Received: 11 April 2023 | Revised: 11 December 2023 | Accepted: 1 January 2024

DOI: 10.1002/aic.18363

AIChE JOURNAL

RESEARCH ARTICLE

Separations: Materials, Devices and Processes

Feature analysis of generic AI models for CO_2 equilibrium solubility into amines systems

Ting Lan[1] | Shoulong Dong[1] | Hui Luo[1] | Liju Bai[1] | Helei Liu[1,2]

[1]International Innovation Institute of Carbon Capture and Utilization (I3CCU), School of Chemistry and Chemical Engineering, Beijing Institute of Technology, Beijing, China

[2]Department of Chemical and Biological Engineering, The University of British Columbia, Vancouver, British Columbia, Canada

Correspondence
Helei Liu, International Innovation Institute of Carbon Capture and Utilization (I3CCU), School of Chemistry and Chemical Engineering, Beijing Institute of Technology, Beijing 102488, China.
Email: lhl0925@hotmail.com, hl_liu@bit.edu.cn

Funding information
Beijing Institute of Technology, Grant/Award Number: 2022CX01004; National Natural Science Foundation of China, Grant/Award Number: 22208017

Abstract

Reported models have disadvantages such as poor prediction accuracy and time-consuming. And they can not reflect the impact of chemical reactions on CO_2 solubility. To compensate for these deficiencies, parameters representing operational parameters, physical properties, chemical properties, and molecular properties are introduced as input variables. A series of models are constructed by three algorithms: back propagation neural network, radial basis function neural network, and random forest. The model with the best prediction performance is level OPCM (RBFNN), with the AARE of only 1.52%. By ranking the importance of the features using the RF algorithm, P_{CO2}, was found to be the key parameter affecting the CO_2 loadings, with M being the least important. Using the screened key parameters to model the model, as well as optimizing the structure, can further improve the predictive performance of the model. The full process development and optimization model framework constructed in this article can provide practical guidance for the development of machine learning models.

KEYWORDS
amines, artificial intelligence (AI) models, CO_2 capture, feature analysis, solubility

1 | INTRODUCTION

The Paris Agreement states that the global temperature will increase by 2°C above pre-industrial levels due to a continuous increase in temperature and that countries around the world should aim to limit global warming to 1.5°C.[1] The main reason for this phenomenon is due to the entry of greenhouse gases (especially carbon dioxide [CO_2]) into the atmosphere. Excessive CO_2 emissions can cause a rise in global temperatures, sea level rise, and frequent extreme weather.[2] Therefore, it is urgent to research and develop an effective way to reduce CO_2 emissions into the atmosphere. At present, post-combustion CO_2 capture (PCC) technology is the most mature means to reduce man-made CO_2 emissions from coal-fired power plants, the steel industry, and other industrial plants.[3,4]

PCC technology for removing CO_2 from flue gas usually includes membrane separation, physical adsorption, chemical absorption, micro-algae bio-fixation, and so on.[5,6] Among them, amine-based chemical absorption technology is the most widely used, which is applies to coal-fired flue gas with large volume and low CO_2 partial pressure. It has the advantages of high selectivity for CO_2, fast absorption rate, outstanding economic benefit, and mature operation.[7] However, there are drawbacks such as high energy consumption and equipment corrosion due to amine solvents in the regeneration process.[8,9] Therefore, there is an urgent need to develop new amine solvents with high absorption capacity and reaction rate, low corrosiveness, and low energy consumption for regeneration.[10–12]

As one of the important parameters for evaluating the performance of absorbents, the CO_2 equilibrium solubility has received extensive attention during carbon capture. Because it plays an important role in improving the absorption performance of absorbers and developing vapor–liquid equilibrium theoretical models.[13] The traditional method to obtain CO_2 equilibrium solubility is to carry out a group of experiments and calculate parameters based on experimental data. The experimental complex process is time-consuming and

https://doi.org/10.1002/aic.18363

TABLE 1 Typical thermodynamic models reported in the literature.

Models	Gas	Amine	Key point	Reference
Kent–Eisenberg	CO_2 H_2S	MEA, DEA	The model is simple, the equilibrium constant is only a function of temperature, and the error in predicting high or low load is large.	Kent and Eisenberg[24]
		HMDA + AMP	Equilibrium constant as a function of temperature and CO_2 loading to predict CO_2 solubility with an accuracy of 7.2% for the ARD.	Mondal et al.[25]
		MAE + AEEA	The correction factor (F_k) introduced for the KE model was improved with an ARD of 4.17%.	Pandey and Mondal[26]
		MCA, CHAP	The equation is no longer calculated iteratively, but the hydrogen ion concentration is estimated by polynomial equation, with ARD of 5.7% for MCA and 2.9% for CHAP.	Tzirakis et al.[27]
Hu–Chakma	CO_2 H_2S	AMP	The equilibrium constants depend on the physical and free amine concentrations in the liquid phase and the temperature. However, the authors are theoretically wrong for the reaction of spatially site-resistant amines.	Hu and Chakma[28]
		N-(2-HE)PRLD	In the vicinity of 0.4–0.6 CO_2 solubility, the deviation is very large, and the overall prediction effect is good, with ARD of 4.3%.	Zheng et al.[29]
Li–Shen	CO_2	MEA, MDEA, and their mixture	The equilibrium constant is divided into two parts based on activity coefficient and temperature. However, the correction factor for activity coefficient is not accurate.	Li and Shen[30]
		1DEA2P	The prediction error was large, with an ARD of 12.2%.	Liu et al.[31]
Deshmukh–Mather	CO_2 H_2S	MEA	The extended Debye-Huckel theory is used to calculate the activity coefficient, which is accurate and convenient to describe the amine equilibrium of a single acid gas.	Deshmukh and Mather[32]
		1DMA2P	The short-range interaction parameters between ion-molecule and ion-ion in the activity model are a function of temperature with an ARD of 2.64%.	Afkhamipour and Mofarahi[33]
C_f	CO_2	MPDL	A correction factor (C_f) was applied to the non-idealities of the amine solution, corrected for an ARD of 2%, but the model is less generic.	Xiao et al.[34]
		2DMA2M1P	The C_f model was modified by introducing Δp to reflect the correction for pressure, with an ARD of 2.08%.	Zhang et al.[35]
Helei–Liu	CO_2	1DMA2P	The new model developed by adding the total pressure to the Hu-Chakma model. The applicability is wider, but the limitation of the experimental results of different total pressures may lead to the poor correlation of the new model, ARD of 10%.	Liu et al.[36]
e-NRTL	CO_2, H_2S	MEA, DEA, MDEA/PZ	The e-NRTL activity coefficient model consisted of short-range, long-range, and solvation forces. Nevertheless, the parameters of the binary interactions are different from those of the ternary interactions, leading to the discontinuity of the model.	Moioli and Pellegrini[15] Chen et al.[37–40]
		DPTA	The experimental VLE data were simulated by regressing various interaction parameters (molecule-molecule and molecule-ion pairs) and equilibrium constants for amine deprotonation, carbamate formation, bicarbonate formation, and zwitterion deprotonation reactions. The ARD was 22.3%.	Agarwal et al.[41]

Note: ARD is the average absolute deviation.
Abbreviations: 1DMA2P, 1-dimethylamino-2-propanol; 2DMA2M1P, 2-(dimethylamino)-2-methyl-1-propanol; AEEA, aminoethylethanolamine; AMP, 2-amino-2-methyl-1-propanol; CHAP, N–cyclohexyl-1,3-propanediamine; DEA, diethanolamine; DPTA,dipropylenetriamine; HMDA, hexamethylenediamine; MCA, N-methylcyclohexyamine; MDEA, N-methyldiethanoamine; MEA, monoethanolamine; MPDL, N-methyl-4-piperidinol; N-(2-HE)PRLD, N-(2-hydroxyethyl) pyrrolidine; TETA, Triethylenetetramine.

laborious. However, the established thermodynamic models can be used to analyze and correlate the equilibrium solubility of CO_2 under different conditions using a limited number of experimental data points. These models can generate large amounts of data as data sources for designing and optimizing CO_2 capture, saving time and cost. Table 1 lists the typical thermodynamic models reported in the references. Researchers were committed to improving the Kent–Eisenberg model. They developed a series of models, such as the Hu–Chakma model, making so much effort to improve the sensitivity of the model. However, the accuracy still needs to be higher.

Besides, the scope of application is also limited. Meanwhile the e-NRTL proposed by Chen et al. has received a great deal of attention as one of the most accurate thermodynamic equilibria. The e-NRTL model provides a rigorous thermodynamic framework and adjustable model parameters to correlate experimental data and to interpolate, extrapolate, and predict thermodynamic properties and gas–liquid phase equilibria of multicomponent electrolyte systems.[14] Scholars have used the e-NRTL model to thermodynamically analyze the system (CO_2 + Amine + H_2O). For example, Moioli and Pellegrini calculated and regressed the binary interaction parameters of the e-NRTL model.[15] The effect of pressure on CO_2 solubility at different temperatures in the (MDEA + PZ) system is described. Zhang et al. analyzed the (MEA + CO_2 + H_2O) system and found that the model results were slightly lower than the experimental data at high loadings and high MEA concentrations.[16] Although the model has wider applicability, the regression of model parameters is laborious and tedious for complex electrolyte solution systems. And the use of the e-NRTL model is less satisfactory for electrolyte systems with strongly hydrated high-charge density ions. There is an urgent need to develop a new model to solve the above problems.

