Journal of Thermal Analysis and Calorimetry, Vol. 91 (2008) 1, 203–212

THERMAL PARAMETERS OF PHENYLCARBAMIC ACID DERIVATIVES USING CALCULATED MOLECULAR DESCRIPTORS WITH MLR AND ANN
Quantitative structure-property relationship studies

Jolanta Klos[1*], *P. Nowicki*[1] *and J. Cizmarik*[2]

[1]Department of Inorganic and Analytical Chemistry, Faculty of Pharmacy, University of Medical Sciences, Grunwaldzka 6,
60-780 Poznań, Poland
[2]Department of Pharmaceutical Chemistry, Comenius University, 832 32 Bratislava, Slovakia

The aim of this work was to build MLR and ANN models for predicting certain thermal parameters of phenylcarbamic acid derivatives. For 66 compounds belonging to this group DSC analysis was performed. Based on the DSC curves, nine thermal parameters were calculated. The chemical structure of newly synthesized local anaesthetic drugs was encoded in calculated theoretical descriptors. To build the QSPR models Multiple Linear Regression and Artificial Neural Networks were applied. The variable reduction in the case of MLR was performed by means of visual inspection of the significant loading plots obtained by Principal Component Analysis, but using forward selection. Two models of ANN were built: linear and non-linear, but for the reduction of the variables the genetic algorithm was applied. As a result, MLR and ANN models for predicting some thermal parameters of phenylcarbamic acid derivatives were obtained.

Keywords: *ANN, MLR, phenylcarbamic acid derivatives, QSPR, thermal analysis*

Introduction

Interest in the prediction of certain physico-chemical properties has grown during the last 10 years. Prediction of some properties is especially valuable when experimental determination is difficult or impossible. This concerns drug substances as well as substances included in the formulation. The calculation of some properties before synthesis would be desirable in the drug discovery phase. Elimination of compounds that are likely to possess very unfavourable physicochemical properties is possible.

Various physicochemical properties were predicted from the molecular structure, such as dielectric constants [1], boiling point [2, 3], heat capacity [4], thermal decomposition [5–7], ΔG and ΔH of formation [8], critical temperature and critical pressure [9], enthalpy of sublimation [10], etc. Many papers describe ANN and MLR models trained in parallel using the same descriptors and compound sets. Some comparisons have been published, for instance for $\log P_{\mathrm{oct}}$ [11, 12], boiling point [13], critical temperature and critical pressure [14]. The main advantage of artificial neural network modelling is that a non-linear relationship can be modelled without any assumptions; on the contrary, MLR models assume a linear relationship. The goal of this paper was to build and compare MLR and ANN models for the newly synthesised local anaesthetic drugs – derivatives of phenylcarbamic acid – describing the relationship between the chemical structure encoded in calculated theoretical descriptors and thermal parameters [15–19]. Thermal stability and parameters related to the melting point and thermal decomposition of new compounds are important physicochemical properties. As a result, on the basis of the obtained models we can predict some thermal parameters for compounds belonging to the same group before synthesis and without performing DSC analysis.

Experimental

The derivatives of phenylcarbamic acid with local anaesthetic and anti-arhythmic activity were analysed. These compounds were synthesized by Cizmarik and colleagues at Comenius University in Bratislava. 66 compounds are divided into 5 groups. The samples labelled BK (19 samples) are a series of 1-ethoxymethyl-2-pirolidyno(or piperidino- or azepane-) esters of 2-, 3- and 4-alkoxy(tetraoxy- to heptaoxy-) phenylcarbamic acid [20]. Samples labelled V (24 samples) are a series of 1-methyl-2-piperidinoethyl esters of 2-, 3- and

[*] Author for correspondence: jklos@amp.edu.pl

1388–6150/$20.00

Akadémiai Kiadó, Budapest, Hungary
Springer, Dordrecht, The Netherlands

4-alkoxy(metoxy- to decyloxy-)phenylcarbamic acid [21]. Samples B (12 samples) are a series of 2-dimethyloesters of 2-, 3- and 4-alkoxy(trioxy- to decyloxy-)phenylcarbamic acid [22]. Samples A (5 samples) are a series of 2-piperidinoethylesters of 2-, 3-and 4-alkoxy(trioxy- or hexaoxy)phenylcarbamic acid [23]. Samples labelled Z (6 samples) are a series of pirolidynoethylesters of 2-, 3- and 4-alkoxy(metoxy- or octanoxy)phenylcarbamic acid [24]. All samples (except samples labelled BK which are acetates) are hydrochloric salts.

DSC

A thermal study using DSC was carried out on Setaram Setsys TG-DSC 15 equipment. 2 mg samples were heated from 20 to 500°C at 5°C min^{-1} in corundum crucibles in nitrogen atmosphere.

Based on the TG and DTG curves, stages of thermal decomposition, mass losses and ranges of temperatures for each stage were determined. Based on the heat flow curve, temperatures of onset, the maximum of endothermic peaks and their enthalpy values were determined.

