Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Chromatography A, 1028 (2004) 287–295

JOURNAL OF CHROMATOGRAPHY A

www.elsevier.com/locate/chroma

Prediction of gas chromatographic retention indices of a diverse set of toxicologically relevant compounds

Z. Garkani-Nejad [a], M. Karlovits [b], W. Demuth [b], T. Stimpfl [c], W. Vycudilik [c], M. Jalali-Heravi [d], K. Varmuza [b,*]

[a] *Faculty of Science, Vali-e Asr University of Rafsanjan, Rafsanjan, Iran*
[b] *Laboratory for Chemometrics, Institute of Chemical Engineering, Vienna University of Technology, Getreidemarkt 9/166, A-1060 Vienna, Austria*
[c] *Institute of Forensic Medicine, University of Vienna, Sensengasse 2, A-1090 Vienna, Austria*
[d] *Department of Chemistry, Sharif University of Technology, P.O. Box 11365-9516 Tehran, Iran*

Received 8 July 2003; received in revised form 1 December 2003; accepted 3 December 2003

Abstract

For a set of 846 organic compounds, relevant in forensic analytical chemistry, with highly diverse chemical structures, the gas chromatographic Kovats retention indices have been quantitatively modeled by using a large set of molecular descriptors generated by software *Dragon*. Best and very similar performances for prediction have been obtained by a partial least squares regression (PLS) model using all considered 529 descriptors, and a multiple linear regression (MLR) model using only 15 descriptors obtained by a stepwise feature selection. The standard deviations of the prediction errors (SEP), were estimated in four experiments with differently distributed training and prediction sets. For the best models SEP is about 80 retention index units, corresponding to 2.1–7.2% of the covered retention index interval of 1110–3870. The molecular properties known to be relevant for GC retention data, such as molecular size, branching and polar functional groups are well covered by the selected 15 descriptors. The developed models support the identification of substances in forensic analytical work by GC–MS in cases the retention data for candidate structures are not available.

Keywords: Retention indices; Retention prediction; Molecular descriptors; Regression models; Feature selection; Mathematical modelling; Forensic analysis

1. Introduction

Investigations and developments of mathematical models that are able to predict chromatographic retention data from chemical structures data have found wide interest in studies on quantitative structure–property relationships (QSPRs) [1]. Typical works in this field deal with 50–200 organic compounds, often belonging to a strictly defined class of substances. Aim is usually to create a model by using a small number of well interpretable molecular descriptors, although a great variety of much more than 1000 descriptors has been described and suggested for QSPRs [2,3].

Recently published papers on relationships between molecular descriptors and gas chromatographic retention data, for instance, deal with sets of 149 alkanes [4], 130 methylalkanes [5], 400 alkenes [6], 150 alkylbenzenes [7–10], 200 polycyclic aromatic hydrocarbons [11], 60 polychlorinated naphthalenes [12], up to 100 esters, alcohols, aldehydes and ketones [13–16], 50 terpenes [17], and up to 400 diverse organic compounds [18–20]. Typically, 20–300 molecular descriptors are tested, and the final models contain less than 10 selected ones. Most used multivariate methods are multiple linear regression (MLR), partial least squares regression (PLS), principal component regression (PCR), and artificial neural networks (ANNs). Previously investigated compound classes that are related to this study are a set of about 60 stimulants and narcotics [21], and a set of about 30 chemical warfare agents [22]. A good correlation has been found between boiling points and retention indices based on polycyclic hydrocarbons used as standard compounds [23].

In this study, a set of 846 compounds from a database used in toxicology and forensic GC–MS analyses has been used to investigate the applicability of QSPR approaches and standard methods from chemometrics for the prediction of Kovats retention indices from chemical structure data. The chemical structures of the considered compounds are

[*] Corresponding author. Tel.: +43-1-58801-16060; fax: +43-1-58801-16091.
E-mail address: kvarmuza@email.tuwien.ac.at (K. Varmuza).

0021-9673/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.chroma.2003.12.003

highly diverse, and most of them contain several functional groups.

Forensic toxicological analysis is often confronted with the identification of initially undetermined substances in biological material. Finding a poison that is a priori unknown in the case of a suspected intoxication is a difficult task, because many compounds have to be considered (toxicity depends on dose) but identification criteria of only about 7000 drugs or pesticides are presently available. Furthermore, new compounds—such as designer drugs—may be present. The most used analytical technique in this field is GC–MS. It has been shown that a combination of sophisticated sample preparation [24], strictly defined experimental conditions for GC–MS, an appropriate database [25] and new software concepts [26] for selection and comparison of mass spectra greatly facilitates the identification of poisons in biological materials. Although mass spectra are indispensable data for substance identification and structure elucidation, chromatographic retention data are very useful too [27]; for instance isomeric compounds often have very similar mass spectra, and some classes of compounds show non-specific mass spectra. Because many spectroscopic databases do not contain chromatographic retention data, an automatic prediction of retention indices from chemical structures would be helpful for the identification of a priori unknown compounds [28].

The strategy applied in this study is in some aspects different from previous works on retention index modeling. A relative large set of 846 organic compounds with very diverse chemical structures was used, and the initial number of molecular descriptors was 1497. Selection of subsets of descriptors was guided by mathematical principles but not by chromatographic experiences. Easily available software has been applied for the generation of molecular descriptors (*Dragon*) and for a conversion of two-dimensional (2D) structures into three-dimensional (3D) structures (*WebLab Viewer*); both were offered for free download at the time of this work. For multivariate regression a widely used chemometric software (*Unscrambler*) was applied for a comparison of the routinely used linear calibration methods PLS, PCR, and MLR.

