Monatsh Chem (2011) 142:1111–1119
DOI 10.1007/s00706-011-0528-0

ORIGINAL PAPER

In silico cytotoxicity estimation of ionic liquids based on their two- and three-dimensional structural descriptors

Mohammad H. Fatemi · Parisa Izadiyan

Received: 23 November 2010 / Accepted: 3 May 2011 / Published online: 31 May 2011

Abstract The cytotoxicity of a series of ionic liquids containing ammonium, pyrrolidinium, imidazolium, pyridinium, and piperidinium cations against leukemia rat cell line IPC-81 was estimated from their structural parameters using quantitative structure–activity relationship methodology. Linear and nonlinear models were developed using genetic algorithm multiple linear regression and multilayer perceptron neural network approaches. Robustness and reliability of the constructed models were evaluated by internal, external, and Y-randomization procedures. Furthermore, the chemical applicability domain was determined via a leverage approach for each model. The results of this study revealed that the contribution of structural characteristics of the anionic parts of the studied ILs were fewer than of the cationic parts.

Keywords Cytotoxicity · Multiple linear regression · Quantitative structure–activity relationship · Artificial neural network · Ionic liquids

Introduction

An ionic liquid (IL) is a salt with a melting temperature below the boiling point of water [1]. Typical ILs consist of an organic cation with delocalized charges and a small inorganic anion such as Cl, BF_4, or PF_6, which is weakly coordinated to the organic cation. Ionic liquids are not a new class of materials, but they do have a set of physical properties that has sparked increased attention in recent years. The expanding interest in ionic liquids refers to their ability to be used in diverse applications such as sensors, fuel cells, batteries, capacitors, ionogels, extractants, and solvents in analysis, synthesis, catalysis, and separation [2, 3]. Due to their salt-like structures, ionic liquids usually exhibit a negligible vapor pressure up to very high temperatures for which they are often known as "green solvents" [4]. In spite of all the advantages of ILs, it should be noted that continued development and further use of these compounds may lead to accidental discharge and contamination. Although ILs can lessen the risk of air pollution due to their low vapor pressure, they do have significant solubility [5] as well as high stability [6] in water. As a result, this is the most likely medium through which ILs will be released into the environment.

To date, some studies have investigated the toxicity of ILs on human or rat cell lines [7, 8], which have revealed that there are some ionic liquids with low to high hazard potential for humans and the environment. According to these results, the "greenness" of ionic liquids strongly depends on their substructures (e.g., head group or side chain of the cation). Therefore, having information about the biochemical activity of ionic liquids based on their structures before putting them into wide use will provide a good insight into their environmental effects. Although in vivo or in vitro risk assessments have nowadays significantly improved, these methods are very time consuming. Besides this fact, they are not cost effective and do not respond to the large numbers of different chemicals.

As a good alternative, quantitative structure–activity/property relationships (QSAR/QSPR) have been successfully established. These approaches provide information that is useful for molecular design and medicinal chemistry [9]. The QSAR/QSPR models are mathematical equations which relate chemical structure of compounds to a wide

M. H. Fatemi (✉) · P. Izadiyan
Chemometrics Laboratory, Faculty of Chemistry, University of Mazandaran, Babolsar, Iran
e-mail: mhfatemi@umz.ac.ir

variety of their physical, chemical, biological, and technological properties and activities. The main task of QSAR/QSPR is to obtain a reliable statistical model for the prediction of activities/properties of new chemical substances and analytical systems. These relationships also take an approach to the identification and isolation of the most important structural descriptors that affect physicochemical properties. Nowadays, QSAR/QSPR models are rapidly developing and have been widely used by chemists for predicting different chemical and physical properties of different types of molecules. In the case of toxicological studies, successful QSAR investigations can be found in the literature; for example, Nowaczyk and Modzelewska-Banachiewicz [10] described the activity of fungicide agents containing a quinazolinone ring using the QSAR approach. Their models displayed good fits with the experimental in vitro data, with correlation coefficients of 0.923 and 0.854 for the activity against yeast and filamentous fungi, respectively. In other work, Gosav et al. [11] estimated the toxicity of novel amphetamines using neural networks (NNs) and their constitutional characteristics. In the case of ionic liquids, there are many QSPR studies on the correlation or prediction of their physical properties, such as melting point [12], viscosity [13], conductivity [13], molar volumes [14], and density [14, 15], while less attention has been paid to their toxicity. Only recently, Torrecilla et al. [16] have estimated the toxicity of a series of ionic liquids using their empirical formulas (elemental composition) and molecular weights as descriptors. This attempt resulted in linear and nonlinear models which were analyzed by statistical parameters, analysis of residuals, and statistical dispersion tests. The successful development of such models will help to understand the relationship between ionic liquids structure and their toxicity. As a cellular test system, promyelocytic leukemia rat cell line IPC-81 has been frequently used in cytotoxicity assays of ionic liquids, with the reduction of the WST-1 dye as an indicator of cell viability [7]. In the present work, we propose new externally predictive quantitative structure–toxicity relationship (QSTR) models based on two- and three-dimensional (2D and 3D) structural descriptors for the prediction of cytotoxicity of ionic liquids to leukemia rat cell line IPC-81.

