Prediction of Thermodynamic Parameters in Gas Chromatography from Molecular Structure: Hydrocarbons†

Matevž Pompe,*,‡ Joe M. Davis,§ and Clint D. Samuel§

Faculty of Chemistry and Chemical Technology, University of Ljubljana, Aškerčeva 5, 1000 Ljubljana, Slovenia, and Department of Chemistry and Biochemistry, Southern Illinois University, Carbondale, Illinois 62901

Received December 17, 2003

Theoretical prediction of gas-chromatographic retention times could be used as an additional method for a more accurate identification of organic compounds during GC/MS analysis. Two separate quantitative structure−property relationship models were introduced for the calculation of thermodynamic values ($\Delta H°$, $\Delta S°$) for aliphatic and aromatic hydrocarbons. These values are required for the calculation of retention times in temperature programmed gas chromatography. Seven-descriptor and five-descriptor MLR models were selected for the calculation of $\Delta H°$ and $\Delta S°$ values, respectively, based on the best cross-validation abilities. The final prediction capabilities of the models were evaluated by a test set procedure. RMS errors calculated from the test set were 207 cal mol^{-1} and 0.58 cal mol^{-1} K^{-1} for $\Delta H°$ and $\Delta S°$ prediction models, respectively. To evaluate the error of the models represented in the time scale, several chromatograms were simulated using experimental Pro ezGC and theoretically calculated thermodynamic data. Afterward a standard deviation of retention time residuals was calculated. It was found out that, although the standard deviation varies from one chromatographic condition to another, the ratio between the standard deviation and the maximum available separation space for the particular set of organic compounds remains constant and was around 5% of the maximum separation space available at selected chromatographic conditions. Our prediction model was able to accurately differentiate between the retention times of the consecutive compounds in the *n*-alkanes, 1-alkenes, and 2-alkenes homological series.

INTRODUCTION

Gas chromatographic separation followed by mass spectrometric detection is one of the most widely used methods for the quantification and identification of volatile organic compounds in different matrices. However, even when a mass spectrometer is used as a detector the unspecific fragmentation of homologous compounds in the ion source can hinder their identification. In such cases the identification of the compounds can be achieved using their retention properties; therefore, the substance is positively identified if there is a match in mass spectra as well as the retention time between the pure standard and the suspected compound. However, it is not always possible to obtain samples of a pure standard material for such comparisons especially in cases of nontarget analysis of flavors, environmental samples, etc. It is obvious that a development of a theoretical model for the prediction of retention properties would be beneficial.

In quantitative structure-retention relationship (QSRR) studies a chemical structure is transformed to the computer readable form, and afterward a quantitative correlation with the retention property is found. Using this technique several models were developed that predict isothermal gas chromatographic retention indices for the individual classes of the organic compounds[1−13] or for the large variety of them.[14−17] Other studies tried to build a predictive model for the evaluation of gas chromatographic retention times.[18−22] Whether we use theoretical models for prediction of retention indices or retention times a major problem arises when we want to use these data in order to identify organic compounds in the temperature programmed gas chromatography (TPGC). The limitations of using the retention indices in TPGC are described in the literature;[23] on the other hand, predicted retention times are restricted to fixed chromatographic conditions. This standardization is not convenient, since the variable temperature program is the key factor of the optimization in gas chromatographic separation.

It is clear that in order to use the developed prediction model as a helping tool in the real analytical laboratory the predicted retention quantity must be temperature independent and easily transferred from one analytical system to another. Thermodynamic parameters such as standard-state changes of enthalpy ($\Delta H°$) and entropy ($\Delta S°$) between mobile and stationary phases are temperature independent, at least within the temperature limits used in TPGC.[24] Both parameters can be transferred from one column to another having different column dimensions as long as they contain the same stationary phase. Several articles explain the procedure of how to use thermodynamic parameters in order to calculate gas chromatographic retention times[25−28] or peak widths[25,29] for organic species in TPGC.

Therefore it seems obvious to us that if we want to create a theoretical model, which will be used in an ordinary

* Corresponding author fax: ++ 386 61 125 8220; e-mail: matevz.pompe@uni-lj.si.

† Dedicated to Dr. George W. A. Milne a former long-term Editor-in-Chief of *JCICS*.

‡ University of Ljubljana.

§ Southern Illinois University.

10.1021/ci0304268 CCC: $27.50
Published on Web 02/27/2004

analytical laboratory, it must be able to predict the thermodynamic values of the individual organic compounds. These values can be further used for the calculation of basic chromatographic parameters such as retention times and peak widths. Since this is the first study of this kind we have selected the simplest case when only hydrocarbons are involved in the model. The main goal of our study was therefore a creation of the prediction models for the calculation of thermodynamic parameters such as $\Delta H°$ and $\Delta S°$ for aliphatic and aromatic hydrocarbons purely from the chemical structure.

EXPERIMENTAL SECTION

Calculation of $\Delta H°$ and $\Delta S°$ Values. Thermodynamic parameters $\Delta H°$ and $\Delta S°$ were obtained from the database included in the software Pro ezGC for WINDOWS (Restek Co., Bellefonte, PA). Software contains experimental thermodynamic retention values. Since these values are not readily accessible from the software we had to use the following procedure to extract the necessary information. The procedure uses the same thermodynamic equations that are used by Pro ezGC software; therefore, we were able to calculate back the initial thermodynamic values. Because we do not know the origin of the experimental data included in the Pro EzGC software we have used the expression "Pro ezGC" thermodynamic values for these data although it is known that they are somebody's experimental data.

Isothermal retention times (t_r) of 280 aliphatic and aromatic hydrocarbons were calculated at 40, 60, 80, 100, and 120 °C for the nonpolar stationary phase DB-1. It is known that in isothermal GC analysis, the retention factor (k'), the phase ratio (β), the retention time (t_r), the void time (t_0), and the thermodynamic distribution coefficient (K) are related by

$$k' = (t_r - t_0)/t_0 = \beta K \quad (1)$$

and the thermodynamic distribution coefficient is further related to standard state change of free energy ($\Delta G°$) between the mobile and the stationary phases by

$$K = \exp(-\Delta G°/RT) \quad (2)$$

where R is the gas constant and T is the absolute temperature. By replacing $\Delta G°$ in eq 2 with the expression $\Delta H° - T\Delta S°$ obtained from the definition of Gibbs energy, the following relation exists between the retention factor and enthalpy and entropy changes

$$\ln k' = \ln \beta + \ln K = \ln \beta - \Delta H°/RT + \Delta S°/R \quad (3)$$

A plot of $\ln k'$ vs T^{-1} is a line having slope $-\Delta H°/R$ and an intercept $\ln \beta + \Delta S°/R$. The phase ratio β was calculated from the geometry of the separation column by using eq 4

$$\beta = 4d_f/d_c \quad (4)$$

where d_f and d_c were the stationary phase film thickness (0.25 μm) and the column internal diameter (250 μm) of a 30-m capillary, respectively. Both calculated thermodynamic parameters ($\Delta H°$, $\Delta S°$) are shown in Table 1.

Calculation and Selection of the Structural Descriptors. Molecular structures were created by HyperChem. Afterward, MOPAC software[30] was used for the geometry optimization and calculation of net atomic charges. Using CODESSA software[31,32] 367 different topological, geometric, informational, electrostatic, electrotopological, and quantum-chemical descriptors necessary for the creation of the models were calculated. The descriptors employed in the study contain the information about the connections between atoms, symmetry, shape, branching, distribution of charge, and quantum-chemical properties of the molecules. It is obvious that we cannot use all these descriptors for the creation of a single prediction model, because most of the descriptors do not encode any structural feature that is responsible for the retention properties of the molecules and such a model would be most probably coincidental. Therefore a descriptor reduction procedure should be applied. Before we have performed the selection of structural indices we have divided our data set into two subsets by using a random number generator. A training set contained approximately 70% of organic compounds (195) and was used during descriptors selection as well as during the creation of the multiple linear regression (MLR) prediction models. The remaining 85 compounds were placed into the test set and were used for the final evaluation of the prediction capabilities of the developed MLR model. The compounds that formed the test set are marked in Table 1 by footnote a.

The selection of structural descriptors was accomplished by applying the heuristic optimization search in the CODESSA software as described below. An optimal subset of descriptors was selected by minimizing a cross-validation error of the MLR model. At the beginning all descriptors were omitted from the further study that show no variation between structures or were not defined for each compound included in the study or had squared correlation coefficient smaller than 0.01. The remaining descriptors were pairwise correlated. One of the descriptors was removed from the descriptor's set if the pairwise correlation coefficient exceeded 0.995. This procedure reduced the number of descriptors offered by CODESSA to approximately half. The final number of descriptors used for the stepwise selection of the best subset of structural indices for the calculation of $\Delta H°$ and $\Delta S°$ values were 118 and 143, respectively.

