 00	37.000	47	19.068	28	3209	1	8
11m2pc3	105. 9	67	135	86	207	15.5 67	139	82.8 09	22.3 67	36.733	47	20.495	32	2860	1	8

Table 1. (continued)

	distance index[b]															
hydrocarbon[a]	BP[a]	W	WW	ω	$\omega\omega$	H	P	V	WS	ws	TNP	ln π	Z	twc	n_r	n ω
ib2mc3	110	69	142	86	209	15.3 17	136	84.6 22	23.0 83	35.650	45	21.189	33	2190	1	8
1nepec3	106	65	125	78	175	15.6 67	115	78.5 37	24.1 33	34.167	41	20.049	28	2116	1	8
1e2pc3	108	72	156	91	233	15.0 67	149	88.7 66	22.6 40	36.233	47	22.000	40	2176	1	8
b2mc3	124	74	165	91	235	14.9 33	144	89.6 00	23.1 74	35.517	45	22.510	39	1997	1	8
1spc3	117. 7	69	141	82	191	15.2 33	119	81.2 68	24.1 00	34.233	41	21.306	38	1912	1	8
5msbc3	115. 5	64	120	77	168	15.6 67	112	76.4 88	24.0 33	34.333	41	19.879	34	2172	1	8
1pec3	128	78	183	91	239	14.6 00	134	91.5 12	24.3 14	33.900	41	23.609	42	1493	1	8
1133 mc4	86	60	108	84	184	16.3 33	136	76.1 60	21.3 33	44.000	50	18.257	23	3018	1	8
1234 mc4	114. 5	60	106	92	218	16.1 67	148	79.6 92	19.7 33	46.667	52	18.493	34	2780	1	8
12ec4	119	66	129	98	257	15.5 33	156	87.3 60	20.3 81	44.733	50	20.390	42	2018	1	8
1sbc4	123	66	129	90	221	15.5 33	138	84.0 00	22.0 67	41.667	46	20.390	41	1837	1	8
p3mc4	117. 4	71	153	95	241	15.2 33	155	88.5 71	21.8 67	42.333	49	21.594	38	1829	1	8
123mc5	117	58	99	109	290	16.4 17	164	86.6 42	16.2 83	58.750	53	17.800	36	2100	1	8
124mc5	115	59	103	108	284	16.3 33	164	86.6 42	16.5 83	58.000	53	18.087	34	2033	1	8
1e1mc5	121. 5	59	103	103	267	16.3 33	152	85.1 20	17.7 83	55.333	50	18.087	37	2071	1	8
1e2mc5	124. 7	61	110	111	307	16.0 83	167	89.9 23	16.8 55	57.083	52	18.781	40	1820	1	8
1e3mc5	121	63	119	109	294	15.9 50	167	89.9 23	17.4 17	55.783	52	19.291	39	1736	1	8
1mc7	134	61	109	142	451	16.0 00	202	102. 836	13.3 88	83.333	55	18.898	42	1279	1	8
bcprm	129	72	157	98	269	15.4 67	208	100. 414	20.7 24	39.933	58	21.635	52	2385	2	8
bcb	136	64	124	112	324	16.2 00	265	101. 644	17.0 48	55.333	73	19.291	58	2571	2	8
s25o	125	58	101	131	403	16.8 33	279	102. 789	13.4 43	75.500	76	17.394	52	3123	2	8
s34o	128	57	98	125	378	17.0 00	290	101. 500	13.6 60	70.000	80	16.989	53	3043	2	8
1223 mbcb	105	54	87	93	222	17.3 33	220	80.0 00	17.4 83	50.167	77	16.008	33	8886	2	8
2244 mbcb	104	59	106	93	223	16.8 33	247	85.3 83	18.7 00	47.833	81	17.564	32	7613	2	8
33mbc310hx	115	56	95	124	361	17.1 67	288	97.1 57	14.1 67	73.500	83	16.583	39	3864	2	8
1mbc410h	125	56	94	146	473	17.0 83	320	104. 186	11.8 88	90.250	86	16.701	47	3772	2	8
7mbc410h	138	58	101	150	499	16.8 33	347	107. 956	11.8 45	90.000	90	17.394	52	3452	2	8
bc510o	141	58	100	165	583	16.7 50	369	116. 090	10.7 98	104 .91 7	89	17.512	55	2692	2	8
2mbc320h	130. 5	56	94	150	496	17.0 83	338	105. 156	11.1 60	90.250	90	16.701	52	3122	2	8
bc420o	133	58	101	168	600	16.8 33	372	113. 217	10.2 00	105.000	92	17.394	57	2593	2	8
14mbc211hx	91	53	85	115	310	17.5 83	277	91.2 47	13.9 50	68.583	85	15.485	37	4414	2	8
bc33 0o	137	55	91	171	617	17.2 50	373	112. 301	9.31 7	107.917	93	16.296	56	2571	2	8
1mbc221h	117	52	81	135	410	17.6 67	324	100. 800	11.5 17	82.667	90	15.197	46	3280	2	8
2mbc221h	125	54	88	141	447	17.4 17	349	105. 075	11.2 67	83.250	93	15.890	50	2962	2	8
7mbc221h	128	53	84	142	455	17.5 00	355	105. 745	11.0 14	83.833	94	15.603	52	3068	2	8
ds2121o	103	58	104	108	292	17.3 33	370	99.6 15	15.2 00	57.333	104	16.871	56	6014	3	8
ds2022o	115	54	88	112	320	17.6 67	370	99.6 15	14.2 48	60.667	104	15.720	60	6682	3	8
tc510024o	149	54	88	170	615	17.6 67	599	120. 662	9.61 4	112.000	139	15.720	65	4938	3	8
tc510035o	142	55	92	170	615	17.5 83	598	121. 333	9.80 2	111.500	138	16.008	64	4670	3	8
tc3210o	136	51	79	166	588	18.1 67	612	117. 370	8.68 6	105.667	146	14.504	66	4943	3	8
1mtc2210 h	111	49	73	149	489	18.5 00	592	110. 507	9.75 7	94.333	150	13.693	57	6679	3	8
3mtc 2210 h	120. 5	50	76	152	509	18.3 33	627	114.000	9.68 1	94.333	154	14.099	62	6006	3	8
tc3300o	125	50	76	180	674	18.3 33	688	120. 400	7.94 3	120.333	160	14.099	69	5162	3	8
tec330o	137. 5	48	71	191	749	19.0 00	118 9	129. 541	7.02 4	129.000	257	13.000	82	8457	4	8

