Prediction of Gas-Chromatographic Retention Indices Using Topological Descriptors

Matevž Pompe

Faculty of Chemistry and Chemical Technology, University of Ljubljana, Aškerčeva 5, 1001 Ljubljana, Slovenia

Marjana Novič

National Institute of Chemistry, Hajdrihova 19, 1001 Ljubljana, Slovenia

Received April 3, 1998

Theoretical prediction of gas-chromatographic retention indices could be used as an additional method for the identification of organic substances during gas-chromatographic separation. Our previously developed model, based on artificial neural networks, has been extended with the additional topological structural descriptors to improve prediction capabilities. The topological indices were selected for the representation of chemical structures because of their simplicity; therefore they could also be used for solving identification problems by chromatographers who are not experts in structural representation. An extensive data set of 381 simple organic compounds with known retention indices taken from the literature served as a training and test set. Sixteen informational and topological structural descriptors were selected for the description of molecular structure. The same data set was used for the prediction of gas-chromatographic retention indices using a multiple linear regression model and back-propagation of error and counterpropagation artificial neural network. The average root mean squared error values of a 10-fold cross-validation procedure were 22.5, 19.2, and 36.1, respectively.

INTRODUCTION

In quantitative structure retention relationship (QSRR) research, chemical compounds must be presented in computer-readable form, and afterward the quantitative correlation between a chemical structure and its property can be obtained using different statistical and learning procedures, that is, multiple linear regression (MLR) and several types of artificial neural networks (ANNs). But the major problem in QSRR studies still remains a mathematical representation of molecular structure, that is, translation of molecular structure to the computer-readable form retaining as much structural information as possible.

In QSRR studies the chemical compounds can be presented by empirical physicochemical parameters[1] or nonempirical structural descriptors.[2] Because it is not always possible to obtain physicochemical parameters for each compound of interest, the development of nonempirical structural representation that is able to give a satisfactory correlation with the studied property would be beneficial. Several topological, geometric, electronic, and quantum chemical indices have been already used in quantitative structure property relationship research.[3,4] Because topological indices can be easily designed and calculated and they still give good correlation with retention properties, their use in QSRR studies instead of more complicated geometric, electronic, and quantum chemical descriptors is promising. It makes modeling of chromatographic properties easier and well suited for practical chromatography.

The idea of representing chemical structure by means of a graph is very old. Wiener[5] was the first who tried to represent it by a numerical number. Until now many different topological indices have been developed; among them the best known are those of Hosoya,[6] Randić,[7] Bonchev and Trinajstić,[8] Balaban,[9] and Kier and Hall.[10] Many investigators have reported a good correlation between the experimental gas-chromatographic retention indices and structural characteristics of molecules by using different topological indices as structural descriptors.[4,11−13] However, all of them tried to predict retention properties only for a limited number of organic compounds.

Because our research group is involved in investigation of abundance of different volatile organic compounds in the atmosphere, we have tried to develop a single model for the prediction of gas-chromatographic retention indices for the large variety of organic compounds. Only a very limited number of similar studies have been reported.[14] They used complicated structural coding that is not suitable for practical applications.

The aim of our work was to extend our previously developed model[15] with additional structural descriptors and with an enlarged data set. Plainly topological indices have been selected for the representation of molecular structure to obtain a prediction model as simple as possible. Besides the MLR model, back-propagation of error (BPE) and counterpropagation (CP) ANNs have also been tested for building the prediction model.

METHOD

Data Set. The data set has been taken from the literature.[14,16] It contains acyclic and cyclic alkanes, alkenes, dienes, alcohols, esters, ketones, aldehydes, ethers, and aromatic compounds with three to eleven carbon atoms and up to two oxygen atoms. This leads to the final number of 381 compounds used for the prediction of retention data.

10.1021/ci980036z CCC: $18.00
Published on Web 12/18/1998

Experimental gas-chromatographic retention indices for nonpolar stationary-phase squalene at 70 °C are presented in Table 1 (column 1).

Before the data set was used for creation of the different models, the descriptors as well as retention indices were normalized; each done separately. Their minimal values were set to zero, and maximal values to one.

Calculation and Selection of Structural Descriptors. Molecular structures have been created by HyperChem™. Afterward, 68 informational and topological descriptors were calculated with Codessa software.[17] Informational descriptors included molecular weight, gravitation index, the counts of atoms, rings, aromatic bonds, aromatic rings, substructures, and so forth. Topological descriptors included Wiener, Randić, Kier and Hall, Kier shape, information content, and Balaban indices. Molecular connectivity descriptors (Randić, Kier and Hall) encoded the size and the degree of branching in a molecule, Kier shape indices encoded topological molecular shape by using graph theoretical approach, and information content indices used information theory for characterization of the structural features present in the molecule, that is, atoms, bonds, and so forth. Because such a large amount of chemical descriptors is not suitable for the creation of any models, especially not for MLR, the reduction of variables is obligatory. The selection of structural descriptors is accomplished by applying the heuristic optimization search in Codessa software[18] as described below.

At the beginning, all descriptors were checked to ensure that values of each descriptor are available for each structure and that there was a variation in these values. At the same time each descriptor had to have a correlation coefficient with the inquired property > 0.01. Descriptors that did not meet these requirements were discarded and only information-rich descriptors were retained. The remaining description set has been further examined. After checking all possible pairs of descriptors, one descriptor was eliminated from each pair exhibiting a high pairwise correlation ($r > 0.995$). This led to the final number of 57 descriptors that were used for the stepwise selection of best subset of structural indices.

The selected descriptors have been listed in decreasing order according to the correlation coefficient of the one-factor correlation equation. All possible two-parameter correlations were calculated where the first parameter was one of the best 10 descriptors according to one-parameter linear regression model (MLR), whereas the second parameter was one of the 56 remaining descriptors. The pairwise correlation coefficient between both parameters of MLR had to be < 0.99. Ten best correlations were chosen from the set of calculated two-parameter linear equations. The criterion for the selection of the top 10 *n*-parameter models was correlation coefficient (r).

After the 10 best two-parameter linear regression models were selected, the third parameter was added from the set of remaining descriptors. Again the pairwise correlation coefficient between any pair of parameters present in the model did not exceed 0.99. These series of steps were repeated until we got a model with the prescribed number of parameters. During each series of steps the dimension of the model was increased by one.

At the end we must stress that very low limits for the elimination of the structural descriptors were selected, that is, one parameter and pairwise correlation coefficient were < 0.01 and > 0.99, respectively. Such low limits were selected because it had been already shown that linear combinations of well-correlated parameters, which individually do not correlate with the studied property, give very good correlation.[19]

Creation of Prediction Models. The best MLR model was selected according to the prediction capabilities of the leave-one-out (LOO) cross-validation procedure. The same set of structural descriptors was used for the creation of BPE and CP ANN models, since it has already been shown[20,21] that once the key structural features are identified through MLR technique, ANNs provide better prediction capabilities.

Back-propagation neural network models were composed of 16 neurons in the input layer, one hidden layer, and one neuron in the output layer.[22] The influence of ANN architecture on the learning results were tested by changing the number of neurons in the hidden layer. At the same time, other learning parameters such as learning rate, momentum, and number of epochs were also optimized. During all optimization steps the performance of different models were tested by a 10-fold cross-validation procedure described in the next section.

In our study the third modeling approach was based on the CP ANN. The influence of architecture on prediction results was tested by changing the dimensions of CP ANN, where neurons were organized in a map in a two-dimensional plane[23,24] with $n \times n$ neurons where $n = 5, 10, 15, 20, 25, 30, 40, 50$. For the best three models the optimization of number of learning cycles was performed. Again the optimal parameters were selected according to the results of a 10-fold cross-validation procedure. The remaining parameters of the ANN were the same during all learning steps. Each neuron had 16 weights in the Kohonen layer plus one output weight. At the beginning, all weights were initialized randomly. Triangular form of the ANN correction function was used, and the remaining parameters of the ANN were the following: $r_{max} = n$, $a_{max} = 0.50$, $a_{min} = 0.01$, and toroid condition was not used.