In the second step, all remaining descriptors were sorted in the decreasing order of the squared correlation coefficient of the corresponding simple linear regression model. All two-parameter MLR models were calculated, where the pairwise correlation coefficient between descriptors which were included in the same model did not exceed 0.99. The best 10 two-parameter MLR models were selected showing the highest F-values and were used as the working sets. Each of the chosen working sets was extended to the prescribed number of parameters included in the model by the stepwise selection procedure, where new descriptors were added to the model if they show correlation with the descriptors already included below 0.99 and if they improve the significance of the model, that is, the new F value was above $n/(n+1)\bullet F_{old}$. Here n is a number of descriptors in the new working set. A more detailed description of a stepwise selection procedure can be found elsewhere.[32,33] As the final result of the described stepwise addition procedure 10 correlations were selected with the highest r^2. The derived correlations were tested for their cross-validation capabilities by the leave-one-out cross validation procedure. The model

Table 1. Pro EzGC and Calculated Values for ΔH^0, ΔS^0, and t_r Obtained for DB1 Stationary Phase

		$\Delta H°$ (cal mol^{-1})		$\Delta S°$ (cal mol^{-1}K^{-1})		$t_{r\,(min)}$	
no.	compound name	Pro ezGC	calc	Pro ezGC	calc	Pro ezGC	calc
2	*n*-decane	−10539	−10568	−16.9	−16.6	20.4	21.6
3	*trans*-2-decene	−10609	−10529	−17.1	−16.9	20.6	20.4
4	*cis*-2-decene[a]	−10577	−10493	−16.9	−16.7	20.9	20.4
6	1-menthene	−10059	−10023	−15.3	−15.7	21.0	19.7
8	s-butylcyclohexane	−10349	−10063	−16.1	−15.4	21.1	20.6
9	*cis*-1-methyl-2-propylbenzene[a]	−10128	−10320	−15.5	−15.7	21.0	21.8
11	*n*-butylcyclohexane	−10229	−10210	−15.7	−15.9	21.2	20.5
13	indene	−9843	−9938	−14.6	−14.3	21.1	22.7
14	2,5-dimethylnonane	−10299	−10467	−15.9	−16.5	21.3	21.2
16	1-methyl-2-isopropylbenzene	−10136	−10155	−15.4	−15.7	21.3	20.6
17	4-phenyl-2-butene [a]	−10217	−10457	−15.6	−15.6	21.4	23.1
18	*m*-allyltoluene	−10224	−10201	−15.6	−15.5	21.4	21.2
19	2,2,3-trimethyloctane	−10278	−10242	−15.7	−16.0	21.5	20.8
20	butylcyclohexane[a]	−10221	−10230	−15.6	−15.9	21.5	20.7
21	*p*-allyltoluene	−10247	−10242	−15.6	−15.6	21.6	21.4
25	3-methyl-4-ethyloctane	−10436	−10465	−16.1	−15.9	21.9	22.5
27	1-methyl-3-*n*-propylbenzene[a]	−10313	−10298	−15.7	−15.9	21.8	21.2
28	4-ethylnonane[a]	−10073	−10710	−15.0	−16.5	21.8	23.1
29	5-methyldecane	−10073	−10766	−15.0	−16.8	21.8	22.9
30	*trans*-1,2-dimethylcyclooctane[b]	−9583	−9768	−13.7	−15.1	21.6	18.7
32	2,2,7,7-tetramethyloctane[a]	−10486	−10390	−16.2	−16.2	22.0	21.5
33	*o*-allyltoluene[a]	−10296	−10387	−15.6	−15.3	22.0	23.5
34	*p*-diethylbenzene	−10232	−10401	−15.4	−16.0	22.0	21.7
36	*n*-butylbenzene	−10267	−10293	−15.5	−15.9	22.0	20.8
37	1,4-diethylbenzene	−10302	−10401	−15.6	−16.0	22.0	21.7
39	*cis*-1,2-dimethylcyclooctane[a,b]	−9579	−9768	−13.6	−15.1	21.8	18.7
41	1,3-dimethyl-5-ethylbenzene	−10442	−10271	−16.0	−16.0	22.2	20.9
43	*o*-diethylbenzene	−10227	−10337	−15.4	−15.7	22.2	21.9
45	1,3-dimethyl-3-ethylbenzene[a]	−10349	−10138	−15.7	−15.7	22.4	20.6
46	5-ethylnonane[a]	−10772	−10652	−16.9	−16.4	22.5	22.7
48	1-phenyl-2-butene[a]	−10432	−10592	−15.9	−15.6	22.5	24.5
49	1-methyl-2-*n*-propylbenzene	−10257	−10320	−15.4	−15.7	22.4	21.8
51	2-methyldecane[a]	−10334	−10724	−15.6	−17.1	22.5	21.6
52	neopentylbenzene	−9961	−10494	−14.5	−15.5	22.4	24.0
53	4-methyldecane	−10479	−10758	−16.0	−16.8	22.6	22.6
54	1-*tert*-butyl-4-methylcyclohexane	−10474	−10083	−15.9	−15.1	22.6	21.9
55	methylcyclononane[b]	−9680	−9832	−13.7	−15.0	22.3	19.3
56	2,3-dimethylnonane	−10659	−10580	−16.5	−16.4	22.7	22.4
57	3-methyldecane[a]	−10363	−10826	−15.6	−16.8	22.7	23.2
58	ethylcyclooctane[a]	−9739	−10012	−13.8	−15.2	22.5	20.4
59	2,6-dimethylstyrene	−10271	−10482	−15.3	−15.4	22.7	24.2
60	1,2-dimethyl-4-ethylbenzene	−10434	−10382	−15.7	−15.8	23.0	22.1
61	1,3-dimethyl-4-ethylbenzene	−10405	−10324	−15.6	−15.9	23.0	21.5
62	*m*-ethylvinilbenzene	−10656	−10392	−16.3	−15.7	23.1	22.3
63	3-ethylnonane[a]	−10915	−10702	−17.0	−16.7	23.3	22.3
64	1-methylindane[a]	−10458	−10348	−15.7	−15.2	23.1	23.7
65	1,4-dimethyl-2-ethylbenzene[a]	−10500	−10317	−15.8	−15.7	23.2	21.9
66	1,2-dimethyl-4-ethylbenzene	−10428	−10382	−15.6	−15.8	23.2	22.1
67	*p*-ethylvinilbenzene	−10664	−10419	−16.2	−15.8	23.4	22.5
68	1-*tert*-butyl-3-methylbenzene	−10464	−10488	−15.6	−15.8	23.4	23.3
70	*tert*-penthylbenzene	−10268	−10336	−15.0	−15.7	23.5	22.0
72	1-methyl-3,5-diethylbenzene[b]	−10487	−10920	−15.6	−16.1	23.6	26.2
75	2,5-dimethylstyrene	−10461	−10499	−15.5	−15.5	23.6	24.1
76	*o*-ethylvinylbenzene	−10635	−10298	−16.0	−15.3	23.7	22.4
77	2,4-dimethylstyrene	−10443	−10389	−15.5	−15.6	23.8	22.8
79	4,5-diethyloctane[b]	−10526	−10851	−15.7	−15.7	23.9	26.4
80	1-ethyl-3-isopropylbenzene[a]	−10616	−10738	−15.9	−16.2	24.0	24.1
81	(1-methylbutyl)benzene[a]	−10557	−10771	−15.7	−15.8	24.0	25.1
82	2,3-dimethyl-1-ethylbenzene[a]	−10554	−10449	−15.7	−15.6	24.0	23.4
84	1,2-dimethyl-3-ethylbenzene	−10598	−10453	−15.8	−15.6	24.1	23.3
85	(2-methylbutyl)cyclohexane	−10711	−10562	−16.1	−15.7	24.1	24.1
88	1-methyl-2-butylcyclohexane	−10819	−10568	−16.4	−15.7	24.3	24.2
89	1-ethyl-2-isopropylbenzene	−10597	−10696	−15.7	−15.9	24.3	24.6
91	n-undecane[a]	−11142	−11146	−17.2	−16.7	24.6	25.9
92	1-ethyl-4-isopropylbenzene	−10709	−10786	−16.0	−16.2	24.5	24.5
94	1,2,4,5-tetramethylbenzene	−10543	−10382	−15.5	−15.7	24.5	22.5
95	1,4-divinylbenzene[a]	−10514	−10466	−15.4	−15.2	24.5	24.4
96	1,2-divinylbenzene	−10587	−10579	−15.6	−15.0	24.6	6.1
99	2-methyldecalin[a]	−10217	−10470	−14.5	−14.8	24.6	25.9
100	2-methyllindene	−10138	−10410	−14.1	−14.9	25.2	25.0
101	(1-methylethyl)cyclopentane	−9030	−9052	−15.6	−15.8	11.8	11.1
102	*n*-octane	−9381	−9449	−16.8	−16.7	11.6	12.2