2. Data and software

2.1. Database

The database used is a subset of the database “Mass spectral and GC data of drugs, poisons, pesticides, pollutants and their metabolites” [25] containing 4367 entries. A set of 846 compounds has been selected for this work, with molecular masses between 109 and 491 (median 260), and non-hydrogen elements in the ranges $C_{4–32}$ $N_{0–7}$ $O_{0–11}$ $S_{0–4}$ $P_{0–2}$ $F_{0–6}$ $Cl_{0–10}$ $Br_{0–2}$ $I_{0–1}$. Table 1 lists prominent categories of forensic relevant substances, demonstrating the high structural diversity of the used data set. For reference, *n*-alkanes with 14 to 30 C-atoms have been included. The Kovats retention indices (*I*) [29,30] of the compounds are between 1110 and 3870 with a median of 2000. Packed and capillary columns with apolar stationary phases, such as methylpolysiloxanes (e.g. HP1, OV-101) were used in building the retention index database [25]. Therefore, an inherent variability of the listed retention indices is present. Based on laboratory experiences the differences between experimental and listed values have to be expected in a range of about plus/minus 50 retention index units. That is higher than reported for homogeneous groups of compounds, as for instance plus/minus 3 units for alkylbenzenes [7] or alkanes [31].

Table 1
Prominent substance categories in the used data set of toxicologically relevant compounds

No.	Category	No. of compounds
1	Insecticides	88
2	Herbicides	74
3	Hypnotics	57
4	Antihistamines	40
5	Neuroleptics	35
6	Fungicides	26
7	Tranquilizers	25
8	Antidepressants	22
9	Chemicals	21
10	Fatty acids	19
11	Potent analgesics	19
12	Hydrocarbons	17
13	Stimulants	15
14	Analgesics	14
15	Biomolecules	14
16	Others	360
	Sum	846

2.2. Software

Molecular descriptors have been generated by software *Dragon*, version Web 3.0 [32]. Conversion of 2D structures into 3D structures was performed with software *WebLab Viewer*, version 3.50 [33]. Statistical evaluation of data and multivariate data analysis have been performed mainly by the software products *Unscrambler*, version 7.8 [34], and *Systat*, version 10.0 [35]. Additional programs have been developed in Matlab 6.0 [36] for handling of chemical structures and for data analysis. All work has been performed on personal computers running under operating system Microsoft Windows 2000; the reported computing times refer to a machine with 2 GHz.

3. Methods

3.1. Conversion of 2D structures into 3D structures

Software *Dragon* requires chemical structures with all H-atoms given explicitly, and a part of the molecular descriptors generated are 3D descriptors requiring 3D coded structures. The chemical structures available were 2D structures encoded in Molfile/SDF format [37]. Software *WebLab Viewer* was applied as an easy to use and convenient tool for 3D conversion and for adding missing H-atoms. The resulting 3D structures, however, are not energy optimized and may be in some cases only crude approximations. A Matlab program, that calls the *WebLab Viewer*, has been developed for an automatic conversion of all structures in an SDF-file; average computing time was 0.2 s per converted structure. The output file is again in Molfile/SDF format, appropriate for input into software *Dragon*.

The practicability of 3D descriptors is doubtful, especially for a set with very diverse chemical structures as used in this study [38]. Energy optimization methods are very time-consuming and may introduce artifacts depending on the substance class. Furthermore, flexibility of the molecules is not considered by the used descriptors. In a recent quantitative structure–activity relationship (QSAR) toxicity study was found that molecular descriptors, computed from 2D and 3D structures, gave models with very similar predictive power [39]. Within the data-driven strategy applied in this study, the 3D descriptors have not been eliminated a priori.

3.2. Generation of molecular descriptors

The used software *Dragon* is capable of generating 1497 molecular descriptors; the descriptors are divided into 18 groups as shown in Table 2. Average computing time for all descriptors was 2.5 s per structure; input file was in SDF-format (see Section 3.1); output file was in text format and was imported into software *Unscrambler, Systat*, and Matlab programs.

Table 2
Groups of the 1497 molecular descriptors generated by software *Dragon* [3,32]

Group no.	Group name	Dimensionality	No. of descriptors
1	Constitutional descriptors	0	47
2	Topological descriptors	2	266
3	Molecular walk counts	2	21
4	BCUT descriptors	2	64
5	Galvez topological charge indices	2	21
6	2D autocorrelation descriptors	2	96
7	Charge descriptors	3	14
8	Aromaticity indices	3	4
9	Randic molecular profiles	3	41
10	Geometrical descriptors	3	70
11	RDF descriptors	3	150
12	3D-MoRSE descriptors	3	160
13	WHIM descriptors	3	99
14	GETAWAY descriptors	3	197
15	Functional groups	1	121
16	Atom-centered fragments	1	120
17	Empirical descriptors	1	3
18	Properties	1	3
	Sum		1497

3.3. Regression models

Aim of the work was the development of a mathematical model that uses molecular descriptors, x_j with $j = 1 \dots p$, as input variables (features) and is capable of producing an output, I^*, that is a good estimation of the corresponding experimental retention index, I. A linear model is given by:

$$I^* = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p \quad (1)$$

with b_j being the regression coefficient for descriptor j, and b_0 the intercept. The regression methods compared are multiple linear regression, (MLR: ordinary least squares regression), principal component regression (PCR), and partial least squares regression (PLS) [40–42]. The last two methods are capable to handle large sets of variables, are tolerant to co-linearities, and can be optimized for maximum prediction. The 846 compounds have been divided into two random samples, a training set including 700 compounds, and a prediction set with 146 compounds. The number of descriptors used was between 15 and 529 selected by the methods described in Section 3.5. For PCR and PLS the descriptor values have been autoscaled.

