Intelligent prediction model of ammonia solubility in designable green solvents based on microstructure group contribution

Tianxiong Liu\textsuperscript{a}, Xiaojun Chu\textsuperscript{a}, Dingchao Fan\textsuperscript{a}, Zhaoyuan Ma\textsuperscript{a}, Yasen Dai\textsuperscript{a}, Zhaoyou Zhu\textsuperscript{a}, Yinglong Wang\textsuperscript{a} and Jun Gao\textsuperscript{b}

\textsuperscript{a}College of Chemical Engineering, Qingdao University of Science and Technology, Qingdao, People’s Republic of China; \textsuperscript{b}College of Chemical and Environmental Engineering, Shandong University of Science and Technology, Qingdao, People’s Republic of China

ABSTRACT

The rapid selection of environmentally friendly and efficient solvents is critical for improving the safety, environmental protection, and efficiency of a process. In this study, a deep neural network structure was proposed to predict the solubility of ammonia in ionic liquids based on molecular structure, combined with support vector machine (SVM), random forest (RF) and deep neural network (DNN) algorithm. In this study, a group-based quantisation method for ionic liquids was proposed. On this basis, a feature preprocessing method integrating feature selection and data standardisation was proposed. Then, the eigenvectors extracted from the molecular structure were used to predict the solubility of ammonia in ionic liquids using SVM, RF and DNN models. Based on the cross-validation optimisation model structure, three models were evaluated. Results showed that the three models yielded high prediction accuracy, and that the prediction accuracy of the MLP model was higher than those of the SVM and RF models. For the MLP model, the coefficient of determination was 0.992. The model has good prediction performance and generalisation ability. Therefore, it can be used to select the best ionic liquid ammonia absorbent accurately and efficiently.

1. Introduction

Ionic liquids are considered efficient and safe solvents in many industries because of their low vapour pressure and high chemical stability [1]. In recent years, ionic liquids have been widely used in extraction [2], absorption [3] and other separation processes as a substitute for organic solvents. With the widespread use of fossil fuels, the discharge of such as ammonia into the atmosphere has serious adverse effects on human beings and the environment [4]. It is found that the use of environment-friendly solvents such as ionic liquids to achieve ammonia removal has great advantages [5]. However, it is difficult and time-consuming to select ionic...
liquids by experiment in the face of complex environmental conditions and different requirements for deamination in actual production. Therefore, there is an urgent need for a simple and effective method to predict the absorption capacity of ammonia gas from the molecular structure of simple and readily available ionic liquids, to achieve efficient screening of deamination solvents [6].

In recent years, with the rapid development of machine learning models, they have been widely used in many fields, such as thermal conductivity ratio and dynamic viscosity between nanofluids [7,8], the performance of CO₂ foam injection in oil reservoirs [9], the efficiency of chemical flooding in oil reservoir [10], the porosity and permeability of oil reservoirs [11,12], the heating load of buildings' energy efficiency [13], and the solubility of gases in ionic liquids [14–16]. Using machine learning methods to predict the properties of various compounds has gradually become a research hot spot. Up to now, a lot of solubility data in ionic liquids have been obtained experimentally [17, 18]. Based on these data, various methods have been tried to predict the absorption capacity of ionic liquids [19]. Dai et al. [20] to predict the absorption capacity of ionic liquids to ammonia. Both UNIFAC and GCLF-EOS methods have achieved high accuracy. Marcus et al. [21] investigated the effects of dipole interaction and hydrogen bonding on CO₂ absorption in ionic liquids. The accurate calculation of the solubility of carbon dioxide in a non-volatile solvent such as ionic liquids is realised by using Henry's law constant. Deng et al. [22] combined the critical properties of ionic liquids with commonly used deep learning models such as convolutional neural network and deep neural network to achieve a prediction ability of carbon dioxide solubility that exceeds the traditional thermodynamic model. Ouaer et al. [23] used bayesian regularisation and other back propagation methods to optimise the multilayer perceptron. An optimised multilayer perceptron is used to predict the solubility of carbon dioxide in ionic liquids. After verification, the multilayer perceptron model and the equation of state model achieve similar high-accuracy results. Ahmadi et al. [24] achieved an accurate prediction of hydrogen sulphide solubility in ionic liquids based on gene expression programming model. The results show that the proposed model is more accurate than the thermodynamic model. [25] proposed a multilayer perceptron model to predict the solubility of hydrogen sulphide in a variety of different ionic liquids. Critical pressure, critical temperature, centrifugal factor, temperature and pressure are selected as inputs and the natural logarithm of solubility as outputs. The model achieves high accuracy in a wide range. AliShafiei et al. [26] studied the ability of artificial neural network (ANN) trained by back propagation (BP) and particle swarm optimisation (PSO) to correlate the solubility of H₂S in 11 kinds of different ILs, and proposed the optimisation architecture of neural layer. Finally, through PSO training algorithm, better results than BP training programme based on statistical standards were obtained. Shamshirband et al. [27] studied the machine learning model (MLP, PSO) and equation of state (PR, SRK) to predict ammonia solubility in ionic liquids using molecular weight, temperature, pressure, and other physical properties as inputs. The results showed that the machine learning model had satisfactory accuracy. However, the reason for its ability to absorb ammonia was not analyzed from the perspective of ionic liquid structure.

