Engineering Applications of Artificial Intelligence 167 (2026) 113770

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

A quantum chemistry-driven machine learning model for predicting solubility of carbon dioxide in ionic liquids☆

Tianxiong Liu [a,1], Wenguang Zhu [a,1], Ying Gao [b], Runqi Zhang [a], Yusen Chen [a], Chao Guo [c], Hongru Zhang [a], Jianguang Qi [a], Yinglong Wang [a,*], Peizhe Cui [a]

[a] College of Chemical Engineering, Qingdao University of Science and Technology, Zhengzhou Road 53, Shibei District, Qingdao, Shangdong, 266042, People's Republic of China

[b] School of Date Science, Qingdao University of Science and Technology, No. 99 Songling Road, Laoshan District, Qingdao, Shangdong, 266042, People's Republic of China

[c] College of Materials and Chemistry & Chemical Engineering, Chengdu University of Technology, Erxianqiao East Third Road 1, Chenghua District, Chengdu, Sichuan, 610059, People's Republic of China

ARTICLE INFO

Keywords:
Ionic liquids
Deep neural network
Quantum chemical
Solubility of carbon dioxide

ABSTRACT

Ionic liquids (ILs) are promising eco-friendly solvents for carbon dioxide (CO_2) dissolution and capture. Utilizing deep neural network modeling to accelerate the design and screening of ILs can contribute to promoting green and sustainable development. In this study, a quantitative structure-property relationship (QSPR) model was constructed to link the structure of ionic liquids with their CO_2 solvation ability. The deep neural network model was driven using two environmental descriptors, temperature and pressure, as well as 16 quantum chemical descriptors calculated from the Conductor-like Screening Model for Real Solvents (COSMO-RS). This model uniquely utilizes the sparsity of IL σ-profile curves for descriptor classification. The study explored the impact on machine learning modeling using two data splitting methods: "point-based" and "component-based". The former randomly divides the entire dataset into a training set and a test set, yielding Coefficient of Determination(R^2), Root Mean Square Error(RMSE), and Mean Absolute Error(MAE) values of 0.9904, 0.0216, and 0.0133, respectively, on the test set. The latter splits the dataset based on the type of ILs into set1 and set2, yielding R^2, RMSE, and MAE values of 0.9297, 0.0631, and 0.0450, respectively, on the test set. The model was further validated and explained using Applicability Domain (AD) and SHapley Additive exPlanations (SHAP) methods. This model provides accurate predictions of CO_2 solubility in ILs and offers guidance for designing ILs for CO_2 capture.

1. Introduction

Significant quantities of greenhouse gases in the atmosphere, particularly CO_2 released from industrial activities, are recognized as the primary driver of global warming. The escalating CO_2 emissions lead to rising global temperatures and sea levels, causing a range of environmental issues. To mitigate these challenges, it is crucial to find efficient methods for CO_2 capture and to reduce the cost of regeneration, which is of utmost importance in today's society. In addressing the issue of rising atmospheric CO_2 concentrations, Pancione et al. (2024) established a baseline for TSA based ocean carbon capture applications and promoted further research to design optimized carbon dioxide adsorption processes.

ILs are garnering significant attention due to their unique molecular structures (including anions and cations), specialized functional groups, and outstanding properties. Hydrogen bond donors (HBD) and hydrogen bond acceptors (HBA) in ILs play crucial roles in determining their physical and chemical properties (Dikki et al., 2024). The HBD groups in cations and HBA groups in anions can interact with target molecules (such as CO_2) via hydrogen bonding, thereby influencing the solubility

☆ Additional Supporting Information may be found in the online version of this article.

* Corresponding author.

E-mail addresses: tianxliu0717@163.com (T. Liu), zwgizwgi@163.com (W. Zhu), gaoying@qust.edu.cn (Y. Gao), zrq19806059693@163.com (R. Zhang), yusenchenmail@126.com (Y. Chen), guochao19@cdut.edu.cn (C. Guo), zhanghr103@163.com (H. Zhang), 03578@qust.edu.cn (J. Qi), wangyinglong@qust.edu.cn (Y. Wang), cpzmagic@qust.edu.cn (P. Cui).

[1] These authors contributed equally to this work and should be considered co-first authors.

https://doi.org/10.1016/j.engappai.2026.113770
Received 7 April 2024; Received in revised form 12 November 2025; Accepted 6 January 2026
Available online 8 January 2026

and selectivity of ILs. Strong HBD cations can enhance the attraction of ILs to electron-dense molecules (such as the oxygen atom in CO_2), while strong HBA anions can form hydrogen bonds with HBD-characteristic molecules (such as the carbon atom in CO_2), thus enhancing the solubility and capture capacity of ILs (Sistla and Sridhar, 2021). Compared to traditional organic solvents, ILs exhibit promising potential in CO_2 capture and separation.

ILs have emerged as novel media for CO_2 capture, with extensive research conducted on their mechanisms and performance. Both experimental and computational studies have demonstrated that certain ILs exhibit exceptional CO_2 capture capabilities under specific temperature and pressure conditions. For example, Torralba-Calleja et al. (2013) reviewed CO_2 absorption data for various ILs, including imidazolium-, pyrrolidinium-, pyridinium-, quaternary-ammonium-, and tetra-alkyl-phosphonium-based ionic liquids, affirming the versatility and inherent advantages of ILs in CO_2 capture, thus highlighting their promising potential. Aghaie et al. (2018) provided a concise review of various CO_2 capture technologies, detailing the processes and mechanisms of CO_2 capture using ILs, including the effects of molecular-level interactions, CO_2 solubility, and selectivity in ILs. Their review offers a comprehensive guide for researchers and engineers, addressing both thermodynamic and mass transfer aspects of CO_2/ILs processes. Additionally, Ramdin et al. (2012) focused on the experimental data regarding CO_2 solubility, selectivity, and diffusivity in different ILs, emphasizing the impact of anions, cations, and functional groups on CO_2 solubility, biodegradability, and toxicity. This body of research provides critical theoretical foundations for developing new and efficient CO_2 capture materials. The evidence supports the potential of ILs as effective CO_2 capture media, which could play a significant role in addressing global climate change and reducing greenhouse gas emissions.

However, there are countless types of ILs, and experimental synthesis is costly and time-consuming (Fan et al., 2023; Farahipour et al., 2016; Wang et al., 2022). Despite extensive efforts in experimental research, achieving effective CO_2 capture and storage remains a formidable task, making it highly inefficient to solely rely on experimentation for screening suitable ILs. To address this, Cho et al. (2017) explored the solubility of CO_2 in ILs with varying cyanide content in the anion. CO_2 solubility in these ILs was assessed through the measurement of the bubble point pressure of CO_2-IL mixtures across an extensive spectrum of temperatures and pressures. The study revealed that CO_2 solubility in ILs rose with pressure, declined with temperature, and was influenced by the cation's alkyl chain length. Recent studies, such as those by Panjapornpon et al. (2024), demonstrated the integration of experimental data with machine learning to predict CO_2 solubility in tubular reactors. Bardeeniz et al. (2024) used physics-guided methods to improve CO_2 equilibrium models for solubility prediction under varying pressures and temperatures. Both studies emphasized the importance of integrating physics-guided neural networks and state-of-the-art machine learning techniques to address the challenges in CO_2 solubility prediction.

In recent years, various thermodynamic models, such as E-NRTL, UNIQUAC, Peng-Robinson equation of state, PC-SAFT, and Redlich-Kwong(Asadi et al., 2020; Hassanpouryouzband et al., 2019; Kamgar and Rahimpour, 2016; Shiflett and Maginn, 2017; Yazdani et al., 2023), have shown successful applications in estimating CO_2 solubility in ILs systems. However, a notable drawback of these methods is their reliance on experimental data for fitting the mixing parameters that describe interactions between molecular components. This limitation poses challenges in developing novel ILs systems with broader applicability. In recent studies, Wang et al. (2023) conducted an investigation into the solubility behavior of imidazolium ILs using molecular dynamics (MD), highlighting the potential of this approach for solubility prediction. Furthermore, quantum chemical (QC) calculations, which consider IL-IL and IL-CO_2 interactions, demand significant computational resources. Surprisingly, the first-principles-based QC thermodynamic model COSMO-RS shows promise in predicting gas solubility and other thermodynamic properties, as it only requires molecular structure information for its calculations. Mohan et al. (2023a) combined the COSMO-RS model with machine learning to predict CO_2 solubility in various DESs. Using COSMO-RS predicted solubility and temperature-pressure parameters, they developed a multiple linear regression model with an AARD of 12 %. An ANN-based ML model using COSMO-RS features achieved an AARD of 2.72 %, closely matching experimental results. Li et al. (2024) established three ML models. The predictive molecular descriptors were derived from a combination of the Group Contribution methods, the COSMO-RS sigma-moments, and energy descriptors. The results indicated that both the ANN and XGBoost Regressor models provided more reliable predictions for Hansen solubility parameters of cocrystal coformers. In summary, machine learning models can serve as useful tools for designing and selecting DESs for CO_2 capture and utilization, and they also show excellent potential in predicting other solubilities. The combination of machine learning and advanced quantum chemistry methods can play a significant role in carbon capture and materials design.

With the rise of artificial intelligence technology, developing Quantitative Structure-Activity Relationship (QSAR) models to predict CO_2 solubility in ILs has garnered significant interest as a potential solution (Venkatraman et al., 2019). Leveraging powerful computational capabilities and large databases, these models can not only predict the properties of new IL structures but also provide physical insights that experimental methods might not reveal. Among machine learning (ML) models, artificial neural networks (ANNs) have received considerable attention for their ability to simulate complex situations and predict fluid properties and phase equilibria (Bakhbakhi, 2011; Liu et al., 2023). Numerous studies demonstrate the exceptional predictive performance of ANN models based on molecular descriptors. For instance, Daryayehsalameh et al. (2021) employed six different artificial intelligence techniques to predict CO_2 solubility in 1-butyl-3-methylimidazolium tetrafluoroborate. They found that a cascade feedforward neural network was the optimal model for the substances considered, achieving a prediction accuracy of AARD = 6.88 %, R^2 = 0.98808 for the entire experimental dataset. Caprio et al. (2025) proposed a mixed physics information model for predicting the solubility of carbon dioxide in absorption mixtures. This model aims to predict the behavior of novel mixtures by characterizing individual absorption mechanisms and combining physics based insights to evaluate the contribution of each mechanism based on the type of mixture. These studies highlight the superior predictive capabilities of ANN-based ML models for determining solvent thermodynamic properties. Given the diversity of IL systems, developing ANN models to predict CO_2 solubility in ILs is crucial for efficiently screening structurally diverse ILs. Such models hold significant promise for accelerating the discovery and design of ILs with desired CO_2 solubility properties, thus advancing various fields including carbon capture and other sustainable technologies.

The development of QSAR models relies on quantifying ILs' structures into data suitable for machine learning. Quantum chemical calculations provide effective COSMO-RS descriptors to quantify molecular structural changes, including the Sigma profile charge value (Sσ-profile), which represents surface charge density fragments of molecules (Abranches et al., 2022). The reliability of Sσ-profile in predicting ILs' properties has been demonstrated. Recently, Mohan et al. (2022) developed a machine learning model using Sσ-profile derived from COSMO-RS as input for an ANN, predicting the infinite dilution activity coefficients of ILs. Similarly, Boublia et al. (2022) used analogous descriptors to characterize DESs and employed an ANN model to predict their conductivity. Cao et al. (2018) developed an Extreme Learning Machine model using Sσ-profile descriptors to assess ILs' toxicity to leukemia rat cell lines, outperforming multiple linear regression and multilayer perceptron methods. The success of COSMO-RS derived descriptors in various machine learning applications indicates their potential for predicting CO_2 solubility in ILs.

In this study, a DNN model with multiple hidden layers was

developed to predict CO_2 solubility in ILs. The objective was to explore the physical characteristics of CO_2 solubility in ILs, adhering to Henry's law, and to investigate its relationship with pressure, temperature, and the structural configuration of cations and anions. Experimental data on CO_2 solubility in ILs were collected from the literature, and COSMO-RS software was used to generate new solubility data, which were then compared with experimental results. COSMO-RS was employed to produce surface charge density distributions (Sσ-profile) for cations and anions in various ILs, from which PSσ-profile descriptors were derived. These descriptors, along with pressure and temperature, were used as input features for the DNN model to develop a QSPR model for predicting CO_2 solubility in ILs. To ensure the model's reliability and accuracy, AD and SHAP methods were used to further validate and interpret the model results. By analyzing the contribution of different input features to the predictions, valuable insights were obtained. These insights provide significant guidance for designing new ILs with enhanced CO_2 solubility and offer substantial progress in environmental technologies such as carbon capture and storage.

2. Methodology

2.1. Data set

All data in this study were mainly from the review article by Lei et al. (2014). The original dataset contained over 10,000 data points, but there were duplicate data points, and some data points had very similar temperature and pressure values. Therefore, we removed these duplicate and anomalous data points. Ultimately, we retained 6173 data points. The 6173 data points were mainly obtained from experimental measurements of 79 ILs consisting of 35 cations and 18 anions at an extensive spectrum of temperatures and pressures from 243.2 to 453.15 K and 1–49990 kPa. The ILs nomenclatur, temperatures, pressures, and mole fraction of solubility were listed in Table S1. The cationic family consisting mainly of imidazolium, pyrrolidinium, pyridinium, phosphonium, and ammonium, and the anionic family consisting mainly of halides, sulfonates, sulfates, borates, and nitrates. Measurements may vary due to differences in experimental methods (isovolumic saturation method, synthetic vesicle method, magnetic levitation balance method, microbalance method) as well as the presence of impurities, and no stringent data filtering procedures were implemented other than limiting the singularity of data points for ILs, and the deletion of too many data points not only reduces the diversity of the data, but may also have a negative impact on the model(Glavatskikh et al., 2019). In addition, this dataset has been critically examined and widely used for QSPR modeling in previous studies(Jian et al., 2022; Liu et al., 2021; Zhang et al., 2021). Therefore, high quality data can hold promise for the next ML and molecular modeling.