Machine learning (ML) has been applied to the carbon capture field and can develop solubility prediction models with superior performance. It can make up for the shortcomings of the traditional thermodynamic model.[17] Artificial neural network, as a branch of ML, can reflect the complexity of the system and the internal relationship between the inputs and outputs with high confidence and accuracy. Abooali et al. developed a model to predict the CO_2 solubility of DEA, MDEA and their blended solutions using both genetic programming (GP) and stochastic gradient boosting (SGB) algorithms, where average absolute relative deviation (ARD_{SGB}) was 0.95628%.[18] Dashti et al. used four neural networks (genetic algorithm-adaptive neuro-fuzzy inference system [GA-ANFIS], particle swarm optimization ANFIS [PSO-ANFIS], coupled simulated annealing-least squares support vector machine [CSA-LSSVM], and radial basis function [RBF]) to predict the equilibrium CO_2 absorption capacity in 12 amine solutions, with the LSSVM model having the best accuracy of mean square error (MSE) and average regression coefficient (R^2) of 2% and 0.9338, respectively.[19] Liu et al. proposed a new XGBoost ML model to predict the solubility of CO_2 in aqueous solutions of three amines with average absolute relative error (AAREs) of 3.77%, 0.29%, and 0.70%, respectively.[20] Moreover, Tellagorla et al. studied the (TAEA + MDEA) and (TAEA + AMP) systems and built feedforward neural network models with average absolute deviation (AAD) of 1.538% and 2.696%, respectively.[21]

The developed model based on ML mentioned above obtained some outstanding prediction results. However, the output of the model was difficult to achieve satisfactory accuracy due to insufficient consideration of the chemical reaction process affecting the CO_2 solubility. Meanwhile, it did not provide a reasonable selection method of the input parameters to provide judgment for optimizing the model. Therefore, this work aims to develop generic models with wide applicability, excellent prediction accuracy, and strong reaction background. For such a goal, the experimental data of CO_2 equilibrium solubility of 17 different types of amine solutions were selected. Furthermore, three algorithms of back propagation neural network (BPNN), radial basis function neural network (RBFNN), and random forest (RF) were used to construct ML models at different levels. Firstly, the model for the adsorption of CO_2 in specific amines was developed. The performance of the model was also evaluated by the average absolute relative error (AARE), root mean square error (RMSE), and correlation coefficient (R). Second, generic Artificial intelligence (AI) models for the CO_2 equilibrium solubility of amine solutions were established by considering the operating conditions, physical properties, chemical properties, molecular characteristics, and so on. Finally, this article utilized the RF algorithm to rank feature importance to obtain key influence factors. This process could provide theoretical guidance for subsequent screening of parameters and optimization of the AI model. The developed model balanced fitness and accuracy, providing a preliminary selection and evaluation of the CO_2 equilibrium solubility of potential amine solvents under various conditions.

2 | ESTABLISHMENT OF AI MODEL

In this work, there are four steps to build AI models, as shown in Figure 1. It includes input parameters selection, data collection, model building, and model evaluation.

2.1 | Input parameters selection

The process of equilibrium is complex in all cases, including chemical reactions, physical diffusion, and vapor–liquid equilibrium. All parameters that affect the CO_2 equilibrium solubility should be considered in the modeling.

The operational parameters, such as CO_2 partial pressure (P_{CO2}), temperature (T), and concentration (C), can affect the CO_2 absorption of amine solutions, which is used to predict the CO_2 loading in the previous model. However, it can only represent the operation process and cannot give the basic theoretical relationship. Therefore, the previous model only considers these three parameters, which cannot fully reflect the influence factors of CO_2 solubility. Moreover, physical diffusion is another influencing factor, which is related to viscosity (η), density (ρ), and molecular weight (M). With different amine solutions, some molecular and chemical properties of amines also have a very strong influence on the whole process of the reaction, including the dipole moment (μ) and the dissociation coefficient (pKa).

These factors were divided into different types of input parameters. Extrinsic operational factors (P_{CO2}, T, and C) that affect CO_2 solubility were used as a set of input descriptors. There is no doubt that the concentration of amine can significantly affect the absorption of CO_2. Similarly, higher temperatures lead to faster molecular motion, which also promotes CO_2 absorption into the amine solutions. Increasing CO_2 partial pressure also promotes the number of CO_2 molecules entering the solvent. Meanwhile, physical properties that significantly affect the

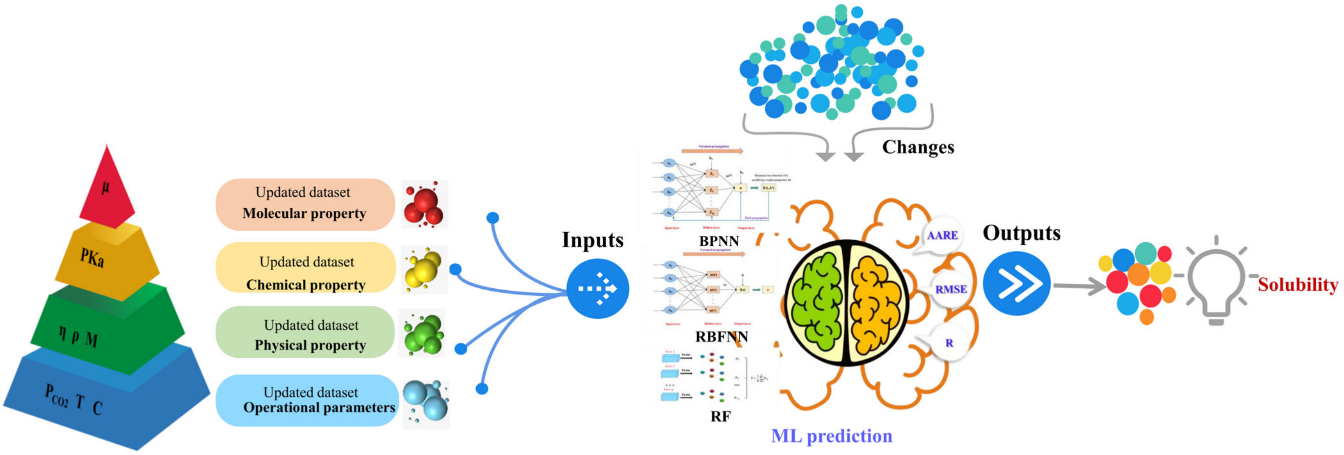


FIGURE 1 Model development flow chart.

diffusion of CO_2 molecules are treated as another set of input variables. In addition, since the aqueous amine solution is alkaline and CO_2 is an acidic gas that dissolves more easily in a more alkaline environment, the chemical nature of the aqueous amine solution (*pKa*) has the same effect on the overall performance of the absorption process, *pKa* was used as the third set of input parameters. Finally, an extra parameter is introduced to describe the nature of the molecule, and according to the principle of similarity of phase solubility, dipole moment also affects the solubility of CO_2 in amine-water solvents.

As shown in Figure 1, the selected parameters were classified into four groups according to their properties and combined with each other to form different levels of the model. The model containing only operating parameters (P_{CO2}, T, and C) as input descriptors is named level O. Besides these three parameters, parameters representing amine chemistry (*pKa*), physical properties (η, ρ, M), and amine molecular structure were added to form level OC, level OP, and level OM, respectively. To further improve the model, the model with seven parameters (P_{CO2}, T, C, η, ρ, M, and *pKa*) will be called level OPC. However, the parameters considered in the model created above are not comprehensive enough to describe the reaction at the molecular level, so on the basis of level OPC, the parameter μ was further introduced to constitute level OPCM.

In brief, in this work, the models and the generic model were built on the basis of different levels of parameters, involving the operating conditions, chemical properties, physical properties, and molecular properties.

2.2 | Data collection and models building

The more uniformly distributed data and the more data inputs will make the model training better and make the models more reliable. The data should be able to reflect the underlying nonlinearities, complexities, and intricacies of the targeted system behavior. In this work, 17 amine-based systems were considered, including nine single amine systems and eight blended amine systems, involving several amine classes (i.e., primary, secondary, and tertiary amines), as shown in Table 2.

The experimental data used in this article was from published reference, the solubility covers the range of 0.06–1.23 (mol CO_2/mol amine) at a certain P_{CO2}, T, and C. Simultaneously, the data was collected for ρ, η, and M, as well as *pKa* and μ for different amine molecules. In order to obtain a large number of experimental data points for the model to be adequately trained, the same interpolation method as in the literature[22] was used to generate more training data, so 2049 data points were obtained.

The calculation of M, ρ, and η for the blended amine system applied the mixing rule and used the following Equations (1)–(3). X_A is the mass fraction of A and X_B is the mass fraction of B.

$$M_{mix} = X_A M_A + X_B M_B \tag{1}$$

$$\frac{1}{\rho_{mix}} = \frac{X_A}{\rho_A} + \frac{X_B}{\rho_B} \tag{2}$$

$$\eta_{mix}^{1/3} = X_A \eta_A^{1/3} + X_B \eta_B^{1/3} \tag{3}$$

The BPNN is one of the most widely used neural network models in chemical applications. The RBFNN has a simple network structure and good nonlinear fitting ability. Furthermore, the RF has good resistance to overfitting and was chosen as the modeling method for predicting the equilibrium solubility of CO_2.[23] In the process of building the model, the input data were divided into a train set and a test set with a ratio of 70% and 30%. Such an allocation can be a good way to train the neural network, avoid overfitting, and achieve the optimal network structure, as well as a good way to compromise the robustness and accuracy of the established model.

2.3 | Model evaluation

Three evaluation metrics, namely RMSE, AARE, and R, were used to evaluate the prediction performance of the established BPNN, RBFNN, and RF models. The models could correlate the chosen input parameters and the experimental solubility values and reflect the

TABLE 2 Data ranges collected from the literature for 17 amine-based systems.

