JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

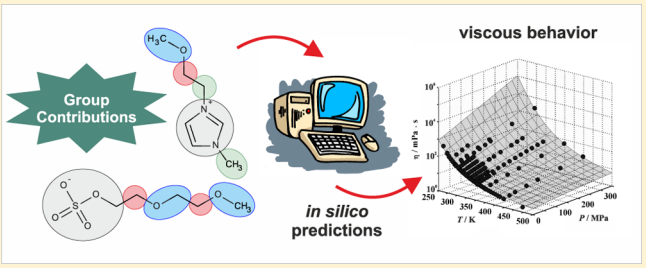
Viscosity of Ionic Liquids: An Extensive Database and a New Group Contribution Model Based on a Feed-Forward Artificial Neural Network

Kamil Paduszyński* and Urszula Domańska

Department of Physical Chemistry, Faculty of Chemistry, Warsaw University of Technology, Noakowskiego 3, 00-664 Warsaw, Poland

Supporting Information

ABSTRACT: A knowledge of various thermophysical (in particular transport) properties of ionic liquids (ILs) is crucial from the point of view of potential applications of these fluids in chemical and related industries. In this work, over 13 000 data points of temperature- and pressure-dependent viscosity of 1484 ILs were retrieved from more than 450 research papers published in the open literature in the last three decades. The data were critically revised and then used to develop and test a new model allowing in silico predictions of the viscosities of ILs on the basis of the chemical structures of their cations and anions. The model employs a two-layer feed-forward artificial neural network (FFANN) strategy to represent the relationship between the viscosity and the input variables: temperature, pressure, and group contributions (GCs). In total, the resulting GC-FFANN model employs 242 GC-type molecular descriptors that are capable of accurately representing the viscosity behavior of ILs composed of 901 distinct ions. The neural network training, validation, and testing processes, involving 90, 5, and 5% of the whole data pool, respectively, gave mean square errors of 0.0334, 0.0595, and 0.0603 log units, corresponding to squared correlation coefficients of 0.986, 0.973, and 0.972 and overall relative deviations at the level of 11.1, 13.8, and 14.7%, respectively. The results calculated in this work were shown be more accurate than those obtained with the best current GC model for viscosity of ILs described in the literature.



INTRODUCTION

Low-melting organic salts, commonly known as ionic liquids (ILs), have been of considerable interest to both scientific and industrial communities conducting their research in different fields of pure and applied chemistry since the beginning of the 1990s.[1−3] This is mainly due to the unusual physical and chemical properties of these chemicals and all of the related advances making ILs competitive against conventional organic solvents.[4] In particular, ILs exhibit extremely low volatility and good thermal stability. Those features significantly reduce exposure of ILs to the environment, and that is why they are frequently perceived as eco-friendly media or solvents, falling under the general rules of green chemistry. The most attractive feature of ILs is their tunability, which allows different properties of ILs to be tailored by a proper selection of the cation and/or anion core, side chains, and task-specific groups. Finally, another characteristic of major importance is their ability to dissolve and partition a great variety of materials ranging from greenhouse gases[5] and inorganic salts[6] to small biomolecules[7] and renewable biomass feedstocks.[8−11]

Viscosity (in particular the dynamic viscosity, commonly denoted by η) is a quantity describing a fluid's internal flow resistance. In liquids, the viscosity is a result of momentum transfer between molecules, which is dominated by collisions and different molecular interactions, including repulsion and hydrogen bonding, short-range van der Waals interactions, and long-range electrostatic forces. From the point of view of chemical technology and engineering, the viscosity is a key input variable for effective and reliable design, analysis, and optimization of a great majority of processes (e.g., mixing, separations) and equipment (e.g., heat exchangers, pipelines, distillation columns) employed in numerous branches of chemical industry.[12,13] In particular, a knowledge of the temperature and pressure dependence of the viscosity (i.e., the $P-\eta-T$ surface) of liquids is of primary importance. Furthermore, because of the large number of substances being used in chemical industry and the lack of experimental data for some new chemicals (e.g., cheaper, safer, "greener" ones), there is still a growing need for reliable and comprehensive mathematical tools that can be used in a purely in silico mode and exhibit enhanced predictive capacity.

In general, ILs are highly viscous fluids. Depending on the type of cation−anion combination, their viscosities can be even several orders of magnitude higher than those of water and widely utilized organic solvents. According to the common opinion of the scientific community working in the field of ILs, this undesirable property poses one of the major obstacles to

Received: April 3, 2014
Published: April 24, 2014

dx.doi.org/10.1021/ci500206u | J. Chem. Inf. Model. 2014, 54, 1311−1324

successful applications of these novel solvents. For this reason, research focused on the discovery of new low-viscous ILs is crucial for the future development of ionic liquid science and consequently cleaner technologies.

Searching for the new ion structures can be performed by means of two methodologies. First, one can systematically (or by "hit-or-miss") check each possible cation−anion combination by direct experimental procedures, namely, syntheses followed by measurements. Because of the huge number of ILs that one can imagine, such an approach has to be time-consuming and thus expensive. Across the pages of this paper we will show that taking into account cations and anions published to date, one can produce about 150 000 binary ILs, while the number of actual salts for which some viscosity data have been reported is on the order of 1500. Therefore, the data available cover only about 1% of all potential ILs. Of course, experimental studies on the remaining 99% would be difficult, and that is why one should seriously consider the second methodology for studying the domain of ILs. In contrast to the first one, it is based on computer-aided molecular design (CAMD) methods.[14] First of all, by using CAMD one can rapidly carry out comprehensive screenings of huge numbers of compounds without performing any experiments. Furthermore, the methodologies and models of CAMD can be built up on the basis of different ideas and mathematical tools. They can be derived either from theoretical ideas having physical foundations (e.g., molecular-based, thermodynamic),[15,16] simple empirical correlations between viscosity and other thermophysical properties,[17,18] generalized correlations,[19] quantitative structure−property relationships (QSPRs),[20−29] or group contribution (GC) methods.[30−34] The models can adopt simple linear or more complex nonlinear relationships between inputs (chemical information) and outputs (properties)[20−24,26−34] as well as sophisticated machine-learning algorithms such as artificial neural networks (ANNs).[19,25,35] In particular, QSPR correlations are developed on the basis of molecular descriptors obtained directly from chemical structure or input data that are experimental in nature. In turn, GC models use schemes to decompose the cation and anion moieties into arbitrarily defined functional groups. The set of structures used in the model development usually covers a certain fraction of the domain and is used to fit some parameters/constants that can eventually be adopted to predict the properties of other compounds. The quality of such calculations and the range of possible chemical structures that can be captured by the model represent its accuracy, predictive capacity, and versatility. All three of these features are equally essential during development of a new model and have to be taken into consideration during the evaluation process.

Since this paper presents another contribution to this particular field, we decided to carry out a short survey of some significant (in our opinion) papers dealing with different methods for predicting the viscosities of ILs.[15−35] However, we emphasize that the following review covers methods for pure ILs only. Some methods for binary and ternary systems composed of ILs and molecular solvents have been also reported, but their discussion is baseless in the context of this work.

Hole theory is the only theoretical approach that has been adopted to represent the dynamic viscosity of ILs to date.[15,16] This simple model presumes that in the microstructure of ILs there are vacancies (called holes) that are present because of thermal fluctuations of the local density. According to the theory, the mobility of ions toward holes of adequate size is directly related to the fluid viscosity. The idea was first applied a decade ago by Abbott[15] to the viscosities of several ILs at room temperature. It was revisited in 2011 by Bandrés and co-workers,[16] who additionally incorporated quantum-chemical calculations to finally obtain very promising predictions (within ±7%) for a series of eight pyridinium-based ILs with various anions. The main disadvantage of the hole theory is that it requires experimental surface tension data.

Slattery et al.[17] attempted to relate viscosity to the molecular volumes of ions determined from X-ray crystal structures. Within the framework of this so-called volume-based approach, they concluded that for a given homologous series of ILs a linear relationship between the natural logarithm of the viscosity and the molecular volume holds. However, the fit coefficients are anion-specific, and what is even worse, they are also distinct depending on the presence of task-specific functional groups attached to the cation side chains. This serious limitation makes the methodology useless when designing new ILs. A similar approach called the scaling exponent method, although having a single adjustable parameter implemented, seems to be a more promising approach of this type, as previously shown by Pádua and co-workers.[18] Finally, Dutt et al.[19] tested a few generalized correlations of viscosity as a function of temperature, $\eta/\eta_0 = f(T/T_0)$, where the subscript "0" corresponds to the reference temperature (323 K). In their study, the authors used simple Arrhenius-type equations as well as a three-layer feed-forward ANN. It was shown that the correlation obtained with the neural network reproduced the experimental data (73 ILs, 654 data points) most accurately (within ±6%).[19] Unfortunately, this method is not predictive at all, since it requires a reference value of the viscosity.

QSPR correlations are another family of predictive tools that have been widely investigated in recent years.[20−31] In this short review, we will not describe all of the published models but will focus only on several QSPR models proposed by the research team of Tochigi, Yamamoto, and co-workers,[20,21,30,31] primarily because of some similarities between those models and the new one proposed in this paper. An essential common feature is that they both employ the GC idea. On the other hand, the remaining models[22−26,28,29,35] were established mainly on the basis of descriptors calculated from molecular topology (as in the paper by Valderrama et al.,[25] where mass connectivity indices for the cation and anion were used) or ab initio methods. Of course, such models may provide some molecular and physical insight into the viscosity of ILs. However, from the point of view of practical utilization, their use is much more difficult because they usually require quantum-chemical software and expertise. Furthermore, many of those correlations do not predict the temperature-dependence of the viscosity, and what is even worse, they are valid only for specific families of ILs (e.g., numerous QSPR equations reported by Chen et al.[24,26,29]). One may expect that those models suffer from poor versatility and a limited range of applications.