[a] Notation and boiling points (BP) are taken from Rücker and Rücker.[9] [b] W = Wiener index; WW = hyper-Wiener index; ω = detour index; $\omega\omega$ = hyper-detour index; H = Harary index; P = Pasaréti index; V = Vérhalom index; WS = Wiener−sum index; ws = reverse Wiener-sum index; TNP = total number of paths; π = product−form version of the Wiener index; Z = Hosoya index; twc = total walk count index; n_r = number of rings; n = number of atoms.

models" which were selected using the standard CROMRsel approach[32,33] according to the best fitted and cross-validated statistical parameters. Previously, the CROMRsel approach was applied on several data sets which earlier had been modeled by neural networks (NNs), and, for each data set, models obtained using CROMRsel were better.[32,34,35] This modeling approach is also briefly described in the companion paper.[35]

After that, we tried to select a model(s) on the DS-76 set which will have a good performance in the prediction for molecules from the DS-104 set, and vice versa (to select model(s) on DS-104 with good predictive performance for DS-76 molecules). For this purpose the CROMRsel procedure was modified (the new approach is called CROMRsel-pred) to be able to select, among several possibilities, the best fit and cross-validated models, which will have the best predictive performance when applied to an external (test) data set (in this paper DS-104 is the test set for the model developed on DS-76 and vice versa). In the text that follows the term "calculated" refers only to the fit or cross-validation procedure, while the term "predicted", only to the prediction for an external (test) data set. Here we have to point out that the term "predicted" used in this study has limited meaning because the predictive quality parameters of the models were monitored during the model developing process and used for selecting the best models.

Statistical Parameters. For measuring the model quality, we used fit statistical parameters: correlation coefficient (R) and standard error of estimate (S), corresponding leave-one-out cross-validated parameters (Rcv and Scv), and predictive parameters (Rpred, Spred). Standard errors of estimate between experimental and fitted (S), leave-one-out cross-validated (Scv), and predicted (Spred) values were calculated using M (the number of compounds) in the denominator, as was done in ref 34.

Table 2. Linear Multivariate Regression Models of Boiling Points for DS-180, DS-76, and DS-104 Sets

N	*R*	Rcv	*S*/°C	Scv/°C	descriptors[a]
180	0.9863	0.9855	7.28	7.47	*H*, ws, *n*
180	0.9897	0.9881	6.322	6.82	*H*, ws, n_r, *n*
180	0.9910	0.9890	5.90	6.58	$\omega\omega$, *H*, ws, *Z*, *n*
180	0.9931	0.9913	5.19	5.84	$\omega\omega$, *H*, *P*, ws, TNP, *Z*, n_t, *n*
76	0.9956	0.9944	5.62	6.34	*H*, ws, n_r, *n*
76	0.9960	0.9944	5.32	6.35	ω, *P*, WS, *Z*, *n*
76	0.9969	0.9959	4.67	5.38	*H*, *P*, ws, *Z*, n_t, *n*
104	0.9668	0.9603	6.84	7.47	*Z*, twc, *n*
104	0.9758	0.9730	5.86	6.18	*H*, WS, *Z*, *n*
104	0.9800	0.9768	5.34	5.73	$\omega\omega$, *H*, ws, *Z*, *n*

[a] See Table 1 and text for explanation of descriptors.