The number of components used in PCR and PLS models has been determined by a cross-validation within the training set (see Section 3.4). After cross-validation, a new model has been built from the whole training set with the determined number of components. Finally, this model was applied to the prediction set. Computing time for linear models was between 5 and 50 s, depending on the number of used descriptors. All methods have been applied four-fold with different randomly selected compositions of training and prediction set.

3.4. Evaluation of regression models

The prominent goal of the regression models is a good performance for the prediction of retention indices for compounds not used in the training. The standard error of prediction (SEP) is a measure for this quality, defined as [41]:

$$\mathrm{SEP} = \left[\frac{1}{n-1}\sum(I_i^* - I_i - \mathrm{bias})^2\right]^{0.5} \qquad i = 1 \cdots n \quad (2)$$

$$\mathrm{bias} = \left(\frac{1}{n}\right)\sum(I_i^* - I_i) \qquad i = 1 \cdots n \quad (3)$$

with I_i^* being the estimated retention index of compound i in the prediction set, I_i the corresponding database value, and n the number of compounds. The bias corresponds to the mean of the prediction error, and SEP is equivalent to the standard deviation of the bias-corrected prediction error.

The standard error of prediction for cross-validation ($\mathrm{SEP_{CV}}$) is calculated in the same way but the estimated retention indices are those predicted during cross-validation within the training set. The standard error of calibration (SEC) refers to the estimated retention indices in the training set using a model that was calculated from the whole training set after cross-validation.

Correlation coefficients r_p and r_c characterize the linear relationship between experimental and calculated retention indices for the prediction set and the training set, respectively.

The importance, IMP_j, of a descriptor j in a linear model has been measured by the standardized regression coefficient [43]:

$$\mathrm{IMP}_j = \frac{(b_j s_j)}{s_y} \tag{4}$$

with b_j being the regression coefficient for descriptor j in the model (Eq. (1)), s_j the standard deviation of descriptor j, and s_y the standard deviation of the response (I); all calculated from the training set. IMP is identical to the regression coefficient obtained from autoscaled data.

For PCR and PLS an appropriate number of components has been determined from a plot showing $\mathrm{SEP}_{\mathrm{CV}}$ versus the number of components used, obtained by a cross-validation with 10 segments. Because this curves not always have a clear minimum, a heuristic strategy has been applied as follows. The maximum number of components considered was 40, or the number of descriptors if less than 40. If $\mathrm{SEP}_{\mathrm{CV}}$ shows a clear minimum then the minimum determines the number of components used for the model (even if it is the maximum number of components tested). If no clear minimum appears, the lowest number of components has been chosen that gives a $\mathrm{SEP}_{\mathrm{CV}}$ approximately 1% above the found minimum.

3.5. Feature selection

A working subset of 529 descriptors was selected from the 1497 generated descriptors by routine methods as follows: (1) descriptors which are constant have been eliminated (107 descriptors eliminated); (2) descriptors which are almost constant—that means all but a maximum of five values are constant—have been eliminated (52 descriptors eliminated); (3) descriptors containing very low or very high values have been eliminated. The thresholds used were defined as the 5 and 95% quantiles of the distribution given by the values of all remaining descriptors (472 descriptors eliminated); (4) for all pairs of remaining descriptors the correlation coefficient was determined. If a correlation coefficient was higher than 0.95 then the descriptor with the larger sum of correlation coefficients with the other descriptors has been eliminated. In this step, 337 descriptors have been eliminated resulting in the working set with $p = 529$ descriptors.

Three further methods for feature selection have been tested for the working set data.

(a) Selection of descriptors possessing highest absolute correlation coefficients with the retention index calculated from the training set. Resulting subsets with 200, 100, and 15 descriptors have been tested.
(b) Elimination of descriptors with small absolute regression coefficients b_j (Eq. (1)) obtained for a PCR or a PLS model from a training using all 529 descriptors. With the remaining descriptors a new model has been computed.
(c) A forward stepwise feature selection based on F-statistics together with MLR, as implemented in software *Systat*, has been used. For a significance level of 1% a subset with 15 descriptors has been obtained.