The very first model published by Yamamoto[20] in 2006 concerns ILs based on imidazolium, pyridinium, and ammonium cations combined with several popular anions. A product of exponential functions was used to fit experimental temperature-dependent viscosity data for 62 compounds, while seven molecular descriptors derived from the sizes and electrostatic properties of the cations were adopted.[20]

Interestingly, the presence of a given type of anion was treated as a separate molecular descriptor of the IL. The authors noted that such an approach is simple and very effective at the same time. Therefore, the method was revised a year later.[21] In their subsequent work, the authors performed a study of a QSPR model that employs GC-type molecular descriptors exclusively.[30,31] They defined a number of cation cores, with anions as separate groups. Additionally, a few small groups building up cation side chains were introduced (including simple alkyl chains R attached at different positions of the cation ring as well as some other "special" substituents). It has turned out that such an approach exhibits more or less the same predictive capacity as previous QSPR methods. However, QSPR involving only GC descriptors is easy to use because all of the occurrences of groups can be easily identified from the chemical structures of the cation and anion. Additionally, the reverse design of ILs, which exhaustively searches for the structure taking into account the target values of viscosity, was attempted. A proposal to cover a set of candidate structures was made possible by using the model.[30]

A GC approach based on a similar "cation + anion + side chain" idea has also been considered in the papers of Gardas and Coutinho,[32,33] who proposed and developed a predictive method based on the Vogel−Tammann−Fulcher(VTF) equation. The main working equation of the model is of the common VTF form, namely,

$$\ln(\eta/\mathrm{Pa{\cdot}s}) = A_\eta + \frac{B_\eta}{T - T_{0\eta}} \quad (1)$$

The coefficients A_η and B_η in eq 1 are calculated through a linear combination of respective group contributions, while the value of $T_{0\eta}$ is a fixed constant irrespective of the IL. Despite the relatively small number of cations and anions that can be captured by this GC model, it is perceived as the most reliable predictive tool for temperature-dependent viscosities of ILs reported in the literature to date.[36] What is very important is the fact that the Gardas−Coutinho GC model has been systematically tested and extended over the past few years.[37−39] The current version of the model allows experimental viscosity data to be reproduced quite accurately (within ±10%).

Gharagheizi et al.[34] also developed a GC model for viscosity at atmospheric pressure. The mathematical formulation of this model is as follows:

$$\log(\eta/\mathrm{mPa{\cdot}s}) = \eta_0 + \Delta\eta_c + \Delta\eta_a + AT + \frac{B}{T} + CT^2 + \frac{D}{T^2} \quad (2)$$

where $\Delta\eta_c$ and $\Delta\eta_a$ are the cationic and anionic components of the viscosity, respectively, calculated from group contributions, whereas the coefficients A, B, C, D, and η_0 are constants of the model independent of the IL. The authors abandoned the idea adopted by former researchers[30,32,33] and defined 46 small building blocks to describe 443 ILs (including 17 cationic families). According to standard procedures, the whole pool of data was divided into a training set (1336 data points) used to determine the model parameters and a testing set (336 data points) used to evaluate the model. It was shown that accuracies of correlation and prediction are similar, confirming the internal consistency of the model itself as well as the experimental data used during its development. The average absolute relative deviation values reported for the whole data set were at the level of 6.3%. At first sight, this is a surprisingly good result, taking into account the diversity of both the training and testing sets and the relatively small number of model parameters. However, such a low value is misleading, since the calculated deviations refer not to the differences between the calculated and experimental viscosities but rather to differences between their decimal logarithms. The real overall deviation in the actual viscosity computed on the basis of supplementary material provided by the authors is much higher (around 31%). Taking a closer look at this model, we also observed that it suffers from a very unclear and awkward definition of groups that in our opinion is unsuitable for such complex fluids as ILs. On the basis of this definition, some fragments of cations/anions are omitted when assigning the groups. For example, there is no group that can be assigned to the negatively charged nitrogen atom located at the center of common amide-based anions. The same can be noticed for cations. For example, 1,1-dimethylimidazolium cation is represented only by two groups defined as hydrogen atoms attached to carbon side chains (!), as if the rest of molecule did not exist. The authors did not provide any explanation and justification for that, and hence, a typical user (a chemist or a chemical engineer) may be confused when trying to apply the model.

This paper is a continuation of our study on GC methods for different thermophysical properties of ILs. In 2012, we reported a method for predicting density[40] that was recently acknowledged as the best currently available method of this type.[36] Here we propose another predictive method, this time for the estimation of viscosities of pure ILs. To elaborate the model, we compiled and critically revised a comprehensive set of literature data on viscosity as a function of temperature and pressure for 1484 ILs. The approach employed in the new model is based on the GC strategy and an ANN-based machine-learning algorithm. To our best knowledge, this is the very first paper in which such a combination has been considered for the viscosities of pure ILs. Nevertheless, it has recently been shown in the literature[41−44] that such a GC-ANN methodology allows general and satisfactory relationships between the functional groups and the various physical properties of a large number of organic and inorganic compounds to be established. That is why we attempted to transfer it to prediction of the viscosities of ILs. Moreover, we underline that accounting for the pressure dependence of the viscosity is also a novelty of this work.

It is noteworthy that an ANN was applied by Suzuki et al.[45] to represent temperature-dependent viscosity data of several hundred organic chemicals. Their methodology was based on 10 physicochemical and structural molecular descriptors such as molar refraction, magnetic susceptibility, critical temperature, and energy of vaporization and indicator variables representing the presence or absence of some relevant functional groups such as OH and C≡N.[45] In the case of ILs, such an approach is not possible to apply because the corresponding data are lacking. In our opinion, GC is the only methodology that can be used to model ILs at this stage of research and development of these interesting fluids.

In the rest of this paper, we present in detail the model development process, including the computational tools utilized, the database, the data evaluation procedure, and the functional group assignments. Then we present, discuss, and analyze the results of viscosity calculations. Finally, the calculations with the new GC method are compared with those obtained using the Gardas−Coutinho model.[33]

DEVELOPMENT OF THE NEW MODEL

The main objectives of the new model development were (1) to achieve an extended range of applicability comprising a wider diversity of ILs and (2) to increase the accuracy of predictions over broader ranges of temperature and pressure compared with other similar GC methods described in the literature. To achieve these goals, a comprehensive experimental data compilation was used, a new set of functional groups was defined, and a new way of treating $P-\eta-T$ data by using ANN was proposed.

Database. Viscosity data for 1484 ILs were used in the model development. The data included 2271 data sets and 13 470 data points and were extracted from more than 450 articles published in the open literature in the last three decades (1984–2014). A detailed summary of the database is given in the Supporting Information.

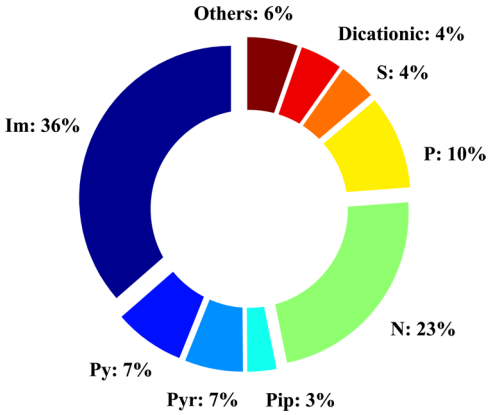


Figure 1. Viscosity database used in the development of the new model. Classification with respect to the cation type is shown. A key is given in footnote *a* of Table 1.

It can be seen in Figure 1 and Table 1 that a great variety of families of ILs (with respect to their cations) are contained in the database. As in the case of the density database published by us previously,[40] imidazolium-based ILs predominate remarkably. A number of ILs based on novel cations (e.g., quinolinium, isoquinolinium, quinolizinium, azepanium, thiolanium, or thianium) together with 66 dicationic ILs were included as well. In total, 722 distinct cations were collected. In the case of the anions, 179 different structures were identified, including inorganic as well as organic ones. In particular, halides, nitrates, chlorates, borates, phosphates, sulfates, sulfonates, imides, methides, metal complexes, carboxylates, amino acids, acesulfamates, and saccharinates are the anions that mostly occur in the database. All of the cationic and anionic structures and their preferred IUPAC names are listed in the Supporting Information.

The database was organized in the same manner as in our previous work.[40] For each IL, an independent data file was created. Within the file, data sets from different literature sources were collected and described, including information on sample preparation and purification, supplier, and purity (the water content was published by a great majority of researchers) as well as on the experimental technique used to determine the viscosity. Such information is crucial from the point of view of data revision and evaluation. Finally, data points of viscosity were listed, including corresponding conditions of temperature and pressure. A pressure of 0.1 MPa was assigned to data points measured at "ambient" or atmospheric pressure or in cases where an actual value of the pressure was not specified.

Data Revision and Evaluation. The collected experimental $P-\eta-T$ data reported by different authors usually exhibit large discrepancies. It can be speculated that the observed scatter might be related to impurities in the samples (mainly water and halide content), purification procedures, or the experimental techniques employed to measure viscosity. Therefore, all of the available experimental data were subjected

Table 1. Summary of the Viscosity Database Used in This Work: Numbers of Ionic Liquids, Data Sets, and Data Points; Temperature, Pressure, and Viscosity Data Ranges; and Classification with Respect to the Cation Type

	Number of data collected					
IL family[a]	ILs	sets[b]	points[b]	T (K)	P (MPa)	η (mPa·s)
Im	537	1013 (185)	6071 (607)	253–573	0.1–350	0.841–257000
Py	109	169 (12)	1703 (64)	273–373	0.06–65.9	5.49–19610
Quin	8	8 (0)	45 (0)	298–363	0.1–0.1	16.1–2388
Pyr	97	188 (37)	1348 (204)	258–573	0.1–150	1.14–12220
Ox	12	12 (0)	12 (0)	298–298	0.1–0.1	90–731
Pip	49	66 (3)	300 (3)	283–413	0.1–0.1	5.76–1669.87
Mo	30	31 (0)	281 (0)	283–368	0.1–0.1	5.84–3278.8
Azp	13	20 (0)	156 (0)	293–368	0.1–0.1	6.27–798
Guan	22	23 (1)	182 (1)	293–353	0.1–0.1	8.76–2260
N	336	431 (38)	1712 (103)	268–573	0.1–0.1	0.8–28508
P	147	179 (13)	1103 (78)	253–433	0.1–0.1	0.41–123500
S	58	61 (3)	388 (3)	253–363	0.1–0.1	4.2–2287
dicationic	66	70 (1)	169 (2)	297–353	0.1–0.1	61–15540
overall	1484	2271 (293)	13470 (1065)	253–573	0.06–350	0.41–257000

[a]Key: Im, imidazolium (including benzimidazolium, pyrazolium, and tetrazolium); Py, pyridinium; Quin, quinolinium (including isoquinolinium and quinolizinium); Pyr, pyrrolidinium (including pyrrolidonium and pyrrolinium); Ox, oxazolidinium; Pip, piperidinium (including piperazinium); Mo, morpholinium; Azp, azepanium (including ε-caprolactam-based cations); N, ammonium; P, phosphonium; S, sulfonium (including thiolanium, thianium, thiazolium, and benzthiazolium). [b]Numbers of data sets and data points excluded from the model development process are given in parentheses.

dx.doi.org/10.1021/ci500206u | J. Chem. Inf. Model. 2014, 54, 1311–1324

into the following evaluation procedure before the new GC method was developed.

(1) All of the data reported at elevated pressures (P > 0.1 MPa) were accepted. First of all, this choice was made because the number of such data points was relatively small (merely 25 ILs, 42 data sets, and 1443 data points). Moreover, we believed that rejecting some part of these data would negatively affect predictions at high pressure.

