



# Prediction of Henry's law constants of CO<sub>2</sub> in imidazole ionic liquids using machine learning methods based on empirical descriptors

Ting Wu<sup>1</sup> · Wan-Li Li<sup>1</sup> · Meng-Yao Chen<sup>1</sup> · Yan-Mei Zhou<sup>1</sup> · Qing-You Zhang<sup>1</sup>

Received: 29 June 2020 / Accepted: 3 November 2020 / Published online: 19 November 2020  
© Institute of Chemistry, Slovak Academy of Sciences 2020

## Abstract

In this study, a total of 160 experimental data points of Henry's law constant of CO<sub>2</sub> in 32 imidazole ionic liquids (ILs) were collected, with the temperatures range from 283 to 350 K. Herein intuitive and explanatory descriptors related to Henry's law constant (HLC) were suggested from the 2D structural features of the ILs according to experimental experience and laws. Temperature was used as another variable due to its significant effect on Henry's law constant. Three machine learning methods were used to construct models to fast predict the HLC based on suggested descriptors. Multi-layer Perceptron was mainly used to build the model and compared with the results of Random forest and Multiple Linear Regression after investigating the outliers and variable selection. In addition, if only one data point was left at a similar temperature and the reduced dataset was also used to build models in the same procedure, the results were not as good as those of the full dataset but still satisfactory.

**Keywords** Henry's law constant · Ionic liquids · CO<sub>2</sub> · Machine learning · QSPR

## Introduction

Carbon dioxide, whose increase in concentration is the main cause of the greenhouse effect, is one of the main greenhouse gases. As people pay more attention to environmental issues, the technologies of carbon dioxide capture arousing more and more interests (Orhan et al. 2016). There are already some methods for capturing CO<sub>2</sub> (such as adsorption, absorption, cryogenic methods, separation with membranes, etc.), and the amine-based solvents method is currently the most commonly used method (Zhang et al. 2013). However, due to its volatility and corrosivity, this method is potentially environmentally threatening and its operation is unstable (Bahadur et al. 2015; Dutcher et al. 2015).

Ionic liquids, as a green alternative to traditional volatile solvents, have excellent properties, such as low vapor pressures, recyclable and adjustable characteristics, good thermal stability and so on (Welton 1999; Rogers and Seddon 2003; Hallett and Welton, 2011). Thus, ILs have the potential to become a new option for capturing CO<sub>2</sub> (Zhang et al. 2013). Since the first discovery that CO<sub>2</sub> was soluble in ILs was made by Blanchard et al. (1999), then some experimental and theoretical studies on the dissolution of CO<sub>2</sub> in different ILs have been performed (Marcus 2018; Baghban et al. 2015; Aghaie et al. 2019; Deng et al. 2019). Among them, imidazole ILs are favored by most researchers because they are easy to synthesize and have high adjustability.

ILs as a promising solvent for capturing CO<sub>2</sub> in the future, it is important to measure the Henry's law constant of CO<sub>2</sub> in it. Henry's law states that the amount of dissolved gas is proportional to its partial pressure in the gas phase and the proportionality factor is called the Henry's law constant (Sander 2015). It is a physical constant describing the distribution ability of a compound in the gas–liquid two-phase and it can affect the direction and velocity of migration in two phases of gas and liquids (Bhangare et al. 2018). As a ubiquitous metric, it is particularly important in chemical processing and environmental science (Wang et al. 2017; Kapelner et al. 2016). The theoretical methods for obtaining

**Electronic supplementary material** The online version of this article (<https://doi.org/10.1007/s11696-020-01415-8>) contains supplementary material, which is available to authorized users.

✉ Qing-You Zhang  
zhqingyou@aliyun.com

<sup>1</sup> Henan Engineering Research Center of Industrial Circulating Water Treatment, Henan Joint International Research Laboratory of Environmental Pollution Control Materials, Henan University, Kaifeng 475004, Henan, China

the HLCs include currently thermodynamic method, kinetic method, quantum chemistry calculations and quantitative structure–property relationship (QSPR) research (Abraham et al. 2019; Odabas and Adali 2016; Chao et al. 2017; Gorji et al. 2017; Zhao and Rafiqul 2017).

Some studies have been done about HLCs of CO<sub>2</sub> in ILs by some researchers. However, traditional experiments are expensive and time-consuming due to the large variety of ILs and thermodynamic or kinetic methods are difficult to deal with large amounts of property data. Therefore, the QSPR research is another better choice. Some properties of ILs have been studied in this way, such as toxicity, melting point, viscosity, density, molecular orbitals and so on (Barycki et al. 2018; Cerecedo Cordoba et al., 2019; Koi et al. 2019; Chen et al. 2019; Pereira et al. 2017). QSPR models have also been applied to the prediction of HLCs of CO<sub>2</sub> in ILs. Diako Ghaslani et al. developed two models with Multiple Linear Regression (MLR) and Least Squares Support Vector Machines (LS-SVM) to estimate the HLC of CO<sub>2</sub> in 32 ILs at 298 K (including imidazolium, pyridinium, pyrrolidinium and thiouronium families) (Ghaslani et al. 2017). Descriptors in most of these studies were calculated by Dragon software, and then were selected by the genetic algorithm and lacks intrinsic meaning to target property and is not intuitive enough.

In this study, a dataset consisting of 160 H's law constants of CO<sub>2</sub> in 32 imidazole ILs at different temperatures was collected. It is worth mentioning that intuitive and meaningful descriptors were manually extracted from the structures of ILs, which were similar as the descriptors in our prior research to predict the toxicity of ILs (Wu et al. 2020). The ILs in the test set were not included in the training set, i.e., the constructed models were assessed by the ILs not be used to train models. Three machine learning methods, Multi-layer Perceptron (MLP), Random forest (RF), and MLR, were used to build model, respectively, and the model established by MLP has obtained best results. The importance of variables was also studied, and some linear models were developed with MLR using just temperature variable, cation and anion descriptors, respectively, to investigate their contribution to HLC.

