Predicting the melting point of imidazole-based ionic liquids using QSPR model based on SMILES optimal descriptors

Wenzhen Wang\textsuperscript{1}, Bo Yang\textsuperscript{1}, Xingang Jia\textsuperscript{1, 2, *}

\textsuperscript{1} School of Chemistry and Chemical Engineer College, Xi'an Shiyou University, Xi'an, 710065, China
\textsuperscript{2} School of Chemistry and Chemical Engineer, Northwestern Polytechnical University, Xi'an, 710129, China

*Corresponding author e-mail: jiaxingang76@xsyu.edu.cn

Abstract. Recently, ionic liquids (ILs) are known as potential green materials with unique properties such as low vapor pressure, high thermal stability, and excellent solution ability of organic or inorganic materials. In this paper, the quantitative structure property relationship (QSPR) models of 111 common imidazole-based ionic liquids were established by using CORrelation and Logic (CORAL) software to predict the melting point of imidazole-based ionic liquids. The molecular structure of imidazole-based ionic liquids is represented by simplified molecular input line entry system (SMILES). The balance of correlations and the classic scheme were compared by building the models of QSPR based on the Mont Carlo method. SMILES descriptors were randomly divided into three data sets, and the results showed that the model of classic scheme was more reliable. The best correlation coefficient (R\textsuperscript{2}) values for the validation sets of two models were 0.9512 and 0.9219, respectively. Their best internal validated correlation coefficient (Q\textsuperscript{2}) values were 0.9414 and 0.8961, respectively.

1. Introduction
As we all know, ionic liquids are liquid substances composed of ions at room temperature. Ionic liquids were first discovered in 1914, and their properties are not stable. Through many years of research and exploration of ionic liquids by many scientists, a breakthrough has finally been made [1]. A series of ionic liquids have been developed, which provide a large number of valuable achievements and ideas for the later researchers. More and more people are engaged in the research of ionic liquids, because more ionic liquids have been developed and synthesized, many properties, such as strong thermal stability, excellent solubility, low vapor pressure, non-volatile, wide stable temperature range and wide electrochemical stable potential window [2-8], that can be widely used in the fields of catalysis, synthesis and analysis [9-14]. At present, the main research fields of ionic liquids are focused on the preparation of new ionic liquids, characterization of physical and chemical properties of ionic liquids, research on ionic liquids as electrolytes and solvents, etc [15-29].

Ionic liquids are called green media and green solvents [30-35]. Because of its excellent physical and chemical properties, it has a very wide range of applications in various fields. For example, because of its low melting point and good thermal conductivity, ionic liquids can be used as heat transfer fluids for solar energy collection, as adsorption media for refrigeration (heating), and as phase change heat storage
Due to the variety and quantity of ionic liquids, there are some difficulties in screening, which also limits the rapid development of ionic liquids. In order to avoid the waste of resources and time, the prediction of the melting point of ionic liquids can be used as a basis and reference method for screening ionic liquids.

Melting point is a very important physical property for ionic liquids. Quantitative structure/activity/property relationships (QSAR/QSPR) provide an accurate and reliable mathematical equation for the prediction of melting point. The QSPR model has been well used to predict the properties of compounds. Such as melting point [36,37], toxicity [38,39] and viscosity [40,41].

CORrelation And Logic (CORAL) software has been successfully used to build QSAR/QSPR models [42-45].

The molecules are uniquely represented by simplified molecular input line entry system (SMILES) to participate in the modeling process. The linear relationship between the endpoint and the SMILES descriptors is calculated by the Monte Carlo method in the CORAL software. Yan et al. described that QSAR method and genetic algorithm were used to predict the melting points of two kinds of imidazolium ionic liquids. The correlation coefficients ($R^2$) of internal validation of two models of imidazolium ionic liquids were 0.82 and 0.77, respectively. Their results display that two predictive models are true and effective [46]. Ghaedi A. showed that the QSAR model based on SMILES descriptor was constructed by CORAL software to predict the cytotoxicity of 225 kinds of ionic liquids to leukemia rats. The external test set with correlation coefficients ($R^2$) of and cross validated correlation coefficients ($Q^2$) were 0.7315-0.8760 and 0.7062-0.8490, respectively [47].
Fig. 1 The graph performance for balance of correlations and classic scheme of QSPR models of three splits.

Mehrkesh et al. introduced that new quantum chemical descriptors are used to predict the melting point and viscosity of ionic liquids. The relative average error of melting point is 3.16% [48]. Toropov et al. considered that the prediction ability of QSAR model is the index of ideality of correlation (IIC) [49]. Recently, Lotfi S et al. explained that the QSAR model was used to predict the toxicity of ionic liquids to Staphylococcus aureus. The correlation coefficient ($R^2$) of the test set is 0.8809-0.9240 [50].

CORAL software was used to build a QSPR model based on the Monte Carlo method to predict the melting points of 111 imidazole-based ionic liquids in this study. With three splits, the data randomly is divided into three groups to inspect the accuracy and reliability of QSPR models.

Table 1. Comparison of predictability of the balance of correlations and the classic scheme.

| Threshold | Probe 1 | Probe 2 | Probe 3 | Average | Dispersion |
|-----------|---------|---------|---------|---------|------------|
| $R^2_{\text{test} \ 1}$ | 0.8757  | 0.9095  | 0.9108  | 0.8987  | 0.0162     |
| $R^2_{\text{test} \ 2}$ | 0.8011  | 0.9180  | 0.8664  | 0.8618  | 0.0478     |
| $R^2_{\text{test} \ 3}$ | 0.8564  | 0.9037  | 0.8965  | 0.8855  | 0.0208     |
| $R^2_{\text{test} \ 4}$ | 0.8966  | 0.9224  | 0.7757  | 0.8649  | 0.0639     |
| $R^2_{\text{test} \ 5}$ | 0.9044  | 0.9009  | 0.8708  | 0.8920  | 0.0151     |
| $N_{\text{epoch} \ 1}$ | 19      | 8       | 20      | 15.67   | 5.44       |
| $N_{\text{epoch} \ 2}$ | 20      | 10      | 10      | 13.33   | 4.71       |
| $N_{\text{epoch} \ 3}$ | 20      | 19      | 6       | 15.00   | 6.38       |
| $N_{\text{epoch} \ 4}$ | 14      | 19      | 20      | 17.67   | 2.62       |
| $N_{\text{epoch} \ 5}$ | 18      | 15      | 13      | 15.33   | 2.05       |