Compared with using the physical and chemical properties of molecules as descriptors to build a constitutive relationship model, using molecular structure to build a constitutive relationship model has stronger stability and a better ability to recognise isomers. Wang et al. [28] used feedforward neural network and support vector machine to predict the cytotoxicity of ionic liquid to rats directly from molecular structure. The model has strong generalisation ability and prediction performance. Uesawa et al. [29] took three-dimensional molecular conformation images as input and used Alexnet deep learning model to build a constitutive relationship model. The model has high predictive power. Su et al. [30] designed normalised molecular representations. On this basis, molecular bonds and atoms are vectorised by tree-structured long and short-memory network. Finally, the back propagation neural network is used to predict the molecular properties. The model has high accuracy and can cover more diverse molecular structures. Chen et al. [31] developed autoencoder models based on molecular structure. The encoded molecular vector is combined with the convolutional neural network. The accurate prediction of surface charge density is realised. The model’s generalisation capability is tested and applied in case studies. In conclusion, the molecular vector obtained from molecular structure has a more essential representation of molecular structural characteristics and a more effective prediction of isomerism.

Considering the nonideality of the ionic liquid system and the complexity of absorption mechanism, a reliable ammonia solubility prediction deep learning model (DNN) is established by combining the molecular structure of ionic liquid with support vector machine (SVM), random forest (RF) and deep neural network (DNN). In this work, the solubility of ammonia gas in ionic liquid under different conditions is first collected, and then the molecular characteristic vector is established to characterise the molecular structure of the ionic liquids. Based on the molecular vector optimised by feature
engineering, DNN model, SVM model and RF model are used to accurately predict the absorption capacity of ionic liquids to ammonia. The model structure is optimised by cross validation to improve the robustness of the model.

2. Experimental data

The solubility data of 10 ionic liquids to ammonia gas at different temperatures and pressures are collected from the literature [32–35]. The total number of data points is 320. A simplified molecular structure string (SMILES) is obtained from PubChem [36] to describe the structure of ionic liquids. Detailed data and corresponding strings are listed in Table S1. The combination of different ionic liquids and temperature and pressure is shown in Figure 1.

In order to convert the collected textual ionic liquid structures into numerical forms acceptable to the model. The ionic liquid is regarded as a macromolecular organic matter composed of cationic and anionic [37] smiles by analyzing ionic liquids. The molecular structure and chemical information of ionic liquid are obtained. Finally, the ionic liquid is split into molecules composed of substructures. The molecular feature vector of ionic liquid is constructed according to the frequency of the substructure. By analyzing the data, a total of 19 different molecular substructures are obtained. The corresponding substructures are provided in Table S2. The number of groups in all molecular structures is shown in Figure 2.

