Reliable and robust f(T,P,I)-QSPR models for ionic liquids enabled by balancing data distribution and LOIO-CV

Xiao Liu[1], Mengxian Yu[1], Qingzhu Jia[1], Fangyou Yan[1], Yin-Ning Zhou[2], and Qiang Wang[1]

[1]Tianjin University of Science and Technology
[2]Shanghai Jiao Tong University

August 16, 2022

Abstract

The thermodynamic properties at variable temperature and pressure, such as density (ρ) and viscosity (η) are necessary in chemical process design. The quantitative structure-property relationship (QSPR) is a quick and accurate method to obtain the properties from a large number of potential ionic liquids (ILs). The QSPR models for ρ and η may have "pseudo-high" robustness validated by leave-one-out cross-validation (LOO-CV) and weakened stability with the unbalanced data point distribution. A rigorous model evaluation method named the leave-one-ion-out cross-validation (LOIO-CV) was proposed to evaluate robustness of ILs QSPR models. Balancing the distribution of data points in ILs, two f(T,P,I)-QSPR models were developed with norm index (I) to predict ρ and η of ILs at variable temperature and pressure. LOIO-CV method can enhance the stability QSPR model in predicting the properties of ILs with new cations and anions, which is essential for data driven design of ILs.

1. Introduction

Ionic liquids (ILs), composed of organic cations and organic/inorganic anions, have been diffusely utilized in absorption and separation[1, 2], synthesis[3], catalysis[4, 5] and electrochemistry[6, 7] owing to their superior properties as gas solubility, thermal stability and low volatility. Density(ρ) and viscosity (η) are key process parameters required in a significant amount of applications such as chemical process simulation, equipment sizing, lubrication and refrigeration[8, 9]. On the other hand, some properties are often estimated in relation to these two basic attributes, such as heat capacity, speed of sound and surface tension. In face of the vast number of ILs, it is a hard task to experimentally measure the ρ and η at variable temperature and pressure of all ILs. Accordingly, computational tools are particularly important to fill the gap of ILs property database. Furthermore, computational methods are also valuable for the property-directed design of ILs.

Quantitative structure-property relationship (QSPR) is one of the commonly used approaches to calculate the physical characteristics of chemical substances[10, 11]. Up to now, QSPR has been widely applied to the field of ILs, especially in the temperature and pressure-dependent property[12-14]. By combining group contribution (GC), Kamil Paduszynski[15] built a QSPR model for estimating the ρ at different temperatures of ILs with the most comprehensive collection of data reported to date. In Das et al.'s investigation[16], based on the multilayered variable selection strategy, a QSPR model was developed for predicting the η of ILs with $Q\ ^{2}{}_{\mathrm{LOO}}= 0.713$. Mirkhani and Gharagheizi[17] proposed a linear QSPR model for predicting the η of 293 ILs using Genetic function approximation for the model's parameter selection with$R\ ^{2}{}_{\mathrm{training}} = 0.8096$. In the process of establishing the QSPR model, validation is inevitable. The external validation and internal validation are usually adopted in most studies[18, 19]. Although a few QSPR models have been developed with external validation and internal validation for the temperature and pressure-dependent properties of ILs, the stability and reliable of QSPR are challenged.

Due to the special nature of ILs - consisting of both anions and cations - ILs with both cations and anions in the training set can be obtained directly from the contributions of both cations and anions. However, most studies ignored the criterion that both cation and anion of one IL in testing set should not both reappear in training set, otherwise, it will lead to "pseudo-high" accuracy for external and internal (Leave-one-out cross-validation, LOO-CV) validations. Recently, Makarov et al.[20] analyzed the published QSPR models[21, 22] for the melting point of ILs by the five-fold cross-validation (5-CV) and found that traditional validation method has "pseudo-high" accuracy. Nevertheless, for the temperature and pressure-dependent properties of ILs as ρ and η , the data points of ILs vary considerably with variable temperature and pressure. K-fold cross-validation (K-CV) is difficult to balance the distribution of data points under different types of ions. So, it is therefore necessary to establish an easy-to-implement internal validation method to efficiently and accurately evaluate the QSPR model. Furthermore, the stability of the QSPR model is also related by the veracity and distribution of the dataset. The authenticity of experimental data was evaluated in most previous studies[15, 23-25], while the distribution of data points was ignored. For example, in the QSPR model of heat capacity developed by Sattari et al.[26], 1528 data points were used for [C4mim][PF6], which account for 41% of the total dataset. Similarly, in our previous work on heat capacity QSPR model[23], [C4mim][PF6] accounted for 21% of the total dataset. The QSPR model is usually established by the least square method, whose objective function is the minimum sum of error squares[10]. Thus, a balanced distribution of data points should be selectively collected. Further, in the case of temperature and pressure-dependent properties, the temperature and pressure terms are usually treated as constant terms for all ILs[27, 28]. The temperature and pressure terms are affected by the structure of IL based on the analysis of our previous works[23, 29, 30], so it is necessary to introduce descriptors to temperature and pressure terms.

In this contribution, two $f\ (T\ ,P\ ,I\ )$-QSPR models for ρ and η were established by a method for a balanced distribution of data points and the treatment of temperature and pressure effects according to the structures of ILs. A novel internal validation method namely the leave-one-ion-out cross-validation (LOIO-CV) was proposed to handle the "pseudo-high" accuracy of LOO-CV for ILs. These models were also validated by the external validation, which follow the principle that cation and anion do not appear in the training set and testing set simultaneously. Analysis of the statistical results showed that two models achieved good predictive power as well as stability, which is an excellent guide for future rapid screening and design of functional ILs.

2. Methodology

2.1. Database

The ILs data were collected from the National Institute of Standards and Technology (NIST)[31]. In total, 19335 ρ data points for 972 ILs and 9238 η data points for 832 ILs were included in the dataset. For ρ and η , the temperature and pressure ranges were 221.314 ~ 473.15 K and 0.0815 ~ 251.5 MPa, 253.15 ~ 438.15 K, and 0.06 ~ 300 MPa, respectively. The total dataset contains 501 cations, including imidazolium (im), pyridinium (py), pyrrolidinium (pyr), ammonium (N), phosphonium (P), piperidinium (pip), morpholinium (mor), sulfonium (S), triazolium (Trl), propylpyrazolium (pyra), etc. It contains 154 anions, such as bis[(trifluoromethyl)sulfonyl]imide [(N(SO$_2$CF$_3$)$_2$)$^-$], tetrafluoroborate [(BF$_4$)$^-$], hexafluorophosphate [(PF$_6$)$^-$], dicyanamide [(N(CN)$_2$)$^-$], tetracyanoborate [(B(CN)$_4$)$^-$], trifluoroacetate [(C(CN)$_3$)$^-$], tris(pentafluoroethyl)trifluorophosphate [(PF$_3$(C$_2$F$_5$)$_3$)$^-$], halogen [(X)$^-$], thiocyanate [(SCN)$^-$], alkoxy-alkylsulfates [(RSO$_3$)$^-$], alkyl-sulfate [(RSO$_4$)$^-$], and so on. In particular, geminal dicationic ILs (GDILs) were also collected in this work (E.g. 1-methyl-3-(3-(trimethylammonio)propyl)-1H-imidazolium bis(dicyanamide)). The information about these ILs together with corresponding experimental values of ρ and η are shown in Tables S1 ~ S2 of Supporting Information (exp-cal-values.xlsx).