4. Results

4.1. Introduction

The results obtained with different regression methods and different sets of descriptors are summarized in Table 3, and are discussed in the following subsections. Note that SEP and SEC are means of four experiments with different random compositions of training and prediction set. A value of 82 retention index units for SEP is considered as a reference for comparisons of methods and datasets, since this value has been obtained for the chemometric standard method PLS directly applied to the working data set with all 529 descriptors. The prediction errors, SEP, of the four experiments were 75, 88, 85, and 79, demonstrating that a SEP obtained from a single experiment may be misleading

Table 3
Results obtained for different regression methods and different subsets of descriptors

Method	p	Feature selection	SEP	SEC
PLS	529	Reference	82	54
	200	Correlation coefficient	122	92
	100	Correlation coefficient	119	109
	15	Correlation coefficient	154	157
	15	Stepwise	79	81
	301	Excluding 3D descriptors	79	58
	202	Regression coefficient	79	58
PCR	529	–	114	111
	200	Correlation coefficient	145	140
	100	Correlation coefficient	150	141
	15	Correlation coefficient	158	164
	15	Stepwise	79	81
MLR	200	Correlation coefficient	115	74
	100	Correlation coefficient	125	108
	15	Correlation coefficient	156	157
	15	Stepwise	79	81
	13	Stepwise, then excluding 3D descriptors	92	95

p: number of descriptors; SEP: standard error of prediction; SEC: standard error of calibration. SEP and SEC are means from four experiments with different random samples for training (700 objects) and prediction (146 objects).

because of an accidental composition of training and prediction set. The corresponding calibration errors, SEC, were 55, 53, 53, and 55; as expected SEC is lower then SEP and less fluctuating. The same trend was found for other data sets and methods; in a few cases, SEP and SEC were similar. It is instructive to compare the results with those obtained after a random assignment of the retention indices to the 846 compounds; for instance PLS with 529 descriptors yielded a SEC of 322 and a SEP of 875.

4.2. Comparison of methods

Using all 529 descriptors, PLS regression resulted in much better models (averaged SEP = 82) than PCR (averaged SEP = 114); MLR was not applicable because of high co-linearities of the descriptors. Feature selection by using a set of descriptors possessing maximum correlation coefficient with the retention index was not successful; subsets with 200, 100 or 15 descriptors selected by this method gave models with a much higher SEP than models calculated from all 529 descriptors. Again PLS was in general better than PCR but similar to MLR; for instance for a subset with 100 descriptors, SEP was 119, 150, and 125 for PLS, PCR, and MLR, respectively.

The elimination of descriptors possessing small absolute regression coefficients is demonstrated by an example. Application of PLS with all 529 descriptors and using 15 components resulted in a model with regression coefficients b_j (j = 1–529) between −0.039 and 0.083. For a selection of descriptors with close to zero regression coefficients five threshold between 0.005 and 0.03 have been applied, and descriptors with a regression coefficient higher than the threshold were used for a new PLS model. Fig. 1 shows that $\mathrm{SEP_{CV}}$ is lowest for absolute thresholds between 0.01 and 0.02, corresponding to 63 to 197 selected descriptors. The corresponding values for SEP are between 77 and 81 which is better than 88 obtained with all descriptors. Feature selection by deleting descriptors with close to zero regression coefficients was capable of improving the models, however, an optimization of the applied threshold is required. Table 3 contains the average of four experiments with a mean of 202 selected descriptors and a mean of 79 for SEP which is much better than the SEP of 122 for 200 descriptors selected by maximum correlation coefficient with the retention index. However, the improvement is small in comparison with models calculated from all 529 descriptors (SEP = 82). Also PCR showed an increase in the prediction performance when applying this feature selection method but could not reach the performance of PLS models. For MLR, this feature selection method has been applied to the previously selected 200 descriptors (maximum absolute correlation coefficient with the retention index); the results showed a decrease in the prediction performance.

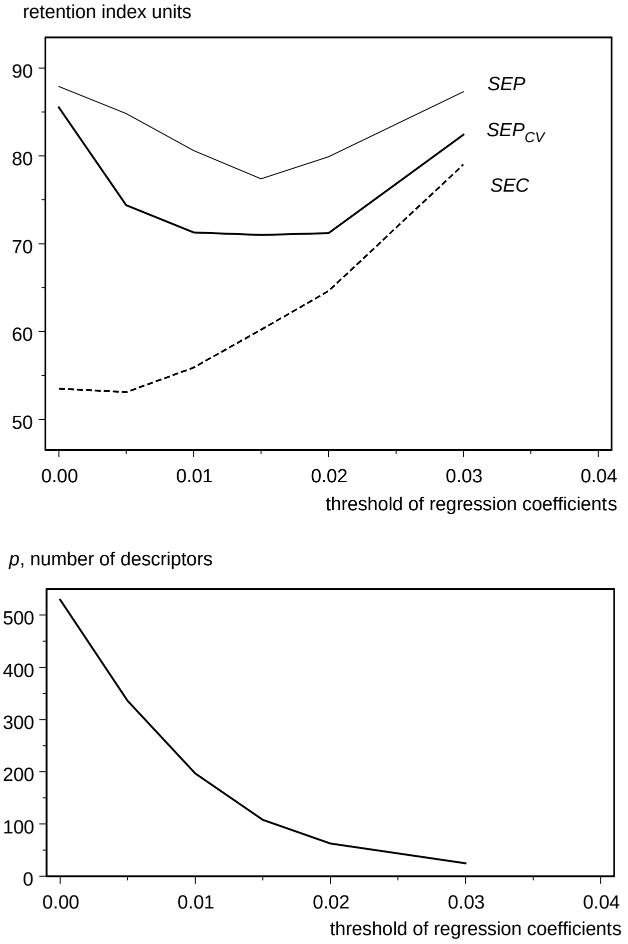


Fig. 1. Performance of PLS models with descriptors selected by a threshold of the regression coefficients as obtained in a PLS model with all 529 descriptors.