| $R^2_{\text{test} \ 1}$ | 0.7917  | 0.8426  | 0.8867  | 0.8403  | 0.0388     |
| $R^2_{\text{test} \ 2}$ | 0.8132  | 0.9113  | 0.8382  | 0.8542  | 0.0416     |
| $R^2_{\text{test} \ 3}$ | 0.7145  | 0.8930  | 0.8310  | 0.8128  | 0.0740     |
| $R^2_{\text{test} \ 4}$ | 0.8412  | 0.8478  | 0.7670  | 0.8186  | 0.0366     |
| $R^2_{\text{test} \ 5}$ | 0.8334  | 0.8424  | 0.8166  | 0.8308  | 0.0107     |
| $N_{\text{epoch} \ 1}$ | 13      | 18      | 20      | 17.00   | 2.94       |
| $N_{\text{epoch} \ 2}$ | 18      | 20      | 20      | 19.33   | 0.94       |
| $N_{\text{epoch} \ 3}$ | 20      | 16      | 19      | 18.33   | 1.70       |
| $N_{\text{epoch} \ 4}$ | 20      | 20      | 20      | 20.00   | 0.00       |
| $N_{\text{epoch}}$ | 13 | 20 | 18 | 17.00 | 2.94 |
|------------------|---|---|---|---|---|
| **Split 3** | | | | | |
| $R^2_{\text{test}}$ 1 | 0.8971 | 0.9120 | 0.8837 | 0.8976 | 0.0116 |
| $R^2_{\text{test}}$ 2 | 0.8685 | 0.8911 | 0.8871 | 0.8822 | 0.0098 |
| $R^2_{\text{test}}$ 3 | 0.8538 | 0.8842 | 0.9268 | 0.8883 | 0.0300 |

| Threshold | Probe 1 | Probe 2 | Probe 3 | Average | Dispersion |
|-----------|--------|--------|--------|---------|------------|
| $R^2_{\text{test}}$ 4 | 0.8748 | 0.7677 | 0.8583 | 0.8336 | 0.0471 |
| $R^2_{\text{test}}$ 5 | 0.8005 | 0.8061 | 0.6500 | 0.7522 | 0.0723 |
| $N_{\text{epoch}}$ 1 | 6 | 17 | 20 | 14.33 | 6.02 |
| $N_{\text{epoch}}$ 2 | 14 | 20 | 12 | 15.33 | 3.40 |
| $N_{\text{epoch}}$ 3 | 17 | 19 | 17 | 17.67 | 0.94 |
| $N_{\text{epoch}}$ 4 | 17 | 19 | 19 | 18.33 | 0.94 |
| $N_{\text{epoch}}$ 5 | 19 | 17 | 20 | 18.67 | 1.25 |

**Classic scheme**

| Split 1 | | | | | |
|---------|---|---|---|---|---|
| $R^2_{\text{test}}$ 1 | 0.7303 | 0.8652 | 0.8166 | 0.8040 | 0.0558 |
| $R^2_{\text{test}}$ 2 | 0.7081 | 0.8330 | 0.8102 | 0.7838 | 0.0543 |
| $R^2_{\text{test}}$ 3 | 0.8664 | 0.8929 | 0.7422 | 0.8338 | 0.0657 |
| $R^2_{\text{test}}$ 4 | 0.8587 | 0.8993 | 0.7780 | 0.8453 | 0.0504 |
| $R^2_{\text{test}}$ 5 | 0.8363 | 0.8020 | 0.7050 | 0.7811 | 0.0556 |
| $N_{\text{epoch}}$ 1 | 20 | 16 | 20 | 18.67 | 1.89 |
| $N_{\text{epoch}}$ 2 | 13 | 20 | 20 | 17.67 | 3.30 |
| $N_{\text{epoch}}$ 3 | 20 | 20 | 19 | 19.67 | 0.47 |
| $N_{\text{epoch}}$ 4 | 20 | 19 | 20 | 19.67 | 0.47 |
| $N_{\text{epoch}}$ 5 | 20 | 20 | 20 | 20.00 | 0.00 |

| Split 2 | | | | | |
|---------|---|---|---|---|---|
| $R^2_{\text{test}}$ 1 | 0.9403 | 0.9211 | 0.9168 | 0.9261 | 0.0101 |
| $R^2_{\text{test}}$ 2 | 0.9234 | 0.9302 | 0.9114 | 0.9217 | 0.0078 |
| $R^2_{\text{test}}$ 3 | 0.9494 | 0.9188 | 0.9230 | 0.9304 | 0.0135 |
| $R^2_{\text{test}}$ 4 | 0.9330 | 0.9254 | 0.9505 | 0.9363 | 0.0105 |
| $R^2_{\text{test}}$ 5 | 0.9060 | 0.9227 | 0.9436 | 0.9241 | 0.0154 |
| $N_{\text{epoch}}$ 1 | 17 | 18 | 20 | 18.33 | 1.25 |
| $N_{\text{epoch}}$ 2 | 20 | 19 | 17 | 18.67 | 1.25 |
| $N_{\text{epoch}}$ 3 | 17 | 14 | 20 | 17.00 | 2.45 |
| $N_{\text{epoch}}$ 4 | 20 | 20 | 19 | 19.67 | 0.47 |
| $N_{\text{epoch}}$ 5 | 20 | 20 | 19 | 19.67 | 0.47 |

| Split 3 | | | | | |
|---------|---|---|---|---|---|
| $R^2_{\text{test}}$ 1 | 0.8220 | 0.8774 | 0.8496 | 0.8496 | 0.0226 |
| $R^2_{\text{test}}$ 2 | 0.8015 | 0.7669 | 0.8339 | 0.8008 | 0.0274 |
| $R^2_{\text{test}}$ 3 | 0.8183 | 0.8505 | 0.7864 | 0.8184 | 0.0262 |
| $R^2_{\text{test}}$ 4 | 0.7877 | 0.7809 | 0.8519 | 0.8068 | 0.0320 |
| $R^2_{\text{test}}$ 5 | 0.7659 | 0.7887 | 0.8071 | 0.8072 | 0.0169 |
| $N_{\text{epoch}}$ 1 | 13 | 20 | 20 | 17.67 | 3.30 |
| $N_{\text{epoch}}$ 2 | 10 | 17 | 20 | 15.67 | 4.19 |
| $N_{\text{epoch}}$ 3 | 20 | 19 | 18 | 19.00 | 0.82 |
| $N_{\text{epoch}}$ 4 | 20 | 20 | 20 | 20.00 | 0.00 |
| $N_{\text{epoch}}$ 5 | 20 | 20 | 19 | 19.67 | 0.47 |
2. Data and Methods

2.1. Data set

The structures, melting points and formulas of 111 imidazole-based ionic liquids were obtained from Ionic Liquids Database-ILThermo (v2.0). The chemical structures of imidazole-based ionic liquids are drawn and transformed into SMILES descriptors by ChemSketch [51].

