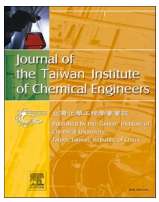
Journal of the Taiwan Institute of Chemical Engineers 171 (2025) 106054

Contents lists available at ScienceDirect

Journal of the Taiwan Institute of Chemical Engineers

journal homepage: www.journals.elsevier.com/journal-of-the-taiwan-institute-of-chemical-engineers



Leveraging spatial charge descriptor in deep learning models: Toward highly accurate prediction of vapor-liquid equilibrium

Hsiu-Min Hung, Ying-Chieh Hung *

Department of Chemical Engineering and Biotechnology, National Taipei University of Technology, Taipei 10608, Taiwan

ARTICLE INFO

Keywords:
Vapor-liquid equilibrium
Sigma profile
MLP-ANN
MACCS keys
Molecular descriptor

ABSTRACT

Background: Vapor-liquid equilibrium (VLE) data is essential for separation processes, but traditional models such as UNIFAC require extensive experimental parameters. To improve the efficiency of VLE modeling and reduce the dependence on experiments, we developed machine learning models using COSMO-based σ-profiles and MACCS keys.

Methods: Two deep learning models, MLP-COSMO and MLP-MACCS, are developed. MLP-COSMO, based on σ-profiles, allows high-precision phase equilibrium predictions using only molecular structures, eliminating the need for experimental interaction parameters. The pressure range spans 0.947 to 817 kPa and the temperature range spans 199.93 to 548.15 K.

Significant findings: By utilizing σ-profiles, our developed model MLP-COSMO, which surpasses COSMO-SAC (2010) in accuracy and achieves a level comparable to UNIFAC, with specific results of R^2-y = 0.9926, R^2-P = 0.9889, $AAD-y(\%) = 1.04\%$ and $AARD-P(\%) = 2.88\%$, evaluated on a test dataset excluded from training process. This study successfully demonstrated that high-precision VLE predictions can be achieved using only a molecular structure, effectively addressing the challenge of missing experimental parameters. Furthermore, the results indicate that the spatial charge descriptor (σ-profile), encapsulating molecular polarity information, is considered to be more suitable as input data for machine learning models in VLE prediction than structural fingerprints of MACCS.

1. Introduction

In recent advances in chemical engineering, the study of essential properties of materials has become indispensable spanning diverse fields such as separation, materials design, environmental engineering, and the transition to renewable energy. Among the studies of physical properties in chemical engineering, phase equilibrium has been a long-standing area of research. Phase equilibrium refers to the balance of chemical species between different phases, such as gas and liquid, liquid and liquid, or liquid and solid in systems composed of one or multiple substances. For example, understanding the vapor-liquid equilibrium (VLE) is fundamental to designing and operating the distillation process.

To reduce experimental costs associated with phase equilibrium measurement, the development and construction of excess Gibbs free energy models remains a significant part of phase equilibrium research. Whether empirical or theoretical, the parameters of molecular interactions are key factors in determining the accuracy of the model, and these parameters are typically determined by fitting them into experimental data. Empirical excess Gibbs free energy models, such as the Non-Random Two Liquid (NRTL) Model [1], the Wilson Model [2], and the Universal Quasi-Chemical (UNIQUAC) Model [3], operate using empirically determined parameters. For systems or conditions lacking experimental data, the simulation of phase equilibrium becomes impossible. To address this issue, the UNIFAC (UNIQUAC Functional Group Activity Coefficients) model [4,5] based on group contribution method is developed and widely used. In UNIFAC, molecules are divided into functional groups, and each functional group contributes to the overall properties of the mixture. By summing the contributions of all functional groups present in a molecule, we estimate various thermodynamic properties, including activity coefficients, vapor-liquid equilibria, and liquid-liquid equilibria.

As a first-principle approach, Klamt proposed the conductor-like screening model for Real Solvents (COSMO-RS) method [6–10], and Lin et. al. proposed the COSMO segment activity coefficient (COSMO--SAC) method [11–15]. By combining the molecular surface screening charge distribution obtained through quantum chemical calculations

* Corresponding author.
E-mail address: hungyc@ntut.edu.tw (Y.-C. Hung).

https://doi.org/10.1016/j.jtice.2025.106054
Received 11 November 2024; Received in revised form 21 February 2025; Accepted 21 February 2025
Available online 5 March 2025

and the relationship equations of statistical thermodynamics, it is possible to calculate the activity coefficients of components in mixed systems using only molecular structure information. These methods do not require empirically determined parameters, allowing for the prediction of various thermodynamic properties, including phase equilibria [13,16,17], solubility [18,19], CO_2 capture [20–23], and partition coefficients [18,24,25]. In the thermodynamics models based on first-principles mentioned above, the probability distribution of the surface charge segment of a molecule was obtained by optimizing the geometry of molecules in the conductor like screening model (COSMO) continuum solvation model of conductor-like screening (COSMO), called the sigma profile (σ-profile). Since the σ-profile represents the distribution of surface charge densities in a molecule. When a molecule's σ-profile shows a higher probability distribution in regions of high charge density, it implies that the molecule has hydrogen bond-accepting characters. Conversely, if the molecule exhibits a higher probability distribution in regions of low charge density, it often suggests that the molecule has hydrogen bond-donating characters.

Although many conventional thermodynamic models have demonstrated good predictive capability for VLE, the development of model parameters has always been a tedious process. In recent years, leveraging machine learning (ML) and artificial intelligence (AI) to assist in thermodynamic model development or directly predict thermodynamic properties has emerged as a promising solution. ML and AI have made notable strides in various chemical research domains in the past decade [26–34]. Recently, there has been an increasing amount of research attempting to use AI to predict VLE of substance mixtures. Mohanty applied AI network for the estimation of VLE for the binary systems of carbon dioxide–ethyl caproate, ethyl caprylate, and ethyl caprate. The results demonstrated a lower percentage deviation compared to the VLE estimated by the Soave-Redlich-Kwong or Peng Robinsons equation of state [35]. Vaferi et al. [36] built up a ANN based model to predict bubble and dew point pressures of various CO_2–refrigerant binary systems. In the 9 binary systems studied, the predicted results showed the average absolute relative deviation (*AARD* − *P*(%)) and root mean square error (RMSE) of 2.79 % and 0.1153, respectively. Sun et al. [37] utilized machine learning models of ANN and RF, with input layers based on component thermodynamic properties to predict the phase behavior of VLE. The model demonstrated high efficiency for a quick pre-estimation of the azeotropic distinction or a relative volatility estimation. Mesbaha et al. [38] developed a ML model based on the Least-Square Support Vector Machine (LSSVM) algorithm to simulate 425 VLE data from seven CO_2/hydrocarbon binary mixtures in supercritical or near critical conditions, showed the *AARD* − *P*(%) of 3.61.