For all samples the thermal decomposition proceeds in three or four stages. The first stage is always connected with desorption of a small amount of adsorbed water (small mass losses observed on TG and DTG curves), but on the heat flow curve an endothermic peak corresponding to the melting of the compound is observed (temperature of onset, maximum of the melting peak and enthalpy was determined). The second (in the case of the samples with three steps of thermal decomposition) or the second and the third stage (in the case of the samples with four steps of thermal decomposition) occur with large mass losses due to the destruction of these organic compounds (TG and DTG curves), while at the same time a heat flow curve indicates one or two endothermic effects, respectively. The last stage of the thermal decomposition is linked to small mass losses on the TG and DTG curves; on the heat flow curve no effect is observed.

The samples for which the thermal decomposition proceeds in four stages are: samples BK (except sample BK 166); 26V; 28V; 29V; 30V; XXIV Z; XX B; XXI B; XXIV B; XXV B; XXVI B and XXVIII B. In the case of samples: BK166; samples 0-25V; A samples; I Z; II Z; III Z; XXII Z; XXIII Z, VII B; VIII B; IX B; XIII B; XIV B; XV B there are three stages of thermal decomposition.

For the interpretation of the recorded curves it was necessary to select suitable parameters for all stages. Therefore, the endothermic effects illustrated by the heat flow curve of each sample required the designation of the onset (P1o, P2o and P3o), the maximum (P1m, P2m and P3m) of the peak temperature and their enthalpy (P1E, P2E and P3E) (Fig. 1).

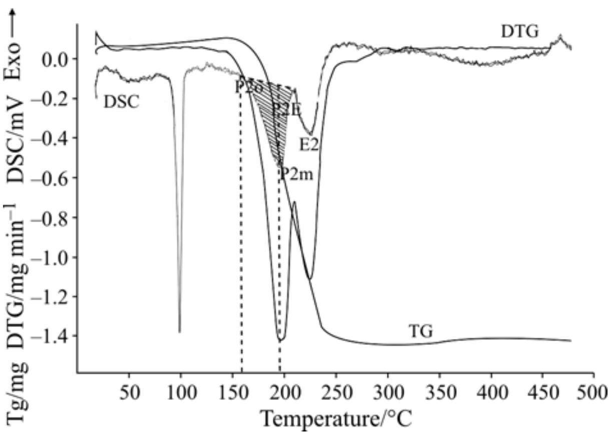

Fig. 1 The example of DSC curve for investigated compounds

Molecular descriptors

The molecular descriptors used consist of 0D, 1D, 2D and 3D theoretical descriptors [25]. For all molecules the geometrical structure was optimised using HyperChem Release 7.0 Professional software. Geometry optimisation was obtained by the semiempirical method AM1 (Austin Model 1) using the Polak-Ribiere conjugate gradient algorithm with an RMS gradient of 0.01 kcal Å^{-1} mol^{-1}) as a stop criterion. The Cartesian coordinate matrices of the positions of the atoms in the molecule were used for the calculation of 1264 molecular descriptors using Dragon 5.3 software. The following groups of descriptors were calculated: constitutional descriptors, topological descriptors, walk and path counts, connectivity indices, information indices, 2D autocorrelations, edge adjacency indices, Burden eigenvalues, topological charge indices, eigenvalue-based indices, Randic molecular profiles, geometrical descriptors, RDF descriptors, 3D-MoRSE descriptors, WHIM descriptors, GETAWAY descriptors, functional group counts, atom-centred fragments, charge descriptors and molecular properties. Additionally, some QSAR properties using HyperChem 7.0 were calculated, such as the approximate solvent accessible surface area, grid solvent accessible surface area, molecular volume, refractivity, polarizability and molecular mass. Furthermore, the following energies and gradient calculated using HyperChem software were added to the dataset: total energy, binding energy, isolated atomic energy, electronic energy, Core-Core interaction, heat of formation and gradient.

Regression models

MLR modelling

Variable reduction consists of selecting a subset of variables able to preserve the essential information contained in the whole dataset, but eliminating redundancy, a too highly intercorrelated variable, etc. For variable reduction visual inspection of the significant loading plots obtained by Principal Component Analysis was used. This is a useful tool to select the most relevant variables to preserve the most important information contained in the original data [26, 27]. Using this method, the number of variables was reduced to about 70 yielding the most important information [28].

The next step was variable selection in order to reach optimal model complexity in predicting the response variable. Regression models with a few predictor variables are simple and stable in the statistical sense, have high predictive power and can be easily interpreted. To avoid selection of variables characterised by chance correlation validation techniques were applied.

Forward selection was used to select a variable. This technique starts with no variables in the model and one variable is added at a time until the stopping criterion is satisfied. The variable considered for inclusion at any step is the one yielding the largest single degree of freedom F-ratio among the variables eligible for inclusion, and this value is larger than the fixed value F_{in}. At each step the j variable is added to a k-size model if

$$F_{\mathrm{j}} = \max{}_{\mathrm{j}}\left(\frac{RSS_{\mathrm{k}} - RSS_{\mathrm{k+j}}}{s^2_{\mathrm{k+j}}}\right) > F_{\mathrm{in}}$$

where RSS is the residual sum of squares. The subscript $k+j$ refers to quantities computed when the j^{th} variable is added to the current k variables already in the model.