The models of balance of correlations and classic scheme are constructed by CORAL software. In the balance of correlations model, the data is divided into three parts. They are training set, calibration set and testing set. In the classic scheme model, the data is divided into two parts. They are training set and testing set.

The data characteristics of QSPR model of balance of correlations and classic scheme and the structures, melting points and formulas of 111 imidazole-based ionic liquids are shown in Table S1 and Table S2 in supplementary material, respectively.

2.2. Methods

2.2.1. Molecular descriptors. There are three optimal descriptors in CORAL software, which are Graph descriptor, SMILES descriptor and hybrid descriptor. In the descriptor of Graph, there are three kinds of molecular graphs: hydrogen-suppressed molecular graph (HSG), hydrogen filled graph (HFG), and graph of atomic orbitals (GAO). As we all know, HSG is generally chosen to participate in modeling when running CORAL software.

Next, the explanation of optimal descriptors of Graph are shown in equation (1) [47].

\[
\begin{align*}
\text{SMILES}_{\text{DCW}}(\text{Threshold}, N_{\text{epoch}}) &= \sum \text{CW}(S_k) + \sum \text{CW}(S_{SS_{k}}) \\
&= \sum \text{CW}(S_{SS_{k}}) + \text{CW}(\text{PAIR}) + \text{CW}(\text{NOSP}) + \text{CW}(\text{HALO}) \\
&= \sum \text{CW}(\text{BOND}) + \text{CW}(\text{C}_{\text{max}}) + \text{CW}(\text{N}_{\text{max}}) + \text{CW}(\text{O}_{\text{max}}) \\
&= \sum \text{CW}(\text{H}_{\text{max}})
\end{align*}
\]  

Where, \( C_5 \) and \( C_7 \) are the existence of five member rings and seven member rings. \( e_0_k \) is vertex degree (morgan’s extended connectivity of zero order) \( e_1_k, e_2_k \) and \( e_3_k \) are morgan’s extended connectivity of first, second and third orders, respectively. \( p_{2k}, p_{3k} \) and \( p_{4k} \) are the numbers of paths of length 2, 3 and 4 which started from a given vertex in Graph, respectively. \( s_{2k} \) and \( s_{3k} \) are the valence shell of second and third orders, respectively. \( n_{nk} \) is the nearest neighbour code.

There is a matrix of 10 times 10 in CORAL software. If \( e_1_k \) of the second row and \( e_2_k \) of the third column are selected, it is \( e_1_k + e_2_k \). If \( e_2_k \) of the third row and \( e_1_k \) of the second column are selected, it is \( | e_1_k - e_2_k | \). \( p_{2k}, p_{3k}, p_{4k}, s_{2k}, s_{3k} \) and \( n_{nk} \) are like this.

The expression of SMILES optimal descriptor display in equation 2.

\[
\begin{align*}
\text{SMILES}_{\text{DCW}}(\text{Threshold}, N_{\text{epoch}}) &= \sum \text{CW}(C_5) + \sum \text{CW}(C_7) \\
&= \sum \text{CW}(S_k) + \sum \text{CW}(e_0_k) + \sum \text{CW}(e_1_k) + \sum \text{CW}(e_2_k) + \sum \text{CW}(e_3_k) + \sum \text{CW}(p_{2k}) + \sum \text{CW}(p_{3k}) + \sum \text{CW}(p_{4k}) \\
&= \sum \text{CW}(s_{2k}) + \sum \text{CW}(s_{3k}) + \sum \text{CW}(s_{ss_{k}}) + \sum \text{CW}(n_{nk})
\end{align*}
\]

Where, \( S_k, S_{SS_k} \) and \( S_{SS_{k}} \) are a single, double and triple component property in SMILES attribute, respectively. NOSP and HALO indicate whether some chemical elements exist. (For example: N, O, S, P, F, CI, Br, I). BOND believes whether there are double bonds (note ‘=’), triple bonds (note ‘#’ ) and stereo-chemcial bonds (note ‘@’ and ‘@@’ ) exist, respectively. PAIR is the presence of any
pair of molecular features from BOND, NOSP, HALO together. \( C_{\text{max}} \) is the total number of rings. \( N_{\text{max}}, O_{\text{max}} \) and \( S_{\text{max}} \) are the total number of nitrogen, oxygen and sulphur atoms in the molecular structure, respectively [47].

The representation of hybrid optimal descriptors are interpreted in equation 3.

\[
\text{SMILES}_{\text{DCW}}(\text{Threshold} N_{\text{epoch}}) = \text{Graph}_{\text{DCW}}(\text{Threshold} N_{\text{epoch}})
\]

Where, the hybrid descriptors are made up of SMILES descriptors and graph descriptors. The Threshold (T) and \( N_{\text{epoch}} \) in equation (1), (2) and (3) are parameters of the Monte Carlo optimization [47]. The T is a standard for classification of components of the presentation of the molecular structure. The \( N \) is the number of epochs of the Monte Carlo optimization.

The purpose is to predict the melting point of imidazole-based ionic liquids in this paper. The following equation is the melting point of calculated imidazole-based ionic liquids.

\[
T^* = C_0 + C_1 \times \text{DCW}(\text{Threshold} N_{\text{epoch}})
\]

Where, the unit of \( T^* \) is centigrade, which is also the final calculated data. \( C_0 \) is the intercept, \( C_1 \) is the slope and DCW (Threshold, \( N_{\text{epoch}} \)) is HybridDCW (Threshold, \( N_{\text{epoch}} \)).

