PCCP

Accepted Manuscript

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/pccp

Journal Name

RSCPublishing

ARTICLE

A structure-activity relationship study of the toxicity of ionic liquids using an adapted Ferreira-Kiralj hydrophobicity parameter

Cite this: DOI: 10.1039/x0xx00000x

Received 00th January 2012,
Accepted 00th January 2012

DOI: 10.1039/x0xx00000x

www.rsc.org/

Eduardo Borges de Melo[a]

The Ferreira-Kiralj hydrophobicity parameter W_c is a number fraction of hydrophobic carbon atoms and can be regarded as a constitutional descriptor since its calculation depends only on the number of polar and nonpolar carbons in a compound. Hydrophobicity is important to the toxicity of ionic liquids (ILs), which are salts by nature. Herein, a descriptor for this property was calculated using a simple adaptation of the type of polar carbon atoms included (W_cAdap) to explore the possibility of its use in quantitative structure-activity relationship (QSAR) studies of ILs. The resulting model was tested using a database of ILs with toxicity against the Leukemia rat cell line IPC-81. Two other models were constructed using CrippenLogP and MannholdLogP descriptors, which are both available in the free program PaDEL. The use of W_cAdap led to a better and more indicative model. Thus, W_cAdap may be a suitable molecular descriptor for the hydrophobicity of ILs in QSAR studies.

Introduction

Ionic liquids (ILs) are a class of chemicals that has recently emerged as alternatives to environmentally damaging volatile organic compounds (VOCs). These are comprised of a very large number of chemical compounds entirely made up of ions (*i.e.*, salts) with melting points below that of water, and thus, are liquid at, or close to, room temperature. The 1:1 molar ratio mixture of an organic cation and an organic or inorganic anion is called a *true ionic liquid.* The organic cation is generally imidazolium, pyridinium, quaternary ammonium, or quaternary phosphonium; the anion can be halide, triflate, trifluoroborate, or hexafluorophosphate. However, ILs are not simple mixtures because they do not retain the "identities" of the cation and anion. In addition, if two ILs are mixed together, the ionic associations found in each one are lost; *i.e.,* it is not possible to differentiate ions according their original ILs or to identify the unique interactions of individual ILs.[1-4]

These are considered as "green" alternatives to conventional solvents because of their outstanding properties, especially the negligible vapor pressure, and these do not contribute to air pollution.[3-4] Thus, ILs are an attractive medium for various types of chemical processes such as organic synthesis, catalysis or biocatalysis, protein purification, CO_2 capture, preparation of liquid crystals, drug synthesis and delivery, in batteries and solar panels,[5] absorption refrigeration systems,[6] membrane preparation,[3] biodiesel production,[7] and hydrogen sulfide and thiol scavenging.[8] Furthermore, their physicochemical properties (such as viscosity and density) can be properly adjusted by varying the ionic structure.[3] Thus, this class of compounds has attracted considerable interest in the chemical industry.

Although ILs can lessen the risk of air pollution owing to their insignificant vapor pressure, they do have significant solubility in water, which is the most probable route for the flow of ILs into aquatic ecosystems. In addition, their non-volatility, as well as high chemical and thermal stability (which are also of industrial interest) suggest potential problems with degradation or persistence in the environment.[5,9] Pham *et al*[5] reviewed several toxicological aspects and environmental fate of ILs. The environmental behavior of these compounds would be determined by their hydrophobicity: hydrophobic ILs can be attenuated by sediments and become persistent contaminants in the environment, while hydrophilic ILs are likely to enter aquatic ecosystems.[10]

Studies indicate that ILs may cause cellular and subcellular alterations in bacteria, human, and mammalian cell lines. Aquatic toxicity tests show that these compounds are capable of causing acute toxicity to animals and plants,[11-14] depending on the chemical structure.[15] Although ILs are being considered as green solvents, their toxicity can be many orders of magnitude greater than that of organic solvents.[16] Because of this, methods to derive quantitative structure-activity relationship (QSAR) models have been successfully applied to the prediction of various endpoints of ILs, especially physicochemical properties.[17]

However, considering the importance of hydrophobicity, it is important to note that many of the algorithms currently available are not parameterized for calculating LogP values of ionized molecules, salts, and chemical substances consisting of disconnected structures. Thus, models built on these conditions cannot be truly predictive, which can also be said for other types of non-parameterized molecular descriptors. Therefore, the objective of this study was to derive a useful molecular descriptor for characterizing the hydrophobicity of ionic liquids, independent of the experimental values of the partition coefficient and group or atom contributions. The results are presented as a QSAR study on the toxicity of ILs using simple descriptors[18] obtained from the PaDEL 2.2 program,[19] combined with the proposed descriptor, which is an adapted version of the Ferreira-Kiralj hydrophobicity parameter Wc.[20] This descriptor can be used as an alternative non-LogP-type hydrophobic descriptor for the QSAR studies of ILs.

Experimental

Data set

Recently, Zhao *et al.*[17] carried out QSAR studies using a database of 100 highly diverse ILs with toxicity (EC_{50} in µmol) against the Leukemia rat cell line IPC-81. This database, available in the UFT/Merck Ionic Liquids Biological Effects Database (http://www.il-eco.uft.uni-bremen.de), was selected to test the hypothesis of this study. In this work, the Simplified Molecular Input Line Entry System (SMILES) strings of each IL were also taken from this database, using the corresponding Chemical Abstracts Service Registry Number of each compound. Thus, the range of toxicity varied from -0.24 (high) to 4.58 (low). The database was split into a training set (ILs **1-80**) and a test set (ILs **81-90**), as done in the original reference.[17] A few representatives from this database are presented in Fig. 1. The SMILES strings of all compounds and dependent variables provided as in the original reference[17] ($LogEC_{50}$) are available in the Supplementary Information, Table S1.

