Laakso, Juho-Pekka; Ebrahimpoor Gorji, Ali; Alopaeus, Ville; Uusi-Kyyny, Petri

Machine learning modeling of the CO_2 solubility in ionic liquids by using σ-profile descriptors

Published in:
Chemical Engineering Science

DOI:
10.1016/j.ces.2025.121226

Published: 15/03/2025

Document Version
Publisher's PDF, also known as Version of record

Chemical Engineering Science 307 (2025) 121226

Contents lists available at ScienceDirect

Chemical Engineering Science

journal homepage: www.elsevier.com/locate/ces

Machine learning modeling of the CO_2 solubility in ionic liquids by using σ-profile descriptors

Juho-Pekka Laakso *, Ali Ebrahimpoor Gorji, Petri Uusi-Kyyny, Ville Alopaeus

Aalto University, School of Chemical Technology, Department of Chemical and Metallurgical Engineering, P.O. Box 16100, FI-00076 Aalto, Finland

ARTICLE INFO

Keywords:
Machine learning
Quantitative structure–property relationship
Carbon dioxide solubility prediction
Ionic liquids
σ-profile

ABSTRACT

The solubility of carbon dioxide (CO_2) in solvents is important for carbon capture and utilization technologies, with ionic liquids (ILs) being promising due to their ability to capture CO_2. Since the number of possible ILs is huge, predicting CO_2 solubility during solvent screening is essential. In this work, various machine learning (ML) models including multiple linear regression, artificial neural network, and random forest, were developed by using 9864 data points covering 124 ILs and descriptors from the σ-profile for predicting CO_2 solubility in ILs. The random forest model produced the best performance ($R^2 = 0.9754$ and MAE = 0.0257). We estimated the importance of the descriptors, highlighting that those with non-polar characteristics of the σ-profile are important. Lastly, we predicted CO_2 solubilities for 1444 unstudied ILs. The combination of ML with the σ-profile descriptors offers great generalizability for predicting CO_2 solubility in ILs. This enables IL screening for CO_2-related applications.

1. Introduction

Carbon capture and utilization have gathered extensive attention recently (Aghaie et al., 2018). Well-known large-scale carbon capture technologies such as solvent-base chemisorption, carbonate looping, and oxyfuel processes already exist (MacDowell et al., 2010). However, there is room for improvement regarding these technologies. An interesting strategy for reducing CO_2 emissions is to integrate carbon capture and utilization, which offers many advantages compared to carbon capture alone, such as the decreased need for CO_2 transportation, storage, and purification (Sun et al., 2021).

Ionic liquids (ILs) have recently drawn attention for their potential applications in carbon capture (Aghaie et al., 2018) and utilization (Chen and Mu, 2019). ILs consist of a combination of anions and cations while having a melting point below 100°C. ILs are known for their distinct properties, such as negligible vapor pressure, high thermal stability, and capacity to dissolve gases (Lei et al., 2014; Zeng et al., 2017). A unique feature of ILs is their tunability, where properties such as viscosity, melting point temperature, and CO_2 solubility can be tailored by changing the combination of cations and anions. However, there are at least a million ion combinations (Rogers and Seddon, 2003), which emphasizes the importance of the ability to predict optimal IL for various applications.

The CO_2 solubility in ILs can be predicted using thermodynamic models. The idea of the group contribution (GC) method has been combined with thermodynamical theory in the UNIFAC model (Fredenslund et al., 1975), which is extended to include ILs (Chen et al., 2020; Zhou et al., 2021). Phase equilibrium predictions can be made based on quantum chemical calculations without any additional parameter fitting by using COSMO-RS (Klamt, 1995), which has been utilized for the prediction of CO_2 solubility in ILs (Diedenhofen and Klamt, 2010; Islam et al., 2022). The equation of state (EOS) is one approach for modeling CO_2 solubility in ILs, whereas advanced EOS such as PC-SAFT (Chen et al., 2012), SAFT-VR Mie (Beraldo et al., 2024), and CPA (Wang et al., 2023) have been applied as modeling options for CO_2 solubility. The benefit of thermodynamic theories is that they provide a theoretical background for predictions. Nevertheless, making predictions can be challenging when several parameters are unavailable.

The machine learning (ML) method used together with the quantitative structure–property relationship (QSPR) approach has gained increasing attention in modeling the properties of ILs (Koutsoukos et al., 2021). The ML modeling approach is suitable for screening purposes (Sun et al., 2023). Screening could be performed using standard EOS models such as PC-SAFT. However, the ML modeling approach can be better for screening a wider range of candidates than PC-SAFT. Obtaining parameters for a large number of candidates can be

* Corresponding author.
E-mail address: juho-pekka.laakso@aalto.fi (J.-P. Laakso).

https://doi.org/10.1016/j.ces.2025.121226
Received 28 October 2024; Received in revised form 13 December 2024; Accepted 14 January 2025
Available online 16 January 2025

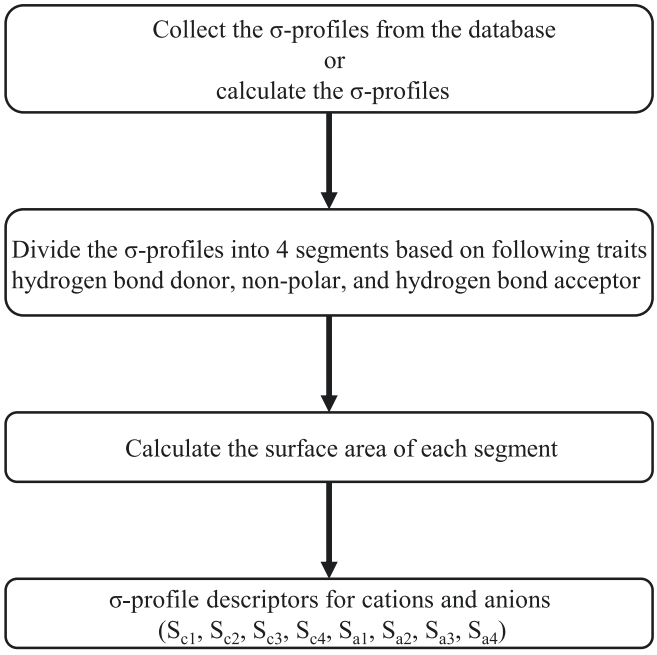

Fig. 1. The method for calculating descriptors from the σ-profile.

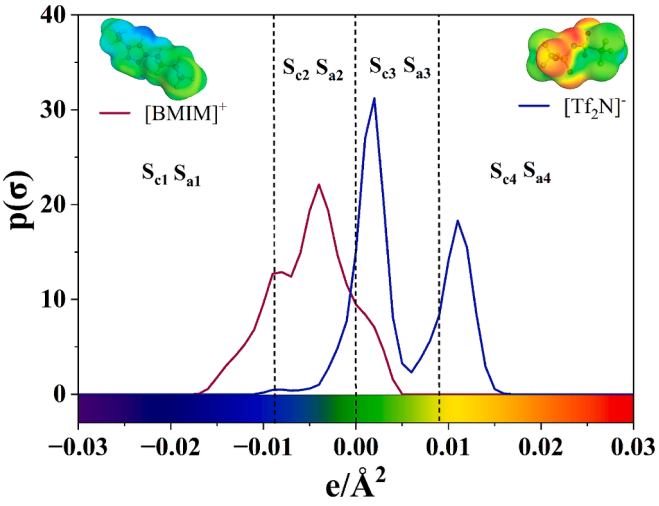

Fig. 2. The σ-profile segmentation for [BMIM] cation and [Tf_2N] anion.

difficult, especially when the necessary experimental data is missing. On the other hand, PC-SAFT is based on the theory of thermodynamics, which might give confidence when extrapolating predictions. However, obtaining optimal parameters for a large number of ILs can be challenging. The benefit of the ML method is its ability to recognize patterns between inputs and large amounts of data. Thus, the ML method can give accurate estimations within the range of the used data in model training. This methodology can extrapolate predictions, but the extent to which it can do so depends on multiple factors. Typically, this is evaluated by estimating the applicability domain (AD) (Netzeva et al., 2005). The AD can be estimated by calculating the leverage and principal component analysis or by checking the range of the used data. The importance of descriptors can also indicate the AD, while the importance of descriptors could be estimated by calculating the SHAP (SHapley Additive exPlanations) values, which have been used for understanding the behavior of ML models developed for predicting the properties of ILs (Lemaoui et al., 2024; Liu et al., 2023; Yang et al., 2024).

In addition to ML modeling, the deep learning modeling approach has been used successfully for predicting CO_2 solubility in ILs. The GC descriptors have been used in deep learning modeling together with deep neural networks and long short-term memory network methods (Ali et al., 2024). The solubility has been modeled by incorporating different optimization methods together with the standard deep learning models (Davoodi et al., 2024). In addition, graphical information of molecular structures has been used for predicting CO_2 solubility in ILs using graph neural networks (Jian et al., 2022).

The GC approach, combined with the ML method, is widely used. In the GC method, the molecular structure is divided into various functional groups, which are used as descriptors for model regression. Song et al. (2020) developed artificial neural network (ANN) and support vector machine (SVM) models for predicting CO_2 solubility in 124 ILs by using 10,116 data points. GC descriptors have also been used in deep ANN models (Ali et al., 2024). Recently, Yang et al. (2024) have developed models for the same dataset using hybrid descriptors, combining new descriptors with the GC descriptors. Both GC-based models have a large number of descriptors (53 and 133). However, the GC method has limitations, as it relies on predefined functional groups. If a specific compound has groups that are not present in the training procedure, then predictions cannot be made. The GC method might have problems when the number of certain functional groups is high or complex geometrical orientations of atoms are important, for example, in the case of isomers (Gani, 2019). There is no strict limit for the number of descriptors. Generally, it is recommended that the number of descriptors does not exceed half the number of data points (Le et al., 2012) to avoid overfitting and maintain generalization. In addition, it is important that descriptors not only boost the performance but also give meaningful insight into the underlying phenomena.