IIC is a measure of the predictive power of a model.

\[
\text{IIC} = R \times \frac{\min |\Delta_k|}{\max |\Delta_k|}
\]

Where, IIC is the index of ideality of correlation. \( R^2 \) is a correlations coefficient related to the data set. MAE is calculated from equation (6).

\[
\pm \text{MAE} = \frac{1}{N} \sum_{k=1}^{N} |\Delta_k|
\]

Table 2 Data set of SMILES and corresponding experimental and calculated values of melting point.

| SMILES | DCW   | Expr  | Calc  | Expr-Calc |
|--------|-------|-------|-------|-----------|
| +C[n+]1cen(CC)c1^FC(F)(F)S(=O)(=O)[N-]S(=O)(=O)C(F)F | 48.74567 | -1.2500 | -6.3086 | 5.0586 |
| +C[n+]1cen(CCC)c1^FC(F)(F)S(=O)(=O)[N-]S(=O)(=O)C(F)F | 46.05325 | -13.1200 | -9.8155 | 3.3045 |
| +C[n+]1cen(CCC)OC1^FC(F)(F)S(=O)(=O)[N-]S(=O)(=O)C(F)F | 97.49175 | -6.1000 | -16.1142 | 10.0142 |
| +C[n+]1cen(CCC)OC1^FC(F)(F)S(=O)(=O)[N-]S(=O)(=O)C(F)F | 41.21727 | -13.1200 | -9.8155 | 3.3045 |
| +C[n+]1cen(CCC)OC1^FC(F)(F)S(=O)(=O)[N-]S(=O)(=O)C(F)F | 110.92123 | 83.4700 | 74.6742 | 8.7958 |
| +C[n+]1cen(CCC)OC1^FC(F)(F)S(=O)(=O)[N-]S(=O)(=O)C(F)F | 89.14285 | 37.9500 | 46.3081 | 8.3581 |
| +C[n+]1cen(CCC)OC1^FC(F)(F)S(=O)(=O)[N-]S(=O)(=O)C(F)F | 115.90393 | 78.9500 | 81.1642 | 2.2141 |
| +C[n+]1cen(CCC)OC1^FC(F)(F)S(=O)(=O)[N-]S(=O)(=O)C(F)F | 115.90393 | 78.9500 | 81.1642 | 2.2141 |
| +C[n+]1cen(CCC)OC1^FC(F)(F)S(=O)(=O)[N-]S(=O)(=O)C(F)F | 123.14708 | 102.9500 | 90.5982 | 12.3518 |
| +C[n+]1cen(CCC)OC1^FC(F)(F)S(=O)(=O)[N-]S(=O)(=O)C(F)F | 128.30201 | 91.8500 | 97.3124 | 5.4624 |
| +C[n+]1cen(CCC)OC1^FC(F)(F)S(=O)(=O)[N-]S(=O)(=O)C(F)F | 116.91589 | 85.8400 | 82.4821 | 3.3579 |
| +C[n+]1cen(CCC)OC1^FC(F)(F)S(=O)(=O)[N-]S(=O)(=O)C(F)F | 110.05156 | 51.9500 | 73.5414 | -21.5914 |
| +C[n+]1cen(CCC)OC1^FC(F)(F)S(=O)(=O)[N-]S(=O)(=O)C(F)F | 77.99475 | 31.7879 | 31.7879 | 0.1621 |
| +C[n+]1cen(CCC)OC1^FC(F)(F)S(=O)(=O)[N-]S(=O)(=O)C(F)F | 82.40791 | 47.9500 | 37.5360 | 10.4140 |
| +C[n+]1cen(CCC)OC1^FC(F)(F)S(=O)(=O)[N-]S(=O)(=O)C(F)F | 84.37500 | 20.8500 | 14.0844 | 6.8016 |
| +C[n+]1cen(CCC)OC1^FC(F)(F)S(=O)(=O)[N-]S(=O)(=O)C(F)F | 107.48003 | 74.4300 | 70.1921 | 2.2141 |
| +C[n+]1cen(CCC)OC1^FC(F)(F)S(=O)(=O)[N-]S(=O)(=O)C(F)F | 94.98369 | 52.3300 | 53.9158 | -1.5858 |
| +C[n+]1cen(CCC)OC1^FC(F)(F)S(=O)(=O)[N-]S(=O)(=O)C(F)F | 74.05927 | 13.9500 | 26.6620 | -12.7120 |
| +C[n+]1cen(CCC)OC1^FC(F)(F)S(=O)(=O)[N-]S(=O)(=O)C(F)F | 123.14708 | 102.9500 | 90.5982 | 12.3518 |
| +C[n+]1cen(CCC)OC1^FC(F)(F)S(=O)(=O)[N-]S(=O)(=O)C(F)F | 128.30201 | 102.9500 | 90.5982 | 12.3518 |
| +C[n+]1cen(CCC)OC1^FC(F)(F)S(=O)(=O)[N-]S(=O)(=O)C(F)F | 128.30201 | 102.9500 | 90.5982 | 12.3518 |
| +C[n+]1cen(CCC)OC1^FC(F)(F)S(=O)(=O)[N-]S(=O)(=O)C(F)F | 128.30201 | 102.9500 | 90.5982 | 12.3518 |
| +C[n+]1cen(CCC)OC1^FC(F)(F)S(=O)(=O)[N-]S(=O)(=O)C(F)F | 128.30201 | 102.9500 | 90.5982 | 12.3518 |
2.3. Validation of the best model

The data set is randomly divided into three groups. The search of various T and N by associating the data Graph and SMILE descriptors, which the prediction ability of the model for predicting the melting point of imidazole-based ionic liquids based on Monte Carlo calculation is successfully improved.

In this paper, we mainly adopt internal verification, external verification and some good verified parameters to find the best predictive model.

2.3.1. Internal-Validation. Internal validation uses Leave-one-out cross validation method (LOO) to verify the predictive performance of the model. Internal validation is consisted of training set in data set. The accuracy of predictive model is determined the size of Q2. Generally, Q2 no less than 0.7 is a reasonable model [52].

\[
Q^2 = 1 - \frac{\sum(y_{obs} - y_{pred})^2}{\sum(y_{obs} - \bar{y}_{tr})^2}
\]
Where, $y_{obs}$ is observed value, $y_{pred}$ is predicted value. $\bar{y}_t$ is average observed value [47].

