Front. Energy 2025, 19(6): 815–838
https://doi.org/10.1007/s11708-025-1011-7

REVIEW ARTICLE

Machine learning-based structure–property modeling for ionic liquids design and screening: A state-of-the-art review

Yijia Shao[1], Ziyu Wang[1], Lei Wang[2], Yunlong Kuai[1], Ruxing Gao (✉)[1], Chundong Zhang (✉)[2]

1 School of Energy Science and Engineering, Nanjing Tech University, Nanjing 211816, China
2 State Key Laboratory of Materials-Oriented Chemical Engineering, College of Chemical Engineering, Nanjing Tech University, Nanjing 211816, China

Abstract With the growing emphasis on sustainable development, the demand for environmentally friendly solvents in green chemical processes and carbon dioxide capture is increasing. Ionic liquids (ILs), as promising green solvents, offer significant potential but face considerable challenges, particularly in solvent selection. To overcome the limitations of traditional screening methods, machine learning (ML) techniques have recently been applied, offering a more efficient and data-driven approach. This review provides an overview of key ML methods used in solvent screening and compares them with traditional experimental and theoretical techniques. It examines the role of descriptor selection in structure–property-based methods, such as quantitative structure-activity relationships (QSAR) and quantitative structure–property relationships (QSPR), which are critical for predicting IL properties. The review also explores the application of these methods to screen IL properties, including toxicity, viscosity, density, and CO_2 solubility. Additionally, it discusses challenges in selecting appropriate models based on data scale and task complexity, integrating physical information for model interpretability, and achieving multi-objective optimization to balance key properties in ionic liquid (IL) design. Finally, it summarizes the achievements, limitations, and prospects of ML applications in ILs research, offering insights into how these methods can advance the development of sustainable ILs.

Keywords machine learning (ML), ionic liquid (IL), structure–property, molecular descriptors, physical property

1 Introduction

As environmental issues become increasingly severe, achieving sustainable development in the chemical industry has emerged as a critical global challenge [1]. Although the widespread application of petroleum-based organic solvents has driven technological progress, their high volatility and toxicity pose significant threats to both the environment and human health [2]. Developing alternative solvents that are both efficient and environmentally friendly has become a pressing concern for scientists and engineers. Among various alternatives, ILs have stood out as a new class of solvents due to their low volatility, broad liquid temperature range, and excellent thermal stability [3]. These properties make ILs an ideal choice for green chemistry and sustainable technologies.

Unlike traditional solvents, ILs consist of organic cations and inorganic or organic anions, giving them a unique molecular structure and the characteristic of "designability" [4]. By selecting different ion combinations or introducing functional groups, researchers can tailor the physicochemical properties of ILs to meet specific requirements. For instance, in carbon capture, ILs based on imidazolium cations have been designed to enhance CO_2 absorption capabilities [5], while amino-functionalized ILs exhibit higher selectivity and solubility [6]. In catalytic reactions, ILs systems based on phosphate anions have significantly improved the efficiency of acid-catalyzed reactions [7]. Furthermore, due to their high conductivity and low volatility, ILs are employed in the design of battery electrolytes, greatly enhancing the stability and safety of electrochemical devices [8]. However, the immense design freedom of ILs presents a vast chemical space: theoretically, the combination of cations and anions can

Received Dec. 29, 2024; accepted Apr. 9, 2025; online May 30, 2025
Correspondences: Ruxing Gao, grxing@njtech.edu.cn;
Chundong Zhang, zhangcd@njtech.edu.cn

yield up to 1018 potential ILs [9], posing significant challenges designing and screening.

Reviewing the development of ILs, molecular design approaches have evolved from experimental trial-and-error to theory-assisted methods. In previous studies, ILs design primarily relied on extensive laboratory experimentation, exploring combinations of cations and anions to investigate their properties [10,11]. However, this empirical approach was costly, inefficient, and insufficient for addressing the complex and diverse demands of practical applications. With the advent of theoretical computational tools, researchers have employed methods such as molecular dynamics (MD) simulations [12], quantum chemical calculations [13], and thermodynamic predictions [14] to gain deeper insights into the relationship between the structure and properties of ILs at the molecular level. While these computational methods have significantly improved design precision, they remain constrained by computational complexity and high costs, particularly for large-scale screening.

To effectively study the physicochemical properties of ILs, researchers have developed the structure–property relationship (SPR) approach, which summarizes experimental findings and empirical knowledge [15]. This method analyzes the qualitative relationship between molecular structures (e.g., functional groups and molecular configurations) and physicochemical properties, providing an intuitive approach for preliminary IL design and screening. However, SPR is limited to identifying trends and guiding experimental design. It lacks the capability for quantitatively analysis. To address this, Hansch and Fujita [16] proposed the quantitative structure–property relationship (QSPR) model, which transforms molecular structural features into molecular descriptors (such as molecular weight, topological indices, and polarity) and employs regression algorithms to quantify relationships between these descriptors and properties. This development not only expanded the scope of IL design but also laid the foundation for data-driven methods.

In recent years, emerging technologies such as ML have further advanced QSPR methodologies [17]. ML enables the construction of complex nonlinear models that can accurately predict IL properties, handle high-dimensional data, and uncover hidden relationships. By integrating experimental data with theoretical computations, ML has significantly enhanced IL design efficiency and optimization, offering a novel pathway for developing multifunctional and high-performance ILs. However, current research still primarily focuses on optimizing single functionalities, while systematic summaries of design principles for multifunctional ILs remain scarce. Furthermore, the organic integration of experimental data, theoretical computations, and ML to achieve cross-scale, multi-objective collaborative optimization presents an ongoing challenge.

Therefore, this review begins with an overview of various ML methods applied in QSPR modeling (see Fig. 1), followed by a discussion on molecular design strategies, highlighting the evolution and advancements of IL design methods and their applications. It particularly emphasizes the application of ML-based QSPR methods in solvent design and screening. Additionally, it compares the strengths and limitations of different ML methods across varying data scales and task complexities. It also proposes an interpretative approach that integrates MD and density functional theory (DFT)

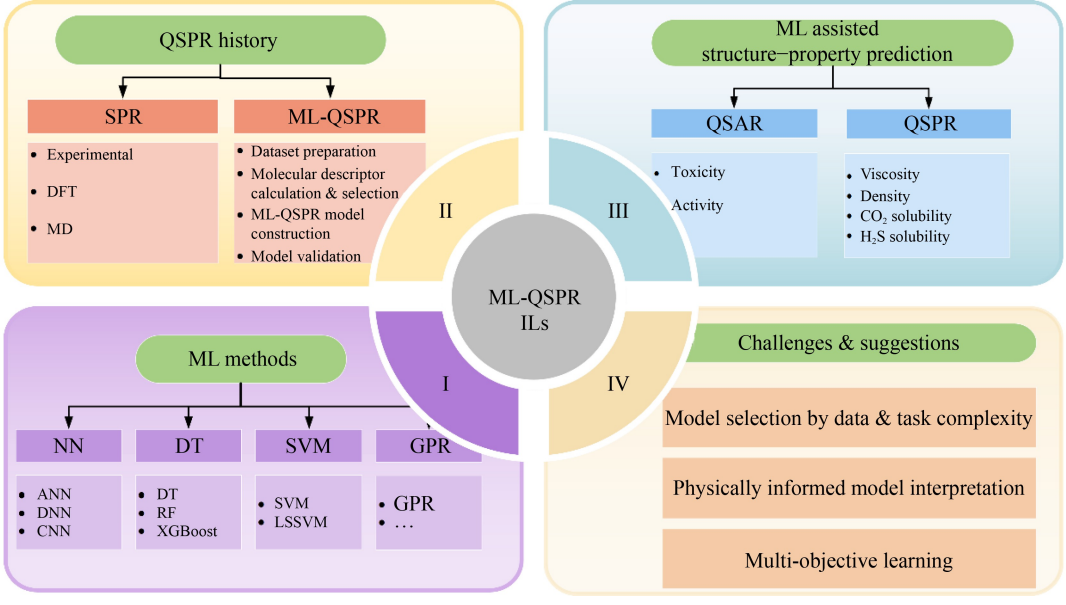

Fig. 1 Framework of this paper.

with ML models to provide a more comprehensive and in-depth physicochemical explanation. Finally, it discusses multi-property prediction for practical engineering applications, offering forward-looking insights for the precise design of novel ILs. This review aims to establish a systematic framework and provide insights for advancing IL-based solvent design and screening across various applications.

2 Machine learning methods

At present, ML methodologies have been extensively employed across numerous conventional disciplines, provi ding novel insights that have facilitated innovation and advancement within these traditional domains. In particular, the integration of ML techniques—ranging from the prediction of physicochemical properties to process optimization—has demonstrated distinct advantages in the field of solvent design and screening. This section presents a comprehensive overview of the application of various ML techniques in solvent research. A search of the Web of Science database using the keywords "machine learning" and "ionic liquid" over the past years returned hundreds of relevant papers (as illustrated by recent publication trends in Fig. 2). The most frequently employed ML methods are neural network algorithms, including artificial neural networks (ANNs), deep neural networks (DNNs), and convolutional neural networks (CNNs), collectively cited in 55 articles. These are followed by decision tree-based algorithms, such as decision trees (DTs) and random forests (RFs), which appear in 32 articles. Support vector machine (SVM) algorithms are referenced in 24 articles, while Gaussian process regression (GPR) is mentioned in 8 articles.

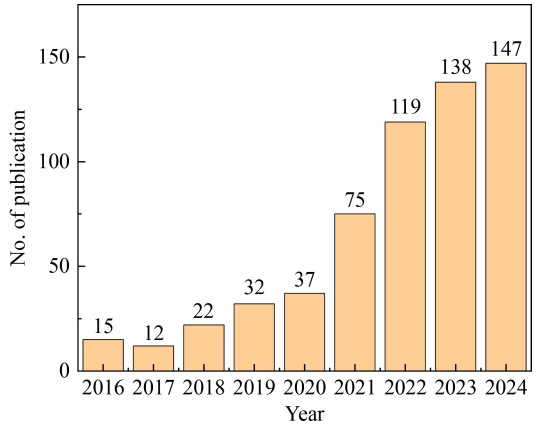

Fig. 2 Web of science search of "machine learning" and "ionic liquids".

2.1 Neural network methods

Neural network algorithms (as show in Fig. 3), the most popular ML algorithms, are extensively employed in prediction of complex properties of various solvents due to their exemplary nonlinear fitting capabilities. ANNs typically consist of one input layer, one output layer, and a few hidden layers [18], making them well-suited for problems with limited number of parameters. However, their predictive accuracy may decline when addressing more complex problems. To overcome this limitation, learning algorithms have emerged. DNN algorithms, which feature multiple hidden layers, can efficiently extract features from complex raw data and are particularly suited to addressing high-dimensional and nonlinear issues.

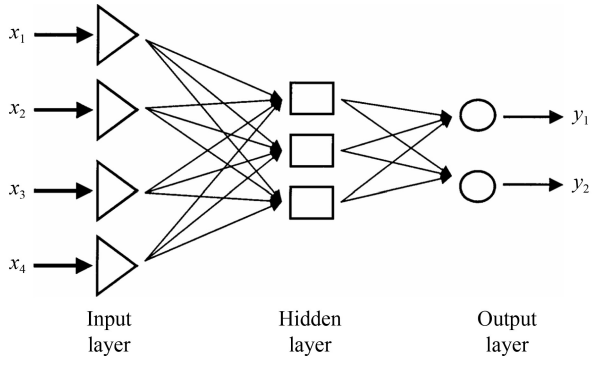

Fig. 3 Neural network algorithms (adapted with the permission from Agatonovic-Kustrin and Beresford [19], copyright 2000, Elsevier).

For example, Ardeshiri and Rashidi [20] modeled the CO_2 removal efficiency of dilute ethanolamine solvents using an ANN approach. They optimized the number of neurons and tested different transfer functions in multilayer perceptron (MLP) and radial basis function (RBF) networks to identify the most effective ANN model. By using experimental parameters for training, they effectively captured the relationship between input variables (e.g., solvent concentration, temperature, and flow rate) and output performance.

Zhu et al. [21] employed a five-enhanced ANN approaches to estimate and compare the solubility of sulfur dioxide (SO_2) in various deep eutectic solvents (DESs). The models employed molecular weight, solvent water content, temperature, pressure, and SO_2 absorption data as inputs. The comparison revealed that the single hidden layer multilayer perceptron method optimized with the Levenberg-Marquardt (MLP-LM) algorithm demonstrated the highest accuracy in predicting the DES-SO_2 phase equilibrium data, achieving an R^2 of 0.97936.

To improve the prediction accuracy, Zhang et al. [22] proposed an accurate and interpretable DNN (AI-DNN) model for predicting the lipid solubility properties of organic compounds. Compared to previous DNNs based on QSPR methodologies, the AI-DNN approach utilizes a hybrid molecular representation that integrates

descriptors from the chemistry development kit (CDK) and features learned through directed message passing neural networks (D-MPNNs). This hybrid approach captures both local and global molecules features, thereby enhancing model performance in predicting physicochemical properties.

Although ANN and DNN algorithms have been widely used in solvent selection, they still present limitations. These algorithms rely on a set of hyperparameters that must be carefully tuned [23]. Additionally, as network depth increases, model interpretability tends to decrease, and the p risk of overfitting becomes more pronounced. In this context, CNNs, as deep learning models specialized in spatial data, can effectively extract complex features through local connectivity and parameter sharing [24], often without the need for human supervision [25]. CNNs demonstrate enhanced robustness when handling high-dimensional data, effectively mitigating overfitting while maintaining interpretability.

For instance, Bahmaninia et al. [26] used a CNN model to predict CO_2 solubility in a variety of physical solvents, including methanol, ethanol, propylene glycol, *n*-pentanol, *n*-butanol, *n*-propanol. They also analyzed the effects of non-mass factors such as critical temperature and critical pressure on the solubility of CO_2 in physical solvents. The CNN model achieved an R^2 of 0.9943, outperforming other models in predictive accuracy.

Chen et al. [27] trained a transformer model on the ChEMBL database to extract molecular fingerprints, which were then used as inputs for a CNN model to predict the surface charge density profile (σ profile) and cavity volume (VCOSMO). The results demonstrate that the model could rapidly predict these properties of millions of molecules within just a few minutes. Leveraging this high computational efficiency, the researchers performed high-throughput solvent screening, offering promising implications for solvent selection and process design in chemical engineering.

