Journal of the Serbian Chemical Society
JSCS-info@shd.org.rs • www.shd.org.rs/JSCS
J. Serb. Chem. Soc. 79 (9) 1111–1125 (2014)
JSCS–4651
UDC 547.546:615.9+532.74:
512.763+519.233.5
Original scientific paper

QSAR studies for assessing the acute toxicity of nitrobenzenes to *Tetrahymena pyriformis*

DAN-DAN WANG[1], LIN-LIN FENG[1], GUANG-YU HE[2] and HAI-QUN CHEN[1]*

[1]*School of Environmental and Safety Engineering, Changzhou University, Jiangsu Province, Changzhou, 213164, China and* [2]*Key Laboratory of Advanced Catalytic Materials and Technology, Changzhou University, Jiangsu Province, Changzhou, 213164, China*

(Received 10 September, 2013, revised 11 November, accepted 6 December 2013)

Abstract: Quantitative structure–activity relationship (QSAR) models play a key role in finding the relationship between molecular structures and the toxicity of nitrobenzenes to *Tetrahymena pyriformis*. In this work, a genetic algorithm along with partial least square (GA–PLS) was employed to select the optimal subset of descriptors that significantly contribute to the toxicity of nitrobenzenes to *T. pyriformis*. A set of five descriptors, namely G2, HOMT, G(Cl···Cl), Mor03v and MAXDP, was employed for the prediction of the toxicity of 45 nitrobenzene derivatives and then they were used to build the model by the multiple linear regression (MLR) method. It transpired that the built model, the stability of which was confirmed using the leave-one-out validation and external validation test, showed high statistical significance (R^2 = 0.963, Q^2_{LOO} = 0.944). Moreover, the *y*-scrambling test indicated there were no chance correlations in the model.

Keywords: quantitative structure–activity relationship; multiple linear regressions.

INTRODUCTION

Nitrobenzenes are important fine organic intermediates that are widely used in many fields, such as the synthesis of pharmaceuticals, dyestuffs and explosives.[1,2] Since most of nitrobenzenes and their derivatives are hazardous and have high potential to pollute the environment, it is of great significance to study their acute toxicity. With the rapid development of industry and agriculture, a growing number of nitrobenzenes leak into the environment, especially in aquatic ecosystems.[3] There is little possibility of testing the acute toxicity of each compound, since it would be time-consuming and expensive. Hence, great attention is being paid to finding tools capable of assessing the acute toxicity of nitroben-

* Corresponding author. E-mail: hqchenyf@hotmail.com
doi: 10.2298/JSC130910025W

zenes, among which the quantitative structure–activity relationship (QSAR) method is the most powerful one.

The quantitative structure–activity relationship (QSAR) method focuses on the motto that the properties of chemical compounds are determined by their molecular structures.[4] Thus, based on accurate experimental data of only some of the chemicals in one group, the activities of chemicals in the whole group can be predicted using suitable models, including compounds that have not yet been experimentally synthesized.[5–9]

For many years, QSARs have been efficiently used in the study of toxicity mechanisms of various reactive chemicals. Dearden *et al*. reviewed the attempts to model the acute toxicity of nitrobenzenes.[10] Due to the reactive electrophilic nature of nitrobenzenes, the global electrophilicity and local maximum philic power ($\omega^{+}_{C_{max}}$ / $\omega^{-}_{C_{max}}$), along with the total Hartree–Fock energy (E_{HF}), were used as independent variables, while the toxicity of 174 selected aromatic compounds to *T. pyriformis* were considered as the dependent variable.[11] Roy and Ghosh introduced extended topochemical atom (ETA) indices to model the toxicity of nitrobenzene derivatives to *T. pyriformis*.[12] Furthermore, quantum chemical methods were used to calculate molecular descriptors by Gu *et al*., making it easier to understand the built models.[13,14] Estrada and Uriarte obtained a good quantitative structure–toxicity model of 42 nitrobenzenes by their original topological sub-structural molecular design (TOPS–MODE) approach.[15] Artemenko *et al*. studied the toxicity of 95 diverse nitroaromatics to *T. pyriformis* and discussed possible modes of action by hierarchical technology for quantitative structure–activity relationships (HiT QSAR).[16] In some of the above-mentioned models, external validation and proof of passing the *y*-scrambling test were absent, causing serious doubts about the reliability of the interpretation and making it hard to assess their predictive power, which is a very important part of QSAR studies.

The selection of the molecular descriptors most relevant to acute toxicity is the key problem involved in the QSAR method, as well as the application of appropriate techniques for constructing the models. At present, the genetic algorithm (GA) is well known as an interesting and more widely used variable selection method.[17–19] GA is a stochastic method to solve the optimization problems defined by fitness criteria, applying the evolution hypothesis of Darwin and different genetic functions, *i.e.*, crossover and mutation.[20] Nowadays, many different techniques, such as multiple linear regression (MLR), partial least squares (PLS) and different types of artificial neural networks (ANN), have been widely used in building QSAR models.[21,22]

The aim of this study was to develop a reliable and predictive QSAR model using the MLR method for identifying the factors governing the acute toxicity of nitrobenzenes to *T. pyriformis* and to predict their acute toxicity from their mole-

cular structures. For this purpose, a group of 45 nitrobenzene compounds having the structure of a single nitrobenzene ring with different substituent groups, such as nitro-, halogens (fluorine, chlorine, bromine), was chosen as the sample set. The leave-one-out cross-validation, a *y*-scrambling test and outer samples prediction were performed to validate the developed model.