Table 3. Multivariate Regression Models Selected from Both Initial and the Natural Logarithm of Initial Descriptors for DS-180, DS-76, and DS-104 Sets

N	*R*	Rcv	*S*/°C	Scv/°C	descriptors[a]
180	0.9867	0.9861	7.16	7.34	$\ln(W+1)$, $\ln(Z+1)$
180	0.9925	0.9919	5.40	5.60	$\ln(\omega+1)$, $\ln(Z+1)$, $\ln(n_r+1)$
180	0.9936	0.9929	5.00	5.25	TNP, $\ln(\omega+1)$, $\ln(Z+1)$, $\ln(n_r+1)$
180	0.9943	0.9937	4.69	4.94	*P*, TNP, $\ln(\omega+1)$, $\ln(\omega\omega+1)$, $\ln(Z+1)$
180	0.9952	0.9944	4.29	4.64	WW, *P*, WS, TNP, $\ln(\omega+1)$, $\ln(Z+1)$
76	0.9958	0.9951	5.41	5.86	$\ln(W+1)$, $\ln(Z+1)$
76	0.9978	0.9975	3.89	4.17	n_t, $\ln(\omega+1)$, $\ln(Z+1)$
76	0.9987	0.9983	3.02	3.50	$\omega\omega$, *V*, WS, $\ln(V+1)$, $\ln(Z+1)$
104	0.9808	0.9790	5.22	5.46	$\ln(\omega+1)$, $\ln(\mathrm{TNP}+1)$, $\ln(Z+1)$
104	0.9827	0.9804	4.96	5.27	*H*, ws, *n*, $\ln(Z+1)$

[a] See Table 1 and text for explanation of descriptors.

RESULTS AND DISCUSSION

Linear Models. First, the best linear multivariate models containing *I* ($I = 3-8$) descriptors, which have similar or better fit performances than the published models (for the same sets),[7] were selected starting from the initial descriptors. Statistical parameters of these models, as well as the descriptors involved, are given in Table 2.

It is interesting that, at this linear level, only one distance-related index, that is, the Harary index, is included in all but two of the best models presented in Table 2. Additionally, only the number of atoms is also present in all linear models. Both the *Z* and ws indices are involved in seven out of 10 models.

Nonlinear Models. After that we enlarged the initial set containing 15 initial descriptors with natural logarithms of 14 descriptors (because ln π was used as initial descriptor, we did not use its logarithmic transformation). Then, the best nonlinear multivariate regression models were selected from these 29 descriptors. Obtained models are given in Table 3.

In these (nonlinear) models the Harary index and the ws index are involved only in one model. In all models descriptor $\ln(Z+1)$ is present. However, the distance-related indices are involved in generating the best models. For example, the best six-descriptor model for the DS-180 set contains WW, *P*, WS, TNP, $\ln(\omega+1)$, and $\ln(Z+1)$, the best five-descriptor model for the DS-76 set contains, $\omega\omega$, *V*, WS, $\ln(V+1)$, and $\ln(Z+1)$, and the best four-descriptor model for DS-104 set contains *H*, ws, *n* and $\ln(Z+1)$. One can also see that the logarithmically transformed descriptors occur more frequently in the models presented in Table 3 than the linear ones, indicating the presence of the logarithmic nonlinear relationships between calculated indices and boiling points.

In the next step the data set contained the logarithm of initial (14) descriptors (except ln π) and the logarithm of their cross-products (105 descriptors), having 120 descriptors, altogether. Selected models are given in Table 4.

One can see in the case of the DS-180 and DS-76 sets that the Hosoya (*Z*) is involved in each of the obtained best models, mostly in combination with other indices (detour or Wiener).

Best Predictive Models. Before we started to generate models with a good predictive performance, we investigated the distribution of BP values over DS-76 and DS-104 sets. Results are presented in Figure 1 and indicate that DS-76 and DS-104 sets have very different boiling point distributions, far broader for DS-76 than for DS-104.

Because of that, starting from the best fit models developed on DS-76 compounds one cannot expect to predict correctly boiling points for DS-104 compounds. Having this in mind, we had to select the best model for the DS-76 and DS-104 sets according to the fit, cross-validated, and predictive statistical parameters of the models. In this, predictive parameters were used as the most important ones. We expect that such a model would possess very good properties and, accordingly, could be considered as the best model in general. Unfortunately, there are only a few such models. The best one is the two-descriptor model involving $\ln[(\omega+1)(Z+1)]$ and $\ln[(P+1)(n_r+1)]$. The model obtained for DS-76 is as follows:

$$\mathrm{BP} = -(138.94 \pm 2.1) + (42.05 \pm 0.72)\ln[(\omega+1)(Z+1)] - (15.36 \pm 1.0)\ln[(P+1)(n_r+1)] \quad (12)$$

$$N = 76, \quad R = 0.9970, \quad \mathrm{Rcv} = 0.9965, \quad S = 4.61\ ^\circ\mathrm{C}, \quad \mathrm{Scv} = 4.96\ ^\circ\mathrm{C}$$