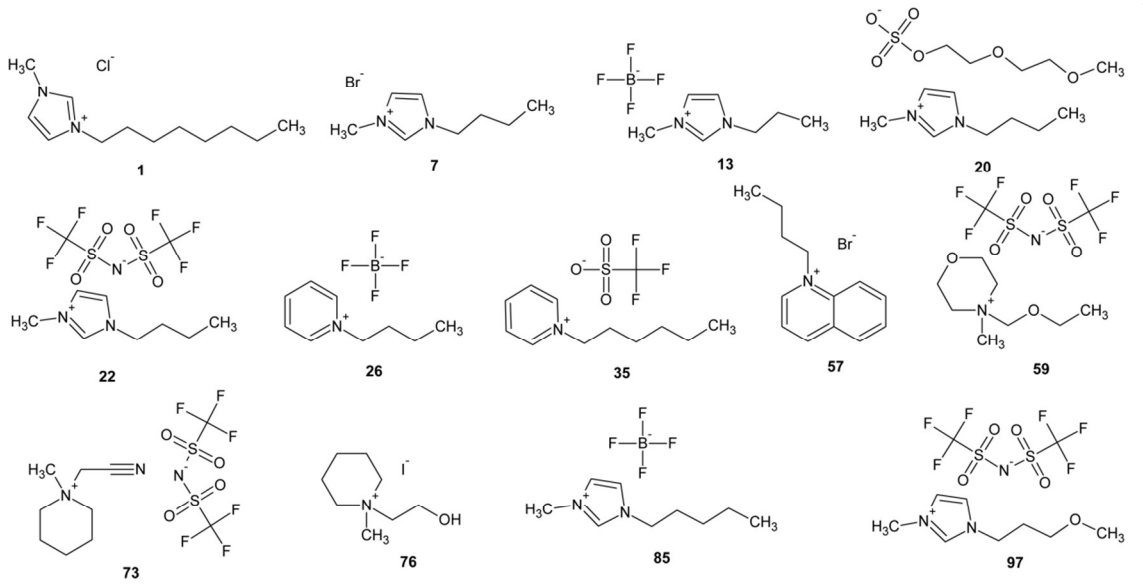

Fig. 1 Examples of ILs from the database in Zhao *et al.*[14]

Adapted Ferreira-Kiralj hydrophobicity parameter (W_cAdap) and other molecular descriptors

Considering the importance given to ILs in recent years, it is highly desirable to understand the basic factors affecting their behavior in different biological systems. One of these factors is hydrophobicity and its relationship to toxicity. According to Segundo Ranke *et al.*,[21] some theoretical methods to calculate LogP may work well for certain groups of ILs, but may fail if the necessary fragment constants for the structure of ILs under study are not available.

The Ferreira-Kiralj parameter W_c is a simple descriptor related to hydrophobicity. It is constitutional in nature because it is based only on the number of specific types of atoms that compose a molecule. Thus, W_c can be calculated using the equation:

$$W_c = \frac{N_C^{hyd}}{A - N_H}$$

where N_C^{hyd} is the number of hydrophobic carbon atoms, A is the total number of atoms, and N_H the number of hydrogen atoms. According to the original proposal of Ferreira and Kiralj,[20] hydrophobic carbon atoms pertain to all carbon atoms except those in C=O, C-O$^-$, and C≡N groups. As W_c is a number fraction of hydrophobic carbon atoms, the equation makes it clear that the greater the number of hydrophobic carbons in the

structure, the higher the value of W_c, ranging from 0, if the molecule has no hydrophobic carbon (*e.g.*, $H_2C=O$), to 1, if all carbons are hydrophobic (*e.g.*, $H_3C\text{-}CH_3$). This proposal was originally used for a set of β-lactam antibiotics, in a situation that can be considered similar to this study, where lipophilicity parameters do not necessarily contain the same information related to relationships between the efflux activities of the AcrAB-TolC pump of gram-negative bacteria and resistance to lipophilic and amphiphilic drugs, since some of these antibiotics have charged and delocalized functional groups.[20] Since ILs are salts, this situation can be considered similar. Another advantage of using a parameter such as W_c is that it can be applied to ILs with anions having an alkyl chain. Many ILs such as **1** (Fig. 1) are formed by cations with long chains and simple anions (Cl^-, Br^-, BF_4^-, or PF_6^-), and a number of studies, as well as other properties, are based only on the cationic structure. However, many ILs also have an anionic structure with long chains (*e.g.*, **20**, Fig. 1), which influences the physicochemical properties of ILs. In this context, a descriptor based only on the counting of specific types of carbon may allow assessment of the overall hydrophobicity of ILs, both those with simple and complex anions. However, it is important to note that it is generally accepted that anion influence on the toxicity of ILs is subordinate to the cation effect.[10]

Thus, to obtain W_cAdap, non-hydrophobic carbons connected to charged atoms, both in cationic and anionic structures, were considered aside from those originally proposed.[20] Table 1 shows the values of N_C^{hyd}, N_C^{hyd}Adap, A, N_H, W_c and W_cAdap of all compounds in the dataset. The values were obtained by simply counting the atoms in the two-dimensional structures that are also available in the UFT/Merck Ionic Liquids Biological Effects Database and in the molecular formulas.

