



Molecular fingerprint-based machine learning assisted QSAR model development for prediction of ionic liquid properties

Yi Ding^{a,1}, Minchun Chen^{b,1}, Chao Guo^{a,1}, Peng Zhang^c, Jingwen Wang^{a,*}

^a Department of Pharmacy, Xijing Hospital, Fourth Military Medical University, Xi'an 710032, China

^b Department of Pharmacy, Xi'an No.3 Hospital, the affiliated Hospital of Northwest University, Xi'an, 710018, China

^c Department of Urology, Wuhan Third Hospital/Tongren Hospital of Wuhan University, Wuhan 430071, China

ARTICLE INFO

Article history:

Received 21 August 2020

Received in revised form 6 November 2020

Accepted 24 December 2020

Available online 28 December 2020

Keywords:

Ionic liquid

QSARs

Machine learning

Refractive index

Viscosity

ABSTRACT

Ionic liquids (ILs) have many applications in, for example, organic synthesis, batteries and drug delivery. In this study, molecular fingerprint (MF) was used to represent ionic liquids (ILs) and was combined with machine learning (ML) to develop quantitative structure-activity relationship (QSAR) models for predicting the refractive index and viscosity of ILs. To demonstrate the effectiveness of this approach, four datasets with different sizes containing different numbers of ILs' refractive indexes and viscosity, which were previously used to develop QSAR models by molecular descriptor (MD)-based method and group contribution method (GCM), were employed to develop QSAR models by MF-ML method. The results showed that the models developed by MF-ML showed comparative predictive performance with the MD-based method and GCM for these four datasets, but MF-ML can more quickly obtain the representations of IL within milliseconds. Moreover, the MF-ML models were interpreted by the recently developed shapely additive explanation (SHAP) method. The results showed that the models made the predictions based on the reasonable understanding of how different features affect the related properties of IL, thus building the trustworthiness of MF-ML models. This study offered a new approach with theoretical support to rapidly developing trustful QSAR models to predict the properties of ILs.

© 2021 Elsevier B.V. All rights reserved.

1. Introduction

Ionic liquids (ILs) are composed of anions and cations, both of which can be either organic or inorganic ions [1,2]. They have wide applications such as organic synthesis [3], batteries [4], catalysis [5], pharmaceuticals [6], and CO₂ adsorption [7]. The different applications require different desirable properties of ILs, such as high CO₂ adsorption capacity for CO₂ capture [8] and high refractive index for optical materials [9]. However, experimentally screening desirable ILs with the target property one by one is unrealistic because the number of potential ILs is estimated to be ~10 [8]. An alternative approach is to develop accurate quantitative structure-activity relationships (QSARs) models based on the existing experimental data points. In this way, one can quickly predict the desirable properties of new ILs without time-consuming and expensive experimental tests.

To develop QSARs models, several approaches were developed, mainly including the group contribution method (GCM) [10–12] and molecular descriptor (MD)-based methods [13–16]. In GCM, we

manually decomposed IL into several groups, and linearly combined their contributions toward the specific property, which is time-consuming. The contribution of each group was calibrated by minimizing the difference between the predicted results and the experimental ones in the training dataset. Although the GCM method can achieve a rather accurate prediction for a specific dataset, it becomes problematic if new ILs that need to be predicted contain groups that are not included in this specific dataset. GCM method sometimes needs to be combined with the physicochemical properties of ILs (e.g., density) to develop the QSAR model, which will also be problematic if these properties for new IL are not available. Moreover, GCM works based on the assumption that the atom groups in different ILs contributed equally to the properties of ILs, which is sometimes too arbitrary. When more ILs are involved, the same atom groups in different ILs may contribute differently. Such a non-linear relationship cannot be captured well by the GCM.

The MD-based method uses MD to represent ILs to correlate the property of IL, in which MD is the values of different physicochemical properties, such as HOMO/LUMO energies, charge partial surface areas (CPSA), and geometrical indices [17]. The main limitation of the MD-based method is the expensive calculation cost, which is time-consuming especially when more and more ILs are involved in the

* Corresponding author.

E-mail address: wangjingwen909@gmail.com (J. Wang).

¹ These authors contributed equally to this work.

dataset. MD is often used as input in methods such as conductor-like screening model for real solvents (COSMO-RS) [18,19], or artificial neural network (ANN) [20,21].

Molecular fingerprint (MF) encodes the chemical structural features of compounds into binary vectors containing only 0s and 1s [22], which are commonly used in tasks such as virtual screening [23], similarities searching [24], and clustering [25]. For the binary vectors, 0 means no certain chemical structure (e.g., -OH) is present in the compound while 1 means its presence. Different chemical structural features (e.g., -OH and -Br) occupy different positions in the binary vectors. As compared with MD, MF is much easier to obtain and understand. For instance, MFs of over thousands of compounds can be obtained within 1 s, which is impossible for obtaining their MDs in such a short time. Recently, MFs have been combined with machine learning (ML) to successfully develop the QSAR model to predict the ligand biological activity [23], toxicity [26,27], and the rate constants of OH radical toward organic contaminants [28,29]. Given this, in this study, we transferred this method to develop QSAR models to predict the properties of ILs.

To demonstrate the efficiency of this method, QSAR models for predicting two properties of ILs, i.e., refraction index and viscosity, were developed by using MFs of ILs as inputs into a traditional ML algorithm—extreme gradient boosting (XGBoost) [30]. XGBoost is a tree-based machine learning algorithm with a gradient boosting design. Gradient boosting develops trees in a step-wise way, in which the latter tree tries to minimize the error of the former tree. Details of XGBoost can refer to Qi et al.'s paper [31]. The refractive index and viscosity are two important properties of ILs, in which several related properties can be estimated once the refractive index of the material is known [17] while viscosity has a great influence on the transfer performance of the IL containing system [16]. This study used four datasets of different sizes and each of them were used to develop QSAR models by the traditional MD-based method or GCM. We then built QSAR models on these four datasets with the MF-XGBoost method and compared their predictive performance. Then, we used the recently developed shapely additive explanation (SHAP) method [32] to interpret the developed models, i.e., showing how the features such as temperature or pressure or atom groups affect the predictions, which is important for trusting our model when “black box” ML methods are used.