MATERIAL AND METHODS

Dataset

The QSAR modeling was applied on a set of nitrobenzenes (their molecular structures are given in Fig. S-1 of the Supplementary Material to this paper). The dataset used in this study was extracted from a single literature source.[23] It consists of 45 nitrobenzene compounds based on a nitrobenzene ring structure with different halogen substituents. Herein, –log IGC_{50} means the inverse logarithm of the concentration causing 50 % growth inhibition of *T. pyriformis*, which was used as a measure of the toxicity of the compounds. The experimental acute toxicity values of the nitrobenzenes to *T. pyriformis* are listed in Table I, as well as the corresponding names, molecular formulas and CAS numbers. The bigger the value of –log IGC_{50}, the higher is the acute toxicity of the compound.

TABLE I. List of the 45 compounds considered in the study, including corresponding names, molecular formulas, CAS numbers and –log IGC_{50} values

No.	Compound	Formula	CAS	–log (IGC_{50} / mmol mL^{-1})
1	1,3-Dinitrobenzene	$C_6H_4N_2O_4$	99-65-0	0.89
2	1-Bromo-4-nitrobenzene	$C_6H_4BrNO_2$	586-78-7	0.38
3	1,3,5-Trimethyl-2-nitrobenzene	$C_9H_{11}NO_2$	603-71-4	0.86
4	1-Methyl-2,4-dinitrobenzene	$C_7H_6N_2O_4$	121-14-2	0.87
5	1,2-Dichloro-3-nitrobenzene	$C_6H_3Cl_2NO_2$	3209-22-1	1.07
6	1,2-Dinitrobenzene	$C_6H_4N_2O_4$	528-29-0	1.25
7	1,4-Dinitrobenzene	$C_6H_4N_2O_4$	100-25-4	1.30
8	1,3-Dimethyl-2-nitrobenzene	$C_8H_9NO_2$	81-20-9	0.30
9	1,2-Dimethyl-3-nitrobenzene	$C_8H_9NO_2$	83-41-0	0.56
10	1,3,5-Trichloro-2-nitrobenzene	$C_6H_2Cl_3NO_2$	18708-70-8	1.43
11	1,2,3-Trichloro-4-nitrobenzene	$C_6H_2Cl_3NO_2$	17700-09-3	1.51
12	4-Chloro-1-methyl-2-nitrobenzene	$C_7H_6ClNO_2$	89-59-8	0.82
13	1,4-Dichloro-2-nitrobenzene	$C_6H_3Cl_2NO_2$	89-61-2	1.13
14	1-Chloro-2,4-dinitrobenzene	$C_6H_3ClN_2O_4$	97-00-7	1.98
15	1,2,3,4-Tetrachloro-5-nitrobenzene	$C_6HCl_4NO_2$	879-39-0	1.78
16	1-Methyl-4-nitrobenzene	$C_7H_7NO_2$	99-99-0	0.17
17	1,3,5-Trichloro-2,4-dinitrobenzene	$C_6HCl_3N_2O_4$	6284-83-9	2.19
18	1-Bromo-2,4-dinitrobenzene	$C_6H_3BrN_2O_4$	584-48-5	2.31
19	1,5-Dichloro-2,3-dinitrobenzene	$C_6H_2Cl_2N_2O_4$	28689-08-9	2.42
20	1,3-Dichloro-5-nitrobenzene	$C_6H_3Cl_2NO_2$	618-62-2	1.13
21	1-Fluoro-3-nitrobenzene	$C_6H_4FNO_2$	402-67-5	0.20
22	1-Fluoro-2-nitrobenzene	$C_6H_4FNO_2$	1493-27-2	0.23
23	1-Ethyl-4-nitrobenzene	$C_8H_9NO_2$	100-12-9	0.43
24	1,2-Dimethyl-4-nitrobenzene	$C_8H_9NO_2$	99-51-4	0.59
25	1-Chloro-2-nitrobenzene	$C_6H_4ClNO_2$	88-73-3	0.68
26	1-Chloro-2-fluoro-3-nitrobenzene	$C_6H_3ClFNO_2$	21397-07-9	0.80

TABLE I. Continued

No.	Compound	Formula	CAS	–log (IGC_{50} / mmol mL^{-1})
27	1-Chloro-3-nitrobenzene	$C_6H_4ClNO_2$	121-73-3	0.84
28	1-Bromo-3-nitrobenzene	$C_6H_4BrNO_2$	585-79-5	1.22
29	1,2,4,5-Tetrachloro-3-nitrobenzene	$C_6HCl_4NO_2$	117-18-0	1.47
30	1-Fluoro-2,4-dinitrobenzene	$C_6H_3FN_2O_4$	70-34-8	1.71
31	1,2,3,5-Tetrafluoro-4-nitrobenzene	$C_6HF_4NO_2$	314-41-0	1.87
32	1,5-Difluoro-2,4-dinitrobenzene	$C_6H_2F_2N_2O_4$	327-92-4	2.03
33	1,2,3,4,5-Pentafluoro-6-nitro-benzene	$C_6F_5NO_2$	880-78-4	2.43
34[a]	Nitrobenzene	$C_6H_5NO_2$	98-95-3	0.14
35[a]	1-Chloro-4-nitrobenzene	$C_6H_4ClNO_2$	100-00-5	0.43
36[a]	2,4-Dichloro-1-nitrobenzene	$C_6H_3Cl_2NO_2$	611-06-3	0.99
37[a]	1,2-Dichloro-4-nitrobenzene	$C_6H_3Cl_2NO_2$	99-54-7	1.16
38[a]	1,4-Dibromo-2-nitrobenzene	$C_6H_3Br_2NO_2$	3460-18-2	1.37
39[a]	1-Chloro-2-methyl-3-nitrobenzene	$C_7H_6ClNO_2$	83-42-1	0.68
40[a]	1,2,4-Trichloro-5-nitrobenzene	$C_6H_2Cl_3NO_2$	89-69-0	1.53
41[a]	1,2-Dichloro-4,5-dinitrobenzene	$C_6H_2Cl_2N_2O_4$	6306-39-4	2.21
42[a]	1,2,4,5-Tetrachloro-3,6-dinitro-benzene	$C_6Cl_4N_2O_4$	20098-38-8	2.74
43[a]	1-Fluoro-4-nitrobenzene	$C_6H_4FNO_2$	350-46-9	0.25
44[a]	1-Bromo-2-nitrobenzene	$C_6H_4BrNO_2$	577-19-5	0.86
45[a]	1,2,3-Trifluoro-4-nitrobenzene	$C_6H_2F_3NO_2$	771-69-7	1.89