Among these, the selection of molecular descriptors plays a crucial role in the predictive capability of the machine learning model. Taking the σ-profile as a molecular feature input layer in the machine learning model can provide numerous advantages over other conventional molecular fingerprints. First, the σ-profile depicted the probability distribution of spatial surface charge densities of a molecule. Unlike traditional molecular descriptors that focus only on atom types and connectivity, this method incorporates spatial information, earning it the designation of a spatial charge descriptor [39]. Due to these additional physical characteristics, it is better suited for predicting properties related to the three-dimensional structure. Another significant advantage is that the format of σ-profiles is customizable and remains consistent regardless of the size or complexity of the molecules, making it suitable for machine learning input. In a recent study, the σ-profile has been claimed to be universal molecular descriptors capable of predicting various physicochemical properties, highlighting its potential superiority over traditional molecular descriptors [40].

Numerous successful machine learning models have been implemented taking σ-profile as a molecular fingerprint. Hao et al. [39] integrated the COSMO-based σ-profile into machine learning modeling, creating the COSMO-SVM model to validate cocrystal formation. They reported that the σ-profile, which contains intrinsic spatial charge information, significantly improves the prediction accuracy of the ML model. Kataoka et al. [20] incorporated σ-profile into various machine learning models, successfully described the phase change of the biphasic solvent in the application of CO_2 capture. Li et al. [41] took σ-profiles as molecular fingerprints to predict drug solubility in super-critical CO_2. Their results demonstrated that deep neural networks (DNN) and random forest (RF) provided high accuracy in the prediction of solubility using σ-profile as input features. Mohan et al. [42] proposed an artificial neural networks to predict the solubility of CO_2 in deep eutectic solvents (DES), using σ-profiles to represent DES molecules.

In this work, the multilayer perceptron artificial neural network (MLP-ANN) algorithm was selected to build predictive machine learning models. The algorithm had been successfully validated for the prediction of various thermodynamic properties [41,43–46]. Previous work using machine learning methods primarily used thermodynamic properties from real experiments of individual components as input information. By contrast, our study aims to develop predictive machine learning models using first-principle information as input for VLE behavior prediction, with the goal of reducing experimental costs and addressing the issue of missing thermodynamic parameters.

2. Method

2.1. Multilayer perceptron-artificial neural network (MLP-ANN)

In this work, the MLP-ANN models were developed using TensorFlow [47], an open-source Python package. Among various artificial neural network (ANN) models, the multilayer perceptron (MLP) has the greatest practical importance. MLP-ANN is a type of feedforward network with a structure that includes an input layer, an output layer, and several hidden layers. Each layer consists of multiple neurons, each of which is connected to neurons in the preceding and following layers, allowing information to flow forward and back during training. Compared to deep neural networks (DNNs), MLP-ANNs typically have a limited number of hidden layers, making them relatively simple but still effective for many applications.

2.2. Hyperparameter optimization

Hyperparameter optimization is a process that aims to find the optimal set of hyperparameters for machine learning models. The commonly used algorithms include grid search, random search, particle swarm optimization, tree-structured Parzen estimator, hyperband, genetic algorithms, and Bayesian optimization. In this work, Bayesian optimization was used for hyperparameter optimization in the MLP-ANN model. The range of the hyperparameter is guided by previous experience and literature reference [41]. Based on the consideration of a relatively larger dataset, we expanded the range of neurons, epochs, and batch size during the optimization process. The hyperparameter optimization items and range are listed in Table 1. During the optimization process, hyperparameters were optimized to minimize the objective function, which will be introduced in the later section. The

Table 1
List of hyperparameter ranges in the MLP-ANN model during Bayesian optimization.

MLP-ANN	
Hyperparameters	Range
Number of hidden layers	(1, 4)
Neurons	(16, 256)
Epoch	(200, 1500)
Batch size	(16, 240)
Learning rate	(0.0001, 0.01)
Activation function	(ReLU)

hyperparameter values obtained from Bayesian optimization in the developed MLP-MACCS and MLP-COSMO models are summarized in **Table S1** of the Supplementary Materials.

2.3. Data transformation

Data transformation is a critical preprocessing step to address skewed data distributions. Data skewness can be understood as the asymmetry of a data distribution, in which the data points may cluster more on one side of the distribution than on the other. As is known, handling skewed or imbalanced datasets is a frequent and challenging issue in machine learning [48]. Classification models often misclassify the minority class. Consequently, if skewed data are not properly handled, the model is prone to high misclassification rates [49]. The use of logarithmic data transformation in handling skewed datasets also demonstrated an effective improvement in the predictive performance of the ANN model [50]. Inspired by previous works, we implemented root and logarithmic transformations for the correction of data skewness.

3. Data processing

The experimental data for VLE were obtained from the DECHEMA Chemistry Data Series [51]. The Fig. 1 shows distribution of VLE data including vapor composition (y_A) and pressure (P). **Fig. 1(a)** and **Fig. 1 (b)** are the original distribution of y_A (y_A^{ori}) and P (P^{ori}) . In the original dataset, the distribution of y_A shows negative skewness with skewness = -0.6439. In contrast, the original P shows extremely high asymmetry with skewness = 4.3526. Therefore, we need to restructure the data based on the positive skewness of P to facilitate machine learning models. Fig. 1 **(c)** to **1(f)** represent various data transformation methods in the pressure dataset, including square-root transformed pressure ($P^{1/2}$) (Fig. 1 **(c)**), cubic-root transformed pressure ($P^{1/3}$) (Fig. 1 **(d)**), 15th-root transformed pressure ($P^{1/15}$) (Fig. 1 **(e)**) and logarithmic transformed pressure (log P) (Fig. 1 **(f)**), respectively. As illustrated in Fig. 1 **(c)** to **1(e)**, the root transformation effectively reduces the skewness of the P distribution. Skewness decreases progressively with higher root orders. Fig. 1 **(f)** shows the logarithmic transformation of P, resulting in the smallest skewness value of 0.0768.