2.2. Data pre-processing

In the NIST database, the vast data points at variable temperature and pressure were included for one IL. Some ILs would represent a large percentage of the dataset if all these points were collected for modeling. According to the principle of the least square method[32, 33], a large percentage of some ILs could reduce the reliability of the QSPR model. Therefore, the criteria were adopted in the process of data collection for which data points were collected at 5 K temperature and 2.5 MPa pressure intervals.

2.3. $f\ (T\ ,P\ ,I\ )$-QSPR model

$f\ (T\ ,P\ ,I\ )$-QSPR models were established to describe the relationship of ρ and η with structure, temperature and pressure[23]. The preliminary$f\ (T\ ,P\ ,I\ )$-QSPR models are shown as Eqs. (1)-(2).

ρ is the density of the ILs in units of kg[?]m^3,η is the viscosity of the ILs in units of Pa[?]s, T is the temperature in K, and P is the pressure in kPa. α is a variable related to the ILs structures. In most studies, the parametersβ , γ , and χ , are treated as constant terms for all ILs[27, 34]. From our previous works[23, 29], treating these three coefficients as variables for each IL makes the model more accurate. This strategy has hence been continued in the present work.

2.4. Proposed norm descriptors

The step matrix (MS), such as the full step matrix (MS $_{\mathrm{F}}$), the adjacent step matrix (MS $_{\mathrm{A}}$), the adjacent-interphase step matrix (MS $_{\mathrm{AB}}$) and the adjacent-interphase-jump step matrix (MS $_{\mathrm{ABC}}$) are used to reflect the connection relationship of atoms, as Eqs. (3)-(6). On this basis, two step matrices (MS $_{\mathrm{ABC\_cyc}}$ andMS $_{\mathrm{bon\_cyc}}$), given by Eqs. (7)-(8), are defined to present the interaction of adjacent-interphase-jump atom on the ring and the interaction of atoms on different bonds on the ring, respectively. To better reveal the properties of atomic in molecules, the property matrices (MP) are used as shown in Table 1. The properties of each atom were shown in S1 of Supporting Information (atom properties.xlsx).

Table 1 . The property matrices (MP).

MP	Notes
	relative atom mass
	electronegativity
	ionization energy
	atomic radius (Å)
	number of outermost electrons
	number of the electron shell
	branched degree

where s_{ij} is the step between atom i and$j$. b_{ij} is the type of chemical bond between atom i and j (The single, double, triple and benzene ring bonds are 1, 2, 3, and 1.5 respectively).

The atomic distribution matrices (MA) are grouped by $MS$and MP , which reflect the relationship between atoms and the special contribution of each atom. In addition, MP and MScorresponding to the H-suppressed structure are also obtained theMA in the same way. The norm indexes (I) are the norm of atomic distribution matrices as listed in Eqs. (9)-(14). The MAused for ρ and η are shown in Table C1-C2 of Supporting Information (atomic-distribution-matrix.docx). An example for the prediction process with two ILs generating norm indexes (I) and applying the ρ (T ,P ,I)-QSPR model is shown in E1 of Supporting Information (example.xlsx).

Where λ is the eigenvalue of matrix, M^H is the Hermite matrix.

2.5. Model validation

LOIO-CV method. The implementation process of LOO-CV is shown in Figure 1. LODPO-CV and LOILO-CV are two execution methods belonging to LOO-CV, which are often used to evaluate the robustness of IL-QSPR models. LODPO-CV is widely used because of the ease implementation. For LODPO-CV, a data point was removed to implement model validation process and the remaining data points serve as training set. The interpolating process for LODPO-CV leads to the “pseudo-high” accuracy. For LOILO-CV, all data points of one IL were removed to implement model validation process. While LOILO-CV is a better method than LODPO-CV, LOILO-CV will also produce "pseudo-high" accuracy because both cation and anion of the removed IL may appear in remaining set. In both internal and external validation of ILs QSPR models, an important criterion has been ignored: both cation and anion of one IL in testing set cannot reappear in training set. If the cation and anion of one IL in the testing set reappear in the training set at the same time, the contributions of the anion and cation have been present in the training set, so the predicting ability of the model cannot be reflected. Hence, to enhance the accuracy of model evaluation and verify the robustness of the model, the internal validation method of LOIO-CV was proposed as presented in Figure 1. It mainly includes two processes: (1) leave-one-cation-out cross-validation (LOCO-CV), in which ILs with the same cation are treated as testing set and the remaining ILs are used as training set; (2) leave-one-anion-out cross-validation (LOAO-CV), all ILs with the same anion as the validation set and the remaining ILs are used as the training set.

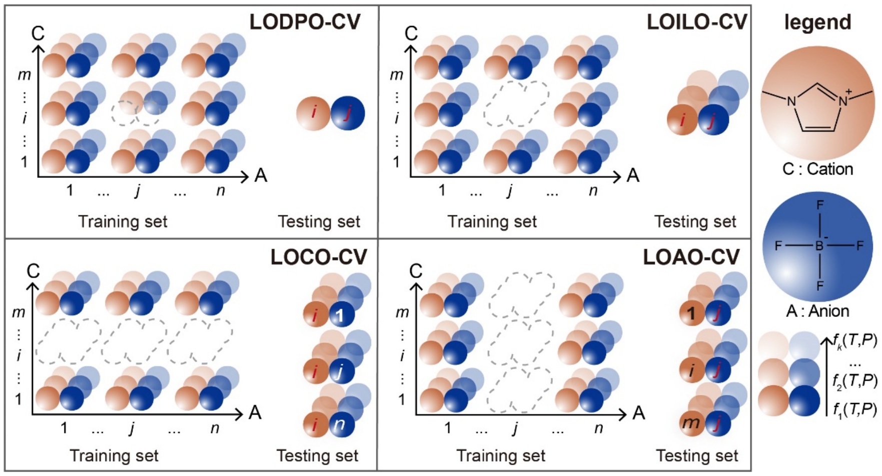

Figure 1. The comparison of leave-one-out cross-validation (LOO-CV) and leave-one-ion-out cross-validation (LOIO-CV) methods: orange ball with C_i stands for cations, blue ball with A_i stands for anions and C_i tightly close to the A_i stand for ILs, and the balls in different shades with $f_k\ (T\ ,P\ )$ represent data points at variable temperatures and pressure.

*Other validation protocols***.** The correlation coefficient ($R\ ^2$) and mean absolute error (MAE) were used to evaluate the QSPR model quantitatively[35]. The definitions of these parameters are listed in Table S1 of Supporting Information (Statistical parameters.docx). Through external validation, the predicting ability of these models was fully evaluated by$R\ ^2{}_{\text{training}}$ of the training set and $R\ ^2{}_{\text{testing}}$ of the testing set. To avoid cations and anions in the testing set reappearing in the training set in the meantime, the dataset was divided into training set (80%) and testing set (20%) by the ion structures and the proportion of data points. Y -randomization test was used to avoid the possibility of chance correlation in modeling[36].