[a]Test set

The compounds in Table I were sorted from low value to high value of their acute toxicity. Then, the first, the fifth, the ninth sample, *etc*. were chosen to create a test set, and the remaining 33 samples were regarded as the training set. The 33 training samples were utilized to construct the model. The other 12 samples were utilized to evaluate the predictive ability of the obtained model.

Molecular descriptor calculation and selection

Constructing numerical descriptors of a set of molecules is necessary for QSAR models. Descriptor reflects some of molecular properties, which can then be related with biological activity. For the compounds studied in this article, up to 1644 molecular descriptors were calculated using DRAGON software, which is a sophisticated program for the calculation of molecular descriptors.[24,25] To date, a wide variety of descriptors have been reported for QSAR analysis, such as topological descriptors, constitutional descriptors, geometric descriptors and charge related descriptors.[26,27] The geometries of all molecules were optimized by MM+ force field and then by the AM1 semi-empirical method with an *SCF* convergence of 10^{-5} and a *RMS* gradient of 10^{-2} kcal* Å^{-1}·mol^{-1}.[28] The DRAGON software users' guide can be referred to for a detailed description on the types of the molecular descriptors that DRAGON can calculate as well as the calculation procedures.[29]

The molecular descriptors that remained constant or near constant for all molecules were removed from the descriptor pool, since such descriptors could not encode the structural differences between the compounds, which account for the differences between their acute

* 1 kcal = 4184 J

toxicity. Further reduction of the descriptor pool was attained by examining pair-wise correlations between descriptors so that only one descriptor was retained from a pair contributing similar information (correlation coefficient > 0.95 in this study). Finally, a total set of 521 remaining descriptors was achieved and used to select the optimal subset of descriptors that significantly contribute to the acute toxicity.

The selection of molecular descriptors plays a significant role in QSAR analysis. With hundreds of descriptors remaining, a more powerful optimization method was required to find the optimum quantitative relationships between the molecular descriptors and the acute toxicity. The genetic algorithm (GA), which was developed to simulate processes observed in natural evolution, is a popular solution to solve this problem.[30] In this study, the GA, a powerful optimization method, along with the partial least square method (PLS), which is a robust statistical method for variable selection, was used to find the molecular descriptors closely related to acute toxicity. The GA–PLS programs were implemented using the software package PLS–Algorithm Toolbox written by Leardi and Lupiáñez.[31] A detailed description of how to use GA–PLS and the parameters required can be found in the literature.[31] In this work, all calculation programs implementing GA–PLS were written in M-file using the MATLAB package.[32]

MLR method

MLR is a widely-used statistical analysis method to model the relationship between a scalar dependent variable Y and several explanatory variables denoted X, which can build a simple and interpretable model.[33] In multiple linear regression, n compounds with a known dependent variable (acute toxicity) and independent variables (molecular structure descriptors) are used for building the model. It is assumed that the acute toxicity of each compound can be represented as:

$$y_i = b + b_1 x_{i,1} + b_2 x_{i,2} + ... + b_m x_{i,m} + \varepsilon_i \quad (1)$$

where y_i is the i^{th} acute toxicity (i = 1,2, …, n), $x_{i,k}$ is the value of k^{th} descriptor for compound i (k = 1,2,…m) and ε_i is the i^{th} residual and with b as the vector of the regression coefficients. In matrix notation:

$$\mathbf{y} = \mathbf{X}b + \boldsymbol{\varepsilon} \quad (2)$$

where $\mathbf{y}$ is the vector of the toxicity values for different compounds, $\mathbf{X}$ is the matrix of descriptors for different compounds and $\boldsymbol{\varepsilon}$ is the vector of the residuals.

When the matrix $\mathbf{X}^{\text{T}}\mathbf{X}$ is non-singular, the least square estimate of b is therefore obtained as:

$$b = (\mathbf{X}^{\text{T}}\mathbf{X})^{-1}\mathbf{X}^{\text{T}}\mathbf{y} \quad (3)$$

The estimated acute toxicity value can be calculated as:

$$\hat{y} = \mathbf{X}\hat{b} \quad (4)$$

In this work, the multiple linear regressions were performed using the statistics software SPSS.[34] The linear relationship between the acute toxicity data of the compounds and their structure parameters was fitted by the multiple stepwise regression method in 95 % confidence intervals. The qualities of the statistics of the MLR equation were judged by parameters such as the R^2 value (coefficient of determination), the F value (Fischer statistics) and the S value (standard deviation).