2.3.2. External-Validation. The method of external validation and internal verification are virtually identical. But the data of external validation are composed of test set.

$$Q^2_{F1} = 1 - \frac{\sum (y_{pred(test)} - y_{obs(test)})^2}{\sum (y_{obs(test)} - \bar{y}_t)^2}$$

(8)

$$Q^2_{F2} = 1 - \frac{\sum (y_{obs(test)} - y_{pred(test)})^2}{\sum (y_{obs(test)} - \bar{y}_{ext})^2}$$

(9)

Where, $y_{pred}$ is the predicted value. $y_{obs}$ is the observed value. $\bar{y}_t$ is the average observed value. The values in above mention are very similarly in $Q^2_{F1}$ and $Q^2_{F2}$, but they are a little different. $\bar{y}_t$ is the average observed value of the test set compounds, $\bar{y}_{ext}$ is the average observed value of the training set compounds [50]. The difference of between $Q^2_{F1}$ with $Q^2_{F2}$ is smaller, the predictive ability of the model is more reliable.

2.3.3. Data randomization (Y-randomization). Data randomization test is used to inspect the robustness of the model. For a reliable model, from the perspective of equation, $R^2$ should be greater than $R^2r$, and from the perspective stability of model, $R^2$ should also be greater than $R^2r$ [53].

$$C_{R^2_p} = R \sqrt{R^2 - R^2_t}$$

(10)

Where, $R$ is the correlation coefficient of non-randomized model. $R^2$ is the determination coefficient of the non-randomized model. $R^2_t$ is the average determination coefficient of the randomized model. $C_{R^2_p}$ links $R^2$ and $R^2_t$ together to evaluate the reliability of a model. The value of $C_{R^2_p}$ should be greater than 0.5 for QSAR model to be accepted [54].

| Table 3 Criteria predictability for best QSPR models for three splits. |
|--------------------------------------------------|
| **Balance of correlations**                        |
| **x-experimental values**                           |
| $r^2_{v(x,y)}$                                      |
| $11$  | $16$  | $18$  | $15$  | $20$  | $33$  |
| $0.9108$  | $0.9113$  | $0.9219$  | $0.8993$  | $0.9512$  | $0.8774$  |
| $r^2_0$                                      |
| $0.9038$  | $0.9113$  | $0.7210$  | $0.7316$  | $0.8585$  | $0.8668$  |
| $\bar{y}_t$                                          |
| $0.9099$  | $0.99082$  | $0.8436$  | $0.8420$  | $0.9126$  | $0.773$  |
| $x$-calculated values                              |
| $r^2$                                     |
| $0.0077$  | $0.0000$  | $0.2179$  | $0.1865$  | $0.0688$  | $0.0120$  |
| $\bar{y}_{ext}$                                          |
| $0.0010$  | $0.0034$  | $0.0850$  | $0.0637$  | $0.0407$  | $0.0000$  |
| $k$                                               |
| $0.9307$  | $0.8142$  | $0.7423$  | $0.7844$  | $0.8140$  | $0.8751$  |
| $\bar{y}_{ext(test)}$                              |
| $1.0062$  | $1.1526$  | $1.2470$  | $1.1892$  | $1.1486$  | $1.0811$  |
| $R^2_{ad(test)}$                                   |
| $0.8347$  | $0.9080$  | $0.5087$  | $0.5310$  | $0.70880$  | $0.7873$  |

| **Classical scheme**                                |
|--------------------------------------------------|
| **x-experimental values**                           |
| $r^2_{v(x,y)}$                                      |
| $11$  | $16$  | $18$  | $15$  | $20$  | $33$  |
| $0.9108$  | $0.9113$  | $0.9219$  | $0.8993$  | $0.9512$  | $0.8774$  |
| $r^2_0$                                      |
| $0.9038$  | $0.9113$  | $0.7210$  | $0.7316$  | $0.8585$  | $0.8668$  |
| $\bar{y}_t$                                          |
| $0.9099$  | $0.99082$  | $0.8436$  | $0.8420$  | $0.9126$  | $0.773$  |
| $x$-calculated values                              |
| $r^2$                                     |
| $0.0077$  | $0.0000$  | $0.2179$  | $0.1865$  | $0.0688$  | $0.0120$  |
| $\bar{y}_{ext}$                                          |
| $0.0010$  | $0.0034$  | $0.0850$  | $0.0637$  | $0.0407$  | $0.0000$  |
| $k$                                               |
| $0.9307$  | $0.8142$  | $0.7423$  | $0.7844$  | $0.8140$  | $0.8751$  |
| $\bar{y}_{ext(test)}$                              |
| $1.0062$  | $1.1526$  | $1.2470$  | $1.1892$  | $1.1486$  | $1.0811$  |
| $R^2_{ad(test)}$                                   |
| $0.8347$  | $0.9080$  | $0.5087$  | $0.5310$  | $0.70880$  | $0.7873$  |

| $\Delta R^2_{ad}$                                  |
|--------------------------------------------------|
| $0.0485$  | $0.0472$  | $0.1552$  | $0.1530$  | $0.0562$  | $0.0864$  |
\[
\frac{(r^2-r_0^2)}{r^2} \text{ should be } <0.1 \quad [52]
\]
\[
\frac{(r^2-rr_0^2)}{r^2} \text{ should be } <0.1 \quad [52]
\]
K should be 0.85 < k < 1.15 [52]
kk should be 0.85 < kk < 1.15 [52]
\[
R^2_m(\text{test}) \text{ and } R^2m(\text{test}) \text{ should be } >0.5 \quad [52]
\]
\[
R^2m(\text{average}) \text{ should be larger 0.5} \quad [55]
\]
\[
\Delta R^2m \text{ should be lower 0.2} \quad [54]
\]

3. Result and discussion
The best QSPR model of balance of correlations and classic scheme is established by searching the best of Threshold and Nepoch in CORAL software. The values of the following equation are calculated according to equation (4).