(2) In the case of ILs for which a number of data sets at ambient pressure (P = 0.1 MPa) were available, all of the data points were simultaneously fitted using the following equation:

$$\eta(T, P = 0.1\ \text{MPa}) = \frac{A}{\sqrt{T}} \exp\left(\frac{B}{T - C}\right) \quad (3)$$

For this purpose, the *lsqnonlin* function in MATLAB (R2013a; The Mathworks, Inc.), which employs the Levenberg–Marquardt algorithm,[46,47] was used. The determination of the coefficients A, B, and C in eq 3 was repeated several times, and the results obtained were revised after each calculation. Between consecutive calculations it was possible to reject and reinclude whole viscosity data sets taken from individual literature sources. The calculations were repeated until a final fit was obtained wherein the maximum relative deviation between the retained experimental values and the smoothed values was in the range from ±5 to ±10% while average absolute deviation was at the level of 3%.

(3) In the case of ILs for which a few single-data-point sets were provided, the purity of the samples used by the authors as well as the experimental method used to measure the viscosity were the essential criteria of rejection or inclusion of a particular data set. In some cases, there was a possibility to analyze and check the quality of the data by comparing them to data for similar compounds (e.g., within a homologous series).

(4) For ILs for which only a single data point was reported (689 ILs, i.e., almost a half of the ILs stored!), an assessment of the data quality was not actually possible. Nevertheless including such data significantly extended the range of applications of the GC model, and hence, we decided to accept all of them as reported.

Figure 2 shows a final result of the described selection procedure (particularly step 2) for the representative IL 1-butyl-3-methylimidazolium tetrafluoroborate, $[C_4C_1Im][BF_4]$. In total, 39 data sets and 234 data points at P = 0.1 MPa can be found in the viscosity database. However, only 20 data sets and 157 data points were accepted to be used in the next steps of GC model development. The relative deviation between the accepted data and those fitted with eq 3 varies from −8.8 to +9.5% while the overall absolute deviation is 2.5%.

The total numbers of 2271 data sets and 13 470 experimental data points contained in the database were reduced to values of 1978 and 12 405, respectively, after the evaluation procedure. A complete list of accepted and rejected data sets is given in the Supporting Information.

Functional Groups. The treatment of ILs in terms of the GC method proposed in this paper is the same as in our previous contribution concerned with density.[40] However, the number of functional groups is higher because of the greater number and diversity of structures deposited in the viscosity database.

In terms of the presented model, each IL is composed of three basic fragments: the cation core, the anion core, and substituents attached to the cores. This idea has been applied in other literature GC models as well.[30−33] The total set of

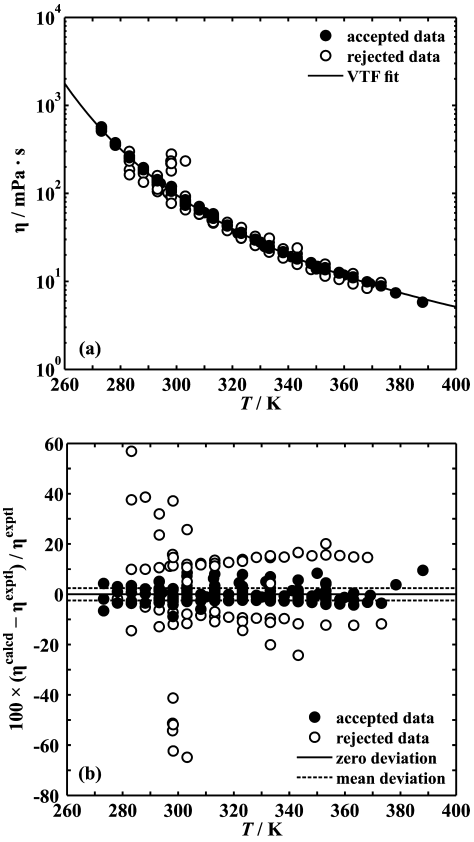


Figure 2. Atmospheric-pressure viscosity data for $[C_4C_1Im][BF_4]$: (a) experimental viscosity vs temperature and the corresponding VTF correlation (see eq 3) with A = 2.7156 mPa·s·K$^{1/2}$, B = 836.34 K, and C = 169.69 K; (b) relative deviations between the experimental and correlated data. A full list of references for accepted and rejected experimental data can be found in the Supporting Information.

functional groups defined in this work contains 242 units divided into three main classes: 59 cation core functional groups, 92 anions/anion core functional groups, and 91 substituted functional groups. We stress that our methodology is different than that proposed by Gharagheizi et al.,[34] who used much smaller and thus chemically simpler groups. There is no doubt that the use of small and simple groups is more convenient. However, it may adversely affect the performance of the GC method. It is especially significant for ILs, which are very complex fluids with different kinds of molecular interactions. Therefore, larger (and hence more numerous) groups accounting for so-called "proximity effects" are preferred for ILs. Moreover, the groups defined should allow differentiation among positional isomers, especially in ILs based on aromatic cations. In our previous paper[40] we successfully addressed all of these points, and that is why we recall them in this paper.

The chemical structures of the cation cores are shown in Figure 3. As can be seen, more than 20 families of ILs were included. To take into account the positional isomerism of aromatic ILs (mainly those based on imidazolium and pyridinium cations but also some pyrrolidinium and piperidinium cations), special cores were defined for the respective isomers. Moreover, unique cationic cores for protic ILs were distinguished (not shown in Figure 3).

The set of anionic functional groups comprises a variety of both organic and inorganic anions. In the case of small anions

dx.doi.org/10.1021/ci500206u | J. Chem. Inf. Model. 2014, 54, 1311−1324

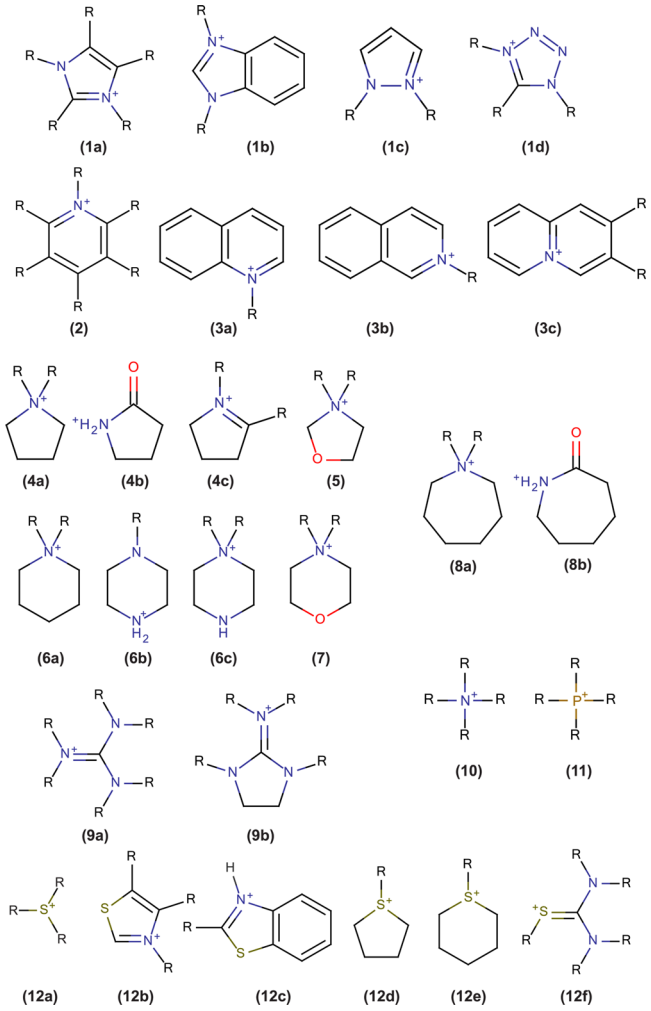


Figure 3. Chemical structures of cations constituting ILs taken into consideration in this study. Key: (1a) imidazolium, (1b) benzimidazolium, (1c) pyrazolium, (1d) 1,4,5-tetrazolium, (2) pyridinium, (3a) quinolinium, (3b) isoquinolinium, (3c) quinolizinium, (4a) pyrrolidinium, (4b) pyrrolidonium, (4c) pyrrolinium, (5) oxazolidinium, (6a) piperidinium, (6b, 6c) piperazinium, (7) morpholinium, (8a) azepanium, (8b) ε-caprolactam-based cation (2-oxoazepanium), (9a) guanidinium, (9b) cyclic guanidinium, (10) ammonium, (11) phosphonium, (12a) sulfonium, (12b) thiazolium, (12c) benzthiazolium, (12d) thiolanium, (12e) thianium, (12f) thiourea-based cations.

or anions for which experimental data are not abundant, the whole anion is considered to be an individual group. This is the same protocol as in the GC model of Gardas and Coutinho.[32,33] For larger anions or anions for which experimental data for a given homologous series are available, an appropriate anion core was identified and assigned. For example, there is a significant amount of viscosity data for ILs based on *n*-alkylsulfate anions $[C_kSO_4]$ (where $C_k \equiv C_kH_{2k+1}$, k = 1−8). Therefore, we defined two groups constituting this family of ILs, namely, the anion group $[CH_3SO_4]$ corresponding to methylsulfate anion and the anion core group $[CH_2SO_4]$ assigned for ILs with a longer side chain attached to the sulfate group. We distinguished those two groups instead of one (e.g., $[SO_4]$) in order to highlight the different properties of methyl and methylene groups attached directly to the oxygen atom of the sulfate anion. Similar reasoning was carried out for several other series of anions (e.g., carboxylates and dialkylphosphates).

The set of substituted groups includes standard building blocks of side chains attached to the cores: CH_3, CH_2, CH, CH_2═CH, OH, C≡N, $CH_2C(O)O$ (ester), aromatic rings, and so on. The proposed set of units allows a representation of aromatic compounds and enables a differentiation between groups occurring in cyclic and acyclic structures (e.g., the model distinguishes between the CH_2 group in a linear alkyl chain, the c-CH_2 group incorporated in aliphatic rings, and the aC−CH_2 group attached to an aromatic ring). Moreover, we defined individual units for the groups substituted directly on heteroatoms occurring in cation rings because it is well-known that the presence of a heteroatom (N, P, S) modifies the structural characteristic and electric charge distribution of groups in its close proximity. For example, a separate group, denoted as aN−CH_2, was defined for a methylene group connected directly to a nitrogen atom incorporated into an aromatic ring. Such a group occurs in a variety of cations, including imidazolium, pyrazolium, triazolium, pyridinium, pyridazinium, quinolinium, and isoquinolinium.

