J Mol Model (2006) 12: 417–421
DOI 10.1007/s00894-005-0062-2

ORIGINAL PAPER

Kaido Tämm · Peeter Burk

QSPR analysis for infinite dilution activity coefficients of organic compounds

Received: 28 February 2005 / Accepted: 21 September 2005 / Published online: 7 December 2005

Abstract A quantitative structure–property relationship study of the infinite-dilution activity coefficients for a set of 38 organic compounds in ionic liquids such as 1-methyl-3-ethylimidazolium bis((trifluoromethyl)sulfonyl)imide, 1,2-dimethyl-3-ethylimidazolium bis((trifluoromethyl)-sulfonyl) imide, and 4-methyl-N-butylpyridinium tetrafluoroborate. QSPR study was carried out using the CODESSA PRO program. A general three-parameter QSPR model was obtained. Three orthogonal theoretical molecular descriptors satisfactorily correlate with the activity coefficients. The descriptors, such as the complementary information content, the fractional partial negative surface area and the count of hydrogen donor sites describe the dilution process in ILs.

Keywords Ionic liquids · QSPR · CODESSA PRO · Infinite dilution activity coefficients

Introduction

Ionic liquids (ILs) are liquids composed entirely of ions. They have garnered increasing interest in the last few years as novel solvents for electrochemistry, [1] biochemistry [2], and for synthesis and catalysis [3–6]. Organic ILs have been known for almost a century, but only during the last decade have they emerged as important materials with a growing applications base sufficient to sustain interest in their development. Inorganic liquids, also known as molten or fused salts, have an even longer history and are better characterized, but their high melting points, reactivity and poor solvation properties limit applications involving organic compounds. Organic ILs generally have lower melting points and favorable solvation properties for supporting a wider range of chemical applications involving organic compounds. ILs are generally defined as organic salts with a melting point below 150°C [7]. ILs have many useful properties with benefits described as follows: (1) they are good solvents for a wide range of both inorganic and organic materials, and allow unusual combinations of reagents to be brought into the same phase [8, 9]; (2) they are often composed of poorly coordinating ions, so they have the potential to be highly polar, yet non-coordinating solvents exhibiting solvatochromic properties rendering them similar to short chain alcohols [10]; (3) they are immiscible with many organic solvents and many cations combined with anions such as PF6- and $N(SO_2CF_3)_2$ can sustain a biphasic system where the IL is the extracting phase instead of a traditional organic solvent [9–12]; (4) they have relatively low viscosity, high thermal stability, and a wide temperature range for the liquid phase which encompasses that for water and ammonia [13]; (5) many ILs are nonvolatile and hence may be used in high-vacuum systems.

Until recently, there were few known room temperature ILs, which (because of their limited number) were treated as chemical curiosities. This picture has changed dramatically in a short time with over 250 room temperature ILs known today [14]. In addition, they are considered possible replacements for conventional organic solvents, which more often than not are liquid at room temperature. Recently Wilkes [15] published a review of properties of IL solvents for catalysis. Poole's [16] review covers the chromatographic and electroscopic methods for the determination of solvent properties of room temperature ILs.

For ILs to be used effectively as solvents, it is essential to know their interaction with different solutes. A quantitative measure of this property is given by the activity coefficient, γ_i, which describes the degree of nonideality for species i in a mixture. The infinite dilution activity coefficient, γ_i^{inf}, is especially important because it describes the extreme case in which only solute–solvent interactions contribute to nonideality. In addition to its theoretical importance, γ_i^{inf} has practical implications [17]. Separation processes for removing dilute impurities as encountered in many environmental applications, require knowledge of γ_i^{inf} for design purposes. Values of γ_i^{inf} are

K. Tämm (✉) · P. Burk
University of Tartu,
2 Jakobi str.,
Tartu, 51014, Estonia
e-mail: kaido@alfanet.ee
Tel.: +372-7275258

also important for evaluating the potential uses of ILs in liquid–liquid extraction and extractive distillation. Moreover, as shown in Eq. 1, Henry's law constants are directly related to γ_i^{inf}. In Eq. 1, H_i is the Henry's constant for solute i in the solvent of interest and P_i^{sat} is the vapor pressure of solute i at the temperature for which γ_i^{inf} is valid.

$$\gamma_i^{\mathrm{inf}} = \frac{H_i}{P_i^{\mathrm{sat}}} \tag{1}$$

This current study attempts to correlate experimentally measured infinite dilution activity coefficients with theoretical molecular descriptors.