3. Result and discussion

3.1. The results for the fitted $f\ (T\ ,P\ ,I\ )$ model

Unlike treated the β , γ and χ as constant terms in the prior work, this work introduces descriptors to correct the effects of temperature and pressure terms on the$f\ (T\ ,P\ ,I\ )$-QSPR model. To vividly show the difference between the two methods, the data points of ρ andη were fitted by using Eq. (1) and (2), and the scatterplot as shown in Figure 2. When β , γ and χ are constants for all ILs, the parameters in Figures 2 (a) and 2(c) (0.9991 and 0.9675 for ρ and η , respectively) are significantly lower than that of Figures 2 (b) and 2(d) (0.9999 and 0.9994 for ρ andη , respectively) where β , γ, and χ are variables for each IL. Therefore, it is necessary to introduce descriptors to modify the temperature and pressure terms for improving the model's reliability and predictive capability.

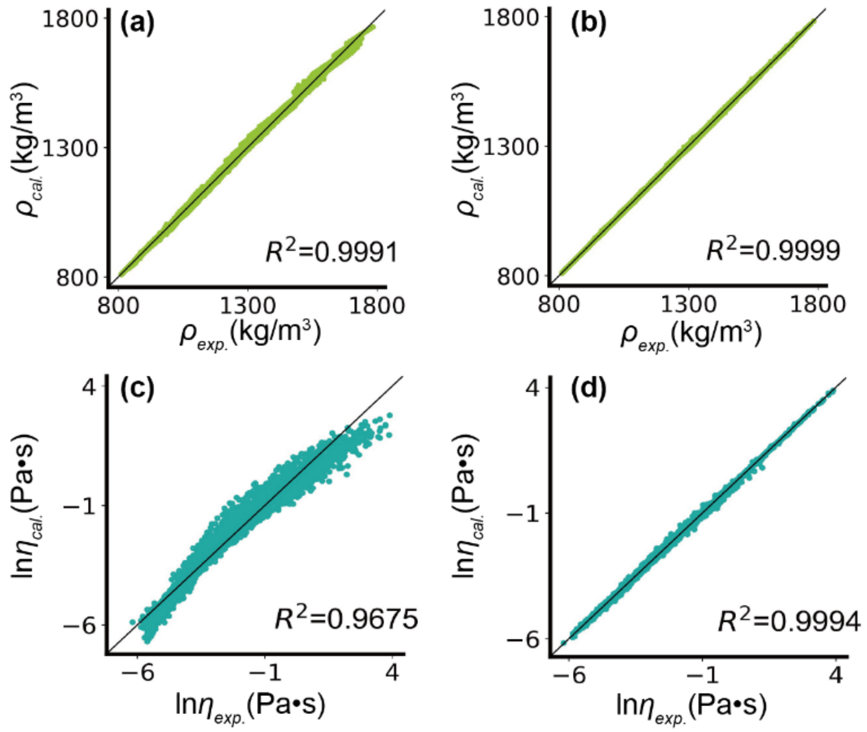

Figure 2. The correlation between experimental and calculated for ρ and η :β , γ , and χ are constants in (a) and (c); β , γ , and χ are variables in (b) and (d).

3.2. ρ (T ,P ,I)-QSPR model

The ρ (T ,P ,I)-QSPR model was proposed as Eq. (15). Detailed parameter values are shown in Table C1 of Supporting Information (atomic-distribution-matrix.docx).

n = 19335, R ^{2}= 0.9922,Q ^{2} $_{\text{LOCO}}$ = 0.9905,Q ^{2} $_{\text{LOAO}}$ = 0.9894;

n $_{\text{training}}$ = 15015; R $^2{}_{\text{training}}$= 0.9922; MAE$_{\text{training}}$ = 9.3290 kg/m^3;

n $_{\text{testing}}$ = 4320; R $^2{}_{\text{testing}}$ = 0.9921; MAE$_{\text{testing}}$ = 9.3606 kg/m^3;

Where, I $_{\text{IL}}$, I $_{\text{C}}$ and$I$ $_{\text{A}}$ represent norm index (I) of ILs, cation and anion, respectively.n $_{\text{C}}$ and$n$ $_{\text{A}}$ are the number of cations and anions (eg.n $_{\text{C}}$ and n $_{\text{A}}$ of 1-methyl-3-(3-(trimethylammonio)propyl)-1H-imidazolium bis(dicyanamide) are 1 and 2, respectively.).

The high R ^{2} and low MAE show that theρ (T ,P ,I)-QSPR model has a good ability to calculate the ρ of ILs. The experimental and calculated ρvalues from the model expressed in Eq. (15) were shown in Table S1 of Supporting Information (exp-cal-values.xlsx).

3.3.1. Model validation

Internal validation. The distribution of cations and anions for ρ are shown in Figures 3a-b. It is apparent that the type distribution of cations in the ρ dataset is more equal than that of anions. As can be seen in Figure 3(a), the cations with large data points are [C4mim] = 9.97%, [emim] = 7.57%, [C6mim] = 5.75%, [meim] = 4.16% and [mC4pyr] = 3.93%. What is noteworthy is that the ILs containing [emim] are in the testing set. Although there are more ILs containing [C4mim] in the ρdataset, the results validated by LOCO-CV are acceptable, with MAE of 7.2524 kg/m^3. Similarly in Figure 3(b), the anions with large data points are [N(SO2CF3)2] = 28.49%, [BF4] = 11.83%, [PF6] =4.62%, [N(CN)2] = 3.57% and [N(SO2F)2] = 3.38%. Although the proportion of [N(SO2CF3)2] in the anion is relatively high, there are as many as 265 cations in the ILs with [N(SO2CF3)2] as well. So, the MAE for ILs containing [N(SO2CF3)2], as verified by LOAO-CV, is quite optimistic at 6.7168 kg/m^3. The validation results of the $\rho(T\ ,P\ ,I\ )$ model are illustrated in Figure 4. The scatter diagram results of the LOCO-CV and LOAO-CV are shown in Figures 4a-b. Clearly, the internal cross-validation results for LOCO-CV and LOAO-CV are 0.9905 and 0.9894, respectively, which fully demonstrated the high stability of the model in predicting ρ of ILs containing novel cations and anions. Analogously, Figure 4d shows the absolute error distributions diagram for the $\rho(T\ ,P\ ,I\ )$-model, LOCO-CV and LOAO-CV. From Figure 4d the error range distribution of LOCO-CV has more points concentrated in the range of 0 ˜ 10 kg/m^3 than LOAO-CV, which further indicates that this model has greater stability in predicting ILs with new cations. The detailed statistical parameters of internal validation are listed in Table 2. It is worth reminding that the model validation results for LOO ($Q\ ^2{}_{\text{LOILO}}$ = 0.9907 and$Q\ ^2{}_{\text{LODPO}}$ = 0.9921) are generally higher than those for LOIO ($Q\ ^2{}_{\text{LOCO}}$ = 0.9905 and$Q\ ^2{}_{\text{LOAO}}$ = 0.9894), especially for LODPO-CV, as can be seen in Table 2. In addition, the "pseudo-high" accuracy of LOO-CV is more evident in the results of the MAE. The MAE of LOILO-CV is 10.2623 kg/m^3 which is lower than that of the LOAO-CV (MAE_{LOAO} = 11.3498 kg/m^3). These facts suggest that the LOILO does not accurately evaluate the stability of QSPR models for new anion and cation, producing a "pseudo-high" accuracy. This provides a more straightforward demonstration that the use of LOO to validate ILs property models leads to an "pseudo-high" accuracy of model stability. There is a strong need to use LOIO-CV to evaluate the ILs QSPR model to obtain a more realistic and stable model. Moreover, the absolute error (AE) distributions of the LOIO-CV is consistent with the training set of the $\rho\ (T\ ,P\ ,I\ )$-model, and most of the errors are within the range of 0 ˜ 10 kg/m^3. Therefore, it is further confirmed that the model is feasible to predictρ of ILs.