A full list of groups defined within the proposed model (including their abbreviations and chemical structures) is recapitulated in the Supporting Information. All of the groups were defined in such a manner that they describe the entire cation or anion. Moreover, when the groups are assigned, it is required that no group is allowed to overlap any other group (no atom of the molecule can be included in more than one group). Finally, if the same fragment of a given compound is related to more than one group, the heaviest group is chosen to represent it. Such an approach of defining the functional groups for GC model is consistent with the definition of so-called first-order groups according to well-known rules proposed in 2001 by Marrero and Gani.[48]

An exemplary group assignment is shown in Figure 4 for a representative IL, namely, 1-(2-methoxyethyl)-3-methylimidazolium 2-(2-methoxyethoxy)ethylsulfate. The cation of this IL is composed of the imidazolium cation core substituted at positions 1 and 3 ($[Im_{13}]^+$; group no. 4) with a methyl group attached directly to a nitrogen atom in the aromatic ring (aN−CH_3; no. 234), a methylene group attached directly to a nitrogen atom in the aromatic ring (aN−CH_2; no. 235), a

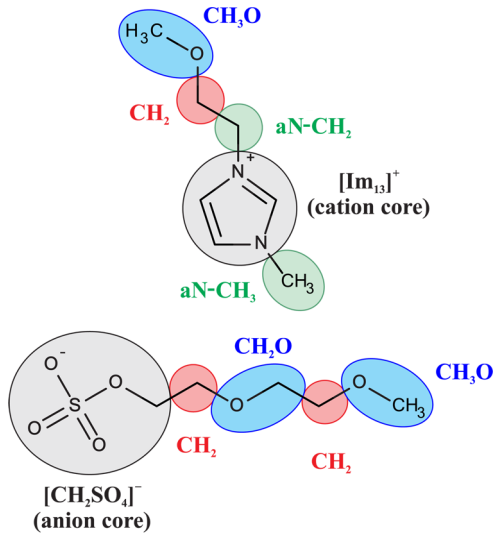


Figure 4. An exemplary group assignment in terms of the proposed GC method.

dx.doi.org/10.1021/ci500206u | J. Chem. Inf. Model. 2014, 54, 1311−1324

methylene group (CH_2; no. 153), and a methoxy group (CH_3O; no. 175). It is noteworthy that the $CH_2CH_2OCH_3$ side chain of the cation could also be decomposed as follows: (1 × aN−CH_2) + (1 × CH_2O) + (1 × CH_3). However, the former assignment contains the heavier group CH_3O, and that is why it is correct in terms of the adopted rules. Finally, the anion consists of a sulfate core ($[CH_2SO_4]^-$; no. 95), a methyleneoxy group (CH_2O; no. 176), a methoxy group (CH_3O; no. 175) and two methylene groups (CH_2; no. 153).

Application of the Feed-Forward ANN. In order to develop the new GC model for the $P-\eta-T$ behavior of ILs, a relationship between group contributions, temperature, pressure, and viscosity had to be established. In this work, ANN was employed for that purpose. Of course, we tried to capture all of the viscosity data by using simpler methods, in particular those proposed by other authors.[30,32−34] Those methods, however, turned out to have insufficient flexibility to accurately reproduce such a huge database as that presented herein.

ANN is a machine-learning algorithm inspired by biological central nervous systems, especially by the brain.[49,50] The main similarities between biological and artificial networks are the layered structure of the system of units of which they consist as well as the fact that these units perform functions collectively and in parallel. Each ANN consists of simple processing units called nodes (or artificial neurons, or simply neurons) that are connected with each other to form a network. In general, artificial neurons from a given layer collect the input data from the previous layer, process them using selected weight and transfer functions, and then feed them as the final output or the input for the next layer of neurons. The adjustable model parameters of each ANN, called weights and biases, are assigned to each node and node−node connection. Generally, they are fitted to some target values of the output variable of interest during the so-called training process. In particular, several types of ANNs can be distinguished depending on the network architecture.[50]

The feed-forward ANN (FFANN) is a network that has been widely used to approximate very complex dependencies between numerous variables and also in purely chemical and chemical engineering applications.[41−45] In this type of network, no loops can be found (i.e., the information moves forward only). The input goes through the subsequent hidden layers of nodes and finally to the output nodes. Most frequently, two-layer FFANNs are used to estimate a relationship between an input vector containing R elements, henceforth denoted as $\mathbf{x} = [x_1, ..., x_R]^T$ (where ^{T} stands for transpose), and the scalar output, henceforth denoted as y. The first layer is the hidden layer. Its size (the number of nodes, S) can be adjusted arbitrarily. The second layer is the output layer, where the final output y is generated, and the number of nodes in this layer is equal to the size of the output (in this case 1). It is worth noting that some authors refer to the described "$R-S-1$" network as a three-layer network, treating the inputs as the first layer. In this work, we do not follow that designation.

In the following subsections, we briefly describe how the inputs are processed by the FFANN to obtain output, how the weights and biases are determined (i.e., how the network is trained), and finally how the FFANN is implemented to model viscosity in terms of the GC approach. All of the presented ANN calculations were carried out using the Neural Network Toolbox of the MATLAB integrated development environment.[50]

Input/Output Processing in the FFANN. First, the $R \times 1$ input vector $\mathbf{x}$ is passed through the hidden layer (denoted as 1), and the intermediate output vector $\mathbf{o}^1 = [o_1^1, ..., o_S^1]^T$ is produced. The rule according to which each neuron produces a single component o_j^1 of the vector $\mathbf{o}^1$ reads as

$$o_j^1 = f_j^1(n_j^1) \qquad (j = 1, ..., S) \tag{4}$$

where f_j^1 stands for the transfer (or activation) function of neuron j in layer 1 and n_j^1 is the net input for neuron j in layer 1, which is calculated using the common dot-product weight function:

$$n_j^1 = \sum_{i=1}^{R} w_{ji}^1 x_i + b_j^1 \qquad (j = 1, ..., S) \tag{5}$$

where the symbols w_{ji}^1 and b_j^1 denote weights and biases associated with every input-node connection and neuron, respectively. Usually, the same transfer function is used for all of the neurons within a given layer (i.e., $f_1^1 = f_2^1 = \cdots = f_S^1 \equiv f^1$). In vector notation, eqs 4 and 5 can be written in more compact form, namely,

$$\mathbf{o}^1 = \mathbf{f}^1(\mathbf{W}^1\mathbf{x} + \mathbf{b}^1) \tag{6}$$

where $\mathbf{W}^1$ is the $S \times R$ matrix of weights and $\mathbf{b}^1$ is the $S \times 1$ vector of biases. By analogy, the output $\mathbf{o}^2$ is produced by the second layer (the output layer). In this case, however, the nodes are fed by the output vector $\mathbf{o}^1$, so the general formula for $\mathbf{o}^2$ is

$$\mathbf{o}^2 = \mathbf{f}^2(\mathbf{W}^2\mathbf{o}^1 + \mathbf{b}^2) \tag{7}$$

In particular, for the two-layer network yielding a scalar output ($o_1^2 = y$), the dimensions of the weight matrix $\mathbf{W}^2$ are $1 \times S$, while the bias vector is a scalar as well ($\mathbf{b}^2 = b_1^2$). Thus, the overall rule for the $\mathbf{x} \rightarrow y$ transformation is

$$y = f^2[\mathbf{W}^2\mathbf{f}^1(\mathbf{W}^1\mathbf{x} + \mathbf{b}^1) + b_1^2] \tag{8}$$

Equation 8 requires two weight matrices ($\mathbf{W}^1$ and $\mathbf{W}^2$) and two bias vectors ($\mathbf{b}^1$ and $\mathbf{b}^2$) containing a total of $S(R + 2) + 1$ parameters. Moreover, the activation functions f^1 and f^2 have to be specified.

Network Training Procedure. After the FFANN architecture has been specified, all of the weights and biases must be determined. This is done in the so-called supervised training (or learning) process. First of all, the process requires a set of examples of network inputs $\mathbf{x}_i$ and target outputs t_i ($i = 1, ..., N$) representing proper network behavior. In general, the training process involves tuning the values of the weights (the elements of $\mathbf{W}^1$ and $\mathbf{W}^2$) and biases (the elements of $\mathbf{b}^1$ $\mathbf{b}^2$) to minimize the network performance function. Usually, the mean-square error (MSE) between the desired and calculated outputs, defined as

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(t_i - y_i)^2 \tag{9}$$

is employed for that purpose. Several algorithms that can be used to optimize the MSE given in eq 9 have been described. The most popular and simultaneously the most effective one (particularly in function-fitting problems) is the Levenberg−Marquardt (LM) optimization algorithm[46,47] with gradients and Jacobians calculated using a technique called error back-propagation.[49] Each training algorithm (including LM) is

usually combined with a cross-validation method called "early stopping".[50] In this method, all of the available data input−target samples ($\mathbf{x}_i$, t_i) are (mostly randomly) divided into three disjoint subsets. The first subset (usually 80−90%) is the training set, which is used for updating the network weights and biases in the MSE minimization. The second and third subsets (5−10% for each) are the validation and testing sets. The MSE on the validation set is monitored during the training process. Normally, the validation-set MSE decreases during the initial stage of training, as does the training-set MSE. However, when the network begins to overfit the data, the error for the validation set typically starts to rise. When the validation error increases for a specified number of iterations, the training is stopped, and the weights and biases at the minimum of the validation error are returned as the optimum vales. On the other hand, the learning process can be stopped if too many iterations are run or if the training-set MSE becomes lower than a predefined goal value. Finally, the testing set is not employed during training or validation. It can be used to evaluate the predictive capacity (generalization accuracy) of the model as well as to compare different models.

Implementation. For the physical problem raised in this paper, the FFANN input vector $\mathbf{x}$ and the output variable y are defined as follows:

$$x_1 = T/K \tag{10a}$$

$$x_2 = P/\mathrm{MPa} \tag{10b}$$

$$x_{2+i} = n_i \qquad (i = 1, ..., N_G) \tag{10c}$$

$$y = \ln(\eta/\mathrm{mPa{\cdot}s}) \tag{10d}$$

where n_i denotes the number of occurrences of functional group i in the cation's or anion's chemical structure and the number of defined groups is $N_G = 242$. The full list of ions and corresponding values of n_i is provided in the Supporting Information. The hyperbolic tangent and identity functions were adopted as the activation functions f^1 and f^2 for the hidden and output layers, respectively (see eq 8):

$$f^1(t) = \tanh t = \frac{e^t - e^{-t}}{e^t + e^{-t}} \tag{11a}$$

$$f^2(t) = t \tag{11b}$$

In the following text, we refer to eqs 10 combined with eqs 8 and 11 as the GC-FFANN model. To train the network, the LM algorithm coupled with the early stopping method was utilized. The number of failed validation checks permitted (i.e., the number of epochs with increasing validation-set MSE) was fixed at 10. The learning process was allowed to be performed for 1000 iterations at most, and no optimization goal was specified (i.e., the demanded training-set MSE value was 0). Before the actual network training, some basic data preprocessing steps were performed. Duplicates of input−target pairs (i.e., exactly the same P−η−T data points reported by different authors; in total 149 data points) were excluded from the data pool. Furthermore, both the input and target data were normalized to fall in the range $[-1, 1]$; this is a common practice in ANN modeling because network training based on such scaled data is more effective and converges much faster. Afterward, 12 256 accepted and preprocessed viscosity data points were divided into the training set, validation set, and testing set, made up of 11 031 (90%), 612 (5%), 613 (5%) samples, respectively. In particular, all of the data points for ILs having in their structures "rare" functional groups (defined as those having fewer than 20 occurrences in the database; 612 data points) were included in the training set. This was necessary to avoid underfitting of the model given in eq 8. The remaining data points were distributed randomly.