Cross-Validation Procedures. The LOO cross-validation procedure was used for the evaluation of different MLR models during structural descriptor selection. The MLR model with the best cross-validation parameters was chosen for further studies.

A very large data set was used; it was composed of more than 10 subsets of chemical compounds. Such diverse data set did not allow us to use a standard training/test set approach for the evaluation of prediction results, since there was on average less than 40 compounds in the individual subsets. For example, the smallest subset of ethers containing only 12 compounds were also rather structurally diverse, that is, they contained 4−8 carbon atoms. In such instances, the selection of only one test set could give seriously biased prediction results.

Therefore for the final evaluation of prediction ability of all developed models a 10-fold cross-validation procedure[25,26] was used. The data set was randomly divided into 10 subsets of roughly equal size. Nine of them were used for the creation of the model, the tenth to test the prediction ability. This procedure was repeated 10 times, each time with one of the 10 subsets treated as a test set. During the 10-fold cross-validation procedure each of the data sets was used once for

Table 1. Experimental and Predicted Values of Retention Indices Obtained by Different Models

ID	structure name	experi-mental	MLR	BPE ANN	CP ANN	ID	structure name	experi-mental	MLR	BPE ANN	CP ANN
1	2,2-dimethylpropane	412.32	402.16	409.04	437.60	80	2,4-dimethyl-2-pentanol	762.00	780.24	764.24	695.52
2	2-methylbutane	475.28	504.16	495.84	489.68	81	3,3-dimethyl-1-butanol	764.00	720.08	700.48	657.68
3	2,2-dimethylebutane	536.80	563.04	550.48	529.92	82	3-hexanol	766.00	765.44	784.80	792.40
4	2,3-dimethylebutane	567.28	595.76	563.92	593.76	83	2-methyl-2-hexanol	803.04	804.32	822.16	814.00
5	2-methylpentane	569.68	584.96	571.92	585.68	84	2-methyl-1-pentanol	790.00	767.92	777.92	787.52
6	3-methylpentane	584.24	606.00	600.80	572.88	85	2,4-dimethyl-3-pentanol	811.04	827.60	808.96	835.12
7	2,2-dimethylpentane	625.60	631.92	623.68	632.72	86	4-methyl-1-pentanol	790.96	767.20	779.60	735.52
8	2,4-dimethylpentane	629.84	650.80	618.48	640.24	87	2,3-dimethyl-3-pentanol	810.00	837.92	820.80	839.12
9	2,2,3-trimethylpentane	639.68	653.84	641.04	665.60	88	2-ethyl-1-butanol	801.04	775.68	788.48	792.40
10	3,3-trimethylpentane	658.88	673.12	663.12	679.84	89	3-methyl-1-pentanol	798.00	777.84	794.72	775.44
11	2-methylhexane	666.56	677.36	670.00	714.16	90	5-methyl-3-hexanol	823.04	832.64	831.36	822.00
12	2,3-dimethylpentane	671.68	680.24	666.56	668.72	91	3-Ethyl-3-pentanol	834.00	852.40	832.56	813.76
13	3-ethylpentane	686.00	688.00	685.44	674.16	92	1-hexanol	823.04	794.00	823.60	680.56
14	2,2,4-trimethylpentane	689.92	683.44	673.76	719.76	93	4-heptanol	860.00	860.88	868.56	883.44
15	2,2-dimethylhexane	719.36	714.88	722.00	731.44	94	2,2,4-trimethyl-3-pentanol	868.00	869.44	853.44	884.88
16	2,5-dimethylhexane	728.40	738.80	730.64	723.12	95	3,5-dimtehyl-3-hexanol	870.00	888.00	887.28	912.88
17	2,4-dimethylhexane	731.92	750.32	747.20	769.44	96	2-methyl-2-heptanol	911.04	896.32	922.32	946.88
18	2,2,3-trimethylpentane	737.12	735.84	743.68	757.52	97	6-methyl-2-heptanol	936.96	925.04	937.44	901.04
19	3,3-dimethylhexane	743.52	746.40	755.68	757.52	98	4-ethyl-3-hexanol	938.96	940.00	945.84	944.32
20	2,3,4-trimethylpentane	752.40	762.64	745.44	757.04	99	4-octanol	960.00	956.88	967.68	933.12
21	2,3,3-trimethylpentane	759.36	755.44	761.28	785.12	100	3-octanol	968.00	958.00	982.64	970.80
22	2-methylheptane	764.88	769.20	763.28	801.52	101	3,6-dimethyl-3-heptanol	970.00	976.56	978.48	932.64
23	4-methylheptane	767.20	774.96	775.76	768.16	102	3-methyl-2-butanone	614.00	618.88	609.76	641.20
24	3,4-dimethylhexane	770.56	782.40	758.56	750.72	103	2-pentanone	639.04	615.44	628.64	635.84
25	3-methylheptane	772.32	781.60	774.96	750.40	104	3-pentanone	650.96	621.04	624.96	613.84
26	2,2,4,4-tetramethylpentane	772.72	710.40	741.68	798.00	105	3,3-dimethyl-2-butanone	674.00	679.44	688.40	711.92
27	3-methyl-3-ethylpentane	774.00	779.92	776.80	754.40	106	4-methyl-2-pentanone	698.00	692.56	698.08	692.32
28	2,2,5-trimethylhexane	776.32	774.00	794.32	820.08	108	4-methyl-3-pentanone	708.00	702.96	700.00	715.20
29	2,2,4-trimethylhexane	789.12	780.56	810.16	816.56	109	3-methyl-2-pentanone	711.04	718.48	727.76	717.52
30	2,4,4-trimethylhexane	808.72	801.20	815.28	824.96	110	3-hexanone	740.96	713.12	724.80	792.40
31	2,3,5-trimethylhexane	812.00	825.20	818.00	824.80	111	2-hexanone	742.00	712.88	743.20	753.92
32	2,2-dimethylheptane	815.36	803.60	822.08	853.20	112	2,4-dimethyl-3-pentanone	763.04	775.60	796.88	799.52
33	2,2,3,4-tetramethylpentane	819.60	805.20	810.16	827.04	113	5-methyl-3-hexanone	794.00	779.60	793.52	835.28
34	2,2,3-trimethylhexane	821.60	812.88	821.20	831.60	114	2-methyl-3-hexanone	798.00	789.68	804.96	762.32
