Electronic Supplementary Material (ESI) for Physical Chemistry Chemical Physics

Estimation of the refractive indices of imidazolium-based ionic liquids using their polarisability values

Pablo Díaz-Rodríguez,[1] John C. Cancilla,[1] Natalia V. Plechkova,[2] Gemma Matute,[1] Kenneth R. Seddon,[2] and José S. Torrecilla[1,*]

[1] Department of Chemical Engineering, Complutense University of Madrid, E–28040 Madrid, Spain. [2] QUILL Centre, School of Chemistry and Chemical Engineering, The Queen's University of Belfast, Belfast, BT9 5AG, UK.

Supplementary Information

Table SI-1. Refractive indices (measured at 298.15 K) of the 72 imidazolium-based ionic liquids found in the literature employed to create the multiple linear regression and artificial neural network models.

IL	RI	Reference
1-Ethyl-3-methylimidazolium tris(pentafluoroethyl)trifluorophosphate	1.3691	[1]
1-Butyl-3-methylimidazolium bis{(nonafluorobutyl)sulfonyl}amide	1.3880	[1]
1-[2,2,2-trifluoroethyl]-3-methylimidazolium bis{(trifluoromethyl)sulfonyl}amide	1.4090	[2]; [3](ILT)
1-Butyl-3-methylimidazolium hexafluorophosphate	1.4090	[4]
1-Ethyl-3-methylimidazolium tetrafluoroborate	1.4117	[1]
1-Propyl-3-methylimidazolium tetrafluoroborate	1.4165	[5]
1-Hexyl-3-methylimidazolium hexafluorophosphate	1.4179	[6]; [7] (ILT)
1-Butyl-3-methylimidazolium tetrafluoroborate	1.4210	[8]
1,3-Dimethylimidazolium bis{(trifluoromethyl)sulfonyl}amide	1.4220	[4]
1-Ethyl-3-methylimidazolium bis{(trifluoromethyl)sulfonyl}amide	1.4230	[6]; [3] (ILT)
1-Methyl-3-octylimidazolium hexafluorophosphate	1.4230	[1]
1-Octyl-3-methylimidazolium hexafluorophosphate	1.4235	[6]
1-Pentyl-3-methylimidazolium tetrafluoroborate	1.4238	[5]
3-Butyloxymethyl-1-methylimidazolium tetrafluoroborate	1.4239	[9]
3-Pentyloxymethyl-1-methylimidazolium tetrafluoroborate	1.4257	[9]
1,3-Diethylimidazolium bis{(trifluoromethyl)sulfonyl}amide	1.4260	[4]
1-Butyl-3-methylimidazolium bis{(trifluoromethyl)sulfonyl}amide	1.4269	[10] (ILT)
1-Hexyl-3-methylimidazolium tetrafluoroborate	1.4270	[5]; [11] (ILT)
1-Ethyl-3,5-dimethylimidazolium bis{(trifluoromethyl)sulfonyl}amide	1.4275	[4]
1-Butyl-3-ethylimidazolium bis{(trifluoromethyl)sulfonyl}amide	1.4285	[4]
1-Isobutyl-3-methylimidazolium bis{(trifluoromethyl)sulfonyl}amide	1.4289	[4]
1-Methoxiethyl-3-methylimidazolium bis{(trifluoromethyl)sulfonyl}amide	1.4293	[4]
3-Hexyloxymethyl-1-methylimidazolium tetrafluoroborate	1.4293	[9]
1,3-Diethyl-5-methylimidazolium bis{(trifluoromethyl)sulfonyl}amide	1.4300	[4]
1-Hexyl-3-methylimidazolium bis{(trifluoromethyl)sulfonyl}amide	1.4302	[12]
1-Ethyl-2,3-dimethylimidazolium bis{(trifluoromethyl)sulfonyl}amide	1.4305	[4]
3-Butyloxymethyl-1-methylimidazolium bis{(trifluoromethyl)sulfonyl}amide	1.4309	[9]
3-Heptyloxymethyl-1-methylimidazolium tetrafluoroborate	1.4313	[9]
3-Pentyloxymethyl-1-methylimidazolium bis{(trifluoromethyl)sulfonyl}amide	1.4320	[9]
1,3-Dibutylimidazolium bis{(trifluoromethyl)sulfonyl}amide	1.4324	[13]
1-Butyl-2,3-dimethylimidazolium tetrafluoroborate	1.4330	[14] (ILT)
1-Methyl-3-octylimidazolium bis{(trifluoromethyl)sulfonyl}amide	1.4333	[8]
1-Octyl-3-methylimidazolium bis{(trifluoromethyl)sulfonyl}amide	1.4333	[6]; [15] (ILT)
3-Hexyloxymethyl-1-methylimidazolium bis{(trifluoromethyl)sulfonyl}amide	1.4337	[9]
3-Octyloxymethyl-1-methylimidazolium tetrafluoroborate	1.4338	[9]
1-Methyl-3-nonylimidazolium bis{(trifluoromethyl)sulfonyl}amide	1.4340	[16]
1-Nonyl-3-methylimidazolium bis{(trifluoromethyl)sulfonyl}amide	1.4340	[16]
1-Methyl-3-octylimidazolium tetrafluoroborate	1.4342	[17]; [18] (ILT)
1-Octyl-3-methylimidazolium tetrafluoroborate	1.4342	[17]