Together with other ML methods, QSPR modeling approaches have been used to predict CO_2 solubility in ILs. The capabilities of the QSPR

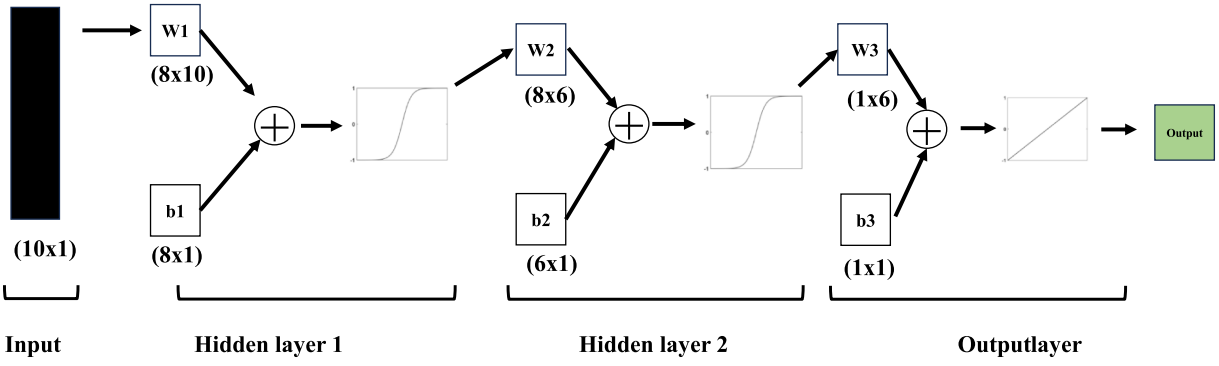

Fig. 3. Schematic representation of the developed ANN architecture.

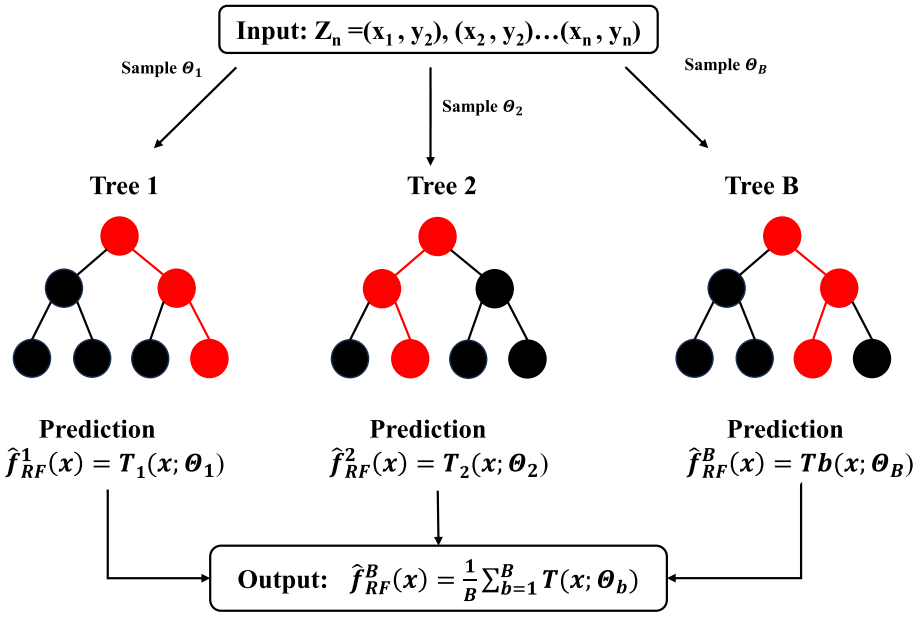

Fig. 4. A schematic representation for RF modelling.

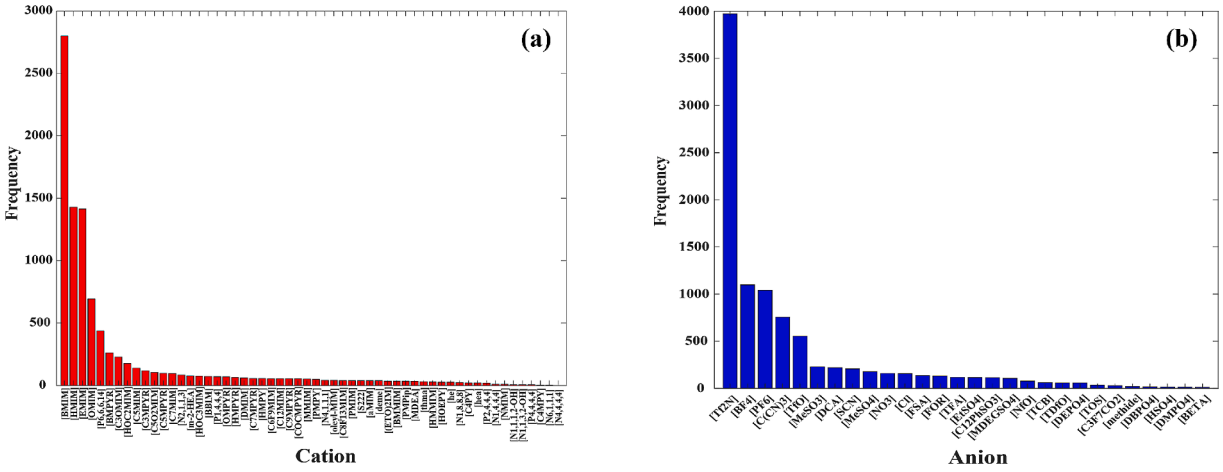

Fig. 5. Frequency of cations (a) and anions (b) in the experimental dataset.

Table 1
MLR model descriptors, parameters, and parameter values.

Descriptor	Parameter	Value
LnT	w_1	−2.0091785
LnP	w_2	1.039453
$\sqrt{P}$	w_3	−0.0148984
S_{c1}	w_4	34.2180538
S_{c2}	w_5	7.9600065
S_{c3}	w_6	−8.6487222
S_{c4}	w_7	−47.0001673
S_{a1}	w_8	6.4377259
S_{a2}	w_9	−1.6507327
S_{a3}	w_{10}	5.0114099
S_{a4}	w_{11}	2.5931122

approach have been studied by using multiple linear regression (MLR) and SVM to model solubility and Henry's law constant of CO_2 in ILs (Ghaslani et al., 2017; Mehraein and Riahi, 2017). Both concluded that SVM is more capable of modeling these systems. Recently, Benmouloud et al. (2024) used the QSPR approach to develop ANN and SVM models for predicting CO_2 solubility, and they used 13 PaDel descriptors. The

Table 2
Statistical parameters (R^2 and MAE) obtained from the developed ML models and the literature.

Model	R^2 (training set)	R^2 (test set)	MAE (training set)	MAE (test set)
MLR (this work)	0.8647	0.8678	0.0531	0.0532
UNIFAC (Zhou et al., 2021)	0.8799[b]	[a]	0.0443[b]	[a]
UNIFAC (Chen et al., 2020)	0.8778	0.6301	0.0477	0.0850
ANN (this work)	0.9757	0.9741	0.0235	0.0240
ANN (Song et al., 2020)	0.9842	0.9836	0.0200	0.0202
ANN (Ali et al., 2024)	0.991	0.986	0.0141	0.0171
ANN (Benmouloud et al., 2024)	0.9807	0.9828	0.0199	0.0195
RF (this work)	0.9853	0.9754	0.0198	0.0257
RF (Liu et al., 2023)	0.9973	0.9817	0.0076	0.0210
RF (Venkatraman and Alsberg, 2017)	0.9200[b]		0.0400[b]	
RF (Ali et al., 2024)	0.974[b]	[a]	0.0232[b]	[a]

a = not reported in the literature.
b = global.

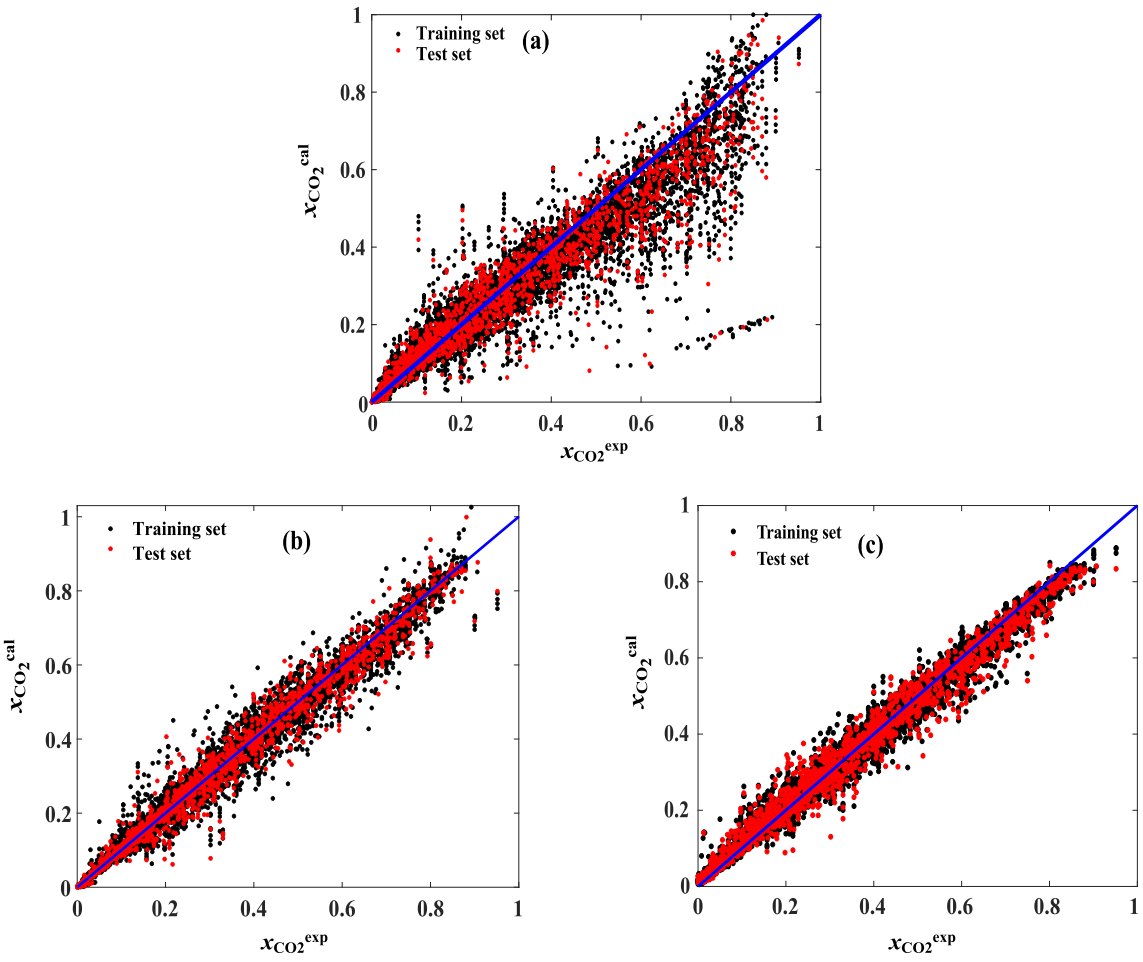

Fig. 6. Comparison between experimental data and (a) MLR, (b) ANN, and (c) RF models.

random forest model (RF) has been applied to predicting CO_2 solubility in ILs by using an extensive dataset (Venkatraman and Alsberg, 2017). The model results were compared to COSMO-RS predictions, in which the QSPR approach with the RF model was more accurate. In addition, ML models have been developed for predicting CO_2 in ILs by using structural encoding (Liu et al., 2023).