No.	Amine	Solvent concentration (mol/L)	Temperature(K)	Pressure(kPa)	Viscosity (mPa·s)	Density (g/mL)	Molecular weight (g/mol)	pKa	Dipole moment(μ)	Solubility (mol/mol)	Reference
1	1DMA2P	2–4	298–333	3.0–121.4	1.33	0.85	103.16	9.20	2.06	0.29–0.97	Afkhamipour and Mofarahi[33]
2	AMP	2–3	293–353	1.6–98.9	99.47	0.93	89.14	9.68	2.95	0.13–0.96	Tontiwachwuthikul et al.[42]
3	DEEA	1–4	293–353	5.9–100.8	4.17	0.88	117.19	9.73	1.16	0.06–0.97	Luo et al.[43]
4	DMEA	2	298–313	8.3–61.1	3.39	0.88	89.14	9.20	2.56	0.57–0.86	Xiao et al.[44]
5	MDEA	2–4	303–323	0.1–98.2	77.32	1.04	119.16	8.54	2.86	0.01–0.88	Benamor and Aroua[45]
6	MEA	2–5	298–353	2.2–101.3	20.19	1.01	61.08	9.45	0.78	0.35–0.80	Shen and Meng[46] Maneeintr et al.[47]
7	PZ	0.1–1	293–323	0.9–95.6	0.97	0.72	86.14	9.71	1.47	0.10–1.23	Aroua and Salleh[48]
8	DETA	1.7–3.6	313–353	1.2–111.0	4.46	0.95	103.17	9.45	2.15	0.63–1.20	Chang et al.[49]
9	DEAB	1–5	298–333	9.0–100.0	3.02	0.86	145.24	10.20	1.65	0.52–1.03	Sema et al.[50]
10	MEA/DEEA	4.0	313.000	15.0–101.5	8.17–12.18	0.90–0.92	89.14–103.16	9.59–9.66	0.97–1.07	0.60–0.82	Xiao et al.[51]
11	1DMA2P/MEA	2.5	313–333	3.1–124.4	2.25–11.26	0.86–0.96	72.31–105.97	9.25–9.40	1.04–1.80	0.07–0.91	Afkhamipour et al.[52]
12	DETA/PZ	2.9–3.1	313–353	1.1–149.8	2.29–3.63	0.82–0.90	99.88	9.50–9.55	1.79–2.02	0.68–1.23	Chang et al.[49]
13	MEA/MDEA	3.4–3.5	313–373	0.9–145.4	48.44	1.03	86.33	9.05	1.68	0.20–0.68	Shen and Meng[46]
14	MEA/AMP	3.8–4.3	313–353	0.9–53.8	37.92–80.34	0.94–0.98	68.24–82.80	9.51–9.58	1.33–2.04	0.45–0.80	Park et al.[53]
15	MDEA/PZ	2.6	303–323	1.0–105.5	38.90–66.04	0.93–1.01	108.11–116.19	8.61–8.84	2.50–2.77	0.11–0.93	Dash et al.[54] Feron et al.[55]
16	AMP/PZ	2.5–4.5	313–353	0.97–139.9	29.87–60.45	0.84–0.88	87.85–88.39	9.68–9.69	2.32–2.65	0.34–0.88	Feron et al.[55] Yang et al.[56]
17	MEA/AMP/BEA	2.1–4.5	313–363	8.1–50.7	36.77–39.46	0.91–0.92	94.75–102.14	9.76–9.83	2.20–2.48	0.25–0.79	Li et al.[57]

Note: The M, ρ, and η of monoamine are derived from Aspen Plus®. The DEAB dipole moment is estimated.
Abbreviations: DEAB, 4-diethylamino-2-butanol; DEEA, N, N-diethylethanolamine; DETA, Diethylenetriamine; DMEA, N, N-dimethylethanolamine; MDEA, N-methyldiethanoamine; PZ, piperazine.

underlying relationships. Smaller values of RMSE and AARE are preferred, while the range of R is between 0 and 1, and R values close to 1 are favorable. The metrics were defined as follows.

$$\mathrm{AARE}=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{p_i-t_i}{t_i}\right|\times 100\%, \tag{4}$$

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{p_i-t_i}{t_i}\right)^2}\times 100\%, \tag{5}$$

$$R=\sqrt{1-\frac{\sum_{i=1}^{n}(p_i-t_i)^2}{\sum_{i=1}^{n}t_i^2}}, \tag{6}$$

where n is the size of data sets; p_i and t_i are the predicted and target values of the ith input data, respectively.

3 | RESULTS AND DISCUSSIONS

3.1 | Generic BPNN models for CO_2 solubility prediction

The existing models need to improve on issues such as needing to be more time-consuming, poor prediction accuracy, and narrow applicability. Moreover, they lack inherent characteristics because they do not reflect the effects caused by chemical reactions, physical diffusion, and chemical properties on CO_2 solubility.

To overcome these shortcomings, this article first developed a BPNN model containing only the operating parameters (P_{CO2}, T, and C) that could significantly affect the equilibrium solubility during CO_2 absorption into amine solutions. Subsequently, in order to optimize the model, the level OC model was created by introducing parameter (pKa) that represented the characteristic properties of the amine itself. At the same time, the parameter (μ) related to the molecular structure is also involved in the construction of the model as input parameters, which is the level OM model. Finally, physical parameters (η, ρ, and M) are also involved in the construction of six-parameter, seven-parameter, and eight-parameter models (level OP, level OPC, and level OPCM). The introduction of parameters reflecting different reaction characteristics in the modeling process can make the generic BPNN model more theoretically interpretable. And the prediction results for different amine systems are more accurate. The BPNN model is constructed by mainly changing the number of hidden layers and the number of hidden layer nodes.

3.1.1 | BPNN models of CO_2 equilibrium solubility based on operating parameters

In order to reasonably describe and predict the solubility, three operating parameters (P_{CO2}, T, and C) were selected as input parameters for the BPNN models. The prediction performance of these BPNN models and the number of hidden layer nodes used by the models are listed in Table 3. In Table 3, we can see that the AAREs of all systems are less than 4%, and 11 of the 17 systems have AAREs <1%. The results prove that the data predicted by the models fit well with the experimental data. In addition, the R values of all models are greater than 0.99, and all RMSE values are less than 13%. The results indicate that the developed BPNN models can be applied to predict the equilibrium solubility of CO_2.

Table 3 also lists the CO_2 solubility prediction error of seven correlations published in the references and compares them with our BPNNs. As shown in Figure 2, the AARE range of these loadings predicted by empirical or semi-empirical models is about 2.64–19.90. Compared with the data of BPNN models with AAREs less than 4%, it further proves that BPNN has significant advantages in predicting the solubility of CO_2 absorption process.

Figure 2 shows the AARE values of the BPNN models for 17 amines. This can be seen in Figure 2 that the prediction performance of MEA/DEEA, MEA/AMP and MEA/AMP/BEA models is the best, and their AAREs are less than 1%. However, the predictions of models based on specific amine solvents are already quite good. But, it is not enough to build a model with only three operating parameters for a larger range of predictions to be applicable. The input properties of the model can be better characterized by incorporating chemical and physical properties of the system as well as properties at the molecular level.

3.1.2 | Generic BPNN models based on different levels of CO_2 solubility

Even though the accuracy of the prediction models for individual systems was already good, they could not be applied to different amine systems. In order to more intuitively illustrate the introduction of parameters related to the reaction characteristics and to better describe the CO_2 absorption process, level O, level OC, level OM, level OP, level OPC, and level OPCM models were established. These models all use a uniform hidden layer containing 10 neurons to facilitate later comparisons.

Table 4 lists the evaluation indicators at different levels. For the level O model established only by operating parameters, AARE, RMSE, and R values are 21.31%, 79.08%, and 0.863, respectively. It means that the prediction accuracy of this model is poor. This deviation is thought to be probably due to the fact that the model only considers three operational parameters: P_{CO2}, T, and C. And does not take into account the differences between different amine systems. Therefore, in the subsequent model development, parameters that can represent the characteristic properties of the amine itself were introduced.

For the level OC model, P_{CO2}, T, C, and pKa were used, and the indexes were significantly improved compared with the level O model. It is confirmed that the absorption degree of CO_2 in amine solution is also affected by the acid dissociation coefficient pKa. The predictive performance of the model can be improved by introducing pKa parameters.

The metrics of level OM model with four parameters (P_{CO2}, T, C, and μ) have no great change compared with level OC. However, its

TABLE 3 Performance evaluation of BPNN models for 17 amine-based systems.

Amine	Data sets	Nodes number of hidden layer	AARE (%) BPNN	AARE (%) Published correlations	RMSE (%)	*R*
1DMA2P	122	10	1.18	2.64	3.57	0.99282
AMP	96	12	0.68	10.50	1.21	0.99941
DEEA	175	12	2.80	4.84	6.25	0.99634
DMEA	80	16	0.30	-	0.72	0.99709
MDEA	113	12	3.97	10.72	12.79	0.99877
MEA	99	12	0.87	4.10	2.03	0.99544
PZ	255	12	2.52	15.70	4.21	0.99920
DETA	95	12	0.28	-	0.64	0.99919
DEAB	104	12	0.52	7.10	0.85	0.99839
MEA/DEEA	48	7	0.01	-	0.02	1.00000
1DMA2P/MEA	175	14	1.33	-	3.18	0.99842
DETA/PZ	159	10	0.49	-	1.01	0.99742
MEA/MDEA	114	10	0.31	-	0.73	0.99959
MEA/AMP	31	12	0.08	-	0.14	0.99998
MDEA/PZ	118	12	1.62	8.11	4.01	0.99846
AMP/PZ	151	13	0.87	19.90	1.28	0.99827
MEA/AMP/BEA	114	10	0.25	-	0.36	0.99996

Abbreviation: AARE, average absolute relative error; BPNN, back propagation neural network; R, correlation coefficient; RMSE, root mean square error.