2. Methods and materials

2.1. Datasets

Four individual datasets containing refractive indices and viscosities of various ILs were employed in this study. Two datasets of refractive index are compiled from studies of Venkatraman et al. (labeled as Refractive index-1) and Sattari et al. (labeled as Refractive index-2) [11,17], respectively, in which Venkatraman et al. used the MD-based method while Sattari et al. used the GCM method to develop QSAR models to predict the refractive indexes for various ILs. The dataset of Venkatraman et al. contained a total of 3147 experimental data points of 467 ILs' refractive indices at different temperatures [17]. The ILs are composed of 240 cations with major classes such as imidazolium, ammonium, pyrrolidinium, and pyridinium, and 86 anions that are dominated by carboxylates, halides, and sulfates. The dataset of Sattari et al. was of a relatively small size, which contained 931 experimental data points of 97 unique ILs constituted from 50 different cations and 33 anions [11].

Two datasets of viscosity are compiled from Zhao et al. (labeled as Viscosity-1) and Chen et al. (labeled as Viscosity-2) [16,33], of which all the experimental data points are collected from IL Thermo and IUPAC Database. In Zhao et al.'s dataset, a total of 1502 experimental viscosity data points of 89 ILs are investigated [16]. The collected viscosity data points (8.28–142,000 cP) cover a wide range of pressures (1–3000 bar) and temperatures (253.15–395.32 K). Chen et al.'s

viscosity data is a relatively small size [33], which contains 304 experimental data points that covered a range of temperatures (258.15–433.15 K) and viscosities (3–2300 cP), and was kept at constant pressure (1.01 bar). Likewise, Zhao et al. used the MDs-based method while Chen et al. used a GCM method to develop QSARs to predict the viscosity for different number of ILs [16,33].

The reason why we chose these four datasets was because (1) two different popular approaches, i.e., MD-based method and GCM, were used to develop models, which can be used to compare with our MF-based method; (2) large and small data volume are both involved in these four datasets, which can be used to test the applicability of our MFs-based methods. A summary of these four data is presented in Table 1. More details about the dataset such as the cations and anions species and the number of data points in each type of IL can be found in their papers [11,16,17,33].

2.2. Machine learning model development

The SMILES of all the cations and anions that constitute ILs were first obtained by the ChemDraw and then converted to MFs by the RDKit package in Python. An MF of IL is thus composed of the MFs of its corresponding cation and anion. An example is illustrated in Fig. 1, in which the MFs of 1-Butyl-3-methylimidazolium cation and hexafluorophosphate anion are combined as MF of IL. Conditions such as the temperature or/and pressure were combined with the MFs of IL as the final inputs to XGBoost, as shown in Fig. 1.

Two types of molecular fingerprints were used in this study: Morgan fingerprint and atom-pair fingerprint, both of which encode the chemical structural information of ILs into binary vectors that are filled with only 0s and 1s. Only 1s represents certain structural features existing in ILs and its position in the vector represents what the specific structural feature is in ILs. The difference between Morgan fingerprint and atom-pair fingerprint is the way the chemical structural features are represented. Morgan fingerprint first localizes a center atom and then includes its neighbor atoms with a certain radius while atom-pair uses substructures composed of two non-hydrogen atoms and an inter-atomic separation. The detailed explanations on how to produce Morgan fingerprint and atom-pair fingerprint are illustrated in Rogers et al.'s and Carhart et al.'s paper [22,34].

Every dataset was split into the training dataset and test dataset, in which the training dataset was used to train ML while the test dataset was used to test the generalizability of the obtained model. The test dataset was not used in the development process of the model, which guaranteed that it had never been exposed to the model. We directly used the same training and test datasets as the studies of Venkatraman et al., Sattari et al., Zhao et al. and Chen et al. to make a fair comparison with their results [11,16,17,33]. However, to control the overfitting problem, we further split the training dataset into a sub-training dataset and a validation dataset, i.e., cross-validation. Overfitting means the model shows high predictive performance on the training dataset but a poor one on the test dataset. In other words, ML only memorizes the data in the training dataset rather than correlating the underlying relationships. This sub-training dataset was used to train the model while the validation dataset was used to choose the hyperparameters of XGBoost to control the complexity of the model and control the overfitting. The validation dataset was not involved in training the model. To fully use the training dataset, we did a 5 cross-validation on the training dataset, in which the training dataset was split 5 times to form 5 sub-training datasets and 5 validation datasets. The optimum hyperparameters were the ones that minimize the average prediction performance on these 5 validation datasets. After obtaining the optimum hyperparameters, the XGBoost was retrained on the whole training dataset to obtain the final model. This final model was then tested on the test dataset as the final evaluation of its predictive performance.

Hyperparameters are the parameters that are pre-set before the training process. The hyperparameters in this study included the

hyperparameters of both XGBoost and MF (i.e., radius and length), in which any values can be taken. It is thus impossible to enumerate every value to obtain the optimum one. Given this, we used the powerful Bayesian optimization algorithm to optimize the hyperparameters, which can choose the next hyperparameter candidate based on the results obtained from the previous ones [35,36]. Hence, the possibility to achieve the optimum values of the hyperparameters was maximized. Table 2 listed the parameter of MF obtained by the Bayesian optimization algorithm, i.e., the length and radius, the positions of cation, anion, temperature, and pressure in the vectors for these four datasets.