A forward stepwise feature selection—as implemented in software *Systat* for MLR—resulted in a subset with 15 descriptors. The MLR models of four experiments with different training and prediction sets yield an averaged SEP of 79, which is slightly better than the SEP for a PLS model using all descriptors. Note that a set of 15 descriptors, selected by maximum absolute correlation coefficients between descriptor and retention index, gives a much larger SEP between 154 and 158. PLS or PCR applied to the set with 15 stepwise selected descriptors gave the same models as MLR because the optimum number of components was 15 in these cases.

As discussed in Section 3.1, the applicability of 3D descriptors is not clear for the compounds used in this study. Excluding the 3D descriptors from the 529 descriptors resulted in a set of 301 descriptors. Application of PLS to these data yielded models with an average SEP of 79; this value is similar to the reference value 82 (PLS with 529 descriptors); so the 3D descriptors have almost no influence. The set of 15 stepwise selected descriptors contains two 3D descriptors (see Section 4.3); elimination of them increases SEP from 79 (15 descriptors) to 92 (13 descriptors).

Furthermore, in a preliminary test artificial neural networks [20,44] have been applied. To avoid a time-consuming feature selection the 15 descriptors found by stepwise selection were used. The final non-linear model obtained by a back propagation training had a similar performance as the MLR model computed from the same descriptors.

4.3. Discussion of a linear model with a small set of descriptors

The model which best combines simplicity and high predictive performance is the MLR model calculated from 15 descriptors obtained by forward stepwise feature selection. The parameters of this model, obtained from the training set with autoscaled data, are given in Table 4, together with the parameters of the corresponding PLS model. The importances, IMP_j, of the descriptors are very similar for MLR and PLS. This result is remarkable because MLR optimizes the fit to the training set, while PLS optimizes the prediction performance by cross-validation; obviously, the data set is large enough to give similar results for both approaches.

A final MLR model has been calculated with the same descriptor set, but using all 846 compounds. The regression coefficients and their standard errors (for unscaled descriptors as generated by software *Dragon*) are given in the last two columns of Table 4. They are very similar to those obtained from the training set with 700 compounds (not shown). Also the values for SEC and r_c (81 and 0.987, respectively) of this model are almost identical with the values obtained for the training set (82 and 0.987, respectively).

Selection of the 15 descriptors was based on mathematical criteria; therefore one cannot expect that all descriptors can be well interpreted in terms of chromatography. Features 1 and 2 are constitutional descriptors (group 1, Table 2). One of them, descriptor RBN is equal to the number of rotatable bonds; large numbers for RBN indicate a high flexibility of the molecules, and are connected with a decrease of the retention index (IMP is −0.146 for the MLR model). The other constitutional descriptor, *n*F, is the number of F-atoms; it has a similar—but less intensive—effect.

Features 3–5 are topological descriptors (group 2); Xt characterizes molecular branching; X2sol describes dispersion interactions in solution. The reciprocal distance index RDCHI increases with molecular size but decreases with molecular branching [32]. Descriptors 4 and 5 show the most prominent positive effect on the retention index of all 15 descriptors with IMP values of 0.387 and 0.572, respectively. For instance an increase of dispersion interactions increases the retention index, which is in agreement with experimental experiences. The retention data used in this work have been obtained on apolar stationary phases; it is known that for this type of phases, dispersion interactions are important. The other descriptor, RDCHI, with a high positive influence on the retention index characterizes size and branching of molecules. Also the influence of this descriptor corresponds to experience because a larger molecular size and less branching increase the retention index.

Descriptors 7 and 8 belong to the GETAWAY group (geometry, topology and atom-weights assembly), recently developed by Todeschini et al. [32]. They are the only 3D descriptors in the model. Descriptor 9, *n*CaR, describes the number of substituted aromatic C-atoms; it shows a distinctive positive effect on the retention index. Descriptor 10, *n*ROR, describes the number of aliphatic ether functions. Descriptors 11–15 give the number of specified atom-centered fragments present in the molecule; three of these fragments contain a hetero atom (R means aliphatic, Ar means aryl). The presence of fragment CH_3 has a small negative effect, presence of the other fragments a small positive effect on

Table 4
MLR model and PLS model (with 12 components) using 15 descriptors selected by a forward stepwise procedure

j	Descriptor code	Descriptor name or variable	*m*	*s*	IMP_j MLR	IMP_j PLS	b_j MLR
0		Intercept	–	–	–	–	−386.8 ± 39.3
1	RBN (1/0)	Number of rotatable bonds	6.177	4.698	−0.146	−0.129	−15.6 ± 1.2
2	*n*F (1/0)	Number of Fluorine atoms	0.080	0.490	−0.051	−0.050	−52.1 ± 7.0
3	Xt (2/2)	Total structure connectivity index	0.010	0.017	0.074	0.070	2149.0 ± 262.0
4	X2sol (2/2)	Solvation connectivity index chi-2	8.091	2.389	0.387	0.424	81.4 ± 2.4
5	RDCHI (2/2)	Reciprocal distance Randic-type index	2.929	0.528	0.572	0.536	543.2 ± 17.6
6	GATS2e (6/2)	Geary autocorrelation-lag 2	0.858	0.373	−0.050	−0.042	−67.6 ± 11.1
7	H5u (14/3)	H autocorrelation of lag 5/unweighted	0.861	0.656	0.155	0.141	118.5 ± 10.4
8	H6u (14/3)	H autocorrelation of lag 6/unweighted	0.627	0.576	0.050	0.054	43.3 ± 13.3
9	*n*CaR (15/1)	Number of substituted aromatic C (sp2)	2.789	2.022	0.209	0.209	51.8 ± 2.1
10	*n*ROR (15/1)	Number of ethers (aliphatic)	0.119	0.373	−0.057	−0.060	−77.0 ± 8.9
11	C-001 (16/1)	CH3R	1.131	1.288	−0.083	−0.087	−32.5 ± 3.5
12	H-050 (16/1)	H attached to hetero atom	0.764	0.903	0.152	0.154	84.5 ± 3.9
13	N-072 (16/1)	RCO–N<	0.351	0.639	0.088	0.089	69.0 ± 5.7
14	N-073 (16/1)	Ar2NH/Ar3N/R..N..R	0.069	0.352	0.062	0.062	88.5 ± 9.1
15	S-107 (16/1)	R2S/RS-SR	0.117	0.367	0.087	0.069	118.3 ± 9.7
I		Kovats retention index	2062	502	–	–	–