Model validation

In order to check the reliability and the stability of QSPR model elaborated by MLR method, both the internal and external validations were conducted. The goodness of the fitting was firstly characterized by the coefficient of determination (R^2) and the root mean squared error (*RMSE*) between calculated and experimental values for the molecules of the training set. The two formulas are given by Eqs. (5) and (6), respectively:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - y_i')^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{5}$$

and:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i' - y_i)^2}{N}} \tag{6}$$

where y_i, y_i' and $\bar{y}$ are the observed property, calculated property and mean value of the property, respectively, and N is the number of observations.

Cross-validation is one of the most popular methods of estimating the robustness of a model. In this work, the internal predictive capability of the model was evaluated by the leave-one-out cross-validation (Q^2_{LOO}), following the mathematic form:

$$Q^2_{LOO} = 1 - \frac{\sum_{i=1}^{\text{training set}}(y_i - y_i')^2}{\sum_{i=1}^{\text{training set}}(y_i - \bar{y})^2}. \tag{7}$$

A good Q^2_{LOO} often indicates a good robustness and high internal predictive power of a QSPR model. The cross-validation coefficient between predicted and observed values of the test set Q^2_{ext} was used to verify the external predictive ability of the MLR model, which can be calculated at the model development step by properly employing a prediction set for validation as follows:

$$Q^2_{\text{ext}} = 1 - \frac{\sum_{i=1}^{\text{test set}}(x_i - x_i')^2}{\sum_{i=1}^{\text{test set}}(x_i - \bar{y}_{\text{tr}})^2} \tag{8}$$

where x_i, x_i', and $\bar{y}_{\text{tr}}$ are the observed property, the calculated property in the test set and the mean value of the property in the training set, respectively.

In addition, the mean absolute error (*AAE*) was also used to assess the obtained model, which was calculated according to the following equation:

$$AAE = \frac{\sum_{i=1}^{n}\left|y_i' - y_i\right|}{N} \tag{9}$$

where y_i, and y_i' are the observed property and calculated property, respectively.

A *y*-scrambling test reveals the robustness of a QSAR model, being a measure of the model overfit. This test is realized by deliberately destroying the connection between the target variable *y* and the independent variables *x* (in QSAR: molecular descriptors). Thus, the *y*-data was randomly permuted, while all *x* data were left untouched. This *y*-scrambling was repeated 100 times. After this procedure, the obtained MLR model must have the minimal R^2 value.

RESULTS AND DISCUSSION

Results of descriptors selection

The GA–PLS procedure was performed to select the optimal set of descriptors. A set of five descriptors were finally selected and used to build the following model of MLR. The correlation matrix for these descriptors used in the present study is shown in Table II, from which it can be seen that no high correlations existed between these descriptors.

TABLE II. Correlation matrix between the selected descriptors

	Mor03v	*G*2	*HOMT*	*G*(Cl···Cl)	*MAXDP*
Mor03v	1	–0.094	0.074	–0.209	–0.124
*G*2		1	0.195	0.354	–0.019
HOMT			1	0.006	0.291
G(Cl···Cl)				1	–0.176
MAXDP					1

The physical meanings of these descriptors are interpreted as follows.[27] *Mor03v* (3D-MoRSE – signal 03/weighted by atomic van der Waals volumes) is a 3D-MoRSE descriptor based on the idea of obtaining information from the 3D atomic coordinates by the transform used in electron diffraction studies for preparing theoretical scattering curves. A generalized scattering function, called the molecular transform, can be used as a functional basis for deriving the specific analytic relationship of X-ray and electron diffraction from a known molecular structure.

G(Cl···Cl) is a 3D atom pair descriptor and one of the primary dimensional features of chemicals. It is mainly related to the sum of the geometrical distances between Cl···Cl. For a compound without a Cl atom, this value will be zero.

*G*2 is a gravitational index (bond restricted), defined as:

$$G2 = \sum_{b=1}^{B} \left(\frac{m_i m_j}{r_{ij}^2} \right)$$

where m_i and m_j are the atomic masses of the considered atoms, r_{ij} is the corresponding interatomic distance and *B* is the number of bonds in the molecule. The *G*2 index is restricted to pairs of bonded atoms and is related to the bulk cohesiveness of the molecules, accounting for both atomic masses and their dis-

tribution within the molecule space and can be extended to any atomic property other than the atomic mass, such as atomic polarizability, atomic van der Waals volume, *etc*. The *G*2 descriptor can be related to the size of the molecule – *G*2 is larger for large molecules.

Harmonic oscillator model of aromaticity index total (*HOMT*) is also a geometrical descriptor. It is based on the degree of alternation of single/double bonds and is used to measure the bond length deviation from the optimal length attributed to the typical aromatic state. In this work, it depended on the variation of the number, position, and nature of the substituents. To illustrate the effects of the substituents on *HOMT*, the title compounds were divided into 17 groups according to the type of ring. Table III presents the values of the *HOMT* indices for the 17 groups of differently substituted benzene rings, labeled in column 3.