One approach for ML modeling is to use the σ-profile-based descriptors. The σ-profile represents the distribution of screening charge density on a molecule's surface, whose geometry is optimized using quantum chemistry (Klamt, 2005). It provides insight into how molecules interact at a molecular level. The σ-profile compresses three-dimensional information to a two-dimensional histogram, which is a key advantage of using σ-profile-based descriptors. Thus, it has better generalizability than the GC descriptors, which are restricted to the functional groups included in the model development. On the other hand, obtaining σ-profile-based descriptors can be challenging and time-consuming since they are calculated via quantum chemistry methods. The σ-profile has been used as a descriptor in ML modeling for predicting a wide range of IL properties such as density, viscosity, and melting point temperature (Lemaoui et al., 2024; Zhao et al., 2015). From an ML modeling point of view, the σ-profile descriptors have a balance between the number of descriptors and the performance (Darwish et al., 2024; Lemaoui et al., 2024). The σ-profile descriptors have been used successfully in ML modeling of CO_2 solubility in deep eutectic solvents (DES) (Lemaoui et al., 2023; Mohan et al., 2024; Wang et al., 2021). As far as we know, the σ-profile descriptors have not been used to predict CO_2 solubility in ILs using an extensive dataset and ML methods.

The purpose of this work is to provide an alternative approach for overcoming the limitations of the GC method for predicting CO_2 solubility across a wide range of ILs. First, we developed an MLR model based on the σ-profile descriptors. Following this, non-linear ML methods including ANN and RF, were applied. Lastly, we tested the predicting abilities of the models by predicting CO_2 solubility for 1568 ILs. The methodology is presented in Section 2, and the model development in Section 3. The results and discussion are in Section 4, and predictions made with the developed model are in Section 5. Lastly, the conclusions are provided in Section 6.

2. Methodology

2.1. Dataset

We obtained the experimental dataset for developing the MLR, ANN, and RF models from the work of Song et al. (2020). The dataset originally had 10,116 data points for 124 ILs, but after removing duplicated points, 9684 data points remained. The complete dataset and descriptors are provided in Table S1, and the removed duplicate points are specified in Table S2. The temperature, pressure, and solubility range from 243.15 K to 453.15 K, from 0.798 kPa to 49990 kPa, and from 0.0000648 to 0.9516 in mole fraction (x_{CO2}) in the dataset. The names and abbreviations of the ions can be found in the Table S3.

2.2. The σ-profile descriptors

The σ-profile is an essential part in the theory of COSMO-RS, where it represents the polarity of the molecule, calculated from quantum chemistry (Eckert and Klamt, 2002; Klamt et al., 1998). The σ-profile is a two-dimensional histogram derived from three-dimensional optimized

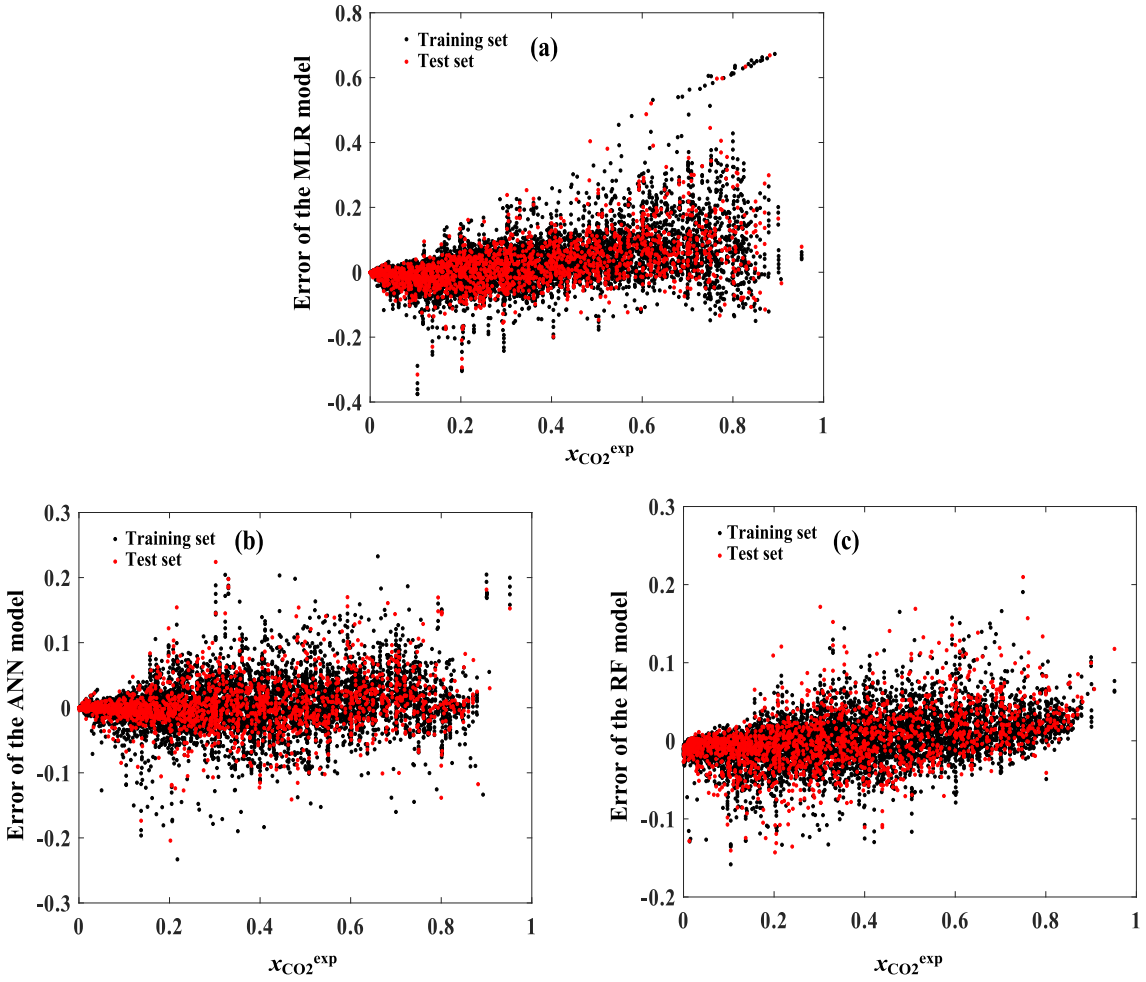

Fig. 7. The deviation between experimental data and (a) MLR, (b) ANN, and (c) RF models.

geometry, where the x-axis corresponds to a specific surface charge with a unit of e/Å^2, and the y-axis tells the likelihood (amount) of finding this surface charge in the molecule. The surface charge is reported at 0.001 e/Å^2 intervals, and its range is typically from −0.03 to 0.03 e/Å^2.

The methodology of deriving the descriptors from the σ-profile for machine learning models is summarized in Fig. 1. First, we collected the σ-profiles of ions from the database (COSMObase 2023). If the σ-profile was not available in the database, we calculated it with the quantum chemical calculations using TURBOMOLE V7.8 and TZVP-basis set. The actual σ-profile calculations were performed in the COSMOtherm 2023 software with BP-TZVP_23.ctd parametrization. The final σ-profile was calculated from the conformation which had the lowest energy.

Following this, the σ-profiles of cations and anions were divided into four segments (S_{c1}-S_{c4} and S_{a1}-S_{a4}) with specific intervals (Fig. 2.). These represent the hydrogen bond donor (HBD), non-polar, and hydrogen bond acceptor (HBA) characteristics (Lemaoui et al., 2024). S_{c1} and S_{a1} are calculated from −0.03 to −0.008 e/Å^2 and correspond to HBD characteristics. S_{c2} and S_{a2} are calculated from −0.008 to 0 e/Å^2, while S_{c3} and S_{a3} are from 0 to +0.008 e/Å^2, which are related to non-polar characteristics. S_{c4} and S_{a4} are calculated from +0.008 to +0.03 e/Å^2 corresponding to HBA characteristics. Finally, the surface area of each segment was calculated by using MATLAB® and the trapz function.

2.3. Multiple linear regression

MLR is a widely used ML method for modeling IL properties such as density, viscosity, surface tension, and melting point temperature (Cao et al., 2024; Koutsoukos et al., 2021; Lemaoui et al., 2024; Zhao et al., 2015). In addition, MLR has been applied to modeling CO_2 solubility in ILs (Mehraein and Riahi, 2017). In the MLR method, linear equations are regressed between data and independent variables, where simplicity is one of its benefits. MLR method is formulated in Eq. (1).

$$y = w_0 + w_1\beta_1 + w_2\beta_2 + w_3\beta_3 + \ldots \quad (1)$$

Where y = predicted property, w = adjustable parameter, and β = descriptor.

We developed the MLR model in MATLAB® by using an algorithm called GlobalSearch. We also constrained this algorithm to have values between 0 and 1 for predicted CO_2 solubility to ensure that predictions are physically meaningful. Without this constraint, the regression produced physically infeasible results without a significant increase in accuracy.