Interestingly, the Pearson correlation coefficient r of the dependent variable with W_c using the compounds in the training set was only -0.01, indicating that the use of the parameter as originally proposed do not encode information relevant to the relationship between the hydrophobicity and toxicological activity of ILs under study. However, using W_cAdap incremented the calculated r to -0.385. Notably, this difference is present even when the r between W_c and W_cAdap is 0.835 (see values in Table 1).

The other descriptors for the study were generated using PaDEL 2.2 (download: http://padel.nus.edu.sg/software/padeldescriptor), a free and open-source JAVA-based software to calculate molecular descriptors and fingerprints.[19] Interestingly, given the objectives for the development of PaDEL and despite its free availability, to our knowledge, only three studies of ILs with QSAR descriptors derived from this program are available in the literature, all of which are recently published.[22-24] SMILES strings were used to generate 1D and 2D descriptors in PaDEL, which facilitated the derivation of descriptors and data reproduction by other researchers. Considering that this study was proposed to be carried out using the complete structure of each IL, simplest possible descriptors that are mainly based on atomic contributions and types of atoms were selected. Owing to the chemical nature of the dataset, the option *remove salts* was disabled. Some of the initially selected descriptors were missing or incomplete, probably because the algorithms are not parameterized for calculations of molecules with charged atoms or chemicals with disconnected structures such as ILs. Constant or near-constant descriptors were also manually discarded. Finally, descriptors having an r with the endpoint lower than 0.2 were removed. Thus, 25 descriptors were obtained. A list with these descriptors is available in the Supplementary Information, Table S2.

Among the calculated descriptors in PaDEL are four LogP parameters: CrippenLogP, MannholdLogP, ALogP and XLogP. The latter descriptor is calculated using a group contribution method parameterized for neutral organic molecules,[18] and thus was not selected for this study. CrippenLogP and ALogP have atomic contributions, according to the literature, justifying their potential use in predictive models for ILs,[18,25] while MannholdLogP is obtained based only on the number of carbons and heteroatoms that form the compound.[26] However, derivation of ALogP from SMILES strings led to some missing data and was therefore removed from the list of descriptors. Thus, only CrippenLogP and MannholdLogP were used for the derivation of QSAR models for comparison with results obtained using W_cAdap. The three parameters were observed to have similar orders of magnitude for r with the dependent variable: -0.417 for W_cAdap, -0.358 for CrippenLogP, and -0.498 for MannholdLogP.

For the process of variable selection, four different matrices, each with 26 descriptors, were built. The only difference between each matrix is the descriptor for hydrophobicity while the remaining descriptors were the same. The list is available in the Supplementary Information, Table S3.

QSAR study

The QSAR study was performed in QSAR Modeling, a free JAVA-based software developed by the research group of the Theoretical and Applied Chemometrics Laboratory (download: http://lqta.iqm.unicamp.br).[27] The final reduction of variables was carried out in this program. Matrices of descriptors were subjected to the method of selection of variables called Ordered Predictors Selection (OPS),[28] an iterative algorithm for building QSAR models.[29-31] This method uses Partial Least Squares, a regression method which reduces the size of the data by transforming them into mutually orthogonal latent variables (LVs),[31] to build models by rearranging the columns of the matrix in such a way that the most important descriptors, classified according to an informative vector (correlation vector, regression vector, and their product), are placed in the first column. In this study, the three vectors were used simultaneously. The models should be classified in descending order of a statistical parameter. In this study, the initial step was carried out using the root mean square error of cross-validation (*RMSECV*) to select the descriptors that can lead to smaller errors, and the subsequent one according to the coefficient of

determination of leave-one-out (LOO) cross-validation (Q^2_{LOO}) to maximize the prediction. As the numerical range of each selected descriptor may be very different, it is necessary to perform the pre-processing scheme known as autoscaling.[27,33,34] The obtained models were refined using Pirouette 4 (www.infometrix.com) by checking the possibility of removal of some of the descriptors to obtain an optimized, simpler, statistically significant, and interpretative model.

In QSAR, it is necessary to apply validation techniques to check the statistical quality of the predicting power of the obtained models. Thus, it is possible to provide a measure of the capacity of the models to perform reliable predictions of the dependent variables under study for compounds not used in the modeling step.[35] The most frequently used approach involves two steps: internal and external validation.

In internal validation, the explained variance of models were evaluated using the coefficient of determination ($R^2 > 0.6$) and the significance of the models was evaluated using the *F*-ratio test at 95% confidence interval ($\alpha = 0.05$). The internal prediction was tested by LOO cross-validation through the Q^2_{LOO} (> 0.5) and RmSquare metrics (Average r_m^2(pred)-scaled > 0.5 and Delta r_m^2(pred)-scaled < 0.2). The robustness of the model was tested by leave-*N*-out (LNO) cross-validation through the systematic removal of a maximum number of elements ($N = 10$) in the training set, in hexaplicate for each *N* value.[33,36-38] The chance correlation was verified by the y-randomization test (20 randomizations) using the approach suggested by Eriksson *et al.*[38] The models were recalculated after randomization. The R^2, Q^2_{LOO}, LNO, and y-randomization tests were carried out in QSAR Modeling.[27] The *F*-ratio test was calculated in Microsoft Excel, as described by Todeschini and Consoni.[18]