## Materials and methods

### Dataset

The dataset about HLC was collected from the ILs Database-IL Thermo v2.0 (updated on June 11, 2019) and related literature (Anderson et al. 2007), including the HLCs of CO<sub>2</sub> in 32 imidazole ILs, for the full names, see the Electronic Supplementary Material (ESM) S1 at different temperatures, with a total of 160 data points, temperature range from 283

to 350 K (see the Table 1). The main structures of cations and anions in this dataset are shown in Fig. 1. To assess the capability of model, the dataset is divided into training set and test set. A total of 25 data points of 6 ILs were randomly selected as the test set (indicated by\* in Table 1), and a total of 135 data points of the remaining 26 ILs were used as the training set.

However, in this dataset, some HLCs of CO<sub>2</sub> in the same IL were obtained at similar temperatures. If the difference between the two temperatures (or more than two) is less than 1.5 degrees, they were regarded as similar temperature here and the two data points were considered as similar data points. The reduced dataset was obtained if one data point was kept for similar data points. For the HLC of the same ionic liquid at similar temperatures, the processing methods are as follows: (1) If there are two HLCs at similar temperatures, one is arbitrarily retained when the difference between the two is less than 1 MPa; (2) For the case where there are more than two HLCs at similar temperatures, if there is one data point with a difference greater than 1 MPa from the average value, the data point is discarded. Then, the median value is retained when the number of remaining data is odd; the data closest to the average value are retained when the number of remaining data is even.

After deleting similar data, a reduced dataset consisting of 98 data points of the HLCs of CO<sub>2</sub> in 32 ILs was formed (see the ESM S2). Similarly, 21 data points of the above 6 ILs were used as the test set, and 77 data points of the remaining 26 ILs were used as the training set.

### Descriptors

The extraction of descriptors is an essential step to establish the relationship between the structures and its properties.

Anions are the main factors affecting the solubility of CO<sub>2</sub> in ILs as well as the HLCs (Kelley et al. 2017; Yim and Lim 2013; Cadena et al. 2004; Aki et al. 2004). The descriptors derived from anions are described as follows. For simple anions, it was recommended to use the entire structure as a descriptor. Specifically, the PF<sub>6</sub>, BF<sub>4</sub> and PF<sub>3</sub>R<sub>3</sub> groups were used as a descriptor ( $a_1$ ), and the Tf<sub>2</sub>N group was suggested as another descriptor ( $a_2$ ). For larger anions, some features of structures were suggested as descriptors. The presence of a cyanide containing anion was observed to have an effect on the HLCs. This may be due to the interaction of Lewis base and Lewis acid, which makes cyanide have an important effect on the solubility of CO<sub>2</sub> in IL. Therefore, a cyano group was suggested as a descriptor, represented by  ${}^a\text{C}\#\text{N}(a_3)$ . At the same time, because the presence of phospholipid and thioester groups has a positive effect on the HLCs, they were also suggested as two descriptors, which are denoted as  ${}^a\text{SO}_4\text{R}(a_4)$  and  ${}^a\text{PO}_4\text{R}_2(a_5)$ , respectively. In addition, it is clear from this dataset that the higher the

**Table 1** Henry's law constant of CO<sub>2</sub> in 32 ILs of 160 data set

ILs	Name	T (K)	HLC (MPa)	ILs	Name	T (K)	HLC (MPa)
1	[Bdimim][PF <sub>6</sub> ]	283.1	4.7			298.15	3.19
		298.15	6.18			298.15	3.7
		323.15	8.85			303	3.34
2	[Bmim][BF <sub>4</sub> ]	283	4.18			303.15	3.46
		283.1	4.08			303.15	3.48
		298	5.9			303.15	3.49
		298	5.27			313	4.12
		298.15	5.65			313	4.87
		303.38	6.16			313.15	4.13
		303.9	6.25			318.15	4.47
		303.93	6.29			318.15	4.44
		313	8.86			318.15	4.46
		313.15	7.735			318.15	4.51
		313.15	7.6			323	4.77
		313.15	7.87			323.15	4.91
		313.99	7.51	8 <sup>a</sup>	[Hmim][Tf <sub>2</sub> N]	283	2.42
		323.15	8.89			293.15	1.244
323.19	8.88			298	3.16		
324.06	9.06			298.1	3.44		
324.18	9.06			298.15	3.5		
334.15	10.48			313	4.36		
342.96	12.24			313	4.56		
343.83	12.58			328.1	5.37		
344.27	12.34			333.15	2.356		
3 <sup>a</sup>	[Bdimim][BF <sub>4</sub> ]	283.1	4.57			343	6.48
		298.15	6.1	9	[Pmim][PF <sub>6</sub> ]	298.15	5.2
		323.15	9.22	10	[Pmim][Tf <sub>2</sub> N]	298.15	3.7
4	[Bmim][PF <sub>6</sub> ]	283	3.88	11	[Omim][Tf <sub>2</sub> N]	298.15	3
		283.1	3.87	12	[Emim][DAC]	298.15	8.21
		283.15	3.783			303	7.9
		283.2	3.49			313.15	9.6
		293.17	4.68			313.15	9.85
		298	5.1	13	[C4H6Ndimim][DCA]	303	13.63
		298.15	5.34			303	13.626
		298.2	5.17			313	15.19
		303	5.98			323	16.29
		303.37	5.71	14 <sup>a</sup>	[C <sub>4</sub> H <sub>6</sub> N][DCA]	303	11.06
		303.38	5.71			303	11.056
303.38	5.74			313	12.72		
313	8.13			323	15.42		
313.15	6.96	15	[C <sub>4</sub> H <sub>6</sub> N][Tf <sub>2</sub> N]	303	4.21		
313.15	7.28			313	5.21		
313.29	6.82			323	5.89		
313.29	6.83	16	[Omim][BF <sub>4</sub> ]	303	5.39		
323.15	8.13			313	6.57		
323.2	8.05			323	7.56		