* To whom correspondence should be addressed. Tel.: +34 91 394 42 44; Fax: +34 91 394 42 43. E-mail address: jstorre@ucm.es (José S. Torrecilla).

3-Heptyloxymethyl-1-methylimidazolium bis{(trifluoromethyl)sulfonyl}amide	1.4346	[9]
1-Decyl-3-methylimidazolium bis{(trifluoromethyl)sulfonyl}amide	1.4356	[6]
1-Methyl-3-decylimidazolium tetrafluoroborate	1.4367	[9]
3-Nonyloxymethyl-1-methylimidazolium bis{(trifluoromethyl)sulfonyl}amide	1.4369	[9]
1-Dodecyl-3-methylimidazolium bis{(trifluoromethyl)sulfonyl}amide	1.4376	[6]
1-Vinyl-3-hexylimidazolium bis{(trifluoromethyl)sulfonyl}amide	1.4430	[16]
1-Vinyl-3-nonylimidazolium bis{(trifluoromethyl)sulfonyl}amide	1.4450	[16]
1-Ethyl-3-methylimidazolium bis(fluorosulfonyl)amide	1.4474	[1]
1-Ethyl-3-methylimidazolium tetracyanoborate	1.4476	[1]
1-Vinyl-3-dodecylimidazolium bis{(trifluoromethyl)sulfonyl}amide	1.4480	[16]
1-Methyl-3-octylimidazolium 2-(2-methoxyethoxy)ethyl sulphate	1.4660	[19] (ILT)
1-Butyl-3-methylimidazolium octylsulfate	1.4699	[20]
1-Ethyl-3-methylimidazolium octylsulfate	1.4710	[19] (ILT)
1-Decyl-3-methylimidazolium dioctylsulfosuccinate	1.4724	[21]
1-Methyl-3-octylimidazolium dioctylsulfosuccinate	1.4724	[21]
1-Hexyl-3-methylimidazolium dioctylsulfosuccinate	1.4738	[21]
1-Butyl-3-methylimidazolium dioctylsulfosuccinate	1.4740	[21]
1-Ethyl-3-methylimidazolium dioctylsulfosuccinate	1.4747	[21]
1-Butyl-3-methylimidazolium methylsulfate	1.4794	[4]
1-Ethyl-3-methylimidazolium ethylsulfate	1.4794	[2]
1-Butyl-3-methylimidazolium 2-(2-methoxyethoxy)ethylsulfate	1.4800	[22] (ILT)
1-Ethyl-3-methylimidazolium methylsulfate	1.4810	[21]
1,3-Dimethylimidazolium methylsulfate	1.4827	[23] (ILT); [24]
1-Decyl-3-methylimidazolium dicyanamide	1.4954	[21]
1-Ethyl-3-methylimidazolium hydrogen sulphate	1.4971	[1]
1-Methyl-3-octylimidazolium dicyanamide	1.4991	[21]
1-Hexyl-3-methylimidazolium dicyanamide	1.5042	[12]
1-Methyl-3-octylimidazolium chloride	1.5050	[9] (ILT)
1-Butyl-3-methylimidazolium dicyanamide	1.5089	[12]
1-Ethyl-3-methylimidazolium dicyanamide	1.5127	[1]
1-Hexyl-3-methylimidazolium chloride	1.5150	[3] (ILT)
1-Butyl-3-methylimidazolium bromide	1.5400	[9] (ILT)
1-Butyl-3-methylimidazolium iodide	1.5720	[3] (ILT)

ILT: Ionic liquids from *IL Thermo* database (NIST Standard Reference Database 147).