Table 1. (Continued)

		$\Delta H°$ (cal mol^{-1})		$\Delta S°$ (cal mol^{-1}K^{-1})		$t_{r\,(min)}$	
no.	compound name	Pro ezGC	calc	Pro ezGC	calc	Pro ezGC	calc
105	*cis*,*cis*,*trans*-1,3,5-trimethylcyclohexane	−9283	−9301	−15.6	−15.5	13.8	14.1
111	*cis*,*trans*,*trans*-1,2,4-trimethylcyclohexane[a]	−9250	−9340	−15.5	−15.4	13.6	14.5
112	1,3-cycloheptadiene	−9277	−8928	−16.4	−15.4	11.8	10.8
113	cycloheptane[b]	−9286	−8807	−16.7	−15.9	11.1	8.9
114	*cis*-1,2-dimethylcyclohexane[a]	−8949	−9066	−15.1	−15.7	12.4	11.6
115	*cis*-1,3-dimethylcyclohexane	−8938	−8995	−15.4	−15.7	11.5	11.0
116	*cis*-1,4-dimethylcyclohexane	−8906	−9054	−15.3	−15.6	11.4	11.6
117	*trans*-1,2-dimethylcyclohexane[a]	−9529	−9044	−17.2	−15.7	11.9	11.4
118	*trans*-1,3-dimethylcyclohexane	−9148	−9002	−16.0	−15.9	11.7	10.7
119	1,3-dimethyl-1-cyclohexene	−9589	−9206	−17.0	−15.9	12.8	12.2
120	1,4-dimethyl-1-cylcohexene[a]	−9589	−9235	−17.0	−15.8	12.8	12.6
121	2,4-dimethyl-3-ethylpentane	−9260	−9311	−15.8	−15.7	13.1	13.5
123	1,3-dimethylbenzene	−9375	−9330	−15.8	−15.4	13.9	14.4
124	1,4-dimethylbenzene[a]	−9374	−9225	−15.8	−15.6	13.9	13.0
125	2,2-dimethylheptane	−9432	−9366	−16.6	−16.3	12.4	12.9
126	2,3-dimethylheptane	−9702	−9450	−16.8	−16.4	14.2	13.1
127	2,4-dimethylheptane	−9460	−9398	−16.6	−16.2	12.6	13.1
128	2,5-dimethylheptane[a]	−9538	−9455	−16.7	−16.4	13.2	13.3
129	2,6-dimethylheptane	−9541	−9373	−16.8	−16.5	12.9	12.2
130	3,3-dimethylheptane	−9406	−9481	−16.2	−16.0	13.2	14.4
131	3,4-dimethylheptane	−9567	−9479	−16.4	−16.2	14.1	13.8
133	3,5-dimethylheptane	−9692	−9462	−16.8	−16.2	14.1	13.5
136	4,4-dimethylheptane	−9344	−9372	−16.2	−16.0	12.6	13.4
137	2,3-dimethyl-1-heptene[a]	−9458	−9612	−16.3	−16.3	13.3	14.5
138	2,3-dimethyl-2-hexene[a]	−9263	−9315	−16.5	−16.5	11.4	11.9
139	ethenylcyclohexane[a]	−9170	−9222	−15.9	−15.8	12.1	12.5
140	4-ethenylcyclohexene	−8985	−9273	−15.2	−15.5	12.4	13.6
141	1-ethyl-1-methylcyclopentane	−9010	−9061	−15.8	−15.7	11.1	11.6
142	ethylbenzene[a]	−9220	−9349	−15.5	−15.4	13.4	14.4
143	ethylcyclohexane	−9059	−9219	−15.3	−15.8	12.6	12.5
144	4-ethylheptane	−9692	−9614	−16.7	−16.3	14.3	14.5
145	ethylidenecyclohexane	−9238	−9395	−15.3	−15.9	14.1	13.7
146	isopropylcyclopentane	−8939	−9032	−15.3	−15.9	11.7	10.9
147	1-methyl-2-ethylcyclopentane	−9053	−9127	−15.5	−15.8	12.2	11.7
148	methylcycloheptane	−9406	−9132	−16.0	−15.6	13.7	12.1
149	5-methylene-2-norbornene[b]	−9370	−9224	−16.8	−15.1	11.6	14.4
150	2-methyl-1-heptene	−9266	−9198	−16.7	−16.6	10.9	10.6
151	2-methyl-2-heptene	−9396	−9269	−16.9	−16.6	11.6	11.1
152	*cis*-4-methyl-2-octene	−9424	−9721	−16.1	−16.5	13.6	14.8
153	*trans*-4-methyl-2-octene	−9518	−9724	−16.2	−16.6	14.0	14.6
154	1,7-octadiene[a]	−9625	−9283	−17.9	−16.3	11.1	11.5
155	1-octene	−9295	−9353	−16.8	−16.6	11.1	11.7
156	*cis*-2-octene	−9422	−9418	−16.7	−16.6	12.1	12.2
157	*cis*-3-octene	−9352	−9409	−16.7	−16.7	11.6	11.7
158	*cis*-4-octene[a]	−9320	−9424	−16.7	−16.7	11.5	11.9
159	*trans*-2-octene	−9440	−9458	−16.9	−16.8	11.8	12.1
160	*trans*-3-octene	−9409	−9453	−16.9	−16.9	11.5	11.8
161	*trans*-4-octene	−9365	−9479	−16.9	−16.9	11.4	12.0
164	pentylcyclopropane	−8991	−9244	−14.9	−16.4	13.0	11.1
166	*n*-propycyclopentane[a]	−9204	−9216	−15.7	−16.1	12.7	11.8
168	i-propylcyclopentane	−9094	−9047	−15.8	−15.9	11.8	11.0
169	styrene	−9264	−9385	−15.2	−14.8	14.5	16.4
170	*cis*-1,1,3,4-tetramethylcyclopentane[a]	−8970	−9158	−15.3	−15.7	11.9	12.8
171	2,2,5,5-tetramethylhexane	−9168	−9353	−15.5	−15.7	13.0	14.7
173	2,2,3,3-tetramethylpentane	−9129	−9147	−15.2	−15.5	13.4	13.2
174	2,2,3,4-tetramethylpentane	−9052	−9124	−15.5	−15.5	12.2	12.8
175	2,3,3,4-tetramethylpentane[a]	−9185	−9220	−15.3	−15.5	13.7	13.6
176	1,1,2-trimethylcyclohexane	−9160	−9337	−14.8	−15.2	14.7	15.2
177	1,1,4-trimethylcyclohexane	−9253	−9260	−15.8	−15.5	13.1	13.9
178	1,2,3-trimethylcyclohexane	−9216	−9443	−15.0	−15.3	14.6	15.7
179	1,2,4-trimethylcyclohexane	−9302	−9393	−15.7	−15.4	13.6	15.0
180	1,3,5-trimethylcyclohexane[a]	−9146	−9296	−15.6	−15.5	12.7	14.0
182	1,3,5-trimethylcyclohexene	−9123	−9304	−14.9	−15.4	14.2	14.3
183	1,2,3-trimethylcyclopentane	−8858	−8984	−15.2	−15.8	11.3	10.8
185	1,1,3-trimethylcyclohexane	−9083	−9258	−15.3	−15.4	12.9	14.0
186	2,2,3-trimethylhexane	−9266	−9235	−16.0	−15.9	12.5	12.8
187	2,2,4-trimethylhexane[a]	−9077	−9134	−16.0	−15.9	11.1	11.9
188	2,2,5-trimethylhexane[a]	−9199	−9117	−16.5	−16.1	10.9	11.5
189	2,3,3-trimethylhexane	−9265	−9290	−15.8	−15.8	13.1	13.4
190	2,3,4-trimethylhexane	−9374	−9353	−15.9	−15.8	13.6	13.7
191	2,3,5-trimethylhexane[a]	−9321	−9226	−16.3	−16.1	12.2	12.2
192	2,4,4-trimethylhexane	−9120	−9174	−15.9	−15.9	11.8	12.3

Table 1. (Continued)