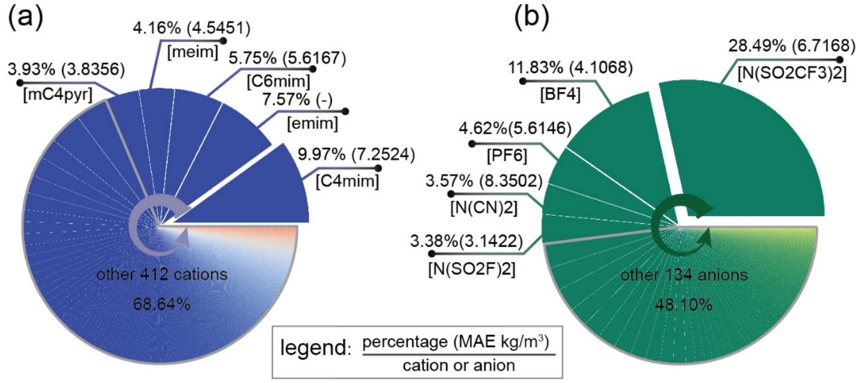

Figure 3. The distribution of cations and anions in theρ dataset. (a) the distribution of cations and (b) the distribution of anions.

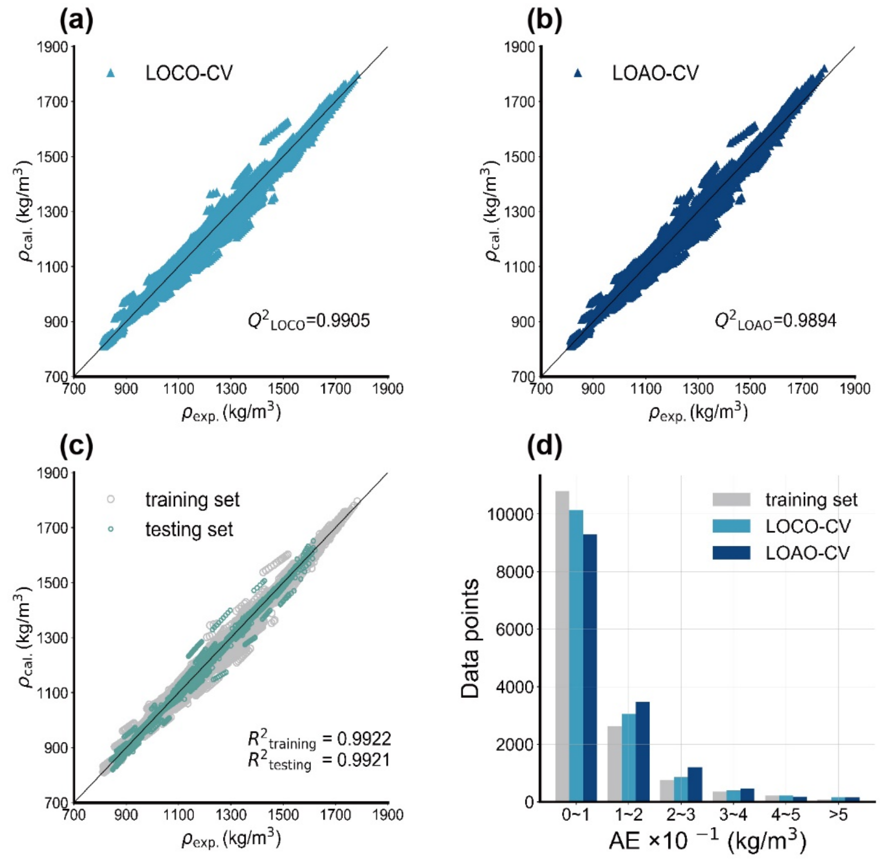

Figure 4. The validation results of the ρ (T ,P ,I)-model: (a-b) the scatter plots of experimental vs. calculated values for LOCO-CV and LOAO-CV, respectively, (c) the external validation, (d) the results for the absolute error distributions diagram of the training set, LOCO-CV and LOAO-CV, respectively.

Table 2 . The detailed results of internal and external validations for density.

Method	Status	Status	D.P.	$R^2(Q^2)$	MAE (kg/m^3)
Internal validation	Internal validation	LOCO-CV	15015	0.9905	10.3470
		LOAO-CV	15015	0.9894	11.3498
		LOILO-CV	15015	0.9907	10.2623
		LODPO-CV	15015	0.9921	9.3723
External validation	External validation	Training set	15015	0.9922	9.3290
		Testing set	4320	0.9921	9.3606
Overall data set	Overall data set	Overall data set	19335	0.9922	9.3361

External validation. For external validation, 757 ILs (15015 data points) from the ρ dataset were used as training set and the remaining 215 ILs (4320 data points) served as the testing set. The detailed results for external validation are listed in Table 2. The$R$ ^{2} of training and testing sets reached 0.9922 and 0.9921, respectively. The MAE values were 9.3290 kg/m^3 and 9.3606 kg/m^3, respectively. The experimental and calculatedρ values of the model for the training/testing set are shown in Figure 4c. It is easy to see that the overall trend of data points in training set and testing set remains roughly the same and both fit near the diagonal, which shows that this model has a good predictive ability for ρ of ILs.

Y-randomized analysis. To evaluate the reliability of the ρ model, Y -random validation was repeated 1000 times. The results of the Y -random validation with and values were less than 0.00248 and 0.00689, respectively, far less thanR^2 (0.9919) of the ρ model. Therefore, the ρ (T, P, I)-QSPR model was not affected by chance correlation.

3.3.2. Model comparison: before and after data pre-screening

To more prominently highlight the importance of data pre-screening in the pre-modelling work, this work implemented LOIO-CV by the data without screening using the same descriptors as the Eq. (15). The detailed LOIO-CV results are shown in Table 3. HigherQ^2($Q^2_{\mathrm{LOCO}} = 0.9919$ and$Q^2_{\mathrm{LOAO}} = 0.9899$) and lower MAE ($\mathrm{MAE}_{\mathrm{LOCO}} = 8.6487$ kg/m^3 and $\mathrm{MAE}_{\mathrm{LOAO}} = 10.2462$ kg/m^3) were obtained when the model was built using the dataset without data pre-screening. However, when Q^2 are recalculated by the dataset selected by following the data pre-screening rules, there was a decrease in Q^2($Q^2_{\mathrm{LOCO}} = 0.9903$ and$Q^2_{\mathrm{LOAO}} = 0.9884$) and an increase in MAE ($\mathrm{MAE}_{\mathrm{LOCO}} = 10.2589$ kg/m^3 and $\mathrm{MAE}_{\mathrm{LOAO}} = 11.8680$ kg/m^3). In addition, the results in Table 3 show that the Q^2($Q^2_{\mathrm{LOCO}} = 0.9905$ and$Q^2_{\mathrm{LOAO}} = 0.9894$) of pre-screened data is higher than the Q^2($Q^2_{\mathrm{LOCO}} = 0.9903$ and$Q^2_{\mathrm{LOAO}} = 0.9884$) post-screened data. It is proved that the model although having a higherQ^2without data pre-screening, is "pseudo-high" in accuracy. Therefore, it is necessary to carry out a pre-screening process before modelling to ensure a balanced distribution of the dataset.