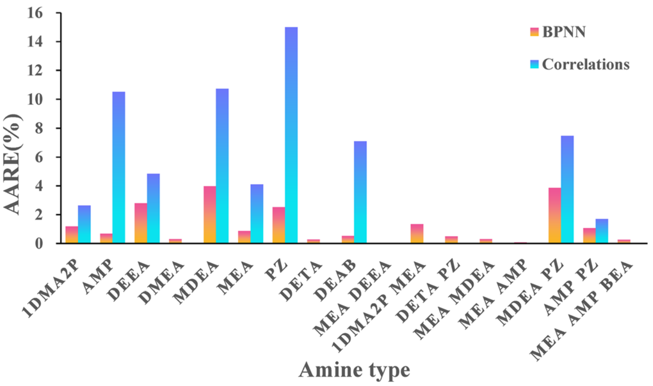


FIGURE 2 AARE values of BPNN for 17 amine based systems and comparison with values published in the literature. AARE, average absolute relative error; BPNN, back propagation neural network.

performance is obviously better than that of level O, indicating that the dipole moment related to molecular properties can also affect the loading of CO_2 in amine aqueous solution, and the introduction of parameter μ can improve the accuracy of the model.

The model performance is further optimized with the further increase of input parameters, which is reflected in level OP. The physical property parameters (η, ρ, and M) that also affect the solubility in the CO_2-Amine-H_2O system are added to the modeling, decreasing the AARE (7.02%) and RMSE (25.15%) of the level OP model. *R*-value (0.987) is closer to 1, and it can be seen from the index that the deviation between the predicted value and the experimental value becomes smaller.

In order to obtain a model with wider applicability and higher accuracy, the level OPC model was constructed. Compared with level OP, level OPC has only one more parameter (*pKa*), but its performance has been further improved. This phenomenon once again verifies that the introduction of appropriate parameters can improve the application range and prediction accuracy of the model.

For the level OPCM model with all the parameter types, the parity diagram of its training set and test set is shown in Figure 3. It can be seen that some of the 2049 data points deviate from the 5% error line. However, this may be due to the fact that only one hidden layer is used for the BPNN model in order to more intuitively compare the influence of different parameters on the model prediction accuracy. The network configuration (e.g., hidden layers, number of neurons in each hidden layer) of the BPNN model was not optimized to be optimal. This article verifies this by optimizing the network structure in the next Section 3.1.3. Among the six generic BPNN models established, this model has the highest prediction accuracy and can reflect the CO_2 absorption process more comprehensively. It can also predict the load value of CO_2 in various amine aqueous solutions.

3.1.3 | Comparison of BPNN models based on different number of parameters

As previously mentioned, a series of generic BPNN models were developed with the same network structure and only one hidden layer. Although the evaluation metrics of the generic models are not

TABLE 4 Comparison of BPNN models based on different number of parameters.

Level	AARE (%)	RMSE (%)	Nodes number of hidden layer	*R*
Level O	21.31	79.08	10	0.863
Level OC	11.02	29.95	10	0.958
Level OM	15.97	45.07	10	0.915
Level OP	7.02	25.15	10	0.987
Level OPC	6.15	21.95	10	0.989
Level OPCM	4.91	16.20	10	0.993

Abbreviation: AARE, average absolute relative error; BPNN, back propagation neural network; R, correlation coefficient; RMSE, root mean square error.

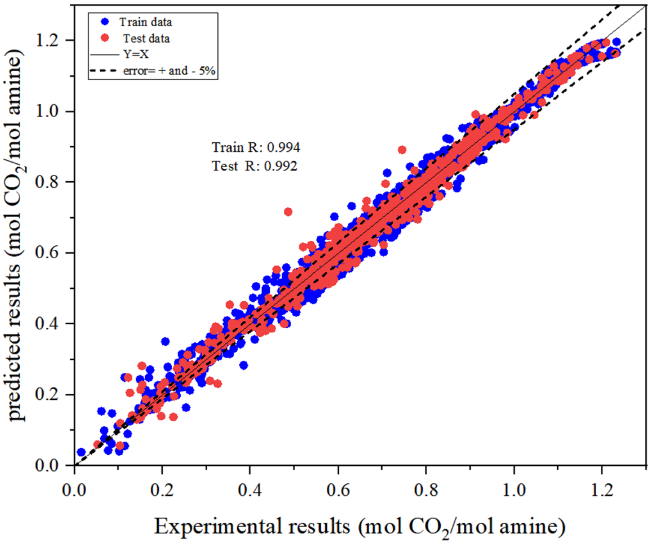


FIGURE 3 Parity plots of BPNN (only one hidden layer) model for level OPCM. BPNN, back propagation neural network.

as effective as those of the individual models, they can provide meaningful guidance to improve the development of new amine solvents.

According to Table 4 in Section 3.1.2 above, for the same BPNN models with only one hidden layer and 10 neurons. It can be seen that for the models with simple network structure, as the input features increase, the closer the *R* value is to 1, and the AARE and RMSE values decrease. This indicates that the introduction of key parameters related to solubility makes the algorithmic logic structure of ML models clearer. To further improve the prediction accuracy on this basis, we have tried to develop BPNNs with a more complex network structure. The specific operating parameters are shown in Table 5.

It is obvious that the BPNN models with more complex structures have better performance. The AAREs of the other five models are below 5% except for level O. Among them, level OPCM, as the model with the best performance, integrates all the four levels of parameters, namely operating parameters, physical property, chemical property, and molecular property, and its AAREA is only 2.03%. The parity plot of the level OPCM with the best performance index is shown in Figure 4. It can be seen that most of the 2049 data sets are distributed near the 45° line. And only a few points deviate from this line, indicating that the predicted values match well with the experimental data. Compared with the single hidden layer BPNN model, this model has better predictive performance. As with the above BPNN models with only one hidden layer, the prediction performance of the model keeps improving as the number of suitable parameters increases. It ensures the validity and wide applicability of the developed models.

3.2 | Generic RBFNN models for CO_2 solubility

In order to further investigate more ML models, RBFNN was developed to predict the CO_2 solubility in different amine solutions. RBFNN has very robust nonlinear approximation capability and fast convergence speed. Moreover, it has obvious advantages over empirical models in terms of time-consuming and accuracy deficiencies. Whereas, in order to better explain the ML model and explore the complexity, structural characteristics and the intrinsic relationship between external conditions and solubility of amine systems. A series of parameters related to chemical reactions, physical diffusion, and vapor–liquid equilibrium have been introduced. The model of level OC, level OM, level OP, level OPC, and level OPCM has been established successively.

During the modeling process, the adjustable configuration parameters (e.g., target parameters and diffusion factors) for training the RBFNN model were continuously added to obtain an optimized network model. A comparison of the prediction results with experimental data obtained from published literature was also performed to assess whether the prediction performance of the model was superior.

3.2.1 | RBFNN models of CO_2 solubility based on operating parameters

Before building the generic RBFNN model for predicting CO_2 loading, it is necessary to study the RBFNN models of individual amine systems. The datasets used for modeling were same as the BPNN model, using operating parameters (P_{CO2}, *T*, and *C*) that significantly affect CO_2 solubility as input features. The developed models' metrics are shown in Table 6.

As shown in Table 6, *R* values are all greater than 0.98 in 17 systems. Meanwhile, the AAREs of all systems were less than 6%, and

TABLE 5 Comparison of BPNN models for more complex network structures based on different parameter numbers.

Level	AARE (%)	RMSE (%)	Nodes number of hidden layer	*R*
Level O	8.49	26.91	1st layer: 10 nodes; 2nd layer: 16 nodes	0.9667
Level OC	2.97	7.75	1st layer: 12 nodes; 2nd layer: 15 nodes	0.9962
Level OM	3.13	7.85	1st layer: 12 nodes; 2nd layer: 12 nodes	0.9960
Level OP	2.79	7.14	1st layer: 10 nodes; 2nd layer: 12 nodes	0.9973
Level OPC	2.27	5.26	1st layer: 12 nodes; 2nd layer: 15 nodes	0.9978
Level OPCM	2.03	4.52	1st layer: 12 nodes; 2nd layer: 15 nodes	0.9983

Abbreviation: AARE, average absolute relative error; BPNN, back propagation neural network; R, correlation coefficient; RMSE, root mean square error.

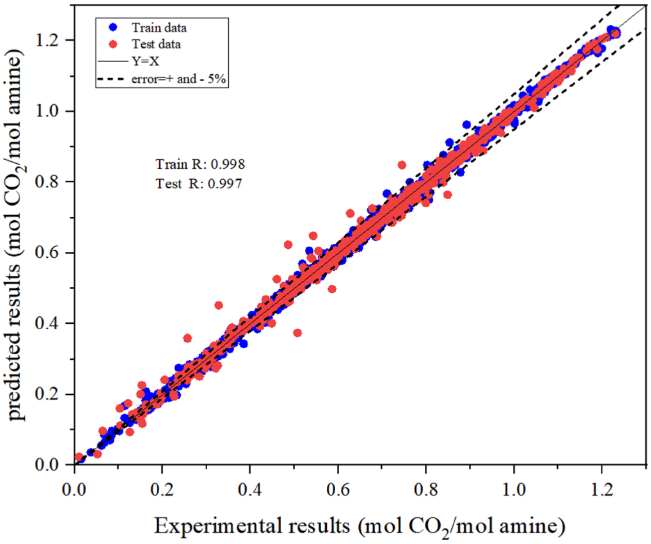


FIGURE 4 Parity plots of BPNN model (multilayer hidden layer) for level OPCM. BPNN, back propagation neural network.

even in 12 systems, the AARE values were below 2%, which was much lower than the value reported in the references. The results show that the developed RBFNN model can predict the CO_2 solubility well in these amine solutions. Meanwhile, it can be seen that the predictions using RBFNN method are better when compared with most empirical or semi-empirical models.

Although the prediction performance of the single-system model based only on the operating parameters is already good, its applicability is still doubtful. So, this work needs to develop generic models for RBFNN in order to subsequently predict the solubility of CO_2 in different amine systems.