2.3. Model interpretation

Model interpretation is an important step for understanding the reason why a model makes a specific prediction especially when a “black box” ML algorithm is used. Here, we used the recently developed SHAP method that is derived from the cooperative game theory [32]. SHAP assigns the Shapley values for the contribution of a feature to the final prediction, in which the higher the absolute Shapley value of a feature, the more important this feature is for the final prediction. A positive or negative Shapley value of a feature represents that this feature can increase or decrease the final prediction. The underlying

working mechanism of SHAP is that the effect of a feature is calculated by checking what the prediction would be if that feature is absent. However, this may lose the interaction information between features because different features may have dependent relationships. To avoid this, we should observe how the predictions change for each possible subset of features with and without a certain feature and then combine these changes to form a unique contribution for each feature, which includes the interaction effect between features. The SHAP method can also show the trend of effect on the final prediction with the change of feature values. The above explanations enable us to evaluate if our model makes predictions based on the knowledge of how different features affect the predictions, even though this model is obtained by a “black box” ML algorithm.

2.4. Evaluation metrics

The root mean square error (RMSE) and R-squared (R^2) were used to evaluate the performance of the developed models. RMSE is the standard deviation of the residuals (prediction errors) (eq. 1) and R^2 gives a fraction of explained variance for a data set (eq. 2). The lower the RMSE and the higher R^2 are, the better the model is.

Table 1
Summary of the experimental data of refractive index and viscosity.

Dataset	Temperature (K)	Pressure (bar)	n_D or η	Data points	Ref
Refractive index-1	283.15–571.15	–	1.355–1.659	3147	17
Refractive index-2	283.15–363.15	–	1.355–1.578	931	11
Viscosity-1	253.15–395.32	1–3000	8.28–142,000	1502	16
Viscosity-2	258.15–433.15	1.01	3–2300	304	33

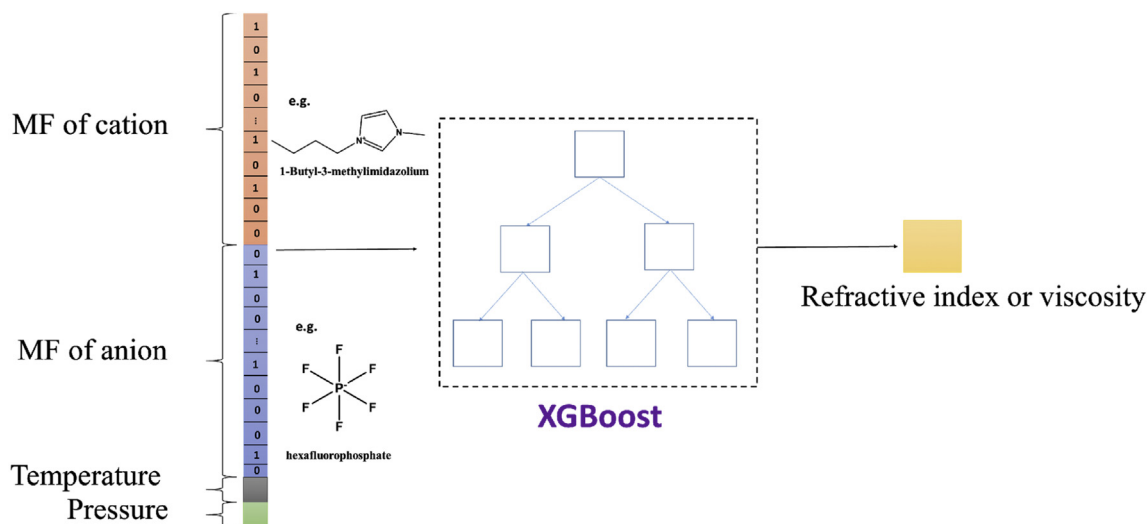


Fig. 1. The scheme of inputs and output used in ML.

Table 2
The optimum parameters of MF obtained by Bayesian optimization algorithm and the positions of cation, anion, and conditions (i.e., temperature and pressure) in the binary vector for all the four dataset.

Dataset	MF (radius, length)	Cation	Anion	Temperature	Pressure
Refractive index-1	(6, 4066)	0–4065	4066–8131	8132	–
Refractive index-2	(2, 2055)	0–2054	2055–4109	4110	–
Viscosity-1	(7, 4632)	0–4631	4632–9263	9265	9264
Viscosity-2	(6, 3752)	0–3751	3752–7503	7504	–

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (V^{exp} - V^{pred})^2}{n}} \quad (1)$$

$$R^2 = \frac{\sum_i (V^{pred} - \overline{V^{exp}})^2}{\sum_i (V^{exp} - \overline{V^{exp}})^2} \quad (2)$$

Table 3

The comparison of predictive performance on the test dataset between Morgan fingerprint-based and atom-pair fingerprint-based XGBoost QSAR models.

Fingerprint	Dataset	RMSE	R ²
Morgan Fingerprint	Refractive index-1	0.017	0.782
	Refractive index-2	0.013	0.853
	Viscosity-1	0.0091	0.97
	Viscosity-2	0.053	0.989
Atom-pair Fingerprint	Refractive index-1	0.016	0.836
	Refractive index-2	0.022	0.568
	Viscosity-1	0.162	0.918
	Viscosity-2	0.065	0.984

where V^{pred} , V^{exp} , $\overline{V^{exp}}$ and n is the predicted, experimental, and average of values and the number of data points.

3. Results and discussion

3.1. The comparison of the predictive performance

Table 3 lists the comparison of predictive performance on the test dataset for these four datasets when different types of molecular fingerprint were used to develop QSAR models. Compared with the atom-pair fingerprint, using the Morgan fingerprint for the representation of ILs is better in terms of RMSE and R² [2]. Except for the dataset of Refractive index-1, QSAR models developed by Morgan fingerprint showed lower RMSE and higher R² for the other three datasets (Table 3) than the atom-pair fingerprint. Hence, Morgan fingerprint was chosen as the representation of ILs in the following study.