The main goal of this work is to discover the most important structural parameters affecting the cytotoxicity of ILs. In addition to good statistical quality, the important aspect of the proposed models is their development by taking into account the fundamental points required by the organization for economic cooperation and development (OECD) principles [17] for regulatory acceptability of QSARs. According to these rules, models must be examined in terms of their validation for predictivity (both by internal and external statistical validation). Furthermore, the possibility of verifying the chemical applicability domain via the leverage approaches of models and, when possible, the mechanistic interpretation of their descriptors must be investigated.

Results and discussion

Interpretation of descriptors

In this work, quantitative relationships between the cytotoxicity of ionic liquids and their structural descriptors were investigated by using linear and non-linear models. Initially, our QSAR modeling effort involved the use of multiple linear regressions. The calculated logEC_{50} values of training and test sets using a linear model are shown in Table 1. Table 2 gives the specifications of the model obtained together with the six descriptors that appeared in the model, which were: heavy atom count (HAC), Moran autocorrelation–lag 8/weighted by atomic Sanderson electronegativities (MATS8e), partial charge weighted topological electronic index ($PCWT^{E}$), 3D-MoRSE-signal 26/weighted by atomic masses (Mor26m), R matrix average row sum (RARS), and R maximal autocorrelation of lag 5/unweighted ($R5u_{+}$). All these descriptors except HAC refer to the cationic part of the ionic liquids. HAC is the heavy atom count in the anions. The negative coefficient associated with this descriptor in the model indicates that an increase of the heavy atom count in the anionic parts leads to a decrease in the logEC_{50} value. It should be noted that a low value of effective concentration (EC_{50}) means high toxicity of the ionic liquid. Therefore, it was concluded that the more heavy atoms are in the anion structure the more toxic is the ionic liquid. Five other descriptors can be divided into four groups: 3D-MoRSE, 2D-autocorrelation, GETAWAY (GEometry, Topology, and Atom-Weights AssemblY), and electronic molecular coding descriptors. MATS8e is one of the 2D-autocorrelation [18] descriptors. These descriptors correspond to 2D-autocorrelations between pairs of atoms in the molecule, and are defined in order to reflect the contribution of a considered atomic property to the experimental observations under investigation (cytotoxicity). The atomic properties (atomic weights) that can be adopted to differentiate the nature of atoms are the mass, polarizability, electronegativity, or the volume. These indices can be readily calculated, i.e. by summing products of the atomic weights of the terminal atoms of all the paths of a prescribed length. For the case of MATS8e, the path connecting a pair of atoms has length 8 and involves the atomic Sanderson electronegativities as weighting scheme

Table 1 Dataset and corresponding observed, MLR, and MLP NN calculated values of logEC_{50}