External validation of the obtained models was performed by predicting pEC_{50} values for the compounds of the test set. Thus, the predictive power was assessed using the coefficient of determination of external validation ($R^2_{pred} > 0.5$), RmSquare metrics of external prediction (Average r_m^2(pred)-scaled > 0.5 and Delta r_m^2(pred)-scaled < 0.2), and Golbraikh-Tropsha statistics (the slopes *k* and *k'* of the predicted versus observed and observed versus predicted response regression lines, respectively, both passing through the origin, where $0.85 < k, k' < 1.15$, and the absolute difference between the coefficient of determination of the predicted versus observed and observed versus predicted responses, both also passing through the origin, $|R^2_0 - R'^2_0| < 0.3$).[33,37,39] Statistical analyses were done using Xternal Validation Metric Calculator 1.0 (download: http://dtclab.webs.com/software-tools).

Table 1 Values of N_C^{hyd}, N_C^{hyd}Adap, A, N_H, and W_cAdap.

Sample	N_C^{hyd}	N_C^{hyd}Adap	A	N_H	W_c	W_cAdap
1	12	9	38	15	0.522	0.391
2	8	5	26	11	0.533	0.333
3	20	17	62	23	0.513	0.436
4	9	6	29	12	0.529	0.353
5	11	8	35	14	0.524	0.381
6	13	10	41	16	0.520	0.400
7	8	5	26	11	0.533	0.333
8	11	8	35	14	0.524	0.381
9	8	5	26	11	0.533	0.333
10	11	8	39	18	0.524	0.381
11	6	3	24	13	0.545	0.273
12	12	9	42	19	0.522	0.391
13	7	4	27	14	0.538	0.308
14	12	9	39	16	0.522	0.391
15	12	9	30	15	0.800	0.600
16	11	8	39	18	0.524	0.381
17	11	8	31	18	0.846	0.615
18	7	4	24	11	0.538	0.308
19	4	1	26	12	0.286	0.071
20	13	10	48	22	0.500	0.385
21	7	4	27	16	0.636	0.364
22	10	7	40	25	0.667	0.467
23	14	11	52	29	0.609	0.478
24	7	4	26	15	0.636	0.364
25	7	4	27	13	0.500	0.286
26	9	6	29	15	0.643	0.429
27	10	7	28	12	0.625	0.438
28	11	7	37	13	0.458	0.292
29	13	9	43	15	0.464	0.321
30	13	9	51	27	0.542	0.375
31	10	7	33	16	0.588	0.412
32	9	6	25	11	0.643	0.429
33	10	7	32	16	0.625	0.438
34	10	7	32	18	0.714	0.500
35	12	9	38	20	0.667	0.500
36	9	6	25	11	0.643	0.429
37	12	9	34	14	0.600	0.450
38	12	9	34	14	0.600	0.450
39	12	9	38	18	0.600	0.450
40	7	4	20	10	0.700	0.400
41	8	5	23	11	0.667	0.417
42	10	7	32	16	0.625	0.438
43	14	11	44	20	0.583	0.458
44	11	8	35	17	0.611	0.444
45	12	9	42	26	0.750	0.563
46	10	7	36	24	0.833	0.583
47	10	7	28	12	0.625	0.438
48	10	7	28	12	0.625	0.438
49	9	4	31	13	0.500	0.222
50	11	6	45	27	0.611	0.333
51	11	8	40	26	0.786	0.571
52	11	7	46	26	0.550	0.350
53	10	6	35	14	0.476	0.286
54	11	8	40	26	0.786	0.571
55	23	19	67	25	0.548	0.452
56	13	10	35	19	0.813	0.625
57	13	10	31	15	0.813	0.625
58	14	10	43	20	0.609	0.435
59	10	6	44	26	0.556	0.333
60	10	6	44	26	0.556	0.333
61	11	7	47	27	0.550	0.350
62	10	6	44	26	0.556	0.333
63	11	7	47	27	0.550	0.350
64	11	7	46	26	0.550	0.350
65	11	7	45	25	0.550	0.350
66	12	8	49	27	0.545	0.364
67	11	7	46	26	0.550	0.350
68	12	8	48	26	0.545	0.364
69	11	7	46	26	0.550	0.350

70	11	7	46	26	0.550	0.350
71	12	8	49	27	0.545	0.364
72	10	6	43	25	0.556	0.333
73	10	5	40	25	0.667	0.333
74	9	5	32	12	0.450	0.250
75	10	6	35	13	0.455	0.273
76	8	4	29	11	0.444	0.222
77	11	8	31	13	0.611	0.444
78	9	5	31	24	1.286	0.714
79	7	3	17	10	1.000	0.429
80	10	6	34	12	0.455	0.273
81	14	11	44	17	0.519	0.407
82	10	7	32	13	0.526	0.368
83	6	3	20	9	0.545	0.273
84	13	10	45	20	0.520	0.400
85	9	6	33	16	0.529	0.353
86	9	6	33	16	0.529	0.353
87	9	6	33	18	0.600	0.400
88	12	9	46	27	0.632	0.474
89	10	7	34	19	0.667	0.467
90	7	4	28	14	0.500	0.286
91	14	11	40	16	0.583	0.458
92	7	4	19	9	0.700	0.400
93	8	3	26	11	0.533	0.200
94	19	15	55	21	0.559	0.441
95	15	12	41	21	0.750	0.600
96	9	4	38	25	0.692	0.308
97	10	7	41	26	0.667	0.467
98	9	4	37	24	0.692	0.308
99	11	7	37	13	0.458	0.292
100	8	5	30	15	0.533	0.333