Table 1 (continued)

ILs	Name	<i>T</i> (K)	HLC (MPa)	ILs	Name	<i>T</i> (K)	HLC (MPa)
		323.2	8.16	17	[C <sub>4</sub> H <sub>6</sub> Ndimim] [Tf <sub>2</sub> N]	303	4.62
		323.28	8.04			313	5.57
		325.64	8.42			323	5.96
		333.07	9.44	18	[Bmim] hexadecanoate	323	7.55
		333.07	9.46	19	[Bmim] octadecenoate	323	5.92
		342.99	10.96	20	[Mmim] [EtSO <sub>4</sub> ]	298.1	13.17
		343.04	10.96			298.15	13.64
		348.2	12.1			313	17.22
5	[Emim][Tf <sub>2</sub> N]	283.1	2.53	21	[Emim][BF <sub>4</sub> ]	298.1	8.11
		298	3.65			313	10.13
		298.1	3.95			328.1	13.17
		298.15	4.3			343	16.21
		298.15	3.56	22 <sup>a</sup>	[Emim][C <sub>4</sub> H <sub>10</sub> PO <sub>4</sub> ]	313.15	6.99
		303	3.95			323.15	8.12
		303.45	4			333.15	9.66
		303.45	4.01	23	[Bmim][C <sub>8</sub> H <sub>18</sub> PO <sub>4</sub> ]	313.15	4.98
		313	5.07			323.15	5.76
		313.55	4.8			333.15	6.85
		313.55	4.83	24	[Mmim] [C <sub>2</sub> H <sub>6</sub> PO <sub>4</sub> ]	313.15	10.64
		323.15	5.15			323.15	12.72
		323.25	5.78			333.15	15.22
		323.45	5.8	25	[Omim][PF <sub>6</sub> ]	298.15	5.14
		328.1	6.38	26	[Emim][ EtSO <sub>4</sub> ]	298.15	10.03
		333.25	6.79	27	[Hmim][eFAP]	298.15	2.52
		343	7.9			333.15	4.2
		343.15	7.7	28 <sup>a</sup>	[Hmim][pFAP]	298.15	2.16
		343.35	7.7			333.15	3.6
6 <sup>a</sup>	[Edimim] [Tf <sub>2</sub> N]	283.1	2.86	29	[P <sub>3</sub> mim][bFAP]	298.15	2.02
		298.15	3.96			333.15	3.29
		323.15	6.05	30	[C <sub>6</sub> H <sub>4</sub> F <sub>9</sub> mim][Tf <sub>2</sub> N]	298.15	2.84
7	[bmim][Tf <sub>2</sub> N]	283	2.53			333.15	4.85
		293.15	2.9	31	[C <sub>8</sub> H <sub>4</sub> F <sub>13</sub> mim][Tf <sub>2</sub> N]	298.15	2.73
		298	3.3			333.15	4.47
		298	3.52	32	[Hmim][ACE]	333.15	11.31

<sup>a</sup>Test set

degree of anion fluorination, the smaller the value of the HLCs. A large number of related studies have shown that the degree of anion fluorination has a great effect on the solubility of carbon dioxide in ILs (Yim and Lim 2013; Aki et al. 2004). Therefore, the amount of fluorine was suggested here as a descriptor, denote as  ${}^n\text{F}$  ( $a_6$ ). The length of the alkyl chain in the anion is also noteworthy. It is observed that the HLC decreases with a significant increase of the length of the anionic chain. Then, another descriptor was extracted, the reciprocal of length ( $a_7$ ).

The cation plays a minor role in the effect of the HLCs (Cadena et al. 2004). Studies have shown that the presence of cationic fluorine substituents will also slightly increase the solubility of  $\text{CO}_2$  in ILs. Therefore, the number of cationic fluorine atoms was also suggested as a descriptor ( $c_1$ ). At the same time, the reciprocal of length of cationic alkyl side chains (Liu et al. 2016), the number of substituents on the imidazolium ring (Aki et al. 2004), and the existence of cyano group (Kim et al. 2014) will also slightly affect the solubility of  $\text{CO}_2$ , in other words, it also affects the HLCs. Therefore, another three descriptors were suggested about cations, represented by  $c_2$ ,  $c_3$  and  $c_4$ , respectively.

As mentioned above, eleven descriptors were suggested from the IL anion and cation structures. And because Henry's law constant changes significantly with temperature, temperature was also considered another variable denoted by  $T$ . In summary, a total of 12 descriptors related to the HLCs of  $\text{CO}_2$  were suggested.

## Methods

### Multi-layer perceptron (MLP)

The main machine learning method used in this study is Multi-layer Perception, which is an artificial neural network with a forward structure. It implements a multilayer network with hidden layers (one hidden layer was used here). In this study, the MLP Regressor package from Weka (version 3.8.3) was used (Hall et al. 2009). To avoid overfitting, according to the number of variables and the number of objects in the training set, the number of nodes in the hidden layer needs to satisfy the following formula (Andrea and Kalayeh 1991) and other parameters are default settings.

$$1.8 < \left( p = \frac{N}{M} \right) < 2.2, M = (I + 1)H + (H + 1)O \quad (1)$$

where  $N$  is the number of samples,  $I$  is the number of nodes of input layers,  $H$  is the number of nodes of hidden layers, and  $O$  is the number of nodes of output layers.