Balance of correlations-Split 1
\[
T^* = -90.0010124(\pm 0.5845760) + 1.6774475(\pm 0.0068237) \times DCW(1,16)
\]
Balance of correlations-Split 2
\[
T^* = -14.7171308(\pm 0.3183496) + 1.4499227(\pm 0.0061842) \times DCW(2,19)
\]
Balance of correlations-Split 3
\[
T^* = -76.1219048(\pm 0.5311769) + 1.6642681(\pm 0.0066326) \times DCW(1,12)
\]
Classic scheme-Split 1
\[
T^* = -15.1446612(\pm 0.2486411) + 1.1326675(\pm 0.0048619) \times DCW(4,20)
\]
Classic scheme-Split 2
\[
T^* = -47.4762362(\pm 0.2702426) + 1.5427009(\pm 0.0043036) \times DCW(4,20)
\]
Classic scheme-Split 3
\[
T^* = -90.3235183(\pm 0.4438837) + 1.2102953(\pm 0.0039759) \times DCW(1,18)
\]
The data set of two models are constructed by 3 probes of the Monte Carlo optimization of balance of correlations and Classic scheme which are randomly divided into three groups, respectively. The preference of model was a model with threshold of 4 for split 2 in classic scheme in Table S1.

Fig.1 shows that models are intuitive, clear and reliable. The figure 1 graphically indicates the best QSPR models for three splits in balance of correlations and classic scheme, respectively, in which Mont Carlo optimization run with best value of \(R^2\) for the test set. The figure displays there is an excellent consensus between experimental and calculated value.

The comparison of predictability of both balance of correlations and classic scheme are showed in Table 1. the research for T and Nepoch displays that preferable T is 4 and Preferable Nepoch is 19. The best data of model are display by bold. These data show that constructing model in Classic scheme is more reliable than others.

It mainly consists of SMILES descriptor of imidazole-based ionic liquids, an example of calculated DCW for best model by Mont Carlo optimize, the melting point of experimental value and calculated value, and the error of experimental value and calculated value in Table 2. These data are built by using the best \(R^2\) in CORAL. The data show that the absolute average error of between experimental value and calculated value is 0.7242.
Table 4 The summery statistical qualities of the QSPR models obtained for Prediction of melting point of imidazole-based ionic liquids for three random splits.

| Split  | n   | R²     | CCC   | IIC   | Q² | s | MAE  | F   | Q²F1 | Q²F2 |
|--------|-----|--------|-------|-------|-----|---|------|-----|------|------|
| Split-1| 51  | 0.9319 | 0.9647| 0.9282| 0.9264| 10.6 | 8.51 | 670  | 0.9185 | 0.8844 |
|        | 49  | 0.9320 | 0.9550| 0.6124| 0.9264| 11.5 | 8.74 | 644  | 0.9118 | 0.8844 |
|        | 11  | 0.9108 | 0.9517| 0.9543| 0.8855| 12.7 | 9.09 | 92   | 0.9118 | 0.8844 |
|        | 96  | 0.8963 | 0.9453| 0.8011| 0.8915| 12.6 | 9.69 | 812  | 0.9118 | 0.8844 |
|        | 15  | 0.8993 | 0.8470| 0.9483| 0.8726| 17.5 | 116  |      |      |      |
| Split-2| 48  | 0.9592 | 0.9792| 0.9010| 0.9552| 7.08 | 5.54 | 1081 | 0.8844 | 0.8436 |
|        | 47  | 0.9338 | 0.9587| 0.8074| 0.9279| 12.3 | 9.71 | 635  | 0.8844 | 0.8436 |
|        | 16  | 0.9113 | 0.9334| 0.9546| 0.8814| 17.3 | 13.1 | 144  | 0.8844 | 0.8436 |
|        | 91  | 0.9401 | 0.9691| 0.8686| 0.99374| 9.09 | 7.34 | 1397 | 0.8844 | 0.8436 |
|        | 20  | 0.9512 | 0.9298| 0.9676| 0.9414| 16.7 | 13.4 | 351  | 0.8844 | 0.8436 |
| Split-3| 46  | 0.9686 | 0.9841| 0.9022| 0.9662| 7.34 | 5.28 | 1357 | 0.7678 | 0.8663 |
|        | 47  | 0.9364 | 0.9662| 0.8522| 0.9308| 10.4 | 7.89 | 1357 | 0.7678 | 0.8663 |
|        | 18  | 0.9219 | 0.8265| 0.9602| 0.8961| 18.7 | 15.7 | 189  | 0.7678 | 0.8663 |
|        | 78  | 0.9348 | 0.9663| 0.8725| 0.9316| 10.4 | 7.64 | 1091 | 0.7678 | 0.8663 |
|        | 33  | 0.8774 | 0.9231| 0.9367| 0.8628| 12.9 | 10.5 | 222  | 0.7678 | 0.8663 |

n is the number of compounds in the set; R² is correlation coefficient; CCC is concordance correlation coefficient; IIC is index of ideality of correlation; Q² is cross-validated correlation coefficient; s is standard error of estimation; MAE is mean absolute error; F is Fischer F-ratio.

The summery statistical qualities of the QSPR models obtained for Prediction of melting point of imidazole-based ionic liquids for three random splits in Table 4. n (the number of compounds in set), R², CCC (concordance correlation coefficient), IIC, Q², F (Fischer F-ratio), s (standard error of estimation) and MAE (mean absolute error) in all data in three splits are the most important. R² and F in the validation set are important for the validity and reliability of QSPR model in each split. Generally, The larger their values, the stronger their prediction ability. The best value of R² and F are 0.9212 and 189 in validation set in balance of correlations-split 3. The best value of R² and F are 0.9512 and 351 in validation set in classic scheme-split 2. The best R² and F are blackened. The best model is a model with a T of 4 for split 2 in classic scheme by comparing the relevant data according to Table S1 and Table 4. These best parameters about R², CCC, IIC, Q², s, MAE and F are 0.9512, 0.9298, 0.9676, 0.9414, 16.7, 13.4, 351.
13.4 and 351, respectively. The results explain that the predictive ability of the best model is classic scheme.

4. Conclusion

The CORAL software was used for building up QSPR models for predicting the melting point of imidazole-based ionic liquids in this study. The software above gives robust predictive models using the global SMILES descriptors which are employed to derive the correlation weights for molecular features by Monte Carlo method. The predictive reliability was estimated by three splits of the data set using balance of correlations and classic scheme methods. The results indicate that the classic scheme gives the enhancement in the predictability of the QSPR models in comparison with the balance of correlations in correlation coefficient (R²) values for the validation sets, internal validated correlation coefficient (Q²) values and other parameters. The classic scheme excellent model can be used for the prediction of melting point of imidazole-based ionic liquids, and provides reference and guidance for the selection of imidazole-based ionic liquids.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (No. 52073228), the Natural Science Foundation of Shannxi Province (NO. 2019JZ-44).