2.2 Decision tree and random forest methods

In the aforementioned studies, neural network algorithms have effectively captured the nonlinear characteristics inherent in solvent selection processes. However, their practical application often encounters challenges related to feature interpretability and model stability. Consequently, researchers often incorporate decision tree models (as show in Fig. 4), which offer greater interpretability, into solvent design and screening workflows. Unlike "black-box" models such as neural networks, decision tree methods can identify key factors influencing solvent performance by recursively selecting optimal split points, from which predictive models are constructed [28,29]. Random forests further enhance precision and robustness by aggregating the results of multiple decision trees, rendering it a formidable tool in addressing complex

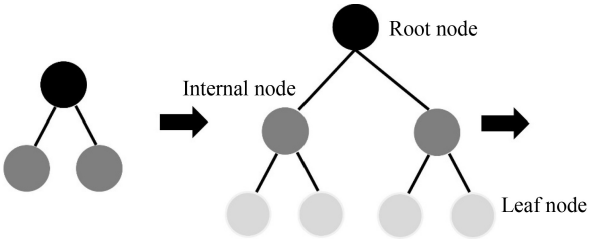

Fig. 4 Decision tree model (adapted with permission from Nakhaei-Kohani et al. [30], copyright 2024, Elsevier).

datasets.

Yin et al. [31] used three ML models, MLP, DT, and Adaboost-DT, to predict the solubility of CO_2 in potassium and sodium amino acid solutions. Their results showed that Adaboost-DT had the highest predictive accuracy with an R^2 of 0.998, followed by the DT model with an R^2 of 0.98.

Huwaimel and Alharby [32] used DT and ensemble methods, including extra tree (ET) and gradient boosting (GB), to predict the solubility of the drug lenalidomide in supercritical carbon dioxide. These models were further optimized using the sine cosine algorithm (SCA), resulting in SCA-DT, SCA-ET and SCA-GB models. The optimized models achieved R^2 values of 0.932, 0.951, and 0.997, respectively.

In a recent study, Wang et al. [33] integrated the RF method with the COSMO-RS method to create a comprehensive database of CO_2 solubility in various DESs, encompassing 1011 data points across a wide range of pressure and temperature conditions. A nonlinear QSPR model was developed using the RF algorithm, enabling accurate predictions of CO_2 solubility in DESs. When compared with traditional thermodynamic calculation methods, the RF-based model demonstrated significantly higher R^2 values, underscoring its predictive superiority.

In a related study, Wang et al. [34] also applied the RF algorithm to predict key process parameters of post-combustion CO_2 capture (PCC), including CO_2 production rate, heat duty, CO_2 absorption efficiency, and dilute solution loading. Through modeling, the RF method successfully captured complex relationships among these parameters, offering valuable insights for optimizing CO_2 separation and improving the economic performance of PCC systems. Moreover, by aggregating predictions from multiple DTs, the RF approach reduced overfitting and improved model generalizability.

2.3 Support vector machine methods

Support vector machine (SVM) is an ML model based on the concept of a hyperplane, which is used to determine the optimal boundary between different classes (as show in Fig. 5). This is achieved by mapping the data to a high-

dimensional space using kernel functions, which enable the identification of a suitable linear partition plane within this space. SVM is particularly effective in addressing binary classification problems, and its generalization ability often surpasses that of neural network algorithms.

In a recent study, Kataoka et al. [35] applied an SVM model, based on molecular surface information, to predict two-phase separation during CO_2 absorption. They investigated the use of a mixture of alkanolamines and organic solvents as an absorber, predicting the CO_2 absorption rate of 61 mixed solvents and their phase states before and after absorption. The SVM method demonstrated exceptional capability in addressing small datasets, maintain an accuracy rate exceeding 90% despite the limited number of mixed solvents in the dataset, which contained only 61 entries.

Abdollahzadeh et al. [37] conducted a comparative analysis of seven ML methods to predict the densities of 149 DES. Their results demonstrated that the least squares support vector regression (LSSVR) method outperformed the other methods, with an R^2 of 0.99798. LSSVR, a variant of the SVM method, employs a Gaussian kernel function, which is particularly effective in capturing nonlinear relationships in high-dimensional spaces.

Boobier et al. [38] employed an ML approach combined with computational chemistry to predict the solubility of organic solvents in water. Nonlinear models, such as SVM and Gaussian process (GP), demonstrated superior performance, with high consistency across five distinct datasets. However, the overall model performance was highly dependent on the quality of the chosen descriptors and the dataset used for training.

2.4 GP regression methods

Gaussian process regression (GPR), Bayesian non-parametric model, assumes that the data follow a Gaussian distribution and uses prior distributions, combined with observed data, to infer posterior distributions (as shown in Fig. 6) [39]. This framework enables GPR to provide stable predictive results even when sample sizes are limited. Consequently, GPR is frequently used in process optimization, especially when handling multivariate and nonlinear objective functions. In Bayesian optimization, GPR is employed to iteratively identify potential optimal conditions, effectively reducing the number of experimental iterations while enhancing the efficiency and accuracy of the optimization process.

Toots et al. [41] employed various ML methods to predict the partition coefficient, log$K$, of hydrocarbon gases in ILs based on cation types. Their results demonstrated that nonlinear methods, such as SVR and

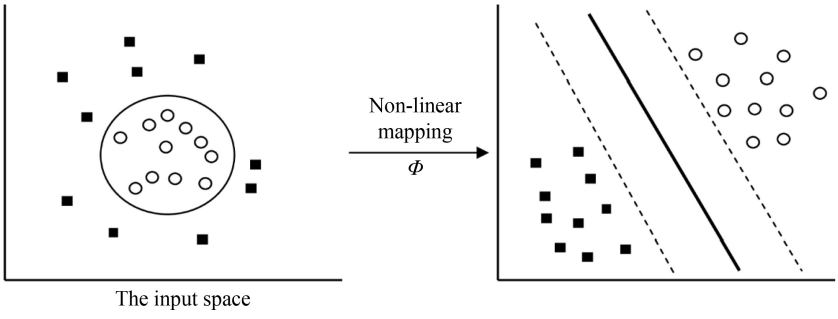

Fig. 5 SVM model generation (adapted from Huang et al. [36] under the terms of CC BY license).

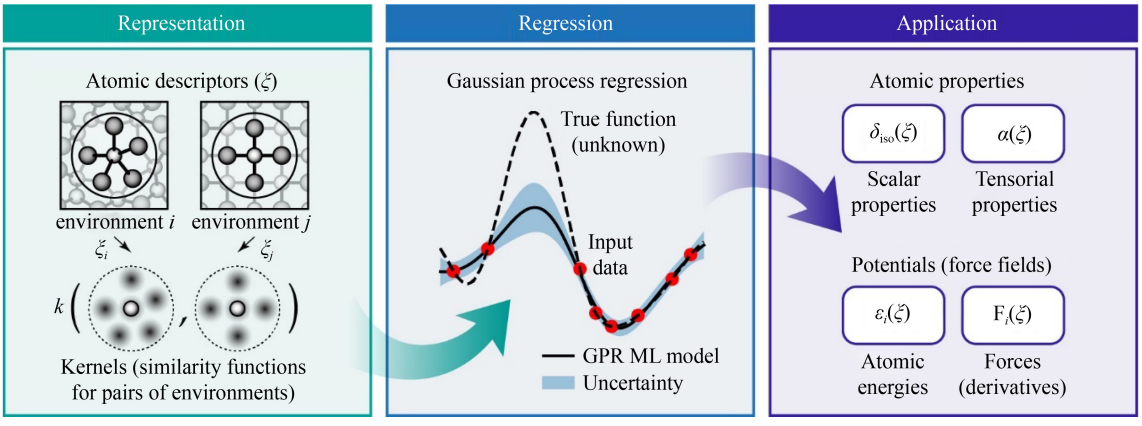

Fig. 6 GPR method applied in materials and molecules (adapted from Deringer et al. [40] under the terms of CC BY 4.0 license).

GPR, exhibited higher predictive accuracy compared to linear methods like multiple linear regression (MLR). Notably, the GPR method provides uncertainty estimates for its predictions, which are critical for assessing model reliability. Moreover, since data availability for ILs is often limited, the GPR method proves particularly advantageous by delivering stable predictions even under small-sample conditions.

The GPR method is also widely applied in the field of process optimization. Zhou et al. [42] applied GPR as a surrogate model in Aspen Plus to optimize a system process involving syngas conversion, the water-gas shift reaction, and monoethanolamine-based carbon capture. Compared to detailed process simulation-based optimization, the GPR-based approach significantly reduced optimization time, from an average of 1218.72 s to just 4.12 s. Furthermore, when compared to other surrogate modeling methods, such as response surface methodology (RSM), high-dimensional model representation (HDMR), and ANN, the GPR model demonstrated superior accuracy (with R^2 values close to 1 and a mean absolute error of 0.027) and optimization efficiency.

In summary, different ML methods have distinct advantages and limitations regarding data requirements, computational cost, accuracy, interpretability, and robustness. To facilitate comparison of these characteristics, common ML models are summarized and compared in Table 1, enabling readers to quickly assess their applicability and selection criteria.

3 Design of ILs

The history of IL design and screening is illustrated in Fig. 7. Since 1992, when Wilkes and Zaworotko [43] synthesized a 1-ethyl-3-methylimidazolium IL with excellent thermal stability, a surge of research into ILs as environmentally friendly solvents was initiated. In the early studies, researchers explored the structures of ILs by experimentally modifying the cations (such as imidazole and pyrimidine) and anions (such as BF4− and PF6−), which initially established the foundational concept of design based on the structure–property relationship.

Around the year 2000s, with advancements in theoretical calculations and molecular simulation techniques (such as COSMO-RS and DFT), the structure–property relationships of ILs began to be progressively quantified. As research data accumulated, investigators started to employ statistical regression methods to build correlations between IL structures and their properties, thereby laying the foundation for future studies involving big data and ML.

In recent years, with the development of artificial intelligence, modern ML algorithms (such as gradient boosting and deep learning) have been widely adopted in the design and screening of ILs, further accelerating the shift from trial-and-error methods to more precise design strategies. Such approaches not only integrate historical experimental data but also predict the performance of novel ILs in specific applications, providing strong

Table 1 Comparison of key characteristics of common ML methods

	Data demand	Computational cost	Accuracy	Interpretability	Robustness
NN	Hight	High	Excellent	Low	Sensitive
RF	Medium	Low	Stable	Medium	Strong
SVM	Medium	Medium	High	Medium	Reliable
GPR	Low	High	Accurate	High	Adaptive

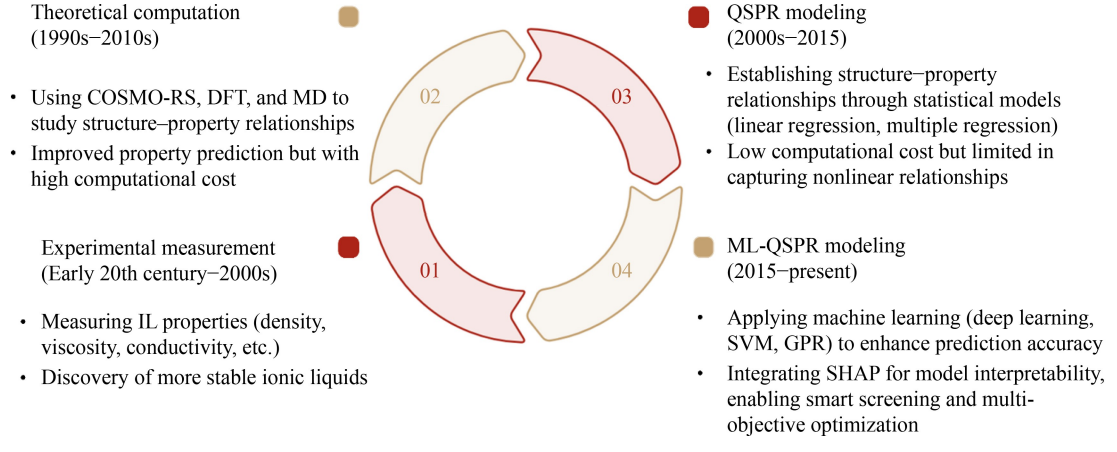

Fig. 7 History of ILs design.

support for the inverse design of ILs.

Therefore, this section is divided into two parts. First, the progress made in elucidating the interrelationship between the structure and properties of ILs through conventional experimental methods and theoretical calculations is reviewed. Then, a detailed introduction to the fundamental steps of the ML-based QSPR (ML-QSPR) approach is provided.

3.1 Experimental and theoretical calculation design methods

In the early studies on the structure–property relationships of ILs, research primarily relied on experimental data accumulation and theoretical calculations to aid modeling. Initially, researchers modified cations, anions, and alkyl side chain lengths, measuring changes in viscosity, density, and melting points through experiments. Zhang et al. [44] analyzed experimental data to summarize the influence of the macroscopic structure of ILs on their properties. Taking the melting point as an example, as the anion size increases, the interaction with the cation weakens, leading to a gradual decrease in the melting point (see Fig. 8). Similarly, as the alkyl side chain lengthens, the melting point decreases. However, when the carbon chain reaches a certain length, the melting point begins to increase again due to enhanced dispersion forces (see Fig. 9).

To explain the structure–property relationships of ILs from a microscopic perspective, Shukla and Saha [45] utilized DFT and natural bond orbital (NBO) analysis to calculate the stabilization energy E(2) and correlate it with experimentally measured melting points. Figure 10 illustrates the NBO charge distribution across different atoms using various colors. The green coloration of the C2-H atom indicates that it is the most electron-deficient among all carbon atoms in the imidazolium ring and alkyl chain and is the only atom that significantly interacts with halide anions. Moreover, for chloride-based ILs, the melting point exhibits a linear correlation with the E(2) value of C2-H (see Fig. 11). In contrast, for iodide-based imidazolium ILs, the relationship between melting point and E(2) follows an exponential fit (see Fig. 12). The

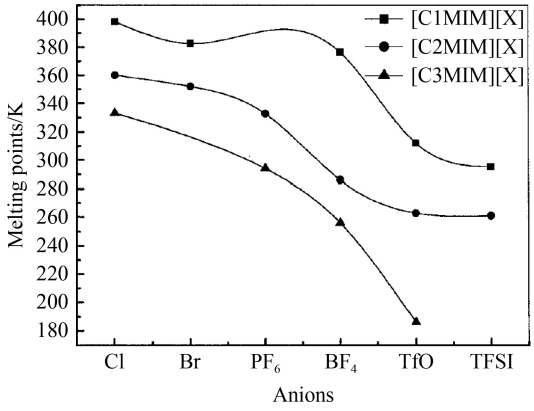

Fig. 8 Melting point variation with anions (adapted with permission from Zhang et al. [44], copyright 2006, AIP Publishing).