### Random forest (RF)

Random forest can be regarded as a set of classification regression trees (but without pruning), it generates a large number of trees in a random manner, and builds a model from the set of trees (Breiman 2001). Random Forests are grown with the Random Forest library in the R Program (version 2.10.1) (Ihaka and Gentleman 1996). In the training process, the training set was randomly divided into two parts for each tree, about  $2/3$  of the data are used for training and the remaining part of the data is used to evaluate the trained model. The process of cross-validation is called Out-of-bag (OOB). The method quantifies the importance of a variable by the decrease in a node's impurity every time the variable is used for splitting. The trained model was used to predict the test set that are not involved in model construction. The number of trees in the random forest is set to 1000, and default parameters were used.

### Multiple linear regression (MLR)

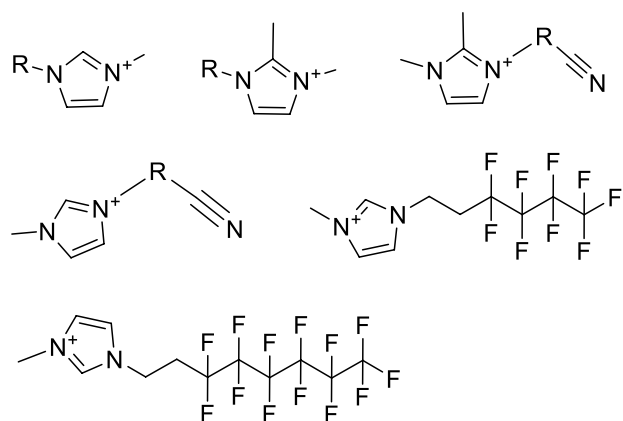
Here, a linear model was established to explore the relationship between the suggested descriptors and Henry's law constants of  $\text{CO}_2$  in ILs. The MLR algorithm used in this study is the MLR software package under Weka (version 3.8.3). The remaining parameters are default values. Select variables could be performed by M5 method. The M5 method using the Akaike criterion for model selection iteratively removes the descriptor with the smallest standardized coefficient until no improvement is observed.

## Results and discussion

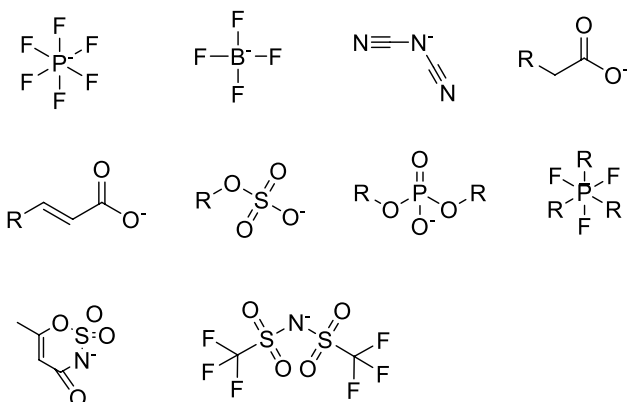
### Model construction with MLP and investigation of outliers

MLP method was used to study the relationship between the structures of imidazole ILs and HLCs. The dataset of 160 data points was divided into a training set and a test set. The result of cross-validation of the training set was  $R$  (the correlation coefficient) = 0.9819, and the result of the test set was  $R$  = 0.9758. Then, all 160 data points were considered as a whole to verify the model. To avoid overfitting, the node of hidden layer  $H$  was set to 6. The results of tenfold cross-validation of the whole dataset obtained were  $R$  = 0.9798, MAE (mean absolute error) = 0.4935 MPa, RMSE (root mean squared error) = 0.7233 MPa.

It is known from experience and related studies that the HLCs would gradually increase as the temperature increases,



(a) The main structures of cation



(b) The main structures of anions

**Fig. 1** The main structures of cations and anions of 32 imidazole ILs

which is consistent with the results of observational experimental data, and the predictions also confirm this. However, by analyzing the data, it is found that there are unreasonable experimental values in the dataset. Taking the ionic liquid 1-Hexyl-3-methylimidazolium bis(trifluoromethylsulfonyl) imide ([Hmim][Tf<sub>2</sub>N]) as an example, the results are shown in Table 2. The experimental HLC at 293.15 K should be between the corresponding HLC of 283 K and 298.1 K (HLC is 2.42 MPa and 3.44 MPa, respectively), while the experimental value was actually 1.244 MPa. This HLC is obviously too low and leading to a large error for the predicted value. The HLC at 333.15 K has a similar situation. It can be seen that the data collected in the database is not always reasonable, and the established model (the high error of these two data points) also implies that such experimental data are possibly abnormal, the two data with high error are in bold in the Table 2. Since most of the data are obtained from experiments, the existence of errors is difficult to avoid. The purity and water content of the ILs purchased by the experimenters, as well as the accuracy of the experimental

methods and experimental instruments and so on, these all will affect the experimental values. This is also a problem of collecting dataset from the database.

In this case, a total of 5 outliers (abnormal data points) found in the whole dataset were deleted and then the remaining 155 data points were used to construct model with MLP. As a result, the correlation coefficient of the tenfold cross-validation of the training set was 0.9844, the correlation coefficient of the test set reaches 0.9897, and MAE=0.4224 MPa, RMSE=0.5095 MPa. It can be clearly seen that the results have improved. The correlation between the prediction values and the experimental values of the test set before and after removing the outliers are show in Fig. 2. And the results are show in Table 3.

In a word, the results show that the predicted value obtained from the model established with the descriptors directly extracted from the structures has a good correlation with the experimental values. It proves once again the influence of the cation and cation structures of ILs on the HLCs, and of course the temperature cannot be ignored.

### Investigation of variables selection

The importance of all the 12 suggested descriptors was calculated by RF to analyze the contribution of different descriptors on HLC, and the results are listed in Table 4. It can be seen that temperature and the number of anionic fluorine atoms have a great contribution to the model to make prediction of HLC, and the cationic chain length and the structure-T<sub>2</sub>N also have a relatively great contribution. On the contrary, the contribution of the PO<sub>4</sub>R<sub>2</sub> group (*a*<sub>5</sub>) and the amount of fluorine of cations (*c*<sub>1</sub>) are relatively small.