3. Deep learning modelling

3.1. Support vector machine

Support vector machine (SVM) is a common machine learning algorithm for data analysis and prediction modelling [38]. In the aspect of regression task, the SVM model can create a hyperplane of the feature vector in high-dimensional space to realise data fitting. The accurate SVM model can make the sample close to the hyperplane. For complex nonlinear mechanism problems, SVM has a strong ability of fitting and has a wide range of applications in molecular structural relations.

SVM framework is implemented by scikit-learn [39] in Python. The model input is the eigenvector generated by GC model, and the output layer gives the predicted solubility value. In order to realise the nonlinear modelling from descriptor to high dimensional space, the gaussian radial basis function (RBF) is selected as the kernel function. For the support vector machine model, the error tolerance of the model is determined by the regularisation parameter C. The most important regularisation parameters C and Gamma of SVM optimised by grid search. As shown in Figure 3, When the regularisation parameter C is set to 30 and gamma is set to 0.1, the SVM model has the highest accuracy.

3.2. Random forest

Random forest (RF), an integrated machine learning algorithm, is composed of decision trees and has strong quantitative analysis ability [40]. Because of the integration algorithm, the random forest almost does not appear as an over-fitting phenomenon, and it has strong adaptability when processing high-dimensional data. It has
a strong advantage for the complex structure of ionic liquids. In the training process, the out of bag error is approximated to the random forest error. Each decision tree in a random forest is built on a different sample set. Approximately one-third of the random sample observations in each sample set are excluded. These missing observations of a given tree are called out of bag samples. In the process of model selection and parameter adjustment, it is usually the key factor to find the parameters that may produce the lowest out of pocket error.

For the random forest model, the number of iterations (the number of decision trees in the RF model) is generally adjusted to control the out of bag error and determine the best model. The model input is the eigenvector generated by the GC model, and the output layer gives the predicted solubility value. As shown in Figure 4, the random forest model performs well in the training set. The number of iterations has a great influence on the results of test and evaluation. When the number of iterations is set to 5 and the maximum depth is set to 50, the model has the highest accuracy.

### 3.3. Multilayer perceptron

Multilayer perceptron (MLP) is a perceptron-based extension with high accuracy in nonlinear model construction [41]. It is a fully connected deep neural network (DNN). The complex deep neural network has a strong application value in molecular structure relations, which can quickly and accurately find the complex nonlinear relations between the molecular structure and properties of ionic liquids. In this work, a four-layer DNN model including the input layer and output layer is established.

The topology of the full connection is shown in Figure 5. The model input is the eigenvector generated by GC model, and the output layer gives the predicted solubility value. The hidden layer neurons receive a group of weighted inputs from the input layer and generate outputs for the next layer until the output layer. Based on the back propagation optimisation model parameters, the accurate prediction of ammonia solubility in ionic liquids is realised. The DNN model is implemented with the keras in Python. ‘Tanh’ is selected as the activation function to realise nonlinear data conversion.

The activation function Tanh is shown in the Equation (1):

\[
tanh(x) = \frac{1 - \exp(-2x)}{1 + \exp(-2x)}
\]  

(1)
Figure 6. MLP model parameter optimisation process.

Adagrad [42] is chosen as the optimiser. The reason is that the adagrad optimisation algorithm can adapt to each parameter of the model independently and provide different learning rates for different variables. In addition, the adagrad optimisation algorithm can update the low-frequency and high-frequency parameters, thus accelerating the convergence speed of the model and improving its robustness of the model. Mean square error is used as a loss function to optimise the model.

The DNN model is composed of many parameters, which can be divided into hyper-parameter and model parameters. The DNN model parameters are fitted from the selected training set by training. The gradient of the loss function is calculated in the model training, and the model parameters are updated by back propagation. In contrast, hyper-parameter is determined before training. Hyper-parameter has a great influence on the results of the model, so it is necessary to find the optimal hyper-parameter through a grid search algorithm. In addition, the alpha is set to 0.0162. The training batch is set to 20. The size of the hidden layer is $10 \times 10$. The optimisation process is shown in Figure 6.