2.2. COSMO-RS

2.2.1. COSMO-RS calculation principle

The COSMO-RS model, as a prior predictive molecular thermodynamic model independent of experimental data, is widely used to calculate various thermodynamic properties of single component and multi-component systems. The COSMO-RS model can need to know the surface charge density distribution (σ-profile) of specific molecules only to predict the thermodynamic properties of solvent systems. In the model, the liquid is considered as an almost tightly packed ideal solvent molecule whole, while the intermolecular interaction are regarded as charge shielding interaction, including van der Waals force interaction (E_{vdW}), Hydrogen Bond interaction (E_{HB}), Electrostatic interaction (E_{misfit}). In this way, the interactions between molecules can be calculated through a set of simple equations in statistical thermodynamics. Although the actual charge distribution on the molecular surface is complex, the COSMO-RS model divides the molecular surface into segments (σ) with fixed charge densities. The statistical distribution of product of σ and the possible range of molecular volume surface area is called σ-profile ($p_s(\sigma)$). The calculation formulas are as follows (Eckert and Klamt, 2002; Klamt, 1995; Klamt and Eckert, 2000; Lin and Sandler, 2002):

$$p_s(\sigma) = \sum_{i \in S} x_s^i p_i(\sigma) \tag{1}$$

where, x_s^i is the mole fraction of component i in system S.

E_{vdW}, E_{HB} and E_{misfit} are calculated according to formulas (2-4) respectively.

$$E_{vdW}(\sigma, \sigma') = a_{eff}\left(\tau_{vdw} + \tau'_{vdw}\right) \tag{2}$$

$$E_{HB}(\sigma, \sigma') = a_{eff} c_{HB} \min\{0; \min[0, \sigma_{donor} + \sigma_{HB}] \max\left[0, \sigma_{acceptor} - \sigma_{HB}\right]\} \tag{3}$$

$$E_{misfit}(\sigma, \sigma') = a_{eff} e_{misfit}(\sigma, \sigma') = a_{eff} \frac{a'}{2}(\sigma + \sigma')^2 \tag{4}$$

for the above equations, σ and σ' represent the net shielding charge density of two different segments, a_{eff} represents the effective contact area, and a' represents the energy factor; τ_{vdW} and τ'_{vdW} are parameters of E_{vdW}; The strength of hydrogen bonds is represented by the symbol c_{HB}; The hydrogen bond energy barrier is expressed as σ_{HB}; Hydrogen bond donors and acceptors are defined as: σ_{donor}=min[σ, σ'], $\sigma_{acceptor}$=max[σ, σ']. Therefore, the overall interaction energy is defined as:

$$E(\sigma, \sigma') = E_{vdW}(\sigma, \sigma') + E_{HB}(\sigma, \sigma') + E_{misfit}(\sigma, \sigma') \tag{5}$$

In addition, in system S, the chemical potential (μ) of the overall system and the chemical potential of specific components can be used to calculate the activity coefficients and other thermodynamic properties of the components. The μ calculation formula for the system is as follows:

$$\mu_s(\sigma) = \frac{RT}{a_{eff}} \ln\left[\int p_s(\sigma') \exp\left(\frac{1}{RT}\left(a_{eff}\mu_s(\sigma') - E_{misfit}(\sigma, \sigma') - E_{HB}(\sigma, \sigma')\right)\right) d\sigma'\right] \tag{6}$$

The chemical potential ${\mu_s}^{x_i}$ of component i in system S is related to the combined chemical potential ${\mu_{cs}}^{x_i}$ of the volume and surface area of the molecule. The calculation formula is as follows:

$${\mu_s}^{x_i} = {\mu_{cs}}^{x_i} + \int p_{x_i}(\sigma)\mu_s(\sigma) d\sigma \tag{7}$$

where, the activity coefficient (γ_s^i) of component i in the system can be calculated based on the chemical potential ${\mu_s}^{x_i}$:

$${\gamma_s}^i = \exp\left(\frac{\mu_s^i - \mu_i^i}{RT}\right) \tag{8}$$

If component i is a gas, the solubility calculation formula in system S is as follows, where p_i and p_i^0 represent the partial pressure and saturated vapor pressure of component i:

$$x_s^i = \frac{{p_i}^0 \times \gamma_s^i}{p_i} \tag{9}$$

It is worth noting that the anions and cations of ionic liquids are considered independent molecules in the solution. According to the COSMO-RS model, the CO_2 ionic liquid system is considered a pseudo ternary system consisting of CO_2, anions and cations. For $[A]^+[B]^-$ type ionic liquids, the formula for calculating the solubility $x_{co_2}^T$ of CO_2 in the pseudo ternary mixture is:

$$x_{co_2}^T = \frac{n_{co_2}}{n_{co_2} + n_{cation} + n_{anion}} = \frac{n_{co_2}}{n_{co_2} + 2n_{IL}} \tag{10}$$

In the formula, n_{IL}, n_{cation}, n_{anion} and n_{co2} are the moles of ionic liquid, cation, anion, and CO_2 in the liquid phase, respectively. For a real binary

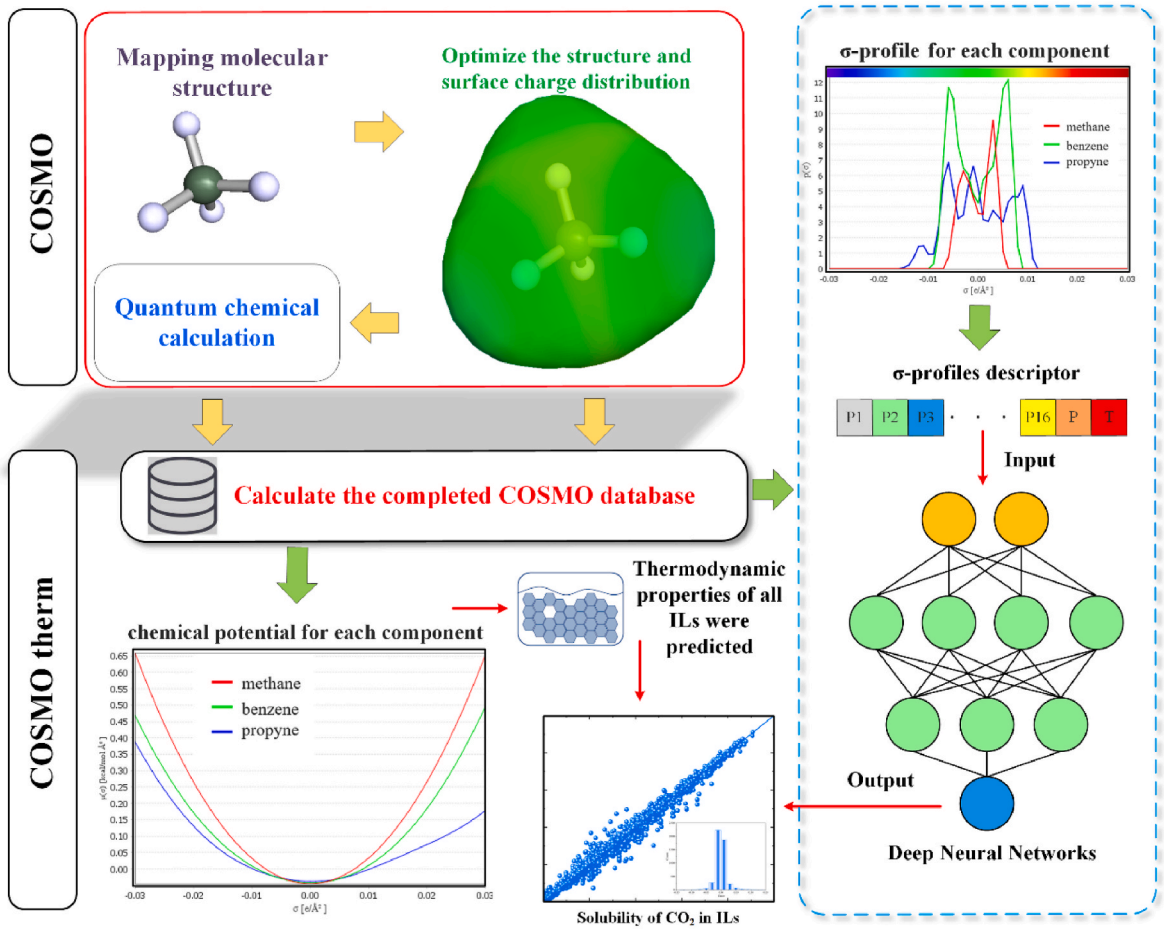

Fig. 1. Flowchart of QSPR model development for correlating CO_2 solubility by QC calculation of characterization information of ILs structure.

CO_2 - IL system, the solubility $x^B_{co_2}$ of CO_2 in the ionic liquid can be calculated using the following formula:

$$x^B_{co_2} = \frac{n_{co_2}}{n_{co_2} + n_{IL}} \tag{11}$$

By combining formulas 10 and 11, the conversion relationship between the real binary CO_2- ionic liquid system and the pseudo ternary system can be obtained:

$$x^B_{co_2} = \frac{2x^T_{co_2}}{x^T_{co_2} + 1} \tag{12}$$

2.2.2. COSMO-RS calculation process

In this study, COSMOthermX (version 2022) software was used to calculate the CO_2 solubility in ILs based on the COSMO-RS model. The computational procedure is shown in Fig. 1. Firstly, QC calculations were performed for different ions using TURBOMOLE. In which the triple zeta valence polarized basis set (TZVP) and Becke-Perdew (BP) function were used to complete the calculation of molecular energy. Following the COSMO calculations, a "COSMO" file was produced for each molecule and stored in the COSMO database. Then the "COSMO" file was loaded into COSMOthermX software to obtain the σ-profile of the molecule. In this study, all calculations were carried out on a desktop PC with Intel i9-13900HX 16-Core 2.20 GHz CPU and 16 GB RAM. More calculation and analysis information about COSMO-RS can be learnt in the software reference manual or in the website (https://www.scm.com/doc/Tutorials/COSMO-RS/Ionic_Liquids.html). All ILs in this study are considered as a combination of anionic and cationic pairs. Finally, the solubility of CO_2 in ionic liquids collected under the same conditions in the literature was calculated. In addition, the σ-profile of each molecule is input as the structural information of the ionic liquid into the DNN model for predicting the solubility of CO_2 in ILs. This process requires dividing the σ-profile information of each molecule to obtain molecular descriptors that can distinguish different ionic liquids. The specific details of the division will be introduced in the next section.

2.3. Division of descriptors

One of the key steps in developing a QSPR model is to quantify the structural information of the studied compounds, called molecular descriptors. Here, the molecular structure of ILs was described as two groups (anion (1), cation (2)). The molecular σ-profile was determined from the results of QC calculations, and the anionic and cationic σ-profiles and 3D screening of charge distribution (σ-surface) for the entire dataset were presented in the supporting information (SI) Table S4 and Table S5. The σ-profiles of HBD and HBA ions describe the probability distribution of molecular surface fragments under different screening charge densities σ (e/Å^2). For standardized nomenclature and ease of modeling, we directly read the ordinate of the σ-profile at predefined σ sampling points (consistent with the segmentation positions in the figure), and denoted the i-th sampling value as P(σ,i) (Å^2) where $i = 1-16$, abbreviated as $P_{\sigma\text{-profile}}$ descriptor. This descriptor quantitatively characterizes the relevant information on the atom/functional group type and relative abundance at a specific σ position and can be directly used as input for machine learning. The good or bad descriptor division plays a very significantly impacts in the performance of the model, Figs. S1 and S2 show the distribution of $P_{\sigma\text{-profile}}$ for the whole dataset among all cations and anions in the dataset, respectively. In the graph it