Finally, the size of the hidden layer, S, must be specified to run the FFANN. Besides some "rules of thumb", there are no unambiguous formulas for selecting the optimum value of S. In general, it is argued that the optimum value depends in a complex way on the dimensionality of the input and output units, the number of training cases, the amount of noise in the targets, the complexity of the function or classification to be learned, the architecture, the type of hidden unit activation function, the training algorithm, and so on. In this work, we followed Klimasauskas,[51] who advanced the rule that the ratio of the number of training cases to the number of network parameters for the two-layer FFANN, defined as

$$r = \frac{N_{\mathrm{train}}}{S(R + 2) + 1} \tag{12}$$

should be at least 5. Since $R = 244$ and $N_{\mathrm{train}} = 11\,031$, it can easily be shown that $r_{\mathrm{min}} = 5$ corresponds to $S_{\mathrm{max}} = 9$. The optimum value of S was determined by trial and error by changing its value from 1 to 12. Networks with $S > S_{\mathrm{max}}$ were trained just to check whether they indeed overfit the input data, producing high generalization errors. Simplicity of the network (i.e., as low a value of S as possible) and overall accuracy of the training/validation/testing results were the criteria imposed to find the optimum architecture.

RESULTS AND DISCUSSION

FFANN Architecture. The training results for a series of 244−S−1 two-layer FFANNs having different numbers of hidden units are presented in Figure 5a, where the optimized MSEs for the training, validation, and testing sets are plotted as functions of S. For each network architecture, the weights and biases were initialized using the Nguyen−Widrow algorithm.[52] As can be seen, the optimum network architecture corresponds to $S_{\mathrm{opt}} = 7$, for which the best performance of training/validation/testing was achieved. Fortunately, $S_{\mathrm{opt}} < S_{\mathrm{max}}$, and thus, Klimasauskas' rule[51] mentioned in previous section holds. In fact, the use of seven hidden nodes corresponds to 1723 weight and bias values and therefore $11031/1723 \approx 6$ training-set data points per parameter. Furthermore, for FFANNs with $S > S_{\mathrm{opt}}$, no significant improvement in the training-set MSE was observed, while the validation- and testing-set MSEs tended to increase.

Figure 5b shows the training process of the selected 244−7−1 network. The minimum in the validation-set MSE was reached after 37 epochs, and the optimization process was stopped after 10 subsequent validation checks. It is noteworthy that the error in the test set did not reach a global minimum. If it did at a significantly different iteration number than the validation-set error, this would be an indicator of a poor division of the data set.[50]

The final values of training-, validation-, and testing-set MSEs are 0.0334, 0.0595, and 0.0603, respectively. The full list of optimized weights and biases is given in the Supporting Information. Additionally, the Supporting Information includes a *mat* file containing all of the relevant information regarding the final FFANN as well as a very simple MATLAB code that

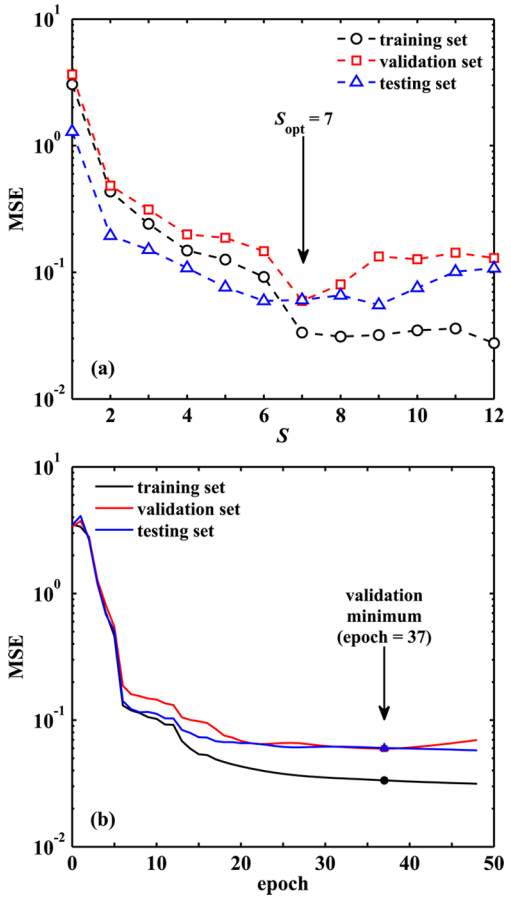


Figure 5. Results of optimization of FFANN weights and biases: (a) influence of the number of hidden nodes (S) on the network performance (expressed as MSE; see eq 9) and (b) training process for the optimum 244−7−1 network.

can be utilized by the reader in his/her own calculations. The instructions for running the program are presented in the Appendix.

Performance of the Proposed Model. In Figure 6a−c, the results obtained using the 244−7−1 FFANN are presented as parity plots comparing experimental (denoted as "exptl") against calculated (denoted as "calcd") viscosities for the training, validation, and testing sets, respectively. In the case of the training set, the calculated viscosity is randomly distributed along the diagonal ($\ln \eta^{calcd} = \ln \eta^{exptl}$). As shown in Figure 6a, the expression given in eq 8 is flexible enough to fit a huge number of diverse data points. The results of predictions (i.e., validation and testing) shown in Figure 6b,c are of similar quality, which confirms the reliability of the mathematical approach used to represent the data as well as the adequacy of the group contributions defined in this work.

To illustrate the accuracy of the model, percent deviations between the experimental and calculated data are given in Figure 7 as histograms. It can be seen that most of the experimental viscosity data were reproduced by the proposed GC-FFANN model within ±25% (approximately 91, 86, and 87% of the training, validation, and testing samples, respectively). More accurate results, within ±5%, were obtained for about 40% of the points in all three sets. Despite the fact that the plots in Figure 7 cover the ±100% deviation range only, there are also some outliers present having deviations higher than +100%. In particular, there are 62 such points in the training set (<0.5%) and merely four and six such points (<1%) in the validation and testing sets, respectively.

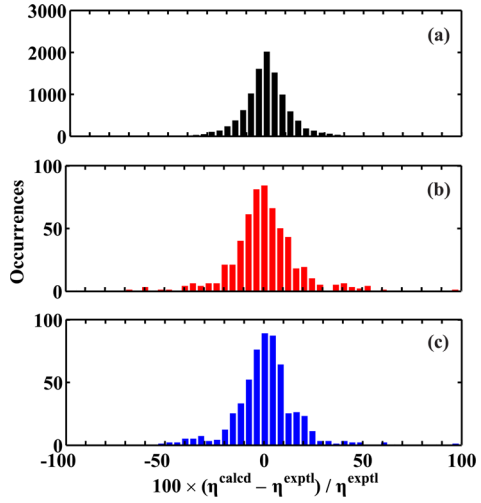


Figure 7. Distributions of relative deviations between the experimental and GC-FFANN-calculated viscosities: (a) training set; (b) validation set; (c) testing set.

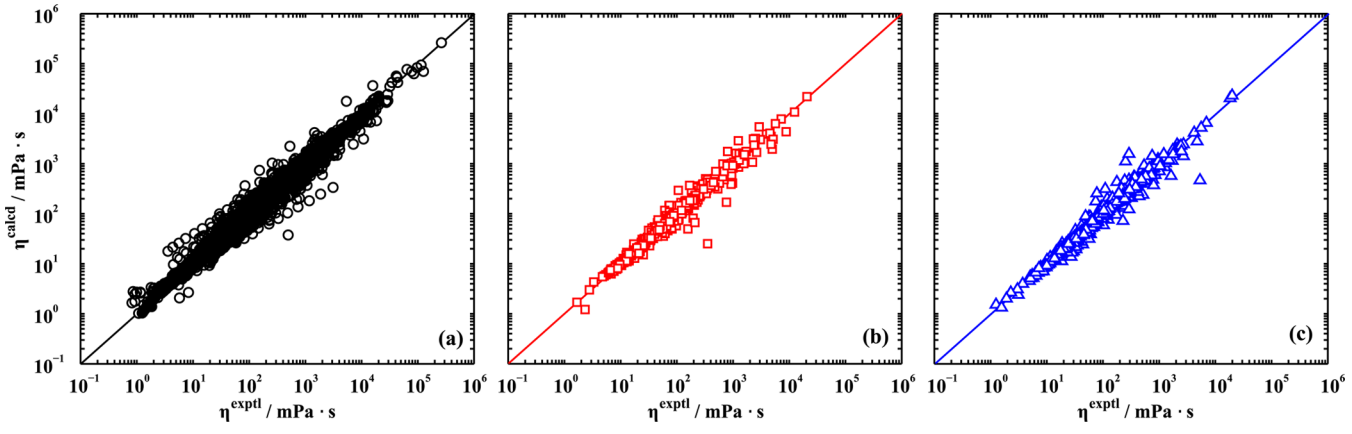


Figure 6. GC-FFANN-calculated viscosities vs experimental viscosities of ILs: (a) training set; (b) validation set; (c) testing set. Solid lines designate the diagonals $\eta^{calcd} = \eta^{exptl}$.

dx.doi.org/10.1021/ci500206u | J. Chem. Inf. Model. 2014, 54, 1311−1324

Overall quantitative analysis of the model accuracy was made in terms of the five statistical measures: average absolute relative deviation (AARD), average absolute deviation (AAD), bias, sum of signs ("±"), and squared correlation coefficient (R^2). The former two measures (AARD and AAD) quantify the magnitude of the scatter, while the bias and "±" give some information on the symmetry of the deviation distribution. They are defined as follows:

$$\mathrm{AARD} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\eta_i^{\mathrm{calcd}} - \eta_i^{\mathrm{exptl}}}{\eta_i^{\mathrm{exptl}}}\right| \times 100\% \tag{13a}$$

$$\mathrm{AAD} = \frac{1}{N}\sum_{i=1}^{N}|\eta_i^{\mathrm{calcd}} - \eta_i^{\mathrm{exptl}}| \tag{13b}$$

$$\mathrm{bias} = \frac{1}{N}\sum_{i=1}^{N}(\eta_i^{\mathrm{calcd}} - \eta_i^{\mathrm{exptl}}) \tag{13c}$$

$$\pm = \sum_{i=1}^{N}\mathrm{sgn}(\eta_i^{\mathrm{calcd}} - \eta_i^{\mathrm{exptl}}) \tag{13d}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(t_i - y_i)^2}{\sum_{i=1}^{N}(\langle t\rangle - y_i)^2} \tag{13e}$$

In eqs 13a to 13e, the summations run over N selected data points. In eq 13e, y_i and t_i refer to the ith calculated and experimental (target) value of $\ln(\eta/\mathrm{mPa{\cdot}s})$. Angle brackets denote the arithmetic mean [i.e., $\langle t\rangle = (1/N)\sum_{i=1}^{N} t_i$].