Descriptor codes (group/dimensionality) and names are from *Dragon* software [32]; *j*, descriptor number; *m*, mean (from training set with 700 compounds); *s*, standard deviation (from training set); IMP_j, importance of descriptor *j* (from training set). The regression coefficients b_j (±standard error) are for a MLR model trained with all 846 compounds using not scaled descriptors as generated by software *Dragon*.

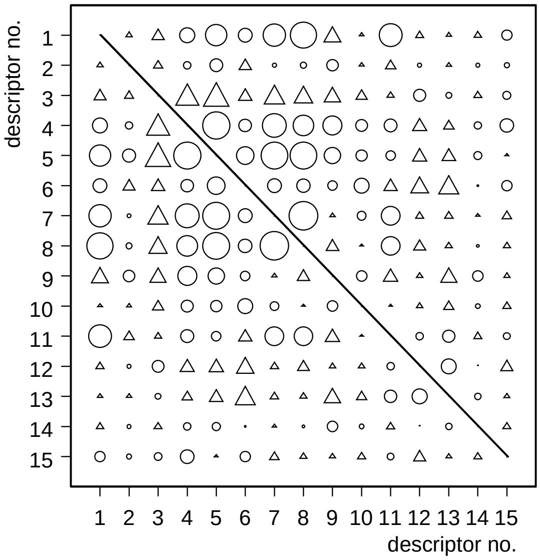


Fig. 2. Graphical representation of the correlation coefficients between the 15 selected molecular descriptors (see Table 4). Size of symbols is proportional to the absolute value of the correlation coefficient; positive values are denoted by circles, negative values by triangles.

the retention index. For instance, descriptor 12 denotes the number of hydrogen atoms attached to a hetero atom and has an importance IMP of 0.152 in the MLR model.

The highest correlation coefficients with the retention index, as calculated from training set data, have descriptors 4 (X2sol, correlation coefficient 0.871), 5 (RDCHI, 0.863), and 7 (H5u, 0.655). These values, however, are considerable lower than the correlation coefficients obtained for the output of multivariate models (for instance 0.987 for a MLR model using 15 descriptors). The correlation coefficient matrix of the 15 selected descriptors is graphically represented in Fig. 2. Maximum correlations occur between the two 3D descriptors number 7 and 8 (H5u and H6u, correlation coefficient 0.823) and between 4 and 5 (X2sol and RDCHI, 0.744). These four descriptors are similar when considering their correlations to the other descriptors. Such similarities of features can be recognized in the plot by a visual comparison of the symbols in the rows (or columns). For instance rows 5 and 7 are similar, also 3 and 12; but the first pair of descriptors is very different from the second pair.

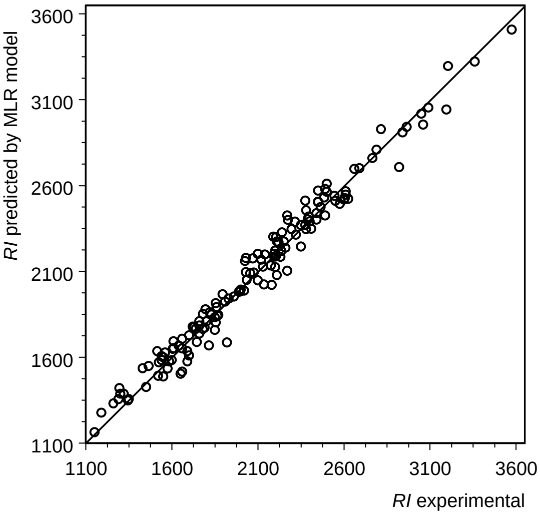
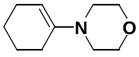
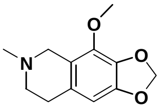
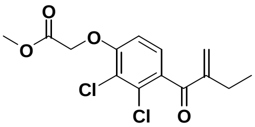
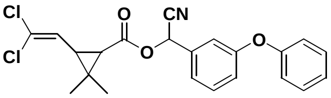


Fig. 3. Retention indices, I^*, predicted by an MLR model, vs. experimental values, I, from database. A subset of 15 descriptors obtained by forward stepwise feature selection has been used; size of training set was 700, size of prediction set 146 (shown); SEP is 74; r_p is 0.988.