2.4. Artificial neural network

ANN is a popular ML technique inspired by biological neural networks. It uses neurons as a computational unit, and each layer is constructed from a set of neurons. The weighted input for neurons is calculated by the activation function, which can be linear or non-linear. The design of the ANN model is constructed by several neurons and layers as well as the choice of activation functions. ANN model can be applied for complex modeling due to jointly fitted parameters (Aggarwal, 2018). However, ANN can be vulnerable to overfitting the data, in which the model learns the noise in addition to patterns from the data (Koutsoukos et al., 2021). ANN has been widely used in predicting

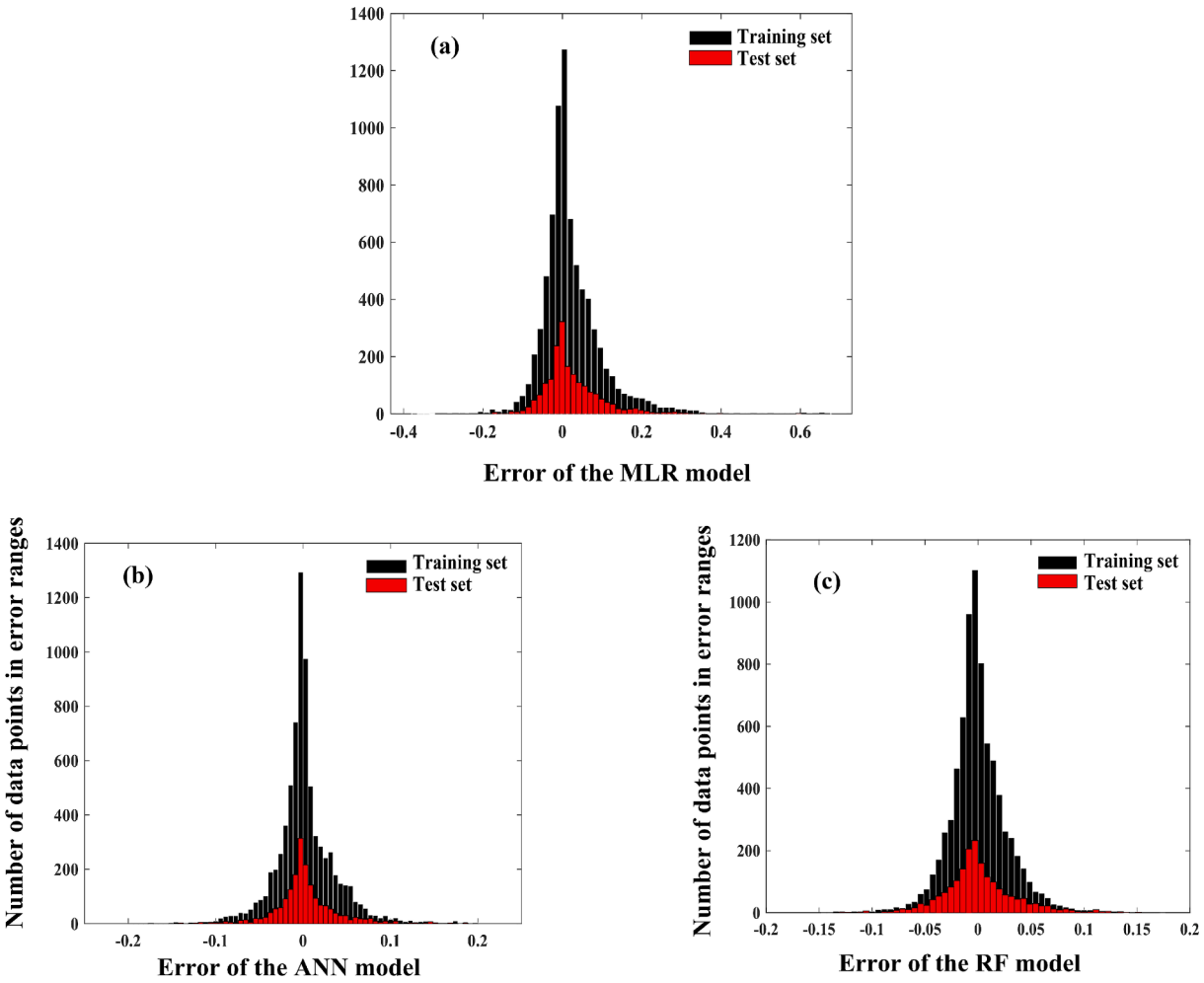

Fig. 8. The distribution of errors for (a) MLR, (b) ANN, and (c) RF models.

pure IL properties (Lemaoui et al., 2024; Paduszyński, 2019). Moreover, ANN modeling has been applied to predict CO_2 solubility in ILs by using GC (Song et al., 2020) and PaDel (Benmouloud et al., 2024) descriptors.

The ANN model developed here has two hidden layers with 8 and 6 neurons. Typically, one hidden layer is sufficient for an ANN model (Aggarwal, 2018). However, it was observed that using two hidden layers significantly decreased the number of unphysical predictions (see Section 5). A schematic representation of the developed neural network architecture is shown in Fig. 3. The input layer transmits descriptors for the first hidden layer without any computations. Weights (w) and biases (b) are added to input matrix (x) containing descriptors (Ln T (K), Ln P (kPa), and segments) as input in Eq. (2).

$$f_1(a_1) = f_1(w_1 \times x + b_1) \tag{2}$$

The activated signal (a) continues to the first hidden layer through activation functions. This is repeated for all hidden layers, as shown in Eq. (3)–(4).

$$f_2(a_2) = f_2(w_2 \times f_1(a_1) + b_2) \tag{3}$$

$$f_3(a_3) = f_3(w_3 \times f_2(a_2) + b_3) \tag{4}$$

Two hidden layers are activated with the hyperbolic tangent sigmoid function (*tansig*) described in Eq. (5). The signal from the last hidden layer is transferred to the output layer where the objective function is calculated. We utilized a linear activation function (*purelin*) to transfer the output signal as shown in Eq. (6), while the target was set to Ln x_{CO2}.

$$f_1(x) = \frac{2}{1 + e^{-2x}} - 1 \tag{5}$$

$$f_3(x) = x \tag{6}$$

2.5. Random forest

The RF method (Breiman, 2001) is a decision tree-based approach for ML modeling. It is a popular approach for predicting IL properties such as density, viscosity, melting point temperature, and CO_2 solubility (Cao et al., 2024; Koutsoukos et al., 2021; Lemaoui et al., 2024; Venkatraman and Alsberg, 2017). In the RF modeling approach, descriptor space is split into sets of rectangles (trees) that can be used to train the model. The advantage of the RF method arises from the idea of training an extensive number of non-correlative (random) trees and then averaging them. This method naturally avoids overfitting, which can be a problem for other tree-based ML models. However, the RF method is likely to perform poorly with a small dataset. For a regression problem, the output is taken from averaging the result of multiple trees. The function for regression problems in the RF method is shown in Eq. (7) (Hastie, 2009). A Schematic representation of the RF concept is shown in Fig. 4.

$$\widehat{f}_{RF}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T(x;\Theta_b) \tag{7}$$

Where $\widehat{f}_{RF}^{B}$ = model prediction, b = 1, 2…B is bootstrap sample, x descriptors, T = prediction from tree after all has been grown, and Θ_b = b:

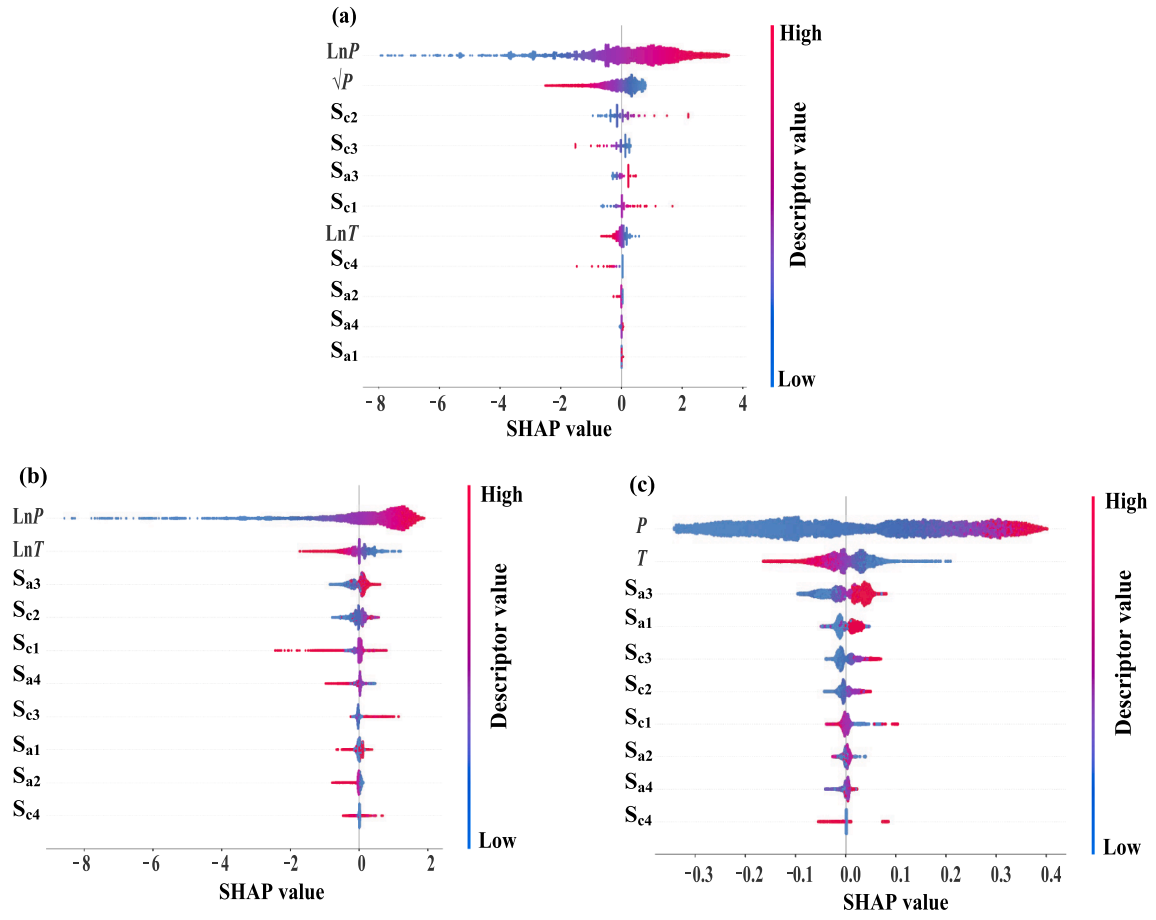

Fig. 9. Evaluation of descriptor importance via SHAP values for the (a) MLR, (b) ANN, and (c) RF models.

th random forest tree.