Data set and methodology

The experimental data of activity coefficients at infinite dilution γ_i^{inf} for a variety of organic solvents i in ILs is taken from the work of Heintz and co-workers [18, 19]. Values for γ_i^{inf} for a total of 38 organic compounds in the

Table 1 Experimental Logarithmic Activity Coefficients at Infinite Dilution ln γ_i^{inf} in [bmpy][BF$_4$], [em$_2$im][N(Tf)$_2$] and [emim][N(Tf)$_2$] at 313 and 343 K[a]

	[bmpy][BF$_4$]		[em$_2$im][N(Tf)$_2$		[emim][N(Tf)$_2$]	
Compound /Temp.	313 K	343 K	313 K	343 K	313 K	343 K
1-butanol	1.288	0.952	1.359	0.989	1.080	0.771
1-hexanol	1.751	1.478	1.756	1.392	1.659	1.306
1-methylcyclohexene	2.992	2.907	2.520	2.363	2.506	2.368
1-pentanol	1.450	1.133	1.755	1.316	1.461	1.088
1-propanol	0.907	0.611	1.015	0.686	0.736	0.466
2,2,4-trimethylpentane	4.589	4.369	3.569	3.359	3.587	3.364
2-methyl-2-butanol	1.242	1.023	1.058	0.798	0.796	0.582
acetone	−0.097	−0.082	−0.772	−0.757	−0.919	−0.879
acetonitrile	−0.545	−0.573	−0.772	−0.821	−0.832	−0.870
a-methylstyrene	1.371	1.672	1.091	1.354	1.110	1.343
benzene	0.494	0.502	0.093	0.100	0.163	0.166
cyclohexane	3.368	3.143	2.699	2.456	2.656	2.465
cyclohexanol	1.206	1.020	1.420	1.157	1.101	0.890
cyclohexene	2.527	2.408	2.040	1.874	2.015	1.869
decane	5.730	5.452	4.879	4.507	5.019	4.617
dichloromethane	−0.290	−0.259	−0.095	−0.153	−0.115	−0.149
ethanol	0.562	0.285	0.714	0.399	0.422	0.171
ethyl acetate	0.977	0.944	0.066	0.068	−0.121	−0.079
ethylbenzene	1.560	1.544	1.035	1.000	1.042	1.018
heptane	4.458	4.290	3.622	3.391	3.647	3.423
hexane	4.102	3.957	3.229	3.050	3.233	3.093
isopropyl alcohol	0.871	0.589	0.925	0.588	0.658	0.369
isopropylbenzene	2.053	2.009	1.471	1.401	1.435	1.391
methanol	0.075	−0.177	0.378	0.073	0.123	−0.137
methyl tert-amyl ether	2.474	2.405	1.541	1.469	1.350	1.341
methyl tert-butyl ether	2.000	1.933	1.054	0.990	0.882	0.874
m-xylene	1.494	1.495	0.964	0.959	1.006	1.012
nonane	5.277	5.006	4.441	4.113	4.521	4.186
octane	4.862	4.646	4.015	3.738	4.069	3.809
o-xylene	1.304	1.325	0.793	0.809	0.874	0.896
p-xylene	1.405	1.423	0.945	0.938	0.995	1.009
sec-butanol	1.150	0.852	1.139	0.802	0.885	0.595
styrene	0.683	0.734	0.399	0.426	0.509	0.519
tert-butyl alcohol	0.994	0.710	0.888	0.591	0.640	0.395
tert-butylbenzene	2.386	2.312	1.722	1.642	1.663	1.612
tetrachloromethane	1.317	1.369	1.220	1.158	1.178	1.147
toluene	0.968	0.989	0.499	0.514	0.551	0.574
trichloromethane	−0.312	−0.184	−0.043	0.006	−0.028	0.029

[a] For some compound the temperatures may vary. The exact temperatures are given in References [18, 19]