Table 2 summarizes the calculations with respect to the whole data set division used in the model development and for different cationic families of ILs. As evidenced by these results, for a great majority of the ILs the viscosity behavior was captured quite accurately, namely, with AARD at the level of 10% or lower and $R^2 > 0.95$, whereas the AARD values obtained for individual ILs vary from <0.1 to 462%. In particular, the number of ILs for which the AARD is lower than 1, 5, 10, and 20% are 101, 387, 718, and 1156, respectively. A full list of all of the statistical measures given in eqs 13a to 13e for each IL separately is presented in the Supporting Information.

Table 2. Summary of Calculations Using the Proposed GC-FFANN Model for Different Cationic Families of ILs: Analysis of the Results by Means of the Statistical Measures Defined in Equations 13a to 13e

IL family[a]	AAD (mPa·s)	AARD (%)	bias (mPa·s)	±	R^2
Im (1a–d)	60.9	10.1	−13.6	209	0.989
Py (2)	25.1	7.85	−13.1	161	0.991
Quin (3a–c)	41.8	10.9	−28.3	3	0.988
Pyr (4a–c)	20	9.45	−10.1	57	0.983
Ox (5)	34.7	13.7	−4.29	4	0.880
Pip (6a–c)	14.4	13.2	−4.71	9	0.962
Mo (7)	18.4	8.52	−5.05	−35	0.991
Azp (8a–b)	9.81	8.17	−3.21	16	0.992
Guan (9a–b)	11.2	7.45	−4.1	5	0.982
N (10)	84.3	19.8	−26.8	93	0.963
P (11)	244	11.5	−69.3	−72	0.988
S (12a–f)	11.9	11.8	−4.52	−27	0.978
dicationic	267	27.2	−88.6	33	0.897
training set	68.5	11.1	−20.9	395	0.986
validation set	67.5	13.8	−17.2	14	0.973
testing set	57.3	14.7	2.83	47	0.972
overall	67.9	11.4	−19.6	47	0.984

[a]The abbreviations for the IL families are explained in footnote *a* of Table 1. The IDs given in parentheses are defined in Figure 3.

The worst accuracy was obtained for ammonium and dicationic ILs. In the case of ammonium ILs (AARD ≈ 20%), this might be associated with the great diversity of compounds present in the database, including cations with different numbers of alkyl groups attached to the nitrogen atom. Particularly, it regards protic ammonium ILs, whose the experimental data are usually subject to very significant scatter. In turn, dicationic ILs (AARD ≈ 26%) are a still insufficiently explored family of molten salts, and experimental errors in the viscosity determinations may be related to such poor predictions. Besides, the obtained relation between the computed results and the experimental data collected to date suggests that there is no clear and simple correlation between the physical properties of monocationic ILs and those of their divalent counterparts. We believe that if there were such a correlation, the values of the AARD (and AAD) would be substantially lower. Similar results were obtained in our previous contribution concerned with the densities of ILs,[40] where the deviations for dicationic ILs were at the level of ±5%, whereas for other families of ILs they were within ±1%.

The negative bias values observed for the majority of families suggest that our new FFANN model tends to under- rather than overpredict the viscosity. This is confirmed by the relatively small (and positive) values of "±". This means that there are more positive than negative deviations; nevertheless, the negative ones are larger in absolute value. Additionally, the observed bias and "±" values indicate that the deviation distributions are closely symmetric for each family of ILs as well as the entire training, validation, and testing sets.

Finally, exemplary applications using the developed GC-FFANN method are presented in Figure 8. In Figure 8a, temperature-dependent viscosity data at ambient pressure are shown for common ILs based on 1-hexyl-3-methylimidazolium cation ($[C_6C_1Im]$) and a few popular anions: bistriflamide ($[NTf_2]$), tetrafluoroborate ($[BF_4]$), hexafluorophosphate ($[PF_6]$), chloride ($[Cl]$), and dicyanamide ($[N(CN)_2]$). It is seen that the FFANN adopted in this work is capable of reliably representing viscosity over a wide range of temperature. The overall AARD values for these ILs are 6.8, 6.4, 4.9, 21, and 6.0%, respectively. In turn, Figure 8b displays the calculated $P-\eta-T$ surface for $[C_4C_1Im][NTf_2]$. We chose this IL for illustrative purposes because numerous high-temperature and high-pressure data (up to T = 573 K and P = 350 MPa) have been reported for it. Again, a satisfactory description is provided by the model despite the fact that the accuracy of the calculations deteriorates as the temperature increases. Nevertheless, it should be pointed out that the calculations shown in Figure 8b are correlations rather than pure predictions. In fact, most of the high-pressure data (more than 90%) were included in the training set during development of the model. Therefore, care must be taken when applying the high-pressure predictions of the proposed model for ILs completely different than those used in the network training process.

Comparison with the Gardas–Coutinho Model. The final step of our new GC-FFANN model evaluation is a

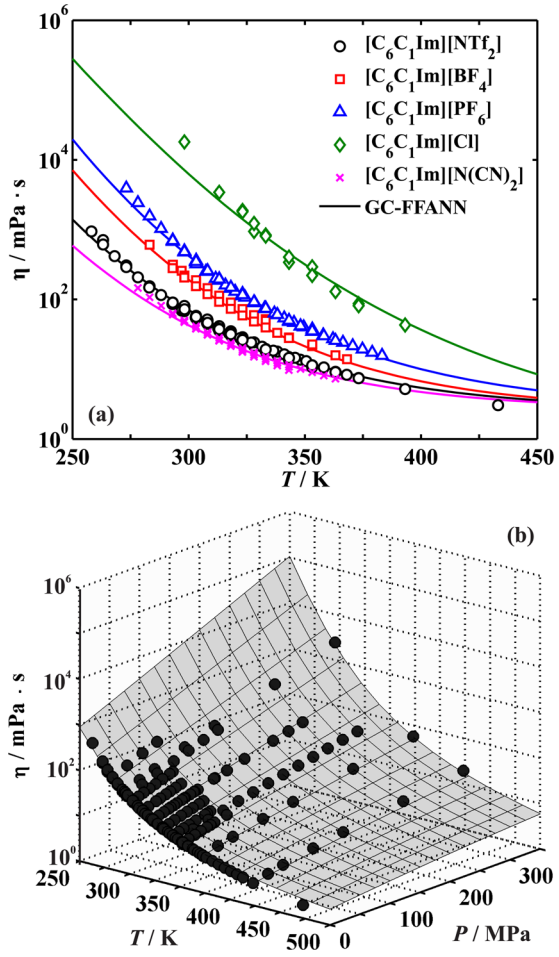


Figure 8. Representative viscosity calculations using the proposed GC-FFANN model: (a) atmospheric-pressure data; (b) $P-\eta-T$ surface for $[C_4C_1Im][NTf_2]$. A full list of references for the experimental data can be found in the Supporting Information.

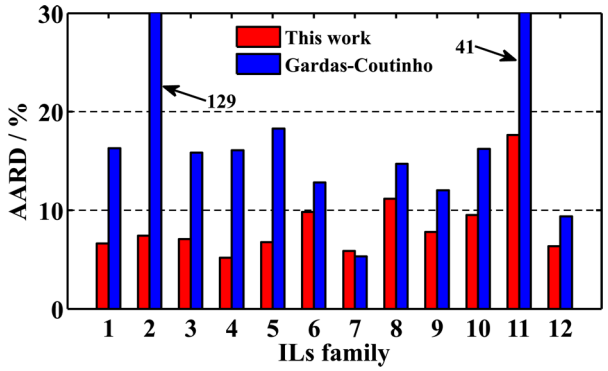


Figure 9. Comparison of AARDs (see eq 13a) in the viscosities for various families of ILs calculated using the proposed GC-FFANN model and the GC correlation developed by Gardas and Coutinho.[33] Key: (1) 1-alkyl-3-methylimidazolium bistriflamides, $[C_nC_1Im][NTf_2]$ (n = 1−14); (2) symmetrically substituted 1,3-dialkylimidazolium bistriflamides, $[C_nC_nIm][NTf_2]$ (n = 2−12); (3) 1-alkylpyridinium bistriflamides, $[C_nPy][NTf_2]$ (n = 2−16); (4) 1-alkyl-1-methylpyrrolidinium bistriflamides, $[C_nC_1Pyr][NTf_2]$ (n = 3−10); (5) 1-alkyl-3-methylimidazolium tetrafluoroborates, $[C_nC_1Im][BF_4]$ (n = 2−10); (6) 1-alkylpyridinium tetrafluoroborates, $[C_nPy][BF_4]$ (n = 3−8); (7) 1-alkyl-3-methylimidazolium hexafluorophosphates, $[C_nC_1Im][PF_6]$ (n = 2−9); (8) 1-alkyl-3-methylimidazolium triflates, $[C_nC_1Im][OTf]$ (n = 2−10); (9) 1-alkyl-3-methylimidazolium alkylsulfates, $[C_nC_1Im][C_kSO_4]$ (n = 1−8, k = 1, 2); (10) 1-alkyl-3-methylimidazolium dimethylphosphates, $[C_nC_1Im][(C_1O)_2PO_2]$ (n = 1−8); (11) 1-alkyl-3-methylimidazolium acetates, $[C_nC_1Im][C_1COO]$ (n = 2, 4); (12) 1-alkyl-3-methylimidazolium thiocyanates $[C_nC_1Im][SCN]$ (n = 2−10), dicyanamides $[C_nC_1Im][N(CN)_2]$ (n = 1−8), tricyanomethanides $[C_nC_1Im][C(CN)_3]$ (n = 2, 4), and tetracyanoborates $[C_nC_1Im][B(CN)_4]$ (n = 2, 6). Individual AARD values for each IL are listed in the Supporting Information.