3.2.2 | Generic RBFNN models based on different levels for CO_2 solubility

Similar to the BPNN method, a series of generic RBFNN models were also developed. The topology of the new model contains 1000 neurons in a single hidden layer.

The performance indicators of the model are shown in Table 7. Compared with the most single-system RBFNN models only based on operating parameters, the values of AARE and RMSE were obviously increased, while the values of R were also decreased. This observation can be attributed to the addition of parameters at the physical property, chemical property, and molecular property levels as additional inputs, which leads to the complexity of the RBFNN model. On the other hand, it may be due to the increase of data sets. It also leads to the complexity of RBFNN model and reduces the accuracy of prediction. However, with the increase of descriptors, the evaluation index of the model becomes better, which is consistent with the previous conclusion.

The parity plot of the level OPCM with the best performance index is shown in Figure 5. It can be seen that the points in the training and test sets are almost all concentrated on the diagonal and the *R* value is very close to 1, which indicates that the modeling effect is well. Comparing level OPC with level OPCM, it is found that the performance improvement is not obvious. There may be two reasons for the observation. First, the last introduced parameter μ might have no significant effect on the solubility of CO_2. Second, too many parameters were introduced, which increased the complexity of the model, so the improvement effect was not obvious. However, all models can help to guide the identification of effective solvents in the future.

3.3 | Generic RF models for CO_2 solubility

In this article, a tree-based CO_2 solubility model was also established using RF. RF has good resistance to overfitting, which is different from RBFNN, so some researchers use it for classification or regression problems. Similar order of establishing model levels as BPNN and RBFNN. During the training of the RF model, the prediction accuracy can be improved by adjusting the number of trees.

3.3.1 | RF models of CO_2 solubility based on operating parameters

Accordingly, a set of RF models with only operating parameters (P_{CO2}, *T*, and *C*) were established for 17 amine solution systems. The solubility values of CO_2 in nine amines published in the literature and detailed results of RF models are listed in Table 8.

As shown in Table 8, except for MDEA, DEEA, PZ, AMP, and MDEA/PZ systems, the AAREs of the other 12 amine systems are less

TABLE 6 Performance evaluation of RBFNN models for 17 amine-based systems.

		AARE (%)			
Amine	Data sets	RBFNN	Published correlations	RMSE (%)	*R*
1DMA2P	122	1.4	2.64	1.995	0.996372
AMP	96	0.61	10.50	1.09	0.999171
DEEA	175	3.1	4.84	4.68	0.995515
DMEA	80	0.77	-	0.986	0.992594
MDEA	113	4.51	10.72	18.793	0.998963
MEA	99	1.34	4.10	2.548	0.992391
PZ	255	2.46	15.70	3.896	0.999386
DETA	95	0.61	-	0.821	0.998266
DEAB	104	1.05	7.10	1.41	0.993643
MEA/DEEA	48	0.33	-	0.402	0.998733
1DMA2P/MEA	175	5.61	-	10.461	0.986028
DETA/PZ	159	0.86	-	1.141	0.995964
MEA/MDEA	114	1.62	-	2.476	0.993466
MEA/AMP	31	0.54	-	0.848	0.998694
MDEA/PZ	118	3.86	8.11	7.468	0.989383
AMP/PZ	151	1.07	19.90	1.703	0.992737
MEA/AMP BEA	114	1.95	-	2.556	0.998448

Abbreviation: AARE, average absolute relative error; R, correlation coefficient; RBFNN, radial basis function neural network; RMSE, root mean square error.

TABLE 7 Comparison of RBFNN models based on different parameter numbers.

	AARE (%)			RMSE (%)			*R*		
Level	All	Train set	Test set	All	Train set	Test set	All	Train set	Test set
Level OC	14.96	12.73	20.15	34.72	23.63	60.58	0.8850	0.9172	0.8010
Level OM	15.04	12.28	21.50	20.41	12.20	39.56	0.8727	0.9171	0.7690
Level OP	1.85	1.01	3.82	4.81	2.46	10.28	0.9935	0.9997	0.9789
Level OPC	1.73	1.02	3.38	3.98	2.01	8.59	0.9937	0.9996	0.9801
Level OPCM	1.52	0.97	0.86	3.79	2.03	7.88	0.9958	0.9998	0.9864

Abbreviation: AARE, average absolute relative error; R, correlation coefficient; RBFNN, radial basis function neural network; RMSE, root mean square error.

than 10%. This indicates that the predicted results agree with the experimental data of most amine systems. The results of correlation comparison between RF models and published AAREs indicate that RF is not good at solubility prediction. The reason may be related to the working mechanism of RF algorithm. It randomly assigns data sets and features to each decision tree. Although this approach can ensure the independence of different trees, it does not fully use of training data sets and features to produce excellent predictive results.

3.3.2 | Generic RF models based on different levels for CO_2 solubility

In this work, the adjustable parameters of the model for predicting solubility based on ML not only included in the operating conditions, but also took advantage of the physical, chemical, and molecular structural properties. In the generic RF models, we optimized the accuracy of model output by regulating the number of decision trees. The number of model evaluation indicators and decision trees at different levels are shown in Table 9. Obviously, the increase of parameters does not help the improvement of RF models. After establishing the four-parameter models level OC and level OM, input parameters continue to be added. However, the performance of level OP, level OPC, and level OPCM is inferior to that of BPNN model and RBFNN model of the same level. The overall prediction accuracy of models constructed based on RF algorithm is relatively poor, so they are not suitable for prediction models. This may be influenced by the characteristics of RF algorithm, which is more suitable to address classification than regression.

3.4 | Comparison of three ML algorithms

A full comparison between the models developed based on the three algorithms was carried out in order to find the model with the best performance. It was found that the prediction accuracy of the models was improved by increasing the parameter types for BPNN and RBFNN models. This may be due to the fact that for the complex process of absorption, the more influencing factors are considered, the more reliable and applicable the model will be, while if only one aspect is considered, it does not reveal the intrinsic relationship between the inputs and outputs well.

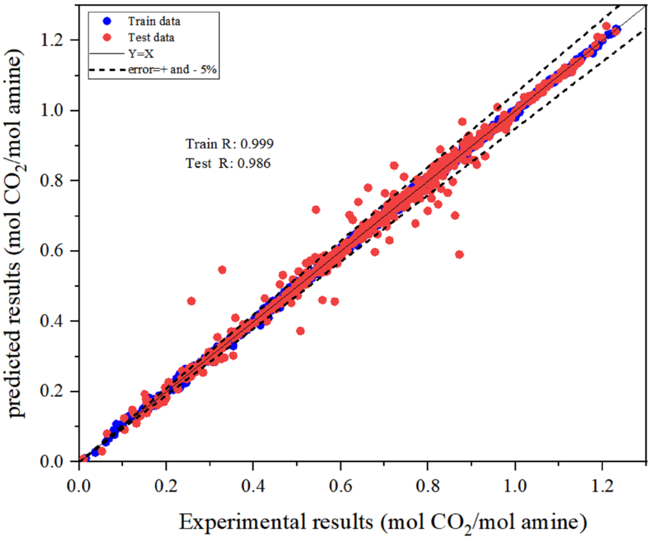


FIGURE 5 Parity plots of RBFNN model for level OPCM. RBFNN, radial basis function neural network.

BPNN showed the best performance when predicting individual amine systems due to the good generalization and fault tolerance of its algorithm. However, RBFNN can best represent the generic model dealing with large amounts of data, and its performance is better than BPNN, because it can be applied to arbitrary precision approximation problems. But RF did not give good prediction in data regression problems, because it cannot generate a continuous output, and it is more suitable for dealing with classification problems to explain the importance of features.

3.4.1 | Comparison of operating parameters-based models constructed using three ML algorithms

In this article, three ML algorithms (BPNN, RBFNN, and RF) are used to develop a series of models at different levels. In order to find a more suitable model for predicting CO_2 solubility, the models built based on operating parameters were first compared, and the results (AARE [%]) of the comparison are shown in Figure 6.

It is easy to conclude that the accuracy of the RF model is much worse than that of the BPNN and RBFNN models, especially for the prediction of the MDEA system. The next best is the RBFNN model, and the best is the BPNN model, which gives very good prediction results. The RF algorithm builds the tree by randomly selecting a subset of features to split the tree, and this feature leads to a large number of training sets that are not involved in the tree generation, which is reflected in the poor prediction accuracy. The BPNN algorithm is

TABLE 8 Performance evaluation of RF models for 17 amine-based systems.

		AARE (%)			
Amine	Data sets	RFNN	Published correlations	RMSE (%)	R
1DMA2P	122	6.78	2.64	10.6191	0.980648
AMP	96	11.57	10.50	21.976	0.975734
DEEA	175	15.01	4.84	32.110	0.985335
DMEA	80	3.88	-	5.299	0.976340
MDEA	113	59.97	10.72	243.105	0.981167
MEA	99	3.94	4.10	5.630	0.981310
PZ	255	13.76	15.70	22.185	0.995104
DETA	95	2.99	-	4.266	0.968302
DEAB	104	3.28	7.10	5.421	0.970503
MEA/DEEA	48	1.80	-	2.454	0.971009
1DMA2P/MEA	175	7.89	-	19.794	0.989557
DETA/PZ	159	2.77	-	3.845	0.979965
MEA/MDEA	114	5.91	-	10.899	0.960492
MEA/AMP	31	4.14	-	5.470	0.966129
MDEA/PZ	118	11.00	8.11	23.057	0.985881
AMP/PZ	151	2.36	19.90	3.403	0.995877
MEA/AMP/BEA	114	5.05	-	6.224	0.991939

Abbreviations: AARE, average absolute relative error; R, correlation coefficient; RF, random forest; RMSE, root mean square error.

TABLE 9 Comparison of RF models for more complex network structures based on different parameter numbers.