TABLE III. *HOMT* values of the compounds with different types of ring substitution

Number of substituents	Type of ring	Class	Compound number	*HOMT*
1	Unsubstituted nitrobenzene	1	**34**	5.844
2	2-Substituted nitrobenzene	2	**6**	5.944
			22	5.959
			25	5.955
			44	5.955
	3-Substituted nitrobenzene	3	**1**	5.944
			21	5.959
			27	5.955
			28	5.955
	4-Substituted nitrobenzene	4	**2**	5.95
			7	5.944
			16	5.944
			23	5.944
			35	5.955
			43	5.961
3	2,4-Disubstituted nitrobenzene	5	**36**	5.96
	2,5-Disubstituted nitrobenzene	6	**4**	5.95
			12	5.955
			13	5.96
			14	5.955
			18	5.953
			30	5.959
			38	5.96
	3,4-Disubstituted nitrobenzene	7	**24**	5.953
			37	5.963
	3,5-Disubstituted nitrobenzene	8	**20**	5.96
	2,3-Disubstituted nitrobenzene	9	**26**	5.962
			39	5.955
	2,6-Disubstituted nitrobenzene	10	**8**	5.953

TABLE III. Continued

Number of substituents	Type of ring	Class	Compound number	*HOMT*
4	2,3,4-Trisubstituted nitrobenzene	11	**11**	5.964
			45	5.973
	2,4,6-Trisubstituted nitrobenzene	12	**3**	5.944
			10	5.961
			19	5.96
	2,4,5-Trisubstituted nitrobenzene	13	**32**	5.969
			40	5.963
			41	5.958
5	2,3,4,5-Tetrasubstituted nitrobenzene	14	**15**	5.97
			29	5.97
	2,3,4,6-Tetrasubstituted nitrobenzene	15	**31**	5.983
	2,3,5,6-Tetrasubstituted nitrobenzene	16	**17**	5.961
6	2,3,4,5,6-Pentasubstituted nitrobenzene	17	**33**	5.988
			42	5.97

Analyzing the data in Table III, several interesting results could be found. 1) Among the compounds with the same nitrobenzene ring, the halogen-substituted nitrobenzene compounds had higher *HOMT* indices than nitrobenzene compounds with alkyl groups; 2) when the H atoms were substituted by F atoms, the *HOMT* index was higher than for compounds with Cl and Br atoms; 3) for the isomers, the variance in the *HOMT* values was subtle.

The relation between the median *HOMT* value of each group and the number of substituents is illustrated in Fig. 1, from which it could be seen that the *HOMT* index, in general, incrementally increases with the number of substituents attached to the ring. As is well known, halogen and nitro groups are electron-withdrawing groups that are responsible for a decrease in the electronic density of a benzene ring to which they are attached.

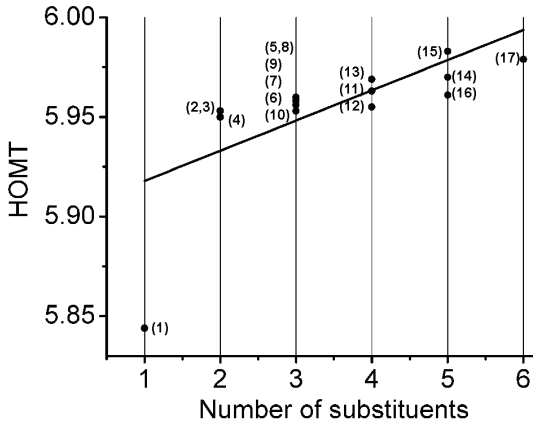

Fig. 1. Plot of the median values of *HOMT* against the number of substituents.

MAXDP is a topological descriptor defined as the maximal electrotopological positive variation, which can be related to the electrophilicity of a molecule.[35] The nitrobenzenes are representatives of electrophilic toxicants in that, depending on the substitution pattern, they may undergo a number of different electrophilic reactions.[36] Due to the reactive electrophilic nature of the nitrobenzenes, it is not surprising that previous modeling efforts focused on the use of electronic molecular descriptors.[37–40] The mechanisms of toxic action have been simplistic in that toxicity was modeled as a function of the ability of the toxicant to reach the active site and its ability to react covalently with some biological macromolecule.[41] Previous studies proposed a number of mechanisms of toxic action.[10,42–44] Despite the lack of knowledge regarding specific mechanisms of toxic action for some compounds, it was recognized that, while it is not easy to qualify, electrophilicity is an important property governing the toxicity of these compounds.

In general, the descriptors that appear in the QSAR model can encode different electronic, steric and electrophilic aspects of the molecules, which affect the acute toxicity of the compound.

Multiple linear regressions were found in SPSS in Analyze/Regression/Linear. The method for the multiple linear regression analysis in this study was “Stepwise”, which is an automated procedure used to select the most statistically significant variables from several explanatory variables. In this study, the experimental acute toxicity values of the 33 compounds in the training set as dependent variables and the *G2*, *HOMT*, *G*(Cl···Cl), *Mor03v* and *MAXDP* as independent variables had to be entered into the multiple linear regression model. The types and definitions of these descriptors are listed in Table IV.

TABLE IV. Molecular descriptors selected by GA–PLS

Descriptor	Independent variable	Type
Mor03v	x_1	3D-MoRSE descriptors
G(Cl···Cl)	x_2	3D Atom pairs
G2	x_3	Geometrical descriptors
HOMT	x_4	Geometrical descriptors
MAXDP	x_5	Topological descriptors

The MLR model built by stepwise regression[45] on the training set is given as Eq. (10):

$$-\log IGC_{50} = -120.042 - 0.538x_1 - 0.026x_2 + \\ +19.745x_3 + 0.462x_4 - 0.216x_5 \\ N = 33,\ R^2 = 0.963,\ F = 140.273,\ S = 0.142 \quad (10)$$

The model was assessed with the R^2 value (coefficient of determination), the *F* value (Fischer statistics), and the *S* value (standard deviation). The number of

observations N was also noted. Generally, the higher the correlation coefficient and the lower the standard error, the more reliable is the model. High values of F indicate the significance of Eq. (10), which reflects the ratio of variance explained by the model and the variance due to the error in the model.