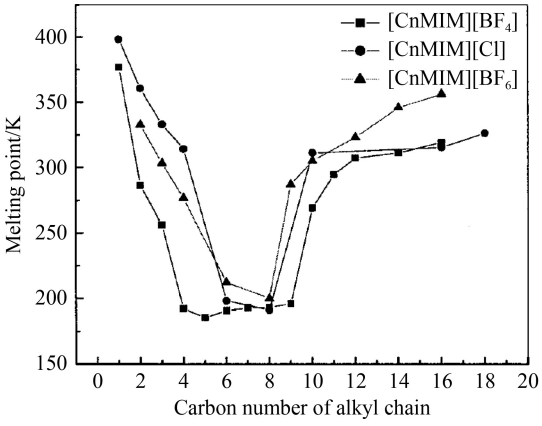

Fig. 9 Melting point variation with carbon number in alkyl chain for CnMIX (adapted with permission from Zhang et al. [44], copyright 2006, AIP Publishing).

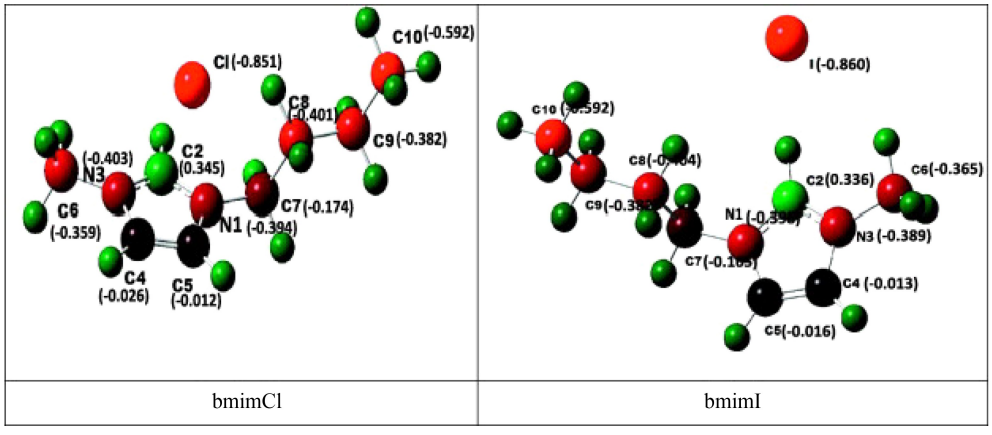

Fig. 10 NBO charge distribution on various atoms of cation and anion (green color representing the electron deficient, while red color indicating electron rich) (adapted with permission from Shukla and Saha [45], copyright 2013, Elsevier).

study also found that for both chloride and iodide anions, a decrease in alkyl chain length increases the E(2) value and raises the melting point.

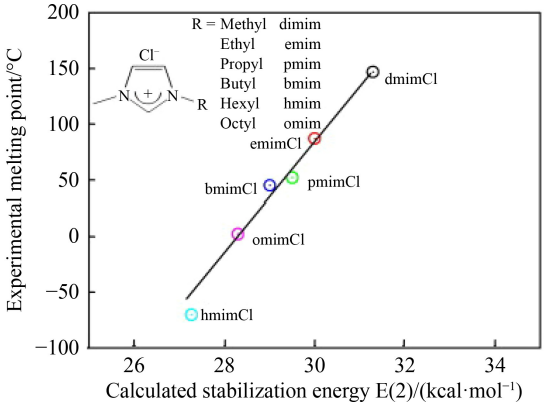

Fig. 11 Correlation of calculated stabilization energy, E(2) with experimental melting points for various chloride derivatives (adapted with permission from Shukla and Saha [45], copyright 2013, Elsevier).

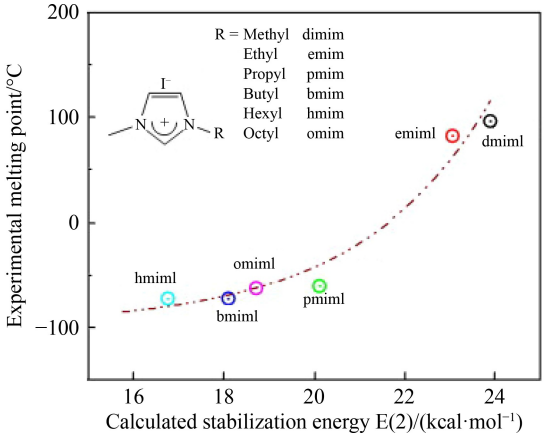

Fig. 12 Correlation of calculated stabilization energy, E(2) with experimental melting points for various iodide derivatives (adapted with permission from Shukla and Saha [45], copyright 2013, Elsevier).

DFT methods can capture the microscopic electronic structure of the IL surface, while MD simulations provide deeper insights into the behavior of ILs. Klahn and Seduraman [46] employed MD simulations to investigate the key factors influencing CO_2 absorption in various ILs. The results (see Fig. 13) indicate that CO_2 absorption is primarily determined by the strength of electrostatic cation–anion interactions. Larger ions exhibit lower ionic densities, resulting in weaker cation–anion attraction and increased average interionic spacing, which facilitates CO_2 incorporation. Additionally, cation functionalization introduces negatively charged regions on the cation surface, which enhances ionic aggregation when in proximity to other cations, thereby reducing CO_2 absorption. Moreover, when the anion size is extremely small, unavoidable cation-cation repulsion occurs, which, to some extent, promotes CO_2 absorption.

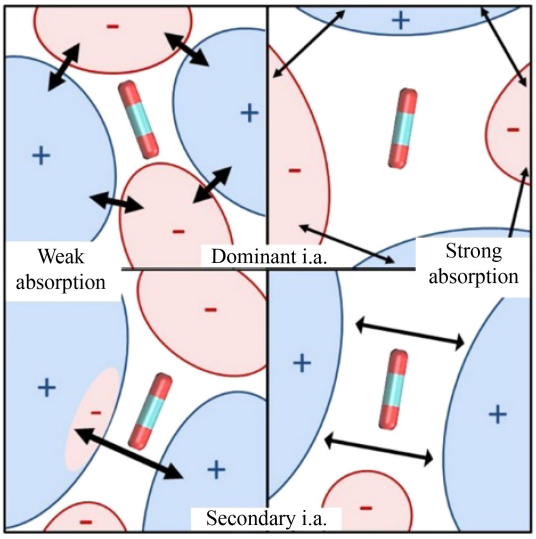

Fig. 13 Revealing the CO_2 absorption mechanism through DFT methods (adapted with permission from Klahn and Seduraman [46], copyright 2015, American Chemical Society).

3.2 ML methods

However, these theoretical computational methods, including the challenges of solving the Kohn-Sham (K-S) equations, are constrained by high computational costs and resource-intensive requirements [47], limiting their application to small systems and short simulation timescales. Meanwhile, experimental methods are time-consuming and labor-intensive. Furthermore, these traditional design and screening methods do not focus on the correlation between structure and properties of ILs. As a result, ML methods, based on statistical principles and data science, have emerged as an efficient screening tool [48]. They not only process large volumes of data but also accurately identify underlying patterns and enable precise classification and prediction. By integrating data from experimental and theoretical computations, ML approaches can significantly enhance the precision of ILs screening, addressing the limitations of traditional methods.

ML methods, as data-driven models that integrate experimental and theoretical approaches, have been widely applied to the screening of ILs. Their primary advantage lies in significantly reducing costs and improving efficiency. In the study of ILs, QSAR and QSPR methods have been extensively utilized.

QSPR methods were first proposed by Hansch and Fujita [16] in their study of the relationship between molecular structure characteristics and properties. These methods have now been widely applied in chemical and pharmaceutical screening and design processes. They are generally categorized into QSAR and QSPR. QSAR

primarily focuses on studying the relationship between molecular structure and biological activity or chemical reactivity. It is widely used in drug design, toxicity assessment, and catalyst screening. In contrast, QSPR focuses on the relationship between molecular structure and physical properties (such as density, viscosity, and solubility) and finds extensive applications in materials science and solvent screening. Early QSPR methods, represented by the model proposed by Katritzky et al. [49] in 1995, follow a five-step process, as shown in Fig. 14(a), i.e., preparation of input data, 3D geometry optimization, calculation of descriptors, statistical analysis, and QSPR report and predictions. However, with the advancement of research and computer science, ML methods have been integrated into QSPR. Figure 14(b) illustrates the basic workflow of modern QSPR methods, which are widely used in the study of ILs.

The main processes of the current ML-QSPR model are as follows:

1) Data set preparation: Currently, there are two main approaches to constructing datasets. One approach involves obtaining data directly through experiments and theoretical calculations, while the other collects data from literature sources and open-access databases.

(2) Molecular descriptor calculation and selection: In this stage, specialized software tools (such as Dragon and RDKit) are used to extract a large number of descriptors from molecular structures. These descriptors provide multidimensional information ranging from 0D to 3D (as shown in Table 2), including molecular composition, topological structure, geometric shape, and electronic properties. Subsequently, methods such as correlation analysis and principal component analysis (PCA) are employed for feature selection and dimensionality reduction, eliminating redundant and low-relevance variables. The result is a refined set of key features that are both physically and chemically meaningful, as well as highly predictive, thus laying a solid foundation for building ML models.

(3) Building the ML-QSPR model: Traditional QSPR models rely primarily on multiple linear regression methods to establish relationships between molecular descriptors and target properties. However, with the introduction of ML techniques, researchers have increasingly adopted nonlinear approaches to construct QSPR models, such as ANN, DT, SVM, and GPR. These machine learning models are more adept at capturing complex, nonlinear relationships between molecular features and physical properties.

(4) Model Validation: In the model validation phase, statistical metrics (see Table 3) are used to comprehensively evaluate the model's performance

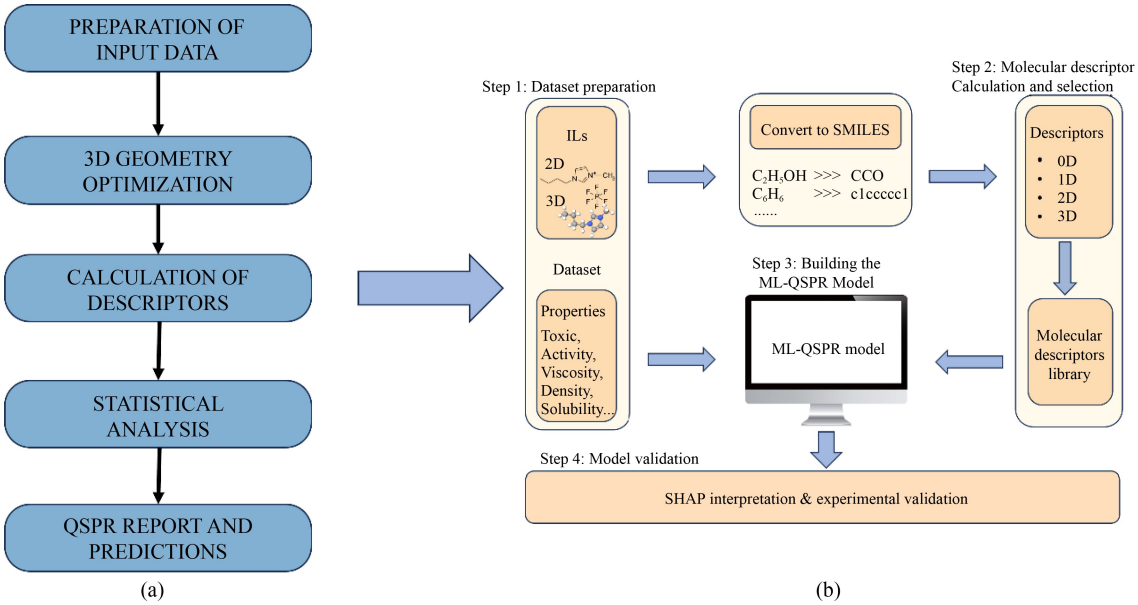

Fig. 14 Development of QSPR model.
(a) Steps of traditional QSPR model; (b) steps of ML-QSPR model.

Table 2 Molecular descriptors classified by dimension

Dimension	Classification basis	Examples
0D	Basic compositional properties	Molecular weight, atom count, bond count, chemical formula
1D	Sequence and functional group counts	Functional group counts (e.g., number of hydroxyl groups), sequence descriptors (e.g., SMILES string)
2D	Topological/graph-based descriptors	Molecular fingerprints (e.g., MACCS keys), topological indices, Randić index
3D	3D geometrical properties	Molecular volume, surface area, shape factors, topological polar surface area

across different datasets, including training, validation, and test sets. Additionally, an external independent data set is employed to assess the model's ability to generalize to new, unseen data. Furthermore, to enhance the interpretability of the model, methods such as SHAP (Shapley additive explanations) and LIME (local interpretable model-agnostic explanations) are applied. These methods provide both local and global explanations, shedding light on the contributions of individual descriptors to the model's predictions. Finally, the model's applicability domain is also assessed to ensure that the ML-QSPR model delivers reliable predictive performance across diverse chemical spaces and conditions.

Among these steps in ML-based QSAR and QSPR modeling, descriptor extraction stands out as a critical process, second only to the algorithm itself. This step directly influences the accuracy of predictions related to the properties and performance of ILs. Descriptor extraction is particularly vital in QSAR and QSPR methods, as it establishes a fundamental link between molecular structure and target properties.

The core objective of molecular structure feature selection and descriptor extraction is to effectively quantify the relationship between molecular structure and the desired properties, depending heavily on the accuracy and diversity of molecular descriptors. Currently, several categories of molecular structure descriptors are widely applied in ILs research:

Group contribution (GC) descriptors: GC methods treat a molecule as a combination of various functional groups, with each group contributing uniquely to the overall molecular properties, such as melting point, boiling point, and solubility. These descriptors leverage statistical characteristics and empirical data of groups to build predictive models for molecular properties. For example, Chen et al. [50] employed GC methods to predict properties of ILs, such as density, heat capacity, viscosity, and surface tension.

Molecular structure-based descriptors: Molecular fingerprints (MF) are one of the most common types of descriptors, encoding molecular structures as binary sequences. Each bit in the sequence represents the presence or absence of a specific structural feature in the molecule. This method is useful for quickly determining whether certain functional groups, which are related to particular properties, exist in the molecule. Simplified molecular input line entry system (SMILES) descriptors, encoded as ASCII strings, provide a detailed representation of molecular composition. SMILES descriptors are often combined with deep learning models to directly extract high-dimensional features from SMILES sequences, facilitating the prediction of complex properties. For example, Ding et al. [51] used ChemDraw to generate SMILES strings for ILs, which were then converted into MF descriptors using Python's RDKit library to predict properties such as refractive index and viscosity.