Results and discussion

The statistically superior model (highest values of *RMSECV* in the first cycle and Q^2_{LOO} in the subsequent one) among those that present the hydrophobicity descriptor of the corresponding matrix was always selected in each step of variable selection by OPS. The whole process was carried out in an iterative manner. The best models obtained were refined using Pirouette 4. To retain the most relevant ones, the hydrophobicity parameter was always kept to assess its importance to each set and ability to generate statistically appropriate models. Performing this process, three models (A to C given below) were obtained and in each one, two LVs were built. Models A and B have four molecular descriptors while model C has five.

Model A:

$LogEC_{50}$ = -5.754 - 0.038*(nBondS) + 0.095*(nO) - 0.150*(nAtomLac) - 2.313*(W_cAdap)

R^2 = 0.813; *RMSEC* = 0.378; F = 167.754; Q^2_{LOO} = 0.792; *RMSEV* = 0.400; Average r^2_m(LOO)-scaled = 0.703; Delta r^2_m(LOO)-scaled = 0.163; Cumulated information: 63.253% (LV1: 39.767%; LV2: 23.486%)

Model B:

$LogEC_{50}$ = +4.800 - 0.033*(nBondsS) + 0.118*(nO) - 0.160*(nAtomLAC) - 0.064*(CrippenLogP)

R^2 = 0.762; *RMSEC* = 0.440; F = 142.013; Q^2_{LOO} = 0.767; *RMSEV* = 0.423; Average r^2_m(LOO)-scaled = 0.688; Delta r^2_m(LOO)-scaled = 0.193; Cumulated information: 80.169% (LV1: 41.165%; LV2: 39.004%)

Model C:

$LogEC_{50}$ = +7.661 - 0.178*(nRotB) - 0.097*(nAromBond) - 0.303*(VAdjMat) - 0.017*(fragC) - 0.765(MannholdLogP)

R^2 = 0.770; *RMSEC* = 0.420; F = 128.715; Q^2_{LOO} = 0.731; *RMSEV* = 0.454; Average r^2_m(LOO)-scaled = 0.623; Delta r^2_m(LOO)-scaled = 0.182; Cumulated information: 82.521% (LV1: 22.154%; LV2: 60.367%)

The maximum variation of the accumulated information among the three models was 19.324%. Models B and C exhibit similar amount of information, although they are probably not related since any selected descriptor is common to both. It is different in the case of models A and B, which are formed by the same descriptors except the hydrophobicity one. The three models obtained can be considered reasonably similar in some of their statistics. The maximum variations between R^2 and Q^2_{LOO} were 5.1% and 6.1% for the explained and predicted variance, respectively. It can be seen that, based on R^2, model A had the highest amount of explained information (81.3%) while model B had the lowest (76.2%). Model A also had the highest amount of predicted information (79.2%); however, model B was better than model C in this parameter (76.7%). The data for the Average and Delta r_m^2(LOO)-scaled metrics show the same trend. However, comparing the difference between R^2 and Q^2_{LOO},[40] model B had the lowest probability of overfitting (0.005), while model C had the highest (0.039); nevertheless, all models presented low probabilities of data overfitting.

The most significant difference between the models was observed when the results of the F test (95% confidence interval, $\alpha = 0.05$) were analyzed. Since all models are formed by the same number of compounds (n = 80) and LVs (p = 2), the tabulated reference value is the same ($F_{p,n\text{-}p\text{-}1}$ = 3.115, p = 2 and n - p - 1 = 77); thus, it is possible to compare the significance of the models. Although all models were considerably higher than the reference value, model A was observed to be the most statistically significant, being 25.741 units higher than model B and 39.039 units higher than model C. Interestingly, this model accumulated the least amount of information from its two LVs.

Table 2 shows the results of robustness and chance correlation tests. The results demonstrate that the models show excellent robustness, with variations in the average Q^2_{LNO} and Q^2_{LOO} being only 0.002 units for the three models. Thus, the models are stable and resistant to small variations. The results of the y-randomization test indicate the absence of chance correlation in all models (*i.e.*, intercepts for the R^2 test are lower than 0.3 and those for the Q^2_{LOO} test are lower than 0.05). The plots for these tests are available in the Supplementary Information, Fig. S1.

Table 2 Results of LNO cross-validation and chance correlation test.

Models	Robustness	Intercept in chance correlation	
	Average Q^2_{LNO}	R^2 vs. $r(y_0,y_r)$	Q^2_{LOO} vs. $r(y_0,y_r)$
A	0.790	-0.025	-0.146
B	0.765	-0.010	-0.140
C	0.729	-0.022	-0.145

Despite the slightly high statistical quality of model A relative to models B and C (except for the significance of the regression), the three models show good internal qualities. However, as the ultimate goal of a QSAR model is the prediction of the endpoint of compounds not originally used for its derivation, such as in the development of new related molecules[41] or for regulatory purposes,[42] it is recommended that externally validated models be considered more realistic and applicable for prediction.[43]

Table 3 presents the results of external validation for the three models (observed and predicted values are available in the Supplementary Information, Table S4). Considering the results of all tests, model C was completely rejected because it has no external predictability. Initially, it was believed that the algorithm could lead to a good model since it was calculated using a simple equation with the number of carbons and heteroatoms taken from a large dataset of 95809 compounds,[26] which could be the reason the values of model C led to better individual r with $LogEC_{50}$. However, Dearden *et al.*[34] noted that in QSAR models, it is not uncommon for individual descriptors that show good individual correlation to lead to bad models when combined with other descriptors and vice versa.