		$\Delta H°$ (cal mol^{-1})		$\Delta S°$ (cal mol^{-1}K^{-1})		$t_{r\,(min)}$	
no.	compound name	Pro ezGC	calc	Pro ezGC	calc	Pro ezGC	calc
193	3,3,4-trimethylhexane	−9264	−9344	−15.6	−15.7	13.6	13.9
194	vinylcyclohexane	−8903	−9214	−15.1	−15.6	12.0	12.7
195	4-vinylcyclohexene[a]	−8999	−9299	−15.2	−15.6	12.4	13.6
197	5-vinyl-2-norbornene[a,b]	−9212	−9676	−15.0	−15.0	14.6	18.4
198	*m*-xylene[a]	−9284	−9335	−15.6	−15.4	13.8	14.4
199	*p*-xylene	−9310	−9225	−15.6	−15.6	13.9	13.1
200	(1-methylethyl)benzene[a]	−9695	−9707	−15.9	−15.5	16.3	17.1
201	*n*-nonane[a]	−9988	−9969	−16.9	−16.6	16.0	16.7
202	(1,1-dimethylethyl)benzene	−10037	−10008	−15.8	−15.4	19.3	20.1
203	(1-methylethyl)cyclohexane	−9577	−9501	−15.4	−15.5	16.5	15.6
204	1-(1-methylethyl)cyclohexene	−9713	−9726	−15.7	−15.9	16.9	16.6
206	*trans*-1-methyl-4-(1-methylethyl)cyclohexane	−10067	−9838	−16.0	−15.4	19.1	18.9
207	1-methyl-*cis*-2-propylcyclopentane	−9593	−9520	−15.5	−15.7	16.5	15.1
208	1-methyl-2-ethylbenzene	−9858	−9908	−15.5	−15.4	18.7	19.2
209	1-methyl-4-ethylbenzene	−9731	−9851	−15.4	−15.8	17.9	17.7
212	5-methyl-3-ethylheptane[a]	−9697	−9999	−15.7	−16.0	16.8	18.6
213	*o*-methylstyrene[a]	−9853	−9895	−15.5	−15.2	18.7	19.6
214	*m*-methylstyrene	−9847	−9839	−15.3	−15.3	19.2	18.8
215	*p*-methylstyrene	−9851	−9851	−15.3	−15.4	19.3	18.8
217	butylcyclopentane	−9930	−9762	−16.2	−15.9	17.3	16.7
222	1,5-cyclooctadiene	−9353	−9384	−14.7	−15.5	16.5	14.5
223	cyclooctane[b]	−9557	−9107	−15.4	−15.5	16.3	11.9
229	2,2-dimethyloctane[a]	−9848	−9929	−16.2	−16.2	16.8	17.7
230	2,3-dimethyloctane	−10007	−9988	−16.1	−16.3	18.4	17.7
231	2,4-dimethyloctane	−10030	−9897	−16.9	−16.3	16.3	17.0
232	2,6-dimethyloctane	−10125	−9932	−16.8	−16.5	17.5	16.7
233	2,7-dimethyloctane[a]	−10130	−9920	−16.8	−16.6	17.4	16.4
234	3,3-dimethyloctane	−9994	−10008	−16.3	−16.0	17.5	18.8
235	3,4-dimethyloctane	−10012	−10033	−16.1	−16.1	18.4	18.5
238	3,6-dimethyloctane	−10109	−10001	−16.6	−16.3	17.8	17.8
240	4,4-dimethyloctane[a]	−10013	−9892	−16.0	−16.0	18.5	17.8
242	2,3-dimethyl-2-octene	−10177	−10341	−16.5	−16.4	18.6	20.5
243	ethenylcyclooctane	−9977	−9952	−15.8	−15.2	18.8	19.8
247	1-ethyl-3-methylbenzene[a]	−9860	−9848	−15.7	−15.7	18.0	18.0
249	3-ethyloctane	−10292	−10264	−16.7	−16.2	19.0	20.1
258	4-isopropylheptane[a]	−9726	−9871	−15.8	−16.0	16.8	17.5
259	methylcyclooctane[a,b]	−9109	−9436	−13.5	−15.3	17.8	15.4
261	3-methyl-3-ethylheptane	−9939	−9940	−16.0	−15.8	18.1	18.5
262	2-methyl-3-ethylheptane	−10063	−9981	−16.5	−15.8	17.8	18.8
263	2-methylnonane	−10338	−10215	−16.9	−16.8	18.9	18.3
264	4-methylnonane	−10310	−10191	−16.9	−16.6	18.8	18.6
265	5-methylnonane	−10314	−10187	−16.9	−16.6	18.7	18.5
266	*cis*-4-methyl-2-nonane	−9913	−10240	−16.0	−16.6	17.9	18.8
267	*trans*-4-methyl-2-nonane[a]	−10003	−10249	−16.1	−16.8	18.4	18.5
268	2-methyl-2-nonene[a]	−10411	−10384	−17.1	−16.8	18.9	19.7
269	*p*-methylstyrene	−9875	−9896	−15.3	−15.3	19.4	19.3
270	*m*-methylstyrene	−9740	−9849	−14.9	−15.3	19.2	19.0
271	*o*-methylstyrene	−9740	−9917	−14.9	−15.2	19.2	19.9
272	a-methylstyrene[a]	−9895	−9836	−15.6	−15.2	18.7	19.1
274	*cis*-2-nonene	−10017	−9934	−16.8	−16.6	16.5	16.2
275	*trans*-2-nonene	−10044	−9974	−17.0	−16.7	16.2	16.4
277	*trans*-octahydro-1H-indene[a]	−9959	−9630	−16.0	−15.2	18.1	17.4
278	1,1,3,3,5-pentamethylcyclohexane[a,b]	−9759	−9774	−15.5	−14.8	17.9	20.3
280	2-propenylbenzene[a,b]	−9766	−9847	−15.8	−15.0	17.1	19.6
283	propylbenzene	−9792	−9771	−15.7	−15.7	17.6	17.0
284	propylcyclohexane	−9730	−9706	−15.7	−15.7	17.1	16.6
285	4-propylheptane	−10204	−10081	−16.8	−16.1	18.0	18.8
287	1,2,3,4-tetramethylcyclohexane[a]	−9882	−9737	−15.8	−15.1	18.1	18.7
290	tert-1,1,3,5-tetramethylcyclohexane[a]	−9560	−9507	−15.4	−15.2	16.3	16.7
293	2,3,4,5-tetramethylhexane[a]	−9697	−9582	−15.7	−15.7	16.8	16.1
294	3,3,4,4-tetramethylhexane	−9780	−9635	−15.9	−15.4	16.9	17.5
295	1,2,4-trimethylbenzene	−9876	−9836	−15.3	−15.6	19.3	18.2
296	1,3,5-trimethylbenzene	−9833	−9732	−15.6	−15.8	18.2	16.9
297	2,2,6,6-trimethylheptane	−9836	−9857	−15.9	−15.8	17.5	18.3
298	3,3,5-trimethylheptane	−9662	−9701	−15.8	−15.8	16.3	16.8
299	3,4,5-trimethylheptane	−9954	−9840	−16.0	−15.7	18.0	17.9
301	dodecane	−11746	−11817	−17.5	−17.1	28.5	30.4
302	1-methyl-2-*n*-butylbenzene[a]	−10865	−10874	−15.7	−15.9	26.5	25.9
303	1-methyl-3-*n*-butylbenzene	−10959	−10850	−16.2	−16.1	26.0	25.0
304	1-methyl-4-*n*-butylbenzene	−10884	−10847	−15.9	−16.3	26.2	24.6
305	1-*tert*-butyl-3,5-dimethylbenzene	−11073	−10931	−16.1	−16.0	27.0	26.4
306	1-*tert*-butyl-4-ethylbenzene[a]	−11090	−11111	−16.1	−16.2	27.2	27.2

Table 1. (Continued)