	Name	logEC_{50} (exp)	logEC_{50} (MLR)	Residual	logEC_{50} (MLP)	Residual
1	1-(Cyanomethyl)-1-methyl-piperidinium chloride	3.82	4.10	0.28	3.83	0.01
2	1-(3-Hydroxypropyl)-1-methylpiperidinium chloride	3.76	3.73	−0.03	3.76	0.00
3	1-(3-Methoxypropyl)-1-methylpiperidinium chloride	3.72	3.41	−0.31	3.70	−0.02
4[a]	1-(3-Hydroxypropyl)-1-methylpyrrolidinium chloride	3.56	3.41	−0.14	3.52	−0.04
5	1-Butyl-1-methylpyrrolidinium chloride	3.55	3.41	−0.14	3.56	0.01
6	Ethyl(2-methoxyethyl)-dimethylammonium chloride	3.53	3.53	0.01	3.55	0.02
7	1-(Ethoxymethyl)-1-methylpiperidinium chloride	3.52	3.36	−0.16	3.62	0.10
8	Butylethyldimethylammonium chloride	3.52	3.57	0.05	3.49	−0.03
9[b]	1-Ethyl-3-methyl-3*H*-imidazolium hydrogensulfate	3.31	3.31	0.01	3.36	0.06
10	1-Butylpyridinium bromide	3.24	2.93	−0.31	3.12	−0.12
11	1-Butyl-3-methylpyridinium chloride	3.12	2.40	−0.72	3.08	−0.04
12	1-Butyl-1-methylpyrrolidinium bromide	3.11	3.16	0.04	3.10	−0.01
13	1-Ethyl-3-methyl-3*H*-imidazolium chloride	3.03	3.45	0.43	3.03	0.00
14[b]	1-Butyl-2-methylpyridinium chloride	3.02	2.51	−0.51	2.95	−0.07
15	1-(Cyanomethyl)pyridinium chloride	2.98	3.61	0.63	2.95	−0.03
16	1-Butylpyridinium tetrafluoroborate	2.95	2.90	−0.05	2.96	0.01
17	1,2,3,4,5-Pentamethylimidazolium iodide	2.91	3.33	0.42	2.87	−0.04
18	1-Butyl-3-methyl-3*H*-imidazolium iodide	2.90	2.68	−0.22	2.91	0.01
19[a]	1-Ethyl-3-methyl-3*H*-imidazolium bis(trifluoromethylsulfonyl)amide	2.85	2.61	−0.24	2.87	0.01
20	Pyridinium chloride	2.83	2.75	−0.08	2.83	0.00
21	1-Butyl-3-methyl-3*H*-imidazolium chloride	2.80	2.64	−0.16	2.81	0.01
22	1-methyl-3-propyl-3*H*-imidazolium tetrafluoroborate	2.78	2.34	−0.44	2.71	−0.07
23	1-Butyl-3-methyl-3*H*-imidazolium bromide	2.77	2.66	−0.11	2.81	0.04
24[b]	1-Ethyl-3-methyl-3*H*-imidazolium tetrafluoroborate	2.73	3.21	0.48	2.62	−0.11
25	1-Butyl-3-methyl-3*H*-imidazolium-*O*-methylsulfate	2.61	2.35	−0.26	2.49	−0.12
26	1-Butyl-4-methylpyridinium chloride	2.59	2.39	−0.20	2.57	−0.02
27	1-Butyl-3-methyl-3*H*-imidazolium tetrafluoroborate	2.47	2.41	−0.06	2.60	0.13
28	1-Hexyl-3-methyl-3*H*-imidazolium tetrafluoroborate	2.39	1.83	−0.56	2.38	−0.01
29[b]	1-Benzyl-3-methyl-3*H*-imidazolium chloride	2.32	2.52	0.20	2.62	0.30
30	1,3-Diethyl-3*H*-imidazolium bromide	2.31	2.74	0.43	2.37	0.06
31	1-Butyl-3-methyl-3*H*-imidazolium bis(trifluoromethylsulfonyl)amide	2.31	2.50	0.20	2.30	−0.01
32	1-Methyl-3-pentyl-3*H*-imidazolium chloride	2.28	2.48	0.21	2.35	0.07
33	1-Butyl-1-methylpyrrolidinium tetrafluoroborate	2.26	2.59	0.33	2.25	−0.01
34[a]	1-Hexyl-1-methylpyrrolidinium chloride	2.24	2.31	0.07	3.00	0.76
35	1-Hexyl-3-methyl-3*H*-imidazolium chloride	2.13	2.05	−0.07	2.13	0.00
36	1-Hexyl-4-methylpyridinium chloride	2.00	1.77	−0.24	2.02	0.02
37	1-Methyl-1-octylpyrrolidinium chloride	1.96	1.46	−0.50	1.95	−0.01
38	1-Heptyl-3-methyl-3*H*-imidazolium chloride	1.87	1.75	−0.12	1.83	−0.04
39[b]	1-Methyl-3-octyl-3*H*-imidazolium chloride	1.38	1.25	−0.13	1.35	−0.03
40	1-Butyl-4-dimethylamino-pyridinium chloride	1.27	1.29	0.01	1.29	0.02
41	1-Methyl-3-octyl-3*H*-imidazolium tetrafluoroborate	1.04	1.07	0.03	1.04	0.00
42	3-Methyl-1-octylpyridinium chloride	0.85	1.06	0.21	0.86	0.01
43	1-Methyl-3-nonyl-3*H*-imidazolium chloride	0.79	0.91	0.13	0.80	0.01
44[a]	1-Butyl-2,3-dimethyl-3*H*-imidazolium tetrafluoroborate	0.75	1.81	1.06	1.90	1.16
45	4-Dimethylamino-1-hexylpyridinium chloride	0.32	0.75	0.43	0.32	0.00
46	Trihexyltetradecylphosphonium tetrafluoroborate	0.23	0.53	0.29	0.25	0.02

Table 1 continued

	Name	logEC_{50} (exp)	logEC_{50} (MLR)	Residual	logEC_{50} (MLP)	Residual
47	Benzyldecyldimethylammonium chloride	0.13	0.26	0.12	0.12	−0.01
48	Benzyldodecyldimethylammonium chloride	−0.18	−0.04	0.14	−0.15	0.03
49[a]	Benzyldimethyltetradecylammonium chloride	−0.27	−0.72	−0.44	0.32	0.59
50	1-Methyl-3-octadecyl-3*H*-imidazolium chloride	−0.42	−0.43	0.01	−0.44	−0.02

[a, b] The internal and external test sets, respectively

Table 2 Details of the constructed GA-MLR model

Descriptor name	Notation	Coefficient	Standard error	*t* value	*p* value
Heavy atom count	HAC	−0.06	±0.020	–2.954	0.006
Moran autocorrelation-lag 8/weighted by atomic Sanderson electronegativities	MATS8e	−1.161	±0.452	−2.572	0.015
Partial charge weighted topological electronic index	$PCET^E$	−0.049	±0.010	−4.953	0.000
3D-MoRSE- signal 26/weighted by atomic masses	Mor26m	2.685	±0.593	4.529	0.000
R matrix average row sum	RARS	9.984	±0.979	10.199	0.000
R maximal autocorrelation of lag 5/unweighted	$R5u_+$	−2.173	±1.095	−1.984	0.056
Constant	–	−3.02	±0.713	−4.238	0.000
$n = 40$ $R^2 = 0.935$ $F = 79.2$ SE $= 0.32$ $\boldsymbol{R}^2_{CV} = 0.87$					