	AARE (%)			RMSE (%)			*R*			
Level	All	Train set	Test set	All	Train set	Test set	All	Train set	Test set	Tree
Level OC	6.27	4.92	9.39	24.96	16.92	43.71	0.9920	0.9947	0.9838	80
Level OM	6.24	5.08	8.94	23.70	17.30	38.63	0.9904	0.9948	0.9853	70
Level OP	11.61	10.49	14.24	43.97	35.42	63.92	0.9742	0.9795	0.9703	50
Level OPC	8.22	7.64	11.58	40.84	32.78	59.65	0.9817	0.9898	0.9803	65
Level OPCM	11.32	10.29	13.7	41.88	34.81	58.37	0.9776	0.9824	0.9729	55

Abbreviations: AARE, average absolute relative error; R, correlation coefficient; RF, random forest; RMSE, root mean square error. OC indicates that the model contains parameters P_{CO2},*T*,*C* and *pKa*; OM indicates that the model contains parameters P_{CO2}, *T*, *C* and *μ*; OP indicates that the model contains parameters P_{CO2}, *T*, *C*, *η*, *ρ* and *M*; OPC indicates that the model contains parameters P_{CO2}, *T*, *C*, *η*, *ρ*, *M*, and *pKa*; OPCM indicates that the model contains parameters P_{CO2}, *T*, *C*, *η*, *ρ*, *M*, *pKa* and *μ*.

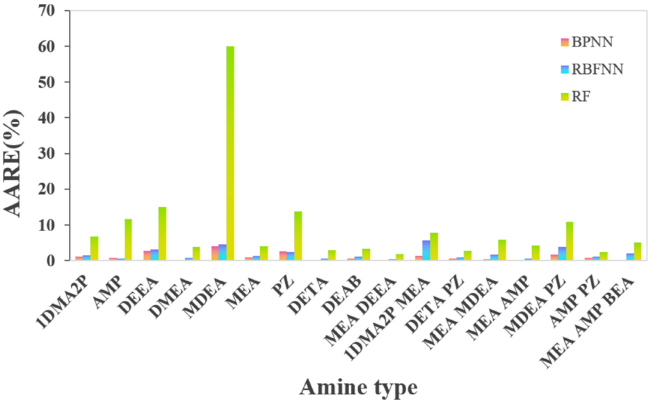
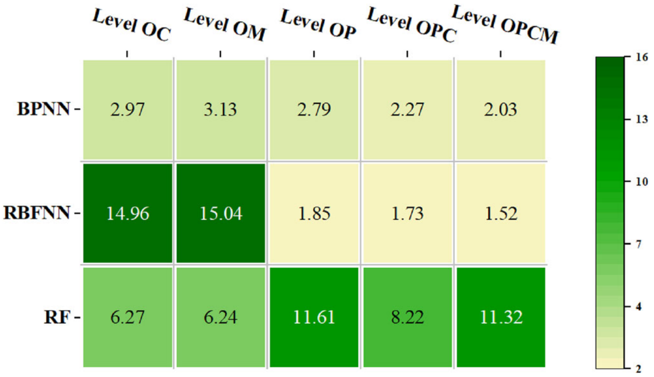


FIGURE 6 Comparison of three ML algorithms for operating parameter-based models. ML, machine learning.



	Level OC	Level OM	Level OP	Level OPC	Level OPCM
BPNN	2.97	3.13	2.79	2.27	2.03
RBFNN	14.96	15.04	1.85	1.73	1.52
RF	6.27	6.24	11.61	8.22	11.32

FIGURE 7 Comparison of ML models based on different input parameters. ML, machine learning.

trained by repeatedly adjusting the weights and thresholds of the network, so that the output value is constantly close to the target value, and has excellent generalization and fault tolerance, which makes the AAREs of MEA/DEEA, MEA/AMP, and MEA/AMP/BEA systems all less than 1%. The reason why RBFNN is slightly worse than BPNN may be that the hidden layer can only perform a fixed nonlinear change due to its simple network structure, and only a small number of data sets in a local area have a significant impact on the output. Compared with the empirical or semi-empirical models, the BPNN and RBFNN models have significantly improved the prediction accuracy.

3.4.2 | Comparison of ML models for multiple amine-based systems

In order to better predict the solubility of CO_2 in various amine solutions, input parameters of different levels were introduced to build the models, and the prediction accuracy was shown in Figure 7.

Comparing the three ML models, it is clear that the addition of parameters does not help the improvement of the RF model. RF can do regression problems, but its advantage is to solve classification problems. The accuracy of level OC and level OM models running by RBFNN algorithm is the worst, which may be caused by its algorithm structure. When the number of training samples is small, the accuracy of the prediction results is low. Therefore, when data sets increase, that is, when level OP, level OPC, and level OPCM models are established, RBFNN performs better than BPNN and RF. The most excellent model performance is the RBFNN model with 8 parameters, whose AARE is only 1.52%, followed by the RBFNN model with 7 parameters, whose AARE is 1.73%. The eight-parameter BPNN model includes operating parameters, physical property, chemical property, and molecular property levels. It also shows its good prediction accuracy with an AARE of 2.03%. Although not as accurate as individual amine systems, generic models take into account more different levels of factors and are more applicable to different amine systems.

Compared with BPNN, RBFNN is more capable of processing a large amount of data and has a faster algorithm speed. However, over-fitting should be avoided. Overall, RBFNN is the best of the three algorithms for generic models that need to process large amounts of data, followed by BPNN and RF.

3.5 | Features analysis of parameters determined by RF algorithm

In general, when developing ML models, the number of features in the data set is relatively large, so it is necessary to select the features that have a greater impact on the results for further modeling.

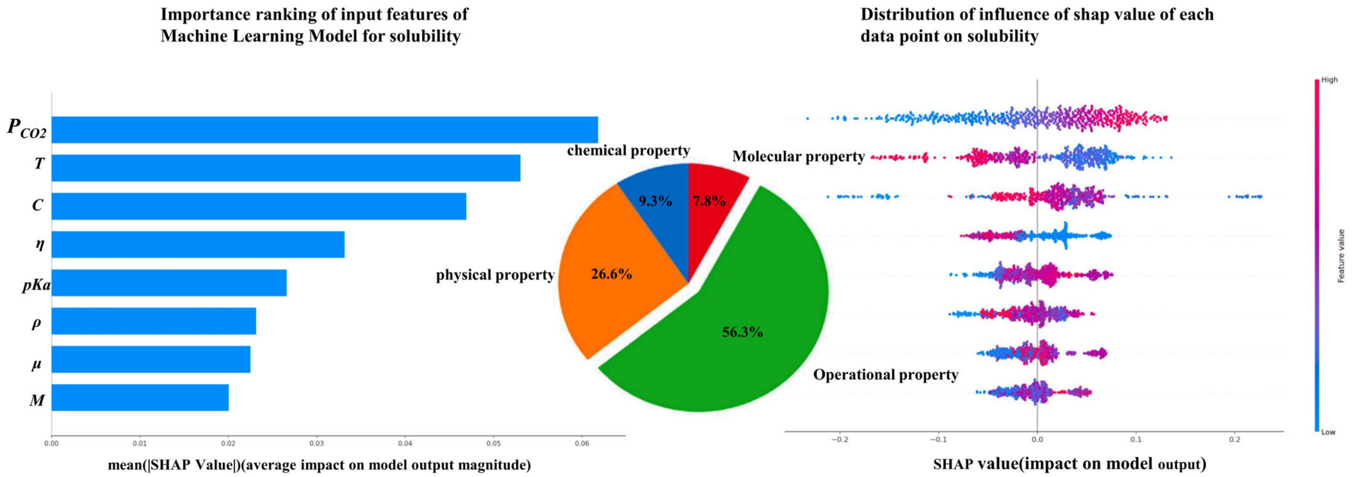


FIGURE 8 The contribution value of each parameter to solubility.

Whereas in RF it is possible to see how much each feature contributes to each tree in the RF, and then take the average. Finally, the contribution between different features was compared. Such an arithmetic feature gives us the feature importance, which is specifically evaluated by the SHAP method.

In the case of level OPCM model, the importance of features is shown in Figure 8. It should be noted that only the test set was used to train the model here. The features are sorted from top to bottom from the most important to the light representing the positive impact of the feature on the model prediction results, and vice versa, are important, with one row representing one feature value. For the rightmost scatter plot, a point represents a sample, and the presentation of the results is shown by color. Red on the rind blue on the right is similar to the negative correlation with the predicted value, and the more dispersed the sample distribution, then the greater the impact of the feature is represented. The contribution of each modifiable parameter to CO_2 loading in descending order is: $P_{CO2} > T > C > \eta > pKa > \rho > \mu > M$. It is clear from Figure 7 that the effect of the three operating parameters on the solubility of CO_2 in aqueous amine solutions is significant. In particular, the pressure, solubility increases with increasing pressure (color changes from blue to red), indicating a positive correlation between solubility and pressure. As P_{CO2} increases, the amount of dissolved CO_2 driving into the solvent increases, resulting in more CO_2 gas dissolving into the liquid phase, so the conclusion is also consistent with our reality. While for M, most of the points diffuse at SHAP = 0, introducing this parameter into the RF model fails to achieve good predictive performance because its contribution to the solubility in the modeling process is negligible. This is probably due to the fact that the molecular weight, does not itself have a significant effect on chemical reactions. While for the leftmost bar graph, although the positive and negative correlation between feature values and input parameters cannot be directly obtained, it is more intuitive to show the influence of each feature on the model output results, further validating the results derived from the scatter plot.

Despite the fact that feature importance can provide useful guidance for parameter selection, the accurate expression of feature importance for each parameter requires further research because of the relatively simple RF structure and poor repeatability.