This model, when applied to the DS-104 set as the external (test) set, gives the following statistical parameters: Rpred = 0.9787, Spred = 5.91 °C. To see the influence of methane on the model quality, we excluded methane from DS-76 and obtained the following model:

$$\mathrm{BP} = -(135.74 \pm 2.3) + (41.30 \pm 0.73)\ln[(\omega+1)(Z+1)] - (14.86 \pm 1.0)\ln[(P+1)(n_r+1)] \quad (13)$$

$$N = 75, \quad R = 0.9965, \quad \mathrm{Rcv} = 0.9960, \quad S = 4.36\ ^\circ\mathrm{C}, \quad \mathrm{Scv} = 4.66\ ^\circ\mathrm{C}$$

In prediction on external data set DS-104 using eq 13 we obtained Rpred = 0.9787, Spred = 5.74 °C, which is significantly better, especially if we keep in mind that only one molecule is excluded. The model obtained for DS-104 is as follows:

$$\mathrm{BP} = -(122.3 \pm 4.8) + (38.7 \pm 1.0)\ln[(\omega+1)(Z+1)] - (13.6 \pm 1.0)\ln[(P+1)(n_r+1)] \quad (14)$$

$$N = 104, \quad R = 0.9788, \quad \mathrm{Rcv} = 0.9776, \quad S = 5.49\ ^\circ\mathrm{C}, \quad \mathrm{Scv} = 5.63\ ^\circ\mathrm{C}$$

Table 4. Multivariate Regression Models for DS-180 and DS-76 Sets Selected from the Natural Logarithm ($\ln(x+1)$) of Initial Descriptors and the Natural Logarithm of Their Cross Products

model 1	$N = 180$, $R = 0.9867$, Rcv = 0.9862, $S = 7.17$ °C, Scv = 7.30 °C
	$BP = (142.1 \pm 3.0) + (33.62 \pm 0.42) \ln[(W+1)(Z+1)]$
model 2	$N = 180$, $R = 0.9927$, Rcv = 0.9923, $S = 5.31$ °C, Scv = 5.48 °C
	$BP = (135.3 \pm 2.1) + (40.99 \pm 0.51) \ln[(\omega+1)(Z+1)] - (14.48 \pm 0.55) \ln[(P+1)(n_r+1)]$
model 3	$N = 180$, $R = 0.9936$, Rcv = 0.9932, $S = 4.99$ °C, Scv = 5.14 °C
	$BP = (161.2 \pm 2.4) + (40.6 \pm 1.9) \ln(Z+1) + (29.22 \pm 0.93) \ln[(WS+1)(ws+1)] - (8.30 \pm 0.55) \ln[(twc+1)(n_r+1)]$
model 4	$N = 76$, $R = 0.9958$, Rcv = 0.9953, $S = 5.45$ °C, Scv = 5.76 °C
	$BP = (145.3 \pm 2.6) + (34.49 \pm 0.37) \ln[(W+1)(Z+1)]$
model 5	$N = 76$, $R = 0.9978$, Rcv = 0.9976, $S = 3.93$ °C, Scv = 4.12 °C
	$BP = (141.1 \pm 1.8) - (25.0 \pm 1.4) \ln(n_r+1) + (33.57 \pm 0.26) \ln[(\omega+1)(Z+1)]$
model 6	$N = 76$, $R = 0.9983$, Rcv = 0.9981, $S = 3.44$ °C, Scv = 3.71 °C
	$BP = (165.8 \pm 2.4) + (154.8 \pm 10.6) \ln[(WW+1)(Z+1)] + (74.0 \pm 2.9) \ln(H+1) - (67.2 \pm 5.5) \ln[(WS+1)(TNP+1)]$

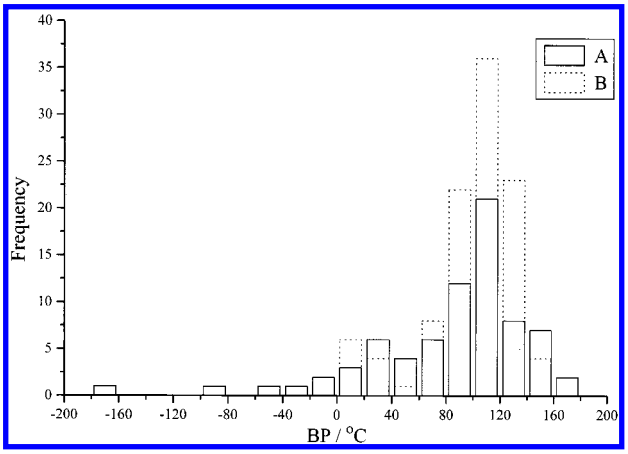

Figure 1. Histogram of the experimental boiling points distribution for the DS-76 set (A, solid line) and the DS-104 set (B, dotted line).