comparison of its correlative/predictive capabilities against those of the best current GC method for viscosity, proposed 5 years ago by Gardas and Coutihno.[33] For the comparative study, several families of ILs that can be represented by both of the models were selected. Exhaustive comparison including all of the ILs collected in the database was not possible, since the parameter table of the Gardas−Coutinho method is still limited (although it is being systematically extended).[37−39] Furthermore, the comparison was carried out only for viscosity data at P = 0.1 MPa, because the Gardas−Coutinho model does not provide the pressure dependence of the viscosity. The set of selected ILs comprises 1,3-dialkylimidazolium, 1-alkylpyridinium, and 1,1-dialkylpyrrolidinium cations and some popular anions such as $[NTf_2]$, $[BF_4]$, $[PF_6]$, [OTf] (triflate), $[C_1SO_4]$, $[C_2SO_4]$, $[(C_1O)_2PO_2]$ (dimethylphosphate), and $[C_1COO]$ (acetate) as well as cyano-based anions such as [SCN], $[N(CN)_2]$, $[C(CN)_3]$, and $[B(CN)_4]$, which have recently been extensively studied.[39]

To sum up, 95 ILs, divided into 12 families, were taken into consideration as the case study. The results are shown in Figure 9, where the comparison is demonstrated in terms of AARDs calculated for each family separately. A more detailed summary is given in tabular form in the Supporting Information. As can be seen, the GC-FFANN method presented in this paper exhibits significantly better accuracy for a great majority of ILs. In particular, the Gardas−Coutinho model fails when applied to $[NTf_2]$-based ILs having symmetrically substituted imidazolium cations (family no. 2 in Figure 9a), with overall deviations of >100%. In fact, this could be expected because structures of this type were not included in the Gardas−Coutinho model parameters estimation procedure. Imidazolium acetates (family no. 11; only two ILs with $[C_2C_1Im]$ and $[C_4C_1Im]$ cations) are also not captured accurately, with AARD = 41%. Nevertheless, we are quite sure that this is due to experimental data scatter and that the results would be substantially better if the authors were to revise the original parameters published in 2009.[33] Worse performance of the GC-ANN model compared with the Gardas−Coutinho method was obtained for 1-alkyl-3-methylimidazolium hexafluorophosphates (family no. 7). However, it should be stressed that in Gardas−Coutinho model, the $[PF_6]$ anion was considered mostly with simple alkyl-substituted imidazolium ILs, whereas in the GC-FFANN model presented in this work, it was combined with many kinds of cations, including those incorporating task-specific functional groups such a C≡N, OH, NH_2, and CH_3O (see the database summary in the Supporting Information).

Finally, the global AARD values (including all families of ILs) were 7.0 and 21% for the GC-FFANN and Gardas−Coutinho models, respectively. If family no. 2 (which causes the biggest bias in AARD) is excluded, the latter value becomes much smaller (13.9%) but is still almost twice as large as that obtained for the method developed in this work.

dx.doi.org/10.1021/ci500206u | J. Chem. Inf. Model. 2014, 54, 1311−1324

CONCLUSIONS

On the basis of numerous and critically revised experimental data on viscosities of pure ILs, a new correlation for predicting $P-\eta-T$ behavior of this class of fluids has been proposed. To our best knowledge, this is the most comprehensive model described in the literature to date, taking into account the size of the database used during its development and testing (1484 ILs, 12 405 data points) as well as the number of functional groups introduced (242 building blocks). The groups enabled us to accurately capture more than 722 cationic and 179 anionic structures described in the literature. What is important is that the groups were constructed in such a manner that they can be used in future studies to estimate viscosities of ILs involving countless numbers of ions, which is not taken into account in some of the analogous models proposed in the literature.

It was confirmed that the FFANN approach combined with the GC strategy serves as an excellent tool for reproducing very complex dependences among any number of input and output variables. In particular, a properly trained network requires a large number of training samples and an appropriate number of hidden nodes. In this work, these conditions were fulfilled, as over 11 000 data points were used to estimate the network weights and biases. The reliability of the finally obtained GC-FFANN model was confirmed by means of an internal cross-validation process during optimization (the "early stopping" method) as well as testing of the true predictive capacity of the network using the data entirely excluded from the parameter optimization process (the testing set). The overall accuracy of the viscosity calculations is at the level of ±10%. Additionally, the superiority of our GC-FFANN model over the competing Gardas–Countinho GC model[33] was demonstrated.

Finally, it should be remarked that there are still some limitations of FFANN-based correlations. Unfortunately, because of the standardization of inputs and targets that is required to run FFANN, a problem may arise in extrapolation. Predictions based on input data falling out of the range of the data employed in the network development should be perceived as somewhat risky, and hence, such results have to be treated carefully. Nevertheless, the database involved in this work covers a broad range of working conditions of temperature (up to 573 K) and pressure (up to 350 MPa). We believe that it is completely sufficient for a typical (or even still not known) applications of ILs in chemical industry.

APPENDIX

A simple (and thus not exactly foolproof) program for calculating viscosities using the proposed GC-FFANN method is included in the Supporting Information as the MATLAB function *ilvisco*. The main code uses several auxiliary files that cannot be modified, moved, or removed. The program performs direct calculations given in eqs 4 to 8 by simple algebraic operations on matrices and vectors. Therefore, it can be executed on any version of MATLAB, even if the Neural Network Toolbox (NNT) is not installed. It also works on GNU Octave free software. Nonetheless, a separate mat file with a MATLAB network object is also provided for users who are familiar with the NNT.

Execution of the program is performed by the command `eta = ilvisco(inputfile)`. It returns a matrix containing calculated values of the viscosity (in mPa·s) along with the corresponding values of temperature (in K) and pressure (in MPa) defined in inputfile.

A typical input file (provided in Supporting Information as *sample.txt*) is organized as follows:

```
# ---------------------------------------
# Viscosity calculation for:
# [C4C1Im][NTf2]
# ---------------------------------------
#
# Group assignment
Im13=1,aNCH3=1,aNCH2=1,CH2=2,CH3=1,NTf2=1
# T-P range
298.15,423.15,5
1,10,10
```

Lines starting with a hashtag (#) are comments. Comments can be used in any part of the file, but they have to start in the first column of a given row. As can be seen, the program requires three lines of input. In the first line, group assignments in the form of a comma-separated list of terms `GroupName=GroupOccurences` are defined. The list of accepted group names is also provided in the Supporting Information (in particular, the code catches and displays unexpected group names). The last two lines contain temperature and pressure ranges specified as `MinVal,MaxVal,NVal`. In the case of the input file *sample.txt* presented above, the program will return viscosities calculated at five temperatures uniformly distributed between 298.15 and 423.15 K, at pressures ranging from 1 to 10 MPa at each temperature. The output will appear as follows:

```
>> eta = ilvisco('sample.txt')

eta =

  298.1500    1.0000   55.2916
  329.4000    1.0000   17.9532
  360.6500    1.0000    8.5555
  391.9000    1.0000    5.3380
  ........   .......   .......
  391.9000   10.0000    5.4816
  423.1500   10.0000    4.0528

>>
```

ASSOCIATED CONTENT

Supporting Information

Table S1 presenting a detailed summary of the viscosity database, including ranges of $P-\eta-T$ data, references to literature sources for each data set, and the decision on acceptance/rejection of a given data set in GC-FFANN calculations; Table S2 listing functional groups defined within the model and their molecular structures; Table S3 containing a full list of weights and biases of the optimized FFANN; Table S4 presenting detailed results of calculations using the developed GC model, including values of all the statistical measures given in eqs 13a to 13e for each data set; Table S5 showing the comparison between calculations obtained with the GC-FFANN model and those obtained with the Gardas–Coutinho GC model; an extensive catalog of ions, where abbreviations, IUPAC preferred names, and chemical formulas are shown separately for almost 900 ions under study; and a simple code allowing one to perform calculations using the MATLAB/GNU Octave platform. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: kpaduszynski@ch.pw.edu.pl.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

Funding for this research was provided by the Warsaw University of Technology. K.P. kindly acknowledges the support of the Foundation for Polish Science within the framework of the START 2013 Program.

REFERENCES

(1) Welton, T. Room-Temperature Ionic Liquids: Solvents for Synthesis and Catalysis. *Chem. Rev.* **1999**, *99*, 2071−2084.

(2) Hallett, J. P.; Welton, T. Room-Temperature Ionic Liquids: Solvents for Synthesis and Catalysis. 2. *Chem. Rev.* **2011**, *111*, 3508−3576.

(3) Plechkova, N. V.; Seddon, K. R. Applications of Ionic Liquids in the Chemical Industry. *Chem. Soc. Rev.* **2008**, *37*, 123−150.

(4) Freemantle, M. *An Introduction to Ionic Liquids*; RSC Publishing: Cambridge, U.K., 2009.

(5) Lei, Z.; Dai, C.; Chen, B. Gas Solubility in Ionic Liquids. *Chem. Rev.* **2014**, *114*, 1289−1326.

(6) Pereiro, A. B.; Araújo, J. M. M.; Oliveira, F. S.; Bernardes, C. E. S.; Esperanca, J. M. S. S.; Canongia Lopes, J. N.; Marrucho, I. M.; Rebelo, L. P. N. Inorganic Salts in Purely Ionic Liquid Media: The Development of High Ionicity Ionic Liquids (HIILs). *Chem. Commun.* **2012**, *48*, 3656−3658.

(7) Freire, M. G.; Louros, C. L. S.; Rebelo, L. P. N.; Coutinho, J. A. P. Aqueous Biphasic Systems Composed of a Water-Stable Ionic Liquid + Carbohydrates and Their Applications. *Green Chem.* **2011**, *13*, 1536−1545.

(8) Zakrzewska, M. E.; Bogel-Łukasik, E.; Bogel-Łukasik, R. Solubility of Carbohydrates in Ionic Liquids. *Energy Fuels* **2010**, *24*, 737−735.

(9) Paduszyński, K.; Okuniewski, M.; Domańska, U. Renewable Feedstocks in Green Solvents: Thermodynamic Study on Phase Diagrams of D-Sorbitol and Xylitol with Dicyanamide Based Ionic Liquids. *J. Phys. Chem. B* **2013**, *117*, 7034−7046.

(10) Paduszyński, K.; Okuniewski, M.; Domańska, U. "Sweet-in-Green" Systems Based on Sugars and Ionic Liquids: New Solubility Data and Thermodynamic Analysis. *Ind. Eng. Chem. Res.* **2013**, *52*, 18482−18491.

(11) Cevasco, G.; Chiappe, C. Are Ionic Liquids a Proper Solution to Current Environmental Challenges? *Green Chem.* **2014**, *16*, 2375−2385.

(12) Nunes, V. M. B.; Lourenco, M. J. V.; Santos, F. J. V.; Nieto de Castro, C. Importance of Accurate Data on Viscosity and Thermal Conductivity in Molten Salts Applications. *J. Chem. Eng. Data* **2003**, *48*, 446−450.