Linear regression models were obtained using software Statistica 6.0 with calculated statistics: goodness of fit (coefficient of determination R^2, multiple correlation coefficient R, F-ratio, standard deviation errors) and goodness of prediction. As a validation procedure cross-validation was applied. 20% of objects excluded from the dataset make up a testing set and responses for them were predicted on the basis of the model obtained. For each model the standard deviation s_{PRESS} (calculated from PRESS, the sum of the squared errors of these predictions, divided by the number of degrees of freedom) was calculated. Outliers were detected by means of a plot of distribution of residuals.

Artificial neural networks

Artificial neural networks were applied. To select the number of variables (theoretical descriptors) a genetic algorithm was used. Genetic algorithms are one of the most effective optimization methods for problems involving a large number of variables [29, 30]. The principle of this method is the evaluation of population models and is loosely based on the Darwinian theory of evolution. In the genetic algorithm terminology, the binary vector I is called chromosome, which is a p-dimensional vector, where each position corresponds to a variable (one if included in a model, otherwise zero). Each chromosome represents a model with a subset of variables [31].

The optimal statistical parameters are defined, along with the model population P and the maximum number (L) of allowed variables in the model, the minimum number of allowed variables is usually equal to one. Moreover, a cross-over-probability p_{C} must also be defined by the user.

Once the leading parameters are defined, an algorithm evaluation starts, based on four main steps:

Random initialization of the population

The model population is initially built by random models with a number of variables between 1 and L, and then models are ordered with respect to the selected statistical parameters – the quality of the models – the best model is in the first place, the worst one at position P.

Crossover step

From the population, pairs of models are selected. Then, for each pair of models, the common characteristics are preserved. For a variable included in one model and excluded from the other, a random number is tried and compared with the cross-over-probability p_{C}: if the random number is lower than the cross-over-probability, the excluded variable is included in the model and vice versa. Finally, the statistical parameters for the new models are calculated: if the parameter value is better than the worst value in the population, the model is included in the population, in the place corresponding to its rank; otherwise it is no longer considered. This procedure is repeated for several pairs.

Mutation step

For each step present in the population (i.e. each chromosome), p random numbers are tried, and one at a time are compared with the defined mutation probability p_{M}: each gene remains unchanged if the corresponding random number exceeds the mutation probability; otherwise, it is changed from zero to one and vice versa. Once the mutated model is obtained, the statistical parameters for the models are calculated. If the parameters value is better than the worst value in the population, the model is included in the population, in the place corresponding to its rank, otherwise it is no longer considered.

Stop condition

The second and the third steps are repeated until some stop condition (mutation and crossover coefficients) is encountered or the process is arbitrarily ended.

It is well known that the genetic algorithm is very flexible and there may be many variations in the traditional genetic algorithm. Using software Statistica 6.0 Holland's genetic algorithm was used with such parameters as: population size, 100; maximum number of generations to execute, 100; mutation coefficient, 0.1; crossover coefficient, 1.0; sampling, 1; unit penalty, 0.0001. The selection was simulated by a simple roulette wheel-based scheme. The goodness of fit is the validation error plus the penalty factor, which is multiplied by the number of features considered. The linear normalization was performed before the selection in such a way so that the quotient of the best and the worst fitness is 2:1 in order to find the global optimum. The assumption in this algorithm is that all strings (with the exception of the best strings) are completely replaced in each new generation.

The multilayer perceptrons (MLP) employed in this study are artificial neural networks with a layered structure and all connections feeding forwards from inputs towards outputs (feed-forward neural networks) [32, 33]. The best architecture, the number of layers and the number of neurons in the hidden layers, training and testing methods were achieved by trial and error. The logistic function was used as an activation function. Two supervised learning algorithms were applied: during the first stage, back propagation of error and in the second conjugate gradient descent. In the back-propagation algorithm a series of input data (e.g. theoretical descriptors after the reduction by means of the genetic algorithm) and their known related output (e.g. one of the 9 thermal decomposition parameters) were presented to the network. The correction of the masses was made iteratively in epochs presenting the training set until the obtained output data were equal to the expected value (target) within a specified threshold. The learning rate determining the size of the change in the masses was 0.01 and the momentum coefficient 0.3. The conjugate gradient descent is a batch update algorithm. The entire training set is fed through the network and used to adjust the network masses at the end of the epoch, not after each case. At the start of the training, the masses were assigned randomly. Additionally, the neurons with the smallest masses (<0.05) were removed.

The set of objects was split into three subsets: the training, validation and testing subset. The training subset was the largest, the validation and testing consist of 13.6% of all objects chosen randomly. During network learning, the RMS error of the training subset is minimized. In order to avoid overtraining, especially in the case of a small number of objects, Gaussian noise was added. Objects belonging to the validation subset are excluded from the training process and used in monitoring the progress of the learning process. The final model was obtained after learning by means of the training subset and checking by means of the validation subset is then tested by means of the testing subset. The objects belonging to the testing subset were totally excluded from the training process. To assess the quality of the models the RMS errors of each subset should be taken into consideration. Equivalence goodness of fit in the linear regression models in MLP are RMS error for the training subset and correlation *R* (Pearson's correlation coefficient for real values and obtained as an output value of a certain model). The equivalence goodness of prediction in linear regression models in the case of MLP are: the quality of validation, the quotient of standard deviations, RMS errors for the testing and the validation subsets.