Based on the results, variables were selected according to the importance of the variables to show the influence of the number of descriptors on the model. 4, 6, 8, and 10 variables were selected, respectively, and four models were built by MLP to compare the results with all 12 variables. The results are shown in Table 5. It can be seen that the cross-validation result of the training set based on 10 descriptors is the best. Then, another method, M5 method in MLR was used to do the variable selection. The linear formula after variable selection is as follows. From the formula, it can be found that descriptors *a*<sub>2</sub>, *a*<sub>5</sub>, *c*<sub>1</sub> were removed in the variable selection process, combined with the results of RF, 10 descriptors (without *a*<sub>5</sub>, *c*<sub>1</sub>) were finally used to describe the structures and build the model. The values of descriptors of the entire dataset obtained from the above rules are shown in the ESM S3.

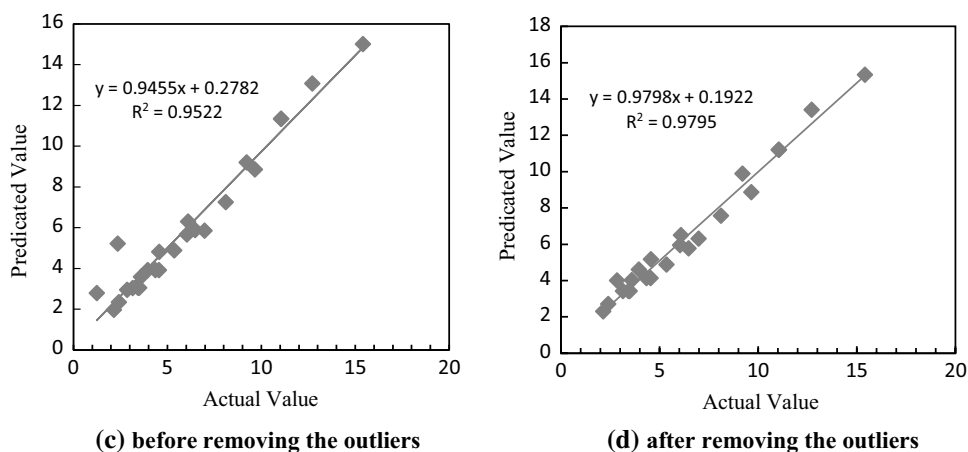
$$\begin{aligned} \text{HLC} = & 3.1653 * a_1 + 5.7563 * a_3 + 3.278 * a_4 + \\ & - 0.2231 * a_6 + 3.7325 * a_7 + 4.6724 * c_2 \\ & + 0.5903 * c_3 + 1.34 * c_4 + 0.1061 * T - 30.1029 \end{aligned} \quad (2)$$

**Table 2** Prediction results for the ionic liquid [Hmim][Tf<sub>2</sub>N]

<i>T</i> (K)	Experimental HLC (MPa)	Predicted HLC (MPa)	Error (MPa)
283	2.42	2.3482	0.0718
293.15	<b>1.244</b>	<b>2.7947</b>	– <b>1.5507</b>
298	3.16	3.0396	0.1204
298.1	3.44	3.0448	0.3952
298.15	3.5	3.0475	0.4525
313	4.36	3.9092	0.4508
313	4.56	3.9092	0.6508
328.1	5.37	4.8881	0.4819
333.15	<b>2.356</b>	<b>5.2225</b>	– <b>2.8665</b>
343	6.48	5.8758	0.6042

The two data with high error are in bold

In addition, in the formula obtained by MLR, the coefficient of each descriptor is positive or negative reflects the contribution of the corresponding descriptor to HLC is positive or negative. For example, the coefficient of  $a_6$  is  $-0.2231$ , which means  $a_6$ , the amount of fluorine in the anion has a negative contribution to HLC, that is, the more fluorine in the anion, the smaller the HLC. On the contrary, the coefficients of the remaining descriptors are all positive, that is, their corresponding structures all contribute positively to the HLC. This again shows that the descriptor in this study is not only intuitive, but also interpretable.

**Fig. 2** The relationship between predicted and experimental values of the test set before and after removing the outliers**Table 3** The results of 160 data points data set with MLP

Set	Original R/MAE/RMSE(MPa)	After deleting outliers R/MAE/RMSE(MPa)
Cross-validation of the training set	0.9819/0.4343/0.6300	0.9844/0.3591/0.5996
Prediction of the test set	0.9758/0.5185/0.7915	0.9897/0.4224/0.5095
Cross-validation of the whole data set	0.9798/0.4935/0.7233	0.9849/0.4273/0.5896

## The effects of temperature, cation and anion descriptors, respectively, on HLC

Some linear models were developed with MLR using just temperature variable, cation and anion descriptors, respectively, to discuss their contribution to HLC based on the modeling results with the entire dataset (as the training set). The results are shown in Table 6. It can be seen that the correlation coefficient between just a temperature variable and HLC is more than 0.5, indicating that temperature does have a great impact. But the coefficient is not higher probably because for different compounds, the influence of temperature is different, which cannot be reflected by a simple linear relationship, and its influence is more reflected in a single compound. In addition, it is obvious from the results that the influence of anions on HLC is indeed better than that of cations, which is consistent with our analysis and the conclusions in the literature. What is more, this also shows that the best results can be obtained by combining these three types of descriptors.