(13) Hendriks, E.; Kontogeorgis, G. M.; Dohrn, R.; de Hemptine, J. C.; Economou, I. G.; Zilnik, L. F.; Vesovic, V. Industrial Requirements for Thermodynamics and Transport Properties. *Ind. Eng. Chem. Res.* **2010**, *49*, 11131−11141.

(14) *Computer Aided Molecular Design: Theory and Practice*; Achenie, L. E. K., Gani, R., Venkatasubramanian, V., Eds.; Computer Aided Chemical Engineering, Vol. *12*; Elsevier: Amsterdam, 2002.

(15) Abbott, A. P. Application of Hole Theory to the Viscosity of Ionic and Molecular Liquids. *ChemPhysChem* **2004**, *5*, 1242−1246.

(16) Bandrés, I.; Alcalde, R.; Lafuente, C.; Atilhan, M.; Aparicio, S. On the Viscosity of Pyridinium Based Ionic Liquids: An Experimental and Computational Study. *J. Phys. Chem. B* **2011**, *115*, 12499−12513.

(17) Slattery, J. M.; Daguenet, C.; Dyson, P. J.; Schubert, T. J. S.; Krossing, I. How to Predict the Physical Properties of Ionic Liquids: A Volume-Based Approach. *Angew. Chem., Int. Ed.* **2007**, *46*, 5384−5388.

(18) Mendonca, A. C. F.; Dörr, N.; Pádua, A. A. H. Predicting Thermophysical Properties of Ionic Liquids as a Function of Temperature and Pressure. *Proc. Inst. Mech. Eng., Part J: J. Eng. Tribol.* **2006**, *226*, 965−976.

(19) Dutt, N. V. K.; Ravikumar, Y. V. L; Rani, K. Y. Representation of Ionic Liquid Viscosity−Temperature Data by Generalized Correlations and an Artificial Neural Network (ANN) Model. *Chem. Eng. Commun.* **2013**, *200*, 1600−1622.

(20) Yamamoto, H. Structure Properties Relationship of Ionic Liquid. *J. Comput. Aided Chem.* **2006**, *7*, 18−29.

(21) Tochigi, K.; Yamamoto, H. Estimation of Ionic Conductivity and Viscosity of Ionic Liquids Using a QSPR Model. *J. Phys. Chem. C* **2007**, *111*, 15989−15994.

(22) Bini, R.; Malvaldi, M.; Pitner, W. R.; Chiappe, C. QSPR Correlation for Conductivities and Viscosities of Low-Temperature Melting Ionic Liquids. *J. Phys. Org. Chem.* **2008**, *21*, 622−629.

(23) Eiden, P.; Bulut, S.; Köchner, T.; Friedrich, C.; Schubert, T.; Krossing, I. In Silico Predictions of the Temperature-Dependent Viscosities and Electrical Conductivities of Functionalized and Nonfunctionalized Ionic Liquids. *J. Phys. Chem. B* **2011**, *115*, 300−399.

(24) Han, C.; Yu, G.; Wen, L.; Zhao, D.; Asumana, C.; Chen, X. Data and QSPR Study for Viscosity of Imidazolium-Based Ionic Liquids. *Fluid Phase Equilib.* **2011**, *300*, 95−104.

(25) Valderrama, J. O.; Muñoz, J. M.; Rojas, R. E. Viscosity of Ionic Liquids Using the Concept of Mass Connectivity and Artificial Neural Networks. *Korean J. Chem. Eng.* **2011**, *28*, 1451−1457.

(26) Yu, G.; Zhao, D.; Wen, L.; Yang, S.; Chen, X. Viscosity of Ionic Liquids: Database, Observation, and Quantitative Structure−Property Relationship Analysis. *AIChE J.* **2012**, *58*, 2885−2899.

(27) Mirkhani, S. A.; Gharagheizi, F. Predictive Quantitative Structure−Property Relationship Model for the Estimation of Ionic Liquid Viscosity. *Ind. Eng. Chem. Res.* **2012**, *51*, 2470−2477.

(28) Chen, B.-K.; Liang, M.-J.; Wu, T.-Y.; Wang, H. P. A High Correlate and Simplified QSPR for Viscosity of Imidazolium-Based Ionic Liquids. *Fluid Phase Equilib.* **2013**, *350*, 37−42.

(29) Yu, G.; Wen, L.; Zhao, D.; Asumana, C.; Chen, X. QSPR Study on the Viscosity of Bis(trifluoromethylsulfonyl)imide-Based Ionic Liquids. *J. Mol. Liq.* **2013**, *184*, 51−59.

(30) Matsuda, H.; Yamamoto, H.; Kurihara, K.; Tochigi, K. Computer-Aided Reverse Design for Ionic Liquids by QSPR Using Descriptors of Group Contribution Type for Ionic Conductivities and Viscosities. *Fluid Phase Equlib.* **2007**, *261*, 434−443.

(31) Matsuda, H.; Yamamoto, H.; Kurihara, K.; Tochigi, K. Prediction of the Ionic Conductivity and Viscosity of Ionic Liquids by QSPR Using Descriptors of Group Contribution Type. *J. Comput. Aided Chem.* **2007**, *8*, 114−127.

(32) Gardas, R. L.; Coutinho, J. A. P. A Group Contribution Method for Viscosity Estimation of Ionic Liquids. *Fluid Phase Equilib.* **2008**, *266*, 195−201.

(33) Gardas, R. L.; Coutinho, J. A. P. Group Contribution Methods for the Prediction of Thermophysical and Transport Properties of Ionic Liquids. *AIChE J.* **2009**, *55*, 1274−1290.

(34) Gharagheizi, F.; Ilani-Kashkouli, P.; Mohammadi, A. H.; Ramjugernath, D.; Richon, D. Development of a Group Contribution Method for Determination of Viscosity of Ionic Liquids at Atmospheric Pressure. *Chem. Eng. Sci.* **2012**, *80*, 326−333.

(35) Billard, I.; Marcou, G.; Ouadi, A.; Varnek, A. In Silico Design of New Ionic Liquids Based on Quantitative Structure−Property Relationship Models of Ionic Liquid Viscosity. *J. Phys. Chem. B* **2011**, *115*, 93−98.

(36) Coutinho, J. A. P.; Carvalho, P. J.; Oliveira, N. M. C. Predictive methods for the estimation of thermophysical properties of ionic liquids. *RSC Adv.* **2012**, *2*, 7322−7346.

(37) Carvalho, P. J.; Regueira, T.; Santos, L. M. N. B. F.; Fernandez, J.; Coutinho, J. A. P. Effect of Water on the Viscosities and Densities of 1-Butyl-3-methylimidazolium Dicyanamide and 1-Butyl-3-methylimidazolium Tricyanomethane at Atmospheric Pressure. *J. Chem. Eng. Data* **2010**, *55*, 645−652.

(38) Freire, M. G.; Teles, A. R. R.; Rocha, M. A. A.; Schöder, B.; Neves, C. M. S. S.; Carvalho, P. J.; Evtuguin, D. V.; Santos, L. M. N. B. F.; Coutinho, J. A. P. Thermophysical Characterization of Ionic

Liquids Able To Dissolve Biomass. *J. Chem. Eng. Data* **2011**, *56*, 4813−4822.
(39) Neves, C. M. S. S.; Kurnia, K. A.; Coutinho, J. A. P.; Marrucho, I. M.; Canongia Lopes, J. N.; Freire, M. G.; Rebelo, L. P. N. Systematic Study of the Thermophysical Properties of Imidazolium-Based Ionic Liquids with Cyano-Functionalized Anions. *J. Phys. Chem. B* **2013**, *117*, 10271−10283.
(40) Paduszyński, K.; Domańska, U. A New Group Contribution Method for Prediction of Density of Pure Ionic Liquids over a Wide Range of Temperature and Pressure. *Ind. Eng. Chem. Res.* **2012**, *51*, 591−604.
(41) Gharagheizi, F.; Eslamimanesh, A.; Mohammadi, A. H.; Richon, D. Group Contribution-Based Method for Determination of Solubility Parameter of Nonelectrolyte Organic Compounds. *Ind. Eng. Chem. Res.* **2011**, *50*, 10344−10349.
(42) Gharagheizi, F.; Eslamimanesh, A.; Mohammadi, A. H.; Richon, D. Use of Artificial Neural Network−Group Contribution Method to Determine Surface Tension of Pure Compounds. *J. Chem. Eng. Data* **2011**, *56*, 2587−2601.
(43) Gharagheizi, F.; Eslamimanesh, A.; Mohammadi, A. H.; Richon, D. Determination of Critical Properties and Acentric Factors of Pure Compounds Using the Artificial Neural Network Group Contribution Algorithm. *J. Chem. Eng. Data* **2011**, *56*, 2460−2476.
(44) Gharagheizi, F.; Eslamimanesh, A.; Ilani-Kashkouli, P.; Mohammadi, A. H.; Richon, D. Determination of Vapor Pressure of Chemical Compounds: A Group Contribution Model for an Extremely Large Database. *Ind. Eng. Chem. Res.* **2012**, *51*, 7119−7125.
(45) Suzuki, T.; Ebert, R.-U.; Schüürmann, G. Application of Neural Networks to Modeling and Estimating Temperature-Dependent Liquid Viscosity of Organic Compounds. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 776−790.
(46) Levenberg, K. A Method for the Solution of Certain Problems in Least Squares. *Q. Appl. Math.* **1944**, *2*, 164−168.
(47) Marquardt, D. An Algorithm for Least-Squares Estimation of Nonlinear Parameters. *SIAM J. Appl. Math.* **1963**, *11*, 431−441.
(48) Marrero, J.; Gani, R. Group-Contribution Based Estimation of Pure Component Properties. *Fluid Phase Equilib.* **2001**, *183−184*, 183−208.
(49) *Artificial Neural Networks: Methods and Applications*; Livingstone, D. J., Ed.; Humana Press: New York, 2008.
(50) Beale, M. H.; Hagan, M. T.; Demuth, H. B. *Neural Network Toolbox User's Guide, R2013a*; The MathWorks, Inc.: Natick, MA, 2013.
(51) Klimasauskas, C. C. In *Neural Networks in Finance and Investing: Using Artificial Intelligence To Improve Real World Performance*; Trippi, R. R., Turban, E., Eds.; Probus: Chicago, 1993; Chapter: Applying Neural Networks, pp 64−65.
(52) Nguyen, D.; Widrow, B. Improving the Learning Speed of 2-Layer Neural Networks by Choosing Initial Values of the Adaptive Weights. *Int. Jt. Conf. Neural Networks, Proc.* **1990**, *3*, 21−26.