We then compared the MD-based method with the MF-XGBoost method on the datasets of refractive index and viscosity. For the refractive index dataset (Fig. 2A), Venkatraman et al. used the MD-based method that correlated the MDs of ILs with their experimental refractive indexes by several regression methods [17], in which the top-2

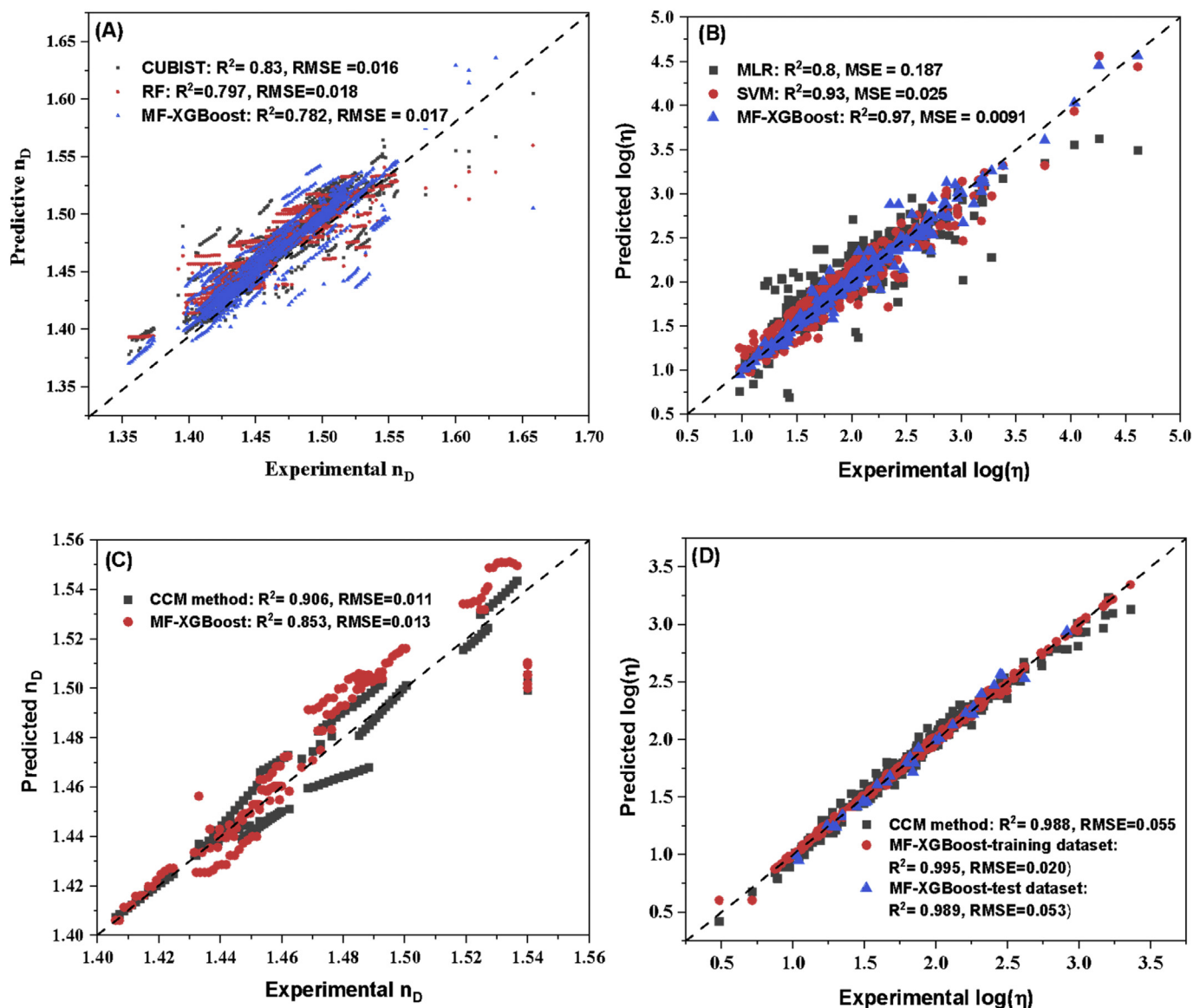


Fig. 2. The plots of experimental refractive index (A) and viscosity (B) versus their corresponding predictive values in the test dataset by different models. (CUBIST: the name of package, Rule- And Instance-Based Regression Modeling; RF: random forest; SVM: supporting vector machine; MLR: multiple linear regression).