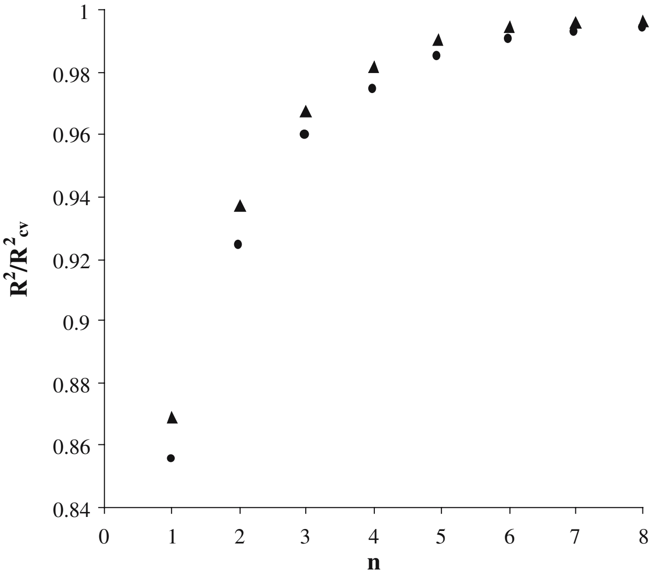

Fig. 1 Number of parameters (n) plotted *vs.* R^2 (▲) and R^2_{cv} (●)

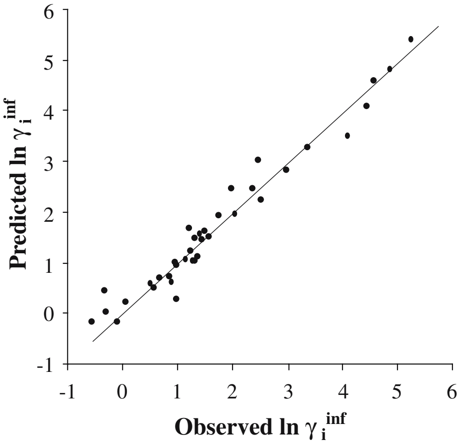

Fig. 2 Observed vs. predicted values of ln γ_i^{inf} for solutes in the IL [bmpy][BF_4] at 313 K

ILs 1-methyl-3-ethylimidazolium bis((trifluoromethyl)sulfonyl)imide ([emim][N($Tf)_2$]), 1,2-dimethyl-3-ethylimidazolium bis((trifluoromethyl)sulfonyl)imide ([em_2im][N($Tf)_2$]), and 4-methyl-*N*-butylpyridinium tetrafluoroborate ([bmpy][BF_4]) were studied at 313 and 343 K (Table 1).

The methodology for a general QSPR approach has been developed and coded as the CODESSA PRO [20] software, which combines different ways of quantifying the structural information about a molecule with advanced statistical analyses for establishing molecular structure–property relationships. CODESSA PRO can calculate a large number of quantitative descriptors solely on the basis of molecular structural information [20]. CODESSA PRO has been used successfully to predict a variety of physical properties of compounds [21–23]. The structures were drawn using ISIS Draw 2.4 [24] and pre-optimized using the molecular mechanics force-field method (MM+) available in HyperChem 7.0 [25]. Final geometrical optimization was performed with a cloned version of MOPAC 7.0 [26] as implemented in the CODESSA PRO software using the AM1 semiempirical method. [27] Thereafter, CODESSA PRO was used to calculate five types of molecular descriptors: constitutional, topological, geometrical, electrostatic and quantum-chemical [28, 29]. The constitutional and topological descriptors were calculated from the 2D structures of the molecules. The geometrical, electrostatic and quantum chemical descriptors were obtained using the geometries optimized with the AM1 method and the corresponding wave functions. Altogether, 656 descriptors were calculated for each of the 38 compounds studied. The correlation analysis to find the best QSPR model was carried out using the BMLR (best multi-linear regression) method in CODESSA PRO. The best multi-linear regression method is based on the (1) selection of orthogonal descriptor pairs, (2) extension of the correlation (saved in the previous step) with the addition of new descriptors until the Fisher-criterion (*F*) [30] becomes less than that of the best two-parameter correlation. The best *N* correlations (by R^2) are saved.