Quantum chemistry-based descriptors: Quantum chemistry-based descriptors are typically derived from molecular electronic structures and energy distributions. One of the most frequently used descriptors is the HOMO-LUMO energy gap, which reflects molecular reactivity and stability. COSMO-RS descriptors, based on quantum chemical calculations, consider molecular geometry and electronic properties and are often used to describe molecular polarity. For example, Venkatraman et al. [52] utilized the HOMO-LUMO energy gap as a descriptor in conjunction with decision with tree-based ensemble learning methods to screen the density of ILs. Their results showed consistency when compared with COSMO-RS predictions.

In summary, molecular descriptors serve as a crucial link between the molecular structure of ILs and their performance predictions in ML models. GC-based descriptors provide a straightforward and efficient way to quantify the contribution of functional groups to the overall properties of a molecule. Structure-based descriptors excel in rapid screening and the extraction of high-dimensional features, while quantum chemistry-

Table 3 Common statistical metrics

Metric	Meaning	Formula
R^2 (coefficient of determination)	Measures the goodness of fit of the model	$R^2 = 1 - \frac{\sum(yi - \widehat{yi})^2}{\sum(yi - \bar{y})^2}$
Q^2 (cross-validated coefficient of determination)	Evaluates the predictive ability of the model	$Q^2 = 1 - \frac{\sum(yi - \widehat{yi,cv})^2}{\sum(yi - \bar{y})^2}$
RMSE (root mean squared error)	Measures the deviation between predicted and actual values	$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum(yi - \widehat{yi})^2}$
MAE (mean absolute error)	Represents the average absolute difference between predicted and actual values	$\mathrm{MAE} = \frac{1}{n}\sum\lvert yi - \widehat{yi}\rvert$
AARD (average absolute relative deviation)	Assesses the relative magnitude of prediction errors, useful for datasets with large value ranges	$\mathrm{AARD} = \frac{1}{n}\sum\left\lvert\frac{yi - \widehat{yi}}{yi}\right\rvert \times 100\%$

Notes: y_i = actual value; $\hat{y}_i$ = predicted value; $\bar{y}$ = mean of actual values; $\hat{y}_{i,cv}$ = cross-validated predicted value; n = number of samples

based descriptors provide insights into the electronic properties and geometric features of molecular interactions. The accuracy and diversity of descriptor selection play a decisive role in the predictive performance of ML models, particularly when applied to QSAR and QSPR modeling.

These descriptors collectively play an indispensable role in the success of ML methods applied to ILs screening and property prediction. By ensuring that descriptors capture relevant molecular features, ML models can effectively predict and optimize IL performance, advancing the field of solvent design and application.

4 ML assisted structure–property prediction

This section provides a detailed overview of the specific applications of ML methods in conjunction with QSAR and QSPR approaches. These methodologies have proven to be invaluable in predicting a range of properties crucial for the design and screening of ILs and other chemical compounds. The properties commonly predicted include toxicity, activity, solubility, viscosity, and density. A summary of relevant studies and results is provided in Table 4.

4.1 Quantitative structure-activity relationship

The use of QSAR modeling in IL research has become increasingly valuable in predicting various biological and environmental properties, particularly toxicity and biodegradability. QSAR is primarily employed to analyze the relationship between molecular structure and biological activity. This method has roots in drug design and bioactivity prediction, where it was initially applied to determine the biological effects of pharmaceutical compounds based on their chemical structures. Early studies often utilized this method in drug screening and bioactivity prediction. For example, Hansch [72] applied QSAR in drug design, using X-ray crystallography and molecular graphics to predict drug molecules' biological activities with higher precision. In the field of IL research, QSAR is mainly used to predict properties such as toxicity and biodegradability. Peric et al. [53] utilized a GC method proposed by Luis et al. [73] to develop a QSAR model that divided IL molecules into three main components: anions, cations, and cation substituents. The study compiled an experimental EC50 value dataset for 13 ILs and combined it with 42 additional EC50 values from literature, establishing an ecotoxicity testing database. The QSAR model effectively predicted the ecotoxicity of ILs, showing agreement between experimental and predicted results. GC analysis revealed that IL anions had a negative contribution to toxicity, while cations had a positive contribution. This model provides guidance for synthesizing more sustainable ILs by identifying structural groups influencing ecotoxicity.

Table 4 Summary of works using ML methods for predicting properties in ILs

Property	Methods	Dataset	Reference
Toxic	MLR	55	Peric et al. [53]
	MLR, ELM	160	Zhu et al. [54]
	RF, XGBoost	160	Wu et al. [55]
Activity	RF, k-NN	47, 83	Hodyna et al. [56]
Viscosity	RF, KNN	13798	Carrera et al. [57]
	ANN-GC	8672	Chen et al. [58]
	AdaBoost, CatBoost, XGBoost, k-NN, RF	2676	Huang et al. [59]
	DNN, CNN	2119	Acar et al. [60]
Density	NLP, PLS, RFR, XGBoost, ASNN, DNN, CP-MPNN, CNF, Trans-CNF, GIN, Trans-CNN	131932	Baskin et al. [61]
	ANN, XGBoost, LightGBM,	34754	Liu et al. [62]
	RF, GBM	–	Venkatraman et al. [52]
CO_2 solubility	GPR	402114	Kuroki et al. [63]
	ANN-GC, SVM-GC,	10116	Song et al. [64]
	GNN	10117	Jian et al. [65]
	SVM, ANN	13055	Tian et al. [66]
	GPR, LightGPR, CatBoost	10116	Yang et al. [67]
	MF, MD, MG, MI	1402	Zhong et al. [68]
H_2S solubility	CNN, RNN, DBN, DJINN	1516	Mousavi et al. [69]
	GMDH, GP, DBN, XGBoost	1516	Mousavi et al. [70]
	DNN, RF	1358	Liu et al. [71]

In addition, some researchers have chosen charge density distribution area (S_σ) and surface electrostatic potential area (S_{EP}) as descriptors for constructing QSAR models. Zhu et al. [54] compared two ML algorithms, MLR and extreme learning machine (ELM), to evaluate IL the toxicity of ILs toward Acetylcholinesterase (AChE). Using a QSAR model, the study selected S_σ and S_{EP} as descriptors. The results indicated that the ELM model had higher accuracy compared to MLR, in predicting IL toxicity to AChE.

Wu et al. [55] further advanced QSAR modeling by investigating the toxicity of ILs toward AChE, identifying key molecular descriptors affecting the model (see Fig. 15). They used ML algorithms such as RF and XGBoost algorithms to analyze the correlation between important molecular descriptors and IL inhibitory capacity on AChE. Through feature importance analysis, they identified 14 key molecular descriptors that influence IL toxicity. Their findings indicate that cation structure plays a crucial role in determining AChE enzyme toxicity, with hydrophobic cations exhibiting stronger inhibitory effects. Other contributing factors

such as branching degree, atomic mass, and partial charges were also found to contribute to toxicity. These findings enhance the understanding of the structure-activity relationship between ILs and AChE inhibition, providing valuable guidance for the design of environmentally friendly ILs.

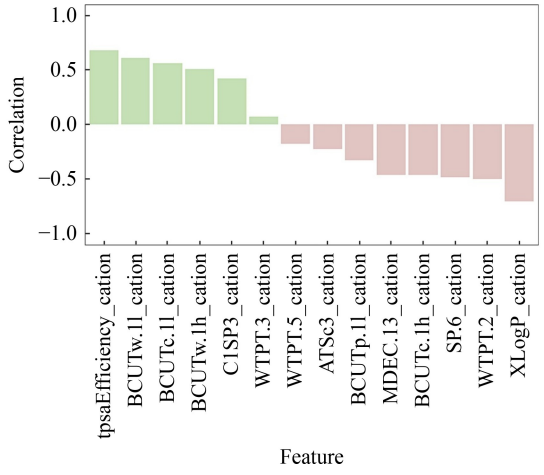

Fig. 15 Analysis of the correlation between 14 important molecular descriptors and the inhibition of AChE enzyme by ILs (adapted with permission from Wu et al. [55], copyright 2024, Elsevier).

Hodyna et al. [56] employed QSAR modeling and experimental studies to predict the antibacterial activity of imidazolium-based ILs. They used a variety of molecular descriptors, both 2D and 3D, to construct QSAR model and adopted RF and k-nearest neighbors (k-NN) algorithms for classification and prediction of IL antibacterial activity. The results showed that RF model based on diverse descriptors achieved an 88% classification accuracy and an 80% prediction accuracy for activity, highlighting the effectiveness of using diverse molecular descriptors in predicting IL antibacterial properties.

4.2 Quantitative structure–property relationship

QSPR methods are widely applied in the study of ILs, focusing on the quantitative relationship between their molecular structures and key physicochemical properties, such as solubility, viscosity, density, and thermal stability. These methods are more commonly used in the study of ILs, particularly in the process of IL design and screening. They enable accurate prediction of performance and stability, providing scientific support for the selection and design of ILs tailored for specific applications.

4.2.1 Viscosity

Among the various properties of ILs, viscosity has been a consistent focus of research. During acid gas desorption, high-viscosity ILs hinder gas diffusion, highlighting the critical role of ML-assisted structure–property relationship approaches in the screening and predicting ILs viscosity. Thus, some studies have adopted the GC method to correlate functional group information with viscosity data. Chen et al. [58] collected 8672 experimental viscosity data points for IL-H_2O mixtures from literature sources. Based on this data, they proposed a nonlinear model combining the GC method and ANN. The model's inputs included IL structural information, composition, and temperature, with output being the mixture's viscosity. The results showed that when the hidden layer contained four or five neurons, the model provided reliable viscosity predictions (see Fig. 16). For instance, with four neurons, the MAE for the training set was 0.0091, and the coefficient of determination (R^2) was 0.9962; for the test set, these values were 0.0095 and 0.9952, respectively. The proposed nonlinear ANN-GC model outperformed linear mixing models in predicting IL-H_2O mixture viscosities.

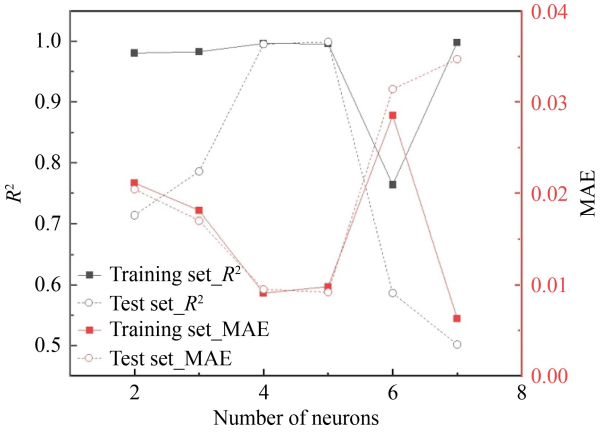

Fig. 16 R^2 and MAE versus the number of neurons in the hidden layer (adapted from Chen et al. [58], under the terms of CC BY 4.0 license).

Compared to GC descriptors, MOLMAP descriptors offer significant advantages. While GC descriptors may struggle to fully capture the intrinsic relationship between structure and viscosity in complex systems, especially when handling ILs viscosity data, MOLMAP descriptors integrate detailed atomic-level information, providing a more comprehensive representation of the structural characteristics of ILs and their mixtures, thereby leading to more accurate viscosity predictions. Carrera et al. [57] combined RF and Kohonen neural network techniques to predict the viscosity of ILs and their mixtures. The research team collected data for 13798 chemical systems from the NIST ILThermo database. Chemical structures were processed using the Chemaxon platform to generate MOLMAP descriptors, which were refined using RF and Kohonen neural networks. This study found that an

increase in cation chain length leads to higher viscosity (see Fig. 17).

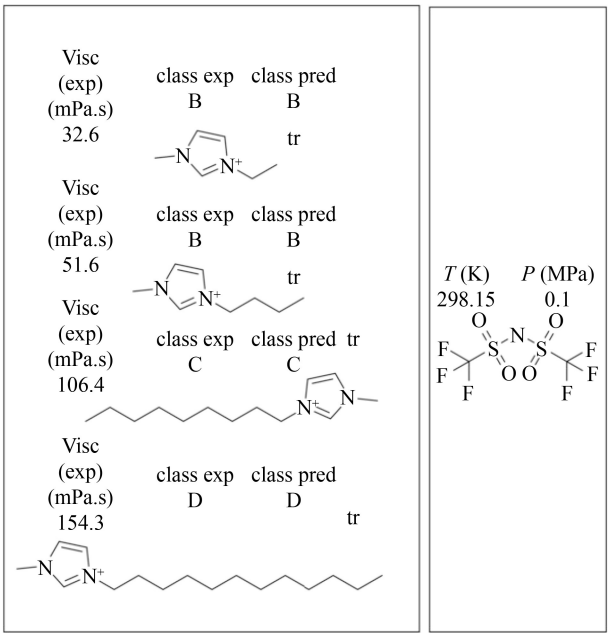

Fig. 17 Effect of the chain length of the cation (adapted from Carrera and Nunes da Ponte [57], under the terms of CC BY license).

In the studies mentioned above, descriptors were directly used as variables in model training, which often resulted in high feature dimensions during the training process. To enhance the performance and efficiency, recent research has increasingly focused on the preprocessing of descriptors. Huang et al. [59] combined the UNIFAC model with ML methods to predict the viscosity of IL-water mixtures. Molecular structures were represented using SMILES, which were then converted into numerical matrices for analysis. The UNIFAC model parameters (R, Q) and Stokes radius (S^+) were used as physical information features, along with SMILES descriptors and experimental data to construct the feature set. Among five ML algorithms tested, the CatBoost model performed the best (see in Fig. 18), and its integration with the UNIFAC model enhanced prediction accuracy. Additionally, SHAP analysis revealed correlations and interactions between different features and viscosity (see Fig. 19). Notably, the molar fraction of ILs (x_{IL}) and temperature (T) have a significant impact on viscosity, with higher x_{IL} and lower T leading to an increase in mixture viscosity.

Acar et al. [60] collected a viscosity dataset of 922 ILs from the ILThermo database, representing each IL's molecular structure using 5272 molecular descriptors. They then performed descriptor cleaning, which involved removing columns with low variance, missing values, or null entries. They normalized all molecular descriptors

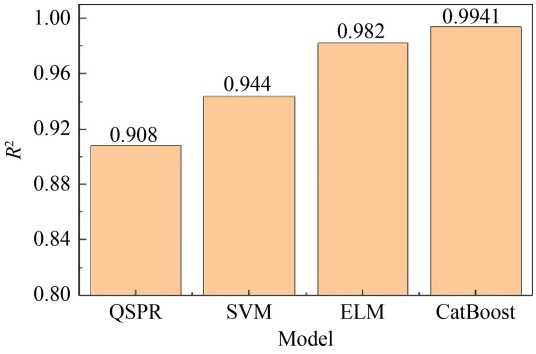

Fig. 18 Comparison of R^2 performance of different models (adapted with permission from Huang et al. [59], copyright 2023, Elsevier).