35	2,2-dimethyl-3-ethylpentane	822.24	810.08	840.00	826.32	115	5-methyl-2-hexanone	814.00	781.12	807.60	828.40
36	3,3-dimethylheptane	835.76	832.96	840.24	833.60	116	4-heptanone	830.96	807.04	820.32	846.40
37	2,4-dimethyl-3-ethylpentane	836.48	834.64	841.36	810.96	117	3-heptanone	842.00	813.92	836.88	833.60
38	2,3,4-trimethylhexane	849.12	851.04	857.68	841.36	118	2-heptanone	844.96	809.04	847.60	815.84
39	2,3,3,4-tetramethylpentane	858.00	849.04	848.64	820.16	119	2,2,4,4-tetamethyl-3-pentanone	878.00	853.84	899.28	953.28
40	3-methyloctane	870.24	874.00	864.72	846.40	120	2,6-dimethyl-4-heptanone	932.00	929.04	922.56	886.40
41	3,3-diethylpentane	877.20	869.76	899.60	835.44	121	2,2-dimethyl-3-heptanone	942.00	917.76	926.40	901.52
42	*n*-butane	400.00	409.52	381.12	407.60	122	3-octanone	944.00	902.48	927.44	930.16
43	*n*-pentane	500.00	512.00	486.32	495.12	123	2-octanone	944.96	907.12	955.20	943.84
44	*n*-hexane	600.00	608.48	595.52	593.52	124	acetic acid, ethyl ester	550.96	545.04	533.12	553.76
45	*n*-heptane	700.00	705.52	707.84	676.88	125	propionic acid, methyl ester	575.04	545.20	577.68	599.76
46	*n*-octane	800.00	799.92	813.60	776.96	126	isobutyric acid, methyl ester	632.00	649.84	665.20	666.96
47	*n*-nonane	900.00	900.64	911.52	872.56	127	propionic acid, ethyl ester	656.96	643.60	652.08	655.12
48	*n*-decane	1000.00	995.44	1006.24	940.08	128	acetic acid, propyl ester	654.00	649.04	649.68	638.40
49	ethylcyclopropane	510.16	468.16	509.92	586.48	129	butyric acid, methyl ester	678.96	657.44	650.08	643.44
50	cyclopentane	565.68	532.08	572.08	531.52	130	isobutyric acid, ethyl ester	718.00	732.64	736.64	738.00
51	ethylcyclobutane	621.12	594.48	608.48	643.44	131	acetic acid, *sec*-butyl ester	712.00	738.40	717.52	710.72
52	methylcyclopentane	627.92	623.68	628.00	645.76	132	acetic acid, isobutyl ester	730.00	723.52	708.80	702.24
53	cyclohexane	662.72	663.04	673.92	623.52	133	isopentanoic acid, methyl ester	732.96	701.12	712.72	721.52
54	1,1,2-trimethylcyclopentane	763.20	780.72	787.12	815.44	134	butyric acid, ethyl ester	751.04	754.16	740.56	727.84
55	1,1,3-trimethylcyclopentane	723.60	743.52	734.24	796.96	135	propionic acid, propyl ester	763.04	749.68	760.88	745.20
56	methylcyclohexane	725.76	744.88	735.52	742.56	136	acetic acid, butyl ester	769.04	756.64	748.96	760.80
57	ethylcyclopentane	733.76	721.44	734.24	698.32	137	butyric acid, isopropyl ester	796.96	822.08	806.16	803.52
58	1,1,dimethylcyclohexane	786.96	795.52	781.60	796.32	138	isopentanoic acid, ethyl ester	813.04	780.00	786.32	768.80
59	isopropylcyclopentane	812.08	800.96	795.92	801.84	139	propionic acid, isobutyl ester	830.00	818.40	838.00	803.52
60	*n*-propylcyclopentane	830.32	820.08	837.12	821.12	140	butyric acid, propyl ester	856.00	851.52	849.84	866.16
61	ethylcyclohexane	834.32	839.60	858.00	844.32	141	acetic acid, 1,3-dimethylbutyl ester	861.04	895.52	859.36	854.72
62	1,1,3-trimethylcyclohexane	840.40	846.96	857.60	910.48	142	propionic acid, butyl ester	873.04	852.96	871.92	869.76
63	1-*cis*-3-dimethylcyclopentane	682.72	697.04	688.96	693.28	143	acetic acid, pentyl ester	878.00	860.48	853.76	867.44
64	1-*trans*-3-dimethylcyclopentane	686.80	692.56	684.56	685.44	144	isobutyric acid, isobutyl ester	873.04	885.92	896.56	912.48
65	1-*cis*-2-dimethylcyclopentane	720.88	714.72	698.64	689.68	145	hexanoic acid, methyl ester	882.96	864.16	863.76	857.12
66	1-*trans*-2-dimethylcyclopentane	689.20	719.44	717.52	717.12	146	butyric acid, isobutyl ester	916.00	915.20	911.12	934.32
67	1-*trans*-2-dimethylcyclohexane	801.84	827.28	813.84	817.52	147	acetic acid, 2-ethylbutyl ester	932.96	933.28	906.96	905.04
68	1-*cis*-2-dimethylcyclohexane	829.28	829.36	814.00	802.24	148	butyric acid, butyl ester	958.00	955.52	943.68	962.96
69	1-*trans*-3-dimethylcyclohexane	805.60	808.32	800.00	796.32	149	hexanoic acid, ethyl ester	962.00	959.04	954.64	952.24
70	1-*cis*-3-dimethylcyclohexane	784.96	809.20	804.96	805.28	150	propionic acid, pentyl ester	970.00	953.12	965.36	960.24
71	2-methyl-2-butanol	612.00	642.24	635.84	592.88	151	acetic acid, hexyl ester	956.00	960.88	966.64	952.24
72	1-butanol	619.04	586.40	604.64	522.32	152	dipropyl ether	673.04	686.00	666.88	753.20
73	3-methyl-2-butanol	645.04	431.20	512.80	680.00	153	butyl ethyl ether	680.96	691.44	696.24	736.16
74	2-pentanol	663.04	667.92	679.20	630.72	154	dibutyl ether	876.00	891.60	893.36	1001.92
75	3-methyl-2-pentanol	765.04	767.92	761.12	744.80	155	1-octene	781.20	785.84	794.56	792.64
76	3-methyl-1-butanol	689.04	671.60	681.84	619.12	156	2-methyl-3-ethyl-2-pentene	778.40	758.80	780.96	760.24
77	4-methyl-2-pentanol	724.96	739.60	722.64	620.48	157	2,3,4-trimethyl-2-pentene	765.92	751.60	776.48	742.24
78	1-pentanol	722.00	691.68	722.00	615.52	158	2,5-dimethyl-2-hexene	749.92	730.40	732.80	704.00
79	2-methyl-3-pentanol	744.96	754.64	759.20	761.92	159	2,3-dimethyl-1-hexene	739.28	742.88	731.84	760.24