After building MLP models, the linear models of artificial neural networks were performed. The inputs were the same variables (theoretical descriptors) as in the case of MLP – after the selection by the genetic algorithm. It was done to compare which function, linear or nonlinear, better reflects the relationship between variables. Artificial neural networks with the linear activation function do not have any hidden layers. Other parameters of training were the same as in the case of MLP.

Results and discussion

MLR analysis

Among the descriptors, the most significant theoretical molecular descriptors after the reduction using PCA were identified by means of the multiple linear regression analysis with a stepwise forward selection method.

The equations obtained for the first endothermic peak (the onset – P1o, the maximum temperature of the peak – P1m and the enthalpy of the process – P1E, respectively) are:

$$\log P1o=19.4369(\pm 9.1540)AROM+ \\ 4.7909(\pm 2.9531)R5m^{+}+ \\ 0.7634(\pm 0.2932)PJI3 -17.6652(\pm 8.9723$$

$$\log P1m=13.6576(\pm 3.8972)RTp^{+}+ \\ 12.2228(\pm 5.4109)AROM+8.0574(\pm 2.2330)R3m^{+}+$$

3.0272(±1.8575)R7m^{+}+0.5773(±0.2951)HATS0m+
0.04662(±0.4100)E1e+0.2048(±0.1437) Mor20v–
1.9779(±0.4611)H0p–22.2319(±7.0304)R2v^{+}–
8.3794(±5.2235)

logP1E=0.8525(±0.7550) MATS8v+
0.3905(±0.2773) PJI3+1.6118(±0.2179)

One outlier was detected in the case of each equation by means of a plot of distribution of residuals – sample BK166. The equations obtained for the second endothermic peak after the outlier – sample BK166 – had eliminated (the onset – P2o, the maximum temperature of the peak – P2m and the enthalpy of the process – P2E, respectively) are:

logP2o=1.6702(±0.4540)BEHm2+
1.3954(±0.4087)X4Av+0.7488(±0.5193)+R3u^{+}+
0.1821(±0.0581)H8m+0.1119(±0.0693)GATS7v+
0.1097(±0.0233) HATS0m+
0.1018(±0.0777)MATS8p+
0.0279(±0.0130G(N..N)+
0.0268(±0.0148) Mor27u–0.4397(±0.0896)Dm–
4.3531(±1.7055)

logP2m=2.4958(±0.4108) R2u^{+}+
0.5761(±0.2604)R2m^{+}+
0.2162(±0.0581) H8m+0.1833(±0.0534)GATS4m+
0.0392(±0.0116)G(N..N)+
0.0092(±0.0080)Psychotic-80–
2.0268(±0.2852)R2e^{+}+1.9093(±0.0767)

logP2E=0.4484(±0.1238) Inflammat-80
+0.2204(±0.1435) Mor15u+2.3209(±0.1014)

The equations obtained for the third endothermic peak after the outlier – sample IZ – had been eliminated (the onset – P3o, the maximum temperature of the peak – P3m and the enthalpy of the process – P3E respectively) are:

logP3o=10.4894(±6.5234)BEHm3+
0.2477(±0.0721) RBN–32.2997(±21.8304)

logP3m=10.5214(±6.5535) BEHm3+
0.2495(±0.0725)RBN–39.4280±21.9311

logP3E=9.4069(±5.7086) BEHm3+
0.1914(±0.0631) RBN–35.0422(±19.1035)