In summary, molecular properties which are known to be relevant for gas chromatographic retention data—such

Table 5

Example compounds with experimental and predicted retention indices, *I*. Compound **1**: *N*-morpholino-1-cyclohexene (a psychedelic designer drug), CAS reg. no. 670804; compound **2**: hydrocotarnine (an ingredient of opium), CAS reg. no. 550107; compound **3**: methyl ethacrynate (a diuretic), CAS reg. no. 6463214; compound **4**: cypermethrin (an insecticide), CAS reg. no. 52315078

No.	Structure	Molecular mass	Experimental, *I*	Prediction error of *I*	
				MLR	PLS-529
1		167.25	1260	71	92
2		221.25	1790	−17	−66
3		317.16	2195	−9	62
4		416.30	2815	113	−97

For the MLR model, a subset of 15 descriptors obtained by forward stepwise feature selection has been used; for the PLS-529 model all descriptors.

as molecular size, branching, polar functional groups—are well covered by the automatically selected 15 descriptors.

4.4. Examples

In Fig. 3, the retention indices as predicted by a MLR model are compared with the experimental values from the database. The subset of 15 descriptors obtained by forward stepwise feature selection has been used; size of the training set was 700, size of the prediction set was 146. Four example compounds have been selected from the prediction set and detailed results are given in Table 5. The mean of the absolute errors for these four compounds is 52.5 for the MLR model and 79.3 for the PLS-529 model. Retention indices for compounds 2 and 3 are better predicted (with absolute errors between 9 and 62) than for compounds 1 and 4 (absolute errors between 71 and 113).

5. Summary and conclusions

Aim of the work was to investigate the applicability of standard software and standard methods—as widely used for the generation of molecular descriptors and for multivariate data analysis—for the prediction of gas chromatographic Kovats retention indices of toxicological relevant organic compounds. The used set of 846 compounds with two-dimensionally encoded molecular structures contains substances from diverse structural and toxicological categories. Software *WebLab Viewer* was applied to add hydrogen atoms as explicitly given atoms and to generate simple 3D structure proposals. The resulting chemical structures, encoded in Molfile format, were directly used as input for the *Dragon* software which calculated 1497 numerical descriptors for each molecular structure. A first step of feature selection eliminated constant descriptors and high correlations between descriptors, and resulted in a working data set with 529 descriptors. An advantage of software *Dragon* is a fast computation of descriptors, avoiding time-consuming quantum-chemical calculations.

Straight forward application of the chemometric standard method PLS by software *Unscrambler* to the data set with 529 descriptors gave linear models, exhibiting an averaged standard error of prediction of 82 retention index units. A drawback of this approach is the large number of descriptors in the model making an interpretation of the model parameters difficult.

For multiple linear regression a subset of 15 descriptors was selected by a forward stepwise variable selection procedure implemented in software *Systat*. Although this feature selection has the goal of best fitting the training data, the resulting linear model has a standard error of prediction of only 79 retention index units. An advantage of this approach is the fact that the influence of the selected descriptors on the predicted retention index can be easily interpreted.

A feature selection based only on maximum absolute correlation coefficients between descriptors and retention index, yielded models with a considerably lower prediction performance than obtained with all descriptors or with the subset of 15 descriptors described above. A feature selection based on maximum regression coefficients resulted in good PLS models with an averaged SEP of 79, however needing about 200 descriptors.

The rather large prediction errors are probably due to the high diversity of the investigated chemical structures, the different GC-column technologies used for establishing the data, and the simple methods applied for 3D structure generation, feature selection, and modeling.

For the identification of a priori unknown compounds as a part of systematic toxicological analyses (often called general unknown analysis), first the chromatographic retention data are obtained. Next step is a comparison of the measured mass spectrum with reference spectra from a database. The resulting hitlist contains the reference spectra most similar to the spectrum of the unknown, but such a hitlist not always allows the identification of an unknown. For instance, the GC–MS database [25] widely used for the evaluation of forensic analyses, contains about 20 compounds with mass spectra exhibiting almost only the base peak at mass 58, originating from ions $(CH_3)_2\ NCH_2{}^+$. These compounds cannot be identified solely by their mass spectra, however, most of them could be distinguished by their retention indices ranging from 1230 (cyclopentadrine, CAS reg. No. 102454) to 2565 (cyamemazine, CAS reg. No. 3546030). The method presented here enables an automatic estimation of retention indices from the molecular structure, using a model derived from about 800 toxicologically relevant compounds. The identification of unknowns in GC–MS analyses can be supported by excluding hitlist structures which give predicted retention indices very different from the experimental values; thus the identification of unknown poisons is facilitated.

Acknowledgements

The authors are grateful to R. Todeschini for providing the *Dragon* software and for helpful discussions, as well as to an unknown reviewer for constructive suggestions. Z.G.-N. thanks the Iranian Ministry of Science, Research and Technology for a scholarship supporting a research stay in Vienna. The work was supported by the Austrian Science Fund, project P14792-CHE.

References

[1] R. Kaliszan, Quantitative Structure–Chromatographic Retention Relationships, Wiley, New York, 1987.

[2] M. Karelson, Molecular Descriptors in QSAR/QSPR, Wiley, New York, 2000.

[3] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley/VCH, Weinheim, 2000.

[4] Q.S. Xu, D.L. Massart, Y.Z. Liang, K.T. Fang, J. Chromatogr. A 998 (2003) 155.

[5] A.R. Katritzky, K. Chen, U. Maran, D.A. Carlson, Anal. Chem. 72 (2000) 101.

[6] Y. Du, Y. Liang, D. Yun, J. Chem. Inf. Comput. Sci. 42 (2002) 1283.