References

[1] Angell C A, Ansari Y, Zhao Z. Ionic Liquids: Past, present and future[J]. Faraday Discussions, 2011, 154.
[2] Kuhn B L, Bárbara F. Osmari, Thaise M. Heinen, et al. Dicationic imidazolium-based dicarboxylate ionic liquids: Thermophysical properties and solubility[J]. Journal of Molecular Liquids, 2020, 308:112983.
[3] A J D P, B A K S, A N S, et al. Estimation of thermodynamic properties of ionic liquids[J]. Journal of Molecular Liquids, 2020.
[4] Rocha M A A, Adriaan V D B, Schröer W, et al. Physicochemical properties of fatty acid based ionic liquids[J]. Journal of Chemical Thermodynamics, 2016, 100:156-164.
[5] Iwasaki K, Yoshii K, Tsuda T, et al. Physicochemical properties of phenyltrifluoroborate-based room temperature ionic liquids[J]. Journal of Molecular Liquids, 2017, 246:236-243.
[6] Gaida B, Brzeczek-Szafran A. Insights into the Properties and Potential Applications of Renewable Carbohydrate-Based Ionic Liquids: A Review[J]. Molecules, 2020, 25(14):3285.
[7] Zhang D, Li B, Hong M, et al. Synthesis and characterization of physicochemical properties of new ether-functionalized amino acid ionic liquids[J]. Journal of Molecular Liquids, 2020, 304:112718.
[8] Rauber D , Heib F , Dier T , et al. On the physicochemical and surface properties of 1-alkyl 3-methylimidazolium bis{(nonafluorobutyl) sulfonyl} imide ionic liquids[J]. Colloids & Surfaces A Physicochemical & Engineering Aspects, 2017, 529:169-177.
[9] Lu P , Zhao Z P , Wang X Y , et al. Understanding effect of molecular structure of imidazolone-based ionic liquids on catalytic performance for biomass inulin hydrolysis[J]. Molecular Catalysis, 2017, 435:24-32.
[10] Yang Z, Li X, Xu F, et al. Synthesis of bio-based polycarbonate via one-step melt polycondensation of isosorbide and dimethyl carbonate by dual site-functionalized ionic liquid catalysts[J]. Green Chemistry.
[11] RukaHirai, TatsukiHibino, TakaichiWatanabe, et al. One-pot synthesis of poly(ionic liquid)s with 1,2,3-triazolium-based backbones via clickable ionic liquid monomers[J]. RSC Advances, 10.
[12] Panchal B , Chang T , Qin S , et al. Optimization of soybean oil transesterification using an ionic liquid and methanol for biodiesel synthesis[J]. Energy Reports, 2019.
[13] Clark K D , Nacham O , Purslow J A , et al. Magnetic ionic liquids in analytical chemistry: A review[J]. Analytica Chimica Acta, 2016, 934:9-21.
[14] Abdulra’Uf L B, Lawal A, Sirhan A Y, et al. Review of Ionic Liquids in Microextraction Analysis of Pesticide Residues in Fruit and Vegetable Samples[J]. Chromatographia, 2019, 83(2).

[15] Fareghi-Alamdari R, Nadiri Niri M, Hazarkhani H. Synthesis and characterization of a new hydroxyl functionalized diacidic ionic liquid as catalyst for the preparation of diester plasticizers[J]. Journal of Molecular Liquids, 2017, 227:153-160.

[16] Yu J, Wheelhouse R T, Honey M A, et al. Synthesis and characterisation of novel nopyl-derived phosphonium ionic liquids[J]. Journal of Molecular Liquids, 2020, 316:113857.

[17] Kurnia K A, Sintra T E, Danten Y, et al. A simple method for preparation for preparation of a novel hydrophobic ionic liquid with a per-fluoro-tert-butoxide anion[J]. New Journal of Chemistry, 2017, 41(1): 47-50.

[18] Chaubey S A, Patra N, Mishra R. Synthesis of novel proline-based imidazolium ionic liquids[J]. Monatshefte für Chemie-Chemical Monthly, 2020, 151(9): 1409-1417.

[19] Reiß M, Brietzke A, Eickner T, et al. Synthesis of novel carbohydrate based pyridinium ionic liquids and cytotoxicity of ionic liquids for mammalian cells[J]. RSC Advances, 2020, 10(24): 14290-14304.

[20] De Sarkar S, Mitra S, Ghosh S K. Relating the physical properties of aqueous solutions of ionic liquids with their chemical structures[J]. The European Physical Journal E, 2020, 43(8): 1-10.

[21] Equihua-Sánchez M, Barahona-Pérez L F. Physical and chemical characterization of Agave tequilana bagasse pretreated with the ionic liquid 1-ethyl-3-methylimidazolium acetate[J]. Waste and Biomass Valorization, 2019, 10(5): 1285-1294.

[22] Ikeda T. Tetra-branched Tetra-cationic Ionic Liquids: Effects of Spacer and Tail Structure on Physical Properties[J]. Bulletin of the Chemical Society of Japan, 2020.

[23] Eftekhar A, Saito T. Synthesis and properties of polymerized ionic liquids[J]. European Polymer Journal, 2017, 90: 245-272.

[24] Thigpen A S, Nestor S T, O'Brien R A, et al. Thioether-functionalized picolinium ionic liquids: synthesis, physical properties and computational studies[J]. New Journal of Chemistry, 2017, 41(4): 1625-1630.

[25] Monti D, Ponrouch A, Palacín M R, et al. Towards safer sodium-ion batteries via organic solvent/ionic liquid based hybrid electrolytes[J]. Journal of Power Sources, 2016, 324: 712-721.

[26] Mejía A, Benito E, Guzmán J, et al. Polymer/ionic liquid thermoplastic electrolytes for energy storage processed by solvent free procedures[J]. ACS Sustainable Chemistry & Engineering, 2016, 4(4): 2114-2121.