Table 3 Results from external validation step.

Statistical parameter	Model A	Model B	Model C
R^2_{pred}	0.809	0.802	-0.554
RMSEP	0.498	0.505	1.417
Average r_m^2(pred)-scaled	0.657	0.623	-0.036
Delta r_m^2(pred)-scaled	0.169	0.190	0.108
k	0.943	0.974	0.889
k'	1.035	0.999	0.892
$\lvert R^2_0 - R'^2_0 \rvert$	0.081	0.108	3.336

On the other hand, both models A and B had good values in their external quality statistics. Their respective R^2_{pred}, besides being well above the minimum, are different by only 0.007 units. Model A presented the higher R^2_{pred} and showed slightly better values for the other parameters. To eliminate any doubt regarding the most appropriate model for the purpose of prediction, the r_m^2(overall)-scaled metrics was analyzed. According to Roy *et al.*,[44] this metrics is based on the prediction of a comparably large number of compounds since both training and test sets were used, and is recommended for the selection of the best predictive model among a set of comparable models. The values of both Average r_m^2(overall)-scaled and Delta r_m^2(overall)-scaled were slightly better for model A (0.694 and 0.178, respectively, compared to 0.659 and 0.196 for model B), and thus, this model can be considered to have the best overall predictive ability. Considering the criteria evaluated, is possible to propose that model A, obtained using W_cAdap, presents sufficient information related to the endpoint of compounds used in the study, and can be used for the prediction of $LogEC_{50}$ for ILs not considered in this study.

The quality of a QSAR model is enhanced if mechanistic interpretation of selected descriptors (Table 4) is possible.[42] For model A, W_cAdap and nAtomLAC show that the increase in cell toxicity of ILs is associated with an increase in the size and hydrophobicity of the aliphatic chains. The derivation of a model with these two descriptors may not be regarded as surprising: it is well documented that the toxicity of an IL can be modulated over several orders of magnitude by altering the hydrophobicity of the alkyl sidechain, especially that of the cation.[10,22,45] The descriptor nBonds is in accordance with this interpretation since the hydrophobicity of ILs under study increased with increasing size of alkyl sidechains, which consequently increases the number of bonds in the compounds. This result is in agreement with the previous proposal of Zhao *et al.*,[17] according to which ILs, because of their structural similarity with surfactants, can damage the cell membrane, leading to an increase in permeability, through a mechanism similar to that of the surfactants. Thus, the increase in hydrophobicity due to increase in the sidechain length of ILs facilitates the occurrence of this damage, and may lead to narcosis, a physicochemical process wherein membrane-bound proteins are disrupted by a chemical.[17]

On the other hand, the descriptor nO indicates that an increase in the number of oxygen atoms in the structure of ILs increases the value of $LogEC_{50}$, and hence decreases the toxicity. This result is also consistent with literature since the presence of oxygen and other heteroatoms increases the number of hydrogen bonds that the molecule can form, which consequently increases its solubility in aqueous medium (*i.e.,* reduces hydrophobicity), and hence causes reduced toxicity profile owing to the reduction in the penetrability of molecules across the biological membrane.[22,23] The values of each descriptor for each compound are also available in the Supplementary Information, Table S2.

Table 4 Selected descriptors and autoscaled coefficients for each model.

Symbol	Definition	Model A	Model B	Model C
fragC	Complexity of the system			-0.114
nAromBond	Number of aromatic bonds			-0.342
nAtomLAC	Number of atoms in the longest aliphatic chain	-0.498	-0.531	
nBonds	Number of bonds (excluding bonds with hydrogen)	-0.368	-0.317	
nO	Number of oxygen atoms	0.245	0.304	
nRotB	Number of rotatable bonds (excluding terminal bonds)			-0.573
VAdjMat	Vertex adjacency information (magnitude)			-0.165
W_cAdap		-0.271		
CrippenLogP			-0.185	
MannholdLogP				-0.554

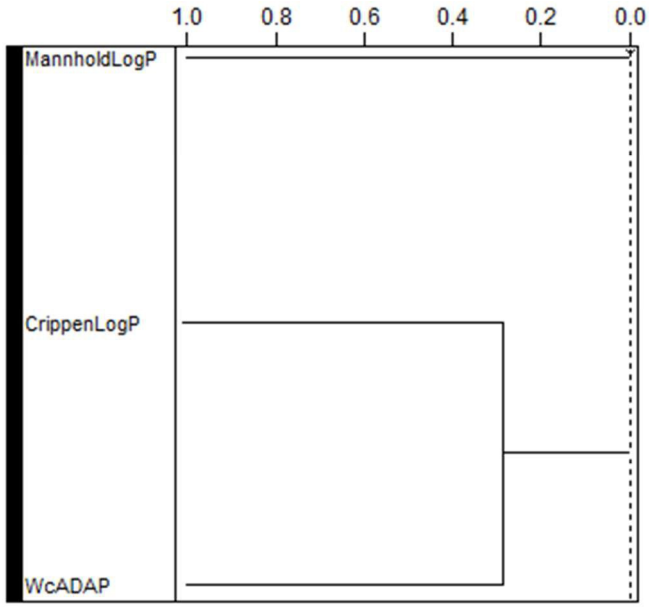

Fig. 2 Dendogram of the analysis based on hydrophobicity variables.