[7] K. Heberger, Chromatographia 29 (1990) 375.

[8] J.M. Sutter, T.A. Peterson, P.C. Jurs, Anal. Chim. Acta 342 (1997) 113.

[9] R. Zhang, A. Yan, M. Liu, Z. Hu, Chemom. Intell. Lab. Syst. 45 (1999) 113.

[10] A. Yan, G. Jiao, Z. Hu, B.T. Fan, Comp. Chem. 24 (2000) 171.

[11] S. Liu, C. Yin, S. Cai, Z. Li, Chemom. Intell. Lab. Syst. 61 (2002) 3.

[12] J. Olivero, K. Kannan, J. Chromatogr. A 849 (1999) 621.

[13] T. Körtvelyesi, M. Görgenyi, K. Heberger, Anal. Chim. Acta 428 (2001) 73.

[14] M.H. Fatemi, J. Chromatogr. A 955 (2002) 273.

[15] B. Ren, Chemom. Intell. Lab. Syst. 66 (2003) 29.

[16] B.S. Junkes, R.D.M.C. Amboni, R.A. Yunes, V.E.F. Heinzen, Anal. Chim. Acta 477 (2003) 29.

[17] M. Jalali-Heravi, M.H. Fatemi, J. Chromatogr. A 915 (2001) 177.

[18] B. Lucic, N. Trinajstic, S. Sild, M. Karelson, A.R. Katritzky, J. Chem. Inf. Comput. Sci. 39 (1999) 610.

[19] M. Pompe, M. Novic, J. Chem. Inf. Comput. Sci. 39 (1999) 59.

[20] M. Jalali-Heravi, Z. Garkani-Nejad, J. Chromatogr. A 945 (2002) 173.

[21] C.G. Georgakopoulos, J.C. Kiburis, P.C. Jurs, Anal. Chem. 63 (1991) 2021.

[22] T.F. Woloszyn, P.C. Jurs, Anal. Chem. 64 (1992) 3059.

[23] W.P. Eckel, T. Kind, Anal. Chim. Acta 494 (2003) 235.

[24] T. Stimpfl, J. Jurenitsch, W. Vycudilik, J. Anal. Toxicol. 25 (2001) 125.

[25] K. Pfleger, H.H. Maurer, A. Weber, Mass Spectral and GC Data of Drugs, Poisons, Pesticides, Pollutants and their Metabolites, second ed., VCH, Weinheim, 1992.

[26] T. Stimpfl, W. Demuth, K. Varmuza, W. Vycudilik, J. Chromatogr. B 789 (2003) 3.

[27] C.H. Lochmüller, S.J. Breiner, in: C.L. Wilkins (Ed.), Computer-Enhanced Analytical Spectroscopy, Plenum Press, New York, 1993, pp. 187–215.

[28] M. Negwer, H.G. Scharnow, Organic-Chemical Drugs and their Synonyms, eight ed., Wiley/VCH, Weinheim, 2001.

[29] E. Kovats, Helv. Chim. Acta 41 (1958) 1915.

[30] D.A. Skoog, J.J. Leary, Principles of Instrumental Analysis, Saunders, Fort Worth, 1992.

[31] Y. Du, Y. Liang, Comput. Biol. Chem. 27 (2003) 339.

[32] R. Todeschini, V. Consonni, A. Mauri, M. Pavan, Software Dragon: Calculation of Molecular Descriptors, Department of Environmental Sciences, University of Milano-Bicocca, and Talete, srl http://disat.unimib.it/chm/Dragon.htm, Milan, Italy, 2003.

[33] Accelrys Inc., Software WebLab Viewer, http://www.accelrys.com, San Diego, CA, 1999.

[34] Camo Process-AS, Software Unscrambler, http://www.camo.no, Oslo, 2002.

[35] SPSS Inc., Software Systat, http://www.systat.com, Chicago, IL, 2000.

[36] The Mathworks Inc., Software Matlab, Natick, MA, 2000.

[37] MDL Information Systems Inc., CT file format, MDL Information Systems Inc., http://www.mdli.com/downloads/literature/ctfile.pdf, San Leandro, CA, 2002.

[38] R. Todeschini, V. Consonni, in: J. Gasteiger (Ed.), Handbook of Chemoinformatics, Wiley/VCH, Weinheim, 2003, pp. 1004–1033.

[39] M. Pintore, N. Piclin, E. Benfenati, G. Gini, J.R. Chretien, QSAR Comb. Sci. 22 (2003) 210.

[40] B.G.M. Vandeginste, D.L. Massart, L.C.M. Buydens, S. De Jong, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics: Part B, Elsevier, Amsterdam, 1998.

[41] R. Kramer, Chemometric Techniques for Quantitative Analysis, Marcel Dekker, New York, 1998.

[42] S. Wold, M. Sjöström, L. Eriksson, in: P.v.R. Schleyer, N.L. Allinger, T. Clark, J. Gasteiger, P.A. Kollman, H.F. Schaefer, III, P.R. Schreiner (Eds.), The Encyclopedia of Computational Chemistry, Wiley, Chichester, 1998, pp. 2006–2021.

[43] I.E. Frank, R. Todeschini, The Data Analysis Handbook, Elsevier, Amsterdam, 1994.

[44] M. Jalali-Heravi, Z. Garkani-Nejad, J. Chromatogr. A 927 (2001) 211.