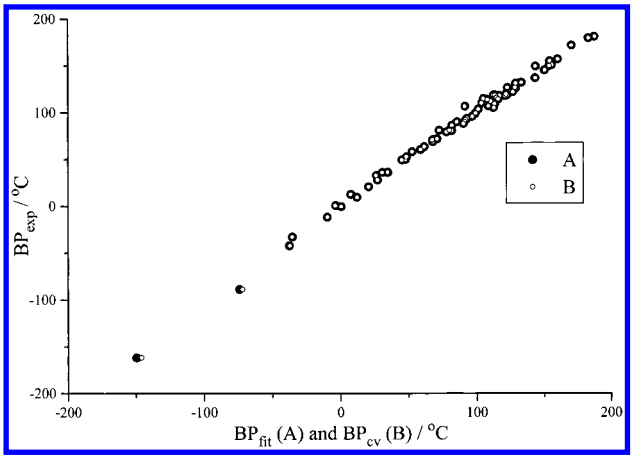

Figure 2. Plot of the experimental versus fit (A) and cross-validated (B) BP values for the DS-76 set.

Model 14, when applied to the DS-76 set, gives Rpred = 0.9970, Spred = 6.30 °C and, without methane Rpred = 0.9965, Spred = 5.33 °C. The quality of models 12 and 14 can be much better seen from the scatter plots of experimental versus calculated (fitted and cross-validated (CV)) BP values (see Figures 2 and 3).

Each fit BP value is close to the corresponding CV value. There is only one exception for methane (see Figure 2). In addition, we calculated the model for DS-180 with the same descriptors (eq 15), and performed the same scatter plot (Figure 4) as was done for DS-76 and DS-104, to see the overall quality in fitting and in cross-validation:

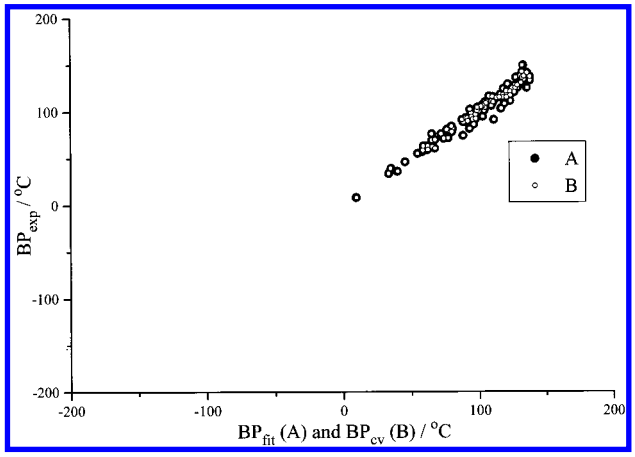

Figure 3. Plot of the experimental versus fit (A) and cross-validated (B) BP values for the DS-104 set.

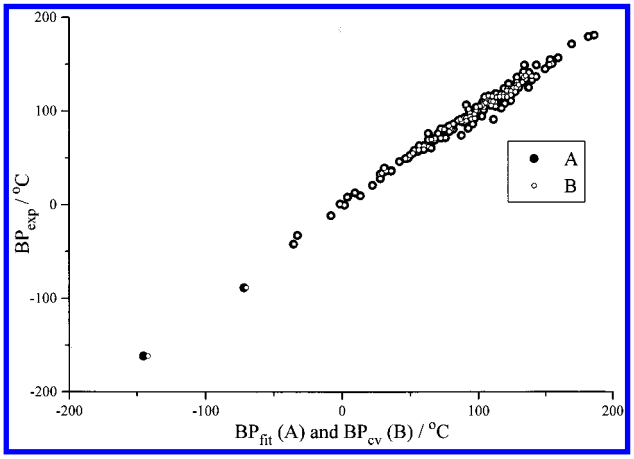

Figure 4. Plot of the experimental versus fit (A) and cross-validated (B) BP values for the DS-180 set.

$$BP = -(135.30 \pm 2.1) + (40.99 \pm 0.51) \ln[(\omega+1)(Z+1)] - (14.48 \pm 0.54) \ln[(P+1)(n_r+1)] \quad (15)$$

$N = 180$, $R = 0.9927$, Rcv = 0.9923, $S = 5.31$ °C, Scv = 5.48 °C

Without methane ($N = 179$), we obtained smaller values of the correlation coefficients ($R = 0.9916$, Rcv = 0.9912), but also the lower standard error of estimates $S = 5.15$ °C, Scv = 5.28 °C. This is so because the methane BP value is far from the mean BP value of the DS-180 set. One can also see from Figure 4 that the difference between the fit and CV BP values is somewhat larger for methane than for any other molecule considered.

Table 5. Statistical Parameters of the One-Descriptor Model Based on Descriptor $(W \times \omega)^{0.12972}$ for DS-76, DS-104, and DS-180 Sets

model 1	$N = 76$, $R = 0.9950$, Rcv = 0.9946, $S = 5.94$ °C, Scv = 6.15 °C
	$BP = (169.09 \pm 3.1) + (93.38 \pm 1.1)(W \times \omega)^{0.12972}$
prediction:	$N = 104$, Rpred = 0.9618, Spred = 7.83 °C
model 2	$N = 104$, $R = 0.9618$, Rcv = 0.9606, $S = 7.33$ °C, Scv = 7.45 °C
	$BP = (160.13 \pm 7.4) + (89.43 \pm 2.5)(W \times \omega)^{0.12972}$
prediction:	$N = 76$, Rpred = 0.9950, Spred = 6.70 °C
model 3	$N = 180$, $R = 0.9876$, Rcv = 0.9873, $S = 6.92$ °C, Scv = 6.99 °C
	$BP = (166.85 \pm 3.1) + (92.08 \pm 1.1)(W \times \omega)^{0.12972}$