### 3.4. Training process

When the model can accurately predict the data that is not in the training set, the model can be regarded as a generalisation model [43]. In the DNN model, the ionic liquid molecules are analyzed by group, and the molecular structure is quantified based on the analysis results. Then MLP, SVM and RF are used for feature correlation. To measure the generalisation capability of the model, external data sets not used for training are used to evaluate the performance of the model. The collected data are randomly divided into 80% training set and 20% test set. In order to improve the robustness of the model, five cross-validation tests are performed. The optimal GC-ML model is established by using the training set, and the three models are evaluated by using test samples.

### 3.5. Model evaluation

Generalisation ability is an important index to evaluating deep learning model [44]. Overfitting will weaken the generalisation ability of the model and reduce the prediction accuracy of the model to the external data set. In this work, the method of stopping ahead of time is used to avoid overfitting. At the same time, the external set is used to measure the generalisation ability of the final model to ensure the satisfactory performance of the final model.

In addition, the prediction ability of the evaluation model to all samples is verified by cross verification [45]. The model performance is quantified by three evaluation indexes Coefficient of determination ($R^2$), Mean square error (MSE), Mean absolute error (MAE). The detailed calculation method is shown in the Equations (2)–(4):

$$R^2 = 1 - \frac{\sum (y_{actual} - \bar{y})^2}{\sum (y_{predict} - \bar{y})^2}$$  \hspace{1cm} (2)

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_{predict} - y_{actual})^2$$  \hspace{1cm} (3)

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_{predict} - y_{actual}|$$  \hspace{1cm} (4)

### 4. Results and discussion

#### 4.1. Feature processing

For a supervised learning model, the quality of input data has a great influence on the prediction ability of the model [46]. Characterisation of molecular structure can be effectively screened out through feature engineering. Data redundancy can be effectively reduced. Based on the feature selector method [47], features with a correlation greater than 0.79, a missing rate greater than 0.6 and unique values are deleted.

There is a large order of magnitude difference between the features of the initial data set. Some abnormally small or large data will mislead the correct training of the model. In addition, if the data distribution is very scattered, it will also affect the training results. In order to solve this problem, all the features of the data set are normalised. The data set is transformed into standard normal distribution by calculating Z-score. Through the standardisation of data, the influence of numerical differences on features is solved, and the solution speed and
accuracy are improved. Feature normalisation is realised by scikit learn toolkit.

4.2. The model performance

The molecular vector of the pre-treated ionic liquid was used to predict the solubility of ammonia gas in the ionic liquid by the subsequent regression model. The results of the final three models are shown in Figure 7. The model can well predict the solubility of ionic liquid ammonia at different temperatures and pressure.

The final GC-SVM model is developed using all training samples, and then evaluated using test samples. The solubility of ammonia predicted by experiments and GC-SVM is shown in Figure 7(a). Based on the cross-validation, $R^2$, MAE and MSE are 0.906, 0.056 and 0.0048, respectively. Through these statistical indexes, it is proved that the GC-SVM model has a good prediction ability for solubility. It can be seen that the fitting ability of the model is good, and the results of the training set are close to those of the test set. Compared with the other two models, the model has a larger error.

After determining the number of random forest iterations, the final GC-RF model is developed using all training samples, and then evaluated using test samples. Based on the cross-validation, $R^2$, MAE and MSE are 0.932, 0.045 and 0.003, respectively. The GC-RF model has been proven to have good prediction ability through these statistical indicators. The prediction results of the GC-RF model are shown in Figure 7(b), which also proves that the GC-RF model can accurately fit the nonlinear relationship between ionic liquid structure and ammonia solubility.

Based on the results of model optimisation, a four-layer deep neural network model is established. The model is used to predict the absorption capacity of ionic liquids for ammonia. The results are shown in Figure 7(c). The model performs well in both the test set and the training set. Based on the cross-validation, $R^2$, MAE and MSE are 0.992, 0.014 and
0.0004, respectively. These statistical indicators prove that the group contribution model coupled with the deep neural network model has a good predictive ability. The results of the training set and test set are very close, which proves that the model has good generalisation ability. Compared with the GC-SVM and GC-RF models, the data points of the GC-MLP model are obviously more closely distributed on both sides of the diagonal. This further proves the accuracy of the model.