We developed the RF model by utilizing the MATLAB® *Treebagger* function with the following characteristics: 50 trees and a minimum node size of 5. We used *P* (kPa), *T* (K), and segments as descriptors, with x_{CO2} as a training target.

2.6. Descriptor importance evaluation

Evaluating the importance of descriptors is a crucial part of model development. The importance of descriptors offers insight into the developed ML models and helps evaluate their predictive capability. The SHAP values offer a robust way to evaluate the complex impacts of descriptors on ML models (Lundberg and Lee, 2017). The SHAP provides a clear and consistent explanation for how descriptors influence the model outcome. This is estimated by calculating SHAP values for each individual descriptor according to Eq. (8). In the SHAP analysis, positive values indicate a positive contribution, while negative values indicate an opposite effect.

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z_i' \tag{8}$$

Where *g* is explanatory model, *z* is telling that does descriptor exist (0,1), *M* is number of inputs, ϕ_i is Shapley values for each descriptor, and ϕ_0 is the constant correction factor.

3. Model development

We developed MLR, ANN, and RF models to predict CO_2 solubility in ILs. We divided training (Table S4) and test (Table S5) sets with an 80/20 ratio. All anions in the dataset were included in training and test sets. This data division was used since the experimental data is unevenly distributed between ions, where [BMIM], [HMIM], [EMIM], and [OMIM] dominated the cation distribution (Fig. 5a) and [Tf2N], [BF4], [PF6], [C(CN)3], and [TfO] dominated the anion distribution (Fig. 5b). This way, we ensured that the variability of cations and anions was high in trained models.

We used two criteria for selecting the final models. The first criterion was the accuracy of regression calculated with R^2 (Eq. (9)) and MAE (Eq. (10)), and the second was to have a minimum amount of unphysical predictions (under 0 or over 1 x_{CO2}). The second criterion was followed by making solubility predictions at 5000 kPa and 323.15 K temperature for all possible cation and anion combinations from the used data, resulting in 1568 ILs and solubility predictions.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left(y_i^{exp} - y_i^{pre}\right)^2}{\sum_{i=1}^{n} \left(y_i^{exp} - \bar{y}_i\right)^2} \tag{9}$$

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} \left|y_i^{exp} - y_i^{pre}\right|}{n} \tag{10}$$

Where y^{exp} is experimental value, y^{pre} is predicted value, $\bar{y}$ mean of experimental values, and *n* is number of data points.

4. Results and discussion

The developed MLR model for predicting CO_2 solubility in ILs is shown in Eq. (11), and values for regressed parameters are presented in Table 1. We used two variables to describe the pressure behavior in the MLR model. This approach was taken since it significantly improved the accuracy of the MLR model.

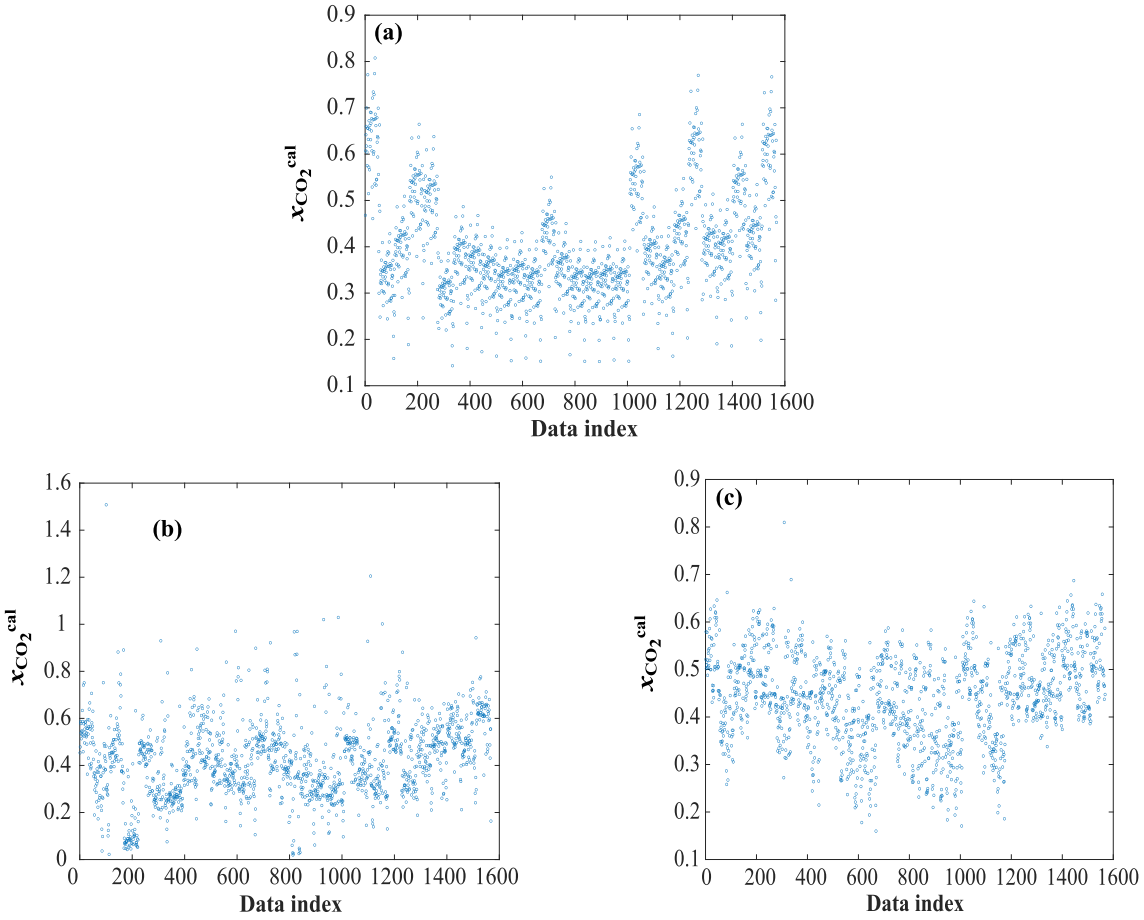

Fig. 10. Predictions for CO_2 solubility at 5000 kPa and 323.15 K for every ion combination in the dataset by using (a) MLR, (b) ANN, and (c) RF models.

$$Ln(x_{CO_2}) = w_1 Ln(T) + w_2 Ln(P) + w_3\sqrt{P} + w_4 S_{c1} + w_5 S_{c2} + w_6 S_{c3} + w_7 S_{c4} + w_8 S_{a1} + w_9 S_{a2} + w_{10} S_{a3} + w_{11} S_{a4} \quad (11)$$

The accuracy of all developed ML models is presented in Table 2, and the models are compared to experimental data in Fig. 6. The performance of the MLR model was compared to the UNIFAC models (Chen et al., 2020; Zhou et al., 2021). The model from the study by Zhou et al. (2021) uses the same dataset as in this work (with duplicated points, see Section 2) and they calculated the activity coefficient for ILs instead of CO_2 solubility expressed as mole fraction (x_{CO2}). Thus, x_{CO2} was calculated using the same approach as in their work (Table S6). Chen et al. (2020) used a significantly smaller dataset consisting of 3428 data points with reported x_{CO2} values. The accuracy of the MLR model is comparable with UNIFAC models. The UNIFAC model is derived from thermodynamics, thus it is surprising that it produces unphysical CO_2 solubility predictions ($x_{CO2} > 1$ or $0 > x_{CO2}$). The UNIFAC models had a similar number of physically unphysical predictions (125 Zhou et al. (2021) and 120 Chen et al. (2020)), corresponding to 1.2 % and 3.5 % from all solubility predictions, respectively. Our MLR model was restricted to a physically meaningful range. Thus, it did not produce any unphysical values.

Among the developed ML models, the ANN and RF models outperformed the MLR model. The performance of the ANN model is compared to experimental data in Fig. 6b. Although the ANN model outperformed the MLR model, the ANN model produced one unphysical prediction in the training set ($x_{CO2} > 1$). These models cannot predict negative mole fractions since the training target is formulated as Ln x_{CO2}.

We compared our ANN model to previously developed ANN models that used the same dataset as in this work (with duplicated points, see Section 2). Song et al. (2020) developed an ANN model by using 51 GC descriptors, and a deep ANN model developed by Ali et al. (2024) used the same 51 GC descriptors and Benmouloud et al. (2024) developed an ANN model by using 13 PaDEL descriptors. The accuracy of our ANN model is similar to previously developed ANN models, where these models differ in terms of number of descriptors. Our model used 10, the model of Benmouloud et al. (2024) used 13, and the Song et al. (2020), and Ali et al. (2024) models used 51 descriptors.

The best performance from the developed model was obtained with the RF model, which is compared to experimental data in Fig. 6c. The accuracy of our RF model is in the same range as previously developed RF models. However, our model has the fewest number of descriptors compared to other RF models from the literature. The highest accuracy found in the literature belongs to tree-based ML models that used so-called hybrid descriptors (Yang et al., 2024) that combine GC with other descriptors. However, the improved accuracy comes at the cost of a high number of descriptors, reaching a total of 133.

Fig. 7 shows the errors of all developed ML models and Fig. 8 shows the distribution of errors. In the case of MLR, the magnitude of errors is the largest (Fig. 7a) and the distribution of errors is the widest (Fig. 8a) compared to the ANN and RF models. The developed ANN and RF models have similar errors and deviations, whereas the magnitude of errors is slightly smaller, and the distribution of the errors is narrower in the case of the RF model.