Table 1 (Continued)

ID	structure name	experi-mental	MLR	BPE ANN	CP ANN	ID	structure name	experi-mental	MLR	BPE ANN	CP ANN
160	2-methyl-3-ethyl-1-pentene	734.96	748.00	754.00	769.68	426	isopropylcyclohexane	918.00	910.72	904.00	919.52
161	2,2-dimethyl-*cis*-3-hexene	716.80	710.88	708.80	698.16	427	*n*-propylcyclohexane	926.00	935.28	962.32	972.40
162	2,2-dimethyl-*trans*-3-hexene	692.80	711.92	727.20	716.08	428	hexylcyclopropane	908.00	934.88	953.36	980.00
163	2,4,4-trimethyl-2-pentene	715.36	687.36	701.12	704.80	429	ethylcycloheptane	965.04	939.44	982.64	930.56
164	2,4,4-trimethyl-1-pentene	704.32	672.48	706.72	716.08	430	1,1-dimethylcycloheptane	911.52	895.52	900.64	922.32
165	2,3-dimethyl-2-pentene	703.36	685.36	692.64	654.56	431	1-*trans*-2-dimethylcycloheptane	936.00	924.24	954.96	930.56
166	3-ethyl-2-pentene	697.20	682.56	679.12	654.16	432	1-*cis*-2-dimethylcycloheptane	956.00	929.92	939.44	946.96
167	2-methyl-2-hexene	691.20	671.12	680.00	664.88	433	1-*trans*-3-dimethylcycloheptane	922.00	907.84	917.76	922.32
168	*cis*-3-heptene	690.40	696.64	686.24	696.72	434	1-*cis*-3-dimethylcycloheptane	915.52	908.72	928.48	917.12
169	*trans*-3-heptene	687.52	696.16	689.84	693.68	435	1-*trans*-4-dimethylcycloheptane	917.04	928.96	933.60	920.72
170	3-methyl-*trans*-3-hexene	691.20	682.88	691.44	678.16	436	1-*cis*-4-dimethylcycloheptane	921.04	925.68	925.36	918.40
171	3-methyl-*cis*-3-hexene	684.56	683.04	688.24	674.24	437	2-butanol	560.96	575.92	581.60	522.32
172	1-heptene	681.84	691.84	689.04	695.20	438	2-methyl-1-propanol	576.00	567.68	558.32	557.12
173	2-ethyl-1-pentene	681.84	662.48	654.24	654.48	439	2-methyl-2-propanol	487.04	491.52	567.52	487.68
174	2-methyl-1-hexene	678.08	659.12	662.80	654.96	440	3-pentanol	664.00	674.56	699.84	674.40
175	3,4-dimethyl-*trans*-2-pentene	678.32	677.52	675.52	685.44	441	2-methyl-1-butanol	695.04	679.52	697.92	671.60
176	3,4-dimethyl-*cis*-2-pentene	670.64	676.56	670.40	654.56	442	2,2-dimethyl-1-propanol	624.96	625.20	580.08	586.88
177	3-methyl-2-ethyl-1-butene	659.12	656.80	627.20	654.56	443	2,2-dimethyl-1-butanol	758.00	748.80	758.96	730.40
178	4-methyl-1-hexene	657.92	675.76	672.24	676.24	444	2,3-dimethyl-2-butanol	710.00	727.60	706.48	694.08
179	4-methyl-*trans*-2-hexene	656.72	681.68	684.56	683.36	445	3,3-dimethyl-2-butanol	711.04	725.12	714.88	713.20
180	4-methyl-*cis*-2-hexene	654.88	681.04	690.00	685.52	446	2-hexanol	764.00	766.64	781.28	711.20
181	2,3-dimethyl-1-pentene	650.40	668.96	659.04	670.00	447	2-methyl-2-pentanol	707.04	721.12	720.88	686.08
182	5-methyl-1-hexene	650.00	664.88	666.16	670.08	448	3-methyl-3-pentanol	730.96	748.80	742.96	761.92
183	3-ethyl-1-pentene	646.88	677.52	659.36	690.80	449	2,2-dimethyl-1-pentanol	844.00	826.64	835.04	812.24
184	3-methyl-1-hexene	644.72	669.68	667.68	670.16	450	3-heptanol	861.04	863.04	875.04	841.92
185	2,4-dimethyl-2-pentene	640.56	652.80	654.56	624.64	451	1-heptanol	928.96	892.88	911.20	861.60
186	2,4-dimethyl-1-pentene	637.68	635.60	639.36	637.92	452	2-heptanol	861.04	860.72	883.12	851.68
187	3,4-dimethyl-1-pentene	636.88	671.28	665.52	655.84	453	2-ethyl-4-methyl-1-pentanol	936.96	928.40	934.16	927.12
188	4,4-dimethyl-*cis*-2-pentene	635.52	628.56	627.12	620.32	454	1-octanol	1021.04	992.96	994.48	918.64
189	4,4-dimethyl-*trans*-2-pentene	614.72	631.84	625.76	627.28	455	2-octanol	959.04	957.04	978.88	943.84
190	3,3-dimethyl-1-pentene	626.08	657.28	662.80	670.00	456	2-ethyl-1-hexanol	984.96	958.40	956.88	930.96
191	2,3-dimethyl-2-butene	625.12	605.84	604.88	585.12	457	2-butanone	530.96	526.24	530.96	531.60
192	4,4-dimethyl-1-pentene	604.64	623.12	632.08	630.00	458	*n*-butyraldehyde	546.00	530.96	546.72	530.40
193	*cis*-2-hexene	603.60	600.16	591.36	590.88	459	isobutyraldehyde	517.04	518.80	514.96	530.40
194	*trans*-2-hexene	596.88	599.68	594.80	595.92	460	2-methylbutyraldehyde	632.00	631.04	631.84	613.84
195	3-methyl-*trans*-2-pentene	612.72	597.44	604.32	570.72	461	isovaleraldehyde	616.00	623.04	639.76	642.32
196	3-methyl-*cis*-2-pentene	602.80	601.84	603.44	585.36	462	valeraldehyde	656.00	639.44	661.20	624.48
197	2-methyl-2-pentene	597.76	588.88	593.36	562.00	463	2,2-dimethylpropanal	568.00	587.76	572.48	626.24
198	*cis*-3-hexene	592.64	604.80	592.40	597.04	464	capraldehyde	763.04	747.36	782.08	734.24
199	*trans*-3-hexene	592.08	602.80	594.08	593.68	465	4,4-dimethyl-2-pentanon	742.96	731.28	725.76	707.76
200	2-ethyl-1-butene	592.00	579.20	559.84	570.72	466	heptanal	866.00	846.64	890.48	883.44
201	1-hexene	582.32	595.04	579.28	597.04	467	2-ethylhexanal	914.00	915.52	946.40	897.28
202	2-methyl-1-pentene	580.08	570.40	562.72	566.16	468	4-octanone	930.00	904.40	923.04	964.16
203	2,3-dimethyl-1-butene	558.80	577.28	553.84	587.84	469	formic acid, *n*-propyl ester	569.04	568.24	543.44	543.84
204	4-methyl-*trans*-2-pentene	561.92	584.72	574.48	586.88	470	formic acid, isopropyl ester	526.00	554.08	545.60	558.00
205	4-methyl-*cis*-2-pentene	556.24	584.72	574.48	586.88	471	acetic acid, isopropyl ester	600.96	630.64	595.28	645.92
206	3-methyl-1-pentene	551.36	588.80	573.68	602.56	472	formic acid, *n*-butyl ester	678.96	690.32	680.56	645.92
207	4-methyl-1-pentene	549.36	575.04	566.24	573.92	473	formic acid, isobutyl ester	639.04	660.56	656.08	662.32
208	2-methyl-2-butene	514.32	510.08	507.60	484.88	474	formic acid, *sec*-butyl ester	632.00	672.16	654.00	666.96
209	3,3-dimethyl-1-butene	506.80	546.56	521.28	569.04	475	formic acid, amyl ester	775.04	798.32	792.32	745.20
210	*trans*-2-pentene	500.00	513.84	502.88	480.64	476	formic acid, 3-pentyl ester	740.96	770.96	745.04	721.52
211	*cis*-2-pentene	504.88	512.72	502.48	489.76	477	formic acid, 2-pentyl ester	736.00	773.12	748.56	750.08
212	2-methyl-1-butene	488.00	492.88	471.84	485.20	478	acetic acid, *tert*-butyl ester	646.00	678.16	693.76	758.08
213	1-pentene	481.76	500.08	474.24	480.64	479	*n*-propionic acid, isopropyl ester	702.00	724.88	729.44	749.84
214	3-methyl-1-butene	450.32	493.12	459.76	501.76	480	*n*-pentanoic acid, ethyl ester	844.00	859.60	854.56	869.52
215	*trans*-2-butene	406.56	423.84	424.40	403.84	481	acetic acid, 2-methyl-2-butyl ester	767.04	796.40	768.56	807.60
216	*cis*-2-butene	416.88	421.84	415.04	394.80	482	acetic acid, *n*-amyl ester	856.00	860.48	853.68	867.44
401	propane	300.00	256.40	300.56	344.32	483	formic acid, hexyl ester	876.00	906.40	922.40	857.12
402	*n*-undecane	1100.00	1098.08	1054.56	962.32	484	acetic acid, 3-pentyl ester	802.00	829.68	818.00	772.00
405	3-methylhexane	676.96	684.72	683.76	682.08	485	acetic acid, 2-pentyl ester	800.00	831.92	809.52	818.00
406	3-ethylhexane	773.04	780.80	774.40	755.44	486	acetic acid, isoamyl ester	836.96	828.48	823.04	867.44
407	2,3-dimethylhexane	761.04	767.12	755.44	753.12	487	isobutyric acid, butyl ester	912.00	921.76	932.64	899.36
408	2-methyl-3-ethylpentane	762.00	766.88	760.40	752.00	488	propanoic acid, amyl ester	964.00	954.16	969.20	952.24
409	2,2,3,3-tetramethylbutane	728.00	711.44	708.88	745.76	489	propanoic acid, isoamyl ester	930.96	918.72	934.24	917.76
410	2-methyloctane	864.00	864.48	869.36	902.48	490	propanoic acid, 2-pentyl ester	892.96	924.96	919.92	921.52
411	3-methyloctane	871.04	873.76	871.12	862.80	491	acetic acid, 4-methyl-2-pentyl ester	854.00	897.52	868.00	872.64
412	4-methyloctane	859.04	870.32	859.92	868.64	492	propyl methyl ether	506.00	491.36	513.04	560.08
413	2,3-dimethylheptane	856.00	856.96	843.04	833.60	493	diethyl ether	486.00	486.88	514.80	565.20
414	2,4-dimethylheptane	822.00	836.24	825.60	830.88	494	methyl butyl ether	610.00	598.72	618.64	661.20
415	2,5-dimethylheptane	832.96	840.88	827.44	824.16	495	methyl *tert*-butyl ether	554.96	539.84	587.44	601.68
416	2,6-dimethylheptane	826.96	832.88	822.80	853.20	496	isobutyl methyl ether	569.04	571.60	580.24	670.88
417	1,1-dimethylcyclopentane	673.04	685.52	666.16	698.48	497	*tert*-butyl ethyl ether	606.96	612.72	612.32	733.12
418	cycloheptane	786.00	793.36	795.92	698.32	498	diisopropyl ether	586.96	633.44	613.20	700.64
419	cyclooctane	915.04	902.24	900.32	842.64	499	isopropyl propyl ether	630.96	658.64	639.28	735.52
420	methylcycloheptane	859.04	863.52	856.88	831.68	500	*tert*-butyl isopropyl ether	644.00	671.76	650.64	753.60
421	pethylcyclopropane	808.00	821.60	789.92	891.28	501	7-methyl-1-octene	840.96	850.40	861.68	882.40
422	1-*trans*-4-dimethylcyclohexane	798.00	807.68	810.32	804.24	502	6-methyl-1-octene	846.96	859.12	852.88	837.44
423	1-*cis*-4-dimethylcyclohexane	805.52	807.60	804.72	800.24	503	5-methyl-1-octene	838.96	855.52	849.36	839.92
424	cyclononane	1034.00	1016.88	1001.04	952.72	504	4-methyl-1-octene	840.00	854.40	845.52	839.84
425	methylcyclooctane	984.0097	1.04	969.60	930.00	505	3-methyl-1-octene	832.96	850.64	836.64	841.92