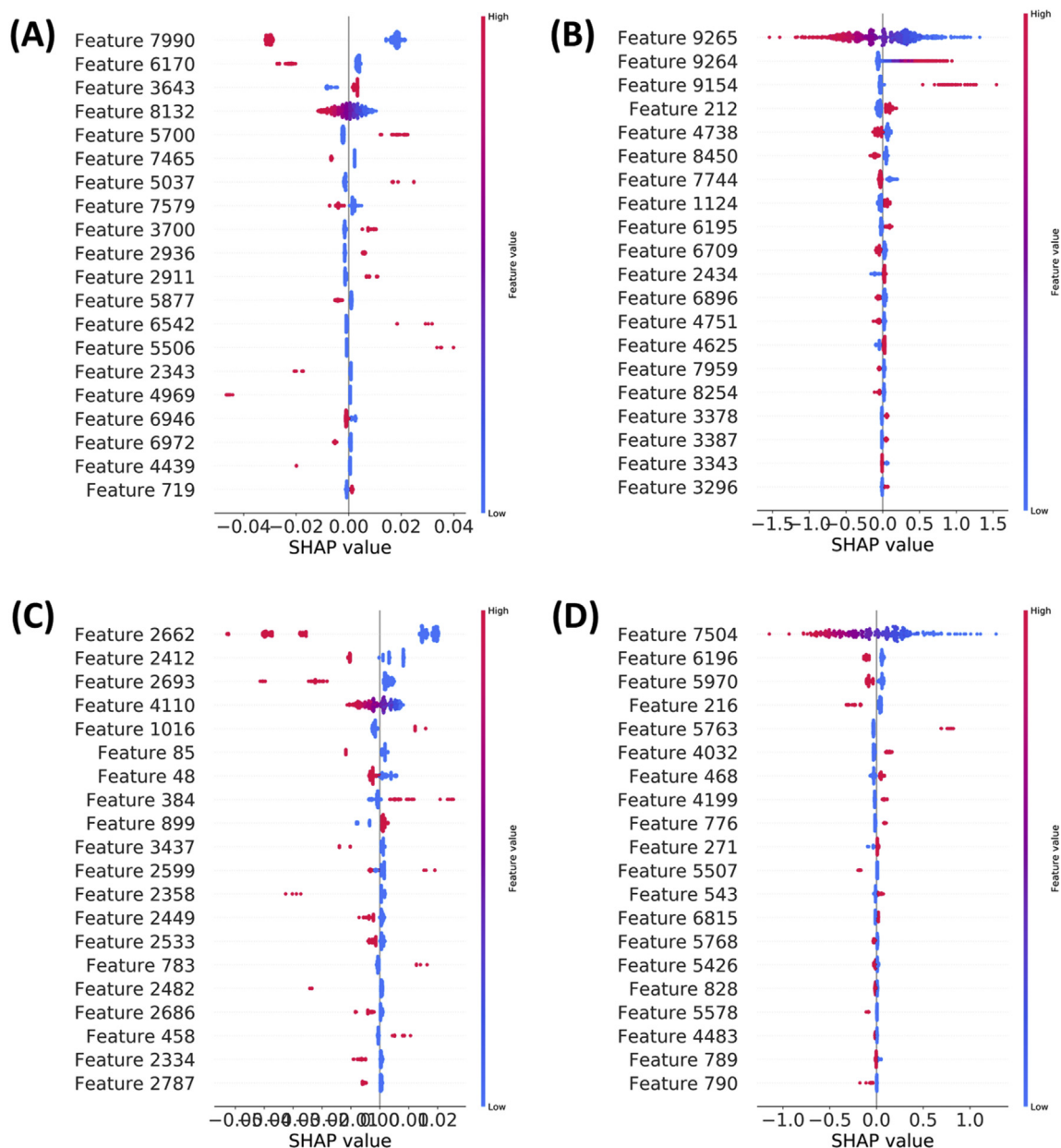


Fig. 3. The SHAP plot of four models obtained on these for the dataset, in which (A) and (C) were for refractive index-1 and refractive index-2 while (B) and (D) were for the viscosity-1 and viscosity-2.

best ones were the CUBIST and RF, as shown in Fig. 2A. We thus chose these two models to compare with the MF-XGBoost model. The predictive performance, R^2 , and RMSE were calculated for the test dataset that has never been “seen” by the model. The MF-XGBoost model showed similar R^2 (0.782) and RMSE (0.017) to CUBIST ($R^2=0.83$, RMSE=0.16) and RF ($R^2=0.797$, RMSE=0.018), indicating that the MF-XGBoost model has a comparative predictive performance to the MD-based models on predicting the refractive indexes of ILs. For the viscosity dataset (Fig. 2B), however, the MF-XGBoost model showed a higher predictive performance ($R^2 = 0.97$, MSE= 0.0091) than that of MLR ($R^2=0.8$, MSE=0.187) and SVM ($R^2=0.93$, MSE=0.025) which were combined with the S_G -profile of ILs in Zhao et al.’s study [16].

For the GCM method, Sattari et al. achieved $R^2 = 0.906$, RMSE = 0.011 on the test dataset for predicting the refractive indexes [11]. The MF-XGBoost method showed a slightly lower but still satisfactory predictive performance ($R^2 = 0.853$, RMSE = 0.013). For the viscosity


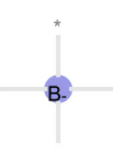
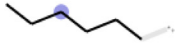

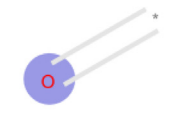

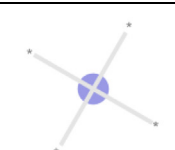

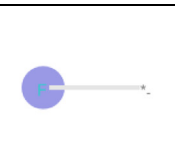
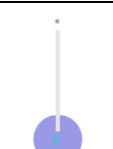
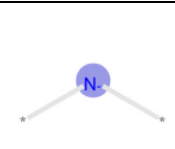
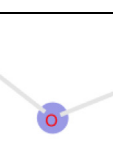
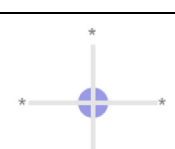
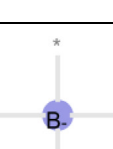
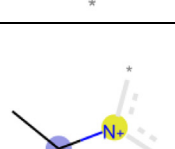

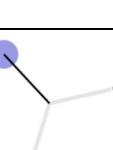
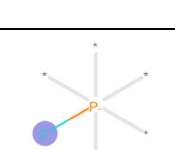
dataset, Chen et al. only offered the predictive performance on the training dataset ($R^2 = 0.988$, RMSE = 0.055) [33]. We have listed predictive performance on the training ($R^2 = 0.995$, RMSE = 0.020) and test dataset ($R^2 = 0.989$, RMSE = 0.053). It should be noted that the test dataset has never been exposed to the model before. All of these results showed that the MF-XGBoost showed comparative predictive performance with both the MD-based method and GCM, indicating the effectiveness of the MF-based method as a reliable approach to developing QSAR models with satisfactory predictive performance for predicting the properties of ILs.

3.2. Model interpretation by SHAP method

We then interpreted these models by the SHAP method because the “black box” XGBoost method was used to develop models. Fig. 3 showed the SHAP plot for these four models interpreted by the SHAP method.

Table 4

The top-5 features in the SHAP plots of Fig. 3 and their represented atom groups and effects on the predictions.