Results and discussion

To find the optimum number of descriptors [31, 32] describing the activity coefficients at infinite dilution in ILs for the current set of structures, we analyzed multi-

Table 2 Three-parameter QSPR models for the ln γ_i^{inf}

				Descriptor's coefficients			Statistical parameters		
Equation	IL	temp. (K)	intercept	^{0}CIC	FNSA-2	H-donors	R^2	R^2_{cv}	s
1	[bmpy][BF_4]	313	0.821	0.043	5.235	−0.083	0.966	0.957	0.308
2	[bmpy][BF_4]	343	0.719	0.042	4.569	−0.101	0.966	0.957	0.298
3	[em_2im] [N($Tf)_2$]	313	0.708	0.036	5.367	−0.036	0.945	0.933	0.332
4	[em_2im] [N($Tf)_2$]	343	0.545	0.034	4.545	−0.056	0.946	0.933	0.311
5	[emim] [N($Tf)_2$]	313	0.589	0.037	5.182	−0.057	0.943	0.932	0.349
6	[emim] [N($Tf)_2$]	343	0.475	0.035	4.444	−0.074	0.945	0.932	0.327

R^2–squared correlation coefficient;
R^2_{cv}–squared cross–validated correlation coefficient,
s–standard error;

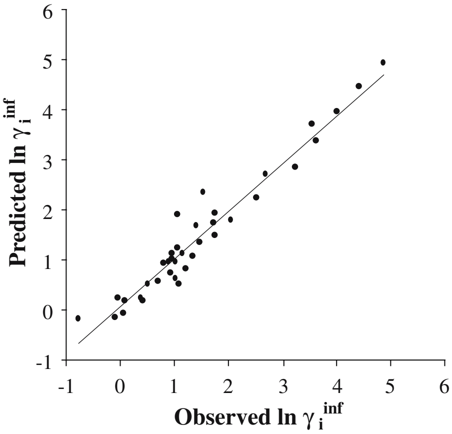

Fig. 3 Observed vs. predicted values of ln γ_i^{inf} for solutes in the IL [em$_2$im][N(Tf)$_2$] at 313 K

parameter correlations containing up to eight descriptors. Fig. 1 shows the relationships of squared correlation coefficient (R^2) and squared cross-validated correlation coefficient ($R^2{}_{\mathrm{cv}}$) of the models with the number of descriptors (n). Fig. 1 was drawn using the one to eight parameter models calculated for the γ_i^{inf} in [bmpy][BF$_4$] at 313 K [(Table 2, Eq. 1] but similar trends were obtained for all six equations shown in Table 2. As can be seen in Fig. 1, R^2 and $R^2{}_{\mathrm{cv}}$ rise steeply as the number of parameters increases from one to eight. In order to avoid the "over-parameterization" of the model, an increase of the R^2 value of less than 0.02 was chosen as the breakpoint criterion. Therefore, we used the best correlation equations with three descriptors for the analysis.

Three orthogonal descriptors were obtained for describing the infinite dilution activity coefficients in three different ILs at 313 and 343 K. The QSPR models with their statistical parameter are shown in Table 2. Figs. 2, 3, 4 demonstrate the correlations between experimental and predicted ln γ_i^{inf} values in three different ILs at 313 K.

Descriptor ^{0}CIC is the complementary information content of zeroth order [33] and is defined by Eq. 2, where n_i is the number of atoms in the ith class, n is the total number of atoms in the molecule, and ^{0}IC is the information content of zeroth order itself defined by Eq. 3. The descriptor ^{0}CIC describes the atomic connectivity in the molecule and encodes the size and the atomic constitution of the compound. These parameters directly affect the intermolecular interaction.