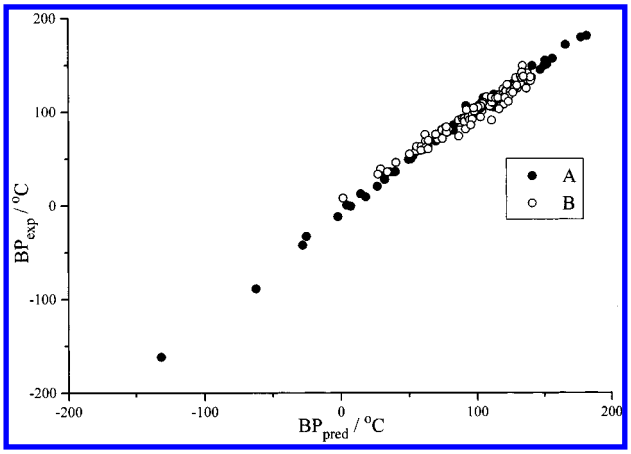

Figure 5. Plot of the experimental versus predicted BP values for the DS-76 (A) and DS-104 (B) sets obtained by eqs 12 and 14, respectively.

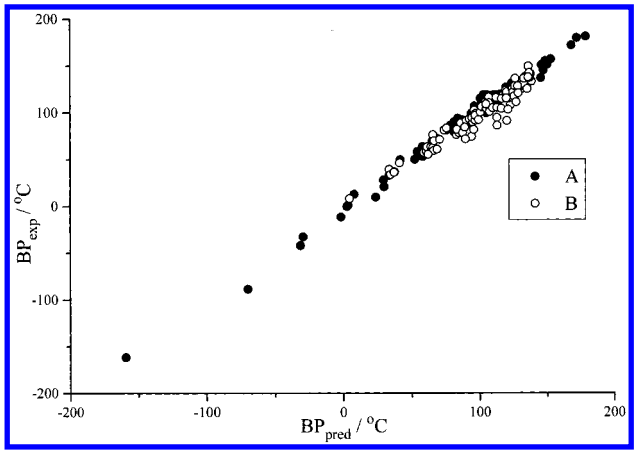

Figure 6. Plot of the experimental versus predicted BP values for the DS-76 (A) and DS-104 (B) sets obtained by $(W \times \omega)^{0.12972}$ descriptor, respectively (using models 1 and 2 in Table 5).

This indicates that the attractive forces between methane molecules, which determine the temperature at which methane boils (methane boiling point), can be the least accurately described with descriptors $\ln[(\omega + 1)(Z + 1)]$ and $\ln[(P + 1)(n_r + 1)]$. This interaction is underestimated in both models developed for DS-76 and DS-180 sets. It is important to note that the regression coefficients of each descriptor in eq 15 have very small errors, implying a good-quality model has been obtained.

Predicted BP values for the molecules in the DS-76 set (by eq 14) and in the DS-104 set (by eq 12) were used to calculate the overall prediction accuracy for all 180 compounds. The scatter plot between experimental and predicted values is given in Figure 5. Predictive statistical parameters for 180 molecules are Rpred = 0.9912, Spred = 6.08 °C.

In addition, we performed the same validation and prediction for molecules from data sets DS-76 (model developed on DS-104) and DS-104 (model developed on DS-76) using a one-descriptor model based on a $(W \times \omega)^{0.12972}$ descriptor. Descriptor $(W \times \omega)^{0.12972}$ was chosen because Rücker and Rücker[8] found that this descriptor gives the best one-descriptor model among the descriptors of the form $(W \times \omega)^x$ (models having such form used in ref 8 ($a + b(W \times \omega)^x$) have three parameters to be optimized: a, b, and x). The models based on this descriptor for all the studied sets and their statistical performances are given in Table 5.

The scatter plot between experimental and predicted values is given in Figure 6.

Overall statistics in the prediction for 180 alkanes (DS-76 plus DS-104) is Rpred = 0.9861, Spred = 7.37 °C.

Comparing the models from Table 5 and the model from eqs 12−15, one can see that the latter models are significantly better according to their fit, CV, and predictive parameters. The most important descriptor involved in the models (eqs 12−15) is $\ln[(\omega+1)(Z+1)]$ (according to the ratio between regression coefficients and its corresponding errors). It was found by Rücker and Rücker[8] that single descriptor $\ln(Z)$ (or Z) is the best descriptor for alkanes having the same number of carbon atoms. They (as well as the other authors) have also found that the detour index is a good descriptor for the alkane BP modeling (especially in combination with the Wiener index). It is important to note that the model containing $\ln(Z + 1)$ and $\ln(\omega + 1)$ terms would have two optimized parameters (one parameter more than when one uses $\ln[(\omega + 1)(Z + 1)]$). The second descriptor is $\ln[(P + 1)(n_r + 1)]$, where P is the Pasaréti index and n_r is the number of rings. For this descriptor, errors in the regression coefficient values are relatively small compared with the regression coefficient values themselves, implying the descriptor significance.