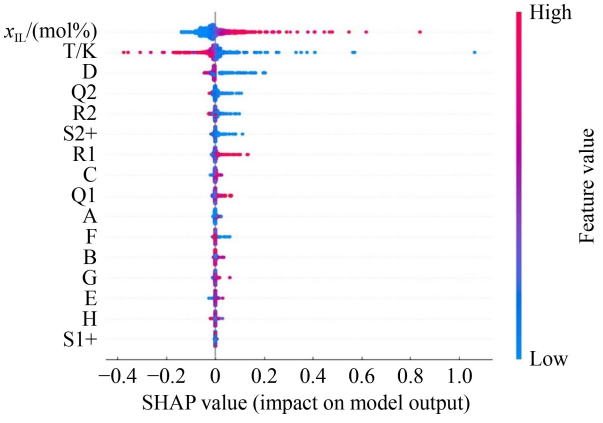

Fig. 19 SHAP value (impact on model output) (adapted with permission from Huang et al. [59], copyright 2023, Elsevier).

and viscosity values using the Standard Scaler function from scikit-learn and further reduced the dimensionality of the original descriptor matrix using Pearson correlation coefficients, ultimately identifying 179 key molecular descriptors. The study employed two deep learning models, DNN and CNN, to predict the viscosities of ILs. The results demonstrated that both the DNN and CNN models achieved outstanding predictive performance on the test dataset. Furthermore, the deep learning models were used to identify the most influential molecular descriptors influencing IL viscosity. The analysis indicated that viscosity reduction could be achieved by decreasing the cation size, shortening the alkyl chains, and lowering the ionization potential/energy. Additionally, for the same cation, further reductions in the size, chain length, and hydrogen bonding of the anion could contribute to additional viscosity reduction.

4.2.2 Density

The density of ILs is a critical property in applications such as energy storage systems and separation technologies, as it directly affects their performance. In the field of IL density prediction, Liu et al. [62] proposed

a hybrid model combining GC methods with ML algorithms to predict the essential properties, including density and heat capacity, of IL-organic solvent binary systems. They gathered isobaric heat capacity data for 17 ILs under varying temperatures and pressures, as well as density data for 129 ILs. Three ML methods—ANN, XGBoost, and LightGBM—were used to develop the model, with SHAP analysis employed to evaluate the importance of each structural feature and parameter in the predictions. The results showed that all three algorithms delivered accurate predictions, with ANN model performing the best. For density, the ANN model achieved an MAE of 0.0049 and an R^2 of 0.9942 (see Table 5). SHAP analysis revealed that the mole fraction of ILs in binary systems had the most significant positive impact on the predicted properties (see Fig. 20).

Additionally, in broader application scenarios, such as screening ILs with potential for sustainable applications, a more comprehensive consideration of various factors is necessary. Venkatraman et al. [52] developed a computational screening strategy to identify ILs with sustainable solvent design potential from a library of 8 million candidate ILs. They collected data for 10 properties of ILs, including density, from over 1500 literature sources and used combinatorial methods to generate a library of approximately 8 million ILs. They employed OpenBabel to generate three-dimensional molecular geometries and optimized them using MOPAC at the semi-empirical PM6 level. Molecular descriptors, including HOMO/

Table 5 Performance of models in density prediction [62]

Model	MAE	R^2	Computation time/s
ANN	0.0049	0.9942	8.7654
XGBoost	0.0117	0.9815	8.1258
LightGBM	0.0156	0.9760	10.3547

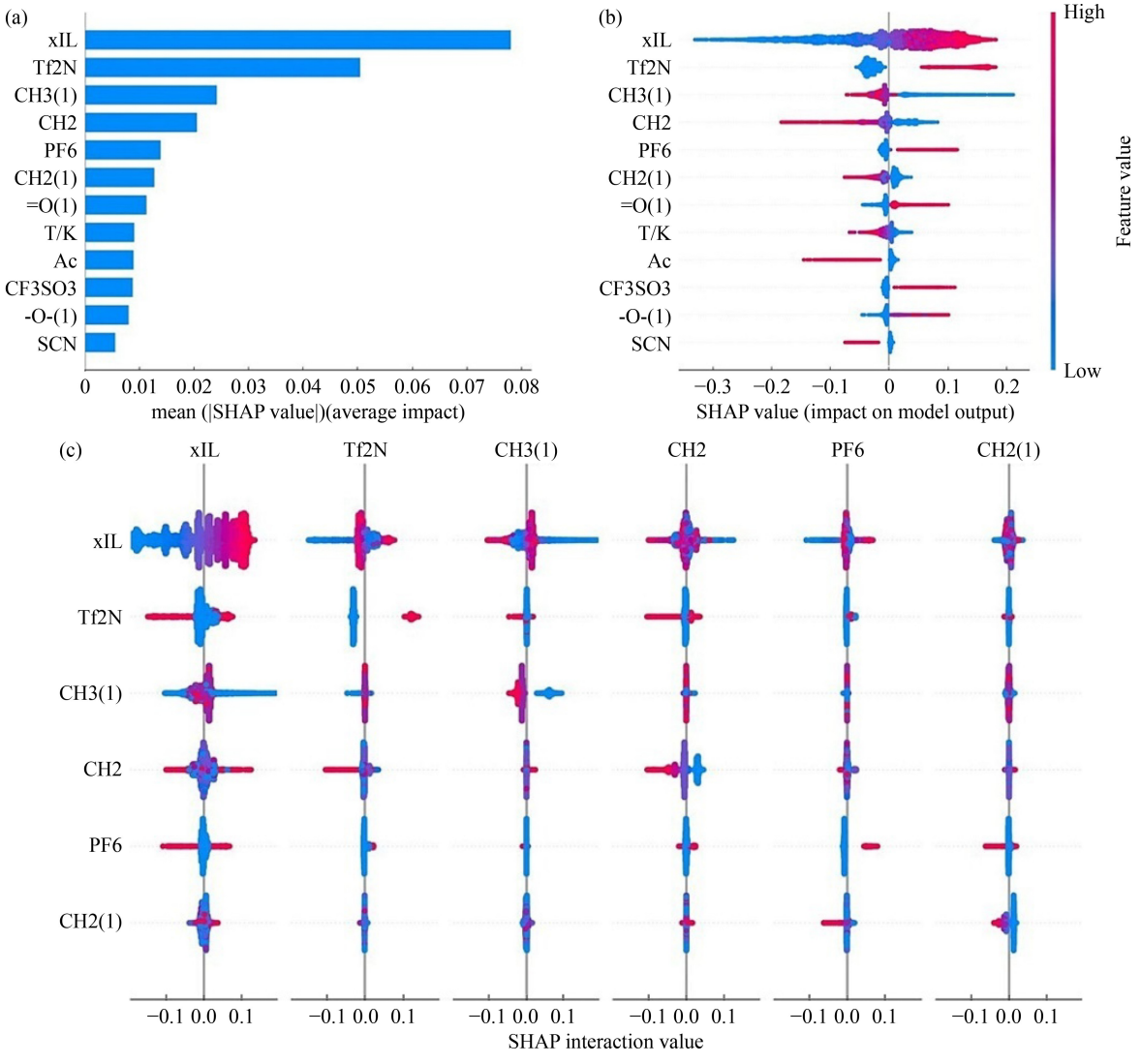

Fig. 20 SHAP analysis plot (adapted with permission from Liu et al. [62], copyright 2023, Elsevier).
(a) SHAP summary plot; (b) SHAP feature importance plot; (c) SHAP interaction plot based on density prediction.

LUMO energy levels, were calculated and used in tree-based ensemble learning methods to develop predictive models, evaluated through 5-fold cross-validation. The research team synthesized 15 ILs to validate the model predictions, with experimentally measured density values ranging from 1060 to 1457 kg/m^3. The results showed consistency between the experimental values, the ML model predictions, and COSMO-RS computational results.

4.2.3 CO_2 solubility

In response to the challenge of global warming, ILs have garnered significant attention in the field of carbon capture due to their unique properties. Many research teams currently employ traditional molecular descriptor methods for quantitative structure–property analysis of CO_2 solubility, aiming to gain deeper insights into the relationship between IL performance and molecular structure. Song et al. [64] compiled 10116 data points of CO_2 solubility in various ILs at different temperatures and pressures. They developed two models, ANN-GC and SVM-GC, using GC methods to predict CO_2 solubility. Comparative analysis revealed that ANN-GC performed slightly better than SVM-GC (see Table 6), demonstrating its ability to make fast and accurate predictions. However, while these ML models are effective in computer-aided molecular design (CAMD), they are data-driven and lack a theoretical foundation. This limitation affects their predictive capability, particularly beyond the temperature and pressure ranges seen in the training data. Tian et al. [66] decomposed ILs into cation fragments, anion fragments, and neutral groups (ionic fragment contribution, IFC), and built QSPR models using SVM and ANN. By training on 13055 CO_2 solubility data points, the IFC-SVM method achieved higher predictive accuracy (R^2 = 0.9855) than the IFC-ANN method (R^2 = 0.9732, see Table 7). While the IFC method yielded satisfactory results using ionic fragments as descriptors, it did not evaluate the contribution coefficients of these fragments. Furthermore, this descriptor-based approach lacked comprehensive molecular information and thermodynamic principles, limiting its interpretability.

Table 6 Statistical indicators for ANN-GC and SVM-GC models [64]

Model	MAE	R^2
ANN-GC	0.0202	0.9836
SVM-GC	0.0240	0.9783

Table 7 Statistical indicators for IFC-SVM and IFC-ANN models [66]

Model	R^2	MSE	MAE
IFC-SVM	0.9855	0.0008	0.0129
IFC-ANN	0.9732	0.0014	0.0252

Traditional molecular descriptor methods provide an effective approach for predicting CO_2 solubility, based on extensive experimental data and classical molecular features. However, as research progresses, there has been a shift toward exploring ILs design from a more microscopic perspective, focusing on electronic structures and atomic-level features to develop more efficient CO_2 capture solvents. Kuroki et al. [63] calculated various physicochemical properties of ILs, including geometric and electronic features, using quantum chemistry and COSMO-RS theoretical calculations. Using the GPR method, they predicted the Henry's constant (HCO_2) for CO_2 solubility in ILs. Based on these predictions, the IL [P_{66614}][PFOS] was synthesized and experimentally verified to have superior CO_2 absorption capabilities compared to the previously best-performing ILs. They found that CO_2 solubility increases in the order of $PF6^-$ < $TFSA^-$ < $PFOS^-$, due to the higher number of fluorine atoms and S = O bonds in $PFOS^-$, which enhance its interactions with CO_2 and result in the most stable Gibbs free energy of absorption.

Designing ILs from the perspective of electronic structures and atomic-level features provides valuable theoretical insights for enhancing the accuracy of CO_2 solubility predictions. However, in practical applications, it is equally important to understand the mechanisms behind model predictions and evaluate their reliability. This necessity has led to the development of interpretability methods that provide deeper insights into the prediction process. Jian et al. [65] innovatively applied graph neural network (GNN) to predict the CO_2 absorption capacity of ILs without relying on traditional molecular descriptors. Using a dataset of 10117 data points, they compared the predictive performance of GNNs with that of traditional ML methods (SVM, RF, XGBoost, and MLP). The GNN model outperformed traditional models, with an R^2 of 0.9884 (see Table 8). Additionally, they developed an IL interpreter based on the GNN model, representing IL molecules as graphs with atoms as nodes and chemical bonds as edges. Node feature vectors encoded atomic properties, and the GNN model aggregated information from neighboring nodes to rank the importance of individual atoms and functional groups. Their study found that in physical absorption, anions play a dominant role, while fluorination of alkyl chains on cations and the presence of longer alkyl chains enhance CO_2 solubility by creating more free volume. Yang et al. [67] employed three ML methods (GPR, LightGBM, CatBoost) and three molecular descriptor types (GC, molecular structure descriptors, and a hybrid GC-MSD) to predict CO_2 solubility in ILs. All models showed good predictive performance (see Table 9), with the CatBoost-GC-MSD model achieving the best results (R^2 = 0.9925). SHAP analysis was used for interpretability (see Fig. 21), identifying pressure, temperature, Chi0, Kappa2, and EState_VSA10 as the

Table 8 Comparison of different ML models [65]

Model	MAE	R^2
SVM-GC	0.0753	0.8240
SVM + FP	0.0655	0.8633
RF-GC	0.0223	0.9774
RF-FP	0.0209	0.9802
XGBoost-GC	0.0182	0.9865
XGBoost-FP	0.0189	0.9847
MLP-GC	0.0170	0.9873
MLP-FP	0.0151	0.9883
GCN	0.0723	0.8197
GAT	0.0253	0.9767
GIN	0.0137	0.9884

Table 9 Comparison of different ML models [67]

Model	Data point	R^2	MAE
GPR-GC	10,116	0.9862	0.0176
GPR-MSD	10,116	0.9866	0.0170
GPR-GC-MSD	10,116	0.9866	0.0170
LightGBM-GC	10,116	0.9839	0.0196
LightGBM-MSD	10,116	0.9876	0.0172
LightGBM-GC-MSD	10,116	0.9878	0.0170
CatBoost-GC	10,116	0.9912	0.0132
CatBoost-MSD	10,116	0.9924	0.0121
CatBoost-GC-MSD	10,116	0.9925	0.0122

top five influential features for CO_2 solubility.

Interpretability methods provide valuable insights into the relationship between molecular structures and CO_2 absorption by analyzing the internal mechanisms of predictive models. Additionally, evaluating uncertainty in model predictions is crucial, leading to research in uncertainty analysis methods to enhance the reliability of screening and design processes. Zhong et al. [68] explored four different molecular representation methods—molecular fingerprints, molecular descriptors, molecular images, and molecular graphs—and introduced a representing uncertainty (RU) method to quantify prediction uncertainty. Compared to traditional model uncertainty (MU) approaches, RU excelled in identifying unpredicted regions. Using this model, they screened 37 ILs with low viscosity, low toxicity, and high CO_2 absorption capacity from 1420 ILs. Experimental validation confirmed the reliability of the model predictions.