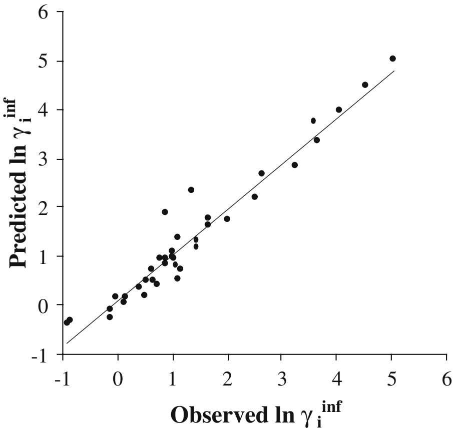

Fig. 4 Observed vs. predicted values of ln γ_i^{inf} for solutes in the IL [emim][N(Tf)$_2$] at 313 K

Table 3 Intercorrelation of the descriptors (R^2 values)

	^{0}CIC	FNSA-2	H-donors
^{0}CIC	–	0.032	0.028
FNSA-2	0.032	–	0.033
H-donors	0.028	0.033	–

Table 4 Validation of the six-parameter model

training sets	R^2 (fit)	s (fit)	predicted sets	R^2 (pred)	s (pred)
B + C	0.965	0.283	A	0.962	0.381
A + C	0.968	0.311	B	0.962	0.379
A + B	0.967	0.338	C	0.965	0.332
average	0.967	0.311	average	0.963	0.365

$$^0CIC = \log_2 n - {}^0IC \tag{2}$$

$$^0IC = -\sum_{i=1}^{0} \frac{n_i}{n} \log_2 \frac{n_i}{n} \tag{3}$$

The second descriptor, the fractional partial negative surface area (FNSA-2) belongs to the class of CPSA (charged partial surface area) descriptors and is defined [(Eq. 4] as the ratio of the total charge weighted partial negative surface area (PNSA2) to the total molecular solvent-accessible surface area (TMSA) [34]. CPSA descriptors are expected to encode the features responsible for polar interactions between molecules.

$$\mathrm{FNSA2} = \frac{\mathrm{PNSA2}}{\mathrm{TMSA}} \tag{4}$$

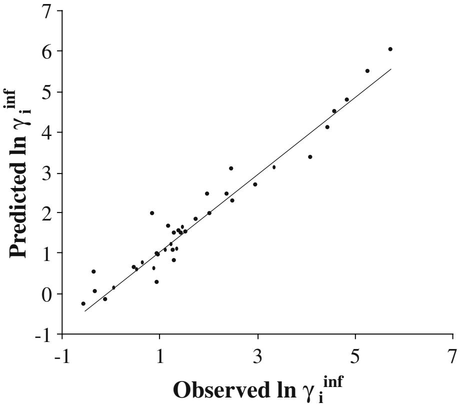

Fig. 5 Cross-validation plot for the six-parameter model

The third descriptor is the count of hydrogen-donor sites. This descriptor directly indicates the hydrogen-donor ability of the molecule—compounds with higher counts of hydrogen donors are more soluble in the ILs.

A major challenge in the development of multiple regression equations is connected with the possible multicollinearity of molecular-descriptor scales. In the case of high mutual correlation of the descriptors, the overall statistical characteristics of the regression may be satisfactory, but the reliability of the descriptor's coefficients, and thus of the whole regression, is low. Multicollinearity can be avoided, at least in part, by examining the correlation coefficients between the descriptor's scales in the QSPR model [29]. Thus, in Table 3 we have listed the correlation coefficients between the descriptors involved in the current three-parameter models. Table 3 demonstrates that all the descriptors are strongly orthogonal, which reflects the statistical reliability of the model.

Validation

To demonstrate the absence of chance correlations, we used the internal validation method. The full set of 38 structures was divided into three groups: structures 1, 4, 7, etc. formed group A, structures 2, 5, 7, etc. formed group B, and structures 3, 6, 8, etc. formed group C. Each subset was predicted by using the other two subsets as the training set. In this procedure, the same descriptors were retained in the correlation equation, but the coefficients were allowed to vary. Similar methods have been used elsewhere. [21, 35, 36]

The results shown in Table 4 disclose an average training quality of R^2=0.967 and an average predicting quality of R^2=0.963, which demonstrates that the proposed model has a good statistical stability and validity. The correlation chart of the validation showing the summary of all three predictions is given in Fig. 5.