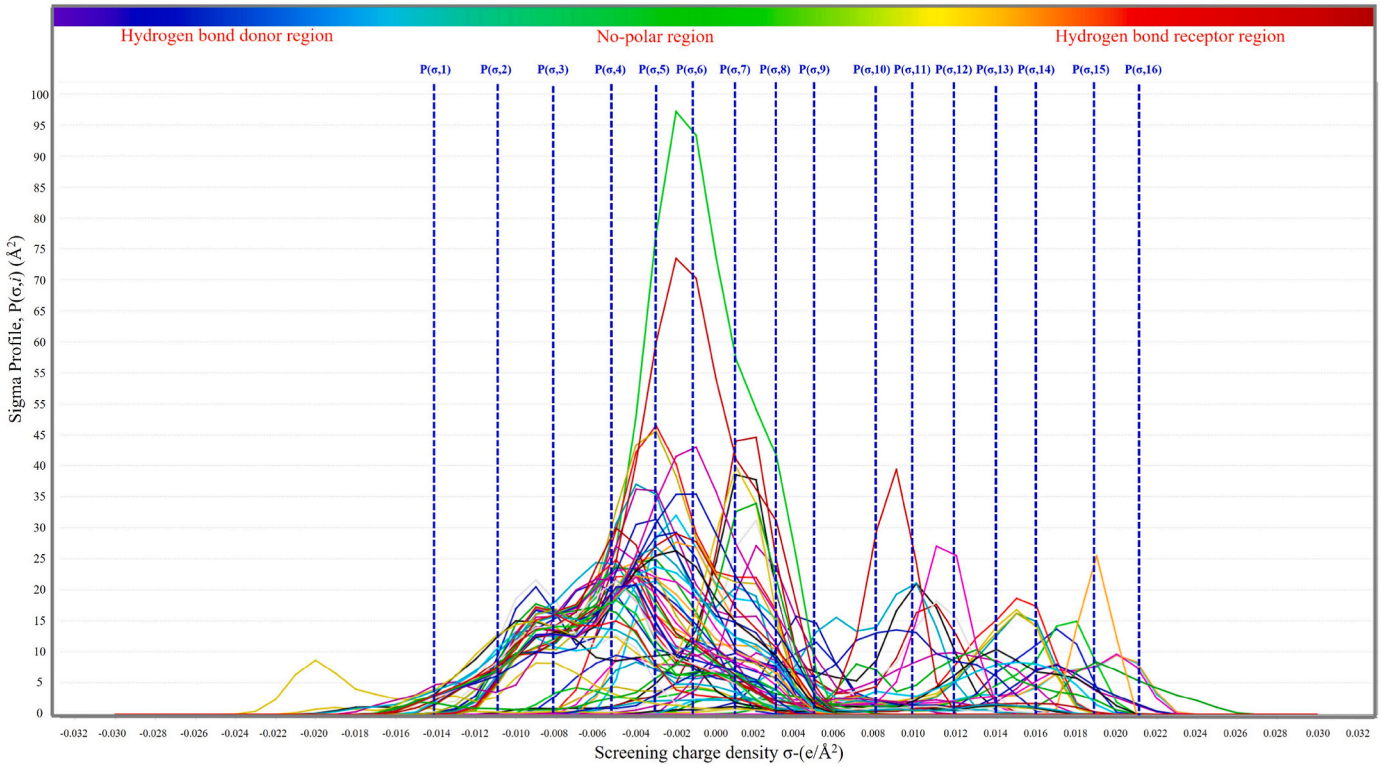

Fig. 2. Distribution of σ-profiles and division of $P_{σ\text{-profile}}$ descriptors for all ILs in the dataset.

is seen that the $P_{σ\text{-profile}}$ distribution of all ions ranged from −0.03 to 0.03 e/Å^2. While the differences in $P_{σ\text{-profile}}$ of different cations were mainly concentrated between −0.025 e/Å^2~0.006 e/Å^2, the differences in $P_{σ\text{-profile}}$ of different anions were mainly concentrated between −0.007 e/Å^2~0.025 e/Å^2. To effectively differentiate between various ILs using $P_{σ\text{-profile}}$ descriptors, it was decided to divide the dense where the difference in $P_{σ\text{-profile}}$ values was large and the sparse where the difference was small. The whole ionic $P_{σ\text{-profile}}$ descriptor division was shown in Fig. 2. 16 descriptors were divided according to the degree of difference in $P_{σ\text{-profile}}$ values, and Table S2 and Table S3 contain the specific values of the sixteen $P_{σ\text{-profile}}$ descriptors divided for different anions and cations. Temperature and pressure as additional environmental descriptors yielded a total of 18 descriptors.

In addition, we calculated the Pearson correlation coefficients between 18 different descriptors and CO_2 solubility in ILs, and presented the results in Fig. 3(a). From the graph, it can be seen that 18 descriptors exhibit different degrees of correlation with CO_2 solubility in ILs. Among them, the pressure parameter showing the strongest correlation.Pearson's coefficient can only identify linear relationships and cannot effectively analyse non-linear or monotonic relationships, so Spearman's rank correlation coefficient was added to detect monotonic patterns among variables for a more comprehensive understanding of feature-target relationships. Compared to the original Pearson correlation analysis (Fig. 3(a)), Spearman's correlation has a significant advantage in capturing variables that are nonlinear but maintain a monotonic trend. Specifically, P already exhibits a strong linear correlation in Pearson's analysis ($r = 0.6576$), and its rank correlation further rises to $\rho = 0.8766$ in Spearman's analysis, suggesting that this feature is more robustly associated with the target variable not only at the linear level, but also at the overall ordinal level. On the contrary, P(σ,7) exhibits a moderately positive correlation in Pearson's ($r = 0.3261$), but its rank correlation drops slightly to $\rho = 0.3003$, suggesting that the variable may be affected by a few extreme values or have non-monotonic fluctuations in local intervals. In addition, the descriptors of P(σ,5), P(σ,6), P(σ,7) and P(σ,8) rank high in Pearson correlation coefficient values with CO_2 solubility in ILs among the 16 $P_{σ\text{-profile}}$ descriptors. From Fig. 2, it can be seen that the $P_{σ\text{-profile}}$ values contained in these four descriptors differ significantly. Therefore, dividing descriptors in the areas of $P_{σ\text{-profile}}$ values which have significant differences, can obtain more descriptors strongly related to CO_2 solubility in ILs.

2.4. DNN architecture and input features

Classification, regression and prediction were desirable applications of DNN(Wang et al., 2019; Wen et al., 2022). DNN is a multi-layer supervised neural network that learns through supervised training, utilizing the previous layer's output features as the input for the subsequent layer. After a layer-by-layer feature process to enhance the representation of input data features(Bürkle et al., 2021). Since DNN have multiple nonlinearly mapped feature transformations, highly complex functions can be fitted. The final output of the selected target value is achieved by nonlinear transfer between different network layers. DNN has demonstrated its strong predictive capabilities and flexibility in various fields of application.

The training of a model is an iterative process, where the weights and biases within the network are adjusted using the backpropagation algorithm. Gradient Descent is employed to minimize the loss function. To prevent overfitting, various regularization techniques are commonly used in DNN models, such as L2 regularization (weight decay), Dropout (randomly dropping neurons), and Early Stopping. Optimization algorithms like Adam, RMSprop, and Adagrad are frequently utilized to accelerate the training process and enhance the model's convergence. The performance of a DNN heavily depends on hyperparameters, such as the learning rate, number of layers, number of neurons per layer, and choice of activation functions. Hyperparameters are typically optimized through methods like cross-validation or grid search. DNN has demonstrated its strong predictive capabilities and flexibility in various fields of application(Ashraf et al., 2021; Bai et al., 2006; Grądziel et al., 2018; Muhammad Ashraf et al., 2020; Skrobek et al., 2023).

In this work, DNN model was performed using the scikit-learn (1.2.2)

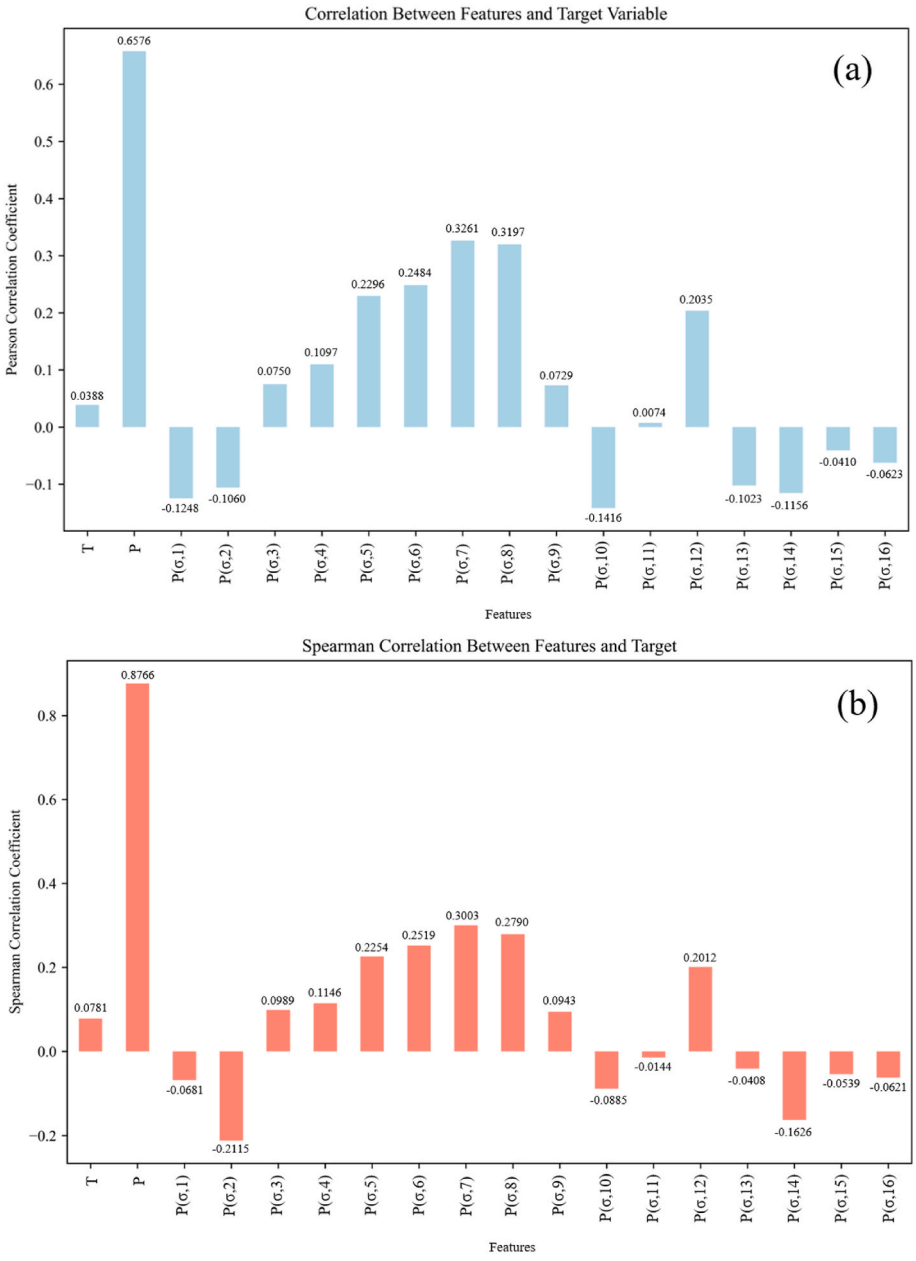

Fig. 3. Histogram of (a) Pearson correlation coefficient and (b) Spearman rank correlation coefficient between 18 descriptors and CO_2 solubility in ionic liquids.

tool in Python (3.8). Sixteen $P_{\sigma\text{-profile}}$ descriptors obtained by dividing the σ-profiles of ionic liquid molecules are used, and two environmental descriptors, temperature and pressure, are attached as inputs. The predicted CO_2 solubility in the ILs was obtained as the output, and the prediction correlations were defined equation (13) as follows(Torrecilla et al., 2010):

$$S = f\left(P1_{\sigma-profile}, P2_{\sigma-profile}, \ldots\ldots, P16_{\sigma-profile}, T, P\right) \tag{13}$$

In the above equation, 'T' is the temperature descriptor (K), 'P' is the pressure descriptor (kPa), 'S' denotes the molar fraction of CO_2 solubility in ILs. The neurons and hidden layers in the neural network are given by the following equation (14)(Adeyemi et al., 2018):

$$H_j = f\left(\sum_{i=1}^{N}\left(W_{j,n}\right)\left(pi_{\sigma-profile}\right) + b_j\right) \tag{14}$$

where '$W_{j,n}$' represents the weights connecting input and the hidden layers, 'j' is the hidden neuron, '$Pi_{\sigma\text{-}profile}$' is the ith σ-profile descriptor, 'b_j' is the bias of neuron 'j' in the hidden layer, and 'f' is the activation function of the neuron.

The SHAP method solves the problems of "lack of global consistency" and "dependency on model structure" in traditional model interpretation. Its calculated SHAP value can simultaneously reflect the local influence of individual features (single sample prediction interpretation) and global importance (cross sample feature ranking) (Liu et al., 2025). This study directly correlates the "model prediction results" with the "structural feature contribution" through this method, transforming the basis for artificial intelligence candidate selection from "black box output" to "quantifiable feature contribution"(Arrieta et al., 2019), which perfectly fits the core goal of eXplainable Artificial Intelligence (XAI) (Bonito et al., 2024) and provides methodological support for the reliability of research conclusions. XAI details can be found in the *3.7 Model Interpretation* section.

Table 1
Hyperparameters of DNN model determined by grid search method based on two dataset splitting schemes.

Hyperparameters	Search space	Results (Point-based)	Results (Component-based)
Activation function	sigmoid, relu, tanh	relu	sigmoid
Optimizer	lbfgs, sgd, adam	lbfgs	adam
Alpha	0.0001, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1	0.005	0.005
Hidden_layer_sizes	(16), (32), (64), (32, 16), (64, 32), (64, 16), (64, 32, 16), (64, 32, 32), (64, 16, 16)	(64, 32, 32)	(64, 16, 16)

2.5. Model training and optimization

In order to fit the best model, it was necessary to optimize the hyperparameters of the DNN model. Hyperparameter values directly influence model performance, and improper selection can lead to significant performance degradation. In order to obtain the optimal hyperparameters, we used grid search to tune the hyperparameters of the model. Grid search was a standard method that attempts to find the optimal values by searching for parameter combinations by specifying a subset of the parameter optimization space within the target algorithm. To avoid test set information leakage, grid search was combined with 5-fold cross-validation. This approach ensures that hyperparameter tuning is performed only on the training and validation sets, and that unseen test sets are only used for model performance evaluation. Table 1 shows the four important hyperparameters and their search space. These hyperparameter combinations totaled 648 candidate combinations, each of which needed to be trained in a 5-fold cross-validation, and a total of 3240 model fits were completed. In this way, the optimal hyperparameters were determined for the DNN model of the two data segmentation schemes. The optimal combination of hyperparameters under these two data segmentation schemes is also different due to the different sample compositions of the training set and the validation set.