to distinguish their nature. Considering the negative coefficient of this descriptor in the model, we may conclude that enhanced values of atomic electronegativities in cations are favorable for toxic effects of the ILs studied. RARS and $R5u_+$ are two GETAWAY descriptors [19]. Such descriptors have shown great potential as powerful variables in QSAR modeling of different biological activities because they encode information about molecular shape, size, and atom distribution [20]. These kinds of descriptors try to match 3D-molecular geometry with chemical information by using different atomic weightings. $R5u_+$ is calculated from R maximal autocorrelation and RARS is the average row sum of the influence/distance matrix which is defined in Eq. 1:

$$\mathrm{RARS} = \frac{1}{A} \cdot \sum_{i=1}^{A} \sum_{j=1}^{A} 1 \frac{\sqrt{h_{ii} h_{jj}}}{r_{ij}} = \frac{1}{A} \cdot \sum_{i=1}^{A} RS_i \qquad (1)$$

where h_{ii} and h_{jj} are the leverages of the two considered atoms, r_{ij} is their geometric distance, A is the number of atoms in the molecule and RS_i is the ith row sum. The row sums of the influence/distance matrix encode some useful information that could be related to the presence of significant substituents or fragments in the molecule. The largest coefficient in the linear model belongs to the RARS descriptor which probably implies the importance of cationic substituents and dimension effects on the cytotoxicity of ILs. The next descriptor is $PCWT^E$, which is discussed by Osmialowski et al. [21]. $PCWT^E$ is defined by Eq. 2:

$$\mathrm{PCWT^E} = \frac{1}{Q_{\min}} \sum_{i<j} \frac{|q_i - q_j|}{r_{ij}^2} \qquad (2)$$

where q_i and q_j are the Zefirov partial charges of the bonded atoms, Q_{min} is the most negative partial charge, and r_{ij} is the corresponding bond length.

The last descriptor is Mor26m, which is one of 3D-MoRSE type descriptors [22]. These types of descriptors as molecular transforming illustrate large capability features for the sign of molecular structures regarding their independent size of the molecule. These descriptors allocate the structural diversity as well as certain allowed alterations that can be represented in different atomic properties such as atomic number, mass, partial charge, and polarizability. Therefore, they can reflect the flexibility in explanation of the molecules for their biological activities. Mor26m represents 3D-MoRSE-signal 26/weighted by atomic masses. The appearance of the 3D-MoRSE signal over the 26th stage weighted by atomic masses might indicate the above-average value distribution on the spatial arrangements of substituents within the molecular environments. In this work, the importance of the atomic masses, which appeared in the 26th stage out of 32-dimensional space, might be taken into account for the ability of the atomic masses information that should be related to the alterations in the spatial

arrangements of the substitution patterns within the electron diffraction properties. The electron distribution regarding the atomic masses of substituents might then be a factor in influencing the cytotoxic ability of the studied ILs. The appearance of these descriptors in the model indicates that atomic masses, electronic properties, and the cation substituents are important structural characteristics which affect the cytotoxicity of ILs. It was demonstrated in this work that cytotoxicity of ILs is closely related to their chemical structure, especially to the special fragments on the cation skeleton. All the above-mentioned parameters could be used for future QSTR investigations about toxicity of ionic liquids.

Nonlinear model

As has been mentioned, the nonlinear model for cytotoxicity estimation of ILs was established by Levenberg–Marquardt multilayer perceptron neural network (MLP NN) analysis on the basis of descriptors selected by a genetic algorithm multiple linear regression (GA-MLR) approach. An improved nonlinear model was developed to predict the logEC_{50} values of the 50 ionic liquids in training, internal, and external test sets. Figure 1 represents the plot of experimental versus calculated logEC_{50} values using the MLP NN model. Inspection of this figure indicates good correlation between experimental and calculated cytotoxicity values.

The correlation coefficients (R^2) between experimental and calculated cytotoxicity values by this model for training, internal, and external test sets were 0.998, 0.954 and 0.917, respectively. The other statistics of the developed MLP NN model were an average error (AE) of

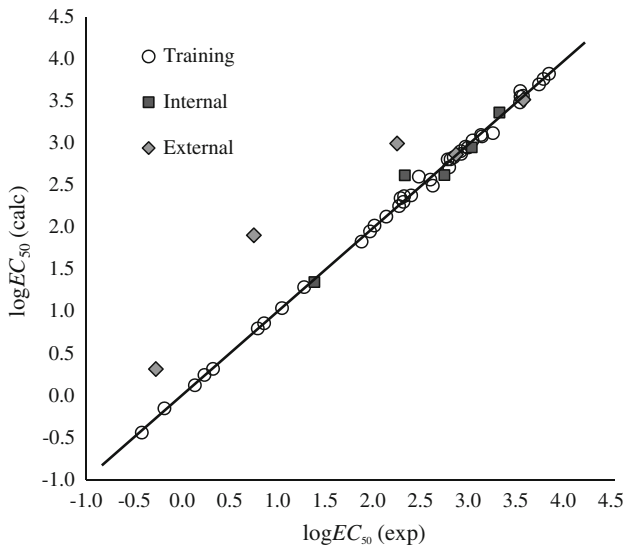

Fig. 1 Experimental versus calculated cytotoxicity values by the MLP NN model

Table 3 Comparative results of linear and nonlinear models

	Model training set		Test set	
	R^2	RMSE	R^2	RMSE
GA-MLR	0.935	0.291	0.862	0.441
MLP NN	0.998	0.045	0.954[a]	0.148[a]
	–	–	0.917[b]	0.672[b]

[a, b] The internal and external test sets, respectively

−0.0005 and average absolute error (AAE) of 0.0295 for the training set, AE = 0.0303 and AAE = 0.1126 for the internal test set, and AE = 0.4951 and AAE = 0.5101 for the external test set. The main parameters of both linear and nonlinear models are represented in Table 3.