Based on Eq. (10), the independent variables x_3 and x_4 were positively correlated with the dependent variable acute toxicity, while the independent variable x_1, x_2 and x_5 were in negative correlations with acute toxicity. As discussed in the section above, this also indicates positive contributions of molecular bulk (size), halogen and additional nitro substitutions in the nitrobenzene ring and negative contributions of –Cl group and molecular electrophilicity to the toxicity.

The relative influence of the various parameters on the targeted value were determined by their standardized regression coefficients in the equation, which were 0.210, –0.282, 0.842, 0.304 and –0.168, respectively. According to these values, the importance of the descriptors involved in the model decreased in the following order: *G*2 > *HOMT* > *G*(Cl···Cl) > *Mor03v* > *MAXDP*. The most significant descriptor is the mass distribution *G*2. The second significant descriptor is *HOMT*.

Model validation

In order to check the reliability and the stability of the QSAR model elaborated by the MLR method, both internal and external validations were conducted. The leave-one out cross-validation correlation coefficient (Q^2_{LOO}) was 0.944, showing the good robustness of the model. Moreover, predictions realized on the test set were in good agreement with the experimental values ($R^2_{\mathrm{ext}} = 0.927$, $Q^2_{\mathrm{ext}} = 0.918$, $RMSE_{\mathrm{ext}} = 0.220$). The value of the cross-validation correlation coefficient (Q^2_{LOO}) was similar to R^2 ($R^2 - Q^2_{\mathrm{LOO}} = 0.019$), which disclosed that the linear modeling method had a good generalization performance.

The dependences between the predicted acute toxicity values *vs*. the observed values for both the training and test sets are shown in Fig. 2, which shows good correlations between the parameters. In addition, the residuals of the predicted values of the acute toxicity against the observed values for the model are shown in Fig. 3. As most of the calculated residuals were distributed on both sides of the zero line, the conclusion could be drawn that there was no systematic error in the development of the developed model.

Moreover, the obtained model was tested for chance correlations by the *y*-scrambling experiment. In this work, the *y*-scrambling was conducted using the "MLR Y-Randomization Test 1.0" java program.[46] This *y*-scrambling was repeated 100 times. Every run yielded estimates of R^2 and Q^2, which are presented in Table S-I of the Supplementary Material to this paper. The obtained mean value of R^2 and Q^2 after a 100-time scrambling of the data set and modeling were

0.1625 (R^2 < 0.3) and –0.2871 (Q^2 < 0.0), respectively. It could thus be concluded that chance correlation had little or no effect in the presented model.

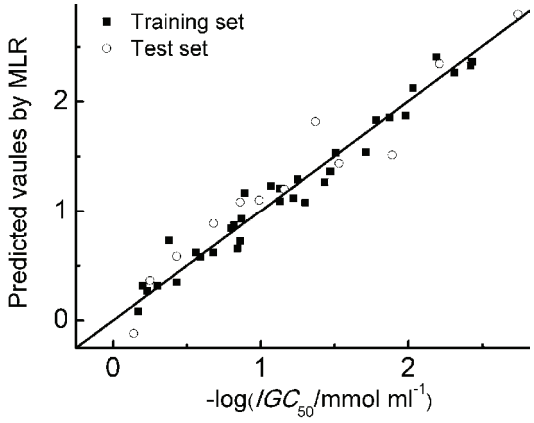

Fig. 2. Comparison between the predicted and the experimental –log IGC_{50} values.

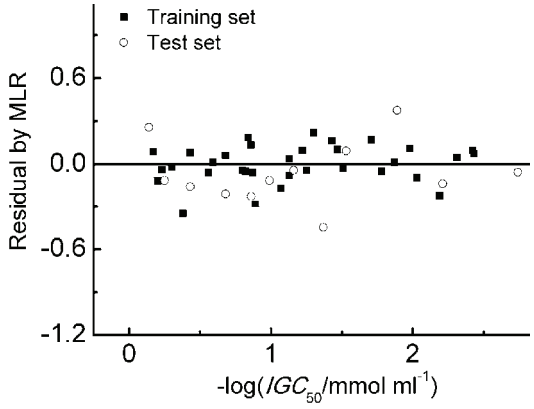

Fig. 3. Plots of the residuals *vs*. the experiment –log IGC_{50} values for the MLR model.

All the results discussed above showed that the presented MLR model could be effectively used to predict the acute toxicity of nitrobenzenes to *T. pyriformis*.

Model comparison

In order to estimate acute toxicity of nitrobenzenes to the *T. pyriformis*, several important relationships were previously proposed, which are reported in Table V.

As seen from Table V, these models were mainly based on three different typologies of descriptors: experimental parameters, quantum-chemical descriptors and molecular structure descriptors. Unfortunately, it is not possible to verify whether one model is better than another is; at most, it is possible to discuss the quality and the drawbacks of each model, because the size and composition of the training sets were usually different. Furthermore, the literature often reports

only the fitting power of a model expressed by R^2, while the predictive power is unknown. A *y*-scrambling test was not always performed. However, the MLR model presented in this paper is derived only from knowledge of molecular structure and it shows good predictive ability and strong robustness.