In this work, two molecular descriptors, the σ-profile and MACCS keys, are introduced as input layer information. The equilibrium molecular geometry is first determined under vacuum, after which the COSMO calculations are performed without any additional geometry optimization. The σ-profile of a given compound is obtained by converting a .cosmo file generated using Gaussian 16 [52] at b3lyp/6–31G (d,p) level [53]. The fundamental and detailed procedure for generating σ-profile is described in the Supplementary Materials. The complete σ-profile comprises 51 surface area values, spanning charge densities from −0.025 to 0.025 e/Å² in increments of 0.001. The MACCS keys descriptor is created by RDkit [54] by converting the SMILES notation of a compound obtained from the website 'PubChem Identifier Exchange Service'. The MACCS keys fingerprint consists of 166 bits; each bit position corresponds to the presence (=1) or absence (=0) of a pre-defined feature, such as atom types, bond types and counts, specific atomic combinations, functional groups, and other structural elements. The conceptual modeling flow chart of this work is shown in the Fig. 2. The detailed architecture of the MLP-ANN in the prediction of VLE behavior is shown in the Fig. 3. The framework contains two MLP-ANNs, one using σ-profile and the other using MACCS keys to represent binary compounds during modeling. In our machine learning model development, all VLE data were randomly shuffled and divided into training and testing datasets in a 4:1 ratio using 5-fold cross-validation to prevent overfitting and ensure consistent results across different data subsets. The prediction precision of the MLP-ANN model was evaluated using the

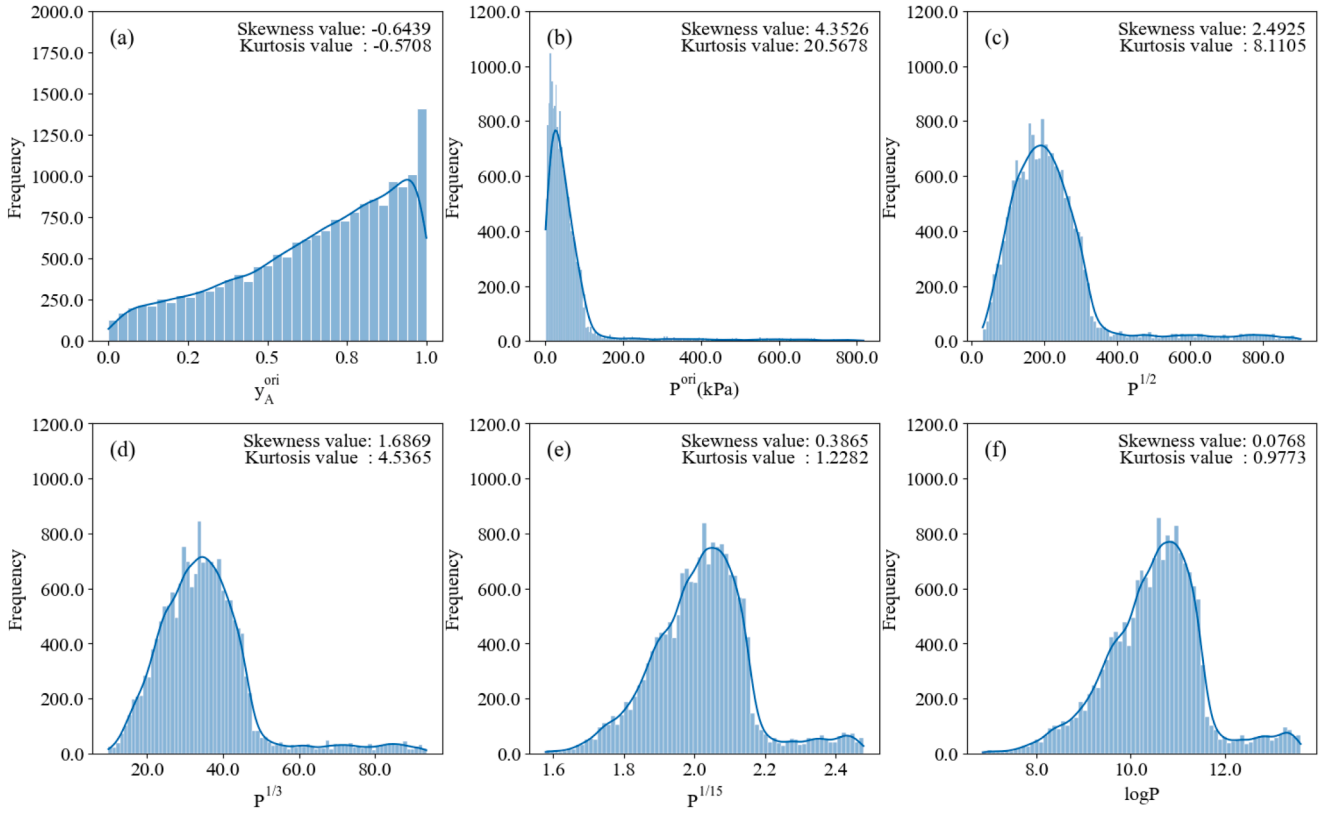


Fig. 1. Histograms of the original and transformed data distributions. (a) Original vapor composition distribution (b) original pressure distribution (c) square-root transformed pressure distribution (d) cubic-root transformed pressure distribution (e) 15th-root transformed pressure distribution (f) logarithmic transformed pressure distribution.