		$\Delta H°$ (cal mol^{-1})		$\Delta S°$ (cal mol^{-1}K^{-1})		$t_{r\,(min)}$	
no.	compound name	Pro ezGC	calc	Pro ezGC	calc	Pro ezGC	calc
308	cycloundecane	−10339	−10530	−13.9	−14.7	27.5	26.1
309	1,3-dimethyl-5-*tert*-butylbenzene	−10899	−10891	−15.7	−16.1	27.0	26.0
312	1,2-diisopropylbenzene[a]	−11420	−11102	−16.6	−15.9	28.6	28.1
313	1,3-diisopropylbenzene[a]	−11044	−11193	−16.4	−16.2	26.1	27.9
314	1,4-diisopropylbenzene[a]	−11086	−11157	−16.2	−16.4	26.9	27.0
315	2,2-dimethyldecane	−11001	−11073	−16.5	−16.6	25.3	25.7
316	2,3-dimethyldecane	−11275	−11093	−16.8	−17.0	26.7	24.9
317	2,6-dimethyldecane	−11011	−11055	−16.6	−16.8	25.3	25.0
318	5,6-dimethyldecane[a]	−10762	−11014	−15.8	−16.6	25.3	25.4
319	4,6-dimethylundecane	−11526	−11628	−17.0	−16.9	28.2	29.3
320	5,7-dimethylundecane	−11508	−11568	−17.0	−16.7	28.1	29.3
321	1-dodecene	−11655	−11691	−17.4	−17.2	28.1	28.8
322	1-ethyldecalin	−10887	−10961	−15.0	−14.9	29.1	30.0
323	2-ethyldecalin[a]	−10881	−11043	−15.0	−14.9	29.0	30.5
324	5-ethylindan	−11079	−11026	−15.5	−15.8	28.9	27.8
325	1-ethyl-2-*n*-propylbenzene	−10850	−10862	−16.0	−15.9	25.7	25.8
326	1-ethyl-3-*n*-propylbenzene[a]	−10836	−10934	−16.0	−16.1	25.4	25.8
327	1-ethyl-4-*n*-propylbenzene	−11038	−10928	−16.4	−16.3	26.1	25.2
328	isopentylbenzene	−10684	−10708	−15.8	−16.0	24.8	24.2
340	1-methyltetralin[a]	−10870	−10826	−15.1	−15.2	28.5	27.7
341	2-methylundecane[a]	−11325	−11363	−16.9	−17.4	26.9	25.8
342	3-methylundecane	−11392	−11449	−16.9	−17.1	27.3	27.5
343	4-methylundecane	−11325	−11393	−16.9	−17.0	26.9	27.2
344	5-methylundecane	−11314	−11336	−16.9	−17.2	26.8	26.2
345	6-methylundecane	−11284	−11342	−16.9	−17.1	26.6	26.5
346	naphthalene	−10389	−10412	−14.3	−14.0	26.7	27.5
349	pentylcyclohexane[a]	−10820	−10815	−15.9	−16.1	25.8	25.2
350	*trans*-1-phenyl-1-butene	−10672	−10490	−15.8	−15.9	24.7	22.5
352	phenylcyclopentane	−10951	−10932	−15.3	−15.2	28.5	28.6
355	1,2,3,4-tetramethylbenzene[a,b]	−10695	−10365	−15.5	−15.8	25.8	22.2
356	1,3,5-triethylbenzene	−11462	−11468	−16.7	−16.4	28.7	29.5
357	*cis*-2-undecene	−11168	−11082	−17.1	−17.0	25.0	24.5
358	*trans*-2-undecene	−11101	−11117	−17.1	−17.0	24.6	24.9
361	tridecane[a]	−12344	−12481	−17.9	−17.5	32.1	34.3
362	cyclododecane[a]	−10843	−11228	−14.2	−14.9	31.0	31.8
363	dicyclohexane[a]	−11600	−11139	−16.0	−15.0	32.0	30.7
364	2,10-dimethylundecane	−11748	−11754	−17.2	−17.4	29.5	28.7
365	2,2-dimethylundecane[a]	−11824	−11728	−17.5	−17.2	29.2	29.4
366	2,3-dimethylundecane	−11900	−11790	−17.3	−17.3	30.4	29.5
367	2,4-dimethylundecane	−11638	−11649	−17.1	−17.3	28.8	28.4
368	2,5-dimethylundecane	−11644	−11645	−17.1	−17.3	28.8	28.2
369	2,6-dimethylundecane	−11644	−11683	−17.1	−17.3	28.8	28.7
370	2,7-dimethylundecane	−11681	−11693	−17.1	−17.3	29.1	28.7
371	2,8-dimethylundecane[a]	−11711	−11664	−17.2	−17.5	29.2	28.0
372	2,9-dimethylundecane	−12036	−11793	−17.9	−17.4	29.9	29.2
373	3,4-dimethylundecane	−11875	−11719	−17.3	−17.2	30.2	29.3
374	3,5-dimethylundecane	−11632	−11671	−17.1	−17.0	28.8	29.2
375	4,5-dimethylundecane	−11772	−11634	−17.2	−17.1	29.6	28.9
376	5,6-dimethylundecane	−11730	−11675	−17.2	−16.9	29.4	29.9
377	1-ethyl-2-butylbenzene	−11540	−11403	−16.7	−16.3	29.4	29.1
384	1-ethyltetralin[a]	−11429	−11414	−15.5	−15.4	32.1	32.1
387	hexylcyclohexane	−11472	−11452	−16.4	−16.4	29.7	29.5
388	methylcycloundecane	−10682	−10727	−14.2	−14.8	29.7	27.6
389	methyldodecane	−12106	−12087	−17.7	−17.7	30.8	30.5
390	3-methyldodecane	−12120	−12103	−17.7	−17.7	31.0	30.7
391	4-methyldodecane	−12085	−12039	−17.7	−17.5	30.7	30.9
392	5-methyldodecane	−12069	−11984	−17.7	−17.7	30.5	29.9
398	1-methyl-2-pentylbenzene[a]	−11695	−11420	−16.8	−16.3	30.3	29.0
399	pentamethyl-benzene[a,b]	−11290	−10974	−15.6	−15.9	30.6	27.2
405	phenylcyclohexane	−11572	−11412	−15.9	−15.4	32.0	32.0
406	5-propylindan	−11618	−11636	−16.0	−16.2	32.2	31.7
407	1-tridecene	−12264	−12373	−17.9	−17.7	31.7	32.8
408	1,2,3-triethylbenzene	−11833	−11494	−16.9	−16.0	31.0	30.8
409	1,2,4-triethylbenzene	−11642	−11472	−16.8	−16.3	29.8	29.7

[a] Test set compound. [b] Outlier.

with the highest q^2 (squared correlation coefficient in cross-validation) for the training set was selected as the best *n*-parameter MLR model. To test the significance of the descriptors selected to the prediction model a *t*-test and partial-*F* statistics were calculated for each parameter of the model.

Very low limits of elimination were chosen, that is, one parameter and pairwise correlation coefficient were <0.01

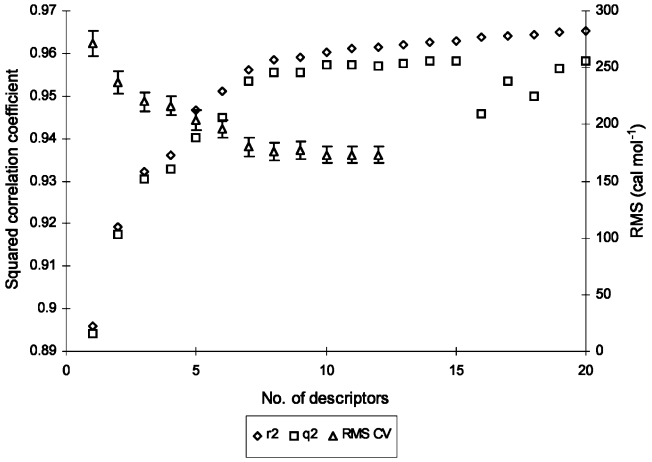

Figure 1. Prediction capabilities of *n*-dimensional MLR models for the prediction of $\Delta H°$ values.

and >0.99, respectively. Such low limits of elimination were selected because it was already reported in the literature that a linear combination of highly correlated descriptors, which individually do not correlate well with the studied property, can also when combined give a very good correlation.[34]

The leave-one-out cross-validation procedure performed on the training set was used for the estimation of cross-validation capabilities of the MLR models during the structural descriptor selection. The root-mean-squared (RMS) error was calculated as well as the squared correlation coefficient (q^2) for the straight line cross-validated vs Pro ezGC values. The model with the best cross-validation parameters was chosen for further studies. The final prediction results of the selected MLR model were estimated from the RMS error value obtained from the test set.

RESULTS AND DISCUSSION

Calculation of $\Delta H°$ Values. Separate QSPR models were developed for the calculation of $\Delta H°$ and $\Delta S°$ values. The influence of the number of selected parameters of the MLR model on its cross-validation capabilities was evaluated for the models with up to 20 descriptors. Straight line retrieved $\Delta H°$ vs Pro ezGC $\Delta H°$ and cross-validated $\Delta H°$ vs Pro ezGC $\Delta H°$ values for the training set were calculated using the least-squares method. The squared correlation coefficients r^2 and q^2 as well as the RMS_{CV} error value were calculated for each *n*-dimensional MLR model and are shown in Figure 1.

The r^2 for the retrieved values increases with the number of parameters of the MLR model. On the other hand, the same parameter for the cross-validated values (q^2) increases only to the point where 11 descriptors are included in the model. Afterward it becomes constant or in some cases even slightly decreases. This result is expected because with the addition of each new descriptor to the model we are improving the mathematical description of the chemical structure, that is, we are introducing structural features that contain some information about the modeled property. This goes to the point where new descriptors cannot improve the mathematical representation any more. The r^2 for the retrieved values still increases, but the cross-validation ability of the models decreases. The model is overfitted, and new structural indices are introducing noise.