TABLE V. Comparison of some models for the prediction of nitrobenzenes' acute toxicity to *T. pyriformis*

Literature	*N*	Model descriptors	Q^2 [a]	R^2 [a]
12	42	$[\eta'_F]_{Cl}$, $[\eta'_F]_{Br/I}$, $[\eta'_F]_{NO_2}$, $[\eta'_F]_{CH_2OH}$	0.88	0.920
15	42	μ_0, μ_1, μ_2, μ_0	0.901	0.910
9	42	$\log K_{OW}$, $\log K_{OW}$, E_{LUMO}	0.866	0.881
13	36	E_{HOMO}, E_{LUMO}, ΔE, P, μ, V, Q_{-NO2}	0.874	0.896
14	20	E_{LUMO}	–	0.889
47	97	$E_{en}^{min}(C-C)$, $^2\chi$, E^{SOMO}, $\bar{V}_0$, $FNSA_{PNSA}^{(2)}$	–	0.815
10	47	$\log D$, E_{LUMO}, $\lvert dC_{ox}\rvert$	0.826	0.858
36	50	ω, $\log E_{LUMO}$, $\log P$	–	0.870

[a]The model with best predictive ability in the literature

CONCLUSIONS

In this study, GA–PLS was used to search for molecular descriptors closely related to the acute toxicity of nitrobenzenes for *T. pyriformis*. A set of five descriptors were finally selected and used to build a model by MLR. The most significant descriptor was the *G*2 molecular descriptor. This descriptor is a geometrical descriptor related to the size of the molecule. Moreover, the number of substituents on the aromatic ring of the nitrobenzenes, connected with the geometrical descriptor *HOMT*, also played a key role in the acute toxicity. Furthermore, internal and external validations were conducted to check the reliability and the stability of the QSAR model elaborated by the MLR method. The results showed the established model had a good predictive ability and strong robustness, with $R^2 = 0.963$, $Q^2_{LOO} = 0.944$ and $Q^2_{ext} = 0.918$. Thus, the presented model could be efficiently employed for estimating the toxicity of nitrobenzene derivatives for which experimental data are unavailable.

SUPPLEMENTARY MATERIAL

The molecular structures of the 45 titled compounds (Fig. S-1) and the results of the *y*-randomization validation (Table S-I) are available electronically from http://www.shd.org.rs/JSCS/, or from the corresponding author on request.

Acknowledgements. Financial support from the National Natural Science Foundation of China (Nos. 51202020 and 51472035), the International S & T Cooperation Program of Changzhou City (CZ20110022), the Science and Technology Department of Jiangsu Province (BY2013024-04, BE2014089) and the Qing Lan Project of Jiangsu Province are gratefully acknowledged.

ИЗВОД

QSAR СТУДИЈЕ АКУТНЕ ТОКСИЧНОСТИ НИТРОБЕНЗЕНА ЗА ПРАЖИВОТИЊУ *Tetrahymena pyriformis*

DAN-DAN WANG[1], LIN-LIN FENG[1], GUANG-YU HE[2] и HAI-QUN CHEN[1]

[1]*School of Environmental and Safety Engineering, Changzhou University, Jiangsu Province, Changzhou, 213164, China и* [2]*Key Laboratory of Advanced Catalytic Materials and Technology, Changzhou University, Jiangsu Province, Changzhou, 213164, China*

Квантитативни модели за релације између структуре и активности (QSAR) имају важну улогу у проучавању структурне зависности токсичности нитробензена за праживотињу *Tetrahymena pyriformis*. У овом раду је примењен генетички алгоритам, заједно са методом парцијалних најмањих квадрата. Токсичност 45 деривата нитробензена описан је помоћу пет дескриптора, наиме *G2*, *HOMT*, *G*(Cl···Cl), *Mor03v* и *MAXDP*. Показало се да је добивени модел статистички значајан (R^2 = 0,963, Q^2_{LOO} = 0,944). Осим тога, одговарајућим статистичком провером (*y-scrambling test*) показали смо да у моделу нема случајне корелације.

(Примљено 10. септембра, ревидирано 11. новембра, прихваћено 6. децембра 2013)