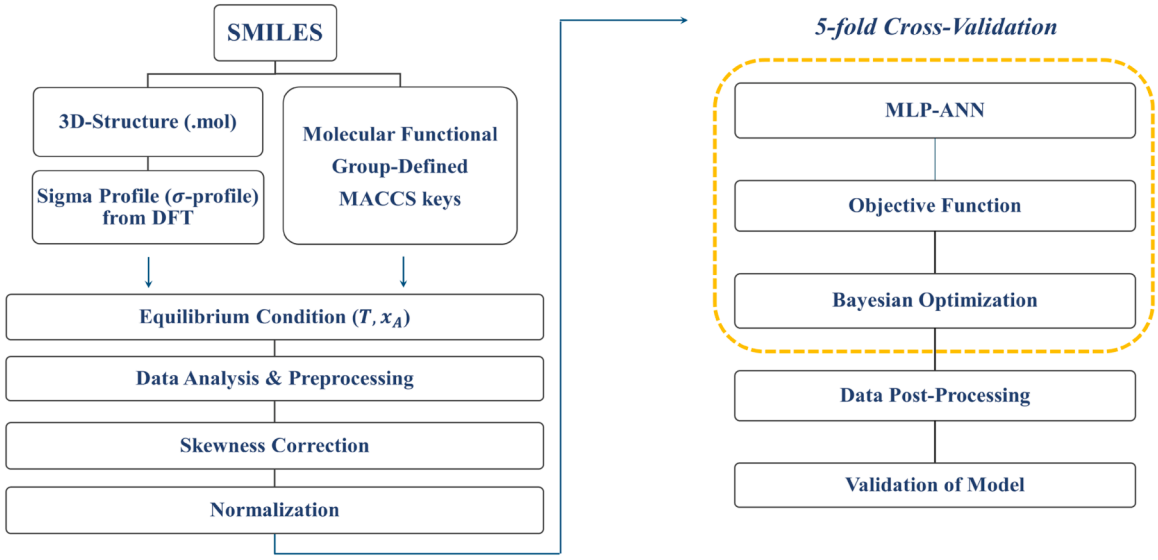


Fig. 2. Conceptual flow chart for the use of the MLP-ANN algorithm with consideration of σ-profile and MACCS keys as input feature in this work.

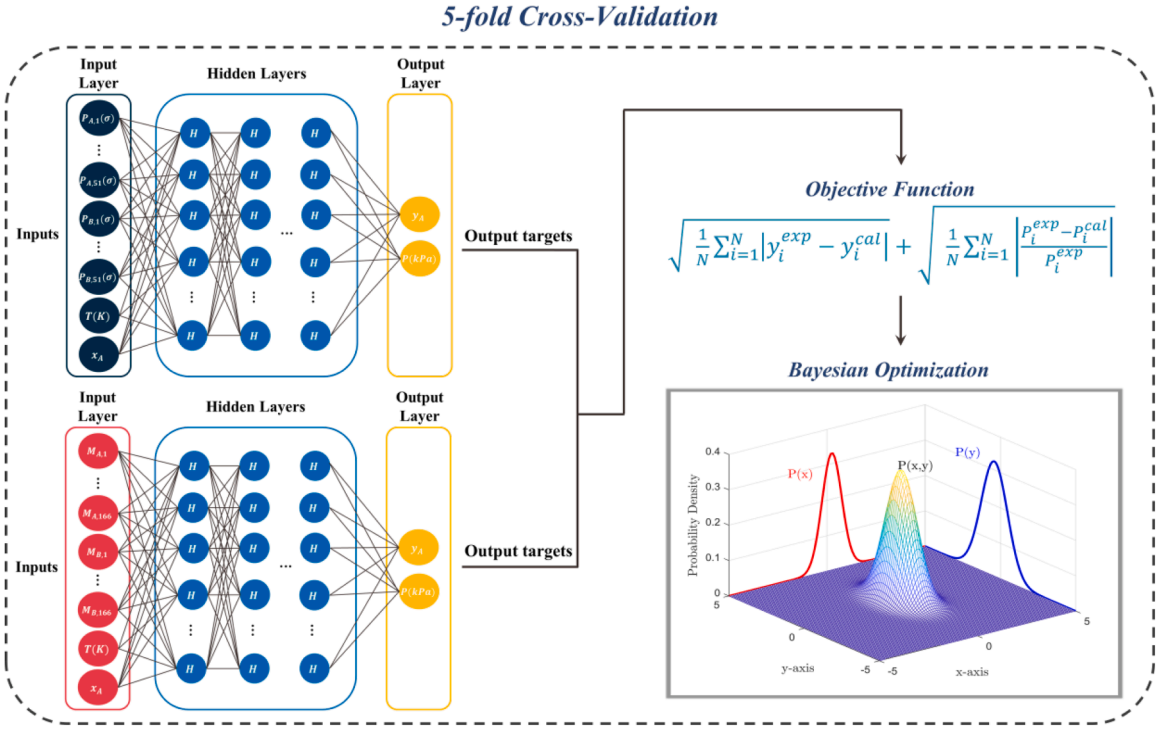


Fig. 3. The architecture of the MLP-ANN framework in the prediction of VLE behavior.

average absolute deviation in mole fraction ($AAD-y(\%)$) and the average absolute relative deviation in pressure ($AARD-P(\%)$). The objective functions used in optimization are defined to minimize the sum of $AAD-y(\%)$and $AARD-P(\%)$. The detailed expressions are as follows.

$$AAD-y(\%)=\left|y_A^{exp}-y_A^{cal}\right|\times\ 100\% \tag{1}$$

$$AARD-P(\%)=\left|\frac{P^{exp}-P^{cal}}{P^{exp}}\right|\times\ 100\% \tag{2}$$

$$\text{obj of VLE}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|y_i^{exp}-y_i^{cal}\right|}+\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|\frac{P_i^{exp}-P_i^{cal}}{P_i^{exp}}\right|} \tag{3}$$

4. Results and discussion

4.1. Model accuracy

As mentioned in the previous section, the skewness of the data distribution plays a crucial role in the construction of machine learning models. First, we discuss the effect of data transformation on model accuracy. Fig. 4 is a combination of a line graph and histograms,