3.6 | Framework of AI models development and Optimization

In this work, a framework of ML model development and optimization was proposed based on the above investigation and discussion, as shown in Figure 9. It was made up of parameter selection, ML models development with model selection and evaluation, feature importance analysis, ML models optimization. It could be briefly described as follows: first, input parameters selections; in order to make the established model widely applicable, the physical and chemical properties of amines, the structural properties of amine molecules, and the influence of operating conditions on absorption were considered comprehensively, and the input parameters were divided into different levels for input. Then, three algorithms with different advantages, BPNN, RBFNN, and RF, were selected to run the model, while three evaluation metrics, AARE, RMSE, and R, were utilized to facilitate visualization of the model performance. The Pearson correlation showed that BPNN and RBFNN performed well in handling the regression problem. The RF algorithm, on the other hand, is less suitable for dealing with regression problems and has an advantage in assessing feature importance. The key parameters can then be analyzed by the RF algorithm, and then modeled using BPNN and RBFNN to further optimize the model. The optimization at this point can be considered in terms of both the structure of the model itself and parameter optimization. The model structure can be adjusted by in terms of the number of hidden layer nodes, and so on. And parameter optimization utilizes the SHAP method to obtain feature importance rankings, and the results can be returned to provide guidance for selecting more appropriate parameters and building new models.

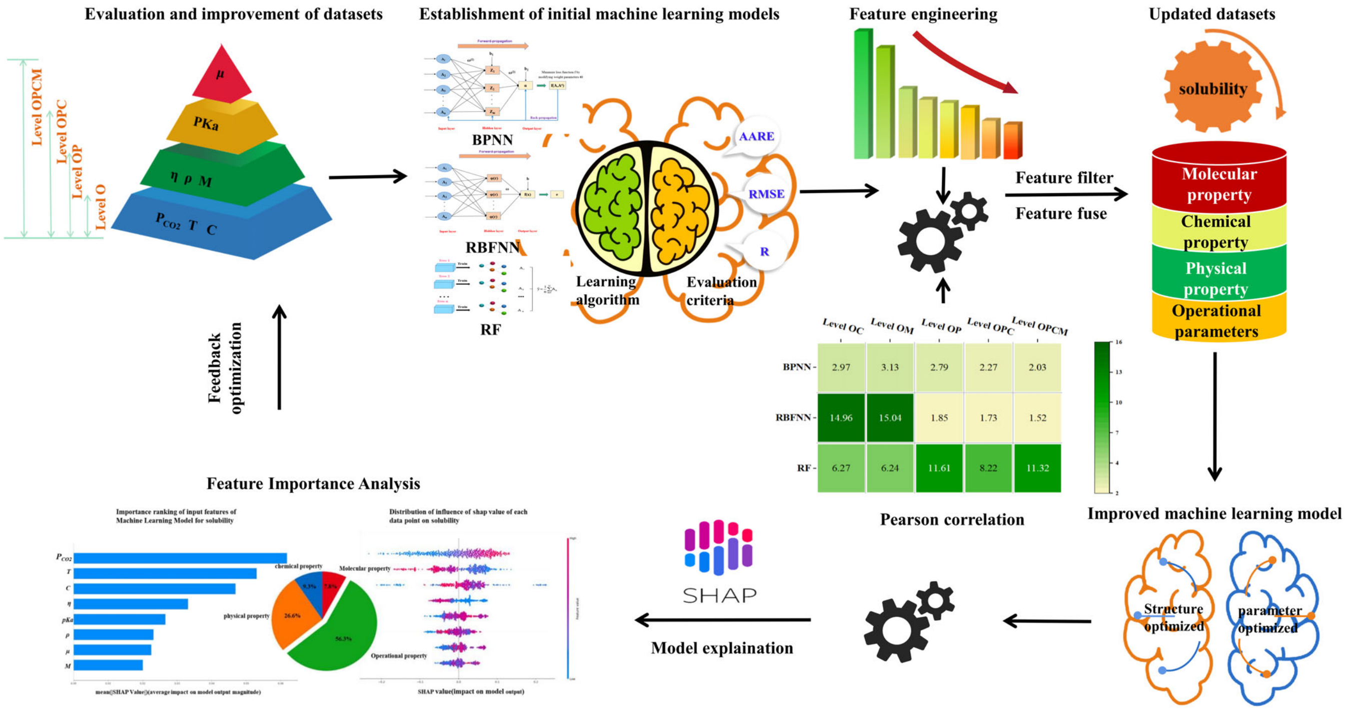


FIGURE 9 Modeling full flow chart.

4 | CONCLUSIONS

In this work, a series of prediction models for CO_2 solubility were developed based on 17 amine systems with 2049 data sets. The prediction data obtained from most of the models trained for a specific amine system basically matched the experimental observations, especially the model developed by the BPNN algorithm, which had 14 amine systems with AARE values of less than 2%. At the same time, in order to greatly extend the scope of application of the model, a general model was established. The generic model developed considered the influence of four levels of operating parameters: physical properties, chemical properties, and molecular properties. It can reflect the effect of chemical reactions on CO_2 solubility to some extent. A comparison of the models shows that for level OC and level OM, BPNN is the best choice, while the developed RBFNN (level OP, level OPC, and level OPCM) models are superior to other generic models. In addition, RF can rank the importance of features for filtering input parameters. It enables the developed generic model to estimate the importance of each feature to the model prediction results. And provide some empirical guidance for subsequent model development. The work in this article can provide theoretical basis and practical guidance for developing new absorbents and identifying new carbon capture alternative molecules in the future.

AUTHOR CONTRIBUTIONS

Ting Lan: Data curation (lead); validation (lead); writing – original draft (lead); writing – review and editing (lead). **Shoulong Dong:** Software (lead); writing – review and editing (supporting). **Hui Luo:** Formal analysis (equal); resources (supporting). **Liju Bai:** Writing – review and editing (supporting). **Helei Liu:** Formal analysis (lead); methodology (lead); supervision (lead); writing – review and editing (lead).

ACKNOWLEDGMENTS

The financial support from the National Natural Science Foundation of China (NSFC-No.22208017) and the talent project from Beijing Institute of Technology (2022CX01004) are gratefully acknowledged. The instrumental support from the Analysis and Testing Center of Beijing Institute of Technology is also highly appreciated.

DATA AVAILABILITY STATEMENT

The numerical data from Figures 2–6 and 8 are tabulated in the Supporting Information. The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Shoulong Dong https://orcid.org/0000-0002-4956-4160
Helei Liu https://orcid.org/0000-0002-8230-5865

REFERENCES

1. Raganati F, Miccio F, Ammendola P. Adsorption of carbon dioxide for post-combustion capture: a review. *Energy Fuel*. 2021;35(16):12845-12868.
2. Goglio P, Williams AG, Balta-Ozkan N, et al. Advances and challenges of life cycle assessment (LCA) of greenhouse gas removal technologies to fight climate changes. *J Clean Prod*. 2020;244:118896.
3. Srisang W, Pouryousefi F, Osei PA, et al. Evaluation of the heat duty of catalyst-aided amine-based post combustion CO_2 capture. *Chem Eng Sci*. 2017;170:48-57.