It is difficult to decide if the small improvements in q^2 values in the constant part of the curve are due to the real improvements in the quality of the model or they are the result of the random error in calculations of the same values. To reduce the possibility to obtain an overfitted model an additional parameter for the estimation of cross-validation capabilities was introduced. The RMS_{CV} error together with the estimate of its standard deviation was calculated. Error bars in Figure 1 represent the calculated standard deviation of the RMS_{CV} error. We can see that RMS_{CV} first decreases and becomes approximately constant from the point where seven descriptors are included in the model, that is, fluctuations beyond this point are within one standard deviation limit. We can conclude that from this point on the improvement of the model is not significant any more. On the basis of these conclusions the seven-parameter MLR model was selected for the calculation of $\Delta H°$ values for nonpolar stationary phase. The selected structural descriptors together with coefficients of the MLR model are presented in Table 2. The regression parameters reported were obtained from the regression model constructed from the training set which contains 195 compounds. Because very liberal values were chosen for testing the pairwise correlation during descriptors the selection procedure extensive model validation was done in order to prevent insignificant variables to participate in the created prediction model. The highest correlation was seen between the Wiener index and the Randic index order 1 as well as between the Kier&Hall index order 0 and the Topographic electronic index. The corresponding values were 0.937 and 0.915, respectively. When high correlations are present in the model, there is a bigger chance that some of the variables are not significant. Therefore *t*-test and partial-*F* statistics[35] were done for every descriptor included in the model. The critical value for the Student's *t*-distribution for $\alpha = 0.01$ and 187 degrees of freedom is around 2.576. We can see from Table 2 that all descriptors are above this value and are therefore retained in the model. A similar conclusion can be made from partial-*F* statistics. The lowest partial-*F* values (Table 2) were calculated for both WNSA-1 descriptors. Both partial-*F* values were still above the critical $F_{0.01}(1, 187) = 6.63$, which points to the significance of all parameters included in the model.

A close inspection of selected descriptors gives us some insight into the structural features that determine the modeled property. In our QSPR study a chemical structure was represented by a seven-dimensional vector, the components of which were four topological, two electrostatic, and one hybrid type of descriptor that combines atomic contributions to the solvent-accessible surface area (SASA) with atomic partial charges.[36,37] Out of the four topological indices three were connectivity indices of different order based on the degree of graph vertices (Randic indices[38]) and the values of the atom valences δ_i (Kier and Hall indices[39]), and the last one was Wiener's topological distance index.[40] Every mentioned topological index encodes the molecular size as well as the branching pattern in its own way. Therefore we can deduct that in our case the molecular size and branching are among major factors that determine the $\Delta H°$ in separation processes. The main disadvantage of the already mentioned descriptors is the lack of ability to encode polar interactions between organic species and stationary phase. Although such interactions are weak in the case when the nonpolar stationary

Table 2. Selected Structural Descriptors and Coefficients of the Best MLR Model for the Prediction of ΔH^0 Values[a]

no.	coeff	SE	*t*-test	partial *F*	name of the descriptor
0	−5250	233	−22.5	504	intercept
1	−641	85.9	−7.45	55.5	Randic index (order 1)
2	13.99	3.46	4.05	16.4	WNSA-1 weighted PNSA (PNSA1*TMSA/1000) [Zefirov's PC]
3	−3.62	0.60	−6.05	36.6	Wiener index
4	2708	253	10.7	115	Topographic electronic index (all pairs) [Zefirov's PC]
5	−891	119	−7.48	56.0	Kier&Hall index (order 0)
6	257	34	7.53	56.8	Randic index (order 2)
7	4.95	1.27	3.88	15.1	WNSA-1 weighted PNSA (PNSA1*TMSA/1000) [Semi-MO PC]

[a] $r^2 = 0.966$, $F = 767$, $S = 161$.

phase is used for the separation of hydrocarbons, they cannot be neglected. One can eliminate this drawback by including two electrostatic[41] and a hybrid class of CPSA descriptor[36,37] to the theoretical model. These descriptors represent or depend on charge distribution in the molecule. In the case of electrostatic descriptors, the descriptors are calculated from the distribution of empirical partial charges in the molecule, which are obtained by an approach proposed by Zefirov.[42,43] On the other hand, MOPAC software was used for the calculation of partial charges and geometry optimization necessary for the creation of the hybrid type of charge partial surface area (CPSA) descriptors. We can see that two WNSA-1 descriptors were included in the model. Since the same conformation was used in both cases to compute the solvent accessible surface area, the indices differ only by the method used for computing the partial atomic charges. It seems that the distribution of charge plays an important role in the processes of the distribution of organic molecules between the mobile and the stationary phases. None of the computational methods alone has adequately presented this distribution. Only the linear combination of indices based on both calculated charge distributions shows a significant impact on distribution processes in gas chromatographic separation.

It must be mentioned that the structural interpretation of the model represents only qualitative information about the factors that are determining the distribution of $\Delta H°$. The main reason the influence of individual factors on modeled property cannot be quantitatively evaluated is mutual correlation of the descriptors involved. We noticed that there is a lot of overlapping in the description of certain structural features such as size, branching, etc., that is, the pairwise correlation coefficients between topological indices which encode these properties varies from 0.591 to 0.937, respectively. The lowest correlation with the remaining descriptors is observed for the electrostatic descriptor *WNSA-1*. Its pairwise correlation coefficient (c) with the other indices varies between −0.248 and 0.189 except with *Topographic electronic index (all pairs)*, where moderate correlation was observed, pairwise correlation coefficient being −0.590. To eliminate these ambiguities orthogonal descriptors should be selected. Because it is usually very difficult to obtain orthogonal descriptors that would at the same time create a very good prediction model an orthogonalization[44,45] of the already chosen structural indices can be performed.

At the end, the test set was used for the evaluation of the prediction capabilities of the MLR model. The graph of predicted versus Pro ezGC $\Delta H°$ is shown in Figure 2 and Table 1. No obvious outliers were detected. The prediction

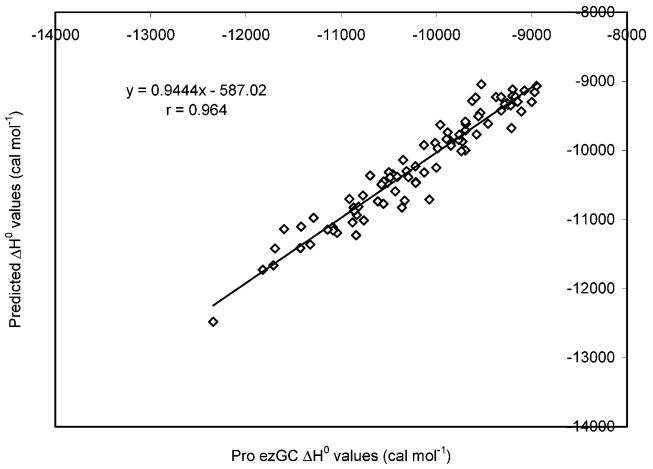

Figure 2. Predicted vs Pro ezGC $\Delta H°$ values.

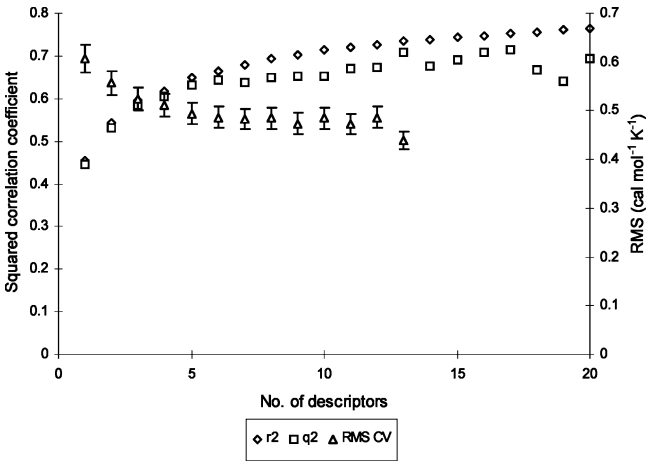

Figure 3. Prediction capabilities of *n*-dimensional MLR models for the prediction of $\Delta S°$ values.

error was estimated with the RMS error, which was 207 cal/mol.

Calculation of $\Delta S°$ Values. The same selection strategy for choosing structural descriptors as was used for the creation of the $\Delta H°$ calculation model was applied for the modeling of the $\Delta S°$ values. Again the cross-validation capabilities of the models with up to 20 parameters were estimated by the leave-one out cross-validation procedure and served as a criterion for the selection of the best MLR model. The influence of the number of selected parameters of the MLR model on r^2, q^2, and RMS_{CV} values for the training set is shown in Figure 3.