Table 3 . Comparison of model stability before and after data pre-screening for density.

method	data screening status	D.P.	Q^2	MAE (kg/m^3)
LOCO-CV	no	36784	0.9919	8.6487
	post	15015	0.9903	10.2589
	pre	15015	0.9905	10.3470
LOAO-CV	no	36784	0.9899	10.2462
	post	15015	0.9884	11.8680
	pre	15015	0.9894	11.3498

here, for data screening method: "no" refers to the model developed by the initial dataset obtained from NIST; "post" means that the dataset used to build the model is consistent with "no", but only the data points selected under the pre-screening rules are used for calculatingQ^2. "pre" refers the model with the dataset after data pre-screening.

3.3.3. Comparison with references

The comparison of ρ (T, P, I)-QSPR model with that in references was mainly carried out through,R^2 and Q^2, and the details are shown in Table 4. Our model has advantages in data set screening, rather than collecting too many data points. Moreover, theρ (T, P, I)-model has higher accuracy withR^2 of 0.9922. The $R^2$is slightly lower than that of our previous studies (Yan et al.[37] and Zhang et. al[29]), which can be attributed to that the data pre-screening process in this work avoided the excessive proportion of data points from the same ILs, so as to effectively avoid the situation of high R^2. In addition, the model was validated by LOCO-CV and LOAO-CV, which proved the stability and robustness of theρ (T, P, I)-model. To sum up, this model is reliable in calculating the ρ values for ILs.

Table 4 . Comparisons of this work with references for the density.

References	Method	D.P.	ILs	T/K	P/MPa	R^2	Q^2_{LOCO}	Q^2_{LOAO}
Paduszynski and Domanska[38]	GC	16830	1028	251 ˜ 473	0.1 ˜ 300	-		
Lazzús[39]	QSPR	3020	163	258 ˜ 473	0.1 ˜ 207	0.9307	-	-
Yan et al. [37]	QSPR	5948	188	253.15 ˜ 473.15	0.1 ˜ 250	0.9980	-	-
Zhang et al.[29]	QSPR	9019	314	253.15 ˜ 473.15	0.1 ˜ 250	0.9970	-	-

References	Method	D.P.	ILs	T/K	P/MPa	R^2	$Q^2{}_{\mathrm{LOCO}}$	$Q^2{}_{\mathrm{LOAO}}$
This work	QSPR	19335	972	221.31˜ 473.15	0.0815 ˜ 251.5	0.9922	0.9905	0.9894

3.3. lnη (T ,P ,I)-QSPR model

The lnη (T ,P ,I)-QSPR model was built to predict η as Eq. (16). The Detailed parameter values are shown in Table C2 of Supporting Information (atomic-distribution-matrix.docx).

n = 9238, R ^{2}= 0.9100,Q ^{2} $_{\mathrm{LOCO}}$ = 0.8863,Q ^{2} $_{\mathrm{LOAO}}$ = 0.8866;

n $_{\mathrm{training}}$ = 7352; R $^2{}_{\mathrm{training}}$ = 0.9091; MAE$_{\mathrm{training}}$ =0.3276 Pas;

n $_{\mathrm{testing}}$ = 1886; R $^2{}_{\mathrm{testing}}$ = 0.9108; MAE$_{\mathrm{testing}}$ =0.3232 Pas;

where, I $_{\mathrm{IL}}$, I $_{\mathrm{C}}$ and$I$ $_{\mathrm{A}}$ represent the norm index of ILs, cation and anion, respectively. n $_{\mathrm{C}}$ and$n$ $_{\mathrm{A}}$ are the number of cations and anions.

The results of statistical parameters showed that this η model has high R ^{2} and low MAE, which demonstrated the advanced accuracy of the model in predicting η . The experimental and calculated η values from the model expressed in Eq. (16) are shown in Table S2 of Supporting Information (exp-cal-values.xlsx).

3.4.1. Model validation

Internal validation. The distribution of cations and anions as shown in Figures 5a-b. The distribution of cations in theη dataset is more varied and more even than anions. The common anion, [N(SO2CF3)2], accounts for 34.49% of the η dataset, and it is worth noting that 255 different cations are combined with [N(SO2CF3)2]. So, the MAE (0.2034 Pas) of LOAO-CV is relatively lower for [N(SO2CF3)2]. In addition, it can be seen that all five cations with a relatively large data points have a low MAE. Figures 6a-b show the experimental versus calculated values scatter plots of LOCO-CV and LOAO-CV. Obviously, the satisfactory results of the$Q$ $^2{}_{\mathrm{LOCO}}$ (0.8863) and$Q$ $^2{}_{\mathrm{LOAO}}$ (0.8866) indicate that the η model has high stability and good prediction performance for ILs containing new cations and anions. The detailed results for internal validation of the η model are given in Table 5. It is clear that the Q ^{2} for the LOO method (Q $^2{}_{\mathrm{LOILO}}$ = 0.8908 and$Q$ $^2{}_{\mathrm{LODPO}}$ = 0.9076) are much higher than the Q ^{2} for LOIO (Q $^2{}_{\mathrm{LOCO}}$ = 0.8863 and$Q$ $^2{}_{\mathrm{LOAO}}$ = 0.8866), and the MAE for the LOO method (MAE$_{\mathrm{LOILO}}$ = 0.3558 Pas and MAE$_{\mathrm{LODPO}}$ = 0.3300 Pas) are lower than the MAE for LOIO (MAE$_{\mathrm{LOCO}}$ = 0.3654 Pas and MAE$_{\mathrm{LOAO}}$ = 0.3677 Pas). This in turn suggests that the stability of the model validated by LOO-CV is limited to predicting ILs with known ionic types, while the model passed by the LOIO is more stable when faced with ILs containing unknown ionic types. In addition, the AE distributions of the lnη (T ,P ,I)-QSPR model, LOCO-CV and LOAO-CV are shown in Figure 6d. Here, the AE ranges for the both internal validations are roughly the same as the lnη (T ,P ,I)-QSPR model, and the errors of most data points are concentrated in 0 ˜ 0.2 Pas, further demonstrating the outstanding stability of the lnη (T ,P ,I)-QSPR model.

