Journal of the Taiwan Institute of Chemical Engineers 000 (2015) 1–16

Contents lists available at ScienceDirect

Journal of the Taiwan Institute of Chemical Engineers

journal homepage: www.elsevier.com/locate/jtice

Prediction of refractive indices of ionic liquids – A quantitative structure-property relationship based model

Mehdi Sattari[a,b,*], Arash Kamari[a], Amir H. Mohammadi[a,c], Deresh Ramjugernath[a]

[a] Thermodynamics Research Unit, School of Engineering, University of KwaZulu-Natal, Howard College Campus, King George V Avenue, Durban 4041, South Africa
[b] Department of Chemical Engineering, Buinzahra Branch, Islamic Azad University, Buinzahra, Iran
[c] Institut de Recherche en Génie Chimique et Pétrolier (IRGCP), Paris Cedex, France

ARTICLE INFO

Article history:
Received 2 November 2014
Revised 27 January 2015
Accepted 1 February 2015
Available online xxx

Keywords:
Refractive index
Ionic liquid
Quantitative structure–property relationship
QSPR
Model

ABSTRACT

In this communication, a reliable Quantitative Structure–Property Relationship (QSPR) model is developed to predict the refractive indices, n_D, of ionic liquids at different temperatures. A dataset comprising 931 experimental data values of refractive index ($\lambda = 589$ nm) for 97 ionic liquids (extracted from the NIST Standard Reference Database) was used to develop and evaluate the model (80% of the data used as a training set and 20% as a test set). In this study, the effects of both anions and cations are considered in the development of the model. Genetic function approximation (GFA) is applied to select the model parameters (molecular descriptors) and develop a linear QSPR model. Statistical analysis of the performance of the model with respect to the dataset indicates an average absolute relative deviation (AARD%) of 0.51, a coefficient of determination (R^2) of 0.935, and a root mean square of error (RMSE) of 1.07×10^{-2}.

1. Introduction

Ionic liquids (ILs) are described as molten salts which are usually liquids at room temperature or close to room temperature (typically below 100 °C) due to their ions being poorly coordinated. ILs have received great interest, mostly in industrial applications, because of their unusual thermophysical properties such as a negligible vapor pressure and high thermal stability [1,2]. Primarily, due to their very low vapor pressure and volatility, most of them are considered as a "green" alternative to volatile organic solvents which are currently commonly used in the chemical industry [3]. ILs can be used in numerous applications such as CO_2 capturing [4–6], catalysis [7,8], extraction and separation processes [9–11], surfactants [12–14], polymer and biopolymer processing [15–17], electroplating [18], solar panels [19,20], fuel cells [21,22], and many more [23–27]. Another reason for their attractiveness and interest shown by researchers is due to the number of permutations in terms of constitution of IL molecules from anion and cation pairs. Consequently, the ionic liquids could theoretically be designed to have a desired thermophysical property by combining different pairs of ions. To explore the "tuneability" and "designability" features of ILs, models have to be developed to relate the thermophysical properties to chemical structure or other physicochemical properties. In this paper, the relationship between the refractive index and chemical structure is determined and an appropriate QSPR model is developed.

Refractive index (n_D) is defined by IUPAC as "the ratio of the speed of light in vacuum to that in a given medium" [28]. Refractive index is a fundamental physical property, especially for optics related research fields, and it is used to verify a material and check its purity, or to determine the concentration of a mixture. It can also provide useful information when studying the forces between molecules or their behavior in solution. It is also related to other properties such as the dielectric constant, density, and surface tension through thermodynamic equations [29,30].

Despite the simplicity of measurement of refractive index, little attention has been given to modeling and even measurement of n_D of ILs. However for other types of compounds, several published works and models are available [31–36]. In the NIST Standard Reference Database #103b [37], experimental data for more than 700 ILs have been collected, but the n_D has been reported for only 97 compounds.

The first model to predict the n_D of ILs was presented by Deetlefs et al. [38] who related the n_D of 9 methylimidazolium based ILs, using Eq. (1), to the surface tension (σ), parachor (P) of the molecule, and molar refraction (R_M).

$$\sigma^{1/4} = \left(\frac{P}{R_M}\right)\left(\frac{n_D^2 - 1}{n_D^2 + 2}\right) \tag{1}$$

All of the parameters of this model should be measured by experiments or correlated by other experimental properties. In case of new ILs, this model thus requires other models to predict the parameters

* Corresponding author.
E-mail addresses: a.h.m@irgcp.fr (A.H. Mohammadi), ramjuger@ukzn.ac.za (D. Ramjugernath).

http://dx.doi.org/10.1016/j.jtice.2015.02.003
1876-1070/

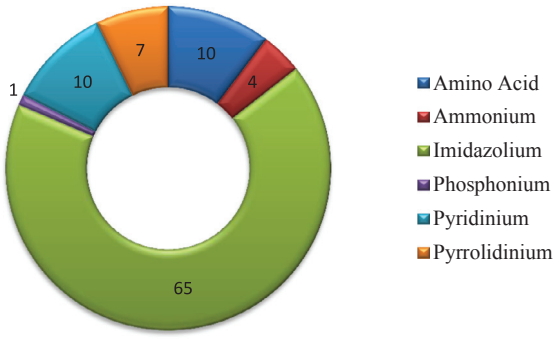

Fig. 1. The number of ionic liquids in different families.

which result in increased errors in the prediction of n_D. In addition, 9 ILs were used to develop this model, and consequently its applicability is limited and its results cannot be generalized. Even with the limited application, the model cannot correlate the data adequately and its average absolute relative deviation (AARD) is close to 8%.