## Comparison of different machine learning methods

So as to compare the modelling results of different machine learning methods, the above 10 descriptors were used to build models through RF and MLR without variable selection, respectively. All of the results obtained by



**Table 4** The variables ranked decreasingly on the basis of variable importance of RF

	Descriptors	Importance		Descriptors	Importance
1	$T$	335.402726	7	$a_1$ (a-PF <sub>6</sub> (BF <sub>4</sub> ) P-)	49.62866
2	$a_6$ (a-nF)	304.083883	8	$c_4$ (c-C#N)	31.907484
3	$c_2$ (c-length)	193.958146	9	$a_4$ (a-SO <sub>4</sub> R)	30.22425
4	$a_2$ (a-Tf <sub>2</sub> N)	163.455462	10	$c_3$ (c-nsub)	24.219884
5	$a_7$ (a-length)	94.108388	11	$a_5$ (a-PO <sub>4</sub> R <sub>2</sub> )	5.872411
6	$a_3$ (a-C#N)	63.857103	12	$c_1$ (c-nF)	2.151513

different methods are listed in Table 7. It can be seen that these results of RF and MLR were not as good as those of MLP, and the MLR results are slightly better than the RF results. In conclusion, the prediction results of this MLP model are closer to the experimental values.

### The validation of the model

Then, for the purpose of verifying the validity of this model, a zero model was tried to establish as follows. The median value of the experimental value was used as the predicted value, and the average absolute error and the root mean square error between it and the experimental values were calculated. The results obtained were MAE = 2.5738 MPa and RMSE = 3.5125 MPa. It can be seen that compared with the results of zero model, the model developed here using 10 descriptors has more impressive results.

Leave-one-compound-out (LOCO) verification was also performed for the dataset of 155 data points to verify the model. That is, each time, one ionic liquid (including all its data points) is used as the test set, and the remaining ILs were used as the training set for modelling to predict the test set. For the dataset of the 155 data points, the results for the LOCO verification of 32 ILs were  $R = 1.2129$ , MAE = 1.6694 MPa.

### Prediction of HLC at specific constant temperature

To investigate the capability of predicting HLC at a specific temperature, a total 19 ILs were selected at 298.15 K in the dataset. Since  $a_2$  was removed in the variable selection

process by M5 method in MLR and descriptors  $a_3$ ,  $a_5$ ,  $c_4$  are all zero according to their structures, 3 descriptors of anion,  $a_1 + a_4$ ,  $a_6$ ,  $a_7$  ( $a_1$  and  $a_4$  were combined) and 3 descriptors of cation,  $c_1$ ,  $c_2$ ,  $c_3$  were finally used. Three machine learning methods were built using the 6 descriptors, separately. For MLP, the tenfold cross-validation results of the whole set were  $R = 0.9333$ , MAE = 0.5949 MPa, RMSE = 1.0019 MPa. And the tenfold cross-validation results of the whole set modelling with MLR were  $R = 0.9855$ , MAE = 0.3717 MPa, RMSE = 0.5126 MPa. Then, the OOB cross-validation with RF was performed, and the results were  $R = 0.9581$ , MAE = 1.1199 MPa, RMSE = 1.6383 MPa. From the results, it can be seen that the capability to predict HLC with the descriptors suggested here at a specific temperature is also satisfactory.

### Investigation of reduced dataset

The research of the reduced dataset was also performed then. The existing outliers in the reduced dataset were removed, that is, the remaining 94 data points were mainly used to study (their descriptors values are shown in the Electronic Supplementary Materials S3 and marked with #). When modelling with MLP, the node of hidden layer H is set to 3. The result of tenfold cross-validation of the entire dataset is  $R = 0.9801$ . The results of cross-validation of 75 data points in the training set were  $R = 0.9788$ , MAE = 0.4919 MPa, RMSE = 0.7985 MPa. For the test set, the results were  $R = 0.9814$ , MAE = 0.6407 MPa, RMSE = 0.8392 MPa. After that, RF and MLR models were also built to compare with MLP, and all the results are shown in Table 8.

Compared with the results in Table 7, it can be seen that the results become slightly worse after removing similar data points. This may be because the existence of similar

**Table 6** Comparison of the linear models using three types of descriptors

	$R$	MAE/RMSE
Just $T$	0.5164	2.0274/2.8705
Just cation	0.5047	2.345/2.8937
Just anion	0.7738	1.6231/2.1234

**Table 5** Comparison of modeling results with different numbers of variables

Variables	Training set ( $R$ /MAE/RMSE)	Test set ( $R$ /MAE/RMSE)
Top 4 important variables	0.8763/1.0838/1.6513	0.9198/0.8909/1.3788
Top 6 important variables	0.9562/0.5775/0.9883	0.9751/0.8185/1.1583
Top 8 important variables	0.9612/0.4943/0.9355	0.9864/0.8222/1.1583
Top 10 important variables	0.9865/0.3369/0.5595	0.9865/0.5171/0.6268
All 12 variables	0.9844/0.3591/0.5996	0.9897/0.4224/0.5095



**Table 7** Comparison of the three methods constructed by MLP, RF and MLR

	MLP R/MAE/RMSE <sup>a</sup>	RF R/MAE/RMSE	MLR R/MAE/RMSE
Cross-validation of the training set	0.9865/0.3369/0.5595	0.9389/0.9049/1.3000	0.9615/0.6852/0.9209
Prediction of the test set	0.9865/0.5171/0.6268	0.9528/0.9196/1.1583	0.9625/0.8809/1.0562
Cvalidation of the whole data set	0.9817/0.4818/0.6534	–	0.962/0.6879/0.9235

<sup>a</sup>The unit of MAE and RMSE is MPa

**Table 8** Results of modeling 98 data points with MLP

	MLP R/MAE/RMSE	RF R/MAE/RMSE	MLR R/MAE/RMSE
Cross-validation of the training set	0.9788/0.4919/0.7985	0.9218/1.2320/1.6825	0.9383/0.9539/1.3225
Prediction of the test set	0.9814/0.6407/0.8392	0.9535/0.9638/1.2039	0.9643/0.8517/1.0313
Cross-validation of the whole data set	0.9801/0.5185/0.7512	–	0.9406/0.9216/1.2695

data points makes the experimental values more reliable (the researchers had more chances to correct their experimental data by comparing similar data points) and the results better than those of only one experimental value. Therefore, it can be considered that if the data of HLCs of CO<sub>2</sub> at different temperatures of ILs in this dataset are more abundant, it will be more conducive to the establishment of the model, and the results will be further improved.