Again r^2 increases with the number of descriptors because we are adjusting the model to our data set. On the other hand, cross-validation parameters q^2 and RMS_{CV} are improving

Table 3. Selected Structural Descriptors and Coefficients of the Best MLR Model for the Prediction of ΔS^0 Values[a]

no.	coeff	SE	*t*-test	partial *F*	name of the descriptor
0	−22.4	1.74	−12.9	166.3	intercept
1	−0.0180	0.00257	−6.7	45.5	PPSA-1 partial positive surface area [Zefirov's PC]
2	36.1	3.09	11.7	135.9	FPSA-2 fractional PPSA (PPSA-2/TMSA) [Zefirov's PC]
3	15.5	2.51	6.2	38.1	relative number of C atoms
4	−0.0123	0.00251	−4.9	24.0	Wiener index
5	−17.48	6.28	−2.8	7.73	FNSA-2 fractional PNSA (PNSA-2/TMSA) [Zefirov's PC]

[a] $r^2 = 0.709$, $F = 92$, $S = 0.445$.

only for the first several descriptors. Afterward the model becomes overfitted, and q^2 and RMS_{CV} become more or less constant. The best cross-validation abilities were observed when 13 structural indices were included in the MLR model. The q^2 and RMS_{CV} values were 0.708 and 0.44, respectively. We have already mentioned that the q^2 and RMS_{CV} values did not change significantly after the first several descriptors were included in the model. To evaluate which changes in the q^2 and RMS_{CV} values were due to an improvement of the model and which were due to the chance correlation or due to the uncertainty in the calculation of a mentioned parameter a precision of the RMS_{CV} value was calculated. The calculated standard deviation for each RMS_{CV} estimate is shown in Figure 3 as an error bar. After five structural descriptors were included in the MLR model the fluctuation of RMS_{CV} values were within one standard deviation. Therefore a five-parameter MLR model was selected for the calculation of $\Delta S°$ values. The list of selected indices and parameters of the MLR model is shown in Table 3.

Again the model validation was done in order to prevent insignificant variables to enter a prediction model. The highest correlation was seen between "FPSA-2 fractional PPSA (PPSA-2/TMSA) [Zefirov's PC]" and "PPSA-1 partial positive surface area [Zefirov's PC]" and was 0.97. All other pairwise correlation coefficients were below 0.9. The partial-*F* statistics and the *t*-test were calculated. For all descriptors test values were above critical ($t_{0.01}(189) = 2.576$, $F_{0.01}(1, 189) = 6.63$), and therefore all parameters were retained in the model.

Out of the five descriptors included in the model one was constitutional, one was topological, and three were electrostatic indices. Selected descriptors encode information about the size and the branching of the molecule as well as the distribution of the charge inside it. All three electrostatic descriptors were calculated from the distribution of the empirical partial charges in the molecule. Therefore we can conclude that besides molecular size and branching the distribution of charge in the molecule plays an important role in determination of the $\Delta S°$ values. The difference between the calculated and the Pro ezGC $\Delta S°$ values for the test set was used for the evaluation of the prediction capabilities of the described MLR model. The predicted $\Delta S°$ vs Pro ezGC $\Delta S°$ data are shown in Figure 4 and Table 1. The quality of the prediction model was worse then in the case of the prediction of the $\Delta H°$ data. The q^2 and RMS values obtained from the test set were 0.491 and 0.58 cal mol^{-1} K^{-1}, respectively.

Accuracy of Calculation of Retention Times in Temperature Programmed Gas Chromatography. Our initial task was to develop a procedure, which will enable chromatographers to calculate retention times of different organic compounds in temperature programmed gas chromatographic separation from their chemical structure. Two individual models were developed for the calculation of the $\Delta H°$ and $\Delta S°$ values, and their prediction errors were estimated using the training/test set procedure. Since these numbers are of little interest to the final users of these models who are rather familiar with chromatography and not QSPR studies, an effort has been made to evaluate an error of calculation of retention times in TPGC.

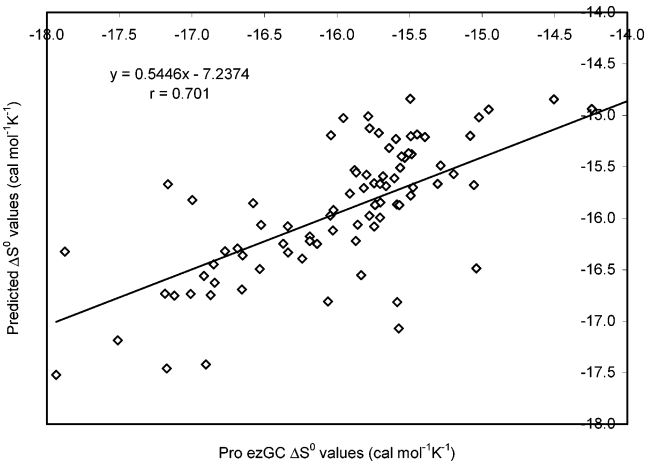

Figure 4. Predicted vs Pro ezGC $\Delta S°$ values.

Thermodynamic data obtained from the Pro ezGC software and those calculated by the developed MLR models were used for the calculation of retention times for the organic compounds from the training and the test set using an already described procedure.[46] Only the simplest case was studied, in which the chromatograms were obtained at a constant heating rate without initial or final isothermal hold times. Temperature programmable gas chromatograms were simulated at various chromatographic conditions where the phase ratio was varied from 0.002 to 0.012, the initial temperature was changed from −20 up to 80 °C, and the heating rate was varied from 2 up to 12 °C/min. An example of the calculated retention times for the phase ratio equals 0.004, the initial temperature is 0 °C, and the heating rate 4 °C/min is shown in Table 1. The difference between the retention times obtained from Pro ezGC and those obtained from the modeled thermodynamic values was calculated for the test set compounds at each chromatographic condition. The RMS error of these retention time residuals served as an estimate of the calculation accuracy of the retention times obtained by the described procedure.

The RMS error of the retention time residuals varied from one set of chromatographic conditions to another; therefore, any kind of normalization procedure would be beneficial. It was found out that the ratio between the RMS error and the maximum separation time remained constant and was 5.4

± 0.1%, where the maximum separation time represents the time difference between the first and the last eluted organic compound. According to this finding it was concluded that errors in the prediction of the thermodynamic data ($\Delta H°$, $\Delta S°$) are responsible for the inaccuracy in the calculation of retention times for around up to 5% of the maximum separation space available at selected chromatographic conditions.

Our last task was to evaluate if our developed prediction models will enable us to accurately distinguish between the components of the homological series where identification based on the mass spectrum library search is disabled due to similar mass spectra. An individual compound was positively identified if its retention time calculated from the predicted thermodynamic values fell in a 95% confidence interval around the true values, which was calculated from the Pro ezGC data. Such comparisons were performed for all simulated chromatograms. At each chromatographic condition the same 15 components were detected as outliers and were marked in Table 1. All except one were cyclic compounds. Although we do not know the accuracy of the Pro ezGC data, which can be the source of error during creation of the model, we have concluded that our model is not the best suited for the prediction of the retention behavior of cyclic compounds. On the other hand, our prediction model successfully distinguished between the homological series of *n*-alkanes with 8 to 12 C atoms, 1-alkenes with 12 and 13 C atoms, and 2-alkenes with 8 to 10 C atoms. In all these cases the calculation error of retention times was substantially better than the time difference between consecutive compounds (see Table 1).

CONCLUSIONS

Two separate quantitative structure−property relationship models were derived for the calculation of the $\Delta H°$ and $\Delta S°$ values in TPGC, which are necessary for the subsequent calculation of the retention times. Seven-descriptor and five-descriptor MLR models were selected for the calculation of the $\Delta H°$ and $\Delta S°$ values, respectively, based on the best cross-validation abilities. The final prediction capabilities of the models were evaluated by the test set procedure. RMS errors calculated from the test set were 207 cal mol^{-1} and 0.58 cal mol^{-1} K^{-1} for the $\Delta H°$ and $\Delta S°$ MLR models, respectively.

An effort has been made to evaluate an error in the calculation of the retention times in TPGC, which arises due to errors in the theoretical prediction of the $\Delta H°$ and $\Delta S°$ values. Several chromatograms were simulated using Pro ezGC and the theoretically calculated thermodynamic data, and the RMS error of retention time residuals was calculated. It was found out that, although the RMS error varies from one chromatographic condition to another, the ratio between the RMS error and the maximum available separation space for the particular set of organic compounds remains constant. Therefore it was concluded that errors in the theoretical calculation of the thermodynamic data ($\Delta H°$, $\Delta S°$) are responsible for the inaccuracy in the calculation of the retention times for around up to 5% of the maximum separation space available at selected chromatographic conditions. Our developed model was able to accurately differentiate between retention times of consecutive compounds in *n*-alkanes, 1-alkenes, and 2-alkenes homological series but failed to give good calculation results for the cyclic components.

ACKNOWLEDGMENT

This work was supported by the Ministry of Education, Science and Sport of the Republic of Slovenia (Grant P1-0153) and the National Science Foundation (Grant CHE-9714328).

REFERENCES AND NOTES

(1) 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.

(2) Sekušak, S.; Sabljić, A. Calculation of retention indexes by molecular topology. 3. Chlorinated dibenzodioxines. *J. Chromatogr.* **1993**, *628*, 69−79.