2.6. Evaluation metrics and validation strategy

Typically, prior to training an ML model, a part of the data is set aside as a test set to evaluate the model's generalization ability, while the remaining data is utilized for model training. This study adopts two data segmentation schemes for data points to explore their impact on ML modeling: (1) "Point based", randomly dividing the entire dataset into a training set and a testing set, with the training set accounting for 80 % and the other 20 % being the testing set, which is commonly used for ML modeling. (2) "Component based", dividing the entire dataset into two parts, set1 and set2, based on the type of ionic liquid, with set1 accounting for 80 % of the total data. In this study, the DNN model under the "Point-based" data segmentation scheme is named P-DNN, and the DNN model under the "Component based" segmentation scheme is named C-DNN. To make the extrapolation setting explicit, we provide compact split statistics in the SI Tables S6–S8 summarize (i) the allocation of IL pairs and the proportion of high-pressure records ($P > 5000$ kPa), and (ii) the distribution of major cation and anion families between the training and test sets. These summaries confirm that, by construction, no IL type appears in both sets and several families occur only in the test set, reflecting a genuine extrapolation scenario for C-DNN. In addition, we also adopted a 5-fold cross validation strategy, which is commonly used to evaluate the performance of DNNs and optimize hyperparameters.

In terms of statistical metrics, RMSE, MAE and R^2 were chosen as the indicators to evaluation the model performance. Among them, the closer the values of MAE and RMSE are to 0, the smaller the error between the predicted and experimental values. R^2 is a decimal number with no unit between 0 and 1. When this value is equal to 1, the predicted and experimental values will be fixed on a straight line with no dispersion, which will be a perfect fitting model. Formulas (15)-(17) for the statistical indicators are given below(Wang et al., 2021):

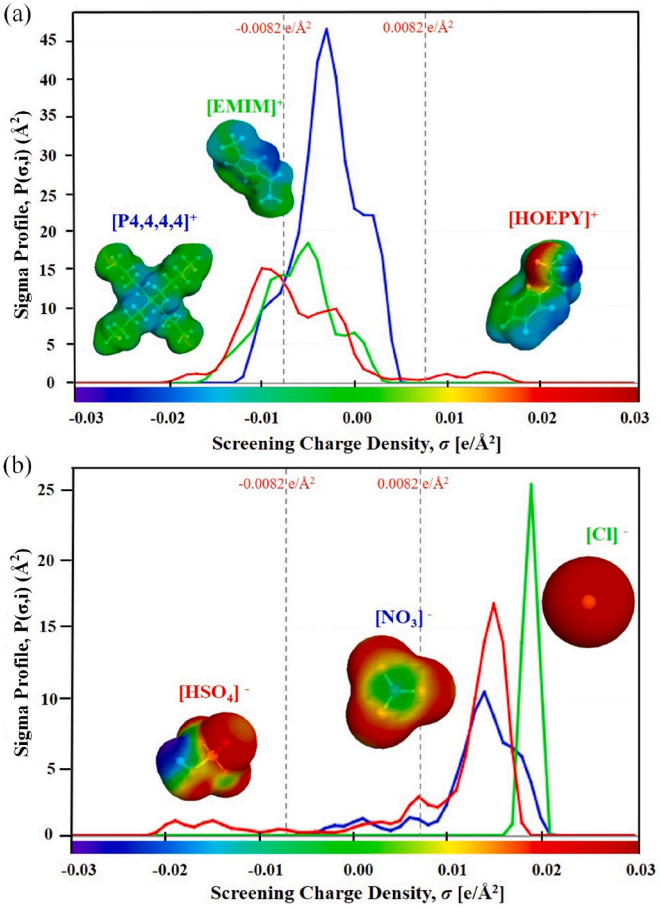

Fig. 4. Representative (a) cation, (b) anion σ-profile and σ-surface in this work.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(s_{\exp} - s_{pred}\right)^2}{n}} \tag{15}$$

$$MAE = \frac{\sum_{i=1}^{n}\left|s_{\exp} - s_{pred}\right|}{n} \tag{16}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(s_{\exp} - s_{pred}\right)^2}{\sum_{i=1}^{n}\left(s_{\exp} - \bar{s}\right)^2} \tag{17}$$

where '$s_{\exp}$', 's_{pred}', and '$\bar{s}$' are the experimental, predicted and mean values of CO_2 solubility in ILs, respectively. n denotes the number of all data points.

3. Results and discussion

3.1. Analysis of σ-profile

The σ-profile curve helps identify the concentration of an atom within a molecule and the polarity of the molecule as a whole. The concentration of the atom is determined by the peak height of the σ-profile curve, while the polarity of the molecule is determined by the position of the peak(Torrecilla et al., 2010). Based on QC methods, the σ-profile of a molecule contains crucial chemical information for predicting its interactions in a fluid. We consider the surface polarization

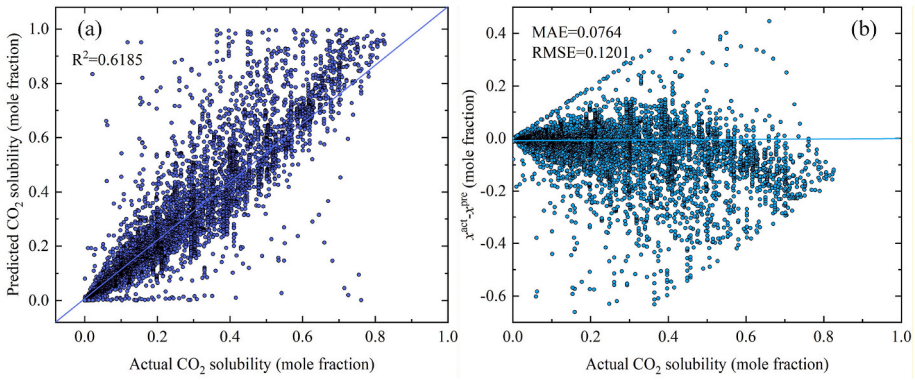

Fig. 5. (a) parity plot and (b) absolute deviation plot of predicted and experimental values based on COSMO-RS model.

charge density and representative σ-profile plots for anions and cations. The σ-profile plot can be divided into three main regions: the HBD region ($P_{\sigma\text{-profile}} < -0.0082$ e/Å^2), the No-polar (N-P) region (-0.0082 e/Å$^2 < P_{\sigma\text{-profile}} < 0.0082$ e/Å^2) and the hydrogen bond acceptor (HBA) region (0.0082 e/Å$^2 < P_{\sigma\text{-profile}}$) (Jiao et al., 2022). The sixty-one averages between −0.03 and 0.03 in this work were reduced to sixteen data points based on the density of the distribution, which included the chemical information needed for model inputs. The polarized charge densities and σ-profile curves of three representative cations of 1-(2-hydroxyethyl) pyridinium ($[HOEPY]^+$), 1-ethyl-3-methyl-imidazolium ($[EMIM]^+$), and tetrabutylphosphonium ($[P,4,4,4]^+$) were shown in Fig. 4(a). From Fig. 4(a), it's evident that the majority of cation charge density peaks are concentrated in both the HBD and N-P regions. Due to the neutral surface of the -CH3, -CH2 and -CH groups in the molecule, these cations show relatively high peaks in the N-P region and the height of the peaks increases with the increase in the number of these groups. It is noteworthy that $[HOEPY]^+$ contains -OH groups that require hydrogen bonding, and thus the corresponding peaks appear in the HBA region. The σ-surface and σ-profile curves of three representative anions of hydrogen sulfate ($[HSO_4]^-$), nitrate ($[NO_3]^-$) and chloride ($[Cl]^-$) were shown in Fig. 4(b). In Fig. 4(b), it is seen that most of the charge peaks of the anions were located in the HBA region, which suggests that the negatively charged surfaces of the anions may form hydrogen bonds with other HBDs. $[Cl]^-$ has a very positive surface polarization charge density, and in addition the H atom in $[HSO_4]^-$ shows a blue color, indicating that this atom was prone to break and provide hydrogen bonding. In summary, the $P_{\sigma\text{-profile}}$ provides a stable representation and description of the molecular structure and can therefore be used as an input for the development of QSPR models for correlating CO_2 solubility in different ILs.

3.2. COSMO-RS model evaluation

The COSMO-RS model is a promising tool for both calculating solvent thermodynamic properties and screening solvents. In earlier work, the effectiveness of COSMO-RS in solubility prediction of cellulose, petroleum asphaltene, and polymers has been demonstrated in the literature and has been widely used for solubility prediction of different gases in various solvents(Chu and He, 2019; Chu et al., 2018; Rashid et al., 2019). Therefore, we used COSMO-RS to predict CO_2 solubility in ILs. In this work, the structural information of the solvent (ILs) and the solute (CO_2) was entered into the COSMOtherm software to calculate the CO_2 solubility within the ILs at a specified temperature and pressure. Since ILs were mixtures of anions and cations, it was common practice to specify the molar concentrations of anions and cations to be half of the mixture(Mohan et al., 2022), and we have adopted the same approach. In order to be able to run the COSMO-RS software successfully, we obtained the structural information from the ILs collected in the literature by QC calculations, and then simulated the calculations using the same experimental conditions (temperature and pressure). Fig. 5(a) shows the results of the comparison between the predicted solubility values of COSMO-RS and the experimental values of CO_2 solubility, and Table S1 shows the specific prediction results of the COSMO-RS model. If the data point was closer to the diagonal slash, it indicates that the COSMO-RS model predicts the data point closer to the experimental value. From Fig. 4(a), it could be seen that when the molar fraction of solubility was less than 0.4, the COSMO-RS model produced reasonably reliable predictions, with most data points closely aligned with the diagonal line. However, when the molar fraction of solubility value was greater than 0.4, the predicted values of most data points were greater than the actual values. By looking at the solubility these high solubility data points were found to be measured at higher pressures. In addition, Among the 6173 data points collected, the COSMO-RS model could not calculate 540 solubility data values, and it was observed that the pressure of these data points was generally greater than 100 bar and the experimental value of solubility was greater than 0.75. The reason for this phenomenon is that the COSMO-RS model overestimates the solubility value at high pressure and this situation would be more and more serious when the pressure was higher. The prediction error of the COSMO-RS model could be seen through Fig. 5 (b) where the value of MAE was 0.0764 and RMSE was 0.1201. Most of the predictions were on the high side and this error was more pronounced when the mole fraction of the experimental values was greater than 0.4.

Generally, the activity coefficient can be used to calculate the solubility of gas at low pressure. As introduced in Section 2.2.1, the activity coefficient can be derived from the COSMO-RS model, enabling the acquisition of relatively accurate gas solubility calculations under low pressure. For the calculation of gas solubility under high pressure, to obtain more accurate calculation results, it is necessary to consider the irrational characteristics of gas, that is, the fugacity coefficient of gas. However, COSMO-RS assumes the incompressible characteristics of the liquid and the ideal characteristics of the gas, and only calculates the activity coefficient independent of pressure, without considering the gas fugacity coefficient. Therefore, it causes a situation of relatively large prediction error under high pressure (Pelaquim et al., 2024), which is the current limitation of this model. To improve prediction results at high pressures, Moity et al. (2012) highlighted the need to combine COSMO-RS with equations of state. For more details, Leonhard et al. (2009) report an example of a combination of COSMO-RS and equations of state. In addition, future research can address these limitations by combining models of non-ideal gas behavior, introducing modifications to fugitivity coefficients, adjusting model parameters, combining more experimental data, and developing new theoretical frameworks. This will help improve the prediction accuracy of the model under high-pressure conditions.

Table 2
R^2, RMSE and MAE performance metrics of DNN models according to two dataset splitting schemes.

	P-DNN		C-DNN	
	Training	Test	Training	Test
R^2	0.9952	0.9904	0.9637	0.9297
RMSE	0.0151	0.0216	0.0406	0.0631
MAE	0.0085	0.0133	0.0296	0.0450

3.3. DNN model evaluation

After completing the hyperparameter optimization, the model was trained using the best hyperparameters. Table 2 records the R^2, RMSE and MAE of the training and test sets in the DNN model. It can be seen that the R^2, RMSE and MAE obtained by the P-DNN model on the test set are close to those obtained on the training set. The R^2, RMSE and MAE on the test set were 0.9904, 0.0216 and 0.0133, respectively. However, there are some differences between the R^2, RMSE and MAE obtained by the C-DNN model on the test set and the R^2, RMSE and MAE obtained on the training set. The R^2, RMSE and MAE on the test set are 0.9297, 0.0631 and 0.0450, respectively. By comparing the performance indicators of the two models, it is found that the "point-based" data segmentation scheme has better prediction effect, which shows the impact of the data segmentation scheme on the performance of the model. This gap is expected because the "component-based" data segmentation scheme enforces compositional extrapolation, see Table S6–S8 for dataset composition and high-pressure coverage. Fig. 6 shows the scatter plots and relative error distribution frequency plot of the predicted CO_2 solubility and the actual CO_2 solubility of the DNN models of the two dataset segmentation schemes, respectively. Fig. 6(a) show the prediction results based on the training and test sets of the P-DNN model. It was seen that the predicted values of the model and the experimental values showed a good agreement, and most of the data points were distributed around the Y = X line. Fig. 6(c) show the prediction results of the training and test sets of the C-DNN mode, and it can be seen that the prediction results of the C-DNN model are more divergent than the actual results compared with the P-DNN model. In addition, Fig. 6(b) and (d) show the relative error distribution frequency plots of the two data segmentation schemes. From the relative error distribution frequency diagram, it can be seen that the prediction errors of the two data segmentation schemes are normally distributed. The prediction error of the P-DNN model is mainly concentrated between ±0.1, while the prediction error of the C-DNN model is larger, and the error distribution is mainly concentrated between ±0.2. The above results show the significant impact of different data set segmentation schemes on ML modeling. The prediction effect of the "Point-based" data segmentation scheme is better than that of the "component-based" data segmentation scheme. In fact, the model performance of the "point-based" data segmentation scheme is overestimated. Specifically, the model using this data splitting scheme demonstrates high accuracy in predicting the CO_2 solubility of ILs within the known data range, but this accuracy is limited to the known dataset. For IL categories outside the dataset, the prediction accuracy of CO_2 solubility may not reach the same level. This is mainly due to information leakage caused by the presence of duplicate IL categories in both the training and testing sets. Because the "Point-based" data splitting scheme is random, and the dataset in this study contains more than 20 data points for each category of ILs, a majority of ILs of the same category end up in both the training and testing sets. This information leakage leads to an inflated prediction performance within the known data range, masking the model's lack of generalization ability. When faced with entirely new IL categories, the model, lacking corresponding training data, struggles to accurately capture the variation in CO_2 solubility, resulting in a decline in prediction accuracy. Although the model with the "Component-based" data splitting scheme may show overall lower performance, it avoids the issue of information