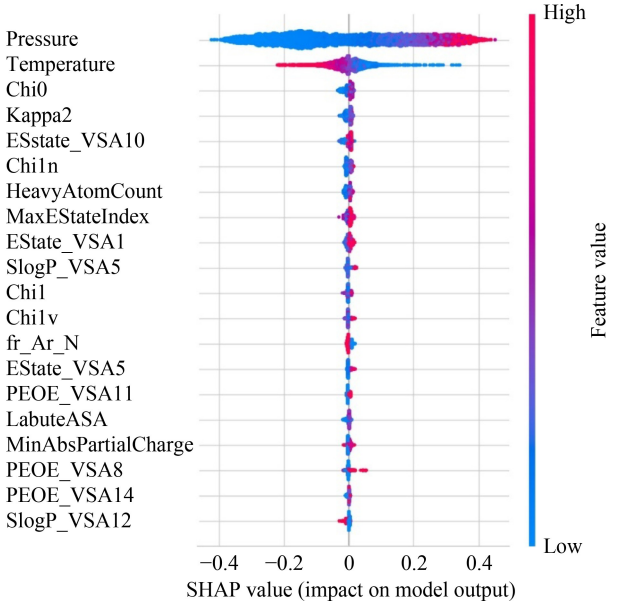

Fig. 21 Significant contribution of input features and output results for the CatBoost-GC-MSD model (adapted with permission from Yang et al. [67], copyright 2024, Elsevier).

4.2.4 H_2S solubility

In industrial removal processes, H_2S and CO_2 often coexist, but their differing physicochemical properties, toxicity, and corrosiveness require distinct treatment methods. While research on H_2S is relatively limited, its unique industrial demands and the challenges of experimental studies make structure–property-based solubility prediction particularly important. This imperative has driven researchers to conduct extensive studies in this area. Mousavi et al. [69] applied several advanced models, including CNNs, recurrent neural networks (RNNs), deep belief networks (DBNs), and DNNs based on initialized decision trees (DJINN), to predict H_2S solubility in ILs. The data set comprised 1516 data points for 37 different ILs, using IL chemical structures, temperature, and pressure as inputs. Among the models, CNN demonstrated the highest accuracy and speed in predicting H_2S solubility. Sensitivity analysis revealed that solubility increases with pressure, while temperature and the presence of –OH groups inversely affect solubility. In a follow-up study, Mousavi et al. [70] employed white-box ML approaches (group method of data handling, GMDH; genetic programming, GP), deep learning models (DBN), and ensemble methods (XGBoost) to further analyze and predict the same dataset. In addition to temperature and pressure, thermodynamic properties such as critical temperature (T_c), critical pressure (P_c), acentric factor (ω), boiling point (T_b), and molecular weight (M_w) were introduced as descriptors to comprehensively characterize ILs properties from a macroscopic thermodynamic perspective, aiming to explore their relationship with H_2S solubility. This study also found that ILs with longer alkyl chains have higher solubility under the same anion conditions (see Fig. 22.). Their larger free volume enhances van der Waals interactions, accommodating

more H2S molecules. Additionally, the fluorine content in the anion significantly affects H_2S solubility (see Fig. 23.). At 313.15 K, increasing the fluorine content improves the solubility of H_2S in ILs under the same cation conditions.

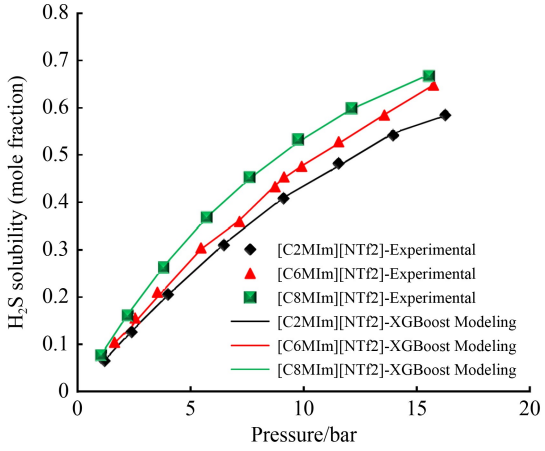

Fig. 22 Influence of cation alkyl chain length on H_2S solubility with the same anion (at 313.15 K) (adapted permission from Mousavi et al. [70] under the terms of CC BY 4.0 license).

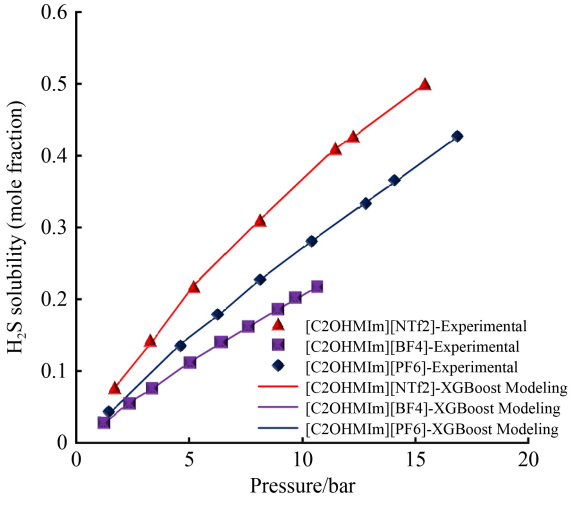

Fig. 23 Impact of fluorine content in anion on H2S solubility in ILs with the same cation (at 313.15 K) (adapted from Ref [70], under the terms of CC BY 4.0 license, copyright 2023).

The above studies employed white-box ML models for prediction but did not offer a detailed quantitative evaluation of the influence exerted by individual variables on H_2S solubility. To address this limitation, some studies adopted the SHAP method for model interpretability analysis. Liu et al. [71] combined DNN and RF methods with QSPR models, using various structural representations (e.g., molecular identifiers MI, MD, and combined MD-MI). They established the relationship between IL structure, temperature, pressure, and H_2S solubility. The results showed that structural representation significantly influenced prediction accuracy, with the ranking of MI > MD-MI > MD (see Table 10). Among the ML models, DNN outperformed RF in terms of predictive accuracy, achieving an R^2 of 0.9923. Moreover, the study utilized the SHAP method to identify and quantify the effect and factors such as pressure, temperature, and specific molecular structural features (e.g., double bonds and carbon chain length) on H_2S solubility. This analysis provided valuable theoretical insights for designing and screening of ILs with high H_2S solubility.

Table 10 Effect of different molecular representation methods on model prediction performance [71]

Model	Representation	R^2	MAE	RMSE
DNN	MD	0.9723	0.0149	0.0288
	MI	0.9923	0.0094	0.0151
	MD_MI	0.9799	0.0149	0.0245
RF	MD	0.9609	0.0247	0.0342
	MI	0.9712	0.0215	0.0293
	MD_MI	0.9637	0.0237	0.0329

5 Challenge of ML methods applied in ILs

Currently, ML methods demonstrate significant potential in the structure–property prediction and design of ILs. However, large-scale applications still face numerous challenges. This section focuses on how to select appropriate ML models based on factors such as data volume and task complexity. Additionally, it emphasizes the importance of integrating ML with traditional experimental methods and molecular simulation techniques to achieve complementary advantages and enhance overall prediction accuracy. In practical engineering applications, multi-property prediction and screening of ILs are essential for achieving high-performance targeted optimization. However, incorporating multi-property prediction within a single model remains a complex issue. To address this, it provides an in-depth analysis of the current challenges and proposes corresponding recommendations, offering both theoretical and practical guidance for the precise design of ILs in the future.

5.1 Model selection based on data scale and task complexity

Based on the current state of research, it is clear that no single ML model can be universally applicable to all prediction tasks. Different models have their distinct advantages and limitations depending on data scale and task complexity. The strengths, weaknesses, and applicable conditions of several widely used ML methods are summarized (see Table 11) to guide model selection

and adjustment in practical applications.

When the data set is large and the task is complex, DNN or GNN models are particularly effective in leverage big data and automatically extract features, leading to high-precision predictions. Baskin et al. [61] compared the performance of various ML methods and molecular representations in constructing QSPR models to predict the physicochemical properties of ILs. Their study identified the optimal models and descriptor methods for different property predictions (see Table 12). The results showed that transformer-based neural networks performed best, nonlinear methods outperformed linear methods, and SMILES-based modeling yielded better results than traditional molecular descriptors. However, these deep learning models typically require significant computational resources and are often difficult to interpret.

When the dataset is limited, SVM, GPR, and ensemble methods may be more suitable. In the study conducted by Mohan et al. [74], the predictive accuracy of various ML methods was compared on a small dataset containing sonic velocity data for 218 ILs (see Table 13). The results showed that SVM and ensemble tree-based methods, particularly XGBoost, exhibited strong predictive performance for small-sample, high-dimensional data. Additionally, in the work of Mazari et al. [75], which focused on predicting the CO_2 solubility in [Bmim][PF6] (see Fig. 24), GPR-based methods not only performed well under sparse data conditions but also provided uncertainty estimates for predictions, facilitating risk assessment.

Therefore, a stepwise strategy is recommended when using ML methods for structure–property prediction. First, a simple linear regression model can be employed to capture overall trends and preliminarily identify key factors. Then, based on the data scale and task requirements, more complex models, such as neural networks or ensemble learning, can be introduced to refine the prediction of critical parameters. This approach not only enhances the reliability of prediction results but also mitigates the risks associated with insufficient data. Additionally, it provides clear guidance for selecting appropriate models in different application scenarios for the design of ILs.

5.2 Model interpretation through the integration of physical information

In traditional ML models, statistical indicators such as R^2 are commonly used to assess predictive accuracy. While these metrics provide an intuitive sense of model performance, they do not reveal how input features are utilized internally, nor do they shed light on the physical mechanisms underlying the prediction outcomes. This "black box" issue limits the deeper physical interpretation of the results. To address this challenge, many researchers have adopted the SHAP analysis method, which quantifies the contribution of each feature to the prediction, clearly demonstrating which structural descriptors play a critical role in predicting the properties of ILs.

However, the interpretability of the SHAP method is highly dependent on the model structure, and different models may yield varying feature importance rankings [76]. Moreover, although SHAP based on conventional structural descriptors can reveal statistical correlations, it

Table 11 Advantages and disadvantages of different ML models

Method	Category	Advantage	Disadvantage
ANN	Non-linear	Avoids the impact of missing local information on the model	Prone to overfitting
SVM	Non-linear	Stable; minor data variations do not significantly affect results	High memory demand; long computation time for large datasets
DT	Non-linear	Can identify outliers; intuitive algorithm	Requires large sample training; limited ability to handle missing values
RF	Non-linear	Less prone to overfitting; good at handling high-dimensional data; retains feature information	Requires large sample training
GPR	Non-linear	Provides uncertainty estimation; works well with small datasets	High computational cost for large datasets
GFA	Hybrid	Can build multi-model systems	Lower prediction accuracy compared to MLR
LSSVM	Hybrid	Faster computation than SVM	Lower prediction accuracy compared to SVM

Table 12 Comparison of different ML models [61]

Property	ML method	Representation	Q_2	RMSE	MAE	Units
Density	DNN	CDK23	0.95	0.042	0.023	g/cm^3
Electrical conductivity	TransCNF	SMILES	0.73	0.49	0.27	lg(S/M)
Refractive index	ASNN	CDK23	0.86	0.016	0.0081	–
Surface tension	XGBoost	ISIDA-Fr	0.75	0.0041	0.0027	N/m
Viscosity	TransCNN	SMILES	0.8	0.33	0.21	lg(megaPa*s)
Melting point	RFR	CDK23	0.6	39	27	degrees K

Table 13 Comparison of different ML models [74]

ML model	Data set	R^2	AARD/%	RMSE/(m·s^{-1})	MAE/(m·s^{-1})
RF	Total	0.990	0.834	20.105	12.273
GBT	Total	0.995	0.581	14.353	8.551
XGBoost	Total	0.994	0.628	15.221	9.236
SVM	Total	0.989	0.527	20.662	7.753
FFNN	Total	0.993	0.626	17.251	9.277
GPR	Total	0.806	4.608	88.668	69.332
Two-factor PR	Total	0.990	0.857	20.432	12.539
MLR	Total	0.873	3.635	71.665	54.256

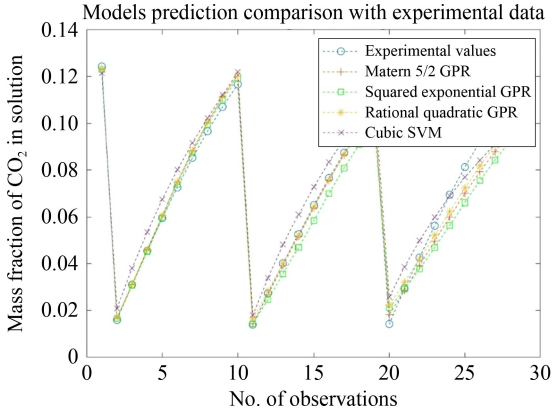

Fig. 24 Comparison of predicted results through ML models against experimental results for testing data set of CO_2 solubility in [Bmim][PF_6] (CO_2 solubility in [Bmim][PF_6] is in mass fraction) (adapted with permission from Mazari et al. [75], copyright 2020, Elsevier).

does not capture the physical mechanisms at the atomic or molecular level, nor can it elucidate how intermolecular interactions, hydrogen bond networks, and electron distributions in ILs influence property predictions.

Therefore, integrating MD and DFT methods, based on theoretical calculations, is recommended for more precise characterizations. Li et al. [77] calculated electronic properties—such as the nucleophilicity index, electrophilicity index, dipole moment, and HOMO and LUMO energies—using DFT as supplementary descriptors. These were combined with experimental conditions and structural descriptors as input features to predict the yield of IL-catalyzed CO_2 cyclization reactions. In the RF model that achieved the highest predictive accuracy, SHAP analysis revealed that four of the five most important descriptors for reaction yield were experimental parameters (see Fig. 25). Notably, the energy gap (ΔE_H–L_1) between the IL's HOMO and the substrate's LUMO had the greatest impact, with a smaller gap facilitating the reaction. This study provides a valuable reference for constructing ML models in multicomponent reactions by incorporating high-accuracy DFT-calculated electronic features with physicochemical interpretability, thereby enhancing overall predictive accuracy.