The next model was proposed by Gardas and Coutinho [29] who developed a group contribution method based on experimental data of 24 imidazolium based ILs comprised of 7 different anions. The model had 10 parameters and its AARD was 0.18%, but it could not predict the n_D of ILs with different cations. The same approach was followed by Soriano et al. [39] and Freireet al. [40] using a few new imidazolium ILs, but their models suffered the same problems as that of Gardas and Coutinho.

The aim of this study was to develop a widely applicable model to predict the n_D of ionic liquids using the Quantitative Structure–Property Relationship method.

2. Model development methodology

2.1. Data preparation

The NIST Standard Reference Database #103b was used to extract n_D data of ILs. The database contained refractive index data for 97 unique ILs with 50 different cations and 33 anions. The number of ionic liquids in different families is presented in Fig. 1. In addition, the structures of the constituent cations and anions are available in Tables 1 and 2, respectively. Some of these ILs had multiple data sources, so it was required to screen and refine the data. In case of multiple data sources, the most recent source with the lowest reported uncertainty was selected. Using this approach, 931 experimental data points were extracted from the NIST database. The data covered a temperature range from 283 to 363.15 K and n_D values ranged from 1.3551 to 1.5778. The names of ILs used in this study, together with the temperature ranges, values, and uncertainties of n_D are listed in Table 3. The complete dataset, including the original source of the experimental data are available upon request from the authors.

2.2. Calculation of the descriptors

In order to determine the relationship between the desired property and the constituent cation and anion combinations, the descriptors of both ions were calculated for each IL. The Dreiding Force field, as explained by Chemaxon's JChem, was employed [41] to optimize the 3D chemical structure of each cation and anion. The most stable conformer of each molecule (the one with the smaller energy) was used as the molecular structure. Thereafter, more than 3000 molecular descriptors were calculated by using the optimized structure of all cations and anions, separately. These descriptors belong to 15 classes of descriptors: Constitutional descriptors; Topological indices; Walk and path counts; Connectivity indices; Information indices; 2D autocorrelations; Burden Eigen values; Edge-adjacency indices; Functional group counts; Atom-centered fragments; Molecular properties; topological charge indices; Eigenvalue-based indices; 2D binary fingerprint; 2D frequency fingerprint; and 3D conformational descriptors.

Ultimately, pair correlation was applied to remove the interrelated descriptors. Accordingly, pair of descriptors with a correlation coefficient greater than 0.9 were removed and remaining used to develop the model.

2.3. Subset selection

In the QSPR approach, it is common to divide the experimental dataset into two subsets. The "training set" is used to develop and train the model, and the "test set" to determine the prediction capability of the model for compounds which were not been used in model development. In this paper, K-means clustering was used to select the training and test sets. K-means clustering is a method of cluster analysis, which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean [42]. As a result, 80% of the experimental data was kept for the training set and the remainder used as the test set to validate the model.

2.4. Genetic function approximation (GFA) for model development

Genetic Function Approximation (GFA) is a fusion of two apparently distinct algorithms: Multivariate Adaptive Regression Splines (MARS) of Friedman [43] and Genetic Algorithm (GA) introduced by Holland [44]. It was originally proposed in the pioneering work of Rogers and Hopfinger [45]. Generally, the target of most QSPR studies is to introduce the linear combination of basic functions $\varphi_k(X)$ of the features $X = [x_1, \ldots, x_m]$ in the training dataset of size M:

$$F(X) = b_0 + \sum_{k=1}^{M} b_k \varphi_k(X) \tag{2}$$

The GFA approach works by generating the initial population of equations by a random selection of descriptors. The fitness function used in GFA during the evolution is Friedman's lack of fit (LOF) score, which is described by the following formula:

$$\mathrm{LOF(model)} = \frac{1}{N} \frac{\mathrm{LSE(model)}}{\left(1 - \frac{(c+1+(d\times p))}{N}\right)^2} \tag{3}$$

In this LOF function, c is the number of non-constant basis functions, N is the number of samples in the dataset, d is a smoothing factor to be set by the user, p is the total number of parameters in the model, and LSE is the least square error of the model. Employment of LOF leads to models with better prediction without the problem of over-fitting.

The initial QSPR models are developed by selection of random sets of descriptors from the pool. The next step is genetic recombination or a crossover process conducted on the linear string of descriptors: Two best models in term of their fitness are selected as parents. Then, each parent is split randomly in two parts from a crossing point, and the first substring of the first parent combined with the second substring of the second parent to create two new children. Next, the best new child model replaces the worst model. This process is continued until no significant fitness improvement of the model is observed in the population. For a population of 300 models, 3000–10,000 genetic operations are normally sufficient to achieve convergence [42,46].

3. Results and discussion

While developing the model, it was found that a linear summation of functional groups could produce a precise and easy-to-use model. In addition, the n_D of ILs showed a linear dependency with regard to temperature. Hence, the resultant model was a 9-parameter linear

Please cite this article as: M. Sattari et al., Prediction of refractive indices of ionic liquids – A quantitative structure-property relationship based model, Journal of the Taiwan Institute of Chemical Engineers (2015), http://dx.doi.org/10.1016/j.jtice.2015.02.003