4.1. Results of the descriptor importance analysis

The importance of descriptors was estimated by calculating SHAP

values for all models, as shown in Fig. 9. In the case of the MLR model (Fig. 9a), the importance of descriptors is listed in descending order based on SHAP values ($\mathrm{Ln}P > \sqrt{P} > S_{c2} > S_{c3} > S_{a3} > S_{c1} > \mathrm{Ln}T > S_{c4} > S_{a2} > S_{a4} > S_{a1}$). Two pressure variables were the most important for the MLR model. However, they have opposite trends. Increasing the value of $\mathrm{Ln}P$ predicted higher CO_2 solubility, while $\sqrt{P}$ had the opposite effect. This contradictory pressure effect does not follow the thermodynamic laws, but it significantly increased the model accuracy. Among all segments, those corresponding to non-polar characteristics were more important than those omitting HBD and HBA characteristics, with the exception of the cation HBD segment (S_{c1}). The two most important segments are associated with cation non-polar characteristics (S_{c2} and S_{c3}), while the two least important segments correspond to anion HBA and HBD characteristics.

Fig. 9b shows the importance of the descriptors for the ANN model. The descriptor importance is listed in descending order based on SHAP values ($\mathrm{Ln}P > \mathrm{Ln}T > S_{a3} > S_{c2} > S_{c1} > S_{a4} > S_{c3} > S_{a1} > S_{a2} > S_{c4}$). Pressure and temperature were the two most important descriptors, and they were consistent with the thermodynamic laws. Regarding segments, the non-polar segments (S_{a3} and S_{c2}) of the cations and anions were the most important descriptors, while the HBA segments of the cation were the least important.

The descriptor importance in the RF model is shown in Fig. 9c, and the descriptor importance is listed in descending order based on SHAP values ($P > T > S_{a3} > S_{a1} > S_{c3} > S_{c2} > S_{c1} > S_{a2} > S_{a4} > S_{c4}$). Once again, pressure and temperature were the two most important descriptors, which is consistent with thermodynamic theory. In the case of cation segments, non-polar segments were more important than HBD and HBA segments. This also applies to the segments of anions, with one exception where the HBA segment (S_{a1}) was more important than the non-polar segment (S_{a2}). In the RF model, the CO_2 solubility increased with non-polar cation and HBA anion segments. In contrast, increasing the values of the cation HBD and HBA as well as the anion HBD segments resulted in a decrease in CO_2 solubility. The behavior of the descriptors in the RF model followed a trend observed in the literature, where the CO_2 solubility follows the non-polar characteristic of ions (Sumon and Henni, 2011).

5. Prediction of CO_2 solubility

The benefit of the ML model lies in its ability to make predictions for unmeasured properties. Thus, it is important to test the predictive capability of the developed ML models. We predicted CO_2 solubility in the ILs with MLR, ANN, and RF models for each ion combination in the dataset. This corresponds to predictions for 1568 ILs, of which 124 are present in model training, resulting in 1444 unstudied ILs. We chose 5000 kPa and 323.15 K conditions for the predictions (Fig. 10). The pressure has the most significant impact on the models, according to SHAP analysis (see Section 4.1). Therefore, the choice of pressure was expected to influence the predictions more than temperature. It is challenging to evaluate the accuracy of predictions for unmeasured ILs. However, a number of unphysical predictions can be detected for CO_2 solubility, which is expressed as mole fractions ($x_{CO2} < 0$ and $1 < x_{CO2}$). We used the number of unphysical predictions together with the accuracy of the models as an indicator for the prediction ability of the models. Furthermore, the number of unphysical predictions was used as a selection criterion for developing the model (see Section 3). The MLR and RF models made no unphysical predictions, while the ANN had 5 $x_{CO2} > 1$ predictions. These unphysical predictions made by the ANN model correspond to 0.3 % of all made predictions. Thus the ANN model produced only a low number of unphysical predictions.

The same CO_2 solubility predictions, which are shown in Fig. 10, have been visualized with heatmaps in Fig. 11. These heatmaps show more clearly the predictions for each individual IL and the differences

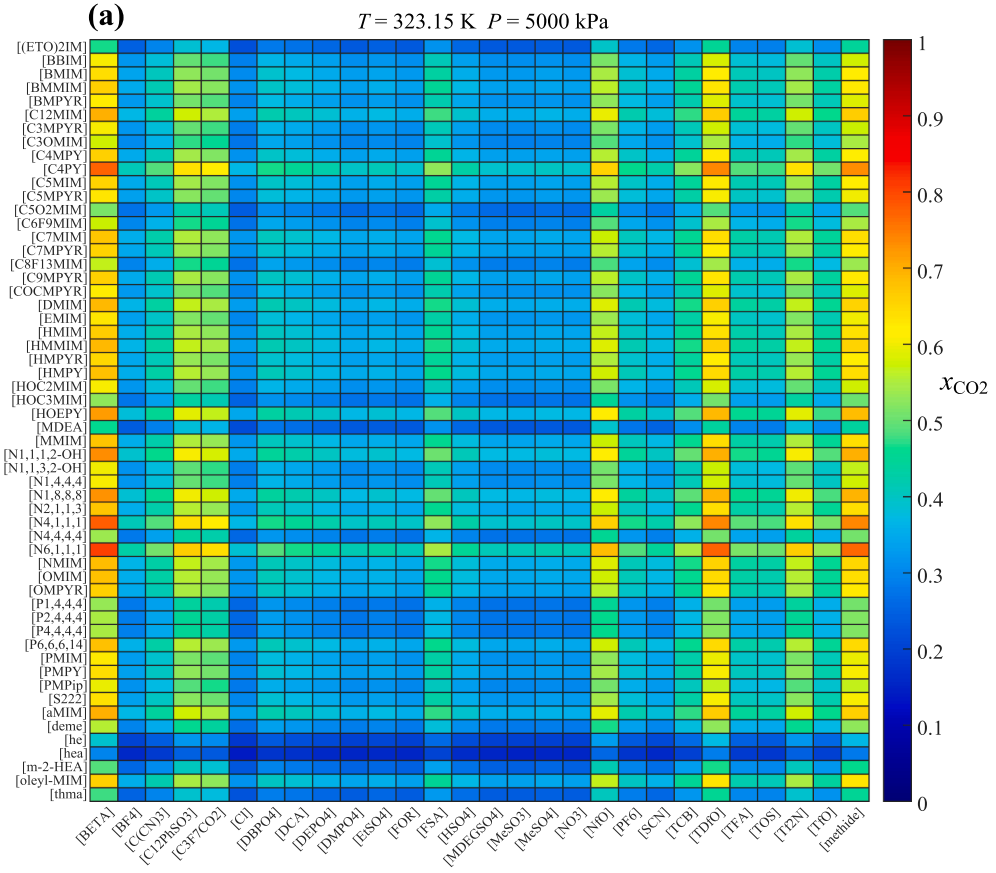

Fig. 11. Predictions for CO_2 solubility shown as heatmaps for (a) MLR, (b) ANN, and (c) RF models.

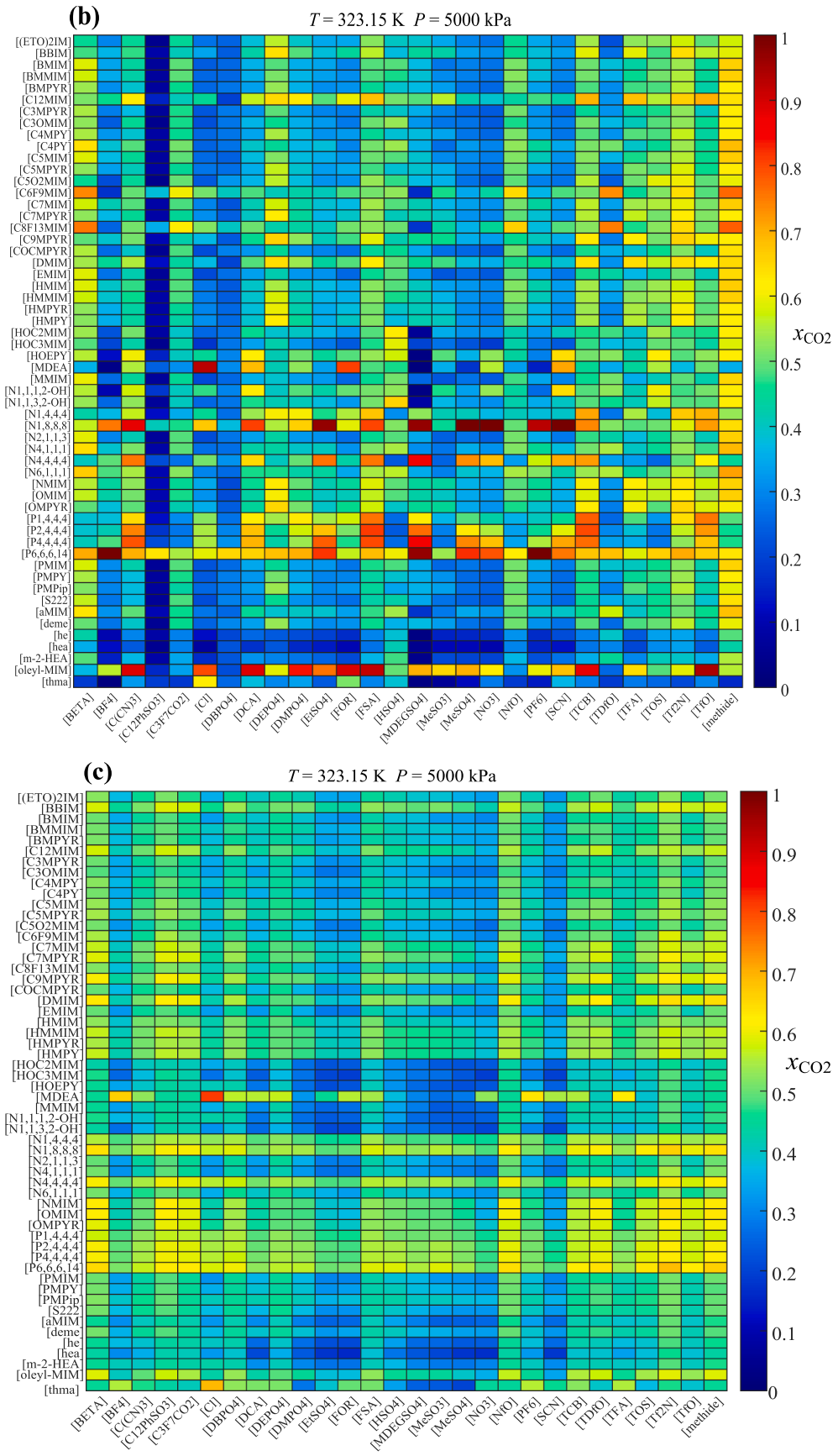

Fig. 11. (*continued*).