A subsequent study utilized an ML model incorporating DFT calculations to screen IL catalysts for CO_2 cycloaddition reactions. Li et al. [78] not only experimentally validated the six selected ILs but also employed DFT calculations to investigate the reaction mechanism after prediction (see Fig. 26). The proposed mechanism consists of three main steps: IL-induced epichlorohydrin ring-opening, CO_2 insertion, and subsequent cyclization to form cyclic carbonate. Among these, the IL-induced substrate ring-opening is identified as the rate-determining step. This study highlights the dual role of DFT calculations, both in providing

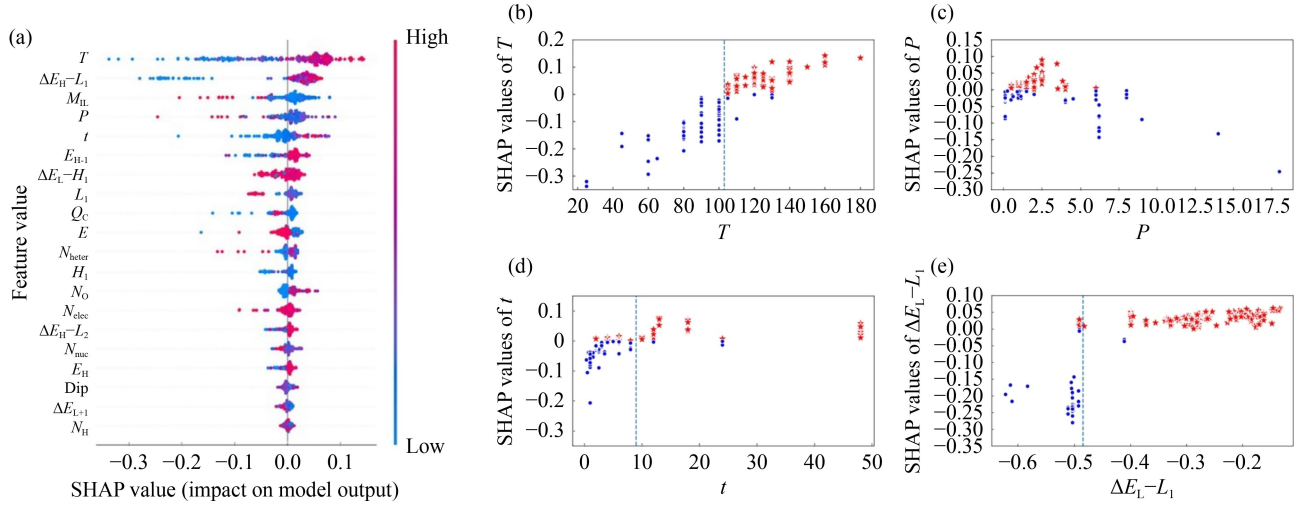

Fig. 25 SHAP analysis plot.

(a) Results of SHAP model with RF algorithm for the imidazole subset; (b) SHAP value of temperature (T); (c) SHAP value of pressure (P); (d) SHAP value of time (t); (e) SHAP value of ΔE_H–L_1 (adapted with permission from Li et al. [77], copyright 2022, Elsevier).

molecular descriptors and elucidating reaction mechanisms, offering a solid theoretical foundation for the rational design of ILs.

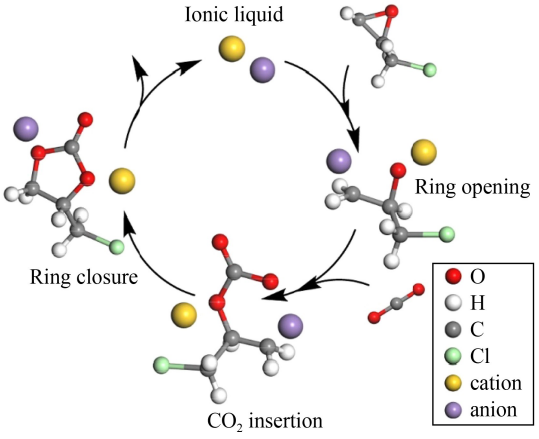

Fig. 26 Reaction mechanism of CO_2 cycloaddition with epichlorohydrin catalyzed by IL (adapted with permission from Li et al. [78], copyright 2024, American Chemical Society).

In summary, an integrated approach combining MD simulations, DFT calculations, and ML models offer a comprehensive explanatory framework that spans from the molecular to the system level. This strategy not only addresses the limitations of traditional "black-box" models but also provides a thorough and precise theoretical foundation for the targeted design of novel ILs.

5.3 Multi-objective property prediction and optimization

In the application of ILs, multiple physical property criteria often need to be met simultaneously. For example, in electrochemical CO_2 reduction, in addition to high CO_2 solubility, factors such as conductivity, viscosity, thermal stability, and safety must also be considered. Traditional single-objective prediction methods often struggle to optimize multiple properties simultaneously, typically improving one characteristic at the expense of others, making it difficult to achieve an overall optimal design. To address this issue, multi-objective property prediction has become an essential aspect of IL design.

In this context, multi-task learning offers an effective strategy. By incorporating a shared feature extraction layer in the model and constructing separate output layers for each target property, multi-task neural networks can simultaneously capture both the common influences and individual differences of IL structures on various property indices [79]. An integrated multi-task neural network can predict multiple properties, such as density, viscosity, surface tension, and melting point simultaneously (see Fig. 27). The shared layer extracts general features that provide complementary information for different properties, while the independent output layers ensure that the optimization of each property occurs without interference from the others, thereby achieving multi-objective collaborative optimization.

In addition, multi-objective optimization based on Pareto Front analysis holds great application potential. Zhang et al. [80] formulated the design problem of ILs as a mixed-integer nonlinear programming (MINLP) problem, aiming to maximize CO_2 solubility under specified conditions. After identifying eight IL combinations with the highest CO_2 solubility, the sigma constraint method was employed by selecting 10 equally spaced sigma values between the minimum and maximum viscosities. This approach converted the multi-objective optimization problem into 10 single-objective optimization problems. Ultimately, this approach resulted in a Pareto-optimal curve, illustrating the trade-off between IL viscosity and CO_2 solubility (see Fig. 28).

This section systematically examines the challenges and strategies for improving the application of ML in the design of IL. First, the choice of an appropriate model should depend on data availability and task complexity. Deep learning and GNNs are effective for capturing nonlinear relationships in complex tasks with abundant data, while SVM, GPR, and ensemble tree methods are more suitable for small datasets or simpler tasks. Second, traditional ML models, though evaluated using statistical metrics like R^2 and RMSE, often act as "black boxes" with limited interpretability. To address this, integrating theoretical computation methods, such as MD and DFT with ML models are recommended. This integration facilitates the calculation of physically meaningful descriptors, which can validate and supplement model interpretations, thereby creating a comprehensive explanatory framework that spans from the atomic-scale details to macroscopic properties. Lastly, multi-objective property prediction is essential for practical applications. This section highlights the role of multi-task learning and Pareto optimization in enabling multi-property optimization of ILs in complex scenarios. In conclusion, this section provides valuable technical guidance and theoretical insights for the precise design and multi-objective optimization of ILs in the future.

6 Conclusions

In the context of the global challenge of climate change, ILs have garnered widespread attention in carbon capture research due to their customizable properties. The integration of ML methods has significantly accelerated the design and screening of IL solvents, providing a data-driven approach to overcome the inefficiencies of traditional experimental and theoretical methods. This review offers an in-depth and comprehensive analysis of the current progress in ML applications for ILs design

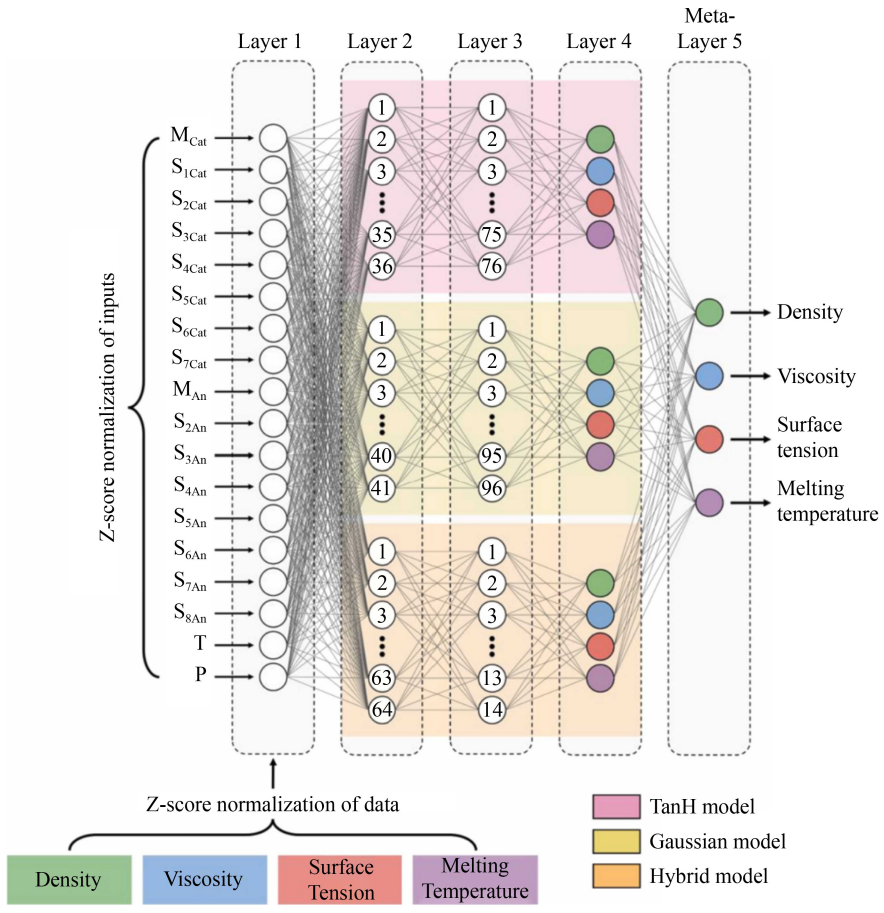

Fig. 27 Architectural schematic of the ensemble of ANNs depicting the TanH model, Gaussian model, hybrid model, and the meta-layer with a total of 11241 parameters (adapted from Lemaoui et al. [79] under the terms of CC BY-NC-ND license).

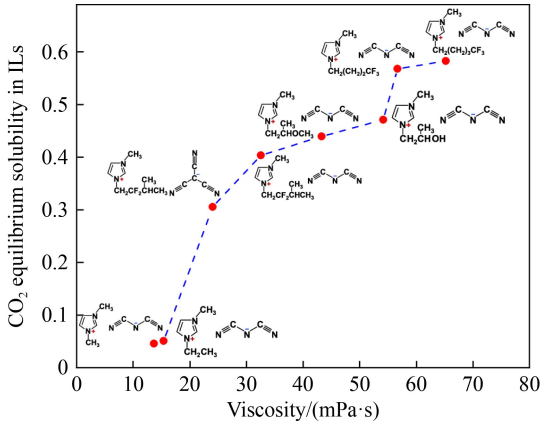

Fig. 28 Pareto-optimal curve representing the trade-off between viscosity and CO_2 equilibrium solubility of ILs (adapted with permission from Zhang et al. [80], copyright 2021, American Chemical Society).

and screening, emphasizing the crucial role of ML in addressing the limitations of conventional approaches and advancing the field.

Recent studies on ILs using ML have increasingly moved away from solely relying on physical properties (e.g., molecular weight, acentric factor, critical temperature, and pressure) as input parameters for property prediction. Instead, the focus has shifted to modern QSAR and QSPR models, which leverage ML to effectively establish relationships between the molecular structures of ILs and their performance. This review highlights the applications of various structure-based descriptors in QSAR and QSPR methods, particularly in predicting key properties such as toxicity, bioactivity, viscosity, density, and gas solubility of ILs. Currently, ML-based research in the structure–property relationships of ILs has evolved from simple models to advanced and hybrid models, such as GNN [65] and VAE-MLP-PSO [81]. These advanced methods have demonstrated significant promise in capturing the nonlinear relationships between complex structures and properties.

This review examines the challenges faced in applying ML to ILs design and screening. It provides an overview of the advantages and limitations of different ML models when dealing with datasets of varying sizes and task complexities. For complex tasks, it recommends a stepwise strategy that not only enhances the reliability of prediction results but also mitigates the risk of overfitting. This approach offers clear guidance on model

selection for IL design across different application scenarios. To further address the challenge of interpreting ML "black-box" models, it introduces a ML interpretability approach that integrates MD and DFT. By incorporating physicochemical information obtained from theoretical calculations as descriptors, this approach provides insights into the role of key molecular structures in property prediction. It also enables an atomic-level explanation of intermolecular interactions in ILs, validating the physicochemical interpretability of ML models.

Finally, this review emphasizes the importance of multi-property prediction in the engineering design of ILs. In applications like post-combustion CO_2 capture, ILs used as absorbents must exhibit low toxicity, low viscosity, and, high CO_2 solubility. To achieve the simultaneous optimization of multiple properties, it recommends using multi-task learning, Pareto optimization, and ensemble learning methods. These approaches provide comprehensive theoretical and practical framework for the targeted design of novel ILs.

In conclusion, ML methods hold great promise in the field of ILs research, with the potential to achieve significant breakthroughs in solvent screening and property prediction. By addressing critical challenges related to model selection, data availability, and interpretability, ML can substantially accelerate the development of ILs as sustainable and multifunctional solvents. This, in turn, will contribute to the transformation of global sustainable energy and chemical industries.

Acknowledgements This work was supported by the "Carbon Upcycling Project for Platform Chemicals" (Grant Nos. 2022M3J3A1045999 and 2022M3J3A1039377) through the National Research Foundation (NRF) funded of the Ministry of Science and ICT, Republic of Korea; the Natural Science Foundation of Jiangsu Province, China (Grant Nos. BZ2023051, BK20200694, and BK20240546); the Science and Technology Project of Changzhou, China (Grant No. CJ20241053), and the Jiangsu Specially-Appointed Professors Program, China.

Competing Interests The authors decalre that they hav no competing interests.