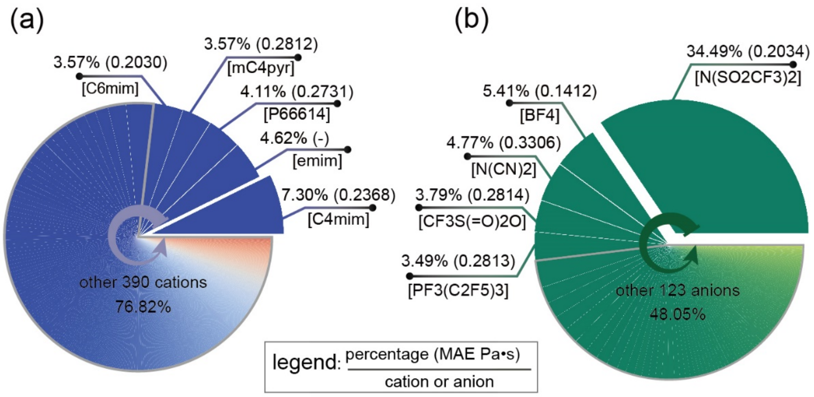

Figure 5. The distribution of cations and anions in theη dataset: (a) the distribution of cations, (b) the distribution of anions.

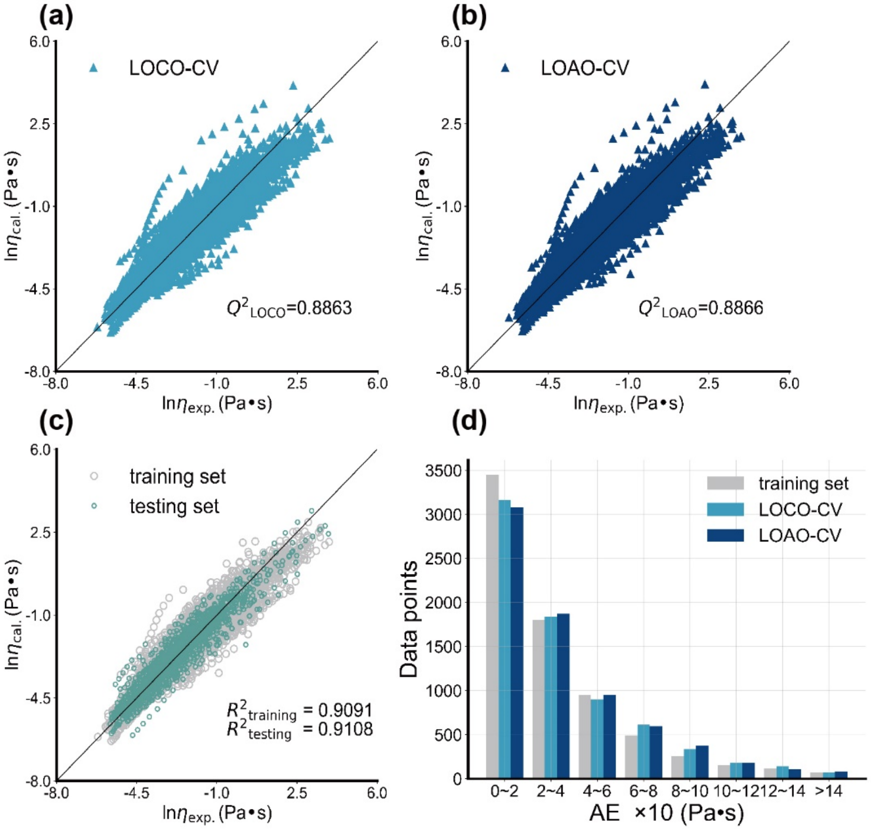

Figure 6. The validation results of the lnη(T ,P ,I)-QSPR model: (a-b) the scatter plots of experimental vs. calculated values for LOCO-CV and LOAO-CV, (c) the external validation, (d) the results for the absolute error distributions diagram of the training set, LOCO-CV and LOAO-CV, respectively.

Table 5 . The results of internal and external validations for viscosity.

Method	Status	Status	D.P.	$R^2(Q^2)$	MAE (Pas)
Internal validation	Internal validation	LOCO-CV	7352	0.8863	0.3654
		LOAO-CV	7352	0.8866	0.3677
		LOILO-CV	7352	0.8908	0.3558
		LODPO-CV	7352	0.9076	0.3300
External validation	External validation	Training set	7352	0.9091	0.3276
		Testing set	1886	0.9108	0.3232
Overall data set	Overall data set	Overall data set	9238	0.9100	0.3267

External validation. In the external validation of theη model, 651 ILs containing 7352 data points are used for training set and 181 ILs containing 1886 data points are treated as testing set. The detailed statistical parameters are listed in Table 4. The $R\ ^2{}_{\text{testing}} = 0.9108$ is close to the $R\ ^2{}_{\text{training}} = 0.9091$. The experimental vs. calculated η values of the model for the external validation is presented in Figure 6c. Seen from that the data points in testing set are consistent with the trend of the training set, indicating that the lnη (T ,P ,I)-QSPR model has quite excellent predictive ability for η of ILs at variable temperature and pressure.

Y-randomized analysis. After 1000 repetitions of$Y$ -random validation, and were lower than 0.00586 and 0.00044, far less than the accuracy of the lnη(T ,P ,I)-QSPR model, indicating that the model was not affected by chance correlation.

3.4.2. Model comparison: before and after data pre-screening

Similar to the analysis of the ρ (T ,P ,I)-QSPR model, this work also carried out a data pre-screening process for theη dataset. A QSPR model was built for the initial ηdataset using the same descriptors as in Eq. (16). The detailed LOIO-CV results for that model are shown in Table 6. The model built without data pre-screening, while having high $Q\ ^2$($Q\ ^2{}_{\text{LOCO}} = 0.8935$ and$Q\ ^2{}_{\text{LOAO}} = 0.8913$) and low MAE ($\text{MAE}_{\text{LOCO}} = 0.3153$ and $\text{MAE}_{\text{LOAO}} = 0.3211$), gets a significant downward trend when the stability of its model is assessed again by post data pre-screening ($Q\ ^2{}_{\text{LOCO}} = 0.8815$ and$Q\ ^2{}_{\text{LOAO}} = 0.8806$; $\text{MAE}_{\text{LOCO}} = 0.3691$ and $\text{MAE}_{\text{LOAO}} = 0.3755$). The model is not as stable as the one obtained after the data pre-screening exercise prior to modelling with$Q\ ^2{}_{\text{LOCO}} = 0.8863$ and$Q\ ^2{}_{\text{LOAO}} = 0.8866$. It is therefore well established that the model built without data pre-screening are less stable. Thus, it is necessary to carry out data pre-processing before building the QSPR model to ensure a balanced and stable distribution of the dataset and to obtain a stable model.

Table 6 . Comparison of model stability before and after data pre-screening for viscosity.

method	data screening status	D.P.	Q^2	MAE (Pas)
LOCO-CV	no	13747	0.8935	0.3153
	post	7352	0.8815	0.3691
	pre	7352	0.8863	0.3654
LOAO-CV	no	13747	0.8913	0.3211
	post	7352	0.8806	0.3755
	pre	7352	0.8866	0.3677

here, for data screening method: “no” refers to the model developed by the initial dataset obtained from NIST; “post” means that the dataset used to build the model is consistent with “no”, but only the data points selected under the pre-screening rules are used for calculating$Q\ ^2$. “pre” refers the model with the dataset after data pre-screening

3.4.3. Comparison with references

The results of comparison between this work and other literatures are detailed in Table 7. The significant difference from other works is that our work has a wider range of datasets and varying pressure conditions. Different from our previous studies and other groups, this work expanded the types of ILs and the number of relevant data points on the basis of increasing the pre-screening process for data points. Thus, the lnη (T ,P ,I)-QSPR model is slightly less accurate than our previous studies, however, it is more robust than the previous models, because this model has gone through a rigorous LOIO-CV model validation method. In addition, the modelling process kind of uses only the topology of the ILs, which has the advantage of simple, fast and efficient preliminary calculations of η .