	Feature	Atom group	Effect on the prediction		Feature	Atom group	Effect on the prediction
Figure 3A	7990		Decrease	Figure 3B	9154	Cl ⁻	Increase
	6170		Decrease		212		Increase
	3643		Increase		4738		Decrease
	5700		Increase		8450		Decrease
	7465		Decrease		7744		Decrease
Figure 3C	2662		Decrease	Figure 3D	6196		Decrease
	2412		Decrease		5970		Decrease
	2693		Decrease		216		Decrease
	1016		Increase		5763	I ⁻	Increase
	85		Decrease		4032		Increase

Taking Fig. 3A as an example to illustrate how to read the SHAP plot, the X-axis is the Shapely values and Y-axis is the feature names. For example, Feature 8132 in Fig. 3A represents the feature in the position of 8132, which is the temperature (Table 2). For every feature, the patterns in the figure are composed of all the data points while the feature values are represented by the color in which color gradually changes from blue to red when the feature values gradually increase. For MF, red color represents 1 while blue color represents 0 because only 0 and 1 are possible values for MF. For comparison, the values of temperature and pressure are the continuous values, corresponding to the continuous change of colors. The feature with positive or negative Shapely values means it can increase or decrease the predictions. For example, feature 8312 represents the temperature, and with increasing temperature (the color is changed from blue to red), its Shapely value is gradually moved from the positive to the negative direction, indicating increasing temperature can decrease the prediction, i.e., refraction index. This is consistent with the fact that high temperature leads to a lower refraction index [9]. In summary, the SHAP plot unveiled how the model made predictions on the target, which we can use to conclude if the model is trustful or not.

3.2.1. The effect of temperature and pressure on the refractive index and viscosity

Based on the model interpretation, we can unveil what effects temperature and pressure have on the refractive index and viscosity the model found. It should be noted that we did not train the model to identify these effects but only correlate MFs with the corresponding refractive index and viscosity. The model “learned” such effects automatically, based on which it makes predictions on the properties.

For the refractive index dataset, i.e., Fig. 3A and C, the effect of temperature (Feature 8132 for Fig. 3A and Feature 4110 for Fig. 3C) on the refractive index is the same, that is, the increasing temperature decreased the refractive index. This is consistent with the experimental fact [9] that the SHAP values gradually changed from negative values to positive values when the temperature was gradually decreased (i.e., the color of feature 8132 in Fig. 3A and 4110 in Fig. 3C changed from red to blue). For the viscosity dataset, i.e., Fig. 3B and Fig. 3D, the effect of temperature on the viscosity (Feature 9265 for Fig. 3B and Feature 7504 for Fig. 3D) is also the same, that is, increasing the temperature decreased the viscosity, which is also consistent with the experimental fact [37]. The relationship between temperature and refractive index or viscosity was easily learned by the model after developing QSAR models. We can also check the transition temperature that starts to decrease the refractive index or viscosity by checking at what temperature the SHAP value is close to 0. For example, 316 °F was found as the transition temperature for Refractive index-1 because the temperature below this value had a negative SHAP value while that above this value had a positive one. Likewise, 313 °F was the transition temperature for Viscosity-1.

The pressure was involved only in the dataset of Viscosity-1 (Feature 9264). As shown in Fig. 3B, the model found that decreasing pressure (Color was changed from red to blue) is beneficial to decreasing the viscosity (SHAP values were changed from positive values to negative ones), which was also consistent with the experimental findings [37]. We can also determine a threshold for pressure: if pressure is below or above this threshold it will decrease or increase the viscosity, by checking at what pressure the SHAP value is close to 0. This threshold value is determined as 10,000 kPa, which means that any pressures below 10,000 kPa decrease the viscosity.

These results indicated that the models correctly “learned” the relationship between temperature or/and pressure on the related properties of ILs. Such a relationship is the knowledge automatically learned by the model after developing QSARs models, which is unveiled by the SHAP method. When the model made the predictions, it used this knowledge to make the predictions. Because this knowledge is

consistent with the experimental fact, we can conclude that our model is trustful.

3.2.2. The effect of atom groups on the refractive index and viscosity

Next, we interpreted how atom groups of IL affect the refractive index and viscosity. When the atom groups are absent, i.e., blue points in patterns, the SHAP values are close to 0 for most of the features, indicating that they contributed negligibly toward the predictions. This is reasonable because atom groups should contribute 0 if they are absent in ILs. Table 4 lists the top-5 atom groups that had the largest effect on the predictions in these four models. Here, the largest effect means the sum of SHAP values for certain atom groups in all the ILs is largest. But, its SHAP value did not have to be the largest for a specific IL. For example, Cl^- had the largest decreasing effect on the viscosity for the dataset of Viscosity-1, because the sum of SHAP values for Cl^- in all the ILs was the largest. But for an individual specific IL, Cl^- did not have to be the largest SHAP value when compared with other atom groups.

The F atom in anions of IL (Feature 7990) was found to decrease the refractive index if it is present in the ILs. This is consistent with the experimental fact that anions containing F atom often show low refractive index [9]. Likewise, B- that is often combined with F atom to form BF_4 also decreases the refractive index if it is present in anions of IL, which is also consistent with the experimental fact [9]. For comparison, the model “thought” that the presence of aromatic carbons in anions and cations of IL (Features 3643 and 5700) can increase the refractive index, as indicated by their positive Shapely values. This is also consistent with the experimental facts that ILs containing aromatic groups often show higher refractive indexes [9]. For the viscosity, the viscosities of ILs generally increase with the length of the alkyl side chain, which is ascribed to the fact that increasing the length of the alkyl side chain will increase van der Waals interaction and thus the viscosity [38,39]. The model also “thought” the presence of alkyl chain can (Feature 212) increase the viscosity, which is consistent with the experimental observations [37]. The F atom in the anions can decrease the viscosity indicated by the negative Shapely value. This is also consistent with the experimental observation that the anions containing F atom often show low viscosity [37]. One interesting observation is that the model thought F-P atom group (Feature 4032) can increase the viscosity, although it contains the F atom, which is different from other anions containing the F atom. This is also consistent with the experimental fact [37]. The above interpretation indicated that our models made the predictions based on the reasonable understanding of how the features affect the properties of IL, which made our models trustful.