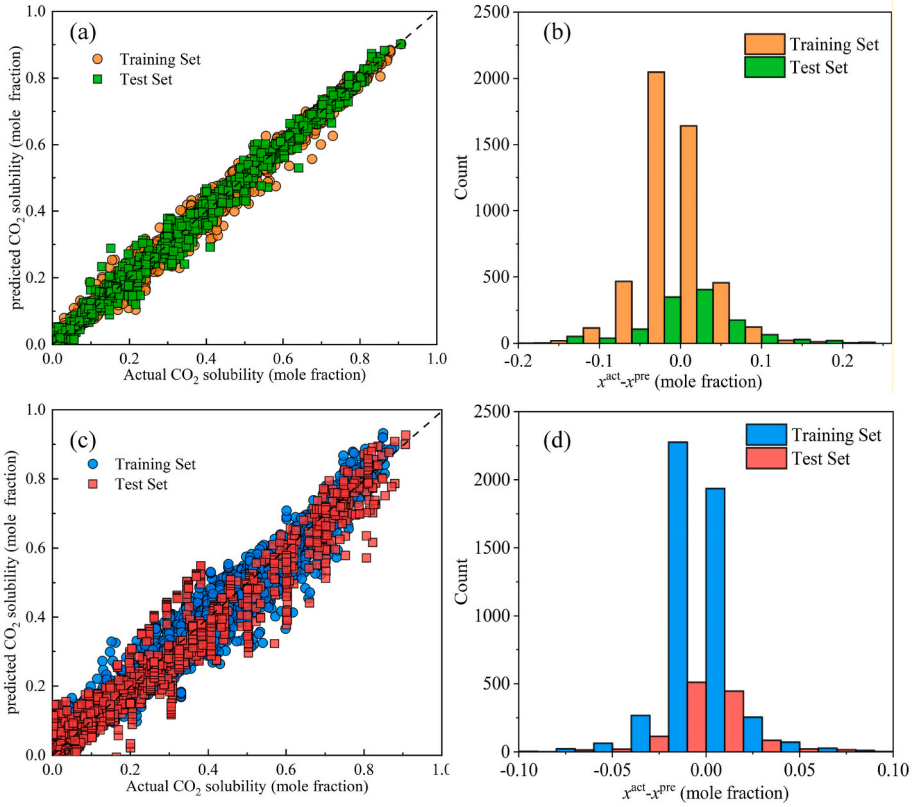

Fig. 6. (a) Scatter plot and (b) of CO_2 solubility predicted by P-DNN model and actual CO_2 solubility, (c) Scatter plot and (d) relative error distribution frequency plot of CO_2 solubility predicted by C-DNN model and actual CO_2 solubility.

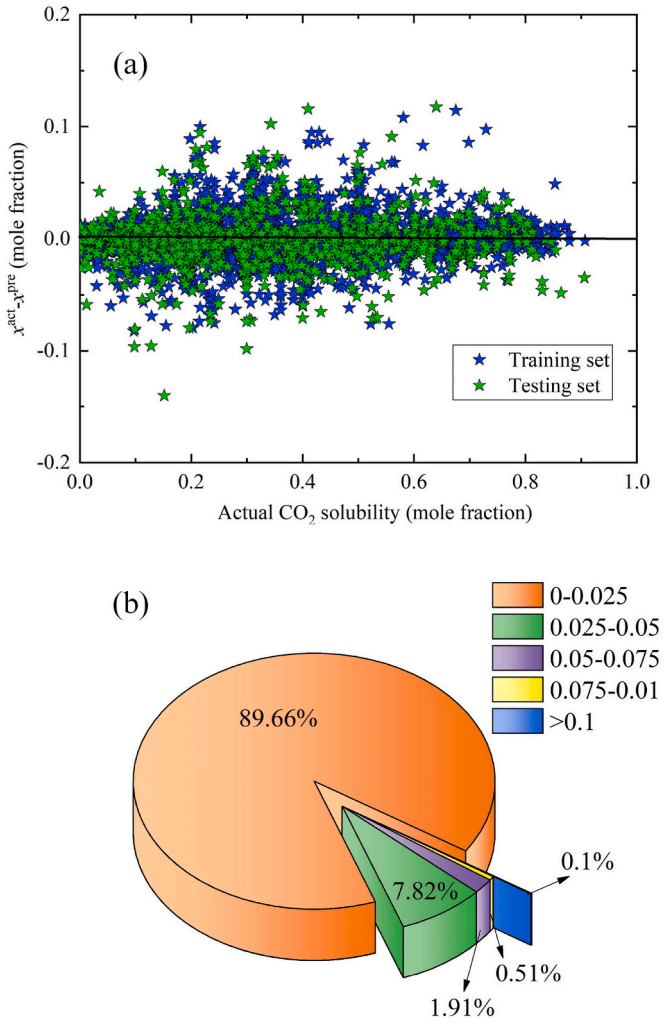

Fig. 7. (a) relative deviation and (b) percentage of different absolute deviations of the predicted and actual values of the P-DNN model for the point-based dataset splitting scheme.

leakage, and its performance on external data is more consistent with its actual ability. Therefore, to further demonstrate that the "Component-based" data splitting scheme can accurately reflect real-world prediction scenarios, it is necessary to validate the external data prediction capabilities of models from both data splitting schemes to assess their ability to predict CO_2 solubility in unseen ILs, thereby showcasing the model's generalization ability.

Compared to the COSMO-RS model, the DNN models for both data splitting schemes demonstrated more reliable predictions. Previous studies have also highlighted the differences between COSMO-RS and other DNN models (Mohan et al., 2023b; Valeh-e-Sheyda et al., 2022). It also shows accurate prediction results under high temperature and high pressure, which effectively solves the limitations of the COSMO-RS model. We also statistically analyzed the distribution of relative deviations and the proportion of different absolute deviations of the P-DNN mode. Statistical analysis of Fig. 7(a) shows that most of the absolute deviations of CO_2 solubility are within 0.1, and only individual relative deviations are slightly higher than 0.1. Furthermore, Fig. 7(a) illustrates that when the solubility value was less than 0.2 or more than 0.6, the error between the model predicted value and the measured value is much smaller, which indicates that the P-DNN model predicts the results more accurately in this solubility value range. Fig. 7(b) shows a high percentage of 89.66 % for absolute deviations within the range of 0–0.025, 7.82 % for absolute deviations within the range of 0.025–0.05, and only 0.1 % for absolute deviations greater than 0.1. These results show that the model can accurately predict the solubility of CO_2 in ILs.

3.4. Validation of external datasets

To validate that the "Component-based" data splitting scheme accurately reflects real-world prediction scenarios, this study evaluates the prediction accuracy of DNN models on external data points under two different data partitioning schemes. The data selected for this study comes from previous experimental results and covers five ILs: 1-(2-hydroxyethyl)-3-methyl-imidazolium tetrafluoroborate ([HEMIM][BF_4]) (Shokouhi et al., 2010), 1-ethyl-3-methyl-imidazolium ethyl sulfate ([EMIM][$EtSO_4$]) (Soriano et al., 2009), 1-butyl-3-methyl-imidazolium thiocyanate ([BMIM][SCN]) (Revelli et al., 2010), 1-ethyl-3-methyl-imidazolium trifluoroacetate ([EMIM][TFA]) (Shiflett and Yokozeki, 2009), and trihexyl(tetradecyl)phosphonium dicyanamide ([$P_{6,6,6,14}$][DCA]) (Ramdin et al., 2013). The data of these five ILs were not used in the training and testing process of the DNN model, thus serving as unbiased external evaluation data. Fig. 8 and Fig. S3 show the distribution of relative prediction errors for CO_2 solubility in the five external IL categories by the P-DNN and C-DNN models. Fig. 8(a) displays the relative prediction error distribution for CO_2 solubility in [HEMIM][BF_4]. The P-DNN model's prediction error ranges from −0.0241 to 0.0077. The small relative error indicates that the model has good prediction accuracy for this IL category. Similarly, the prediction errors for [EMIM][$EtSO_4$], [BMIM][SCN], and [EMIM][TFA] are within the range of −0.08 to 0.036, showing low prediction errors. It is noteworthy that the prediction error for [$P_{6,6,6,14}$][DCA] ranges from −0.188 to 0.093, which is higher than the prediction errors for the other four ILs. This is mainly because the cation in [$P_{6,6,6,14}$][DCA] has strong non-polarity, with values as high as 75.49, 93.38, and 57.03 in P5, P6, and P7, respectively. This promotes the model's prediction results, causing the model's predicted values to be higher than the actual values. Additionally, the training data lacks such highly non-polar ILs, which prevents the model from fully learning the relevant information, resulting in larger prediction errors for [$P_{6,6,6,14}$][DCA]. Fig. 8(e) shows the scatter plot of the model's predicted values against the actual values for these five ILs. Except for [$P_{6,6,6,14}$][DCA], the data points for the other four ILs are close to the Y = X line, with RMSE and MAE values of 0.0522 and 0.0307, respectively. Fig. S3(e) displays that the RMSE and MAE values for the C-DNN model are 0.0622 and 0.0465, respectively. A comparison shows that the prediction performance of the P-DNN and C-DNN models on the external dataset is quite similar. From the perspective of prediction accuracy, the C-DNN model's performance on the external dataset is more aligned with the actual situation, while the P-DNN model's prediction performance is higher than the real performance of the external dataset. This phenomenon aligns with the conclusion that the P-DNN model's performance is overestimated. In fact, the prediction performance exhibited by the P-DNN model is primarily based on the CO_2 solubility in IL solutions of categories covered in the dataset under different temperatures, pressures, and concentrations. Functionally, it is similar to some classical thermodynamic models, which can predict the CO_2 solubility of specific ILs with high precision. In contrast, the performance of the model using the "Component-based" data splitting scheme is mainly reflected in its predictions for CO_2 solubility in unknown ILs. Both models under different data splitting schemes have their own characteristics and can be selected for targeted use depending on the application scenario and prediction needs.

3.5. Comparison with literature reported models

The model developed in this study shows excellent performance in predicting the solubility of CO_2 in ionic liquids, and it is meaningful to compare the model performance with other models reported in the literature. Table 3 summarizes ML results, reporting performance

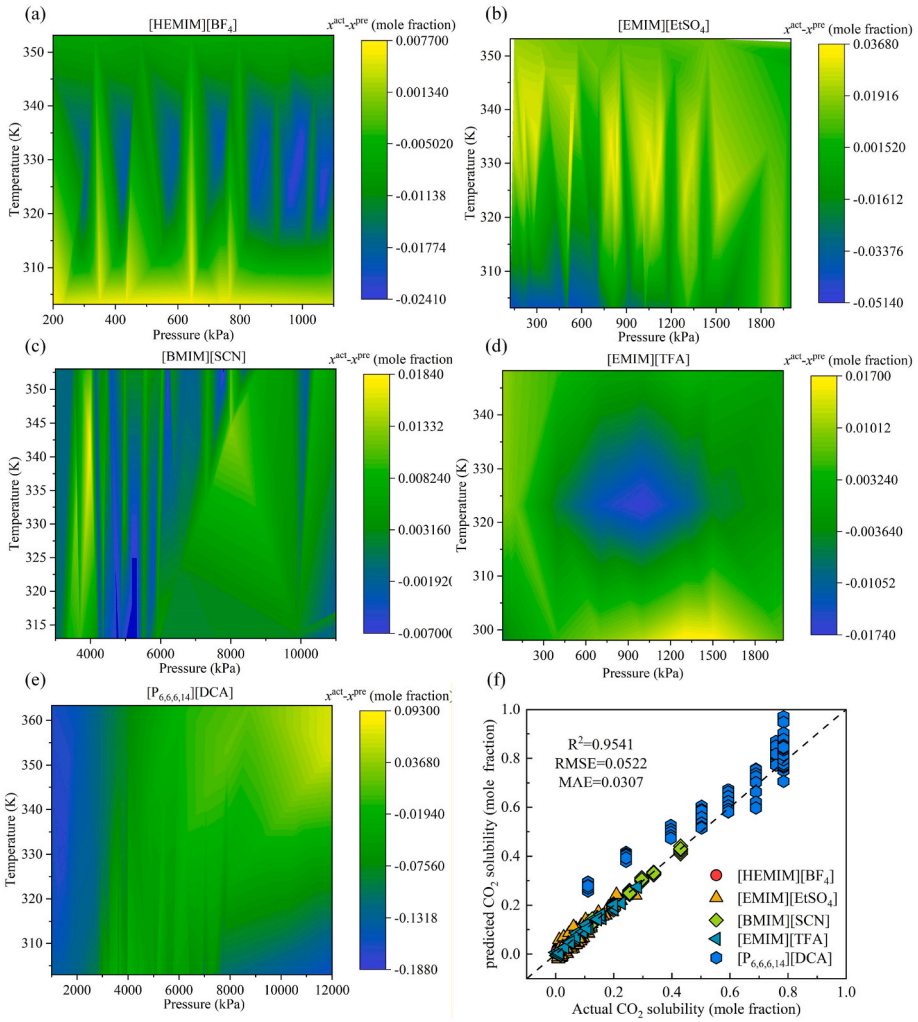

Fig. 8. (a) [HEMIM] [BF_4], (b) [EMIM] [$EtSO_4$], (c) [BMIM] [SCN], (d) [EMIM] [TFA], and (e) [$P_{6,6,6,14}$] [DCA] show two-dimensional contour plots of the relative error between the actual values of CO_2 solubility and the predicted values of the P-DNN model as a function of temperature and pressure, and (f) scatter plots of the predicted and actual values of the P-DNN model for five ILs.