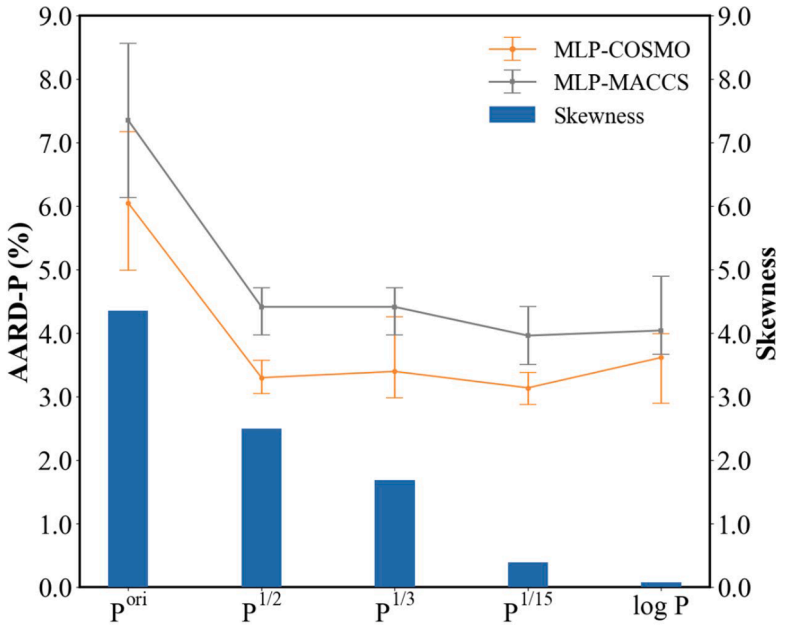


Fig. 4. The relationship between the accuracy of model prediction and the skewness of the pressure data distribution. The line graphs represent $AARD - P(\%)$ using MLP-COSMO and MLP-MACCS for different transformed pressure data distributions. Histograms represent the skewness values for the different transformed pressure data distributions.

demonstrating the relationships between $AARD - P(\%)$ using MLP-COSMO and MLP-MACCS and various transformations of the pressure data distribution (line graph), as well as the skewness values for the different transformed pressure data distributions (histograms). The upper and lower error bars represent the maximum and minimum $AARD - P(\%)$ in the 5-fold cross-validation at various data transformations. It can be seen that the $AARD - P(\%)$ obtained from both developed models is significantly improved with data transformation. In general, lower skewness leads to smaller $AARD - P(\%)$ values, indicating the importance of skewness correction in pressure data distribution. However, even though the logarithmic transformation of P shows the lowest skewness, the $AARD - P(\%)$does not improve further. The variation in the 5-fold cross validation with the log P transformation is even larger than with the root transformation. As a result, the 15th root transformation is chosen for our model construction and data representation.

Table 2 shows the detailed $AARD - P(\%)$, R^2-P, $AAD - y(\%)$ and R^2-y in 5-fold cross validation at $P^{1/15}$ skewness correctionobtained by MLP-COSMO and MLP-MACCS (The performance against other skewness correction was summarized in **Table S2** of the Supplementary Material). The best model for MLP-COSMO and MLP-MACCS shown in the Table 2 are identified by the minimum sums of average $AAD - y(\%)$and average $AARD - P(\%)$ from the 5-fold cross validation. The distribution and range of weights and biases for the model are shown in the **Fig. S1** and **Table S3** of the Supplementary Materials. Table 3 demonstrates the comparison of the precision in predicting vapor composition and pressure in VLE from this work (MLP-COSMO and MLP-MACCS) and the other models found in the literature. Building on the optimized hyperparameters from 5-fold validation, the best MLP-COSMO model in this study achieved an average $AARD - P(\%)$ of 2.16% and an average $AAD - y(\%)$ of 0.64% in the overall dataset. For the test dataset excluded from the training process, the model yielded an average $AARD - P(\%)$ of 2.88% and an average $AAD - y(\%)$ of 1.04% (refer to Table 3). It can be seen that the median $AARD - P(\%)$ and $AAD - y(\%)$ obtained by both developed models are significantly lower than the average values, implying that the dataset contains some outliers. This result also indicates the necessity of considering the mean and the median simultaneously. The median $AARD - P(\%)$ and $AAD - y(\%)$ of the best MLP-COSMO model are both lower than those of the MLP-MACCS model, indicating that the model we trained demonstrates consistency across the entire dataset. Compared to COSMO-SAC (2002) and COSMO-SAC (2010) for VLE prediction, MLP-COSMO demonstrates improvements of 62% and 46% in AARD-P (%) and 65% and 52% in AAD-y (%) for the test dataset, respectively. It is well known that UNIFAC can provide very accurate results for VLE, the best accuracy result of our MLP-COSMO in VLE prediction is comparable to that of UNIFAC. Fig. 5 **(a)** to Fig. 5 **(d)** show parity plots comparing the experimental pressure (P^{exp}) and vapor mole fraction of A (y_A^{exp}) with the predicted values from the MLP-MACCS (Fig. 5 **(a)** and Fig. 5 **(b)**) and MLP-COSMO (Fig. 5 **(c)** and Fig. 5 **(d)**). It can be seen that most of the points are located around the diagonal line through origin, demonstrating the outstanding prediction capability of our developed models. In pressure prediction, the distribution of data points in the high-pressure region (P > 400 kPa) becomes more scattered, indicating a slightly poorer prediction capability in the high pressure region. The corresponding R^2 values for the vapor composition and pressure test data set are 0.9926 and 0.9889, respectively. In the developed models, MLP-COSMO shows exceptionally strong predictive capabilities over other models for the pressure and vapor mole fraction. The results imply that the σ-profiles descriptor containing molecular spatial charge information can significantly improve the prediction accuracy of the equilibrium behavior of the vapor-liquid phase.

To gain deeper insights into the model's performance across various types of molecules, all molecules in the dataset were classified into three main categories: hydrogen bond donors (HB), hydrogen bond acceptors (HB-A), and non-hydrogen bond (non-HB) molecules. The hydrogen bond donor class includes molecules with functional groups such as COOH, NH, and OH. The hydrogen bond acceptor class includes molecules with functional groups such as F, O, and N. In overall, all binary mixtures were categorized into 13 subgroups.

The detailed analysis of the two developed models is shown in Fig. 6. The results reveal that although MLP-COSMO has lower averages and medians in $AARD - P(\%)$ and $AAD - y(\%)$ compared to MLP-MACCS, MLP-COSMO exhibits the biases between systems that form hydrogen bonds (HB) and those that do not. Notably, both models demonstrated lower accuracy for the NH-OH subgroup. Biases in the prediction of VLE for mixtures containing hydrogen bond networks and nitrogen-containing compounds have also been reported in prior literature [25, 55,56]. In COSMO-SAC (2010), the σ-profile of each molecule was

Table 2
The detailed $AARD-P(\%)$, R^2-P, $AAD-y(\%)$ and R^2-y in 5-fold cross validation at $P^{1/15}$ skewness correction obtained by MLP-COSMO and MLP-MACCS.