Table 1. Continued

ID	structure name	experimental	MLR	BPE ANN	CP ANN	ID	structure name	experimental	MLR	BPE ANN	CP ANN
506	2-methyl-1-octene	868.00	843.68	853.60	874.32	539	1-*trans*-3-pentadiene	515.04	500.16	506.56	478.96
507	*cis*-2-nonene	894.00	885.60	892.88	869.12	541	2-methyl-1,3-butadiene	497.04	482.56	493.60	500.64
508	*trans*-2-nonene	890.00	883.52	899.60	883.20	543	1,3-hexadiene	615.04	583.68	596.56	560.00
509	1-nonene	876.96	883.92	890.72	874.32	544	2,3-dimethyl-1,3-butadiene	612.00	553.92	570.32	546.80
510	3,4,4-trimethyl-2-pentene	748.96	725.20	758.48	774.48	545	2-methyl-1,4-pentadiene	558.96	565.20	591.52	590.64
511	6-methyl-1-heptene	790.96	754.96	755.92	755.04	546	3-methyl-1,4-pentadiene	532.00	586.24	617.92	581.60
512	2,3-dimethyl-2-hexene	788.96	756.40	767.52	756.16	547	*cis*-1,5-heptadiene	689.04	688.32	700.96	707.36
513	2-ethyl-1-hexene	776.00	751.68	738.56	749.60	548	*trans*-1,5-heptadiene	682.00	687.76	706.96	709.68
514	2,5-dimethyl-1-hexene	740.96	720.88	730.56	723.36	549	2-methyl-1,5-hexadiene	734.00	650.48	674.72	685.76
515	2,5-dimethyl-*trans*-3-hexene	697.04	733.60	728.24	739.36	550	4-methyl-1,5-hexadiene	736.00	663.28	664.88	676.24
516	2-methyl-*trans*-3-heptene	732.96	760.40	750.64	747.60	551	1,7-octadiene	764.00	774.48	795.36	793.04
517	*trans*-2-octene	798.00	787.20	790.80	792.64	552	benzene	643.04	661.68	688.40	772.72
518	*cis*-2-octene	802.00	787.20	782.96	787.44	553	toluene	750.00	767.76	740.88	790.72
519	*trans*-4-octene	784.00	788.80	776.48	793.04	554	ethylbenzene	840.00	873.44	834.80	882.64
520	2-methyl-1-heptene	776.00	759.68	765.92	747.60	555	1,2-dimethylbenzene	875.04	869.76	894.16	849.92
521	3-methyl-1-heptene	742.00	761.68	749.92	760.80	556	1,3-dimethylbenzene	854.00	851.12	861.92	865.20
522	4-methyl-1-heptene	757.04	764.08	772.80	755.60	557	1,4-dimethylbenzene	856.00	847.76	844.56	846.00
523	5-methyl-1-heptene	750.00	767.12	764.24	758.16	558	isopropylbenzene	913.04	959.60	953.60	971.68
524	2,3,3-trimethyl-1-butene	626.00	629.12	631.84	637.52	559	*n*-propylbenzene	943.52	979.04	937.60	912.40
525	2-methyl-*trans*-2-hexene	648.00	671.12	680.00	664.88	560	*n*-butylbenzene	1042.00	1088.64	1045.68	1042.80
526	*cis*-3-methyl-2-hexene	701.04	676.40	683.28	674.24	561	*tert*-butylbenzene	986.00	1004.40	1019.68	1016.40
527	*trans*-3-methyl-2-hexene	694.00	674.88	686.32	677.92	562	1-methyl-4-isopropylbenzene	1019.04	1014.08	1011.76	1000.72
528	*cis*-5-methyl-2-hexene	670.00	668.96	667.20	666.80	563	1-methyl-3-isopropylbenzene	1010.00	1018.24	1030.24	1004.32
529	*trans*-5-methyl-2-hexene	660.96	668.80	670.40	669.60	564	1-methyl-4-*n*-propylbenzene	1046.48	1041.60	1022.56	1023.12
530	*cis*-3-heptene	704.00	696.32	690.32	689.60	565	1-methyl-3-*n*-propylbenzene	1042.00	1043.28	1025.28	1042.40
531	*trans*-3-heptene	699.04	695.44	687.84	691.04	566	1-methyl-2-ethylbenzene	973.04	959.36	976.56	979.36
532	propane	300.00	261.12	298.32	324.96	567	1,3-diethylbenzene	1038.48	1041.52	1051.12	1043.12
533	2-methylpropene	385.04	380.56	412.32	403.84	568	1,2-diethylbenzene	1051.04	1049.44	1065.92	1023.12
534	1-butene	386.00	402.08	382.16	404.80	569	1,3,4-trimethylbenzene	969.04	919.84	956.16	942.96
535	1,3-butadiene	390.00	398.32	412.72	403.84	570	1,2,4-trimethylbenzene	988.00	934.40	972.48	1009.12
536	1,4-pentadiene	464.00	495.92	476.32	512.24	571	1,2,3-trimethylbenzene	1016.00	949.12	979.20	976.24
538	1-*cis*-3-pentadiene	524.00	500.16	506.56	478.96						

the prediction and nine times for the creation of the model. Average root mean squared error (RMS) and correlation values obtained with the cross-validation procedure were used for the evaluation of prediction capabilities of individual models. The probability of having biased prediction results is minimized because we are using the average of the prediction values of 10 test sets.

RESULTS AND DISCUSSION

Topological indices have been calculated using Codessa software. Afterward an optimal *n*-parameter MLR model was selected with up to 30 descriptors. The influence of the dimension of the MLR model on its prediction capabilities has been tested by the LOO cross-validation procedure. The squared correlation coefficients (r^2) for retrieved and predicted values are shown in Figure 1. In all later model validations the prediction results originate from the 10-fold cross-validation procedure where each object is tested exactly once because the 10 subsets cover the entire data set.