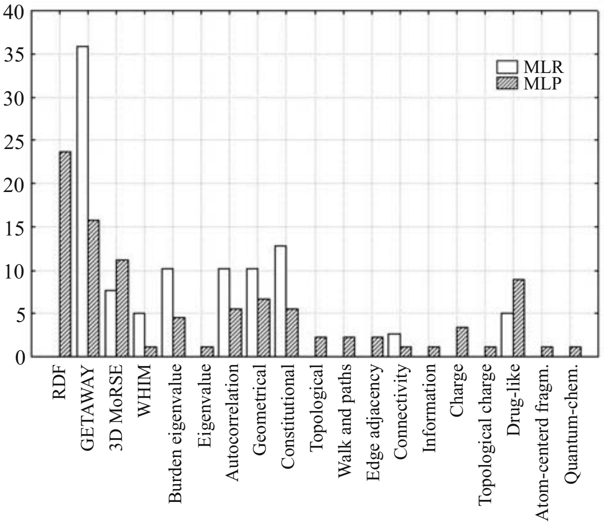

Fig. 2 Contribution of descriptors in MLR and MLP models

Table 1 Statistical parameters for MLR models

Thermal parameters	R	R^2	s	F	FD	s_{PRESS}
P1o	0.5820	0.3387	0.3473	7.8118	4.61	0.2280
	0.6865	0.4713	0.1278	18.127	3.61	0.1086
P2o	0.5258	0.2765	0.3562	7.8992	3.62	0.1086
	0.9727	0.9462	0.0107	94.909	10.54	0.0149
P3o	0.8060	0.6497	1.0127	58.424	2.63	0.9188
	0.8627	0.7443	0.8667	90.231	2.62	0.8827
P1m	0.4837	0.2340	0.3587	9.6231	2.63	0.8827
	0.8961	0.8030	0.6551	24.916	9.55	0.0548
P2m	0.4500	0.2025	0.3768	7.9967	2.63	0.2479
	0.9628	0.9270	0.1292	103.44	7.57	0.0117
P3m	0.8065	0.6394	1.0174	58.622	2.63	0.9211
	0.8630	0.7448	0.8707	90.500	2.62	0.8844
P1E	0.4867	0.2369	0.3407	9.7779	2.63	0.3110
	0.4408	0.1943	0.1202	7.4759	2.62	0.1058
P2E	0.6662	0.4439	0.4169	12.172	4.61	0.3592
	0.7332	0.5376	0.2391	36.039	2.62	0.2507
P3E	0.7892	0.6228	0.8682	52.017	2.63	0.6808
	0.8446	0.7042	0.7585	77.175	2.62	0.6388

The statistical parameters for all equations are shown in Table 1.

Using this comparison of statistical parameters we can stay that the best models were: P2m and P2o. The values of R^2 (0.9270 and 0.9462, respectively) and F values (103.44 and 94.909, respectively) indicate the best fit among all the models. The values of s_{PRESS} (0.0117 and 0.0149, respectively) indicate goodness of prediction.

There is a different number of descriptors in these models. Only models describing the third peak connected with thermal decomposition are three-parameter models. The contribution of descriptors in MLR models is shown in Fig. 2. The selected descriptors in large numbers are: GETAWAY, 3D-MoRSE, geometrical descriptors, Burden eigenvalue descriptors and WHIM descriptors. All these descriptors (except Burden eigenvalue descriptors which are 2D) are derived from the three-dimensional structure of the molecule. Geometrical descriptors are commonly known as topographic indices and are calculated from the graphical representation of molecules but using geometric distances between atoms instead of the topological distances. WHIM descriptors (Weighted Holistic Invariant Molecular descriptors) are geometrical descriptors based on statistical indices calculated from the projections of atoms along principal axes. They are built in such a

Table 2 Descriptors in MLR models

Descriptor	Definition	Descriptor class (%contribution of the class)
R5m+	R maximal autocorrelation of lag 5/weighted by atomic masses	GETAWAY (35.9%)
RTp+	R maximal index/weighted by polarizabilities	
R3m+	R maximal autocorrelation of lag 3/weighted by atomic masses	
R7m+	R maximal autocorrelation of lag 7/weighted by atomic masses	
R2v+	R maximal autocorrelation of lag 2/weighted by atomic van der Waals volumes	
R2u+	R maximal autocorrelation of lag 2 / unweighted	
R2m+	R maximal autocorrelation of lag 2 / weighted by atomic masses	
R2e+	R maximal autocorrelation of lag 2 / weighted by atomic Sanderson electronegativities	
R3u+	R maximal autocorrelation of lag 3 / unweighted	
H8m	H autocorrelation of lag 8 / weighted by atomic masses	
HATS0m	Leverage-weighted autocorrelation of lag 0/weighted by atomic masses	
H0p	H autocorrelation of lag 0/weighted by atomic polarizabilities	
MOR20v	3D-MoRSE signal 20/weighted by atomic van der Waals volumes	3D-MoRSE (7.7%)
MOR27u	3D-MoRSE signal 27 / unweighted	
Mor15u	3D-MoRSE signal 15 / unweighted	
MATS8v	Moran autocorrelation lag 8/ weighted by atomic van der Waals volumes	2D Autocorrelation
MATS8p	Moran autocorrelation lag 8 / weighted by atomic polarizabilities	(10.2%)
GATS4m	Geary autocorrelation lag 4 / weighted by atomic masses	
GATS7v	Geary autocorrelation lag 7 / weighted by atomic van der Waals volumes	
E1e	1st component accessibility directional WHIM index /weighted by atomic Sanderson electronegativities	WHIM (5.2%)
Dm	D total accessibility index / weighted by atomic masses	
PJI3	Petitjean shape index	Geometrical
G(N..N)	Sum of geometrical distances between N..N	(10.2%)
Psychotic-80	Ghose-Viswanaghan-Wendolowski antipsychotic at 80%	Drug-like
Inflammat-80	Ghose-Viswanaghan-Wendolowski antiinflammatory at 80%	(5.2%)
BEHm2	Highest eigenvalue n.2 of Burden matrix/weighted by atomic masses	Burden eigenvalue
BEHm3	Highest eigenvalue n.3 of Burden matrix/weighted by atomic masses	(10.2%)
AROM	Aromacity index	Constitutional
RBN	Number of rotatable bonds	(12.8%)
X4Av	Average valence connectivity index chi-4	Connectivity (2.6%)

way so as to capture relevant molecular 3D information regarding molecular size, shape, symmetry and atom distribution with respect to invariant reference frames. The GETAWAY (GEometry, Topology, and Atom-Weights AssemblY) descriptors have been recently proposed as chemical structure descriptors derived from a new representation of molecular structure, the Molecular Influence Matrix (MIM).