An important observation to be made is that models A and B are very similar in a number of aspects. The OPS approach combined with refinement led to the same descriptors for the two models, except for the hydrophobicity one. In both models, the most important descriptors are nAtomLAC and nBond, as can be observed from the absolute values of the autoscaled coefficients (Table 4). The third and fourth most important descriptors are W_cAdap and nO for model A and nO and CrippenLogP for model B. This degree of similarity prompted a comparison of the three hydrophobicity descriptors using Hierarchical Cluster Analysis (HCA), a classification multivariate method of data analysis that primarily aims to display data in such a way as to emphasize its natural clusters and patterns.[46] This analysis was carried out in Pirouette 4. Although the range of values for W_cAdap is very different from the other hydrophobicity descriptors, the values were also autoscaled.[34] It can be seen in Fig. 2 that MannholdLogP shows no degree of similarity to the other descriptors. On the other hand, W_cAdap and CrippenLogP have sufficient similarity to form a cluster. This may be the reason for the similar results obtained for both models and, more importantly, might imply that W_cAdap, despite being essentially a constitutional descriptor, actually encodes information related to the hydrophobicity of ILs. The lower amount of information that model A accumulated in comparison with model B may suggest that the latter has some "noise," which probably led to the better overall prediction capability of the former. This combination may indicate that information from model A is of better quality. As the only difference between the models is the hydrophobicity descriptor, this is where information would accumulate, at least when considering this database.

Conclusions

The models presented here, validated according to statistical criteria that are stricter than those typically used in QSAR studies, demonstrate that the adapted version of the Ferreira-Kiralj hydrophobicity parameter W_cAdap may be a suitable alternative for modeling the endpoints of ILs. This is particularly true given the salt nature of these compounds, which often cannot be adequately described by traditional algorithms for calculating LogP. Although constitutional in essence, information related to hydrophobicity can actually be encoded using this descriptor as shown by comparative analysis using HCA, thus validating the proposal of Ferreira and Kiralj. Moreover, model A, in addition to presenting the best overall predictive ability, also features descriptors that encode information related to modeling the endpoints of ILs to some extent.

Acknowledgements

The author thanks the *Fundação Araucária* (grant 2010/7354) and the *Programa de Apoio à Pós-Graduação da Coordenação de Aperfeiçoamento de Pessoal de Ensino Superior* (PROAP/CAPES) for financial support.

Notes and references

[a] Theoretical, Medicinal and Environmental Chemistry Laboratory (LQMAT), Department of Pharmacy, Western Paraná State University (UNIOESTE), 2069 Universitária St., 85819110 Cascavel, Paraná, Brazil. E-mail: eduardo.b.de.melo@gmail.com. Tel: +55-45-3220-3256.

Electronic Supplementary Information (ESI) available: Table S1. SMILES strings, Chemical Abstracts Service (CAS) code and LogEC50 for each sample of dataset; Table S2. Molecular descriptors used in this study; Table S3. Values of Crippen and Mannhold LogP; Table S4. Observed and predicted activities for Models 1, 2 and 3 in the external

validation; Fig. S1. Leave-N-out cross-validation and y-randomization plots. See DOI: 10.1039/b000000x/