Artificial Neural Networks Training Characteristics.

The training phase of artificial neural networks (ANNs) is essential for these systems to be functional, because during this process the weighted parameters (known as "weights") which control the connections existing between the different layers of a multilayer perceptron (MLP) are modified. These weights are responsible of the different relative importance of every connection for the fed-forward signals of the network,[25] and their optimisation is essential in order to properly train the MLP and obtain the best estimation of the dependent variables (*vide infra*).[26] To do so, the training dataset is firstly fed-forward through the different layers of the network, and the response is calculated by the neurons (in the hidden and output layers). The mathematical calculations required to obtain a response are fulfilled by each neuron, and have two consecutive steps. The first one determines the sum of the products obtained by multiplying each value by its corresponding weight, and it is performed by an activation function (Equation (SI-1)). The acquired result is then introduced in a transfer function, which restricts the range of the output value of a specific neuron. One of the most commonly employed functions for this task is the sigmoid function (Equation (SI-2)), which provides values between zero and one. It has been the one selected for the designed MLP.

$$x_k = \sum_{jk} w_{jk} \cdot y_j \quad \text{(SI-1)}$$

$$f(x) = \left(\frac{1}{1 + e^{-x_k}}\right) \quad \text{(SI-2)}$$

In the equations above, w_{jk} symbolises the weight which represents the connection between layers j and k, y_j is the fed-forward signal, and x_k and $f(x)$ are the activation and transfer function solutions respectively.[27]

In order to begin the training phase of the MLP, the learning dataset (Table SI-1) was employed to initiate the weight optimisation, and the verification dataset (Table SI-1) was used to grant that the generalisation ability of the MLP is adequate by checking the verification estimation error of the network. Both processes complete a whole cycle or epoch during the ANN training phase. During the learning process, the input values (polarisabilities) of the learning dataset (Table SI-1) are employed by the network to estimate their corresponding output values (refractive index). Once this process has finished, the obtained results are compared to the given targets (theoretical results), resulting in a training prediction error, which is calculated with the mean square error (MSE, Equation (SI-3)). The ANN will optimise the weight values to reduce this error, which should lead towards more accurate outputs when compared to the target values. Once the weight optimisation has been completed, the verification step begins for the active training cycle. In this second step the weights are not modified, and the MLP estimates the output value of the verification dataset (Table SI-1), which has not been used during the learning process, but must be always contained in the range of the training dataset. Once it finishes, a verification prediction error is obtained, and the epoch ends. Additional training cycles occur until this verification prediction error (mean prediction error, MPE, Equation (SI-4)) starts to rise. This is the moment when the weights can be thought of as optimised.[25]

$$MSE = \frac{1}{N}\sum_{N}\left(r - f(x)\right)^2 \quad \text{(SI-3)}$$

$$MPE = \frac{1}{n}\sum_{n}\frac{|r_n - y_n|}{r_n} \cdot 100 \quad \text{(SI-4)}$$

In Equations (SI-3) and (SI-4), N represents the number of neurons in the output layer, r is the real already known value, and f(x) stands for the solution provided by the transfer function for the output value of a given neuron, n represents the number of data from the verification database and r_n and y_n, the real and estimated output values respectively.

The Bayesian Regulation training function (trainBR) was selected as the training function because of the acceptable estimative results it offers when working with relatively small databases.[28,29]

Besides training the network, the optimisation process of the MLP involves the modification of a series of parameters, which were optimised following an exhaustive experimental design (*vide infra*). They are the hidden neuron number (HNN), and other neural network parameters like the learning coefficient (Lc), the learning coefficient decrease (Lcd), and the learning coefficient increase (Lci).[28] The Lc is responsible of the modification degree of the weights during the learning process, while the Lcd and Lci control it.

Once the MLP was optimised, a validation process was performed following the K-fold Cross Validation method, which is a computer intensive technique that employs all available data points as part of the training dataset as well as the verification dataset. Therefore, the algorithm procedure is repeated K times, leaving a fraction ($1/K$ of the data points) for verification purposes in each iteration. Thus, although the computational performance may seem elevated due to the numerous iterations required, this protocol would permit to lower the variability of the estimated values, decreasing the statistical errors in the obtained results. In addition, employing of the K-fold cross validation method would allow proving the usefulness of the designed model for the entire range of the studied dataset.[30] This test or simulation allows defining the level of generalisation the network is able to provide. The main goal of this test is the evaluation of the competence and applicability of the trained MLP when faced to new data, and was designed only with estimating purposes.[26]

MLP parameter optimisation.