	Accuracy							
	Methods							
	MLP-COSMO				MLP-MACCS			
Skewness correction	No.	$AARD-P\ (\%)$	R^2-P	$AAD-y\ (\%)$	R^2-y	$AARD-P\ (\%)$	R^2-P	$AAD-y\ (\%)$	R^2-y
$P^{1/15}$	1	3.08	0.9791	1.20	0.9890	*4.42*	*0.9721*	*1.77*	*0.9755*
	2	**2.88**	**0.9889**	**1.04**	**0.9926**	3.91	0.9651	1.41	0.9850
	3	2.98	0.9752	1.15	0.9887	**3.51**	**0.9908**	**1.40**	**0.9856**
	4	*3.37*	*0.9906*	*1.14*	*0.9904*	3.74	0.9821	1.47	0.9827
	5	3.38	0.9773	1.12	0.9981	4.22	0.9920	1.60	0.9815

*The best model for each skewness correction is identified by the minimum sum of $AAD-y(\%)$and $AARD-P(\%)$, highlighted in blue bold font. Conversely, the worst model is identified by the maximum sum of $AAD-y(\%)$and $AARD-P(\%)$, highlighted in red italics font from the 5-fold cross validation.

Table 3
Comparison of accuracy in predicting the vapor composition and pressure in VLE from this work and other models found in the literature.

Methods	N_{sys}	N_{data}	$AARD-P(\%)$ $(MD-AARD-P(\%))^b$	$AAD-y(\%)$ $(MD-AAD-y(\%))^b$	R^2-P $(P_{corr})^c$	R^2-y $(y_{corr})^c$	Avg. cal. time per data (s)
MLP-COSMO[a]	1553	17,656	**2.16**/*2.86* (**1.32**/*2.02*)	**0.64**/*0.80* (**0.36**/*0.47*)	**0.9929**/*0.9924* (**0.9964**/*0.9962*)	**0.9958**/*0.9943* (**0.9979**/*0.9972*)	**0.152**/*0.179*
MLP-MACCS[a]	1553	17,656	**2.89**/*3.66* (**1.59**/*2.26*)	**1.02**/*1.26* (**0.44**/*0.65*)	**0.9893**/*0.9897* (**0.9946**/*0.9949*)	**0.9896**/*0.9877* (**0.9947**/*0.9938*)	**0.160**/*0.182*
COSMO-SAC(2002) [53]	1222	15,491	7.57	2.94	-	-	-
COSMO-SAC(2010) [53]	1222	15,491	5.30	2.15	-	-	-
COSMO-SAC(DHB) [14]	586	-	5.80	2.53	-	-	-
Mod-UNIFAC [58]	1222	15,491	2.56	1.26	-	-	-

[a]The performance of the best and worst models for MLP-COSMO and MLP-MACCS is highlighted in blue bold and red italics, respectively.
b$MD-AARD-P(\%)$ and $MD-AAD-y(\%)$ represent the median of $AARD-P(\%)$ and $AAD-y(\%)$ for N_{data}, respectively.
[c]P_{corr} and y_{corr} represent correlation coefficient of P and y, respectively.