## Conclusion

Intuitive, simple and interpretable descriptors including 2D structure features and temperature were generated and used to construct the models for prediction of the Henry's law constant of CO<sub>2</sub> in imidazole ILs. Further, the built models were used to predict the ILs of the test set that never appeared in the training set, and satisfactory results were obtained. The different machine learning methods were attempted, and the results indicate that the models developed by MLP are better than those by the MLR and RF. If the experimental data points are more reliable, the model has the potential to be further improved. One advantage of these interpretable descriptors is that it has potential to assist in the mechanism research of targets properties by constructing models using the related structure features.

**Acknowledgements** This work was supported by the National Natural Science Foundation of China [No. 21576071; 21776061]; the Foundation of International Science and Technology Cooperation of Henan Province [No. 162102410012]; the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry [No. 20091001] and the program for Science & Technology Innovation Team in Universities of Henan Province [No.19IRTSTHN029]

## Compliance with ethical standards

**Conflict of interest** On behalf of all authors, the corresponding author states that there is no conflict of interest.

## References

- Abraham MH, Acree WE, Hoekman D (2019) A new method for the determination of Henry's law constants (air-water-partition coefficients). *Fluid Phase Equilib* 502:112300. <https://doi.org/10.1016/j.fluid.2019.112300>
- Aghaie M, Rezaei N, Zendehboudi S (2019) Assessment of carbon dioxide solubility in ionic liquid/toluene/water systems by extended PR and PC-SAFT EOSs: carbon capture implication. *J Mol Liq* 275:323–337. <https://doi.org/10.1016/j.molliq.2018.11.038>
- Aki SNVK, Mellein BR, Saurer EM (2004) High-pressure phase behavior of carbon dioxide with imidazolium-based ionic liquids. *J Phys Chem B* 108:20355–20365. <https://doi.org/10.1021/jp046895>
- Anderson JL, Dixon JK, Brennecke JF (2007) Solubility of CO<sub>2</sub>, CH<sub>4</sub>, C<sub>2</sub>H<sub>6</sub>, C<sub>2</sub>H<sub>4</sub>, O<sub>2</sub>, and N<sub>2</sub> in 1-Hexyl-3-methylpyridinium *Bis*(trifluoromethylsulfonyl)imide: comparison to other ionic liquids. *Acc Chem Res* 40:1208–1216. <https://doi.org/10.1021/ar7001649>
- Andrea TA, Kalayeh H (1991) Applications of neural networks in quantitative structure-activity relationships of dihydrofolate reductase inhibitors. *J Med Chem* 34:2824–2836. <https://doi.org/10.1021/jm00113a022>
- Baghban A, Ahmadi MA, Shahraki BH (2015) Prediction carbon dioxide solubility in presence of various ionic liquids using computational intelligence approaches. *J Supercrit Fluids* 98:50–64. <https://doi.org/10.1016/j.supflu.2015.01.002>
- Bahadur I, Osman K, Coquelet C (2015) Solubilities of carbon dioxide and oxygen in the ionic liquids methyl trioctyl ammonium *Bis*(trifluoromethylsulfonyl)imide, 1-butyl-3-methyl imidazolium *Bis*(trifluoromethylsulfonyl)imide, and 1-butyl-3-methyl imidazolium methyl sulfate. *J Phys Chem B* 119:1503–1514. <https://doi.org/10.1021/jp5061057>