4.3. Model comparison

In order to fully verify the performance of the three models, the absolute errors of each data point of the three models were calculated. The comparison between the absolute errors of the three models and the experimental values is shown in Figure 8. The results show that the GC-SVM and GC-RF models have great errors. The absolute errors of the GC-SVM in all data points are significantly greater than those of GC-MLP. The absolute error of GC-RF model is larger than 0.4 when the experimental value is larger than 0.4. The GC-MLP model achieves small absolute errors in the whole range of experimental values, and can accurately predict the experimental values of ammonia solubility in ionic liquids in the range of 0-1.

The accuracy of the model is further verified by relative error. As shown in Figure 9, compared with the GC-SVM and GC-RF models, the relative errors of most data points in the GC-MLP model are between −0.2 and 0.2. When the experimental value is greater than 0.2, the calculated results of the GC-MLP model almost agree with the experimental value. The results show that the GC-MLP model can accurately predict the absorption capacity of ionic liquid to ammonia, and the accuracy of each data point is high.

4.4. External competitiveness

The predictive power of the model needs to be demonstrated by external competitiveness. By using approximate data sets, the results of the model are compared with the extreme learning machine model established by Kang et al., to test the predictive ability of the model. Baghban et al. proposed the SVM and LSSVM models to predict the ability of ionic liquids to absorb ammonia. The LSSVM model showed high accuracy with the $R^2$ of 0.9915. However, the ability of ionic liquids to absorb ammonia was not considered from the structure of ionic liquids. We compared the model proposed in this work with these models in the literatures. Detailed statistical data are shown in Table 1. In general, the cross-validated GC-SVM and GC-RF models achieved similar accuracy to those in the literature. The GC-MLP model is significantly superior to the models in the literature in determining coefficient, mean square error and mean absolute error. The results show that GC-MLP can accurately predict the solubility of ammonia in ionic liquids, which is more accurate than these models based on descriptors in the literature. It can further satisfy the rapid and accurate screening of ionic liquid ammonia absorbent.
5. Conclusion

Rapid selection of the best ammonia adsorbent can achieve efficient separation of ammonia. In order to accurately predict the absorption capacity of ionic liquid to ammonia, three constitutive models were established. This work provides a reference for the selection and development of solvents for absorbing ammonia. Before the learning process begins, the molecular structure of the ionic liquid is quantified into group interactions. On this basis, vectors representing the molecular structure of ionic liquid are generated. Finally, the prediction model was established by combining MLP, SVM and RF. The GC-MLP model has the highest accuracy and external competitiveness by determining coefficient, relative error, and absolute error evaluation. The results show that the MSE, MAE and $R^2$ of the model are 0.014, 0.0004 and 0.992, respectively, indicating that the method can describe the solubility of ammonia in ionic liquid well. To some extent, this method can replace the experiment of measuring the solubility of ammonia in ionic liquid, which is of great significance to the design of ammonia absorbent in ionic liquid.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by National Natural Science Foundation of China [grant number 22078166].

ORCID

Yinglong Wang http://orcid.org/0000-0002-3043-0891
Jun Gao http://orcid.org/0000-0003-1145-9565