4. Conclusion

This study demonstrated that MF-XGBoost is an effective way to develop QSAR models to predict the refractive index and viscosity of ILs, which should achieve wide applications for predicting other properties of ILs, such as CO_2 adsorption, conductivity, and density. Compared with the MD-based method and GCM, the MF-based method can more quickly obtain the representations of ILs and easily combine with conditions, i.e., temperature and pressure. Although the “black-box” ML algorithm (Here, it was XGBoost) was used, we interpreted our models by the SHAP method and demonstrated that the model made the predictions based on the reasonable understanding of how the features affect the related properties of ILs. For example, the temperature effect on the refractive index and viscosity was correctly “learned” by our model. This study offered a new way to more quickly develop trustful QSAR models to predict the properties of ILs.

Authorship statement

All persons who meet authorship criteria are listed as authors, and all authors certify that they have participated sufficiently in the work

to take public responsibility for the content, including participation in the concept, design, analysis, writing, or revision of the manuscript. Furthermore, each author certifies that this material or similar material has not been and will not be submitted to or published in any other publication before its appearance in the Hong Kong Journal of Occupational Therapy.

Authorship contributions

Conception and design of study: Yi Ding, Jingwen Wang.
acquisition of data: Yi Ding, Chao Guo, Peng Zhang.
analysis and/or interpretation of data: Yi Ding, Chao Guo, Peng Zhang, Jingwen Wang.
Drafting the manuscript: Yi Ding, Chao Guo, Peng Zhang, Jingwen Wang;
revising the manuscript critically for important intellectual content: Yi Ding, Chao Guo, Peng Zhang, Jingwen Wang.
Approval of the version of the manuscript to be published (the names of all authors must be listed): Yi Ding, Minchun Chen, Chao Guo, Peng Zhang, Jingwen Wang.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

All persons who have made substantial contributions to the work reported in the manuscript (e.g., technical help, writing and editing assistance, general support), but who do not meet the criteria for authorship, are named in the Acknowledgements and have given us their written permission to be named. If we have not included an Acknowledgements, then that indicates that we have not received substantial contributions from non-authors.