[27] Lin Z, Barbara D, Taberna P L, et al. Capacitance of Ti3C2Tx MXene in ionic liquid electrolyte[J]. Journal of Power Sources, 2016, 326: 575-579.

[28] Plylahan N, Kerner M, Lim D H, et al. Ionic liquid and hybrid ionic liquid/organoelectrolytes for high temperature lithium-ion battery application[J]. Electrochimica Acta, 2016, 216: 24-34.

[29] Hulsbosch J, De Vos D E, Binnewe K, et al. Biobased ionic liquids: solvents for a green processing industry?[J]. ACS Sustainable Chemistry & Engineering, 2016, 4(6): 2917-2931.

[30] Itoh T, Takagi Y. Laccase-Catalyzed Reactions in Ionic Liquids for Green Sustainable Chemistry[J]. ACS Sustainable Chemistry & Engineering, 2020.

[31] Harsági N, Szőllősi B, Kiss N Z, et al. MW irradiation and ionic liquids as green tools in hydrolyses and alcoholyses[J]. Green Processing and Syntheses, 2020, 10(1): 001-010.

[32] Hospital-Benito D, Lemus J, Moya C, et al. Process analysis overview of ionic liquids on CO2 chemical capture[J]. Chemical Engineering Journal, 2020, 390: 124509.

[33] Yoo C G, Pu Y, Ragauskas A J. Ionic liquids: Promising green solvents for lignocellulosic biomass utilization[J]. Current Opinion in Green and Sustainable Chemistry, 2017, 5: 5-11.

[34] Pacheco-Fernández I, Pino V. Green solvents in analytical chemistry[J]. Current Opinion in Green and Sustainable Chemistry, 2019, 18: 42-50.

[35] Sadjadi S. Magnetic (poly) ionic liquids: A promising platform for green chemistry[J]. Journal of
Molecular Liquids, 2020: 114994.

[36] Keshavarz M H, Pourretedal H R, Saberi E. A novel method for predicting melting point of ionic liquids[J]. Process Safety and Environmental Protection, 2018, 116: 333-339.

[37] Venkatraman V, Evjen S, Knuutila H K, et al. Predicting ionic liquid melting points using machine learning[J]. Journal of Molecular Liquids, 2018, 264: 318-326.

[38] Peng D, Picchioni F. Prediction of toxicity of Ionic Liquids based on GC-COSMO method[J]. Journal of Hazardous Materials, 2020: 122964.

[39] Lan T, Yan X, Yan F, et al. Norm index in QSTR work for predicting toxicity of ionic liquids on Vibrio fischeri[J]. Ecotoxicology and Environmental Safety, 2020, 205: 111187.

[40] Carrera G V S M, Nunes da Ponte M, Rebelo L P N. Chemoinformatic Approaches To Predict the Viscosities of Ionic Liquids and Ionic Liquid-Containing Systems[J]. ChemPhysChem, 2019, 20(21): 2767-2773.

[41] Mirheydari S N, Soleymani J, Jouyban-Gharamaleki V, et al. Viscosity prediction of ionic liquid+ molecular solvent mixtures at various temperatures[J]. Journal of Molecular Liquids, 2018, 263: 228-236.

[42] Skřehota O, Vařeková R S, Geidl S, et al. QSPR designer−employ your own descriptors in the automated QSAR modeling process[J]. Journal of Cheminformatics, 2012, 4(1): 1-1.

[43] Toropova A P, Toropov A A, Leszczyńska D, et al. How the CORAL software can be used to select compounds for efficient treatment of neurodegenerative diseases?[J]. Toxicology and Applied Pharmacology, 2020, 408: 115276.

[44] Kumar P, Kumar A. CORAL: QSAR models of CB1 cannabinoid receptor inhibitors based on local and global SMILES attributes with the index of ideality of correlation and the correlation contradiction index[J]. Chemometrics and Intelligent Laboratory Systems, 2020: 103982.

[45] Toropova A P, Toropov A A, Benfenati E, et al. QSAR modeling of measured binding affinity for fullerene-based HIV-1 PR inhibitors by CORAL[J]. Journal of mathematical chemistry, 2010, 48(4): 959-987.

[46] YAN C Q, WAN H, GUAN G F. Prediction of melting points for 1, 3-disubstituent imidazolium ionic liquids[J]. Acta Physico-Chimica Sinica, 2008, 24(12): 2198-2202.

[47] Ghaedi A. Predicting the cytotoxicity of ionic liquids using QSPAR model based on SMILES optimal descriptors[J]. Journal of Molecular Liquids, 2015, 208: 269-279.

[48] Mehrkesh A, Karunanithi A T. New quantum chemistry-based descriptors for better prediction of melting point and viscosity of ionic liquids[J]. Fluid Phase Equilibria, 2016, 427: 498-503.

[49] Toropov A A, Toropova A P. The index of ideality of correlation: A criterion of predictive potential of QSPR/QSAR models?[J]. Mutation Research/Genetic Toxicology and Environmental Mutagenesis, 2017, 819: 31-37.

[50] Lotfi S, Ahmadi S, Zohrabi P. QSAR modeling of toxicities of ionic liquids toward Staphylococcus aureus using SMILES and graph invariants[J]. Structural Chemistry, 2020, 31(6): 2257-2270.

[51] Österberg T, Norinder U. Prediction of drug transport processes using simple parameters and PLS statistics The use of ACD/logP and ACD/ChemSketch descriptors[J]. European journal of medicinal chemistry sciences, 2001, 12(3): 327-337.

[52] Golbraikh A, Tropsha A. Beware of q2![J]. Journal of molecular graphics and modelling, 2002, 20(4): 269-276.

[53] Ojha P K, Roy K. Comparative QSARs for antimalarial endochins: importance of descriptor-thinning and noise reduction prior to feature selection[J]. Chemometrics and Intelligent Laboratory Systems, 2011, 109(2): 146-161.

[54] Ojha P K, Mitra I, Das R N, et al. Further exploring rm2 metrics for validation of QSPR models[J]. Chemometrics and Intelligent Laboratory Systems, 2011, 107(1): 194-205.

[55] Roy P P, Roy K. QSAR studies of CYP2D6 inhibitor Aryloxypropanolamines using 2D and 3D descriptors[J]. Chemical biology & drug design, 2009, 73(4): 442-455.