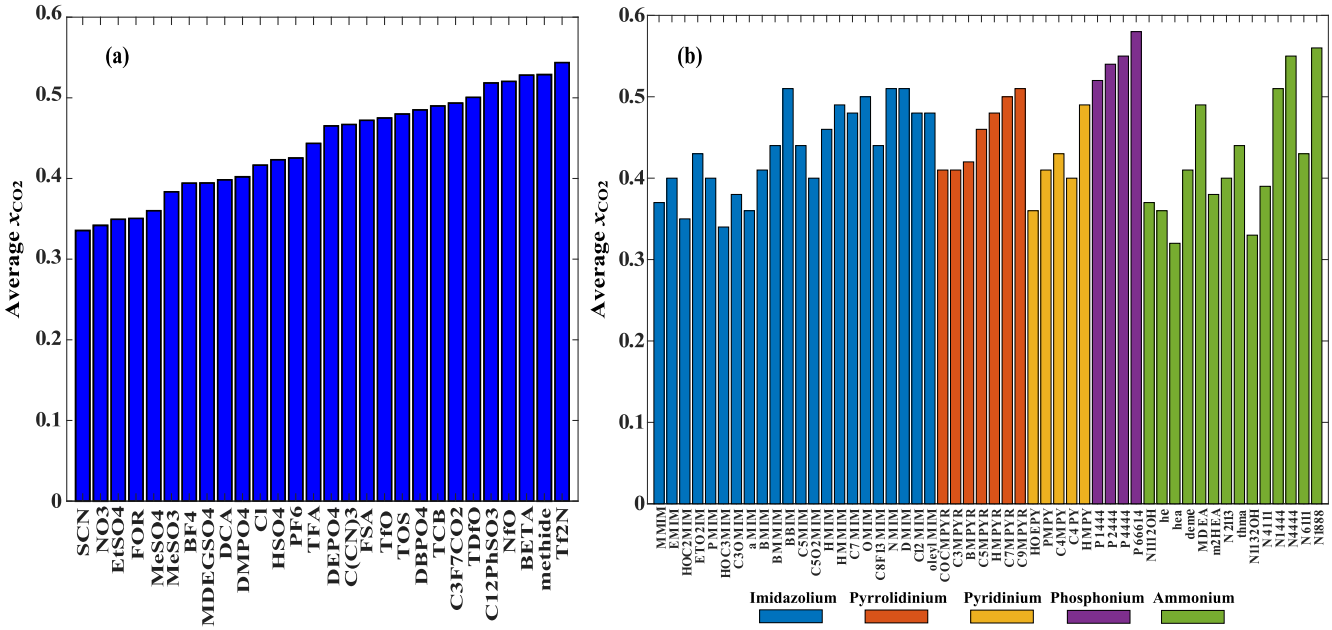

Fig. 12. The average of x_{CO2} prediction at 5000 kPa and 323.15 K made by the RF model while keeping anion constant and using all studied cations (a) and keeping cation constant and using all studied anion (b).

between predictions made by the developed models. In the ANN model, the solubility range for any given ion is large. For example, the solubility range is from close $x_{CO2} = 0$ to near $x_{CO2} = 1$ for anion [MDEGSO4] with different cations. The solubility range of ions is more evenly distributed in the MLR and RF models. The MLR model has higher x_{CO2} variability in certain anions, for example, in the case of [BETA] anion. The RF model had the least variability regarding the CO_2 solubility predictions with a constant anion. However, [Cl] anion produced a large solubility range in the RF model prediction. It is interesting that the statistical parameters in the case of ANN and RF models are close to each other, but the predictions under the studied conditions vary significantly.

The structural trends of the ions were evaluated from the RF predictions at 5000 kPa and 323.15 K since the RF model had the highest accuracy without any unphysical predictions. The structural trends of ions were calculated by keeping either the anion (Fig. 12a) or the cation (Fig. 12b) constant and averaging the CO_2 solubility of all studied cations or anions. Anions have a clear impact on CO_2 solubility without having simple structural trends. The RF model predicts that anion [Tf2N] has the biggest and [SCN] has the lowest positive impact on CO_2 solubility. However, the structural variation in the cation has clear trends (I-IV). (I) An increase in the alkyl chain length of the cations increases the solubility in terms of mole fraction (EMIM < HMIM). (II) The alkylation of phosphonium and ammonium cation increases the solubility (P1444 < P4444 and N1444 < N4444). (III) Branching the cation increases the solubility (BMIM < BMMIM). (IV) The OH-functionalization decreased the CO_2 solubility (EMIM > HOC2MIM). These trends are also reported in the literature (Sumon and Henni, 2011). Furthermore, the cation family influences CO_2 solubility, with the phosphonium family having the highest positive impact, according to the RF model predictions.

6. Conclusions

In this study, MLR, ANN, and RF models were developed to predict CO_2 solubility in ILs. A dataset with 9684 data points from 124 ILs was used to train these models, while descriptors were derived from the σ-profiles. The ANN and RF models outperformed the MLR model in terms of accuracy, with the RF achieving the best performance (R^2 = 0.9754 and MAE = 0.0257) and the MLR achieving the worst (R^2 = 0.8678 and MAE = 0.0532) on the test set. The accuracies of the ANN and RF models were comparable to GC-based ML models reported in the literature.

We estimated the importance of descriptors by calculating the SHAP values for all ML models. The ANN and RF models ranked pressure and temperature as the most essential descriptors, as expected from the thermodynamic theories. In the RF model, SHAP values also highlighted the importance of descriptors, that correspond to the non-polar characteristics of ions. An increase in the non-polar area of cations and anions results in higher CO_2 solubility in the case of the RF models.

We made CO_2 solubility predictions at 5000 kPa and 323.15 K for 1568 ILs, of which 1444 were unstudied. We also tracked the unphysical predictions ($x_{CO2} > 1$ and $x_{CO2} < 0$) within these CO_2 solubility predictions. The MLR and RF models did not predict any unphysical solubilities, while the ANN model predicted five unphysical solubilities. The contribution of ions to the prediction made by the RF model was analyzed. The anion clearly impacts CO_2 solubility, though distinct trends were not observed with structural variation. Anions [Tf2N] and [SCN] had the highest and lowest positive impact on the solubility, respectively. Clear trends were observed in the structural variation of the cations. Increasing alkyl chain length, alkylation of the phosphonium and ammonium cations, and branching the cation increases the solubility. In contrast, OH-functionalization of the cation decreases the solubility. According to predictions made by the RF model, the phosphonium cation family exhibited the highest CO_2 solubility.

Our study demonstrated that the combination of ML tools with the σ-profile descriptors can offer better generalization than using GC descriptors while maintaining similar accuracy to experimental data for predicting CO_2 solubility in ILs. GC descriptors are restricted to functional groups used in model development, whereas the σ-profile descriptors are free from this restriction. Thus, the σ-profile can produce predictions for a wider range of ILs. Moreover, the σ-profile provides insight into the molecular interactions, giving it more physical relevance than the GC approach. Future work could explore other physically relevant descriptors to be combined with the σ-profile to increase the accuracy of this approach.

CRediT authorship contribution statement

Juho-Pekka Laakso: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Ali Ebrahimpoor Gorji:** Writing – review & editing, Validation, Conceptualization. **Petri Uusi-Kyyny:** Writing – review & editing, Supervision, Resources, Project administration. **Ville Alopaeus:** Writing – review & editing, Supervision, Resources, Project administration.

Declaration of competing interest

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

Acknowledgment

This work was financially supported by the Academy of Finland 'In-situ equilibrium shifting in CO_2 utilization reactions by novel absorbents (CO_2Shift)' Project (351113). The authors wish to acknowledge CSC – IT Center for Science, Finland, for computational resources.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ces.2025.121226.

Data availability

Data will be made available on request.

References

Aggarwal, C.C., 2018. Neural networks and deep learning : a textbook. Springer, Cham, Switzerland. https://doi.org/10.1007/978-3-319-94463-0.

Aghaie, M., Rezaei, N., Zendehboudi, S., 2018. A systematic review on CO2 capture with ionic liquids: Current status and future prospects. Renew. Sustain. Energy Rev. 96, 502–525. https://doi.org/10.1016/j.rser.2018.07.004.

Ali, M., Sarwar, T., Mubarak, N.M., Karri, R.R., Ghalib, L., Bibi, A., Mazari, S.A., 2024. Prediction of CO2 solubility in Ionic liquids for CO2 capture using deep learning models. Sci. Rep. 14. https://doi.org/10.1038/s41598-024-65499-y.

Benmouloud, W., Euldji, I., Si-Moussa, C., Benkortbi, O., 2024. Quantitative structure-property relationship techniques for predicting carbon dioxide solubility in ionic liquids using machine learning methods. Int. J. Quantum Chem 124. https://doi.org/10.1002/qua.27450.

Beraldo, C.S., Liang, X., Follegatti-Romero, L.A., 2024. Predicting the solubility of gases in imidazolium-based ionic liquids with SAFT-VR Mie EoS by a novel approach based on COSMO. Chem. Eng. Sci. 285. https://doi.org/10.1016/j.ces.2023.119610.

Breiman, L., 2001. Random Forests. Mach. Learn. 45, 5–32.

Cao, P., Chen, J., Chen, G., Qi, Z., Song, Z., 2024. A critical methodological revisit on group-contribution based property prediction of ionic liquids with machine learning. Chem. Eng. Sci. 298. https://doi.org/10.1016/j.ces.2024.120395.

Chen, Y., Liu, X., Woodley, J.M., Kontogeorgis, G.M., 2020. Gas Solubility in Ionic Liquids: UNIFAC-IL Model Extension. Ind. Eng. Chem. Res. 59, 16805–16821. https://doi.org/10.1021/acs.iecr.0c02769.

Chen, Y., Mu, T., 2019. Conversion of CO2 to value-added products mediated by ionic liquids. Green Chem. 21, 2544–2574. https://doi.org/10.1039/c9gc00827f.

Chen, Y., Mutelet, F., Jaubert, J.N., 2012. Modeling the solubility of carbon dioxide in imidazolium-based ionic liquids with the PC-SAFT equation of state. J. Phys. Chem. B 116, 14375–14388. https://doi.org/10.1021/jp309944t.