Table 7 . Comparisons of this work with QSPR method in references for the viscosity.

References	D.P.	ILs	T/K	P/MPa	R^2	Q^2_{LOCO}	Q^2_{LOAO}
Matsuda et al.[40]	300	-	273 ~ 353	-	0.8971	-	-
Barycki et al.[41]	23	138	298 ~ 343	0.1	0.8260	-	-
Yan et al.[42]	3228	349	253.15 ~ 573	0.06 ~ 300	0.9640	-	-
Zhang et al.[29]	7342	351	253.15 ~ 438.15	0.06 ~ 300	0.9642	-	-
This work	9238	832	253.15 ~ 438.15	0.06 ~ 300	0.9100	0.8863	0.8866

4. Conclusions

Two f (T ,P ,I)-QSPR models based on 19335ρ data points and 9238 η data points were established to calculate the ρ and η of ILs under variable temperature and pressure. In order to accurately verify the stability as well as the robustness of the f (T ,P ,I)-QSPR model, a new internal validation method, LOIO-CV, was proposed, which has more strict evaluation criteria and is more reliable compared with the traditional LOO-CV. Moreover, the stability of the$f$ (T ,P ,I)-QSPR model is improved by using data pre-screening to equalize the distribution of data points of ILs and introducing descriptors to the temperature and pressure terms. Furthermore, using only atomic linkage relationships of ILs, no additional time is required to obtain the optimal structure of the ILs. The results of statistical parameter analysis show that these models have good prediction accuracy and reliability with high$R$ ^{2} and low MAE. In the evaluation of ILsf (T ,P ,I)-QSPR model, both models passed the rigorous LOIO-CV, which indicates that these$f$ (T ,P ,I)-QSPR models have good stability, robustness, and relatively accurate prediction performance. Meanwhile, we also found that the evaluation results of LOIO-CV were affected by the distribution of ion species, that is, the model established by the dataset with a more balanced distribution of data points of different ILs had higher stability and robustness. Therefore, the equilibrium distribution of different ILs data points is particularly important when modeling the properties of ILs by using QSPR method. In one sense, two proposed f (T ,P ,I)-QSPR models are widely applicable to predict the ρ and η properties of ILs, and these models provide an intelligent tool for predicting the design or synthesis of ILs containing novel cations and anions. It is worth mentioning that the strategy can be widely applied to the estimation of other properties of ILs, such as environmental toxicity and other related physicochemical properties.

Supporting Information

The experimental and calculated values of density and viscosity were listed in exp-cal-values.xlsx. The definitions of the statistical parameters were listed in Table S1 of Statistical parameters.docx. The atomic distribution matrix (MA), used for density and viscosity were shown in atomic-distribution-matrix.docx. The properties of each atom were shown in S1 of atom properties.xlsx. Also, an example for calculating the density with the established model was given in example.xlsx.

Conflict of interest

The authors confirm that this article has no conflicts of interest.

Acknowledgments

This work was financially supported by the National Natural Science Foundation of China (21808167), Tianjin Municipal Science and Technology Bureau (20JCQNJC00090).

Data Availability Statement

The ILs data were collected from the National Institute of Standards and Technology (NIST) (*https://ilthermo.boulder.nist.gov/index.html*), and were provided in the Supporting Information (exp-cal-values.xlsx.). The implementation of the LOIO-CV method can be implemented by relevant mathematical software (matlab and mathematica) as well as computer language (e.g. python, R and C++).

References

1. Zhao, Y.; Gani, R.; Afzal, R. M.; Zhang, X.; Zhang, S., Ionic liquids for absorption and separation of gases: An extensive database and a systematic screening method. *AIChE Journal* **2017,** *63* (4), 1353-1367.

2. Zhang, X.; Ding, X.; Song, Z.; Zhou, T.; Sundmacher, K., Integrated ionic liquid and rate-based absorption process design for gas separation: Global optimization using hybrid models. *AIChE Journal* **2021,** *67* (10).

3. Beil, S.; Markiewicz, M.; Pereira, C. S.; Stepnowski, P.; Thöming, J.; Stolte, S., Toward the Proactive Design of Sustainable Chemicals: Ionic Liquids as a Prime Example. *Chemical Reviews* **2021,** *121* (21), 13132-13173.

4. Roberts, N. J.; Lye, G. J., Application of Room-Temperature Ionic Liquids in Biocatalysis: Opportunities and Challenges. **2002,** *818* , 347-359.

5. Vekariya, R. L., A review of ionic liquids: Applications towards catalytic organic transformations. *Journal of Molecular Liquids* **2017,** *227* , 44-60.

6. Wasilewski, T.; Gebicki, J.; Kamysz, W., Prospects of ionic liquids application in electronic and bioelectronic nose instruments. *TrAC Trends in Analytical Chemistry* **2017,** *93* , 23-36.

7. Watanabe, M.; Thomas, M. L.; Zhang, S.; Ueno, K.; Yasuda, T.; Dokko, K., Application of Ionic Liquids to Energy Storage and Conversion Materials and Devices. *Chemical Reviews* **2017,** *117*(10), 7190-7239.

8. Song, Z.; Zhang, C.; Qi, Z.; Zhou, T.; Sundmacher, K., Computer-aided design of ionic liquids as solvents for extractive desulfurization. *AIChE Journal* **2018,** *64* (3), 1013-1025.

9. Huang, Y.; Dong, H.; Zhang, X.; Li, C.; Zhang, S., A new fragment contribution-corresponding states method for physicochemical properties prediction of ionic liquids. *AIChE Journal* **2013,** *59* (4), 1348-1359.

10. Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, II; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; Consonni, V.; Kuz'min, V. E.; Cramer, R.; Benigni, R.; Yang, C.; Rathman, J.; Terfloth, L.; Gasteiger, J.; Richard, A.; Tropsha, A., QSAR modeling: where have you been? Where are you going to? *J Med Chem* **2014,** *57* (12), 4977-5010.

11. Katritzky, A. R., Kuanar, M., Slavov, S., Hall, C. D., Karelson, M., Kahn, I., and Dobchev, D. A., Quantitative Correlation of Physical and Chemical Properties with Chemical Structure: Utility for Prediction. *Chemical Reviews* **2010,** *110(10)* , 5714-5789.

12. Greaves, T. L.; Drummond, C. J., Protic Ionic Liquids: Evolving Structure–Property Relationships and Expanding Applications. *Chemical Reviews* **2015,** *115* (20), 11379-11448.

13. Izgorodina, E. I.; Seeger, Z. L.; Scarborough, D. L. A.; Tan, S. Y. S., Quantum Chemical Methods for the Prediction of Energetic, Physical, and Spectroscopic Properties of Ionic Liquids. *Chemical Reviews* **2017,** *117* (10), 6696-6754.