It can be seen from Figure 1 that r^2 for retrieved values is improved by increasing the number of parameters, whereas the same coefficient for predicted data reaches a maximum when the best 16 descriptors are used for the creation of a MLR model. Additional parameters decrease r^2 of the straight line predicted versus experimental values. It was deduced that only the first 16 descriptors contain structural features that have some influence on gas-chromatographic retention processes and the remaining descriptors do not represent any new information and are introducing noise. Therefore a 16-parameter MLR model was used for the prediction of retention indices for nonpolar stationary phases (Table 1, column 2). The selected structural descriptors together with coefficients of the MLR model are presented in Table 2. The

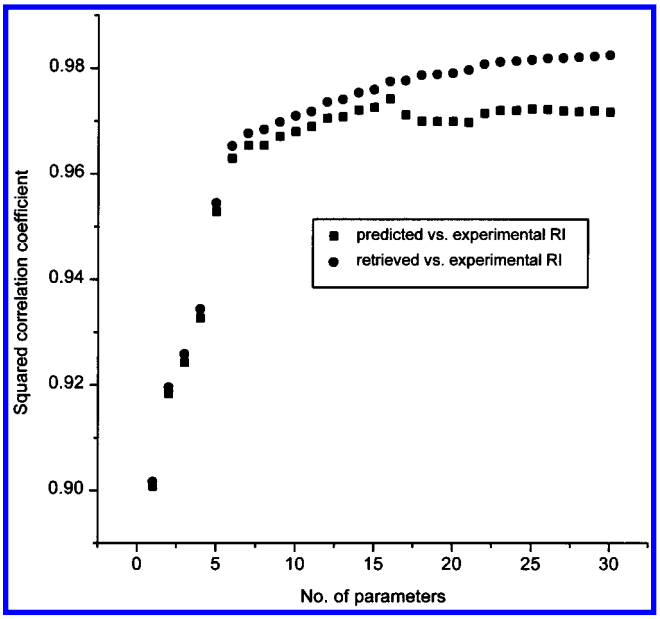

Figure 1. The squared correlation coefficients (r^2) for linear regressions: retrieved RI vs experimental and predicted RI vs experimental values obtained with MLR model.

regression parameters reported in Table 2 have been obtained from the regression model constructed with all 381 compounds.

On the basis of close inspection of Table 2, some conclusions about the constructed MLR model can be made. A chemical structure was represented by a 16-dimensional vector, the components of which were 10 topological and 6 constitutional descriptors. Of 10 topological indices, six were connectivity indices based on the degree of graph vertices (Randić indices[7]) and value of the atom valences δ_i (Kier and Hall indices[10]), two were topological shape indices,[27]

Table 2. Selected Structural Descriptors and Coefficients of Best MLR Model ($r^2 = 0.978$, $F = 992$, $s^2 = 484$)

no.	coefficient	type of descriptor
0	395.12	intercept
1	−46.14	Kier and Hall index (order 1)
2	−82.71	relative molecular weight
3	181.21	number of benzene rings
4	−3335.50	Randić index (order 1)
5	3677.00	number of O atoms
6	44.90	Kier and Hall index (order 3)
7	−3.77	gravitation index (all bonds)
8	−624.18	Randić index (order 2)
9	−121.36	Kier shape index (order 1)
10	89.14	Balaban index
11	−3947.20	relative number of rings
12	3.22	Kier shape index (order 3)
13	−81.32	Randić index (order 3)
14	1.98	complementary information content (order 0)
15	3454.50	number of C atoms
16	−1481.50	Randić index (order 0)

one was derived from the informational theory,[28] and the last one proposed by Balaban was analogous to the connectivity index but is based on the graph distance matrix.[9] Every mentioned topological index describes the molecular graph in its own way, but they all mirror the molecular branching pattern. Since we tried to model the gas-chromatographic retention indices for nonpolar stationary phases, where differences are expected to be due to differences in size and branching of the studied compounds, the selection of the above descriptors was expected.

In addition to the graph-theoretical indices, six constitutional descriptors were selected. They serve for the further discrimination between chemical compounds in cases where plain topological indices did not satisfactorily distinguish between similar chemical structures. The gravitation index (all bonds), the number of rings, of benzene rings, and the numbers of C and O atoms, together with the molecular weight, were used. These parameters presumably encode some information about structural features that are not well described by topological indices used in the study, that is, presence of heteroatoms, multiple bonds, and the aromatic nature of particular compounds.

Since we wanted to check if the use of ANNs could improve prediction results of the models based on classical MLR technique, the same set of structural descriptors were used also for the creation of ANN models. BPE algorithm with one hidden layer was tested first. Optimization of learning rate, momentum, architecture of ANN, and number of learning epochs has been done. During all optimization steps the 10-fold cross-validation procedure was performed to evaluate prediction abilities of the obtained models. The optimal learning rate and momentum remained constant during the whole learning process and were set to 0.1 and 0.2, respectively. Afterward, the influence of BPE ANN's architecture on prediction results was studied by changing the number of neurons in the hidden layer. All ANNs were trained in 100 000 epochs. The average RMS values of the 10-fold cross-validation procedure are shown in Figure 2. Since the optimal parameters for the creation of ANNs were determined according to 10-fold cross-validated prediction results, the overtraining effect was minimized. It was expected that the resulting ANN models were general within the used experimental domain defined by the studied data set.

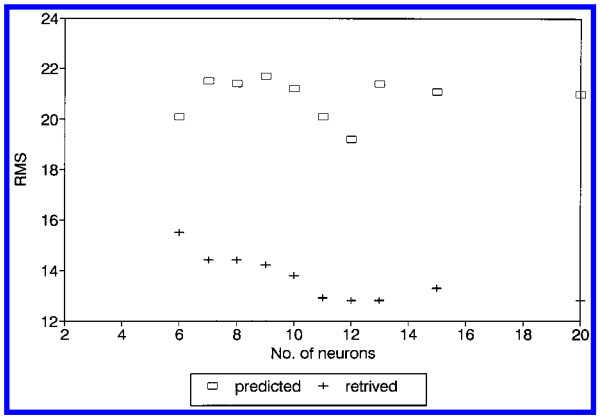

Figure 2. The influence of the number of neurons in the hidden layer of BPE ANN on the average RMS values for the test sets during 10-fold cross-validation procedure.

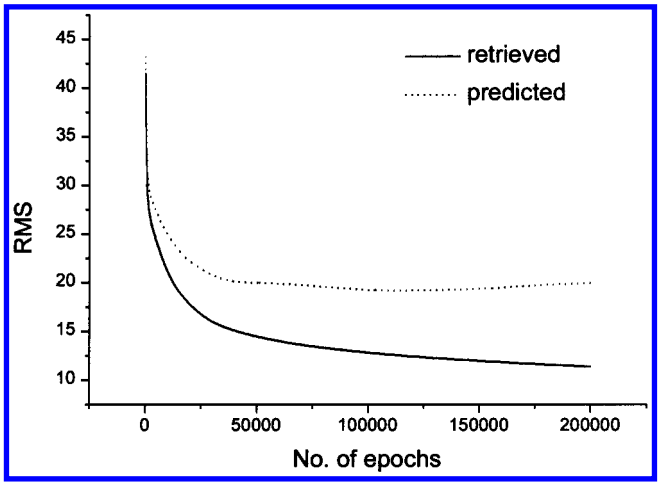

Figure 3. Optimization of number of learning cycles for the BPE ANN with the architecture 16−12−1.

The best prediction results were observed for the ANN with 12 neurons in the hidden layer. The average RMS values obtained by the 10-fold cross-validation procedure were calculated every 100 cycles. The lowest average RMS values for different architecture of BPE ANN were found at the end of the learning procedure, which means that the ANNs were not overtrained.

Since ANNs were not fully trained, the number of learning cycles was increased up to 200 000 epochs for the best three ANNs. Every 50 learning cycles the average training and prediction RMS values were calculated. The average RMS value for the training sets decreased with an increased number of learning cycles, whereas the same value for the prediction set first decreased until it reached its minimum and then started increasing. The best average RMS value was found for the ANN with 12 hidden neurons, which was trained in 120 000 epochs (Figure 3.). The observed result was expected because it is known that increasing the number of learning epochs could result in overtraining of ANNs. Although the minimum average RMS error was found at 120 000 learning cycles, it must be mentioned that deviation of $< 0.5\%$ of minimum average RMS value was found between 100 000 and 135 000 epochs. Therefore we could say that the developed model is robust within the mentioned range of training time. Using the optimized BPE ANN model, the retention indices were predicted by the 10-fold cross-validation procedure (Table 1, column 3).

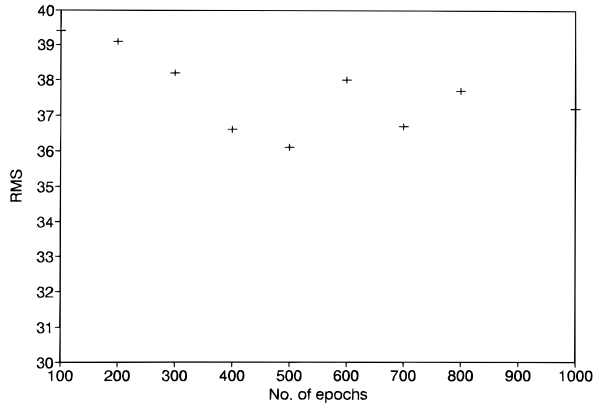

Figure 4. Influence of number of learning cycles on average prediction RMS values for CP ANN with 25 × 25 neurons.