- Barycki M, Sosnowska A, Jagiello K, Puzyn T (2018) Multi-objective genetic algorithm (MOGA) as a feature selecting strategy in ionic liquids' quantitative toxicity-toxicity relationship models' development. *J Chem Inf Model*. <https://doi.org/10.1021/acs.jcim.8b00378>
- Bhangare RC, Ajmal PY, Rathod TD (2018) Experimental and theoretical determination of Henry's law constant for polychlorinated biphenyls: its dependence on solubility and degree of chlorination. *Arch Environ Contam Toxicol* 76:142–152. <https://doi.org/10.1007/s00244-018-0577-z>
- Blanchard LA, Hancu D, Beckman EJ (1999) Green processing using ionic liquids and CO<sub>2</sub>. *Nature* 399:28–29. <https://doi.org/10.1038/19887>
- Breiman L (2001) Random forests. *Mach Learn* 45:5–32. <https://doi.org/10.1023/A:1010933404324>
- Cadena C, Anthony JL, Shah JK (2004) Why is CO<sub>2</sub> so soluble in imidazolium-based ionic liquids? *J Am Chem Soc* 126:5300–5308. <https://doi.org/10.1021/ja039615x>
- Cerecedo Cordoba JA, González Barbosa JJ, Frausto Solís J (2019) Melting temperature estimation of imidazole ionic liquids with clustering methods. *J Chem Inf Model* 59:7. <https://doi.org/10.1021/acs.jcim.9b00203>
- Chen MY, Wu T, Zhao TF, Zhou YM, Zhang QY, Aires-de-Sousa J (2019) Machine learning to predict the specific optical rotations of chiral fluorinated molecules. *Spectrochim Acta Part A* 223:117289. <https://doi.org/10.1016/j.saa.2019.117289>
- Deng T, Liu F, Jia G (2019) Prediction carbon dioxide solubility in ionic liquids based on deep learning. *Mol Phys* 11:1–8. <https://doi.org/10.1080/00268976.2019.1652367>
- Dutcher B, Fan M, Russell AG (2015) Amine-based CO<sub>2</sub> capture technology development from the beginning of 2013—a review. *ACS Appl Mater Interfaces* 7:2137–2148. <https://doi.org/10.1021/am507465f>
- Ghaslani D, Gorji ZE, Gorji AE (2017) Descriptive and predictive models for Henry's law constant of CO<sub>2</sub> in ionic liquids: a QSPR study. *Chem Eng Res Des* 120:15–25. <https://doi.org/10.1016/j.cherd.2016.12.020>
- Gorji AE, Gorji ZE, Riahi S (2017) Quantitative structure-property relationship (QSPR) for prediction of CO<sub>2</sub> Henry's law constant in some physical solvents with consideration of temperature effects. *Korean J Chem Eng* 34:405–415. <https://doi.org/10.1007/s11814-017-0018-0>
- Hall M, Frank E, Holmes G (2009) The WEKA data mining software: an update. *SIGKDD Explor* 11:10–18. <https://doi.org/10.1145/1656274.1656278>
- Hallett JP, Welton T (2011) Room-temperature ionic liquids: solvents for synthesis and catalysis. 2. *Chem Rev* 111:3508–3576. <https://doi.org/10.1021/cr1003248>
- Ihaka R, Gentleman R (1996) R: a language for data analysis and graphics. *J Comput Graph Stat* 5:299–314. <https://doi.org/10.2307/1390807>
- Kapelner A, Krieger A, Blanford WJ (2016) Optimal experimental designs for estimating Henry's law constants via the method of phase ratio variation. *J Chromatogr* 1468:183–191. <https://doi.org/10.1016/j.chroma.2016.08.049>
- Kelley SP, Flores LA, Shannon MS (2017) Understanding carbon dioxide solubility in ionic liquids by exploring the link with liquid clathrate formation. *Chem Eur J* 23:14332–14337. <https://doi.org/10.1002/chem.201703117>
- Kim JE, Kim HJ, Lim LS (2014) Solubility of CO<sub>2</sub> in ionic liquids containing cyanide anions: [c2mim][SCN], [c2mim][N(CN)2], [c2mim][C(CN)3]. *Fluid Phase Equilib* 367:151–158. <https://doi.org/10.1016/j.fluid.2014.01.042>
- Koi ZK, Yahya WZN, Talip RAA (2019) Prediction of the viscosity of imidazolium-based ionic liquids at different temperatures using the quantitative structure property relationship approach. *New J Chem* 43:16207. <https://doi.org/10.1039/C9NJ03436F>
- Liu XY, He MJ, Lv N (2016) Selective absorption of CO<sub>2</sub> from H<sub>2</sub>, O<sub>2</sub> and N<sub>2</sub> by 1-hexyl-3-methylimidazolium tris(pentafluoroethyl)trifluorophosphate. *J Chem Thermodyn* 97:48–54. <https://doi.org/10.1016/j.jct.2016.01.013>
- Marcus Y (2018) The solubility parameter of carbon dioxide and its solubility in ionic liquids. *J Solut Chem* 48:1025–1034. <https://doi.org/10.1007/s10953-018-0816-y>
- Odabas M, Adali M (2016) Determination of temperature dependent Henry's law constants of polychlorinated naphthalenes: application to air-sea exchange in Izmir Bay, Turkey. *Atmos Environ* 147:200–208. <https://doi.org/10.1016/j.atmosenv.2016.10.009>
- Orhan OY, Ume CS, Alper E (2016) The absorption kinetics of CO<sub>2</sub> into ionic liquid—CO<sub>2</sub> binding organic liquid and hybrid solvents. *Energy Eff Solvents CO<sub>2</sub> Capture Gas-Liq Absorpt*. [https://doi.org/10.1007/978-3-319-47262-1\\_11](https://doi.org/10.1007/978-3-319-47262-1_11)
- Pereira F, Xiao KX, Latino DARS, Wu CC, Zhang QY, Aires-de-Sousa J (2017) Machine learning methods to predict density functional theory B3LYP energies of HOMO and LUMO orbitals. *J Chem Inf Model* 57:11–21. <https://doi.org/10.1021/acs.jcim.6b00340>
- Rogers RD, Seddon KR (2003) Ionic liquids-solvents of the future? *Science* 302:792–793. <https://doi.org/10.1126/science.1090313>
- Sander R (2015) Compilation of Henry's law constants (version 4.0) for water as solvent. *Atmos Chem Phys* 15:4399–4981. <https://doi.org/10.5194/acp-15-4399-2015>
- Wang C, Yuan T, Wood SA (2017) Uncertain Henry's law constants compromise equilibrium partitioning calculations of atmospheric oxidation products. *Atmos Chem Phys*. <https://doi.org/10.5194/acp-2017-92>
- Welton T (1999) Room-temperature ionic liquids. *Solvents for synthesis and catalysis*. *Chem Rev* 99(8):2071. <https://doi.org/10.1021/cr980032t>
- Wu T, Li W, Chen M, Zhou Y, Zhang Q (2020) Estimation of ionic liquids toxicity against leukemia rat cell line IPC-81 based on the empirical-like models using intuitive and explainable fingerprint descriptors. *Mol Inf* 39:2000102. <https://doi.org/10.1002/minf.202000102>
- Yim JH, Lim JS (2013) CO<sub>2</sub> solubility measurement in 1-hexyl-3-methylimidazolium ([HMIM]) cation based ionic liquids. *Fluid Phase Equilib* 352:67–74. <https://doi.org/10.1016/j.fluid.2013.05.005>
- Zhang L, Chen J, Lv JX (2013) Progress and Development of Capture for CO<sub>2</sub> by Ionic Liquids. *Asian J Chem* 25:2355–2358. <https://doi.org/10.14233/ajchem.2013.13552>
- Zhao YS, Rafiqul I (2017) Ionic liquids for absorption and separation of gases: an eSxtensive database and a systematic screening method. *AIChE J* 63:1353–1367. <https://doi.org/10.1002/aic.15618>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.