Evidently, these results show improvement of the statistical parameters for the MLP NN model over the linear model, which confirms the nonlinear relationship between structural information and cytotoxicity of ionic liquids. To determine the order of importance of descriptors in the MLP NN model, a sensitivity analysis was performed. According to this method, the differences between the root-mean-square error (RMSE) of the complete network's prediction and the RMSE were obtained when the *i*th variable is excluded from the trained network ($RMSE_i$), and were shown as $Rmdiff_i$ (Eq. 3).

$$Rmdiff_i = RMSE_i - RMSE \quad (3)$$

It is obvious that the most important variable is the one that leads to the highest value of $Rmdiff_i$. The values of $Rmdiff_i$ for the MLP NN model were calculated and plotted in Fig. 2.

As it can be seen in this figure, the order of importance of selected molecular descriptors is RARS > MATS8e > $R5U_+$ > Mor26m > $PCWT^E$ > HAC. According to the sensitivity analysis results, among these six descriptors the MLP NN model has the least sensitivity to the HAC

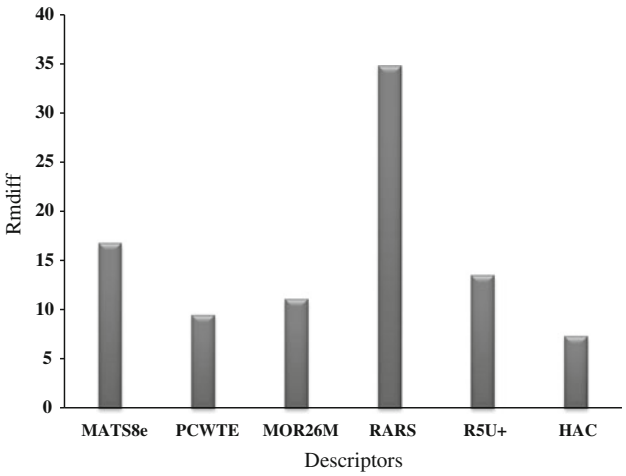

Fig. 2 Sensitivity analysis plot of the MLP NN model

descriptor. This is in agreement with the results of toxicological researches about a lesser contribution of anionic parts of studied ILs to their toxicities [23]. This result can serve as a theoretical and rational support for the experimental researches about toxicity of ionic liquids.

Model validation

In spite of good accuracy and apparent mechanistic appeal, QSAR models should pass rigorous validation tests to be useful as reliable screening tools. The Y-randomization test is a tool used in validation of QSAR models, whereby the performance of the original model in data description is compared to that of models built for permuted (randomly shuffled) response, based on the original descriptor pool and the original model building procedure. The Y-scrambling procedure [24] was performed to ensure that there is not any chance correlation in the data matrix. The mean value of R^2 after 30 times y-scrambling was 0.236, which disapproved the chance correlation probability. The real usefulness of QSTR models is not just their ability to reproduce known data, verified by their fitting power (R^2), but is mainly their possibility of predictive application. For this reason, internal validation, leave one out cross-validation (LOO), was applied on the MLR model which resulted in square cross-validated correlation coefficient $\left(R^2_{CV}\right)$ of 0.87 and RMSE = 0.291, which confirmed good predictive ability of this model. For a QSTR model, internal validation, although important and necessary, does not sufficiently guarantee the predictive ability of a model. Therefore, external validation on a representative number of chemicals must always supplement the internal validation, which will avoid an overoptimistic proposal. This was done through statistical validation of a separate external test set, which was not included in the model development procedure. The results of external validation for both MLR and MLP NN models (which are shown in Table 3) were acceptable and revealed reliability of both models.

Applicability domain

It needs to be emphasized that, no matter how robust, significant, and validated a QSTR model may be, it cannot be expected to reliably predict the modeled activity for the entire universe of chemicals. Therefore, before a QSTR model is put into use for screening chemicals, its domain of application (AD) must be defined [24]. A simple measure of a chemical being too far from the applicability domain of the model is its leverage h_i, which is defined as:

$$h_i = x_i^{\mathrm{T}}\left(X^{\mathrm{T}}X\right)^{-1}x_i \quad (i = 1, \ldots, n) \tag{4}$$

where x_i is the descriptor row-vector of the query compound and X is the $n \times k - 1$ matrix of k model descriptor values for n training set compounds. The superscript T refers to the transpose of the matrix/vector. The warning leverage h^* is, generally, fixed at $3k/n$, where k is the number of model parameters plus one and n is the number of training compounds.

To visualize the applicability domain of the GA-MLR and the MLP NN models, the standardized residuals versus leverage (Hat diagonal) values (William plot) were plotted for an immediate and simple graphical detection of both the response outliers (i.e., compounds with standardized residuals greater than three standard deviation units, $>3\sigma$) and structurally influential chemicals in the model ($h > h^*$). Figures 3 and 4 show the results for the AD analysis of the QSTR models, which were determined by training instances with h values lower than $h^* = 0.525$.