Table 3
Comparison of performance of different models for predicting CO_2 solubility in ILs.

Model	Number of ILs	Number of features	Data points	Testing set			Ref.
				R^2	RMSE	MAE	
ANFIS	5368	5	67	0.9135	0.0726	–	Baghban et al. (2017)
MLP-ANN	5368		67	0.9694	0.0432	–	
FFNN	4397	26	43	0.9458	0.0759	–	Valeh-e-Sheyda et al. (2022)
RBFNN	4397		43	0.8949	0.0818	–	
LSTM	10,116	53	124	0.985	0.0293	0.0174	Ali et al. (2024)
ANN	10,116		124	0.986	0.0273	0.0171	
RF (Point-based)	10,848	13	185	0.96	0.05	0.03	Venkatraman and Alsberg (2017)
RF (Component-based)	10,848		185	0.85	0.01	0.06	
C-DNN	6173	18	79	0.9297	0.0631	0.0450	This work
P-DNN	6173		79	0.9904	0.0216	0.0133	

metrics together with training size and feature counts. Since ANFIS, MLP-ANN, FFNN, and RBFNN (Baghban et al., 2017; Valeh-e-Sheyda et al., 2022) were trained/evaluated on random splits, we benchmark them against our DNN (Point-based) to maintain a like-for-like in-distribution comparison. By comparison, it can be found that the DNN (Point-based) has a higher R^2 and a lower RMSE, and although more data points are used in this work, it also covers a wider variety of ILs. Ali et al. (2024) used a more advanced LSTM model for training, but the model performance results they reported were almost the same as ANN model, both slightly worse than DNN (Point-based). In addition, LSTM training takes longer without achieving better results, a difference attributed to the inherent complexity of LSTM in processing sequential data, which provides guidance for researchers when selecting models. It is worth noting that the RF model reported by Venkatraman et al. also uses two dataset segmentation schemes. The results show that the model performance of the point-based data segmentation scheme is also better than that of the component-based data segmentation scheme, which is similar to the model performance results of the two data segmentation

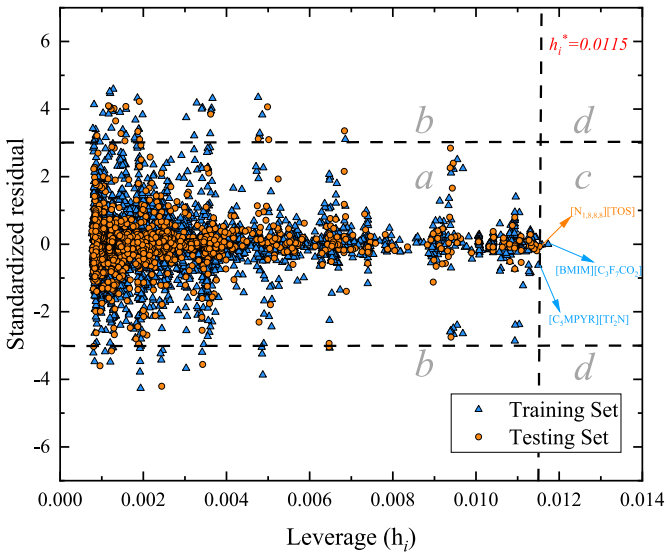

Fig. 9. The AD boundaries of the developed P-DNN are explained by the William diagram.

schemes in this study. In summary, the DNN model developed in this study has shown certain superiority in predicting the solubility of CO_2 in ILs. Comparison with other models shows that different data distribution and model selection have an important impact on the final prediction performance, which provides guidance for similar studies in the future.

3.6. Applicability domain

During the QSPR modeling process, the model's accuracy may be impacted by uncertainties in the experimental data. To mitigate this and evaluate potential outlier effects, AD analysis was used to assess the limitations and range of model predictions in this work. It provides a visual means of identifying the chemical structure of outliers (X outliers) and response outliers (Y outliers) with structural effects in the model. AD is a common technique to detect structural outliers in ML models. Several methods exist for calculating AD, with the leverage method being the most widely used(Zhang et al., 2022). This method was evaluated on the basis of the leverage value (h_i) of each ILs. When the h_i value is higher than the critical leverage value (h*), it indicates the presence of structurally influenced outliers in the molecule, and the model's prediction of them may be unreliable at this point. The standardized residual (Z_{ei}) responds to the change in the corresponding anomalies, and when the Z_{ei} of a compound is between −3 and 3 ($-3 < Z_{ei} < 3$), it indicates that the substance does not have a response anomaly. h_i, h*, and Z_{ei} are calculated by equations (18)–(20) as follows (Tropsha et al., 2003):

$$h_i = v_i \left(V^T V\right)^{-1} \times v_i^T \tag{18}$$

$$h^* = \frac{3(d^* + 1)}{b} \tag{19}$$

$$Z_{ei} = \frac{y_i - \hat{y}_i}{S_e} \tag{20}$$

where'd^*' is the number of variables input to the DNN model, 'b' represents the count of data points in the training set, 'v_i' is a $1\times d^*$ dimensional matrix containing the input parameters, and 'V' is a $b\times d^*$ dimensional matrix. '$\hat{y}_i$' and 'y_i' represent the predicted and experimental values of the data points, respectively, and 'S_e' is the standard deviation of the residuals.

The William plot was used to evaluate the range of the model by plotting Z_{ei} and h_i for each data point. The William plot in Fig. 9 shows the Z_{ei} and h_i for all data points, where the boundary of the AD was formed by three lines with $h^* = 0.0115$ and $Z_{ei} = \pm 3$, dividing the entire William plot into four parts (*a*, *b*, *c*, and *d*). The ILs data points in region '*a*' conform to the model's normal references. It could be seen that most of the ILs were within the AD range, and these predictions were considered reliable and proved the validity and robustness of the developed P-DNN model. The data points in the ILs in region '*b*' have high Z_{ei} values, indicating that the predictions were biased. In this region, there were 1.29 % (64/4939) of data points in the training set and 1.94 % (24/1234) of data points in the test set. This situation might result from the high uncertainty of the experimental data points rather than a false prediction by the model. The data points in region '*c*' were high h_i values ($>h^*$), and this region has just seven data points of three ILs types ([BMIM][$C_3F_7CO_2$], [C_5MPYR][Tf_2N] and [$N_{1,8,8,8}$][TOS]). Although the data points in this part were higher than h^*, the Z_{ei} between its predicted and experimental values are all satisfactory (less than ±3). This further demonstrates the stability and validity of the model, reflecting the good generalization of the model. The ILs in the final region '*d*' were both response anomalies (high Z_{ei}) and structural anomalies ($>h^*$), which have a negligible effect on the model if the data points were slightly above $Z_{ei} = \pm 3$ or h^*. However, if they were far away from $Z_{ei} = \pm 3$ and h^*, the outliers should be removed to minimize the interference with the model. There were no data points in region d in this work, thus indicating that the availability of the data points was all within reasonable limits, and also concluding that the developed P-DNN model is sufficiently accurate.

3.7. Model interpretation

DNN models were developed to predict CO_2 solubility in different ILs. In order to be able to select candidates for ILs with higher solubility, it is necessary to discuss the impact of different molecular characterization information on the final results. SHAP is a comprehensive method for interpreting ML model outputs(Lundberg et al., 2020), offering insights into candidate selection for ILs based on their structural characteristic. The SHAP method was often used to explain the significance of features in traditionally "black-box" or unexplainable models, including neural networks, etc. The SHAP method, rooted in cooperative game theory, calculates SHAP values for various features to quantify their contributions to prediction outcomes. The magnitude of the SHAP value indicates the extent of each feature's contribution to the prediction result, so by comparing the SHAP values of different features, we could effectively determine the features with high contribution value. Fig. 10 (a) shows the average |SHAP| ranking and swarm diagram of each input feature. From a global perspective, pressure (P) is the most important, followed by temperature (T). The swarm diagram shows that higher p (red) corresponds to positive SHAP values, which is beneficial for uplift prediction. Higher T (red) corresponds to negative SHAP values, which is depressurizing prediction, indicating that CO_2 solubility increases with increasing pressure and decreases with increasing temperature, consistent with thermodynamic and experimental patterns(Behera et al., 2023; Ghoderao and Byun, 2024). Regarding structural descriptors, the nonpolar segments P(σ,7)–P(σ,9) rank highly and generally contribute positively, indicating that a larger nonpolar surface area (stronger dispersion/van der Waals interactions and more abundant free volume) is beneficial for CO_2 dissolution. Several segments of the hydrogen bond acceptor region (σ > 0) (such as P(σ,11), P(σ,12), P(σ,14), P(σ,15)) also contribute significantly, suggesting that hydrogen bond acceptor ability (mainly anions) is more crucial for solubility regulation. In contrast, the donor region (σ < 0) is mostly secondary or has inconsistent contribution directions. Overall, the influence of anion-related characteristics on CO_2 solubility is stronger than that of cations.

Fig. 10(b) shows the SHAP force plots of the same ionic liquid [BMIM][BF_4] under three operating conditions, visually presenting the positive/negative contributions and strengths of each characteristic.

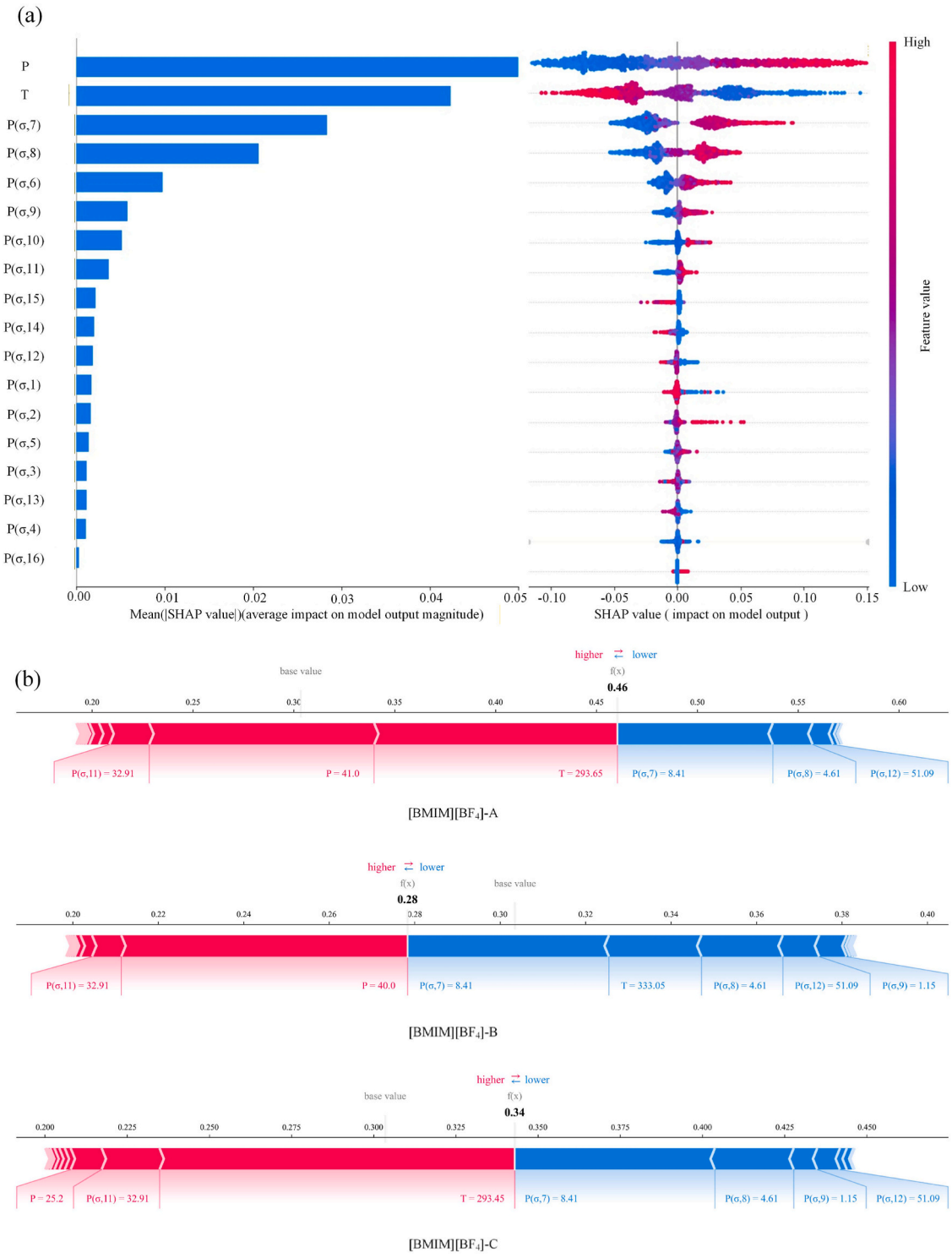

Fig. 10. (a) Significant contribution of input features of P-DNN model and relationship with output results (b) Feature contribution volcano plot of three different [BMIM][BF_4] samples.