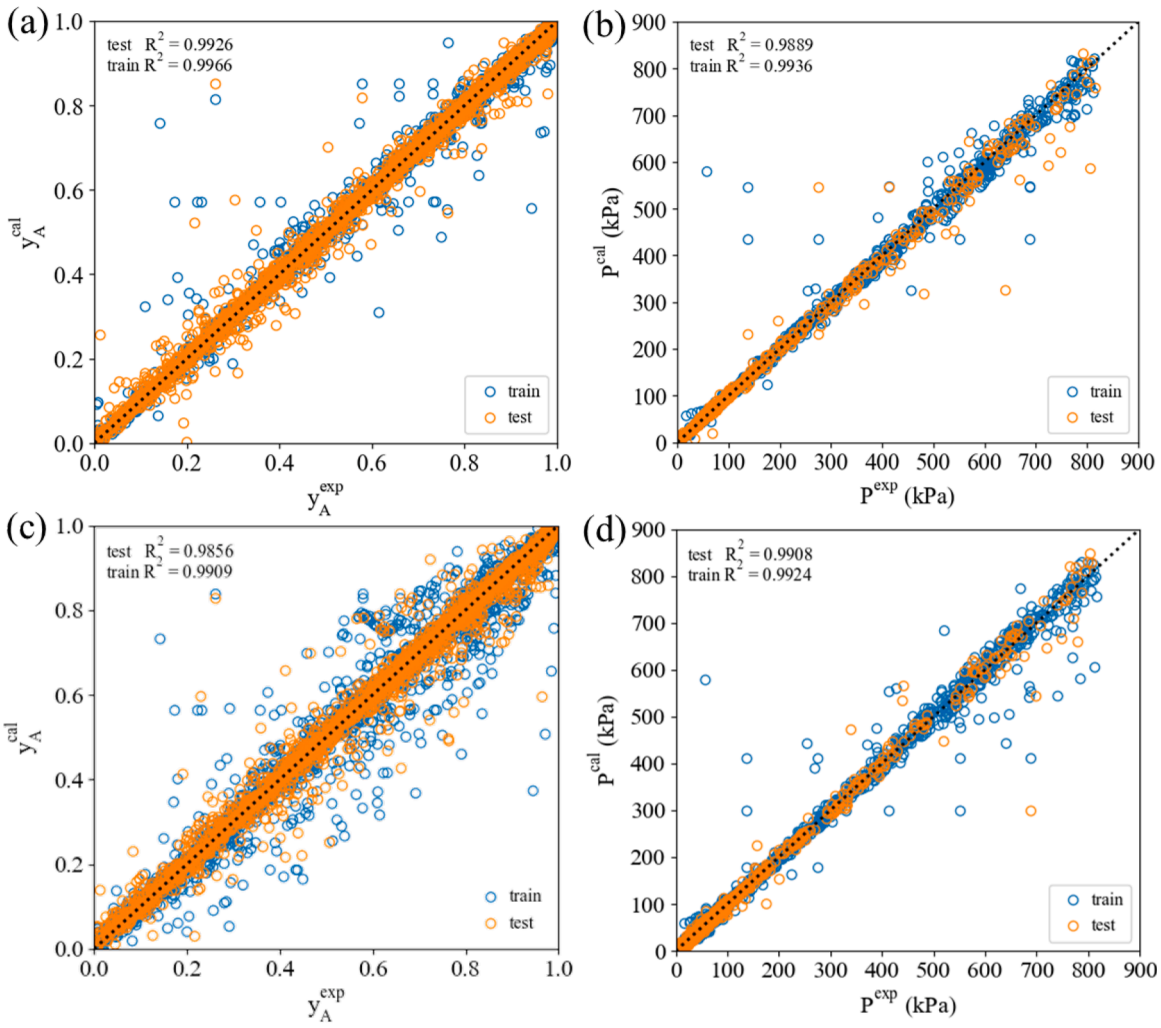


Fig. 5. The parity plot of the experimental and estimated vapor composition and pressure obtained by the best developed model. (a) Vapor composition by using MLP-COSMO (b) pressure by using MLP-COSMO (c) vapor composition by using MLP-MACCS and (d) pressure by using MLP- MACCS.

divided into contributions from non-HB atoms, the hydroxyl group, and from all other types of HB donating and accepting atoms [16]. Incorporating the σ-profile from COSMO-SAC (2010) to enrich hydrogen bond information in molecular descriptors could be a potential strategy for future model improvements.

4.2. *Prediction of VLE phase diagrams*

To evaluate the prediction capability and the application scopes of the developed ML models, the VLE datasets were categorized into the associating and non-associating system. For systems composed of associating molecules, this indicates that, in this binary system, at least one component contains a hydrogen bond (HB) donor group, such as an OH or NH group. For the non-associating binary system, neither component contains an HB donor group. To further validate the predictive capacity of the model, we selected two binary systems that were not present in the original dataset. We compare the predictions of the VLE phase diagram obtained using MLP-COSMO, MLP-MACCS, COSMO-SAC (2002), COSMO-SAC (2010), and Mod-UNIFAC. The predicted P-xy phase diagrams of specific mixtures in the dataset, as determined by MLP-COSMO and COSMO-SAC (2010), are shown in Figs. 7 to 9 The comparisons of the predicted P-xy phase diagrams by MLP-COSMO and MLP-MACCS, MLP-COSMO and COSMO-SAC (2002), and MLP-COSMO and Mod-UNIFAC are provided in **Fig. S2** to **S4** of the Supplementary Materials.

The Fig. 7 **(a)** and Fig. 7 **(b)** present the typical zeotropic mixtures, consisting of water – methanol and trimethylamine – diethylamine, respectively. For both mixtures, the molecules involved were associated through HB (HB–HB), resulting in a positive deviation from Raoult's law. From the results represented, the MLP-COSMO model demonstrated a more accurate predictive capability in the VLE phase diagram for both associating systems with temperature range from 298.15 to 343.15 K. The COSMO-SAC (2010) generally resulted in better prediction for the liquid composition, but larger deviation in the vapor composition. In contrast, the MLP-COSMO exhibited a mote-balanced and accurate prediction for the gas and liquid phases.

For the systems shown in Fig. 8 **(a)** and Fig. 8 **(b)**, which demonstrate the typical behavior of azeotropic mixtures, composed of propanol-undecanol (HB – HB) and hexane–methylethylketone (non-HB – HB), respectively. For the binary mixture propanol - undecanol, both the MLP-COSMO and COSMO-SAC (2010) models provide good prediction for the azeotropic point. However, the prediction of the azeotropic mole fraction for the hexane - methylethylketone system by COSMO-SAC (2010) is slightly overestimated. In contrast, MLP-COSMO demonstrates better predictive ability for systems containing azeotropes, both in terms of the location of the azeotropic point and the temperature dependence of the phase diagram. Finally, the Fig. 9 **(a)** and Fig. 9 **(b)** show the binary systems of water – 2-methoxyethanol (HB – HB) and water – 2-butoxyethanol (HB – HB) [57], the both systems were not included in our dataset. As shown in Fig. 9 **(a)** and Fig. 9 **(b)**, the MLP-COSMO demonstrates superior prediction capability over

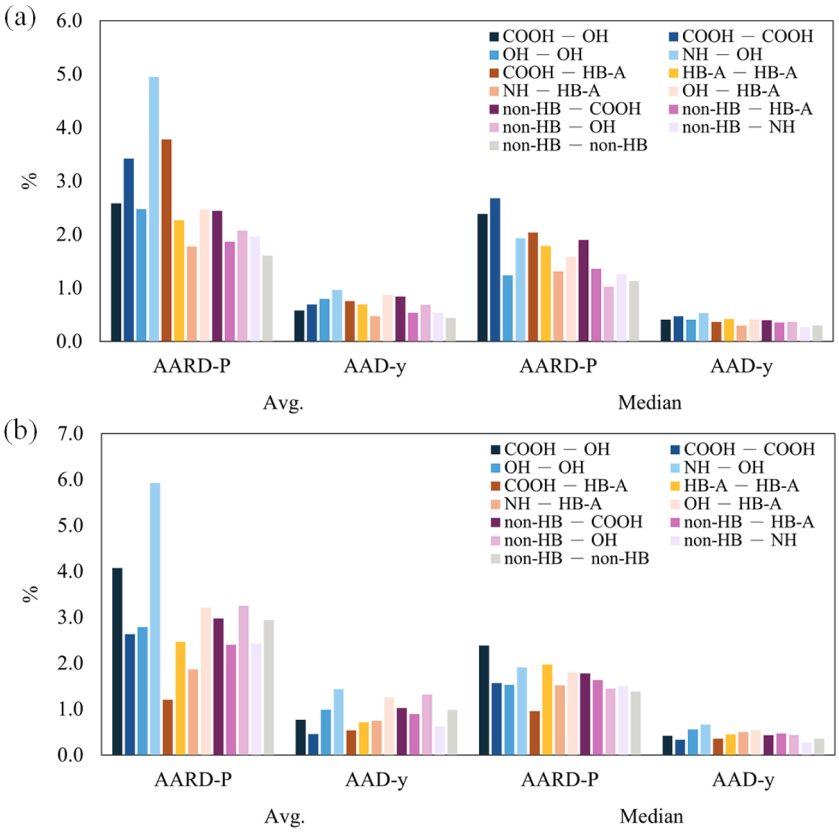


Fig. 6. The performance of (a) MLP-COSMO and (b) MLP-MACCS models on different pairs classified based on the various functional groups of the molecules.