1 M. Zakrewsky, K. S. Lovejoy, T. L. Kern, T. E. Miller, V. Le, A. Nagy, A. M. Goumas, R. S. Iyer, R. E. Del Sesto, A. T. Koppisch, D. T. Fox and S. Mitragotri, *Proc. Nat. Acad. Sci.* USA, 2014, **111**, 13313.
2 G. Chatel, J. F. B. Pereira, V. Debbeti, H. Wang and R. D. Rogers, *Green Chem*. 2014, **16**, 2051.
3 A. Figoli, T. Marino, S. Simone, E. Di Nicolò, X.-M. Li, T. He, S. Tornaghid and E. Driolia, *Green Chem*., 2014, **16**, 4034.
4 G. Cevasco and C. Chiappe, *Green Chem*., 2014, **16**, 2375.
5 T. P. T. Pham, C.-W. Cho and Y.-S. Yun, *Water Res.*, 2010, **44**, 352.
6 K. A. Kurnia, S. P. Pinho and J. A. P. Coutinho, *Green Chem.*, 2014, **16**, 3741.
7 F. Su and Y. Guo, *Green Chem.*, 2014, **16**, 2934.
8 H. Q. N. Gunaratne, P. Nockemann and K. R. Seddon, *Green Chem.*, 2014, **16**, 2411.
9 F. A. e Silva, F. Siopa, B. F. H. T. Figueiredo, A. M. M. Gonçalves, J. L. Pereira, F. Gonçalves, J. A. P. Coutinho, C. A. M. Afonso and S. P. M. Ventura, *Ecotoxicol. Environ. Saf*. 2014, **108**, 302.
10 M. C. Bubalo, K. Radošević, I. R. Redovniković, J. Halambek and V. G. Srček, *Ecotoxicol. Environ. Saf*., 2014, **99**, 1.
11 C. Pretti, C. Chiappe, D. Pieraccini, M. Gregori, F. Abramo, G. Monni and L. Intorre, *Green Chem*., 2006, **8**, 238.
12 C. Pretti, C. Chiappe, I. Baldetti, S. Brunini, G. Monni and L. Intorre, *Ecotoxicol. Eviron. Saf.*, 2009, **72**, 1170.
13 X. Y. Li, S. H. Zheng, X. Y. Dong, J. G. Ma and J. J. Wang, *Ecotoxicology*, 2012, **21**, 253.
14 C. Samorı, D. Malferrari, P. Valbonesi, A. Montecavalli, F. Moretti, P. Galletti, G. Sartor, E. Tagliavini, E. Fabbri and A. Pasteris, *Ecotoxicol. Environ. Saf.*, 2010, **73**, 1456.
15 K. Radošević, M. Cvjetko, N. Kopjar, R. Novak, J. Dumić and V. G. Srček, *Ecotoxicol. Environ. Saf.*, 2013, **92**, 112.
16 C.-W. Cho, Y.-C. Jeon, T. P. T. Pham, K. Vijayaraghavan and Y.-S. Yuna, *Ecotoxicol. Environ. Saf*., 2008, **71**, 166.
17 Y. Zhao, J. Zhao, Y. Huang, Q. Zhou, X. Zhang and S. Zhang, *J. Hazard. Mater.*, 2014, **278**, 320.
18 R. Todeschini and V. Consonni, *Molecular Descriptors for Chemoinformatics*, Wiley-VCH, Weinheim, 2nd edn., 2009, pp. 967.
19 C.W. Yap., *J. Comp. Chem.*, 2011, **32**, 1466.
20 M. M. C. Ferreira and R. Kiralj, *J. Chemom*., 2004, **18**, 242.
21 J. Ranke, A. Muller, U. Bottin-Weber, F. Stock, S. Stolte, J. Arning, R. Stormann and B. Jastorff, *Ecotoxicol. Environ. Saf.*, 2007, **67**, 430.
22 K. Roy and R. N. Das, *J. Hazard. Mater.,* 2013, **166**, 254.
23 R. N. Das and K. Roy, *Ind. Eng. Chem. Res*., 2014, **53**, 1020.
24 K. Roy, R. N. Das and P. L. A. Popelier, *Chemosphere*, 2014, **112**, 120.
25 S. A. Wildman and G. M. Crippen, *J. Chem. Inf. Comput. Sci.*, 1999, **39**, 868.
26 R. Mannhold, G. I. Poda, C. Ostermann, C. and I. V. Tetko, *J. Pharm. Sci.*, 2009, **98**, 861.
27 J. P. A. Martins and M. M. C. Ferreira, *Quím. Nova*, 2013, **36**, 554.
28 R. F. Teófilo, J. P. Martins and M. M. C. Ferreira, *J. Chemom*., 2009, **23**, 32.
29 M. Goodarzi, Y. V. Heyden and S. Funar-Timofei, *TRAC-Trend Anal. Chem*., 2013, **42**, 49.
30 E. B. de Melo, *Chemom. Intell. Lab. Syst*., 2012, **118**, 79.
31 N. B. H. Lozano, R. F. Oliveira, K. C. Weber, K. M. Honorio, R. V. Guido, A. D. Andricopulo and A. B. F. da Silva, *Molecules*, 2013, **18**, 5032.
32 H. Abdi, *Wiley Interdiscip. Rev. Comput. Stat.,* 2010, **2**, 97.
33 R. Kiralj and M. M. C. Ferreira, *J. Braz. Chem. Soc.*, 2009, **20**, 770.
34 J. C. Dearden, M. T. D. Cronin and K. L. E. Kaiser, *SAR QSAR Environ. Res.*, 2009, **20**, 241.
35 A. J. Leo, *Chem. Rev.*, 1993, **93**, 1281.
36 G. Melagraki, A. Afantitis, H. Sarimveis, P. A. Koutentis, J. Markopolus, O. Igglessi-Markopoulou, *J. Comput.-Aided Mol. Des.*, 2007, **21**, 251.
37 K. Roy, P. Chakraborty, I. Mitra, P. K. Ojha, S. Kar and R. N. Das, *J. Comp. Chem.* 2013, **24**, 1071.
38 L. Eriksson, J. Jaworska, A. P. Worth, M. T. Cronin, R. M. McDowell and P. Gramatica, *Environ. Health Perspect.*, 2003, **111**, 1361.
39 A. Golbraikh and A. Tropsha, *J. Mol. Graph. Model*., 2002, **20**, 269.
40 N. Chirico and P. Gramatica, *J. Chem. Inf. Model.*, 2012, **52**, 2044.
41 J. Ranke, S. Stolte, R. Stormann, J. Arning and B. Jastorff, *Chem. Rev.*, 2007, **107**, 2183.
42 E. Benfenati. Theory, guidance and applications on QSAR and REACH, Istituto di Ricerche Farmacologiche, Milan, 1st edn., 2012, pp. 179.
43 E. Papa, J. C. Dearden and P. Gramatica, *Chemosphere*, 2007, **67**, 351.
44 P. P. Roy, S. Paul, I. Mitra and K.Roy, *Molecules*, 2009, **14**, 1660.
45 A. Sosnowska, M. Barycki, M. Zaborowska, A. Rybinska and T. Puzyn, *Green Chem.*, 2014, DOI: 10.1039/C4GC00526K.
46 M. M. C. Ferreira. *J. Braz. Chem. Soc*., 2002, **13**, 74.