As can be seen from these figures, all predictions were reliable for MLP NN and linear models and there is no response outlier compound for either training or prediction sets, which further indicated the reliability of the

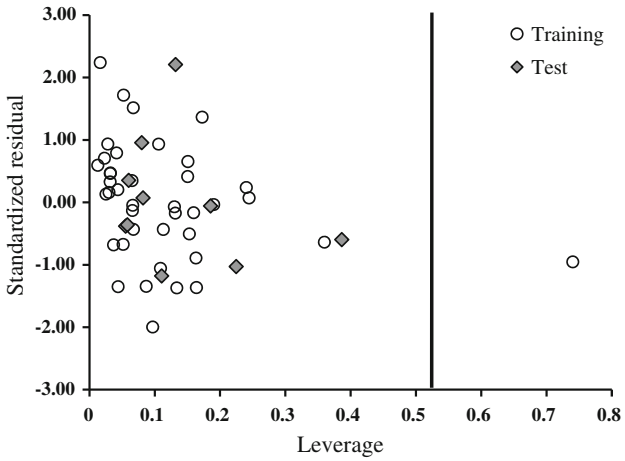

Fig. 3 Application domain plot for the GA-MLR model at $h^* = 0.53$

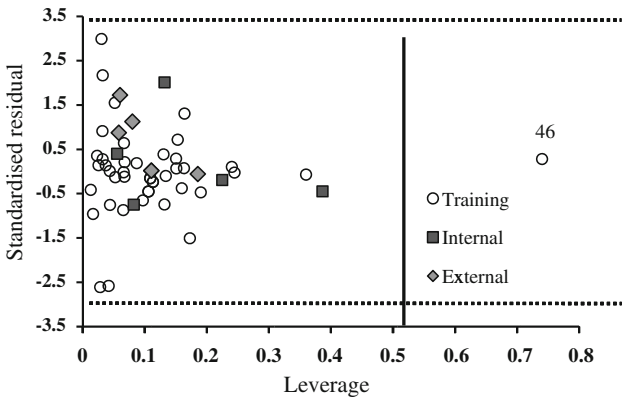

Fig. 4 Application domain plot for the MLP NN model at $h^* = 0.53$

predictions from another aspect. Comparison of the William plot of both proposed models reveals the similarity of chemical applicability domain of both developed models. Moreover, it can be seen that the only chemical influential on the structural domain of both models is number 46. The anomalous behavior of this chemical could be due to the following: (1) incorrect experimental input data, (2) the descriptors selected do not capture some relevant structural features present in this molecule and absent in the others, and (3) its biological mechanism is different from the remaining chemicals. Considering the cationic structure of compound number 46, it could be interpreted that reasons (2) and (3) might be reasonable for this compound. For future predictions, predicted cytotoxicity data must be considered reliable only for those chemicals that fall within the applicability domain on which the model was constructed. New samples with an h value higher than h^* and/or a value of standardized residual higher than $+2.77$ or lower than -2.77 (horizontal dashed lines in Fig. 3) are out of the AD bandwidth of the model and consequently cannot be reliably predicted. Conversely, when the leverage value of a compound is lower than the critical value, the probability of accordance between predicted and actual values is as high as that for the training set chemicals.

Concluding remarks

In this paper, linear and nonlinear QSAR models for cytotoxicity estimation of ionic liquids with high accuracy are presented based on their 2D and 3D structural descriptors**.** The nonlinear model produces better results than the linear model and comprises good predictability. Although the MLP NN model appears statistically more reliable than the GA-MLR model, it needs complex calculations. However, the linear model is simple, transparent, and general with a moderate external predictivity (standard error of 0.48). Therefore, except for conditions where high accuracy is required, the linear model is preferred. There are some other important points are listed. First, the proposed models in this work could identify and provide some insight into structural features which are related to the cytotoxicity of ILs. It was confirmed that structural features of anionic parts in the studied ILs have less effects on the cytotoxicity of these chemicals compared to cations. This result helps to get more useful information about exploring or synthesis of new ionic liquids. Second, the nonlinear relationship can describe accurately the relationship between the structural parameters and the cytotoxicities of the studied ILs. Third, GA-optimization is a good choice for reduction of descriptor numbers and elimination of nonrelevant descriptors and helps to statistically improve the model.

Methodology

Dataset

The structures of a diverse set of 50 ionic liquids as well as their corresponding cytotoxicity against leukemia rat cell line were taken from UFT/Merck ionic liquids biological effects database (Centre for Environmental Research and Sustainable Technology) [25]. The biological endpoint doses (EC_{50} in mg/dm^3) were transformed to the form of the logarithm of half-maximal effective concentration ($\log EC_{50}$). The IUPAC names of ionic liquids as well as the calculated and experimental cytotoxicity values are shown in Table 1.

The maximum value of $\log EC_{50}$ was 3.82 for 1-(cyanomethyl)-1-methylpiperidinium chloride and the minimum value was -0.42 for 1-methyl-3-octadecyl-1*H*-imidazolium chloride. Compounds in the dataset were sorted according to their cytotoxicity values and then on the basis of desired distances from each other. The dataset was divided into training, internal, and external test sets, including 40, 5, and 5, members (y-ranking procedure), respectively. For nonlinear modeling, the training set was used to adjust the model parameters, the internal test set was used to prevent the model from overfitting, and the external test set was used to evaluate the prediction power of developed model. In the case of MLR modeling, internal and external test sets were considered as the test set.