Comparing [BMIM][BF_4]-A (P = 41 bar, T = 293.65 K) and [BMIM][BF_4]-B (P = 40 bar, T = 333.05 K), both have similar pressures and both contribute positively. However, as temperature increases, the SHAP of T changes from positive to negative and jumps to become the main inhibitory characteristic, thus the predicted value of B is significantly lower than that of A. Furthermore, compared with [BMIM][BF_4]-C (T = 293.45 K, P = 25.2 bar), as the temperature decreases, the contribution of T turns positive. Since P decreases, its positive contribution is weakened, so the predicted value of C falls between A and B. The SHAP force plot of a single sample clearly shows the synergistic regulation of solubility by different features. Through model interpretation, to screen ionic liquids with high CO_2 solubility, anions with a higher proportion of nonpolar surface area and stronger acceptor capacity can be preferentially selected.

Analysis by the SHAP method is useful to provide guidance for screening new combinations of ILs. Due to cations such as tetrabutylphosphonium ($[P_{4,4,4,4}]^+$) and1-decyl-3-methyl-imidazolium ($[DMIM]^+$), anions such as bis (pentafluoroethanesulfonyl) amide

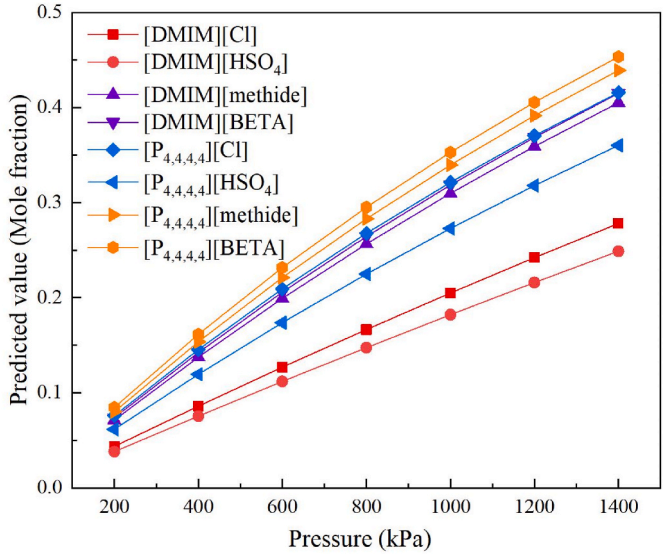

Fig. 11. Using the ML model to calculate the mole fraction of dissolved CO_2 as the pressure changes in the newly combined ILs system at 298.15K.

([BETA]$^-$) and tris (trifluoromethylsulfonyl) methide ([methide]$^-$) have high values of P6, P7, P8, and P9 and low values in the polar region, they can be used as candidates with high CO_2 solubilization capacity. In contrast, cations such as 1,3-dimethyl-imidazolium ([MMIM]$^+$) and 1-ethyl-3-methyl-imidazolium ([EMIM]$^+$), and anions such as [Cl]$^-$ and [HSO_4]$^-$, may exhibit lower CO_2 solubility due to its lower values in the No-polar region (P6, P7, P8, and P9) and higher values in the highly polar region. Considering different combinations of ILs and based on the predictions of the ML model, we recombined eight ILs and calculated their solubilities at different pressures. These combinations include [DMIM][Cl], [DMIM][HSO_4], [DMIM][methide], [DMIM][BETA], [$P_{4,4,4,4}$][Cl], [$P_{4,4,4,4}$][HSO_4], [$P_{4,4,4,4}$][methide], and [$P_{4,4,4,4}$][BETA]. Fig. 11 displays the calculated CO_2 solubility in the newly created ILs at 298.15 K under various pressure. It's evident from Fig. 11 that the solubility follows Henry's law, increasing as the pressure rises. In addition it could be seen that when the cation selects for [DMIM]$^+$ or [$P_{4,4,4,4}$]$^+$, the anion selects for potential candidates [methide]$^-$ and [BETA]$^-$ in combination with the new ILs to dissolve CO_2 were both stronger than those in combination with [Cl]$^-$ and [HSO_4]$^-$. This was due to the larger values of [methide]$^-$ and [BETA]$^-$ compared to [Cl]$^-$ and [HSO_4]$^-$ in the nonpolar region. Thus, having a larger free volume allows CO_2 to interact and have a greater solvation capacity, and the modeling results of the settlement were consistent with the content of the argument. It was noteworthy that although both the anions and cations in [DMIM][methide] and [DMIM][BETA] were potential candidates, since [$P_{4,4,4,4}$]$^+$ has a larger value in the nonpolar region than [DMIM]$^+$ and a zero value in the high polarity region. Therefore, even in combination with the non-candidate anion [Cl]$^-$, the new ILs still have a stronger ability to dissolve CO_2 than the above two. Based on modeling analysis and calculations [$P_{4,4,4,4}$][methide] and [$P_{4,4,4,4}$][BETA] seem to be promising solvents for increasing CO_2 solubility. Moreover, according to the guidance of the conclusions given in this work, the screening efficiency could be improved by considering their σ-profile information as a theoretical basis when screening ILs as trapping agents for gases.

4. Discussion

Although COSMO-RS can rapidly generate σ-profile molecular descriptors and provide initial predictions of CO_2 solubility in ionic liquids under general conditions, it neglects gas non-ideality and liquid compressibility, leading to systematic deviations in high-pressure, high-solubility scenarios (R^2 = 0.6185, RMSE = 0.1201). To overcome this limitation, this study uses the quantum-chemical descriptors generated by COSMO-RS as inputs to a DNN, enabling the deep neural network to capture nonlinear behavior (e.g., the sharp increase in solubility at high pressures) and achieve near-experimental accuracy under a "point-based" data split (R^2 = 0.9904). However, the purely data-driven model lacks physical interpretability grounded in thermodynamic principles, and its descriptors must be computed via the commercial software COSMOthermX; for example, for [EMIM]$^+$, a TZVP basis set and BP functional calculation takes approximately 5 min, imposing a burden on computational resources and costs.

Notably, the model has several limitations that warrant consideration. First, the predicted solubility of newly designed ionic liquids remains unvalidated by experimental data, requiring further experimental verification to confirm their practical applicability. Second, the generalization capability of the model beyond the training descriptor space (e.g., for ionic liquids with extreme structural features not included in the training set) is uncertain, posing risks of unreliable predictions in out-of-domain scenarios. Third, while the model performs well in high-pressure regions, its adherence to fundamental thermodynamic consistency (e.g., phase equilibrium constraints at ultra-high pressures) has not been rigorously validated, which may limit its reliability in extreme operating conditions.

In the future, constructing a publicly available, open-source database could reduce descriptor calculation barriers, and methods that integrate machine learning with thermodynamic theory could balance predictive power and physical meaning. Moreover, SHAP analysis in this study reveals the structural features of ionic liquids that efficiently adsorb CO_2: on the anion side, small-valued strong hydrogen-bond acceptor regions (e.g., [BETA]$^-$, [methane]$^-$) weaken ion aggregation and enhance CO_2 binding; on the cation side, bulky nonpolar functional groups (e.g., [$P_{4,4,4,4}$]$^+$) provide more free volume and promote van der Waals interactions. Through a "component-based" partitioning scheme and a more rigorous evaluation framework, this work not only reduces the time for ionic liquid screening from experimental timescales to minutes but also offers new insights into validating the generalizability of machine-learning models.

5. Conclusions

In this work, using the σ-profile information generated by COSMO software for ILs as a derived descriptor to quantify the structure of ILs, a robust DNN model was developed for predicting the solubility of CO_2 in ILs. A dataset comprising 6173 data points was collected from existing literature for model training. After performing hyperparameter optimization, the P-DNN model with three hidden layers was evaluated. The results showed that the model had an R^2 of 0.9904, an MAE of 0.0133, and an RMSE of 0.0216. In addition, the COSMO-RS model was employed to estimate CO_2 solubility in ILs, yielding evaluation results with an R^2 of 0.6185 and an RMSE of 0.1201. Through error analysis it was seen that the predictive performance of the COSMO-RS model gradually decreases under high-pressure and high-concentration conditions, and the developed P-DNN model can effectively compensate for this limitation and be able to accurately predictions of CO_2 solubility in ILs. Although the overall performance of C-DNN is lower than that of P-DNN, it has more practical value in the "component extrapolation" scenario, so C-DNN should be prioritized for screening "unseen" ILs. Furthermore, the model's range was verified using the leverage method, which showed that only 1.94 % of the data points in the test set were considered structural outliers and all of them were within reasonable ranges. The effect of different molecular features on the model output was further explained by the SHAP method, which revealed that ILs with $P_{\sigma\text{-profile}}$ values that are large in the nonpolar region and small in the polar region have a greater ability to dissolve CO_2. The new ILs combinations of [$P_{4,4,4,4}$][methide] and [$P_{4,4,4,4}$][BETA] proposed accordingly were expected to be solvents for improving CO_2 solubility. In conclusion, the COSMO-derived descriptors are valid descriptors for ILs,

and the developed model was able to train a good performance on these features. Thus, reliable CO_2 solubility values are calculated for ILs and effective guidance is provided for the design of suitable new ILs.

CRediT authorship contribution statement

Tianxiong Liu: Writing – original draft, Methodology, Funding acquisition, Conceptualization. **Wenguang Zhu:** Writing – original draft, Software, Methodology, Data curation. **Ying Gao:** Software, Methodology, Formal analysis. **Runqi Zhang:** Software, Formal analysis, Data curation. **Yusen Chen:** Software, Formal analysis. **Chao Guo:** Writing – review & editing, Supervision. **Hongru Zhang:** Writing – review & editing, Formal analysis. **Jianguang Qi:** Validation, Supervision. **Yinglong Wang:** Writing – review & editing, Supervision, Project administration, Funding acquisition. **Peizhe Cui:** Supervision, Project administration, Funding acquisition.

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.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 22078166 and No. 22178188), Qingdao University of Science and Technology Graduate Student Independent Research and Innovation Program (S2023KY002), and Taishan Scholar Constructive Engineering Foundation (No. tsqn202211163).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.engappai.2026.113770.

Data availability

Data will be made available on request.