4. Chao C, Deng Y, Dewil R, Baeyens J, Fan X. Post-combustion carbon capture. *Renew Sustain Energy Rev*. 2021;138:110490.
5. Abd AA, Naji SZ, Hashim AS, Othman MR. Carbon dioxide removal through physical adsorption using carbonaceous and non-carbonaceous adsorbents: a review. *J Environ Chem Eng*. 2020;8(5):104142.
6. Osman AI, Hefny M, Abdel Maksoud M, Elgarahy AM, Rooney DW. Recent advances in carbon capture storage and utilisation technologies: a review. *Environ Chem Lett*. 2021;19(2):797-849.
7. Meng F, Meng Y, Ju T, Han S, Lin L, Jiang J. Research progress of aqueous amine solution for CO_2 capture: a review. *Renew Sustain Energy Rev*. 2022;168:112902.
8. Mohamadi-Baghmolaei M, Hajizadeh A, Zahedizadeh P, Azin R, Zendehboudi S. Evaluation of hybridized performance of amine scrubbing plant based on exergy, energy, environmental, and economic prospects: a gas sweetening plant case study. *Energy*. 2021;214:118715.
9. Shakerian F, Kim K-H, Szulejko JE, Park J-W. A comparative review between amines and ammonia as sorptive media for post-combustion CO_2 capture. *Appl Energy*. 2015;148:10-22.
10. Zhang Z, Borhani TN, Olabi AG. Status and perspective of CO_2 absorption process. *Energy*. 2020;205:118057.
11. Liu H, Li M, Idem R, Tontiwachwuthikul PP, Liang Z. Analysis of solubility, absorption heat and kinetics of CO_2 absorption into 1-(2-hydroxyethyl) pyrrolidine solvent. *Chem Eng Sci*. 2017;162:120-130.
12. Dutcher B, Fan M, Russell AG. Amine-based CO_2 capture technology development from the beginning of 2013—a review. *ACS Appl Mater Interfaces*. 2015;7(4):2137-2148.
13. Porcheron F, Gibert A, Mougin P, Wender A. High throughput screening of CO_2 solubility in aqueous monoamine solutions. *Environ Sci Technol*. 2011;45(6):2486-2492.
14. Wang S, Song Y, Zhang Y, Chen C-C. Electrolyte thermodynamic models in Aspen process simulators and their applications. *Ind Eng Chem Res*. 2022;61(42):15649-15660.
15. Moioli S, Pellegrini LA. Modeling the methyldiethanolamine-piperazine scrubbing system for CO_2 removal: thermodynamic analysis. *Front Chem Sci Eng*. 2016;10(1):162-175.
16. Zhang Y, Que H, Chen C-C. Thermodynamic modeling for CO_2 absorption in aqueous MEA solution with electrolyte NRTL model. *Fluid Phase Equilib*. 2011;311:67-75.
17. Haghighatlari M, Vishwakarma G, Altarawy D, Subramanian R, Hachmann J. ChemML: a machine learning and informatics program package for the analysis, mining, and modeling of chemical and materials data. *WIREs Comput Mol Sci*. 2020;10(4):e1458.
18. Abooali D, Soleimani R, Rezaei-Yazdi A. Modeling CO_2 absorption in aqueous solutions of DEA, MDEA, and DEA+MDEA based on intelligent methods. *Sep Sci Technol*. 2019;55(2):1-11.
19. Dashti A, Raji M, Alivand MS, Mohammadi AH. Estimation of CO_2 equilibrium absorption in aqueous solutions of commonly used amines using different computational schemes. *Fuel*. 2020;264:116616.
20. Liu H, Jiang X, Idem R, Dong S, Tontiwachwuthikul P. AI models for correlation of physical properties in system of 1DMA2P-CO_2-H_2O. *AIChE J*. 2022;68(9):e17761.
21. Tellagorla R, Balchandani SC, Gumma S, Mandal B. Equilibrium CO_2 solubility of novel tris (2-aminoethyl) amine as a promoter to N-methyldiethanolamine and 2-amino-2-methyl-1-propanol. *Sep Purif Technol*. 2021;279:119705.
22. Quan H, Dong S, Zhao D, Li H, Geng J, Liu H. Generic AI models for mass transfer coefficient prediction in amine-based CO_2 absorber, part II: RBFNN and RF model. *AIChE J*. 2023;69(1):e17904.
23. Dong S, Quan H, Zhao D, Li H, Geng J, Liu H. Generic AI models for mass transfer coefficient prediction in amine-based CO_2 absorber, part I: BPNN model. *Chem Eng Sci*. 2022;264:118165.
24. Kent RL, Eisenberg B. Better Data for Amine Treating. *Hydrocarb. Process*. 1976;55(2):87-90.
25. Mondal BK, Bandyopadhyay SS, Samanta AN. Experimental measurement and Kent-Eisenberg modelling of CO_2 solubility in aqueous mixture of 2-amino-2-methyl-1-propanol and hexamethylenediamine. *Fluid Phase Equilib*. 2017;437:118-126.
26. Pandey D, Mondal MK. Equilibrium CO_2 solubility in the aqueous mixture of MAE and AEEA: experimental study and development of modified thermodynamic model. *Fluid Phase Equilib*. 2020;522:112766.
27. Tzirakis F, Papadopoulos AI, Seferlis P, Tsivintzelis I. CO_2 solubility in aqueous N-methylcyclohexylamine (MCA) and N-cyclohexyl-1,3-propanediamine (CHAP) solutions. *AIChE J*. 2023;69(3):e17982.
28. Hu W, Chakma A. Modelling of equilibrium solubility of CO_2 and H_2S in aqueous amino methyl propanol (AMP) solutions. *Chem Eng Commun*. 1990;94(1):53-61.
29. Zheng W, Xiao M, Liu H, Gao H, Liang Z. Modeling and experiments of equilibrium solubility of carbon dioxide in aqueous N-(2-hydroxyethyl) pyrrolidine solution. *J Taiwan Inst Chem Eng*. 2018;85:132-140.
30. Li M-H, Shen K-P. Calculation of equilibrium solubility of carbon dioxide in aqueous mixtures of monoethanolamine with methyldiethanolamine. *Fluid Phase Equilib*. 1993;85:129-140.
31. Liu H, Xiao M, Luo X, et al. Modeling of CO_2 equilibrium solubility in a novel 1-diethylamino-2-propanol solvent. *AIChE J*. 2017;63(10):4465-4475.
32. Deshmukh R, Mather AE. A mathematical model for equilibrium solubility of hydrogen sulfide and carbon dioxide in aqueous alkanolamine solutions. *Chem Eng Sci*. 1981;36(2):355-362.
33. Afkhamipour M, Mofarahi M. Experimental measurement and modeling study on CO_2 equilibrium solubility, density and viscosity for 1-dimethylamino-2-propanol (1DMA2P) solution. *Fluid Phase Equilib*. 2018;457:38-51.
34. Xiao M, Cui D, Liu H, Tontiwachwuthikul P, Liang Z. A new model for correlation and prediction of equilibrium CO_2 solubility in N-methyl-4-piperidinol solvent. *AIChE J*. 2017;63(8):3395-3403.
35. Zhang R, He X, Liu T, et al. Thermodynamic studies for improving the prediction of CO_2 equilibrium solubility in aqueous 2-dimethylamino-2-methyl-1-propanol. *Sep Purif Technol*. 2022;295:121292.
36. Liu H, Gao H, Idem R, Tontiwachwuthikul P, Liang Z. Analysis of CO_2 solubility and absorption heat into 1-dimethylamino-2-propanol solution. *Chem Eng Sci*. 2017;170:3-15.
37. Chen CC, Britt HI, Boston JF, Evans LB. Local composition model for excess Gibbs energy of electrolyte systems. Part I: single solvent, single completely dissociated electrolyte systems. *AIChE J*. 1982;28(4):588-596.
38. Chen CC, Evans LB. A local composition model for the excess Gibbs energy of aqueous electrolyte systems. *AIChE J*. 1986;32(3):444-454.
39. Chen CC, Mathias PM, Orbey H. Use of hydration and dissociation chemistries with the electrolyte–NRTL model. *AIChE J*. 1999;45(7):1576-1586.
40. Chen CC, Song Y. Generalized electrolyte-NRTL model for mixed-solvent electrolyte systems. *AIChE J*. 2004;50(8):1928-1941.
41. Agarwal NK, Mondal BK, Samanta AN. Measurement of vapour-liquid equilibrium and e-NRTL model development of CO_2 absorption in aqueous dipropylenetriamine. *Environ Sci Pollut Res*. 2021;28:19285-19297.
42. Tontiwachwuthikul P, Meisen A, Lim CJ. Solubility of carbon dioxide in 2-amino-2-methyl-1-propanol solutions. *J Chem Eng Data*. 1991;36(1):130-133.
43. Luo X, Chen N, Liu S, et al. Experiments and modeling of vapor-liquid equilibrium data in DEEA-CO_2-H_2O system. *Int J Greenh Gas Control*. 2016;53:160-168.
44. Xiao M, Liu H, Gao H, Liang Z. CO_2 absorption with aqueous tertiary amine solutions: equilibrium solubility and thermodynamic modeling. *J Chem Thermodyn*. 2018;122:170-182.

45. Benamor A, Aroua MK. Modeling of CO_2 solubility and carbamate concentration in DEA, MDEA and their mixtures using the Deshmukh–Mather model. *Fluid Phase Equilib*. 2005;231(2):150-162.
46. Shen KP, Meng HL. Solubility of carbon dioxide in aqueous mixtures of Monoethanolamine with Methyldiethanolamine. *J Chem Eng Data*. 1992;37(1):96-100.
47. Maneeintr K, Idem RO, Tontiwachwuthikul P, Wee A. Synthesis, solubilities, and cyclic capacities of amino alcohols for CO_2 capture from flue gas streams. *Energy Procedia*. 2009;1(1):1327-1334.
48. Aroua MK, Salleh RM. Solubility of CO_2 in aqueous piperazine and its modeling using the Kent-Eisenberg approach. *Chem Eng Technol*. 2004;27(1):65-70.
49. Chang YC, Leron RB, Li MH. Equilibrium solubility of carbon dioxide in aqueous solutions of (diethylenetriamine + piperazine). *J Chem Thermodyn*. 2013;64:106-113.
50. Sema T, Naami A, Idem R, Tontiwachwuthikul P. Correlations for equilibrium solubility of carbon dioxide in aqueous 4-(diethylamino)-2-butanol solutions. *Ind Eng Chem Res*. 2011;50(24):14008-14015.
51. Xiao M, Cui D, Yang Q, et al. Role of mono-and diamines as kinetic promoters in mixed aqueous amine solution for CO_2 capture. *Chem Eng Sci*. 2021;229:116009.
52. Afkhamipour M, Mofarahi M, Rezaei A, Mahmoodi R, Lee C-H. Experimental and theoretical investigation of equilibrium absorption performance of CO_2 using a mixed 1-dimethylamino-2-propanol (1DMA2P) and monoethanolamine (MEA) solution. *Fuel*. 2019;256:115877.
53. Park SH, Lee KB, Hyun JC, Kim SH. Correlation and prediction of the solubility of carbon dioxide in aqueous alkanolamine and mixed alkanolamine solutions. *Ind Eng Chem Res*. 2002;41(6):1658-1665.
54. Dash SK, Bandyopadhyay SS. Studies on the effect of addition of piperazine and sulfolane into aqueous solution of N-methyldiethanolamine for CO_2 capture and VLE modelling using eNRTL equation. *Int J Greenh Gas Control*. 2016;44:227-237.
55. Feron PHM, Cousins A, Jiang K, Zhai R, Garcia M. An update of the benchmark post-combustion CO_2 capture technology. *Fuel*. 2020;273:117776.
56. Yang ZY, Soriano AN, Caparanga AR, Li MH. Equilibrium solubility of carbon dioxide in (2-amino-2-methyl-1-propanol+ piperazine+ water). *J Chem Thermodyn*. 2010;42(5):659-665.
57. Li T, Yang C, Tantikhajorngosol P, Sema T, Shi H, Tontiwachwuthikul P. Experimental investigations and the modeling approach for CO_2 solubility in aqueous blended amine systems of monoethanolamine, 2-amino-2-methyl-1-propanol, and 2-(butylamino) ethanol. *Environ Sci Pollut Res*. 2022;29(46):69402-69423.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Lan T, Dong S, Luo H, Bai L, Liu H. Feature analysis of generic AI models for CO_2 equilibrium solubility into amines systems. *AIChE J*. 2024;70(5):e18363. doi:10.1002/aic.18363