CONCLUSIONS

(1) Previous selections of the best QSPR models were usually based on fit (or, in the best case, on cross-validated) statistical parameters. We show here that it is better to carry out the model selection according to the predictive than according to the fitted or cross-validated statistics. Moreover, for data sets having property/activity values distributed in different ranges the latter case is the only way for obtaining consistent models.

(2) The use of logarithmic transformation of descriptors gives better models (than the use of initial descriptors). It is also confirmed that the CROMRsel procedure[32,33] is a powerfull method for selecting the best descriptors for designing the multivariate regression models.

(3) It is confirmed that a single descriptor model based on the power of $(W \times \omega)$ descriptor has also very good predictive properties, although this model was selected, in previous papers,[2,7,8] only according to the fit statistics.

(4) The best model for predicting the boiling points of acyclic and cyclic alkanes contains $(Z \times \omega)$ and $(P \times n_r)$ as descriptors. In the $(Z + 1) \times (\omega + 1)$ product, descriptor Z describes variation of BPs for alkanes with the same number of carbon atoms, while ω describes the variation of BPs between the alkanes and cycloalkanes of a different number of carbon atoms (this result was also obtained by Rücker and Rücker).[8,9] The second descriptors involved the Pasaréti index (which has exactly the same values as the ω and W indices for alkanes) corrected with the number of rings. This is the first study in which it is shown that the Pasaréti index can be useful in modeling the BPs of alkanes and cycloalkanes.

ACKNOWLEDGMENT

This work was supported in part by the Ministry of Science and Technology of the Republic of Croatia via Grants No. 00980606 (B.L., S.N., and N.T.) and 106N407 (B.L.), and by the joint research program between the Croatian Academy of Sciences and Arts and the Hungarian Academy of Sciences (I.L. and N.T.). We are thankful to the reviewers for their helpful and detailed comments.

REFERENCES AND NOTES

(1) Ivanciuc. O.; Ivanciuc, T. Matrixes and Structural Descriptors Computed from Molecular Graph Distances. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach Science Publishers: Amsterdam, 1999; pp 221−277.

(2) Nikolić, S.; Trinajstić, N.; Mihalić, Z. The Detour Matrix and the Detour Index. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach Science Publishers: Amsterdam, 1999; pp 279−306.

(3) Estrada, E.; Rodriguez, L.; Gutierez, A. Matrix Algebraic Manipulation of Molecular Graphs. 1. Distance and Vertex-Adjacency Matrices. *MATCH (Commun. Math. Comput. Chem.)* **1997**, *35*, 145−156. Estrada, E.; Rodriguez, L. Matrix Algebraic Manipulation of Molecular Graphs. 2. Harary- and MTI-like Molecular Descriptors. *MATCH (Commun. Math. Comput. Chem.)* **1997**, *35*, 157−167. Estrada, E.; Ivanciuc, O.; Gutman, I.; Gutierez, A.; Rodriguez, L. Extended Wiener Indices. A New Set of Descriptors for Quantitative Structure−Property Studies. *New J. Chem*. **1988**, *22*, 819−822.

(4) Nikolić, S.; Trinajstić, N.; Mihalić, Z. The Wiener Index: Development and Applications. *Croat. Chem. Acta* **1995**, *68*, 105−129.

(5) Gutman, I., Klavžar, S., Mohar, B., Eds. Fifty Years of the Wiener Index. *Comm. Math. Comput. Chem. (MATCH)* **1997**, *35*, 1−259.

(6) Bytautas, L.; Klein, D. J. Mean Wiener Number and Other Mean Extensions for Alkane Trees. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 471−481.

(7) Lukovits, I. The Detour Index. *Croat. Chem. Acta* **1996**, *69*, 873−882.

(8) Rücker, G.; Rücker, C. Symmetry-Aided Computation of the Detour Matrix and the Detour Index. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 710−714.

(9) Rücker, G.; Rücker, C. On Topological Indices, Boiling Points, and Cycloalkanes. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 788−802.

(10) Harary, F. *Graph Theory*; Addison-Wesley: Reading, MA, 1971; 2nd printing.

(11) Trinajstić, N. *Chemical Graph Theory*, 2nd ed.; CRC Press: Boca Raton, FL, 1992.

(12) Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc*. **1947**, *69*, 17−20.

(13) Hosoya, H. Topological Index. A Newly Proposed Quantity Characterizing Topological Nature of Structural Isomers of Saturated Hydrocarbons. *Bull. Chem. Soc. Jpn.* **1971**, *44*, 2332−2339.

(14) Mihalić, Z.; Veljan, D.; Amić, D.; Nikolić, S.; Plavšić, D.; Trinajstić, N. The Distance Matrix in Chemistry. *J. Math. Chem*. **1992**, *11*, 223−258.