Almost the same optimization steps as for BPE ANN were repeated also for the CP ANN. The described 10-fold cross-validation procedure is very time consuming; therefore we could not use it for the optimization of the architecture of CP ANN where training of the ANN is a much longer operation. Different configurations of ANN were formed with the whole data set, and afterward the same data set was used for the retrieval of retention indices. A plot of retrieved versus experimental data has been plotted for each network. The best training results were observed for the CP ANN with the 15 × 15, 20 × 20, and 25 × 25 neurons. All three ANNs were used for further thorough inspection.

To prevent overtraining, the influence of the number of learning cycles on prediction capabilities of selected CP ANNs was tested. The average 10-fold cross-validated RMS values for all three CP ANNs were calculated. The number of learning cycles was changed from 100 up to 1000 epochs. The lowest value was observed for the CP ANN with 25 × 25 neurons at 500 learning cycles (Figure 4.). Predicted retention indices obtained by optimal CP ANN model are shown in Table 1, column 4.

At the end, the prediction capabilities of MLR and both ANN models have been inspected using the 10-fold cross-validation procedure. The graphs of predicted values versus experimental retention indices for all three models are shown in Figure 5.

The best prediction results were observed for BPE ANN model, although the structural descriptors were selected to fit the MLR model. CP ANN gave the worst prediction results, which is due to the sensitivity of the model to large differences in retention indices for similar structure representations. When a created CP ANN model is used for the prediction of retention index for an unknown compound, the predicted retention index is taken from the output layer at the position of the neuron in the Kohonen layer that is most similar (although not equal) to the representation vector of the unknown. Predicted retention index is simply the discrete value of the retention index of the most similar compound included in the training procedure, or an average retention index value of those compounds from the training procedure occupying the same neuron. The prediction error is greater in the parts of CP ANN where differences between weights in the output layer for the neighboring neurons are larger. Only when the unknown compound triggers the neuron that was not occupied during the training can some retention index values different from those in the training set occur.

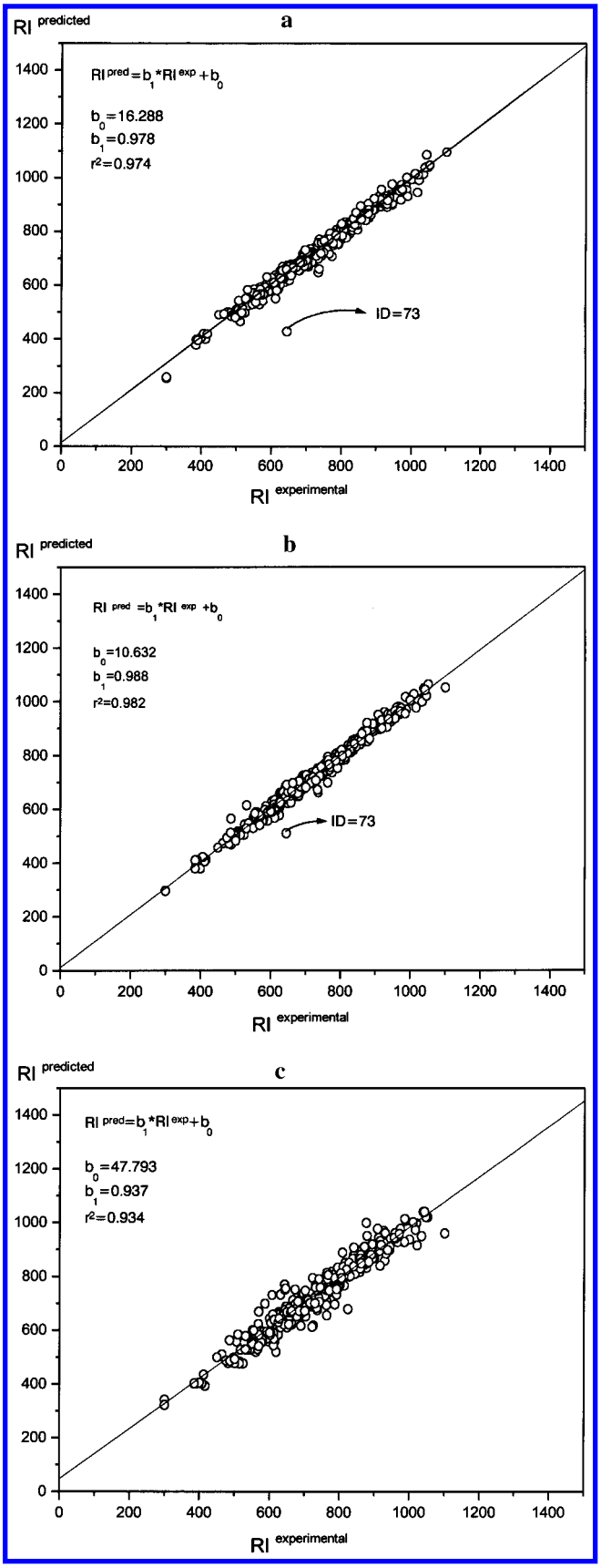

Figure 5. Predicted vs experimental retention indices (RI) for (a) MLR model, (b) BPE ANN model with network architecture 16−12−1, trained by 120 000 training cycles, and (c) CP ANN model having 25 × 25 neurons, trained by 600 training cycles.

Visual inspection of curves predicted versus experimental values obtained by MLR and BPE ANN models discovers

Table 3. Prediction Results Using 10-Fold Cross-Validation Procedure

	MLR				BPE ANN				CP ANN			
	RMS	r^2	b_1	b_0	RMS	r^2	b_1	b_0	RMS	r^2	b_1	b_0
AVG[a]	22.5	0.974	0.978	10	19.2	0.981	0.993	5	36.1	0.938	0.938	29
s [b]	6.0	0.016	0.032	17	3.3	0.008	0.033	16	8.8	0.027	0.034	18
RSD[c] [%]	26.6	1.6	3.3	180	17.1	0.9	3.3	325	24.5	2.9	3.6	64

[a] AVG, average value of 10-fold cross-validation procedure. [b] s, standard deviation of 10-fold cross-validation procedure. [c] RSD, relative standard deviation of 10-fold cross-validation procedure.

one common outlier, 3-methyl-2-butanol (no. 73). The same outlier cannot be observed in the CP ANN model because of larger uncertainty of the model. The root squared difference between predicted and experimental values for 3-methyl-2-butanol exceeded 200 retention time units. When the outlier was removed from prediction and all learning sets, the average RMS values obtained by the 10-fold cross-validation procedure for MLR and BPE ANN decreased to 19.9 and 17.5 retention index units, respectively. The retention index for 3-methyl-2-butanol was obtained from two independent sources. The main reason for large error is probably not of an experimental nature but lies in an inability of selected structural descriptors to describe polar interactions. That is why we did not remove this particular compound from the data set for the final evaluation of developed models. In the presence of polar interactions such large errors in prediction obtained by the described models can be expected. The average values of 10-fold cross validation results are presented in Table 3. For each of the three final models (MLR, BPE ANN, CP ANN) the RMS values and regression parameters, that is, r^2, b_1 and b_0, reflecting linear relation of predicted versus experimental retention indices were determined.

The final prediction results, shown in Table 3, were compared with values reported in the literature from similar studies.[14,15] It must be noted that the data set in the present study was extended compared with our previous one[15] with two new groups of compounds, that is, dienes and aromatic compounds. The proposed 16-component description vector fulfilled complete selectivity of the structural code, which was one of the drawbacks of our previously developed ANN model. Although the complete selectivity of the structural code was obtained, the average cross-validated RMS value for CP ANN did not improve significantly, that is, from 36.6 [15] to 36.1. This unexpected result could be explained by extension of the data set with the new groups of compounds. It is believed that newly introduced dienes and aromatic compounds cannot be appropriately coded by plain topological descriptors, which are able to encode molecular shape and branching. In these two groups of compounds the electronic environment of the molecules, which is not represented by the descriptors used in the study, presumably plays an important role in the intermolecular solute−solute, solute−stationary phase, and solute−mobile phase interactions.