References

[1] M.D. Green and T.E. Long, Polym. Rev. 49, 291–314 (2009). doi:10.1080/1558372090388914.
[2] D. Wang, F. Wei, J.A. Reyes-Labarta, Z.J. Wang, D.M. Xu, J. Gao, L.Z. Zhang and Y.L. Wang, J. Chem. Thermodyn. 155 (6), 106349 (2021). doi:10.1016/j.jct.2020.106349.
[3] J. Palomar, M. Gonzalez-Miquel, J. Bedia, F. Rodriguez and J.J. Rodriguez, Sep. Purif. Technol. 82, 43–52 (2011). doi:10.1016/j.seppur.2011.08.014.
[4] A.R. Bahmani, F. Sabzi and M. Bahmani, J. Mol. Liq. 211, 395–400 (2015). doi:10.1016/j.molliq.2015.07.009.
[5] G.F. Su, J. Zhang, J.Y. Li and P. Ning, Sep. Sci. Technol. 48, 2876–2879 (2013). doi:10.1080/01496395.2013.805781.
[6] M.H. Ahmadi, M. Alhuyi Nazari, R. Ghasempour, H. Madah, M.B. Shafii and M.A. Ahmadi, Colloid Surf. A 541, 154–164 (2018). doi:10.1016/j.colsurfa.2018.01.030.
[7] A. Yokozeki and M.B. Shiflett, Ind. Eng. Chem. Res. 46, 1605–1610 (2007). doi:10.1021/ie061260d.
[8] G.H. Li, Q. Zhou, X.P. Zhang, S.J. Zhang and J.W. Li, Fluid Phase Equilibr. 297, 34–39 (2010). doi:10.1016/j.fluid.2010.06.005.
[34] A. Yokozeki and M.B. Shiflett, Appl. Energ. 84, 1258–1273 (2007). doi:10.1016/j.apenergy.2007.02.005.

[35] Z.J. Li, X.P. Zhang, H.F. Dong, X.C. Zhang, H.S. Gao, S.J. Zhang, J.W. Li and C.M. Wang, RSC Adv. 5, 81362–81370 (2015). doi:10.1039/c5ra13730f.

[36] S. Kim, J. Chen, T.J. Cheng, A. Gindulyte, J. He, S.Q. He, Q.L. Li, B.A. Shoemaker, P.A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang and E.E. Bolton, Nucleic Acids Res. 47, D1102–D1109 (2019). doi:10.1093/nar/gky1033.

[37] M. Chipofya, H. Tayara and K.T. Chong, Int. J. Mol. Sci. 23, 12 (2022). doi:10.3390/ijms23095258.

[38] H. Benimam, C. Si-Moussa, M. Laidi and S. Hanini, Neural Comput. Appl. 32, 8635–8653 (2020). doi:10.1007/s00521-019-04356-w.

[39] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, J. Mach. Learn. Res. 12, 2825–2830 (2011). doi:arxiv-1201.0490.

[40] Ž. Kurtanjek, Chem Biochem Eng Q. 28, 459–463 (2014). doi:10.15255/CABEQ.2014.19399.

[41] A. Barati-Harooni, S. Nasery and A. Tatar, Chem. Eng. Technol. 40, 367–375 (2017). doi:10.1002/ceat.201600110.

[42] R. Ward, X.X. Wu and L. Bottou, J. Mach. Learn. Res. 21, 30 (2020). doi:arxiv-1806.01811.

[43] J. Nilsson, Artificial intelligence: a new synthesis (1999).

[44] H.C. Erythropel, J.B. Zimmerman, T.M. de Winter, L. Petitjean, F. Melnikov, C.H. Lam, A.W. Lounsbury, K.E. Mellor, N.Z. Jankovic, Q.S. Tu, L.N. Pincus, M.M. Falinski, W.B. Shi, P. Coish, D.L. Plata and P.T. Anastas, Green Chem. 20, 1929–1961 (2018). doi:10.1039/c8gc00482j.

[45] M. Tavakoli, A. Mood, D. Van Vranken and P. Baldi, J. Chem. Inf. Model. 62, 2121–2132 (2022). doi:10.1021/acs.jcim.1c01400.

[46] J.B. Yu, X.Y. Zheng and S.J. Wang, J. Process Contr. 79, 1–15 (2019). doi:10.1016/j.jprocont.2019.05.002.

[47] F. Abut, M.F. Akay and J. George, Turk. J. Electr. Eng. Co. 27, 3648–3664 (2019). doi:10.3906/elk-1808-138.

[48] X.J. Kang, Z.P. Iv, Z.B. Chen and Y.S. Zhao, Environ. Res. 189, 9 (2020). doi:10.1016/j.envres.2020.109951.

[49] A. Baghban, M. Bahadori, A.S. Lemraski and A. Bahadori, Ain Shams Eng. J. 9 (4), 1303–1312 (2018). doi:10.1016/j.asej.2016.08.006.