The parameters considered were the input nodes, the hidden neuron number (HNN), the output neuron number and the training function employed.

Selection of input nodes: the eight input nodes correspond with the polarisability values of the structural fragments of the ILs employed (Table SI-1), which means, cation total polarisability value. The polarisability of this ion can be calculated as the sum of its isolated atoms. Additionally, the existing relationship between RI and polarisability is shown in Equation (5).[21]

Even if the aforementioned eight input nodes may seem excessive, the structural information they provide is essential for the MLP to work properly.

HNN optimisation: various tests were developed in order to define the optimal number of neurons in the hidden layer. The hidden neurons are in charge of the converging tendency of the output error function in the learning process. If the HNN is too high, the MLP will over-fit and adjust almost perfectly to the training dataset, leading to a lower applicability and generalisation capability of the network outside this dataset.[26] On the other hand, when the HNN

is too low, the learning capability of the network could be compromised. The HNN tested range was from two to ten, and the results are shown in Table SI-2.

Table SI-2. HNN optimisation results. The selected HNN is shown in bold.

HNN	R^2	MPE (%)
1	0.98	0.29
2	0.99	0.20
3	**0.99**	**0.18**
4	0.99	0.19
5	0.99	0.20
6	0.99	0.22
7	0.99	0.18
8	0.99	0.22
9	0.99	0.20
10	0.98	0.20

Because the R^2 values are very high for all cases, low MPE values were used as the criteria to select the HNN. The lowest MPE values were achieved when three or seven neurons were used. Therefore, three was the HNN chosen because the network complexity is lower and over-fitting is easily avoided.

Output neuron selection: the single output neuron provides the estimated refractive index for a specific imidazolium-based ionic liquid.

Training function selection: the function selected was trainBR (*vide supra*). One of the most relevant characteristics of this function is its capability of determining the number of effective parameters that the network needs to optimise, which is around 24 in this case. Therefore, the use of trainBR is acceptable according to the effective parameter/data point ratio).[28]

Optimisation of neural network parameters: the three parameters which were optimised are Lc, Lcd, and Lci. These MLP parameters have been optimised using a Box-Wilson Central Composite design 2^3 + star points experimental design. This was carried out to select the best values for the network parameters. The Lc and Lcd was tested between 0.001 and 1, and the Lci between 2 and 100. The analysed responses were the MPE (Equation (4)) and correlation coefficient, R^2 (estimated versus target values). In order to select the optimal parameters, a combination of low MPE and high R^2 was defined, giving priority to lower MPE values.

MLP Verification Process

To analyse the level of generalisation of the designed neural network, a verification test was performed employing the obtained data points (Table 1) and, in order to process the results of this procedure, the MPE and the R^2 of the estimated versus target values were studied.

Due to the limited number of values, the *k*-fold cross-validation test was selected.[31] The whole dataset (Table SI-1) was randomly divided into six groups, each composed by 12 data points, creating six MLPs with different training and verification datasets.

Multiple Linear Regression Model Design.

In order to be able to compare the proposed models, six multiple linear regression (MLR) models have been created following the same steps as in their respective MLP models (the database (Table 1) was divided the same way for the ANNs and the MLRs, that is, six different training and verification datasets).

The created MLR model represents the individual contribution of each independent variable (polarisability of the cation, borates, sulfonamides, halides, sulfates, phosphates, dicyanamides, and succinates) to the dependent variable (refractive index value) estimation. The linear relationship between variables is assumed in this model, even when it is known that this linearity cannot be confirmed, but only supposed. However, MLR procedures are not affected by small deviations from linearity to a certain extent.[26] In order to validate this assumption, various statistical parameters have been calculated: correlation coefficient, R^2, and MPE. Additionally, an ANOVA test was performed in order to verify the existence of a statistical correlation between the independent variables and the dependent one, which was confirmed at a 95.0% confidence level.