3D-MoRSE (3D-Molecule Representation of Structures based on Electron diffraction) descriptors are based on the idea of obtaining information from the 3D atomic coordinates by the transformation used in electron diffraction studies for preparing theoretical scattering curves. All descriptors used in MLR models are shown in Table 2. In general, the 3D descriptors can add valuable information to the models. It can be concluded that the geometrical properties of a molecule play a major role in the thermal decomposition of the investigated compounds. Because most of the models include many descriptors belonging to different blocs of descriptors, it is very difficult to state which of them considerably influence thermal decomposition.

The multilayer perceptrons (MLP) obtained after selection by means of the genetic algorithm were artificial neural networks with a layered structure. There is a different number of inputs and one or rarely two hidden layers with different number of neurons. The output is always one, one of the thermal parameters. The architectures of the obtained MLP models and the values of goodness of fit and prediction are shown in Table 3. The RMS errors for testing sets have values from 0.10 to 0.50 in all models. The values for the RMS error of the validation set vary. On the basis of the values of correlation R and RMS errors for training sets, we can state that the best models are P3o and P3E. Taking into consideration the values of RMS errors for testing sets, the quality of validation and the quotient of standard deviations, we can state that the models with the best predictive power are models P2o and P1m. But the high values of RMS errors for validation sets in these models, in comparison with other models, decrease the quality of the models obtained. As we can see, not always do the MLP models with a good predictive power have favourable values of goodness of fit and vice versa.

Table 3 Statistical parameters for MLP models

Thermal parameter	Architecture (input:hidden layer:output)	RMSE of training	RMSE of validation	RMSE of testing	Quality of validation	Quotient of standard deviation	Correlation *R*
P1o	14:14-26-1:1	0.200047	1.058804	0.174822	0.900026	0.896618	0.442805
P2o	7:7-6-1:1	0.143458	7.240020	0.147026	0.972441	0.968661	0.349920
P3o	6:6-57-16-1:1	0.215974	0.286372	0.358701	0.583952	0.509280	0.860611
P1m	11:11-36-1:1	0.171843	1.599036	0.163821	0.962817	0.940619	0.339815
P2m	20:20-7-1:1	0.230894	7.474009	0.242629	0.971847	0.968739	0.333679
P3m	12:12-25-1:1	0.283215	0.265903	0.327617	0.541053	0.588172	0.808741
P1E	10:10-40-1:1	0.521268	1.558770	0.507898	0.910662	1.149848	0.215687
P2E	2:2-53-18-1:1	0.302590	0.865671	0.327639	0.898940	0.890506	0.461233
P3E	7:7-28-1:1	0.139923	0.218448	0.357674	0.408295	0.436268	0.899817

Table 4 Statistical parameters for linear MLP models

Thermal parameter	Architecture (input:hidden layer:output)	RMSE of training	RMSE of validation	RMSE of testing	Quality of validation	Quotient of standard deviation	Correlation *R*
P1o	14:14-1:1	0.148772	1.135043	0.076587	0.960643	0.910822	0.426452
P2o	7:7-1:1	0.142315	7.264731	0.161471	0.976155	0.972079	0.319603
P3o	6:6-1:1	0.255881	0.332012	0.452602	0.679152	0.612108	0.796950
P1m	11:11-1:1	0.133843	1.602268	0.112721	0.955982	0.924589	0.391110
P2m	20:20-1:1	0.048649	7.576125	0.061248	0.977885	0.972966	0.301141
P3m	12:12-1:1	0.868899	0.905274	0.964592	1.731633	1.803086	0.318188
P1E	10:10-1:1	0.158889	1.684620	0.336403	0.993472	0.986841	0.161723
P2E	2:2-1:1	0.220540	0.918225	0.356451	0.907076	0.891769	0.454691
P3E	7:7-1:1	0.211014	0.279716	0.417907	0.554612	0.589754	0.807728