Structural descriptors

To obtain a QSTR model, compounds are represented by theoretical molecular descriptors. In order to compute the structural descriptors, the structures of all cations were drawn using ChemSketch software (v.12) [26], and were optimized by means of the molecular mechanics (MM+) force field of the HyperChem program (v.7) [27]. The final geometries of the minimum energy conformation were obtained by more precise optimization with the AM1 parameterization method by the MOPAC 6.0 package [28]. After geometry optimization, Hyperchem output files were used by the Dragon program [29] as input to calculate descriptors. Furthermore, the program CODESSA [30] was used to compute some additional constitutional, topological, electrostatic, and semi-empirical descriptors. Several descriptors (such as molecular weight, H-bond acceptor, topological polar surface area, heavy atom count, formal charge, etc.) were also considered to characterize the contribution of the anionic part of ILs. In order to reduce redundant and non-useful information, prescreening of descriptors was carried out in the following way: first, constant or near constant descriptors were eliminated, and then among those descriptors whose intercorrelations

exceeded 0.9, the most suitable and interpretable ones were kept while the others were deleted. The remaining 305 descriptors were entered to feature the screening step.

Variable selection

Variable selection is always one of the most important steps in developing a QSTR model, which is especially important when one is required to deal with a large or even overwhelming variable set. It is well known in both chemical and statistical fields that the accuracy of classification and regression techniques is not monotonic with respect to the number of features employed by the model. Therefore, depending on the nature of the regression technique, the presence of irrelevant or redundant features can cause the system to focus attention on the idiosyncrasies of the individual samples and lose sight of the broad picture that is essential for generalization beyond the training set.

Recently, two publications suggested that genetic algorithms (GA) might be useful in data analysis, especially in the task of reducing the number of features for regression models [31, 32]. A genetic algorithm is a powerful optimization method to search for the global optima of solutions. This algorithm is developed to mimic some processes observed in natural evolution. A detailed description of GA can be found in [33]. In the present work, a genetic algorithm along with the stepwise multiple linear regression (stepwise MLR) was applied to determine an optimal subset of variables. It is proper to mention that in this study the genetic algorithm is used as an optimization method to reduce the number of descriptors before application of stepwise MLR. The standard Holland genetic algorithm with elitism and roulette selection was performed using the STATISTICA (Release 7) software [34]. At the end of the GA process, 305 structural descriptors were reduced to the 58 relevant descriptors. In the next step of variable selection, stepwise MLR was applied on the remaining descriptors. In this step, to avoid overcorrelation of the regression equation, the variation of squared correlation coefficient (R^2) and standard error of estimate (SE) in equations by the addition of relevant descriptors to the model were monitored. As shown in Fig. 5, after the addition of six descriptors to the model no significant improvement in the developed model was observed. Therefore, the six parameter equation was selected as the best GA-MLR model.

In order to investigate how efficient GA-MLR is over stepwise MLR, a variable selection procedure was performed separately using just the stepwise multiple linear regression method on the 305 initial descriptors. This investigation resulted in a model which did not possess acceptable statistical parameters and was not practical.

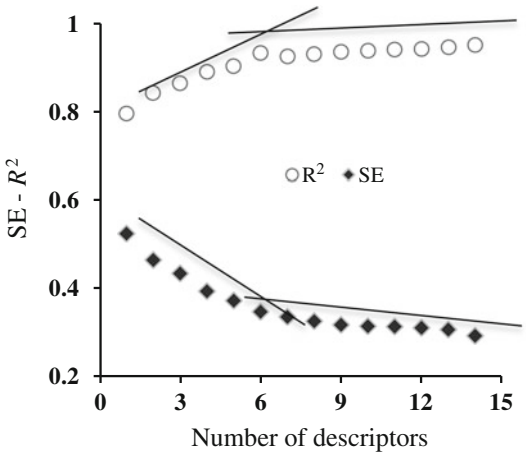

Fig. 5 Variation of R^2 and SE versus number of descriptors in the linear model

Details of the developed GA-MLR model are shown in Table 2. As can be seen, good overall quality of the model in significant terms is in fact indicated by the large F and small p values.

Nonlinear modeling

Artificial neural networks (ANNs) are biologically inspired computer programs designed to simulate the way in which the human brain processes information [35]. An ANN is formed from hundreds of single units, artificial neurons or processing elements (PE), connected by coefficients (weights), which constitute the neural structure and is organized in layers. The ability of ANNs to accurately model nonlinear relationships of input–output pairs of data is well established. In this work, in order to check any nonlinear relationships between structural descriptors and cytotoxicity values, the multilayer perceptron neural network (MLP) [36] was applied using STATISTICA software. A multilayer perceptron is a feed-forward artificial neural network model that maps sets of input data onto a set of appropriate output. It consists of multiple layers of nodes in a directed graph, which is fully connected from one layer to the next. Except for the input nodes, each node is a neuron with a nonlinear activation function.