References

1. S. Seki, S. Tsuzuki, K. Hayamizu, Y. Umebayashi, N. Serizawa, K. Takei and H. Miyashiro, *J. Chem. Eng. Data,* 2012, **57**, 2211−2216.
2. E. Gómez, B. González, N. Calvar, E. Tojo and A. Domínguez, *J. Chem. Eng. Data*, 2006, **51**, 2096-2102.
3. J.G. Huddleston, A.E. Visser, W.M. Reichert, H.D. Willauer, G.A. Broker and R.D. Rogers. *Green Chem.*, 2001, **3**, 156-164.
4. H. Xue, Y. Gao, B. Twamley and J.M. Shreeve, *Chem. Mater*., 2005, **17**, 191–198.
5. W.G. Xu, L. Li, X.X. Ma, J. Wei, W.B. Duan, W. Guan and J.Z. Yang, *J. Chem. Eng. Data,* 2012, **57**, 2177−2184.
6. M. Tariq, P.A.S. Forte, M.F. Costa-Gomes, J.N. Canongia-Lopes and L.P.N. Rebelo, *J. Chem. Thermodynamics*, 2009, **41**, 790–798.
7. A.B. Pereiro, E. Tojo, A. Rodriguez, J. Canosa and J. Tojo, *J. Chem. Thermodyn.*, 2006, **38**, 651-661.
8. W. Xu, L.M. Wang, R.A. Nieman and C.A. Angell, *J. Phys. Chem. B*, 2003, **107**, 11749–11756.
9. K.S. Kim, B.K. Shin and F. Ziegler, *Fluid Phase Equilib*., 2004, **218**, 215-220.
10. A.E. Andreatta, A. Arce, E. Rodil and A. Soto., *J. Solution Chem.*, 2010, **39**, 371-383.
11. M. Wagner, O. Stanga and W. Schroer, *Phys. Chem. Chem. Phys.*, 2004, **6**, 4421-4431.
12. R. Seoane, S. Corderí, E. Gómez, N. Calvar, E. González, E. Macedo and A. Domínguez, *Ind. Eng. Chem. Res.*, 2012, **51**, 2492−2504.
13. O. Russina, L. Gontrani, B. Fazio, D. Lombardo, A. Triolo and R. Caminiti, *Chem. Phys. Lett*., 2010, **493**, 259–262.

14. I.B. Malham and M. Turmine, *J. Chem. Thermodyn.*, 2008, **40**, 718-723.
15. L. Alonso, A. Arce, M. Francisco and A. Soto, *J. Chem. Eng. Data*, 2007, **52**, 2409-2412.
16. R. Kato and J. Gmehling, *Fluid Phase Equilib*., 2004, **226**, 37–44.
17. A. Kumar, *J Solution Chem*., 2008, **37**, 203–214.
18. A. Arce, H. Rodriguez and A. Soto, *Fluid Phase Equilib*., 2006, **242**, 164-168.
19. N. Deenadayalu, K. Ngcongo, T.M. Letcher and D. Ramjugernath, *J. Chem. Eng. Data*, 2006, **51**, 988-991.
20. T. Singh and A. Kumar, *J Solution Chem.*, 2009, **38**, 1043–1053.
21. K. Bica, M. Deetlefts, C. Schröder and K.R. Seddon, *Phys. Chem. Chem. Phys*., 2013, **15**, 2703—2711.
22. F.J. Hernandez-Fernandez, D. Gomez, M. Rubio, F. Tomas-Alonso and G. Villora, *Fluid Phase Equilib.*, 2008, **263**, 190-198.
23. A.B. Pereiro, F. Santamarta, E. Tojo, A. Rodriguez and J. Tojo, *J. Chem. Eng. Data*, **2006, 51**, 952-954.
24. A.B. Pereiro and A. Rodríguez, *J. Chem. Thermodyn.*, 2007, **39**, 978–989.
25. J.C. Cancilla, J.S. Torrecilla and G. Matute, *Curr. Biochem. Eng*., DOI: 10.2174/22127119113019990004.
26. J.S. Torrecilla, J. García, E. Rojo and F. Rodríguez, *J. Hazard. Mater*., 2009, **164**, 182–194.
27. K. Knoerzer, P. Juliano, P. Roupas and C. Versteeg, Innovative Food Processing Technologies: Advances in Multiphysics Simulation. Wiley-Blackwell, Oxford (UK), 2011.
28. H. Demuth, M. Beale and M. Hagan. Neural Network Toolbox for Use with MATLAB® User's Guide. Version 5. Ninth printing Revised for Version 5.1 (Release 2007b); 2007 (online only).
29. A. Oliferenko, P. Oliferenko, J.S. Torrecilla and A. Katritzky, *Ind. Eng. Chem. Res.*, 2013, **52**, 545−546.
30. Y. Bengio and Y. Grandvalet, *J. Mach. Learn. Res.*, 2004, **5**, 1089–1105.
31. A.R. Soleymani, J. Saien and H. Bayat, *Chem. Eng. J*., 2011, **170**, 29–35.