Table 5 Descriptors in MLP models

Descriptor	Definition	Descriptor class (% contribution of the class)
ESpm03u/05r	Spectral moment 03 from edge adj. matrix /spectral moment 05 from edge adj. matrix weighted by resonance integrals edge	Edge adjacency ind. 2.3%
BELv1/BELm4/BELp5	Lowest eigenvalue n. 1 of Burden matrix weighted by atomic van der Waals volume/lowest eigenvalue n. 4 of Burden matrix weighted by atomic van der Waals volume/lowest eigenvalue n.5 weighted by atomic polarizabilities	Burden eigenvalue 4.5%
BEHm5	Highest eigenvalue n. 5 of Burden matrix weighted by atomic masses	
VEA2	Average eigenvector coefficient sum from adjacency matrix	Eigenvalue-based ind.1.1%
RCI	Jug RC index	Geometrical 6.7%
ASP	Asphericity	
DISPe	d COMMA2 value/ weighted by atomic Sanderson electronegativities	
RDF 100u/150u	Radial Distribution Function – 10/15/unweighted	RDF 23.6%
080m/125m	0.80/12.5/weighted by atomic masses	
020p/055p/110p/145p/155p	2.0/5.5/11.0/14.5/15.5/weighted by atomic polarizabilities	
060v/065v/100v/150v/155v	6.0/6.5/10.0/15.0/15.5/weighted by atomic van der Waalsa volume	
020e/085e/125e/135e/155e	2.0/8.5/12.5/13.5/15.5/weighted by atomic Sanderson electronegativities	
Mor04u/18u/29u	3D-MoRSE-signal 04/18/29/unweighted	3D MoRSE 11.2%
Mor10p/25p	3D-MoRSE-signal 10/25/weighted by atomic polarizabilities	
Mor17e	3D-MoRSE-signal 10/weighted by atomic Sanderson electronegativities	
Mor11m/20m/26m/28m	3D-MoRSE-signal 11/20/26/28/weighted by atomic masses	
R6u+/8u+	R maximal autocorrelation of lag 6/8/unweighted	GETAWAY 15.7%
R8e	R autocorrelation of lag 8/ weighted by atomic Sanderson electronegativities	
R2e+/5e+	R maximal autocorrelation of lag 2/ 5/weighted by atomic Sanderson electronegativities	
RTe+	R maximal index/ weighted by atomic Sanderson electronegativities	
R5m+	R maximal autocorrelation of lag 5/weighted by atomic masses	
HATS0m/4m	Leverage-weighted autocorrelation of lag 0/4/ weighted by atomic masses	
HATS3e	Leverage-weighted autocorrelation of lag 3/weighted by atomic Sanderson electronegativities	
HATS0p/1p	Leverage-weighted autocorrelation of lag 0/1/weighted by atomic polarizabilities	
MATS1m/2m	Moran autocorrelation lag 1/2/ weighted by atomic masses	Autocorrela tion 5.6%
MATS4e	Moran autocorrelation lag 4/ weighted by atomic electronegativities	
MATS2v	Moran autocorrelation lag 2/ weighted by atomic van der Waals volumes	
MATS2p	Moran autocorrelation lag 8/ weighted by atomic polarizabilities	
Psychotic-50/80	Ghose-Viswanadhan-Wendolowski: antipsychotic at 50/80%	Drug-like 9.0%
Neoplastic-50	antineoplastic at 50%	
Hypertens-80	antihypertensive at 80%	
Hypnotic-80	hypnotic at 80%	
Infective-80	antiinfective at 80%	
GVWAI-80	alert index at 80%	

Table 5 Continued

RNCG	Relative negative charge	Charge descript. 3.4%
TE1	Topological electronic descriptor	
Q2	Total squared charge	
GGI6	Topological charge index of order 6	Topological charge ind. 1.1%
E3s	3rd component accessibility directional WHIM index/ weighted by atomic electrotopological states	WHIM 1.1%
SRW05/09	Self-returning walk count of order 5/9	Walk and path counts 2.3%
X0A/X2A	Average connectivity index chi-0/chi-2	Connectivity 1.1%
D/Dr07	Distance/detour ring index of order 7	Topological 2.3%
CIC5	Complementary information content	Information 1.1%
C-005	CH3X atom-centred fragments	Atom centred fragm. 1.1%
nR05/07	Number of 5/7-membered rings	Constitutional 5.6%
x4sol	Salvation connectivity index chi-4	
Bind.E	Energy of binding	Quant.-chem d. 1.1%

In order to verify whether the linear function can better reflect the relationship between the thermal parameters (output) and theoretical descriptors (inputs) selected by means of the genetic algorithm, linear models of artificial neural networks were tested. These are models without any hidden layers and the inputs are the same as in the case of MLP models. The statistical parameters for linear models of MLP are shown in Table 4. In nearly all cases, the values of correlation R indicate the weakness of obtained models as regards the goodness of fit. Mostly, the values of RMS errors of training and testing sets are slightly better for linear models than in nonlinear models. One exception is models describing the third process of thermal decomposition (the third peak on the DSC curve). In the case of models: *P3o*, *P3m* and *P3E* the indices of goodness of fit and goodness of prediction are worse in linear models.

Comparing linear and nonlinear MLP models, we can observe that the values of RMS errors of three subsets – training, validation and testing – are not in the same order. Especially the values of RMS errors of validation sets are high. This might result from unequal contribution of samples belonging to different groups: BK, V, A, Z and B to certain subsets. It particularly concerns the groups A and Z. These groups contain only 5 and 6 compounds and because of this, the compounds are not included in each subset. As a result, we obtained models with doubtful quality.

The contribution of descriptors in MLP models are shown in Fig. 2. The selected descriptors (Table 5) in large numbers are: RDF, GETAWAY, WHIM, 3DMoRSE and geometrical. RDF descriptors are based on a radial distribution function, which is described as a probability distribution of finding an atom in the spherical volume of radius *R*. The other descriptors are described in short in the previous section (MLR models).

Conclusions

The aim of this work was to build QSPR models describing the relationship between thermoanalytical parameters and chemical structure of investigated esters of phenylcarbamic acid. A thermal study using DSC for each compound was carried out. Using the DSC curve, 9 thermal parameters were calculated. To build the QSPR models, multiple linear regression and Artificial Neural Networks were applied. The chemical structure was encoded in the calculated theoretical descriptors. The variable reduction was performed by means of visual inspection of the significant loading plots obtained by Principal Component Analysis, but the selection using forward selection. For each thermal parameter, MLR models were calculated with certain indices of goodness of fit and goodness of prediction. The best MLR models are: *P2o* and *P2m*.