(3) Heinzen, V. E. F.; Yunes, R. A. Using topological indices in the prediction of gas chromatographic retention indices of linear alkylbenzene isomers. *J. Chromatogr.* **1996**, *719*, 462−467.

(4) Sutter, J. M.; Peterson, T. A.; Jurs, P. C. Prediction of gas chromatographic retention indices of alkylbenzenes. *Anal. Chim. Acta* **1997**, *342*, 113−122.

(5) Jalali-Heravi, M.; Garkani-Nejad, Z. Prediction of gas-chromatographic retention indexes of some benzene-derivatives. *J. Chromatogr.* **1993**, *648*, 389−393.

(6) Woloszyn, T. F.; Jurs, P. C. Prediction of gas-chromatographic retention data for hydrocarbons from naphthas. *Anal. Chem.* **1993**, *65*, 582−587.

(7) Du, Y. P.; Liang, Y. Z.; Yun, D. Data mining for seeking an accurate quantitative relationship between molecular structure and GC retention indices of alkenes by projection pursuit. *J. Chem. Inf. Comput. Sci*. **2002**, *42*, 1283−1292.

(8) Peterson, K. L. Counter-propagation neural networks in the modelling and prediction of Kovats indexes for substituted phenols. *Anal. Chem*. **1992**, *64*, 379−386.

(9) Gramatica, P.; Navas, N.; Todeschini, R. 3D-modelling and prediction by WHIM descriptors. Part 9. Chromatographic relative retention time and physico-chemical properties of polychlorinated biphenyls (PCBs). *Chemometr. Intell. Lab.* **1998**, *40*, 53−63.

(10) Ivanciuc, O.; Ivanciuc, T.; Klein, D. J.; Seitz, W. A.; Balaban, A. T. Quantitative structure-retention relationships for gas chromatographic retention indices of alkylbenzenes with molecular graph descriptors. *SAR QSAR Environ. Res*. 2001, *11*, 419−452.

(11) Gao, Y. H.; Wang, Y. W.; Yao, X. J.; Zhang, X. Y.; Liu, M. C.; Hu, Z. D.; Fan, B. T. The prediction for gas chromatographic retention index of disulfides on stationary phases of different polarity. *Talanta* **2003**, *59*, 229−237.

(12) Zefirov, N. S.; Palyulin, V. A. Fragmental approach in QSPR. *J. Chem. Inf. Comput. Sci*. **2002**, *42*, 1112−1122.

(13) Basak, S. C.; Nikolic, S.; Trinajstic, N.; Amic, D.; Beslo, D. Fragmental QSPR modeling: Graph connectivity indices versus line graph connectivity indices. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 927−933.

(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.; Novic, M.; Veber, M. Modelling of gas chromatographic retention indices using counterpropagation neural networks. *Anal. Chim. Acta* **1997**, *348*, 215−221.

(16) Pompe, M.; Novic, M. Prediction of gas-chromatographic retention indices using topological descriptors. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 59−67.

(17) Randić, M.; Basak, S. C.; Pompe, M.; Novic, M.; Veber, M. Prediction of gas chromatographic retention indices using variable connectivity index. *Acta Chim. Slov.* **2001**, *48*, 169−180.

(18) Katritzky, A. R.; Ignatchenko, E. S.; Barcock, R. A.; Lobanov, V. S.; Karelson, M. Prediction of gas-chromatographic retention times and response factors using a general quantitative structure−property relationship treatment. *Anal. Chem*. **1994**, *66*, 1799−1807.

(19) 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.

(20) Payares, P.; Diaz, D.; Olivero, J.; Vivas, R.; Gomez, I. Prediction of the gas chromatographic relative retention times of flavonoids from molecular structure. *J. Chromatogr. A* **1997**, *771*, 213−219.

(21) Rayne, S.; Ikonomou, M. G. Predicting gas chromatographic retention times for the 209 polybrominated diphenyl ether congeners. *J. Chromatogr. A* **2003**, *1016*, 213−219.
(22) Mecozzi, M.; Acquistucci, R.; Amici, M. Prediction of gas chromatographic retention times of polychlorinated biphenyls by monodimensional molecular descriptors and multivariate techniques. *Chromatographia* **2003**, *57*, S213−S217.
(23) Gonzalez, F. R.; Nardillo, A. M. Revision of a theoretical expression for gas−liquid chromatographic retention. *J. Chromatogr. A* **1999**, *842*, 29−49.
(24) Hawkes, S. J. Extrapolating gas-chromatographic retention indexes at 2 temperatures to a 3rd temperature. *Anal. Chem.* **1989**, *61*, 88−90.
(25) Snijders, H.; Janssen, H. G.; Cramers, C. Optimization of temperature-programmed gas chromatographic separations 1. Prediction of retention times and peak widths from retention indices. *J. Chromatogr. A* **1995**, *718*, 339−355.
(26) Vezzani, S.; Moretti, P.; Castello, G. Fast and accurate method for the automatic prediction of programmed-temperature retention times. *J. Chromatogr. A* **1994**, *677*, 339−355.
(27) Shrotri, P. Y.; Mokashi, A.; Mukesh, D. Prediction of retention times in temperature-programmed gas solid and gas−liquid-chromatography. *J. Chromatogr. A* **1987**, *387*, 399−403.
(28) Vezzani, S.; Moretti, P.; Castello, G. Automatic prediction of retention times in multi-linear programmed temperature analyses. *J. Chromatogr. A* **1997**, *767*, 115−125.
(29) Al-Bajjari, T. I.; Le Vent, S.; Taylor, D. R. Calculation of programmed-temperature gas-chromatography characteristics from isothermal data. 5. Prediction of peak asymmetries and resolution characteristics. *J. Chromatogr. A* **1994**, *683*, 367−376.
(30) Stewart, J. J. P. Special issue − MOPAC − A semiempirical molecular-orbital program. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 1−105.
(31) Karelson, M.; Lobanov, V. S.; Katritzky, A. R. Quantum-chemical descriptors in QSAR/QSPR studies. *Chem. Rev.* **1996**, *96*, 1027−1043.
(32) Katritzky, A. R.; Lobadov, V. S.; Karelson, M. CODESSA Training Manual, Gainesville, 1995.
(33) Katritzky, A. R.; Mu, L.; Lobanov, V. S.; Karelson, M. Correlation of boiling points with molecular structure 1. A training set of 298 diverse organics and a test set of 9 simple inorganics. *J. Phys. Chem.* **1996**, *100*, 10400−10407.
(34) Randić, M. On characterization of chemical structure. *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 672−687.
(35) Neter, J.; Kutner, M. H.; Nachtshein, C. J.; Wasserman, W. *Applied Linear Statistical Models*, 4th ed.; McGraw-Hill: Boston, MA, 1996; pp 268−268.
(36) Stanton, D. T.; Jurs, P. C. Development and use of charged partial surface-area structural descriptors in computer-assisted quantitative structure property relationship studies. *Anal. Chem.* **1990**, *62*, 2323−2329.
(37) Stanton, D. T.; Egolf, L. M.; Jurs, P. C.; Hicks, M. G. Computer-assisted prediction of normal boiling points of pyrans and pyrroles. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 306−316.
(38) Randić, M. Characterization of molecular branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609−6615.
(39) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure − Activity Analysis*; Wiley: New York, 1986.
(40) Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69*, 17−20.
(41) Osmialowski, K.; Halkiewisz, J.; Kaliszan, R. J. Quantum chemical-parameters in correlation-analysis of gas−liquid-chromatographic retention indexes of amines. 2. Topological electronic index. *J. Chromatogr.* **1986**, *361*, 63−69.
(42) Zefirov, N. S.; Kirpichenok, M. A.; Izmailov, F. F.; Trofimov, M. I. Calculation schemes for atomic electronegativities in molecular graphs within the framework of Sanderson principle. *Dokl. Akad. Nauk. SSSR* **1987**, *296*, 883−887.
(43) Kirpichenok, M. A.; Zefirov, N. S. Electronegativity and geometry of molecules. 1. Principles of developed approach and analysis of the effect of nearest electrostatic interactions on the bond length in organic-molecules. *Zh. Org. Chim.* **1987**, *23*, 673−691.
(44) Randić, M. Orthogonal molecular descriptors. *New J. Chem.* **1991**, *15*, 517−525.
(45) Randić, M. Resolution of ambiguities in structure − property studies by use of orthogonal descriptors. *J. Chem. Inf. Comput. Sci*. **1991**, *31*, 311−320.
(46) Davis, J.; Pompe, M.; Samuel, C. Justification of statistical overlap theory in programmed temperature gas chromatography: Thermodynamic origin of random distribution of retention times. *Anal. Chem.* **2000**, *72*, 5700−5713.

CI0304268