Darwish, A.S., Lemaoui, T., Taher, H., AlNashef, I.M., Banat, F., 2024. High-throughput screening of 2,500 ionic liquids for sustainable furfural recovery: Bridging quantum simulations, machine learning, and experimental validation. Chem. Eng. J. 496. https://doi.org/10.1016/j.cej.2024.153965.

Davoodi, S., Thanh, H.V., Wood, D.A., Mehrad, M., Hajsaeedi, M.R., Rukavishnikov, V.S., 2024. Combined deep-learning optimization predictive models for determining carbon dioxide solubility in ionic liquids. J. Ind. Inf. Integr. 41. https://doi.org/10.1016/j.jii.2024.100662.

Diedenhofen, M., Klamt, A., 2010. COSMO-RS as a tool for property prediction of IL mixtures-A review. Fluid Phase Equilib. 294, 31–38. https://doi.org/10.1016/j.fluid.2010.02.002.

Eckert, F., Klamt, A., 2002. Fast Solvent Screening via Quantum Chemistry: The COSMO-RS Approach. AIChE J 48, 369–385. https://doi.org/10.1002/aic.690480220.

Fredenslund, A., Jones Russel, L., Prausnitz John, M., 1975. Group Contribution Estimation of Activity Coefficients in Nonideal Liquid Mixtures. AIChE J 21, 1086–1099.

Gani, R., 2019. Group contribution-based property estimation methods: advances and perspectives. Curr. Opin. Chem. Eng. 23, 184–196. https://doi.org/10.1016/j.coche.2019.04.007.

Ghaslani, D., Eshaghi Gorji, Z., Ebrahimpoor Gorji, A., Riahi, S., 2017. Descriptive and predictive models for Henry's law constant of CO2 in ionic liquids: A QSPR study. Chem. Eng. Res. Des. 120, 15–25. https://doi.org/10.1016/j.cherd.2016.12.020.

Hastie, T., 2009. The elements of statistical learning : data mining, inference, and prediction, 2nd, ed. ed. Springer series in statistics, Springer, New York.

Islam, N., Warsi Khan, H., Gari, A.A., Yusuf, M., Irshad, K., 2022. Screening of ionic liquids as sustainable greener solvents for the capture of greenhouse gases using COSMO-RS approach: Computational study. Fuel 330. https://doi.org/10.1016/j.fuel.2022.125540.

Jian, Y., Wang, Y., Barati Farimani, A., 2022. Predicting CO2 Absorption in Ionic Liquids with Molecular Descriptors and Explainable Graph Neural Networks. ACS Sustain. Chem. Eng. 10, 16681–16691. https://doi.org/10.1021/acssuschemeng.2c05985.

Klamt, A., 1995. Conductor-like screening model for real solvents: A new approach to the quantitative calculation of solvation phenomena. J. Phys. Chem. 99, 2224–2235. https://doi.org/10.1021/j100007a062.

Klamt, A., 2005. COSMO-RS: From quantum chemistry to fluid phase thermodynamics and drug design. From Quantum Chemistry to Fluid Phase Thermodynamics and Drug Design, COSMO-RS.

Klamt, A., Jonas, V., Bu, T., Lohrenz, J.C.W., 1998. Refinement and Parametrization of COSMO-RS. J. Phys. Chem. A 102, 5074–5085. https://doi.org/10.1021/jp980017s.

Koutsoukos, S., Philippi, F., Malaret, F., Welton, T., 2021. A review on machine learning algorithms for the ionic liquid chemical space. Chem. Sci. 12, 6820–6843. https://doi.org/10.1039/d1sc01000j.

Le, T., Epa, V.C., Burden, F.R., Winkler, D.A., 2012. Quantitative structure-property relationship modeling of diverse materials properties. Chem. Rev. 112, 2889–2919. https://doi.org/10.1021/cr200066h.

Lei, Z., Dai, C., Chen, B., 2014. Gas solubility in ionic liquids. Chem. Rev. 114, 1289–1326. https://doi.org/10.1021/cr300497a.

Lemaoui, T., Boublia, A., Lemaoui, S., Darwish, A.S., Ernst, B., Alam, M., Benguerba, Y., Banat, F., AlNashef, I.M., 2023. Predicting the CO2 Capture Capability of Deep Eutectic Solvents and Screening over 1000 of their Combinations Using Machine Learning. ACS Sustain. Chem. Eng. 11, 9564–9580. https://doi.org/10.1021/acssuschemeng.3c00415.

Lemaoui, T., Eid, T., Darwish, A.S., Arafat, H.A., Banat, F., AlNashef, I., 2024. Revolutionizing inverse design of ionic liquids through the multi-property prediction of over 300,000 novel variants using ensemble deep learning. Mater. Sci. Eng. R. Rep. 159. https://doi.org/10.1016/j.mser.2024.100798.

Liu, T., Fan, D., Chen, Y., Dai, Y., Jiao, Y., Cui, P., Wang, Y., Zhu, Z., 2023. Prediction of CO2 solubility in ionic liquids via convolutional autoencoder based on molecular structure encoding. AIChE J 69. https://doi.org/10.1002/aic.18182.

Lundberg, S., Lee, S.-I., 2017. A Unified Approach to Interpreting Model Predictions. Adv Neural Inf Process Syst.

MacDowell, N., Florin, N., Buchard, A., Hallett, J., Galindo, A., Jackson, G., Adjiman, C. S., Williams, C.K., Shah, N., Fennell, P., 2010. An overview of CO2 capture technologies. Energ. Environ. Sci. 3, 1645–1669. https://doi.org/10.1039/c004106h.

Mehraein, I., Riahi, S., 2017. The QSPR models to predict the solubility of CO2 in ionic liquids based on least-squares support vector machines and genetic algorithm-multi linear regression. J. Mol. Liq. 225, 521–530. https://doi.org/10.1016/j.molliq.2016.10.133.

Mohan, M., Demerdash, O.N., Simmons, B.A., Singh, S., Kidder, M.K., Smith, J.C., 2024. Physics-Based Machine Learning Models Predict Carbon Dioxide Solubility in Chemically Reactive Deep Eutectic Solvents. ACS Omega 9, 19548–19559. https://doi.org/10.1021/acsomega.4c01175.

Netzeva, T.I., Worth Andrew, P., Myatt, G., Nikolova-Jeliazkova, N., Patlewicz, G.Y., Perkins, R., Roberts, D.W., Schultz, T.W., Stanton, D.T., Van de sandt, J.J.M., Tong, W., Veith, G., Aldenberg, T., Yang, C., Benigni, R., Cronin, M.T.D., Gramatica, P., Jaworska, J.S., Kahn, S., Klopman, G., Marchant, C.A.,, 2005. Current status of methods for defining the applicability domain of (Quantitative) structure-activity relationships : The report and recommendations of ECVAM workshop 52. Altern. Lab. Anim. 33, 155–173. https://doi.org/10.1177/026119290503300209.

Paduszyński, K., 2019. Extensive Databases and Group Contribution QSPRs of Ionic Liquids Properties. 1. Density. Ind Eng Chem Res 58, 5322–5338. https://doi.org/10.1021/acs.iecr.9b00130.

Rogers, R.D., Seddon, K.R., 2003. Ionic liquids–solvents of the future? Science 1979 (302), 792–793.

Song, Z., Shi, H., Zhang, X., Zhou, T., 2020. Prediction of CO2 solubility in ionic liquids using machine learning methods. Chem. Eng. Sci. 223. https://doi.org/10.1016/j.ces.2020.115752.

Sumon, K.Z., Henni, A., 2011. Ionic liquids for CO2 capture using COSMO-RS: Effect of structure, properties and molecular interactions on solubility and selectivity. Fluid Phase Equilib. 310, 39–55. https://doi.org/10.1016/j.fluid.2011.06.038.

Sun, J., Sato, Y., Sakai, Y., Kansha, Y., 2023. A review of ionic liquids and deep eutectic solvents design for CO2 capture with machine learning. J. Clean. Prod. 414. https://doi.org/10.1016/j.jclepro.2023.137695.

Sun, S., Sun, H., Williams, P.T., Wu, C., 2021. Recent advances in integrated CO2 capture and utilization: a review. Sustain Energy Fuels 5, 4546–4559. https://doi.org/10.1039/d1se00797a.

Venkatraman, V., Alsberg, B.K., 2017. Predicting CO2 capture of ionic liquids using machine learning. J. CO2 Util. 21, 162–168. https://doi.org/10.1016/j.jcou.2017.06.012.

Wang, Y., Huang, S., Liu, X., He, M., 2023. Thermodynamic model for CO2 absorption in imidazolium-based ionic liquids using cubic plus association equation of state. J. Mol. Liq. 378. https://doi.org/10.1016/j.molliq.2023.121587.

Wang, J., Song, Z., Chen, L., Xu, T., Deng, L., Qi, Z., 2021. Prediction of CO2 solubility in deep eutectic solvents using random forest model based on COSMO-RS-derived descriptors. Green Chem. Eng. 2, 431–440. https://doi.org/10.1016/j.gce.2021.08.002.

Yang, A., Sun, S., Su, Y., Kong, Z.Y., Ren, J., Shen, W., 2024. Insight to the prediction of CO2 solubility in ionic liquids based on the interpretable machine learning model. Chem. Eng. Sci. 297. https://doi.org/10.1016/j.ces.2024.120266.

Zeng, S., Zhang, X., Bai, L., Zhang, X., Wang, H., Wang, J., Bao, D., Li, M., Liu, X., Zhang, S., 2017. Ionic-Liquid-Based CO2 Capture Systems: Structure, Interaction and Process. Chem. Rev. 117, 9625–9673. https://doi.org/10.1021/acs.chemrev.7b00072.

Zhao, Y., Huang, Y., Zhang, X., Zhang, S., 2015. A quantitative prediction of the viscosity of ionic liquids using Sσ-profile molecular descriptors. PCCP 17, 3761–3767. https://doi.org/10.1039/c4cp04712e.

Zhou, T., Shi, H., Ding, X., Zhou, Y., 2021. Thermodynamic modeling and rational design of ionic liquids for pre-combustion carbon capture. Chem. Eng. Sci. 229. https://doi.org/10.1016/j.ces.2020.116076.