(15) Roberts, F. S. *Discrete Mathematical Model*; Prentice Hall: Englewood Cliffs, NJ, 1976; p 58. Bersohn, M. A Fast Algorithm for Calculation of the Distance Matrix of a Molecule. *J. Comput. Chem*. **1982**, *4*, 110−113. Müller, W. R.; Szymanski, K.; Knop, J. V.; Trinajstić, N. An Algorithm for Construction of the Molecular Distance Matrix, *J. Comput. Chem.* **1987**, *8*, 170−173. Brown, R. F.; Brown, B. W. *Finite Mathematics*; Ardsley: New York, 1992; pp 534−545. Mihalić, M.; Trinajstić, N. A Graph-Theoretical Approach to Structure−Property Relationships. *J. Chem. Educ*. **1992**, *69*, 701−712. Mihalić, M.; Nikolić, S.; Trinajstić, N. Comparative Study of Molecular Descriptors Derived from the Distance Matrix. *J. Chem. Inf. Comput. Sci*. **1992**, *32*, 28−37. Schultz, H. P. Topological Organic Chemistry. 13. Transformation of Graph Adjacency Matrixes to Distance Matrixes. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 1158−1159.

(16) Randić, M. Novel Molecular Descriptor for Structure−Property Studies. *Chem. Phys. Lett*. **1993**, *211*, 478−483.

(17) Lukovits, I.; Linert, W. A Novel Definition of the Hyper-Wiener Index for Cycles. *J. Chem. Inf. Comput. Sci*. **1994**, *34*, 899−902.

(18) Klein, D. J.; Lukovits, I.; Gutman, I. On the Definition of the Hyper-Wiener for Cycle-Containing Structures. *J. Chem. Inf. Comput. Sci*. **1995**, *35*, 50−52.

(19) Amić, D.; Trinajstić, N. On the Detour Matrix. *Croat. Chem. Acta* **1995**, *68*, 53−62.

(20) Ivanciuc, O.; Balaban, A. T. Design of Topological Indices. Part 8. Path Matrices and Derived Molecular Graph Invariants. *Comm. Math. Chem. (MATCH)* **1994**, *30*, 141−152.

(21) Trinajstić, N.; Nikolić, S.; Lučić, B.; Amić, D.; Mihalić, Z. The Detour Matrix in Chemistry. *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 631−638.

(22) Nikolić, S.; Trinajstić, N.; Jurić, A.; Mihalić, Z. The Detour Matrix and the Detour Index of Weighted Graphs. *Croat. Chem. Acta* **1996**, *69*, 1577−1591.

(23) Trinajstić, N.; Nikolić, S.; Mihalić, Z. On Computing the Molecular Detour Matrix. *Int. J. Quantum Chem.* **1997**, *65*, 415−419.

(24) Lukovits, I.; Razinger, M. On Calculation of the Detour Index. *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 283−286.

(25) Linert, W.; Lukovits, I. Formulas for the Hyper-Winer and Hyper-Detour Indices of Fused Bicyclic Structures. *Comm. Math. Comp. Chem. (MATCH)* **1997**, *35*, 65−74.

(26) Plavšić, D.; Nikolić, S.; Trinajstić, N.; Mihalić, Z. On the Harary Index for the Characterization of Chemical Graphs. *J. Math. Chem.* **1993**, *12*, 235−250.

(27) Ivanciuc, O.; Balaban, T.-S.; Balaban, A. T. Design of Topological Indices. Part 4. Reciprocal Distance Matrix, Related Local Vertex Invariants and Topological Indices. *J. Math. Chem.* **1993**, *12*, 309−318.

(28) Lukovits, I. An All-Path Vesrion of the Wiener Index. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 125−129.

(29) Plavšić, D.; Trinajstić, N.; Amić, D.; Šoškić, M. Comparison between the Structure-Boiling Point Relationship with Different Descriptors for Condensed Benzenoids. *New J. Chem*. **1998**, *22*, 1075−1077.

(30) Randić, M. On Characterization of Cyclic Structures. *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 1063−1071.

(31) Gutman, I.; Linert, W.; Lukovits, I.; Tomović, Ž. The Multiplicative Version of the Wiener Index. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 113−116.

(32) Lučić, B.; Trinajstić, N. Multivariate Regression Outperforms Several Robust Architectures of Neural Networks in QSAR Modeling. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 121−132.

(33) Lučić, B.; Trinajstić, N.; Sild, S.; Karelson, M.; Katritzky, A. R. A New Efficient Approach for Variable Selection Based on Multiregression: Prediction of Gas Chromatographic Retention Times and Response Factors. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 610−621.

(34) Lučić, B.; Amić, D.; Trinajstić, N. Nonlinear Multivariate Regression Outperforms Several Concisely Designed Neural Networks in QSPR Modeling. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 403−413.

(35) Lučić, B.; Trinajstić, N. Multivariate Regression versus Artificial Neural Network Ensembles in QSAR. *J. Chem. Inf. Comput. Sci.*, submitted.

CI0000777