On the other hand, the prediction results were improved dramatically when MLR or BPE ANN models were used for the modeling of the retention indices. Obtained average RMS values for MLR and BPE ANN models were 40% and 50% better than our previously developed CP ANN model. The prediction results were also better than in a similar study[14] where the authors used a complicated electrotopological description code with 29 descriptors. Their RMS values for the prediction set for MLR and BPE ANN models were 23.4 and 23.6, respectively. The RMS values for the prediction set obtained by our MLR and BPE ANN models (22.5 and 19.2, respectively) are lower, although the data set was larger and more diverse structures were included.

CONCLUSIONS

The aim of our work was to investigate the coding capabilities of topological description indices for QSRR studies. The topological indices were selected for the representation of the chemical structures because of their simplicity and effectiveness. They are easy to calculate and therefore could be used also by chromatographers who are not expert in structural representation. A very heterogeneous data set was selected to test whether our description vector is capable of encoding different types of interactions that are responsible for the retention properties.

Of 57 topological descriptors, 16 were selected on the basis of prediction results of the MLR model. A complete selectivity of the description vector was achieved. Because of time-consuming ANN calculations, the optimization procedure of the structural representation vector was not repeated for the ANN models. Instead, the same set of structural descriptors was used for the creation of BPE and CP ANNs. The implication of structural descriptors optimized on the basis of the MLR model for the creation of the ANN models has also been tested.

The worst results were obtained by the CP ANN model. The MLR model gave better results than, to our best knowledge, any previous similar study without the use of complicated electrotopological and quantum-chemical descriptors or even physical properties such as the boiling points of the molecule. Although the description vector was optimized for the MLR model, the BPE ANN model improved the prediction results. The average RMS value in a 10-fold cross-validation procedure was 19.2 retention time units. The proposed system for the prediction of gas-chromatographic retention indices on a nonpolar stationary phase is robust. Because of its simplicity it is suitable also for routine work. It can solve identification problems where unspecific fragmentation in homologous series occurs in a MS detector and hinders identification of organic compounds during GC-MS analysis.

The worst prediction results were observed for the compounds with heteroatoms and multiple double bonds (aromatic substances). Presumably, in both cases the electronic environment of the molecules plays an import role in solute−solute, solute−stationary phase, and solute−mobile phase interactions, which are responsible for the separation processes. This result was expected because it is known that

one of the main disadvantages of commonly used topological indices is representation of heteroatoms and multiple bonds. These drawbacks could be eliminated by including some electrotopological indices in the description set.

ACKNOWLEDGMENT

The financial support of Slovenian Ministry of Science and Technology through research grants no. J1-7373 (group at Faculty of Chemistry and Chemical Technology) and no. J1-7030 (the group at NIC) is gratefully acknowledged.

REFERENCES AND NOTES

(1) Bernejo, J.; Canga, J. S.; Gayol, O. M.; Guillen, M. D. Utilization of Physicochemical Properties and Structural Parameters for Calculating Retention Indices of Alkylbenzenes. *J. Chromatogr. Sci.* **1984**, *22*, 252.
(2) Kaliszan, R. Quantitative Relationship between Molecular Structure and Chromatographic Retention. Implications in Physical, Analytical, and Medical Chemistry. *CRC Crit. Rev. Anal. Chem.* **1986**, *16*, 323−383.
(3) Katrizky, A. R.; Gordeeva, E. V. Traditional Topological Indices vs Electronic, Geometric, and Combined Molecular Descriptors in QSAR/QSPR Research. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 835−857.
(4) Buydens, L.; Massart, D. L.; Geerlings, P. Prediction of Gas Chromatographic Retention Indexes with Topological, Physicochemical, and Quantum Chemical Parameters. *Anal. Chem.* **1983**, *55*, 738−744.
(5) Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69*, 17.
(6) Hosoya, H. Topological Index. A Newly Proposed Quantity Characterizing the Topological Nature of Structural Isomers of Saturated Hydrocarbons. *Bull. Chem. Soc. Jpn.* **1971**, *44*, 2332−2339.
(7) Randić, M. On Characterization of Molecular Branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609−6615.
(8) Bonchev, D.; Trinajstić, N. J. *Chem. Phys.* **1977**, *67*, 4517.
(9) Balaban, A. T. Highly Discriminating Distance-Based Topological Index. *Chem. Phys. Lett.* **1982**, *89*, 399−404.
(10) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; Wiley: New York, 1986.
(11) Garcia-March, F. J.; Anton-Fos, G. M.; Perez-Gimenez, F.; Salabert-Salvador, M. T.; Cercos-del-Pozo, R. A.; de Julian-Ortiz, J. V. Prediction of Chromatographic Properties for A Group of Natural Phenolic Derivatives by Molecular Topology. *J. Chromatogr. A* **1996**, *719*, 45−51.
(12) Sekušak, S.; Sabljić, A. Calculation of Retention Indices by Molecular Topology. III. Chlorinated Dibenzodioxins. *J. Chromatogr.* **1993**, *628*, 69−79.
(13) Heinzer, V. E. F., Yunes, R. A. Using Topological Indices in the Prediction of Gas Chromatographic Retention Indices of Linear Alkylbenzene Isomers. *J. Chromatogr. A* **1996**, *719*, 462−467.
(14) Bruchmann, A.; Zinn, P.; Haffer, Chr. M. Prediction of Gas Chromatographic Retention Index Data by Neural Networks. *Anal. Chim. Acta* **1993**, *283*, 869−880.
(15) Pompe, M.; Razinger, M.; Novič, M.; Veber, M. Modelling of Gas Chromatographic Retention Indices Using Counterpropagation Neural Networks. *Anal. Chim. Acta* **1997**, *348*, 215−221.
(16) *Gas Chromatographic Data Compilation*; ACTM: Philadelphia, 1971.
(17) Karelson, M.; Lobadov, V. S.; Katrizky, A. R. Qunatum-Chemical Descriptors in QSAR/QSPR Study. *Chem. Rev.* **1996**, *96*, 1027−1043
(18) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. CODESSA Manual. University of Florida: Gainesville, FL, 1995.
(19) Randić, M. On Characterization of Chemical Structure. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 672−687.
(20) Egolf, L. M.; Jurs, P. C. Prediction of Boiling Points of Organic Heterocyclic Compounds Using Regression and Neural Networks Techniques. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 616−625.
(21) Sutter, J. M.; Peterson, T. A.; Jurs, P. C. Prediction of Gas Chromatographic Retention Indices of Alkylbenzene. *Anal. Chim. Acta* **1997**, *342*, 113−122.
(22) Rumelhart, E.; Hinton, G. E.; Williams, R. J. Learning Internal Representations by Error Back-Propagation. In *Distributed Parallel Processing: Explorations in the Microstructures of Cognition*; Rumelhart, D. E.; MacClelland, J. L., Eds., Vol. 1; MIT Press: Cambridge, MA, 1986; pp 318−362.
(23) Hecht-Nielsen, R. Application of Counter-Propagation Networks. *Neural Networks* **1988**, *1*, 131−140.
(24) Zupan, J.; Gasteiger J. *Neural Networks for Chemists*; VCH Verlag: Weinheim, 1993.
(25) Geman, S.; et al. Neural Networks and The Bias/Variance Dilemma. *Neural Computation* **1992,** *4*, 1−48.
(26) Efron, B.; Tibshirani, R. J. *An Introduction to the Bootstrap*; Chapman & Hall: New York, 1993; pp 239−241.
(27) Kier, L. B. In *Computational Chemical Graph Theory*; Rouvray, D. H., Ed.; Nova Science Publishers: New York, 1990; pp 151−174.
(28) Bonchev, D. *Information Theoretic Indices for Characterization of Chemical Structures*; Research Studies Press: New York, 1983.

CI980036Z