References

Abranches, D.O., Zhang, Y., Maginn, E.J., Colón, Y.J., 2022. Sigma profiles in deep learning: towards a universal molecular descriptor. Chem. Commun. 58, 5630–5633.
Adeyemi, I., Abu-Zahra, M.R.M., AlNashef, I.M., 2018. Physicochemical properties of alkanolamine-choline chloride deep eutectic solvents: measurements, group contribution and artificial intelligence prediction techniques. J. Mol. Liq. 256, 581–590.
Aghaie, M., Rezaei, N., Zendehboudi, S., 2018. A systematic review on CO_2 capture with ionic liquids: current status and future prospects. Renew. Sustain. Energy Rev. 96, 502–525.
Ali, M., Sarwar, T., Mubarak, N.M., Karri, R.R., Ghalib, L., Bibi, A., Mazari, S.A., 2024. Prediction of CO_2 solubility in Ionic liquids for CO_2 capture using deep learning models. Sci. Rep. 14, 14730.
Arrieta, A.B., Díaz-Rodríguez, N., Ser, J.D., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F., 2019. eXplainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion 58, 82–115.
Asadi, E., Haghtalab, A., Shirazizadeh, H.A., 2020. High-pressure measurement and thermodynamic modeling of the carbon dioxide solubility in the aqueous 2-((2-aminoethyl)-amino)-ethanol + sulfolane system at different temperatures. J. Mol. Liq. 314, 113650.
Ashraf, W.M., Uddin, G.M., Arafat, S.M., Krzywanski, J., Xiaonan, W., 2021. Strategic-level performance enhancement of a 660 MWe supercritical power plant and emissions reduction by AI approach. Energy Convers. Manag. 250.
Baghban, A., Mohammadi, A.H., Taleghani, M.S., 2017. Rigorous modeling of CO_2 equilibrium absorption in ionic liquids. Int. J. Greenh. Gas Control 58, 19–41.
Bai, L., Zhang, H., Wang, H., Li, J., Lu, L., Zhang, H., Wang, H., 2006. Analysis of ultraviolet absorption spectrum of Chinese herbal medicine–Cortex Fraxini by double ANN. Spectrochim. Acta Mol. Biomol. Spectrosc. 65, 863–868.
Bakhbakhi, Y., 2011. Phase equilibria prediction of solid solute in supercritical carbon dioxide with and without a cosolvent: the use of artificial neural network. Expert Syst. Appl. 38, 11355–11362.
Bardeeniz, S., Panjapornpon, C., Hounkim, W., Dechakupt, T., Tawai, A., 2024. Artificial intelligence-driven control for enhancing carbon dioxide-based wastewater pH regulation in tubular reactor. Comput. Chem. Eng. 192, 108880.
Behera, U.S., Cho, S.-H., Dhamodharan, D., Byun, H.-S., 2023. Phase equilibria of binary and ternary mixtures with poly(styrene-co-hexafluorobutyl methacrylate) and solvents at high pressure and temperature. J. Supercrit. Fluids 205, 106146.
Bonito, L.P.D.D., Campanile, L., Natale, F.D.D., Mastroianni, M., Iacono, M., 2024. eXplainable artificial intelligence in process engineering: promises, facts, and Current limitations. Applied System Innovation 7, 121.
Boublia, A., Lemaoui, T., Abu Hatab, F., Darwish, A.S., Banat, F., Benguerba, Y., AlNashef, I.M., 2022. Molecular-based artificial neural network for predicting the electrical conductivity of deep eutectic solvents. J. Mol. Liq. 366, 120225.
Bürkle, M., Perera, U., Gimbert, F., Nakamura, H., Kawata, M., Asai, Y., 2021. Deep-Learning approach to first-principles transport simulations. Phys. Rev. Lett. 126, 177701.
Cao, L., Zhu, P., Zhao, Y., Zhao, J., 2018. Using machine learning and quantum chemistry descriptors to predict the toxicity of ionic liquids. J. Hazard Mater. 352, 17–26.
Caprio, U.D., Vermeire, F., Gerven, T.V., Leblebici, M.E., 2025. Physics-informed machine learning predicting CO_2 capture performances of organic mixtures. Chem. Eng. Process. Process Intensif. 216, 110410.
Cho, H.-K., Kim, J.E., Lim, J.S., 2017. The effect of cyano groups on the solubility of carbon dioxide in ionic liquids containing cyano groups in anion. Kor. J. Chem. Eng. 34, 1475–1482.
Chu, Y., He, X., 2019. MoDoop: an automated computational approach for COSMO-RS prediction of biopolymer solubilities in ionic liquids. ACS Omega 4, 2337–2343.
Chu, Y., Zhang, X., Hillestad, M., He, X., 2018. Computational prediction of cellulose solubilities in ionic liquids based on COSMO-RS. Fluid Phase Equilib. 475, 25–36.
Daryayehsalameh, B., Nabavi, M., Vaferi, B., 2021. Modeling of CO_2 capture ability of [Bmim][BF4] ionic liquid using connectionist smart paradigms. Environ. Technol. Innov. 22.
Dikki, R., Khokhar, V., Zeeshan, M., Bhattacharjee, S., Coskun, O.K., Getman, R., Gurkan, B., 2024. Composition–property relationships of choline based eutectic solvents: impact of the hydrogen bond donor and CO_2 saturation. Green Chem. 26, 3441–3452.
Eckert, F., Klamt, A., 2002. Fast solvent screening via quantum chemistry: COSMO-RS approach. AIChE J. 48, 369–385.
Fan, D., Xue, K., Liu, Y., Zhu, W., Chen, Y., Cui, P., Sun, S., Qi, J., Zhu, Z., Wang, Y., 2023. Modeling the toxicity of ionic liquids based on deep learning method. Comput. Chem. Eng. 176, 108293.
Farahipour, R., Mehrkesh, A., Karunanithi, A.T., 2016. A systematic screening methodology towards exploration of ionic liquids for CO_2 capture processes. Chem. Eng. Sci. 145, 126–132.
Ghoderao, P.N., Byun, H.S., 2024. Equilibrium solubility and computational modeling of binary system for the 2-(diethylamino)ethyl acrylate and 2-(diethylamino)ethyl methacrylate by pressurized carbon di-oxide. J. Mol. Liq. 397, 124067.
Glavatskikh, M., Leguy, J., Hunault, G., Cauchy, T., Da Mota, B., 2019. Dataset's chemical diversity limits the generalizability of machine learning predictions. J. Cheminf. 11, 69.
Grądziel, S., Krzywanski, J., Grabowska, K., Sosnowski, M., Żyłka, A., Sztekler, K., Kalawa, W., Wójcik, T., Nowak, W., Łopata, S., Sobota, T., Zima, W., 2018. Modeling of a re-heat two-stage adsorption chiller by AI approach. MATEC Web of Conferences 240.
Hassanpouryouzband, A., Farahani, M.V., Yang, J., Tohidi, B., Chuvilin, E., Istomin, V., Bukhanov, B., 2019. Solubility of flue gas or carbon dioxide-nitrogen gas mixtures in water and aqueous solutions of salts: experimental measurement and thermodynamic modeling. Ind. Eng. Chem. Res. 58, 3377–3394.
Jian, Y., Wang, Y., Barati Farimani, A., 2022. Predicting CO_2 absorption in ionic liquids with molecular descriptors and explainable graph neural networks. ACS Sustain. Chem. Eng. 10, 16681–16691.
Jiao, Y., Yin, K., Liu, T., Meng, F., Li, X., Zhong, L., Zhu, Z., Cui, P., Wang, Y., 2022. Process design and mechanism analysis of reactive distillation coupled with extractive distillation to produce an environmentally friendly gasoline additive. J. Clean. Prod. 369, 133290.
Kamgar, A., Rahimpour, M.R., 2016. Prediction of CO_2 solubility in ionic liquids with QM and UNIQUAC models. J. Mol. Liq. 222, 195–200.
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.
Klamt, A., Eckert, F., 2000. COSMO-RS: a novel and efficient method for the a priori prediction of thermophysical data of liquids. Fluid Phase Equilib. 172, 43–72.
Lei, Z., Dai, C., Chen, B., 2014. Gas solubility in ionic liquids. Chem. Rev. 114, 1289–1326.
Leonhard, K., Veverka, J., Lucas, K., 2009. A comparison of mixing rules for the combination of COSMO-RS and the Peng–Robinson equation of state. Fluid Phase Equilib. 275, 105–115.
Li, C., Li, Z., Liu, X., Xu, J., Zhang, C., 2024. Machine learning approach to predict Hansen solubility parameters of cocrystal coformers via integrating group contribution and COSMO-RS. J. Mol. Liq. 408.
Lin, S.-T., Sandler, S.I., 2002. A Priori phase equilibrium prediction from a segment contribution solvation model. Ind. Eng. Chem. Res. 41, 899–913.
Liu, T., Dong, Z., Zhu, W., Chen, Y., Zhou, M., Cui, P., Wang, Y., Zhu, Z., 2023. Prediction of the solubility of acid gas hydrogen sulfide in green solvent ionic liquids via quantitative structure–property relationship models based on the molecular structure. ACS Sustain. Chem. Eng. 11, 3917–3931.
Liu, T., Wen, Q., Sun, Q., Jiang, X., Liang, Z., Gao, H., 2025. Interpretable machine learning model for predicting CO_2 equilibrium solubility in aqueous amine solutions. Chem. Eng. Sci. 310, 121546.

Liu, X., O'Harra, K.E., Bara, J.E., Turner, C.H., 2021. Solubility behavior of CO_2 in ionic liquids based on ionic polarity index analyses. J. Phys. Chem. B 125, 3665–3676.
Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.-I., 2020. From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. 2, 56–67.
Mohan, M., Demerdash, O., Simmons, B.A., Smith, J.C., Kidder, M.K., Singh, S., 2023a. Accurate prediction of carbon dioxide capture by deep eutectic solvents using quantum chemistry and a neural network. Green Chem. 25, 3475–3492.
Mohan, M., Demerdash, O., Simmons, B.A., Smith, J.C., Kidder, M.K., Singh, S., 2023b. Accurate prediction of carbon dioxide capture by deep eutectic solvents using quantum chemistry and a neural network. Green Chem. 25, 3475–3492.
Mohan, M., Keasling, J.D., Simmons, B.A., Singh, S., 2022. In silico COSMO-RS predictive screening of ionic liquids for the dissolution of plastic. Green Chem. 24, 4140–4152.
Moity, L., Durand, M., Benazzouz, A., Pierlot, C., Molinier, V., Aubry, J.-M., 2012. Panorama of sustainable solvents using the COSMO-RS approach. Green Chem. 14, 1132–1145.
Muhammad Ashraf, W., Moeen Uddin, G., Hassan Kamal, A., Haider Khan, M., Khan, A. A., Afroze Ahmad, H., Ahmed, F., Hafeez, N., Muhammad Zawar Sami, R., Muhammad Arafat, S., Gul Niazi, S., Waqas Rafique, M., Amjad, A., Hussain, J., Jamil, H., Kathia, M.S., Krzywanski, J., 2020. Optimization of a 660 MWe supercritical power plant Performance—A case of industry 4.0 in the data-driven operational management. Part 2. Power generation. Energies 13.
Pancione, E., Erto, A., Natale, F.D., Lancia, A., Balsamo, M., 2024. A comprehensive review of post-combustion CO_2 capture technologies for applications in the maritime sector: a focus on adsorbent materials. J. CO_2 Util. 89, 102955.
Panjapornpon, C., Chinchalongporn, P., Bardeeniz, S., Jitapunkul, K., Hussain, M.A., Satjeenphong, T., 2024. Development of physics-guided neural network framework for acid-base treatment prediction using carbon dioxide-based tubular reactor. Eng. Appl. Artif. Intell. 138, 109500.
Pelaquim, F.P., Vilas-Boas, S.M., do Nascimento, D.C., Carvalho, P.J., Neto, A.M.B., da Costa, M.C., 2024. Prediction of greenhouse gas solubility in eutectic solvents using COSMO-RS. Int. J. Thermophys. 45, 70.
Ramdin, M., de Loos, T.W., Vlugt, T.J.H., 2012. State-of-the-Art of CO_2Capture with ionic liquids. Ind. Eng. Chem. Res. 51, 8149–8177.
Ramdin, M., Olasagasti, T.Z., Vlugt, T.J.H., de Loos, T.W., 2013. High pressure solubility of CO_2 in non-fluorinated phosphonium-based ionic liquids. J. Supercrit. Fluids 82, 41–49.
Rashid, Z., Wilfred, C.D., Iyyaswami, R., Appusamy, A., Thanabalan, M., 2019. Investigating the solubility of petroleum asphaltene in ionic liquids and their interaction using COSMO-RS. J. Ind. Eng. Chem. 79, 194–203.
Revelli, A.-L., Mutelet, F., Jaubert, J.-N., 2010. High carbon dioxide solubilities in imidazolium-based ionic liquids and in Poly(ethylene glycol) Dimethyl ether. J. Phys. Chem. B 114, 12908–12913.
Shiflett, M., Yokozeki, A., 2009. Phase Behavior of Carbon Dioxide in Ionic Liquids: [emim][Acetate], [emim][Trifluoroacetate], and [emim][Acetate] + [emim] [Trifluoroacetate] Mixtures. J. Chem. Eng. Data 54, 108–114.
Shiflett, M.B., Maginn, E.J., 2017. The solubility of gases in ionic liquids. AIChE J. 63, 4722–4737.
Shokouhi, M., Adibi, M., Jalili, A., Hosseini-Jenab, M., Mehdizadeh, A., 2010. Solubility and diffusion of H2S and CO_2 in the ionic liquid 1-(2-Hydroxyethyl)-3-methylimidazolium tetrafluoroborate. J. Chem. Eng. Data 55, 1663–1668.
Sistla, Y.S., Sridhar, V., 2021. Molecular understanding of carbon dioxide interactions with ionic liquids. J. Mol. Liq. 325.
Skrobek, D., Krzywanski, J., Sosnowski, M., Uddin, G.M., Ashraf, W.M., Grabowska, K., Zylka, A., Kulakowska, A., Nowak, W., 2023. Artificial intelligence for energy processes and systems: applications and perspectives. Energies 16.
Soriano, A., Doma, B., Li, M.-H., 2009. Carbon dioxide solubility in some ionic liquids at moderate pressures. J. Taiwan Inst. Chem. Eng. 40, 387–393.
Torralba-Calleja, E., Skinner, J., Gutiérrez-Tauste, D., 2013. CO_2Capture in ionic liquids: a review of solubilities and experimental methods. J. Chem. 2013, 1–16.
Torrecilla, J.S., Palomar, J., Lemus, J., Rodríguez, F., 2010. A quantum-chemical-based guide to analyze/quantify the cytotoxicity of ionic liquids. Green Chem. 12, 123–134.
Tropsha, A., Gramatica, P., Gombar, V.K., 2003. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. QSAR Comb. Sci. 22, 69–77.
Valeh-e-Sheyda, P., Heidarian, P., Rezvani, A., 2022. A novel molecular structure-based model for prediction of CO_2 equilibrium absorption in blended imidazolium-based ionic liquids. J. Mol. Liq. 360, 119420.
Venkatraman, V., Alsberg, B.K., 2017. Predicting CO_2 capture of ionic liquids using machine learning. J. CO2 Util. 21, 162–168.
Venkatraman, V., Evjen, S., Lethesh, K.C., Raj, J.J., Knuutila, H.K., Fiksdahl, A., 2019. Rapid, comprehensive screening of ionic liquids towards sustainable applications. Sustain. Energy Fuels 3, 2798–2808.
Wang, J., Song, Z., Chen, L., Xu, T., Deng, L., Qi, Z., 2021. Prediction of CO_2 solubility in deep eutectic solvents using random forest model based on COSMO-RS-derived descriptors. Green Chem. Eng. 2, 431–440.
Wang, N., DeFever, R.S., Maginn, E.J., 2023. Alchemical free energy and Hamiltonian Replica exchange molecular dynamics to compute hydrofluorocarbon isotherms in imidazolium-based ionic liquids. J. Chem. Theor. Comput. 19, 3324–3335.
Wang, Z., Su, Y., Shen, W., Jin, S., Clark, J.H., Ren, J., Zhang, X., 2019. Predictive deep learning models for environmental properties: the direct calculation of octanol–water partition coefficients from molecular graphs. Green Chem. 21, 4555–4565.
Wang, Z., Wen, H., Su, Y., Shen, W., Ren, J., Ma, Y., Li, J., 2022. Insights into ensemble learning-based data-driven model for safety-related property of chemical substances. Chem. Eng. Sci. 248, 117219.
Wen, H., Su, Y., Wang, Z., Jin, S., Ren, J., Shen, W., Eden, M., 2022. A systematic modeling methodology of deep neural network-based structure-property relationship for rapid and reliable prediction on flashpoints. AIChE J. 68, e17402.
Yazdani, M., Salehi, E., Zilabi, S., Nikravesh, G., 2023. An insight into the analogy between solute-solvent binding energy and solubility of acid gases in ionic liquids: thermodynamic modeling versus molecular dynamic simulation. J. Chem. Therm. 184, 107092.
Zhang, J., Wang, Q., Su, Y., Jin, S., Ren, J., Eden, M., Shen, W., 2022. An accurate and interpretable deep learning model for environmental properties prediction using hybrid molecular representations. AIChE J. 68, e17634.
Zhang, X., Wang, J., Song, Z., Zhou, T., 2021. Data-Driven ionic liquid design for CO_2 capture: molecular structure optimization and DFT verification. Ind. Eng. Chem. Res. 60, 9992–10000.