References

- [1] T. Welton, Room-temperature ionic liquids. Solvents for synthesis and catalysis, *Chem. Rev.* 99 (8) (1999) 2071–2084.
- [2] M. Freemantle, An Introduction to Ionic Liquids, Royal Society of chemistry 2010.
- [3] L.C. Branco, J.G. Crespo, C.A. Afonso, Studies on the selective transport of organic compounds by using ionic liquids as novel supported liquid membranes, *Chem. Eur. J.* 8 (17) (2002) 3865–3871.
- [4] M. Galiński, A. Lewandowski, I. Stepniak, Ionic liquids as electrolytes, *Electrochim. Acta* 51 (26) (2006) 5567–5580.
- [5] D. Zhao, M. Wu, Y. Kou, E. Min, Ionic liquids: applications in catalysis, *Catal. Today* 74 (1–2) (2002) 157–189.
- [6] I. Marrucho, L. Branco, L. Rebelo, Ionic liquids in pharmaceutical applications, *Annual review of chemical and biomolecular engineering* 5 (2014) 527–546.
- [7] M. Hasib-ur-Rahman, M. Sijaj, F. Larachi, Ionic liquids for CO₂ capture—development and progress, *Chem. Eng. Process. Process Intensif.* 49 (4) (2010) 313–322.
- [8] D.S. Firaha, O. Hollóczki, B. Kirchner, Computer-aided design of ionic liquids as CO₂ absorbents, *Angew. Chem. Int. Ed.* 54 (27) (2015) 7805–7809.
- [9] S. Seki, S. Tsuzuki, K. Hayamizu, Y. Umembayashi, N. Serizawa, K. Takei, H. Miyashiro, Comprehensive refractive index property for room-temperature ionic liquids, *J. Chem. Eng. Data* 57 (8) (2012) 2211–2216.
- [10] R.L. Gardas, J.A. Coutinho, Group contribution methods for the prediction of thermophysical and transport properties of ionic liquids, *AIChE J.* 55 (5) (2009) 1274–1290.
- [11] M. Sattari, A. Kamari, A.H. Mohammadi, D. Ramjugernath, A group contribution method for estimating the refractive indices of ionic liquids, *J. Mol. Liq.* 200 (2014) 410–415.
- [12] X. Wang, X. Lu, Q. Zhou, Y. Zhao, X. Li, S. Zhang, Database and new models based on a group contribution method to predict the refractive index of ionic liquids, *Phys. Chem. Chem. Phys.* 19 (30) (2017) 19967–19974.
- [13] A. Varnek, N. Kireeva, I.V. Tetko, I.I. Baskin, V.P. Solov'ev, Exhaustive QSPR studies of a large diverse set of ionic liquids: how accurately can we predict melting points? *J. Chem. Inf. Model.* 47 (3) (2007) 1111–1122.
- [14] V. Venkatraman, B.K. Alsberg, Quantitative structure-property relationship modeling of thermal decomposition temperatures of ionic liquids, *J. Mol. Liq.* 223 (2016) 60–67.
- [15] V. Venkatraman, B.K. Alsberg, Predicting CO₂ capture of ionic liquids using machine learning, *Journal of CO₂ Utilization* 21 (2017) 162–168.
- [16] Y. Zhao, Y. Huang, X. Zhang, S. Zhang, A quantitative prediction of the viscosity of ionic liquids using S α -profile molecular descriptors, *Phys. Chem. Chem. Phys.* 17 (5) (2015) 3761–3767.
- [17] V. Venkatraman, J.J. Raj, S. Evjen, K.C. Lethesh, A. Fiksdahl, In silico prediction and experimental verification of ionic liquid refractive indices, *J. Mol. Liq.* 264 (2018) 563–570.
- [18] A. Klamt, F. Eckert, COSMO-RS: a novel and efficient method for the a priori prediction of thermophysical data of liquids, *Fluid Phase Equilib.* 172 (1) (2000) 43–72.
- [19] F. Eckert, A. Klamt, Fast solvent screening via quantum chemistry: COSMO-RS approach, *AIChE J.* 48 (2) (2002) 369–385.
- [20] P. Díaz-Rodríguez, J.C. Cancilla, N.V. Plechkova, G. Matute, K.R. Seddon, J.S. Torrecilla, Estimation of the refractive indices of imidazolium-based ionic liquids using their polarisability values, *Phys. Chem. Chem. Phys.* 16 (1) (2014) 128–134.
- [21] M. Lashkarbolooki, A.Z. Hezave, S. Ayatollahi, Artificial neural network as an applicable tool to predict the binary heat capacity of mixtures containing ionic liquids, *Fluid Phase Equilib.* 324 (2012) 102–107.
- [22] D. Rogers, M. Hahn, Extended-connectivity fingerprints, *J. Chem. Inf. Model.* 50 (5) (2010) 742–754.
- [23] K.-Z. Myint, L. Wang, Q. Tong, X.-Q. Xie, Molecular fingerprint-based artificial neural networks QSAR for ligand biological activity predictions, *Mol. Pharm.* 9 (10) (2012) 2912–2923.
- [24] G. Klopmand, Concepts and applications of molecular similarity, by Mark A. Johnson and Gerald M. Maggiora, eds., John Wiley & Sons, New York, 1990, 393 pp. Price: \$65.00, *J. Comput. Chem.* 13 (4) (1992) 539–540.
- [25] M.J. McGregor, P.V. Pallai, Clustering of large databases of compounds: using the MDL “keys” as structural descriptors, *J. Chem. Inf. Comput.* 37 (3) (1997) 443–448.
- [26] K. Mansouri, A. Abdelaziz, A. Rybacka, A. Roncaglioni, A. Tropsha, A. Varnek, A. Zakharov, A. Worth, A.M. Richard, C.M. Grulke, D. Trisciuzzi, D. Fourches, D. Horvath, E. Benfenati, E. Muratov, E. Wedebye, F. Grisoni, G.F. Mangiatordi, G.M. Incisivo, H. Hong, H.W. Ng, I.V. Tetko, I. Balabin, J. Kancherla, J. Shen, J. Burton, M. Nicklaus, M. Cassotti, N.G. Nikolov, O. Nicolotti, P.L. Andersson, Q. Zang, R. Politi, R.D. Beger, R. Todeschini, R. Huang, S. Farag, S.A. Rosenberg, S. Slavov, X. Hu, R.S. Judson, CERAPP: Collaborative Estrogen Receptor Activity Prediction Project, *Environ. Health Perspect.* 124 (7) (2016) 1023–1033.
- [27] Y. Wu, G. Wang, Machine Learning Based Toxicity Prediction: From Chemical Structural Description to Transcriptome Analysis, *Int. J. Mol. Sci.* 19 (8) (2018).
- [28] S. Zhong, J. Hu, X. Fan, X. Yu, H. Zhang, A deep neural network combined with molecular fingerprints (DNN-MF) to develop predictive models for hydroxyl radical rate constants of water contaminants, *J. Hazard. Mater.* 383 (2020) 121141.
- [29] S. Zhong, K. Zhang, D. Wang, H. Zhang, Shedding light on “black box” machine learning models for predicting the reactivity of HO• radicals toward organic compounds, *Chem. Eng. J.* 126627 (2020).
- [30] T. Chen, C. Guestrin, XGBoost: A Scalable Tree Boosting System, *XGBoost: A Scalable Tree Boosting System*, 2016 785–794.
- [31] T. Chen, C. Guestrin, XGBoost: A Scalable Tree Boosting System, *arXiv* (2016) 785–794.
- [32] S.M. Lundberg, S.-I. Lee, In *A unified approach to interpreting model predictions*, *Adv. Neural Inf. Process. Syst.* 2017 (2017) 4765–4774.
- [33] B.-K. Chen, M.-J. Liang, T.-Y. Wu, H.P. Wang, A high correlate and simplified QSPR for viscosity of imidazolium-based ionic liquids, *Fluid Phase Equilib.* 350 (2013) 37–42.
- [34] R.E. Carhart, D.H. Smith, R. Venkatraghavan, Atom pairs as molecular features in structure-activity studies: definition and applications, *J. Chem. Inf. Comput. Sci.* 25 (2) (1985) 64–73.
- [35] J. Snoek, H. Larochelle, in neural information, A.-R. P., Practical bayesian optimization of machine learning algorithms, *Advances in neural information Processing Systems* 25 (NIPS 2012), 2012.
- [36] I. Dewancker, M. McCourt, S. Clark, Bayesian Optimization for Machine Learning : A Practical Guidebook, *arXiv:1612.04858* 2016.
- [37] G. Yu, D. Zhao, L. Wen, S. Yang, X. Chen, Viscosity of ionic liquids: database, observation, and quantitative structure-property relationship analysis, *AIChE J.* 58 (9) (2012) 2885–2899.
- [38] R. Hagiwara, Y. Ito, Room temperature ionic liquids of alkylimidazolium cations and fluoroanions, *J. Fluorine Chem.* 105 (2) (2000) 221–227.
- [39] Z.B. Zhou, H. Matsumoto, K. Tatsumi, Low-melting, low-viscous, hydrophobic ionic liquids: 1-alkyl (alkyl ether)-3-methylimidazolium perfluoroalkyltrifluoroborate, *Chem. Eur. J.* 10 (24) (2004) 6581–6591.