Conclusions

Three-descriptor QSPR models with good statistical parameters were obtained to correlate with infinite dilution activity coefficients of organic compounds in three different ILs ([emim][N(Tf)$_2$], [em$_2$im][N(Tf)$_2$], and [bmpy][BF$_4$]) in two different temperatures (313 and 343 K). The R^2 values for the models vary from 0.943 up to 0.966. All the descriptors involved were calculated solely from the chemical structures and should describe the dilution mechanism of organic compounds in ILs. The example of the internal validation demonstrate the stability and the reliability of the models.

Acknowledgements Part of this work was supported by the Estonian Science Foundation Grant #5196. The authors thank Dr. Uko Maran for his contribution to the current work.

References

1. McFarlane DR, Sun J, Golding J, Meakin P, Forsyth M (2002) Electrochim Acta 45:1271–1278
2. Cull SG, Holbrey JD, Vargas-Mora V, Seddon KR, Lye GJ (2002) Biotech Bioeng 69:228–233
3. Olivier H (1999) J Mol Catalysis A 146:285–289
4. Fischer T, Sethi A, Welton T, Woolf J (1999) Tetrahedron Lett 40:793–796
5. Lee CW (1999) Tetrahedron Lett 40:2461–2464
6. Welton T (1999) Chem Rev 99:2071–2083
7. Seddon KR (1997) J Chem Technol Biotechnol 68:351–356
8. Thied RC, Seddon KR, Pitner WR, Rooney DW (1999) World Patent 41752
9. Blanchard LA, Brennecke JF (2001) Ind Eng Chem Res 40:287–292
10. Huddleston JG, Visser AE, Reichert WM, Willauer HD, Broker GA, Rogers RD (2001) Green Chem 3:156–164
11. Huddleston JG, Willauer HD, Swatloski RP, Visser AE, Rogers RD (1998) Chem Commun 16:1765–1766
12. Visser AE, Swatloski RP, Reichert WM, Griffin ST, Rogers RD (2000) Ind Eng Chem Res 39:3596–3604
13. Freemantle M (1998) Chem Eng News 13:32–37
14. http://www.ionicliquids-merck.de
15. Wilkes JS (2004) J Mol Catal A 214:11–17
16. Poole CF (2004) J Chromatogr A 1037:49–82
17. Sandler SI (1996) Fluid Phase Equilib 116:333–342
18. Heintz A, Kulikov DV, Verevkin SP (2001) J Chem Eng Data 46:1526–1529
19. Heintz A, Kulikov DV, Verevkin SP (2002) J Chem Eng Data 47:894–899
20. http://www.codessa-pro.com
21. Katritzky AR, Lomaka A, Petrukhin R, Jain R, Karelson M, Visser AE, Rogers RD (2002) J Chem Inf Comput Sci 42:71–74
22. Tämm K, Fara DC, Katritzky AR, Burk P, Karelson M (2004) J Chem Phys A 108:4812–4818
23. Katritzky AR, Tämm K, Kuanar M, Fara DC, Oliferenko A, Oliferenko P, Huddleston JG, Rogers RD (2004) J Chem Inf Comput Sci 44:136–142
24. http://www.mdl.com
25. http://www.hyper.com
26. Stewart JJP (1993) MOPAC 7.0 QCPE #455, http://qcpe.chem.indiana.edu
27. Dewar MJS, Zoebisch EG, Healy EF, Stewart JJP (1985) J Am Chem Soc 107:3902–3909
28. Katritzky AR, Lobanov VS, Karelson M (1995) Chem Soc Rev 24:279–287
29. Karelson, M (2000) Molecular Descriptors in QSAR/QSPR. Wiley, New York
30. Myers RH (1989) Classical and Modern Regression with Applications. PWS-KENT, Boston
31. Katritzky AR, Gordeeva EVJ (1993) Chem Inf Comput Sci 33:835–857
32. Katritzky AR, Petrukhin R, Jain R, Karelson M (2001) J Chem Inf Comput Sci 41:1521–1530
33. Basak SC, Harriss DK, Magnuson VR (1984) J Pharm Sci 73:429–437
34. Stanton DT, Jurs PC (1990) Anal Chem 62:2323–2329
35. Katritzky AR, Maran U, Karelson M, Lobanov VS (1997) J Chem Inf Comput Sci 37:913–919
36. Katritzky AR, Tatham DB, Maran U (2001) J Chem Inf Comput Sci 41:1162–1176