REFERENCES

1. S. Ikeno, C. Ogino, T. Ito, N. Shimizu, *Biochem. Eng. J.* **15** (2003) 193
2. H. Feuer, A. T. Nielsen, *Nitro Compounds: Recent Advances in Synthesis and Chemistry*, VCH Publishing, New York, 1990, p. 35
3. Y. H. Zhao, X. Yuan, G. D. Ji, L. X. Sheng, L. S. Wang, *Chemosphere* **34** (1997) 1837
4. D. J. W. Blum, R. E. Speece, *Ecotoxicol. Environ. Safety* **22** (1991) 198
5. F. R. Burden, D. A. Winkler, *Chem. Res. Toxicol.* **13** (2000) 436
6. E. Estrada, *SAR QSAR Environ. Res.* **11** (2000) 55
7. D. Ivan, L. Crisan, S. Funar-Timofei, M. Mracec, *J. Serb. Chem. Soc.* **78** (2013) 495
8. M. H. Fatemi, H. Malekzadeh, *Bull. Chem. Soc. Jpn.* **83** (2010) 233
9. M. T. D. Cronin, B. W. Gregory, T. W. Schultz, *Chem. Res. Toxicol.* **11** (1998) 902
10. J. C. Dearden, M. T. D. Cronin, T. W. Schultz, D. T. Lin, *Quant. Struct.-Act. Relat.* **14** (1995) 427
11. D. R. Roy, R. Parthasarathi, V. Subramanian, P. K. Chattaraj, *QSAR Comb. Sci.* **25** (2006) 114
12. K. Roy, G. Ghosh, *QSAR Comb. Sci.* **23** (2004) 99
13. Y. L. Gu, J. Q. Tao, Z. H. Fei, G. C. Zhang, *Chin. J. Struct. Chem.* **29** (2010) 86
14. X. F. Yan, H. M. Xiao, *Chin. J. Struct. Chem.* **26** (2007) 7
15. E. Estrada, E. Uriarte, *SAR QSAR Environ. Res.* **12** (2001) 309
16. A. G. Artemenko, E. N. Muratov, V. E. Kuz'min, N. N. Muratov, E. V. Varlamova, A. V. Kuz'mina, L. G. Gorb, A. Golius, F. C. Hill, J. Leszczynski, A. Tropsha, *SAR QSAR Environ. Res.* **22** (2011) 575
17. U. Depczynski, V. J. Frost, K. Molt, *Anal. Chim. Acta* **420** (2000) 217
18. B. K. Alsberg, N. M. Geneste, R. D. King, *Chemom. Intell. Lab. Syst.* **54** (2000) 75
19. D. Jouanrimbaud, D. L. Massart, R. Leardi, O. E. de Noord, *Anal. Chem.* **67** (1995) 4295
20. G. V. Dijck, M. M. V. Hulle, *Chemom. Intell. Lab. Syst.* **107** (2011) 318
21. M. Goodarzi, M. P. Freitas, R. Jensen, *Chemom. Intell. Lab. Syst.* **98** (2009) 123
22. Q. Shen, J. H. Jiang, C. X. Jiao, G. L. Shen, R. Q. Yu, *Eur. J. Pharm. Sci.* **22** (2004) 145
23. M. P. Gonzalez, H. G. Diaz, M. A. Cabrera, R. M. Ruiz, *Bioorg. Med. Chem.* **12** (2004) 735

24. V. Consonni, R. Todeschini, A. Mauri, M. Pavan, *Dragon software: Calculation of molecular descriptors*, Department of Environmental Sciences, University of Milano-Bicocca and Talete, srl, Milan, Italy, 2003, http://www.vcclab.org/lab/edragon/ (accessed in Sep. 2014)
25. I. V. Tetko, J. Gasteiger, R. Todeschini, A. Mauri, D. Livingstone, P. Ertl, V. A. Palyulin, E. V. Radchenko, N. S. Zefirov, A. S. Makarenko, V. Y. Tanchuk, V. V. Prokopenko. *J. Comput.-Aided Mol. Des.* **19** (2005) 453
26. M. Karelson, V. S. Lobanov, A. R. Katritzky, *Chem. Rev*. **96** (1996) 1027
27. V. Consonni, R. Todeschini, *Methods and Principles in Medicinal Chemistry*, in *Handbook of Molecular Descriptors*. Vol. 11, R. Mannhold, H. Kubinyi, H. Timmerman, Eds., Wiley–VCH, Weinheim, 2000
28. Hypercube HyperChem, Inc., http://www.hyper.com (accessed in Sep, 2014)
29. V. Consonni, R. Todeschini, M. Pavan, *DRAGON. Software for the calculation of and web version 2.1 molecular descriptors*, 2002, http://michem.disat.unimib.it/chm/ (accessed in Sep, 2014)
30. J. H. Holland, *Adaptation in Natural and Artificial Systems*, University of Michigan Press, Ann Arbor, MI, 1975
31. R. Leardi, A. Lupiáñez, *Chemom. Intell. Lab. Syst.* **41** (1998) 195
32. MATLAB 7.4. http://www.mathworks.com/products/matlab/ (accessed in Sep, 2014)
33. M. Mahani, H. S. Ghomi, *Anal. Methods* **4** (2012) 3381
34. SPSS for Windows, Rel. 10.0.0, SPSS Inc., Chicago, IL, 1999
35. P. Gramatica, M. Corradi, V. Consonni, *Chemosphere* **41** (2000) 763
36. K. Bellifa, S. M. Mekelleche, *Arab. J. Chem.* http://dx.doi.org/10.1016/j.arabjc.2012.04.031
37. J. W. Deneer, T. L. Sinnige, W. Seinen, J. L. M. Hermens, *Aquat. Toxicol.* **10** (1987) 115
38. G. D. Veith, O. G. Mekenyan, *Quant. Struct.-Act. Relat.* **12** (1993) 349
39. X. Yuan, G. Lu, P. Lang, *Bull. Environ. Contam. Toxicol.* **58** (1997) 123
40. P. Z. Lang, X. F. Ma, G. H. Lu, Y. Wang, Y. Bian, *Chemosphere* **32** (1996) 1547
41. O. G. Mekenyan, G. D. Veith, *SAR QSAR Environ. Res.* **2** (1994) 129
42. T. W. Schultz, M. T. D. Cronin, *Environ. Toxicol. Chem.* **16** (1997) 357
43. D. W. Roberts, *Chem. Res. Toxicol.* **8** (1995) 545
44. G. Dupuis, C. Benezra, *Allergic Contact Dermatitis to Simple Chemicals*, Marcel Dekker, New York, 1982
45. D. L. Massart, B. G. M. Vandeginste, L. M. G. Buydens, S. Dejong, P. J. Lewi, J. Smeyers-Verbeke, *Handbook of Chemometrics and Qualimetrics: Part A*, Elsevier, Amsterdam, 1977
46. MLR Y-Randomization Test 1.0 java program, http://dtclab.webs.com/ software-tools (accessed in Sep, 2014)
47. A. R. Katritzky, P. Oliferenko, A. Oliferenko, A. Lomaka, M. Karelson, *J. Phys. Org. Chem*. **16** (2003) 811.