References

1. Erythropel H C, Zimmerman J B, de Winter T M, et al. The Green ChemisTREE: 20 years after taking root with the 12 principles. Green Chemistry, 2018, 20(9): 1929–1961
2. Shah P, Parikh S, Shah M, et al. A holistic review on application of green solvents and replacement study for conventional solvents. Biomass Conversion and Biorefinery, 2022, 12(5): 1985–1999
3. Ghandi K. A review of ionic liquids, their limits and applications. Green and Sustainable Chemistry. 2014, 4(1): 44–45
4. Zeng S, Zhang X, Bai L, et al. Ionic-liquid-based CO_2 capture systems: Structure, interaction and process. Chemical Reviews, 2017, 117(14): 9625–9673
5. Wang M, Zhang L, Liu H, et al. Studies on CO_2 absorption performance by imidazole-based ionic liquid mixtures. Journal of Fuel Chemistry & Technology, 2012, 40(10): 1264–1268
6. Liu F, Shen Y, Shen L, et al. Novel amino-functionalized ionic liquid/organic solvent with low viscosity for CO_2 capture. Environmental Science & Technology, 2020, 54(6): 3520–3529
7. Pârvulescu V I, Hardacre C. Catalysis in ionic liquids. Chemical Reviews, 2007, 107(6): 2615–2665
8. Guerfi A, Dontigny M, Charest P, et al. Improved electrolytes for Li-ion batteries: Mixtures of ionic liquid and organic electrolyte with enhanced safety and electrochemical performance. Journal of Power Sources, 2010, 195(3): 845–852
9. Rogers R D, Seddon K R. Ionic liquids–solvents of the future? Science, 2003, 302(5646): 792–793
10. Gardas R L, Coutinho J A. A group contribution method for viscosity estimation of ionic liquids. Fluid Phase Equilibria, 2008, 266(1–2): 195–201
11. Kolbeck C, Lehmann J, Lovelock K, et al. Density and surface tension of ionic liquids. Journal of Physical Chemistry B, 2010, 114(51): 17025–17036
12. Wang Y, Jiang W, Yan T, et al. Understanding ionic liquids through atomistic and coarse-grained molecular dynamics simulations. Accounts of Chemical Research, 2007, 40(11): 1193–1199
13. Kroon M C, Buijs W, Peters C J, et al. Quantum chemical aided prediction of the thermal decomposition mechanisms and temperatures of ionic liquids. Thermochimica Acta, 2007, 465(1–2): 40–47
14. Lei Z, Zhang J, Li Q, et al. UNIFAC model for ionic liquids. Industrial & Engineering Chemistry Research, 2009, 48(5): 2697–2704
15. Seybold P G, May M, Bagal U A. Molecular structure: Property relationships. Journal of Chemical Education, 1987, 64(7): 575
16. Hansch C, Fujita T. p-σ-π Analysis. A method for the correlation of biological activity and chemical structure. Journal of the American Chemical Society, 1964, 86(8): 1616–1626
17. Sun Y, Chen M, Zhao Y, et al. Machine learning assisted QSPR model for prediction of ionic liquid's refractive index and viscosity: The effect of representations of ionic liquid and ensemble model development. Journal of Molecular Liquids, 2021, 333: 115970
18. Zou J, Han Y, So S S. Overview of artificial neural networks. In: Livingstone D J, ed. Artificial Neural Networks: Methods and Applications. Totowa, NJ: Humana Press, 2009
19. Agatonovic-Kustrin S, Beresford R. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. Journal of Pharmaceutical and Biomedical Analysis, 2000, 22(5): 717–727
20. Ardeshiri A, Rashidi H. Performance of water-lean solvent for postcombustion carbon dioxide capture in a process-intensified absorber: Experimental, modeling, and optimization using RSM and ML. Industrial & Engineering Chemistry Research, 2023, 62(48): 20821–20832
21. Zhu X, Khosravi M, Vaferi B, et al. Application of machine learning methods for estimating and comparing the sulfur dioxide

absorption capacity of a variety of deep eutectic solvents. Journal of Cleaner Production, 2022, 363: 132465
22. Zhang J, Wang Q, Su Y, et al. An accurate and interpretable deep learning model for environmental properties prediction using hybrid molecular representations. AIChE Journal, 2022, 68(6): e17634
23. Bischl B, Binder M, Lang M, et al. Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery, 2023, 13(2): e1484
24. Goodfellow I. Deep Learning. Cambridge: MIT Press, 2016
25. Gu J, Wang Z, Kuen J, et al. Recent advances in convolutional neural networks. Pattern Recognition, 2018, 77: 354–377
26. Bahmaninia H, Shateri M, Atashrouz S, et al. Predicting the equilibrium solubility of CO_2 in alcohols, ketones, and glycol ethers: Application of ensemble learning and deep learning approaches. Fluid Phase Equilibria, 2023, 567: 113712
27. Chen G, Song Z, Qi Z. Transformer-convolutional neural network for surface charge density profile prediction: Enabling high-throughput solvent screening with COSMO-SAC. Chemical Engineering Science, 2021, 246: 117002
28. Bouzida Y, Cuppens F. Neural networks vs. decision trees for intrusion detection. In: IEEE/IST Workshop on Monitoring, Attack Detection and Mitigation (MonAM), 28:29
29. Kabari L, Nwachukwu E. Decision support system using decision tree and neural networks. Computer Engineering and Intelligent Systems., 2013, 4(7): 8–20
30. Nakhaei-Kohani R, Amiri-Ramsheh B, Pourmahdi M, et al. Extensive data analysis and modelling of carbon dioxide solubility in ionic liquids using chemical structure-based ensemble learning approaches. Fluid Phase Equilibria, 2024, 585: 114166
31. Yin G, Jameel Ibrahim Alazzawi F, Bokov D, et al. Multiple machine learning models for prediction of CO_2 solubility in potassium and sodium based amino acid salt solutions. Arabian Journal of Chemistry, 2022, 15(3): 103608
32. Huwaimel B, Alharby T N. Development of computational intelligence models for assessment of drug nanonization using green chemistry technique: Improvement of drug solubility. Case Studies in Thermal Engineering, 2023, 45: 103005
33. Wang J, Song Z, Chen L, et al. Prediction of CO_2 solubility in deep eutectic solvents using random forest model based on COSMO-RS-derived descriptors. Green Chemical Engineering., 2021, 2(4): 431–440
34. Wang X, Chan C W, Li T. High accuracy prediction of the post-combustion carbon capture process parameters using the Decision Forest approach. Chemical Engineering Science, 2024, 290: 119878
35. Kataoka T, Hao Y, Hung Y C, et al. Prediction of biphasic separation in CO_2 absorption using a molecular surface information-based machine learning model. Environmental Science. Processes & Impacts, 2022, 24(12): 2409–2418
36. Huang M W, Chen C W, Lin W C, et al. SVM and SVM ensembles in breast cancer prediction. PLoS One, 2017, 12(1): e0161501
37. Abdollahzadeh M, Khosravi M, Hajipour Khire Masjidi B, et al. Estimating the density of deep eutectic solvents applying supervised machine learning techniques. Scientific Reports, 2022, 12(1): 4954
38. Boobier S, Hose D R J, Blacker A J, et al. Machine learning with physicochemical relationships: Solubility prediction in organic solvents and water. Nature Communications, 2020, 11(1): 5753
39. Schulz E, Speekenbrink M, Krause A. A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. Journal of Mathematical Psychology, 2018, 85: 1–16
40. Deringer V L, Bartók A P, Bernstein N, et al. Gaussian process regression for materials and molecules. Chemical Reviews, 2021, 121(16): 10073–10141
41. Toots K M, Sild S, Leis J, et al. Machine learning quantitative structure–property relationships as a function of ionic liquid cations for the gas-ionic liquid partition coefficient of hydrocarbons. International Journal of Molecular Sciences, 2022, 23(14): 7534
42. Zhou J, Liu C, Ren J, et al. Targeting carbon-neutral waste reduction: Novel process design, modelling and optimization for converting medical waste into hydrogen. Energy, 2024, 310: 133272
43. Wilkes J S, Zaworotko M J. Air and water stable 1-ethyl-3-methylimidazolium based ionic liquids. Journal of the Chemical Society. Chemical Communications, 1992(13): 965–967
44. Zhang S, Sun N, He X, et al. Physical properties of ionic liquids: Database and evaluation. Journal of Physical and Chemical Reference Data, 2006, 35(4): 1475–1517
45. Shukla M, Saha S. Relationship between stabilization energy and thermophysical properties of different imidazolium ionic liquids: DFT studies. Computational & Theoretical Chemistry, 2013, 1015: 27–33
46. Klahn M, Seduraman A. What determines CO_2 solubility in ionic liquids? A molecular simulation study. Journal of Physical Chemistry B, 2015, 119(31): 10066–10078
47. del Rio B G, Phan B, Ramprasad R. A deep learning framework to emulate density functional theory. npj Computational Materials. 2023, 9(1): 158
48. Zhong S, Zhang K, Bagheri M, et al. Machine learning: New ideas and tools in environmental science and engineering. Environmental Science & Technology, 2021, 55(19): 12741–12754
49. Katritzky A R, Lobanov V S, Karelson M. QSPR: The correlation and quantitative prediction of chemical and physical properties from structure. Chemical Society Reviews, 1995, 24(4): 279–287
50. Chen Y, Kontogeorgis G M, Woodley J M. Group contribution based estimation method for properties of ionic liquids. Industrial & Engineering Chemistry Research, 2019, 58(10): 4277–4292
51. Ding Y, Chen M, Guo C, et al. Molecular fingerprint-based machine learning assisted QSAR model development for prediction of ionic liquid properties. Journal of Molecular Liquids, 2021, 326: 115212
52. Venkatraman V, Evjen S, Lethesh K C, et al. Rapid, comprehensive screening of ionic liquids towards sustainable applications. Sustainable Energy & Fuels, 2019, 3(10): 2798–2808
53. Peric B, Sierra J, Marti E, et al. Quantitative structure-activity

relationship (QSAR) prediction of (eco)toxicity of short aliphatic protic ionic liquids. Ecotoxicology and Environmental Safety, 2015, 115: 257–262

54. Zhu P, Kang X, Zhao Y, et al. Predicting the toxicity of ionic liquids toward acetylcholinesterase enzymes using novel QSAR models. International Journal of Molecular Sciences, 2019, 20(9): 2186
55. Wu X, Gong J, Ren S, et al. A machine learning-based QSAR model reveals important molecular features for understanding the potential inhibition mechanism of ionic liquids to acetylcholinesterase. Science of the Total Environment, 2024, 915: 169974
56. Hodyna D, Kovalishyn V, Rogalsky S, et al. Antibacterial activity of imidazolium-based ionic liquids investigated by QSAR modeling and experimental studies. Chemical Biology & Drug Design, 2016, 88(3): 422–433
57. Carrera G V S M, Nunes da Ponte M. Machine-learning approaches to tune descriptors and predict the viscosities of ionic liquids and their mixtures. Chemistry Methods, 2020, 1(5): 214–223
58. Chen Y, Peng B, Kontogeorgis G M, et al. Machine learning for the prediction of viscosity of ionic liquid-water mixtures. Journal of Molecular Liquids, 2022, 350: 118546
59. Huang M, Deng J, Jia G. Predicting viscosity of ionic liquids-water mixtures by bridging UNIFAC modeling with interpretable machine learning. Journal of Molecular Liquids, 2023, 383: 122059
60. Acar Z, Nguyen P, Cui X, et al. Room temperature ionic liquids viscosity prediction from deep-learning models. Energy Materials, 2023, 3: 300039
61. Baskin I, Epshtein A, Ein-Eli Y. Benchmarking machine learning methods for modeling physical properties of ionic liquids. Journal of Molecular Liquids, 2022, 351: 118616
62. Liu X, Gao J, Chen Y, et al. Machine learning-assisted modeling study on the density and heat capacity of ionic liquid-organic solvent binary systems. Journal of Molecular Liquids, 2023, 390: 122972
63. Kuroki N, Suzuki Y, Kodama D, et al. Machine learning-boosted design of ionic liquids for CO_2 absorption and experimental verification. Journal of Physical Chemistry B, 2023, 127(9): 2022–2027
64. Song Z, Shi H, Zhang X, et al. Prediction of CO_2 solubility in ionic liquids using machine learning methods. Chemical Engineering Science, 2020, 223: 115752
65. Jian Y, Wang Y, Farimani A B. Predicting CO_2 absorption in ionic liquids with molecular descriptors and explainable graph neural networks. ACS Sustainable Chemistry & Engineering, 2022, 10(50): 16681–16691
66. Tian Y, Wang X, Liu Y, et al. Prediction of CO_2 and N_2 solubility in ionic liquids using a combination of ionic fragments contribution and machine learning methods. Journal of Molecular Liquids, 2023, 383: 112066
67. Yang A, Sun S, Su Y, et al. Insight to the prediction of CO_2 solubility in ionic liquids based on the interpretable machine learning model. Chemical Engineering Science, 2024, 297: 120266
68. Zhong S, Chen Y, Li J, et al. Screening environmentally benign ionic liquids for CO_2 absorption using representation uncertainty-based machine learning. Environmental Science & Technology Letters, 2024, 11(11): 1193–1199
69. Mousavi S P, Atashrouz S, Nakhaei-Kohani R, et al. Modeling of H_2S solubility in ionic liquids using deep learning: A chemical structure-based approach. Journal of Molecular Liquids, 2022, 351: 118418
70. Mousavi S P, Nakhaei-Kohani R, Atashrouz S, et al. Modeling of H_2S solubility in ionic liquids: Comparison of white-box machine learning, deep learning and ensemble learning approaches. Scientific Reports, 2023, 13(1): 7946
71. Liu T, Dong Z, Zhu W, et al. 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 Sustainable Chemistry & Engineering, 2023, 11(9): 3917–3931
72. Hansch C. The physicochemical approach to drug design and discovery (QSAR). Drug Development Research, 1981, 1(4): 267–309
73. Luis P, Ortiz I, Aldaco R, et al. A novel group contribution method in the development of a QSAR for predicting the toxicity (Vibrio fischeri EC50) of ionic liquids. Ecotoxicology and Environmental Safety, 2007, 67(3): 423–429
74. Mohan M, Smith M D, Demerdash O, et al. Predictive understanding of the surface tension and velocity of sound in ionic liquids using machine learning. Journal of Chemical Physics, 2023, 158(21): 214502
75. Mazari S A, Siyal A R, Solangi N H, et al. Prediction of thermo-physical properties of 1-butyl-3-methylimidazolium hexafluoro-phosphate for CO_2 capture using machine learning models. Journal of Molecular Liquids, 2021, 327: 114785
76. Lundberg S M, Lee S I. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 2017, 30: 4765–4774
77. Li J, Dong S, An B, et al. Machine learning for the yield prediction of CO_2 cyclization reaction catalyzed by the ionic liquids. Fuel, 2023, 335: 126942
78. Li J, Qi X, Zhang Z, et al. Screening of ionic liquids for efficient CO_2 cycloaddition catalysis under mild condition: A combined machine learning and DFT approach. ACS Sustainable Chemistry & Engineering, 2024, 12(48): 17512–17522
79. Lemaoui T, Eid T, Darwish A S, et al. Revolutionizing inverse design of ionic liquids through the multi-property prediction of over 300,000 novel variants using ensemble deep learning. Materials Science and Engineering R Reports, 2024, 159: 100798
80. Zhang X, Wang J, Song Z, et al. Data-driven ionic liquid design for CO_2 capture: Molecular structure optimization and DFT verification. Industrial & Engineering Chemistry Research, 2021, 60(27): 9992–10000
81. Liu X, Chu J, Zhang Z, et al. Data-driven multi-objective molecular design of ionic liquid with high generation efficiency on small dataset. Materials & Design, 2022, 220: 110888