14. Wu, K.-J.; Luo, H.; Yang, L., Structure-based model for prediction of electrical conductivity of pure ionic liquids. *AIChE Journal* **2016,** *62* (10), 3751-3762.

15. Paduszyński, K., Extensive Databases and Group Contribution QSPRs of Ionic Liquids Properties. 1. Density. *Industrial & Engineering Chemistry Research* **2019,** *58* (13), 5322-5338.

16. Das, S.; Ojha, P. K.; Roy, K., Development of a temperature dependent 2D-QSPR model for viscosity of diverse functional ionic liquids. *Journal of Molecular Liquids* **2017,** *240* , 454-467.

17. Mirkhani, S. A.; Gharagheizi, F., Predictive Quantitative Structure–Property Relationship Model for the Estimation of Ionic Liquid Viscosity. *Industrial & Engineering Chemistry Research***2012,** *51* (5), 2470-2477.

18. Brauner, N.; St. Cholakov, G.; Kahrs, O.; Stateva, R. P.; Shacham, M., Linear QSPRs for predicting pure compound properties in homologous series. *AIChE Journal* **2008,** *54* (4), 978-990.

19. Su, Y.; Wang, Z.; Jin, S.; Shen, W.; Ren, J.; Eden, M. R., An architecture of deep learning in QSPR modeling for the prediction of critical properties using molecular signatures. *AIChE Journal***2019,** *65* (9).

20. Makarov, D. M.; Fadeeva, Y. A.; Shmukler, L. E.; Tetko, I. V., Beware of proper validation of models for ionic Liquids! *Journal of Molecular Liquids* **2021,** *344* , 117722.

21. Venkatraman, V.; Evjen, S.; Knuutila, H. K.; Fiksdahl, A.; Alsberg, B. K., Predicting ionic liquid melting points using machine learning.*Journal of Molecular Liquids* **2018,** *264* , 318-326.

22. Low, K.; Kobayashi, R.; Izgorodina, E. I., The effect of descriptor choice in machine learning models for ionic liquid melting point prediction. *J Chem Phys* **2020,** *153* (10), 104101.

23. Yan, F.; Shi, Y.; Wang, Y.; Jia, Q.; Wang, Q.; Xia, S., QSPR models for the properties of ionic liquids at variable temperatures based on norm descriptors. *Chemical Engineering Science* **2020,***217* , 115540.

24. Kang, X.; Zhao, Z.; Qian, J.; Muhammad Afzal, R., Predicting the Viscosity of Ionic Liquids by the ELM Intelligence Algorithm.*Industrial & Engineering Chemistry Research* **2017,***56* (39), 11344-11351.

25. Yu, G.; Zhao, D.; Wen, L.; Yang, S.; Chen, X., Viscosity of ionic liquids: Database, observation, and quantitative structure-property relationship analysis. *AIChE Journal* **2012,** *58*(9), 2885-2899.

26. Sattari, M.; Gharagheizi, F.; Ilani-Kashkouli, P.; Mohammadi, A. H.; Ramjugernath, D., Estimation of the Heat Capacity of Ionic Liquids: A Quantitative Structure–Property Relationship Approach. *Industrial & Engineering Chemistry Research* **2013,** *52* (36), 13217-13221.

27. Mirkhani, S. A.; Gharagheizi, F.; Farahani, N.; Tumba, K., Prediction of surface tension of ionic liquids by molecular approach.*Journal of Molecular Liquids* **2013,** *179* , 78-87.

28. Farahani, N.; Gharagheizi, F.; Mirkhani, S. A.; Tumba, K., A simple correlation for prediction of heat capacities of ionic liquids.*Fluid Phase Equilibria* **2013,** *337* , 73-82.

29. Zhang, S.; Jia, Q.; Yan, F.; Xia, S.; Wang, Q., Evaluating the properties of ionic liquid at variable temperatures and pressures by quantitative structure–property relationship (QSPR). *Chemical Engineering Science* **2021,** *231* , 116326.

30. Shi, Y.; Li, J.-J.; Wang, Q.; Jia, Q.; Yan, F.; Luo, Z.-H.; Zhou, Y.-N., Computer-aided estimation of kinetic rate constant for degradation of volatile organic compounds by hydroxyl radical: An improved model using quantum chemical and norm descriptors.*Chemical Engineering Science* **2022,** *248* , 117244.

31. Ionic Liquids Database-ILThermo (v2.0).*https://ilthermo.boulder.nist.gov/index.html*(July 2, 2021).

32. Björck, Å., Least squares methods. In *Handbook of Numerical Analysis* , Elsevier: 1990; Vol. 1, pp 465-652.

33. Barlow, J. L., 9 Numerical aspects of solving linear least squares problems. In *Handbook of Statistics* , Elsevier: 1993; Vol. 9, pp 303-376.

34. Xiong, Y.; Ding, J.; Yu, D.; Peng, C.; Liu, H.; Hu, Y., Volumetric Connectivity Index: A New Approach for Estimation of Density of Ionic Liquids. *Industrial & Engineering Chemistry Research***2011,** *50* (24), 14155-14161.

35. Alexander, D. L. J.; Tropsha, A.; Winkler, D. A., Beware of R2: Simple, Unambiguous Assessment of the Prediction Accuracy of QSAR and QSPR Models. *Journal of Chemical Information and Modeling***2015,** *55* (7), 1316-1322.

36. Rücker, C.; Rücker, G.; Meringer, M., y-Randomization and its variants in QSPR/QSAR. *Journal of Chemical Information and Modeling* **2007,** *47* (6), 2345-2357.

37. Yan, F.; Shang, Q.; Xia, S.; Wang, Q.; Ma, P., Application of Topological Index in Predicting Ionic Liquids Densities by the Quantitative Structure Property Relationship Method. *Journal of Chemical & Engineering Data* **2015,** *60* (3), 734-739.

38. Paduszyński, K.; Domańska, U., A New Group Contribution Method For Prediction of Density of Pure Ionic Liquids over a Wide Range of Temperature and Pressure. *Industrial & Engineering Chemistry Research* **2011,** *51* (1), 591-604.

39. Lazzús, J. A., ρ(T, p) model for ionic liquids based on quantitative structure-property relationship calculations. *Journal of Physical Organic Chemistry* **2009,** *22* (12), 1193-1197.

40. Matsuda, H.; Yamamoto, H.; Kurihara, K.; Tochigi, K., Computer-aided reverse design for ionic liquids by QSPR using descriptors of group contribution type for ionic conductivities and viscosities. *Fluid Phase Equilibria* **2007,** *261* (1-2), 434-443.

41. Barycki, M.; Sosnowska, A.; Gajewicz, A.; Bobrowski, M.; Wileńska, D.; Skurski, P.; Giełdoń, A.; Czaplewski, C.; Uhl, S.; Laux, E.; Journot, T.; Jeandupeux, L.; Keppner, H.; Puzyn, T., Temperature-dependent structure-property modeling of viscosity for ionic liquids. *Fluid Phase Equilibria* **2016,** *427* , 9-17.

42. Yan, F.; He, W.; Jia, Q.; Wang, Q.; Xia, S.; Ma, P., Prediction of ionic liquids viscosity at variable temperatures and pressures.*Chemical Engineering Science* **2018,** *184* , 134-140.