Two models of Artificial Neural Networks were built, linear and nonlinear, in order to find out which function better describes the relationship between variables. But in the case of ANN for reduction the variables, the genetic algorithm was applied. The results are ambiguous. By far the best MLP models were derived for thermal parameters describing the third peak on the DSC curve: *P3o*, *P3m* and *P3E*. But in the case of other models the nonlinear ANN models have better indices of goodness of fit than linear ANN models, but very often worse indices of goodness of prediction. For the thermal parameters describing the first and the second process of thermal decomposition MLR models with assumed linear relationship between thermal parameters and chemical structure of investigated compounds can be recommended.

References

1 S. Sild and M. Karelson, J. Chem. Inf. Comput. Sci., 42 (2002) 360.

2 J. Homer, S. C. Generalis and J. H. Robson, Phys. Chem., 1 (1999) 4075.

3 A. J. Chalk, B. Beck and T. Clark, J. Chem. Inf. Comput. Sci., 41 (2001) 457.

4 D. W. Noid, M. Varma-Nair, B. Wunderlich and J. A. Darsey, J. Thermal Anal., 37 (1991) 2295.

5 R. C. O. Sebastiao, J. P. Braga and M. I. Yoshida, J. Therm. Anal. Cal., 74 (2003) 811.

6 N. Sbirrazzuoli, D. Brunel and L. Elegant, J. Thermal Anal., 49 (1997) 1553.

7 M. Wesolowski and B. Suchacz, J. Therm. Anal. Cal., 68 (2002) 893.

8 A. A. Gakh, E. G. Gakh, B. G. Sumpter and D. W. Noid, J. Chem. Inf. Comput. Sci., 34 (1994) 832.

9 G. Espinosa, D. Yaffe, A. Arenas, Y. Cohen and F. Giralt, Ind. Eng. Chem. Res., 40 (2001) 2757.

10 M. H. Charlton, R. Y. Docherty and M. G. Hutchings, J. Chem. Soc., Perkin. Trans., 2 (1995) 2023.

11 J. Huuskonen, D. Livingstone and I. V. Tetko, J. Chem. Inf. Comput. Sci., 40 (2000) 947.

12 D. J. Livingstone, M. G. Ford, J. J. Huuskonen and D. W. Salt, J. Comput.-Aided Mol. Des., 15 (2001) 741.

13 M. D. Wessel and P. C. Jurs, J. Chem. Inf. Comput. Sci., 35 (1995) 68.

14 B. E. Turner, C. L. Costello and P. C. Jurs, J. Chem. Inf. Comput. Sci., 38 (1998) 639.

15 M. A. S. Silva, R. G. Kelmann, T. Foppa, A. P. Cruz, C. D. Bertol, T. Sartori, A. Granada, F. Carmignan and F. S. Murakami, J. Therm. Anal. Cal., 87 (2007) 463.

16 V. A. Drebushchak, T. P. Shakhtschneider, S. A. Apenina, T. N. Drebushchak, A. S. Medvedeva, L. P. Safronowa and V. V. Boldyrev, J. Therm. Anal. Cal., 84 (2006) 643.

17 D. Kiss, R. Zelko, Cs. Novák and Zs. Éhen, J. Therm. Anal. Cal., 84 (2006) 447.

18 K. Michalik, Z. Drzazga, A. Michnik and M. Kaszuba, J. Therm. Anal. Cal., 84 (2006) 119.

19 A. E. Almeida, A. G. Ferreira, M. S. Crespi, Z. A. Andrade and M. C. Chung, J. Therm. Anal. Cal., 83 (2006) 277.

20 E. Sedlarova, L. Buciova and J. Cizmarik, Ceskoslov. Farm., 45 (1996) 139.

21 M. Pokorna, J. Cizmarik, E. Sedlarova and E. Racanska, Ceskoslov. Farm., 48 (1999) 80.

22 J. Cizmarik, M. Mitosinkova, A. Borovansky and P. Svec, Pharmazie, 33 (1978) 509.

23 J. Cizmarik, A. Borovansky and P. Svec, Ceskoslov. Farm., 25 (1976) 118.

24 J. Cizmarik, E. Polasek, P. Svec and E. Racanska, Ceskoslov. Farm., 45 (1993) 139.

25 R. Todeschini and V. Consonni, Handbook of Molecular Descriptors, Wiley-VCH, Weinheim 2000.

26 I. T. Jolliffe, Principal Component Analysis, Springer-Verlag, New York 1986.

27 J. E. Jackson, A User's Guide to Principal Components, Wiley, New York 1991.

28 S. Derksen and H. J. Keselman, Brit. J. Math. Stat. Psy., 45 (1992) 265.

29 T. Hancock, R. Put, D. Coomans, Y. V. Heyden and Y. Everingham, J. Chem. Inf. Comput. Sci., 76 (2005) 185.

30 S. P. Niculescu, J. Mol. Struct., 622 (2003) 71.

31 D. E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning, Addison Wesley Professional, 1989.

32 J. Zupan and J. Gasteiger, Neural Networks in Chemistry and Drug Design, Wiley-VCH, 1999.

33 C. Ochoa, A. Chana and M. Stud, Curr. Med. Chem. – Central Nervous System Agents, 1 (2001) 247.

Received: January 24, 2007
Accepted: April 24, 2007
OnlineFirst: July 12, 2007

DOI: 10.1007/s10973-007-8354-7