Descriptors which were selected by the GA-MLR procedure were used as inputs of the network. The Levenberg–Marquardt (LM) algorithm [37] is one of the most efficient learning algorithms for neural networks. The advantages of using the LM algorithm are that specifying rate or momentum is not necessary and training processes are much more rapid. Therefore, in this study, the LM algorithm was used to develop a nonlinear model. To obtain better results, the parameters that influence the performance of the MLP NN were optimized. The optimized architecture of MLP

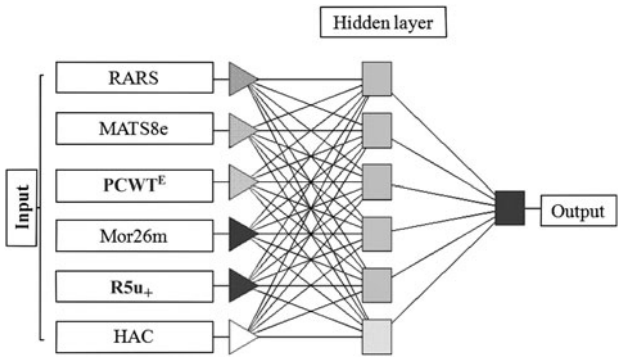

Fig. 6 The architecture of the MLP NN model

network was obtained as 6:6:1, which is shown in Fig. 6. The optimized and trained network was used to calculate the logEC_{50} values of training, internal, and external test sets, which are shown in Table 1.

References

1. Dommert F, Schmidt J, Krekeler C, Zhao YY, Berger R, Delle Site L, Holm C (2010) J Mol Liq 152:2
2. Bourbigou HO, Magna L, Morvan D (2010) Appl Catal A 373:1
3. Lukasik RB (2007) Monatsh Chem 138:1137
4. Earle MJ, Seddon KR (2000) Pure Appl Chem 72:1391
5. Thuy Pham TP, Cho CW, Yun YS (2010) Water Res 44:352
6. Romero A, Santos A, Tojo J, Rodriguez A (2008) J Hazard Mater 151:268
7. Rankea J, Muller A, Weber UB, Stock F, Stolte S, Arning J, Stormann R, Jastorff B (2007) Ecotoxicol Environ Saf 67:430
8. Kumar RA, Papaïconomou N, Lee JM, Salminen J, Clark DS, Prausnitz JM (2009) Environ Toxicol 24:388
9. Schultz TW, Cronin MTD, Netzev TI (2003) J Mol Struct (THEOCHEM) 622:23
10. Nowaczyk A, Banachiewicz BM (2010) Cent Eur J Chem 8:440
11. Gosav S, Praisler M, Dorohoi DO (2007) J Mol Struct 834–836:188
12. Yan C, Han M, Wan H, Guan G (2010) Fluid Phase Equilib 292:104
13. Tochigi K, Yamamoto H (2007) J Phys Chem C 111:15989
14. Jacquemin J, Ge R, Nancarrow P, Rooney DW, Gomes MFC, Pádua AAH, Hardacre C (2008) J Chem Eng Data 53:716
15. Lazzus JA (2009) J Taiwan Inst Chem Eng 40:213
16. Torrecilla JS, Garcia J, Rojo E, Rodriguez F (2009) J Hazard Mater 164:182
17. OECD principles for the Validation, for Regulatory Purpose, of (Q)SAR Models, http://www.oecd.org/document/4/0,3343,en_2649_34379_42926724_1_1_1_1,00.html
18. Moreau G, Broto P (1980) Nouv J Chim 4:359
19. Consonni V, Todeschini R, Pavan M (2002) J Chem Inf Comput Sci 42:682
20. Kušic H, Rasulev B, Leszczynsk D, Leszczynski J, Koprivana N (2009) Chemosphere 75:1128
21. Osmialowski K, Halkiewicz J, Kaliszan R (1986) J Chromatogr 361:63
22. Urra LS, Gonzalez MP, Teijeira M (2006) Bioorg Med Chem 14:7347
23. Docherty KM, Kulpa CF (2005) Green Chem 7:185
24. Tropsha A, Gramatica P, Gombar VK (2003) QSAR Comb Sci 22:69
25. UFT Merck Ionic Liquids Biological Effects Database, http://www.il-eco.uft.unibremen.de/
26. ChemSketch ver. 12, Advanced Chemistry Development, http://www.acdlabs.com/resources/freeware/chemsketch/
27. HyperChem ver. 7, Hypercube Inc, http://www.hyper.com/
28. MOPAC 6.0, http://www.ccl.net/cca/software/MS-WIN95-NT/mopac6/index.shtml
29. Dragon, Talete s.r.l., http://www.talete.mi.it/index.htm
30. CODESSA, Semichem Inc, http://www.semichem.com/
31. Yasri A, Hartsough D (2001) J Chem Inf Comput Sci 41:1218
32. Hasegawa KJ (1999) Chem Inf Comput Sci 39:112
33. Leardi RJ (2007) J Chromatogr A 1158:226
34. STATISTICA rel. 7, Statsoft Inc, http://www.statsoft.com/
35. Kustrin SA, Beresford R (2000) J Pharm Biomed Anal 22:717
36. Daqi G, Yan J (2005) Pattern Recognit 38:1469
37. Kanzow C, Yamashita N, Fukushima M (2004) J Comput Appl Math 172:375