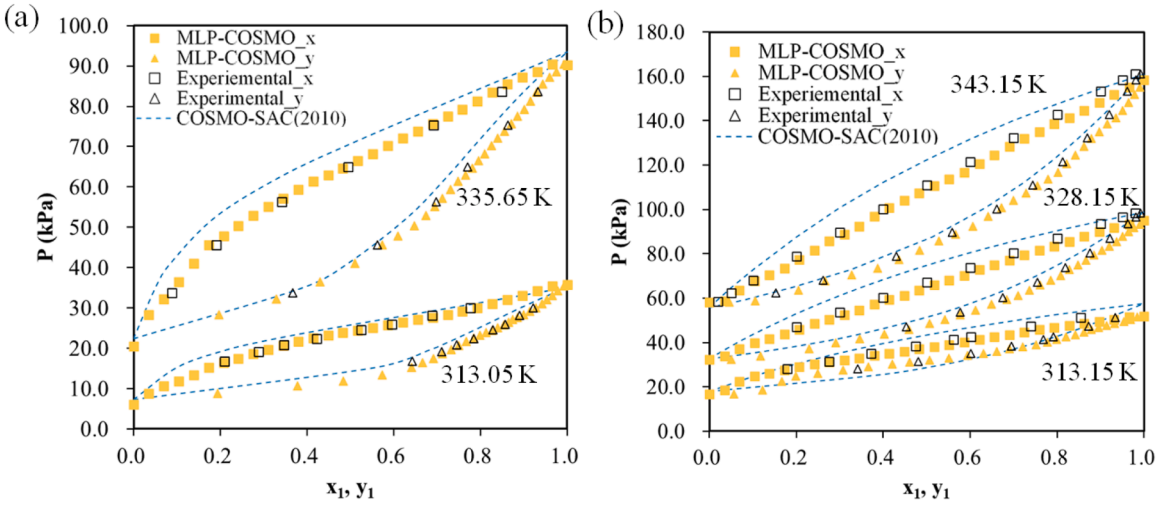


Fig. 7. Vapor-liquid equilibrium diagram of (a) water - methanol, (b) trimethylamine – diethylamine obtained from the best MLP-COSMO and COSMO-SAC (2010) modeling, compared to experimental data.

COSMO-SAC (2010), even for systems not included in the previous ML model training.

5. Conclusions

In this work, deep learning-based MLP-ANN was developed for highly accurate prediction in a binary vapor-liquid equilibrium dataset consisting of 169 compounds, 518 mixtures (the list of the mixtures is shown in the **Table S4** of the Supplementary Materials) and 17,656 raw data, with pressure ranging from 0.947 to 817 kPa and temperature ranging from 199.93 to 548.15 K. Two molecular descriptors, the physically guided σ-profile and the chemical structural MACCS, were selected to represent the features of the input layer during the training of the ML model training. The MLP-COSMO model provides R^2-P of 0.9889, $AARD-P(\%)$ of 2.88% for pressure, and R^2-y of 0.9926, $AAD-y(\%)$ of 1.04% for vapor composition. The $AARD-P(\%)$ and $AAD-y$ (%) obtained from MLP-COSMO show improvements of 46% and 52%, respectively, compared to COSMO-SAC (2010) and are comparable to those of UNIFAC. In the same ML algorithm (MLP-ANN), using the σ-profiles as input features leads to better prediction accuracy of VLE phase behavior over structural MACCS keys. Furthermore, for selected molecular pairs that are not in the dataset but have temperature and

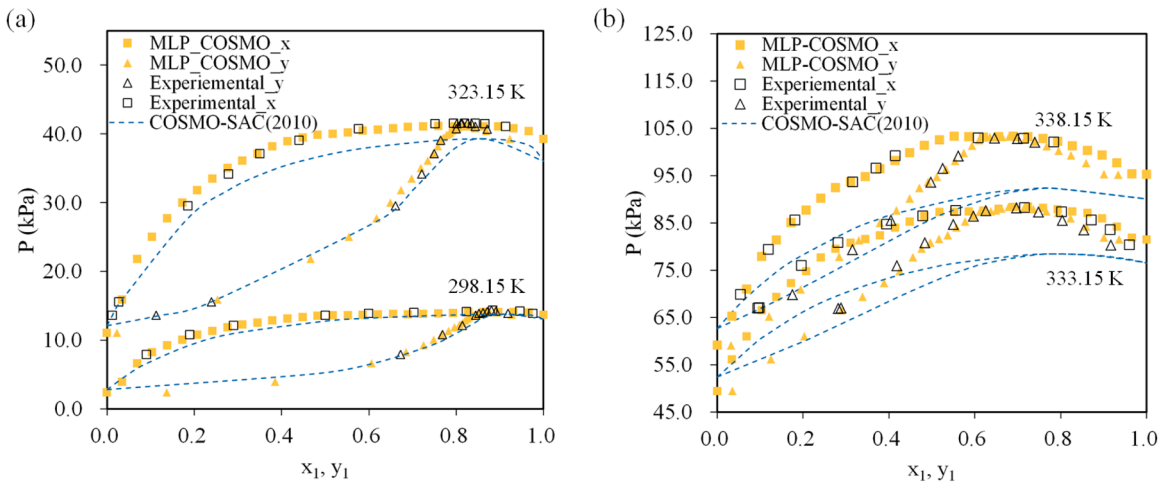


Fig. 8. Vapor-liquid equilibrium diagram of (a) propanol – undecanol and (b) hexane – methylethylketone obtained from the best MLP-COSMO and COSMO-SAC (2010) modeling, compared to experimental data.

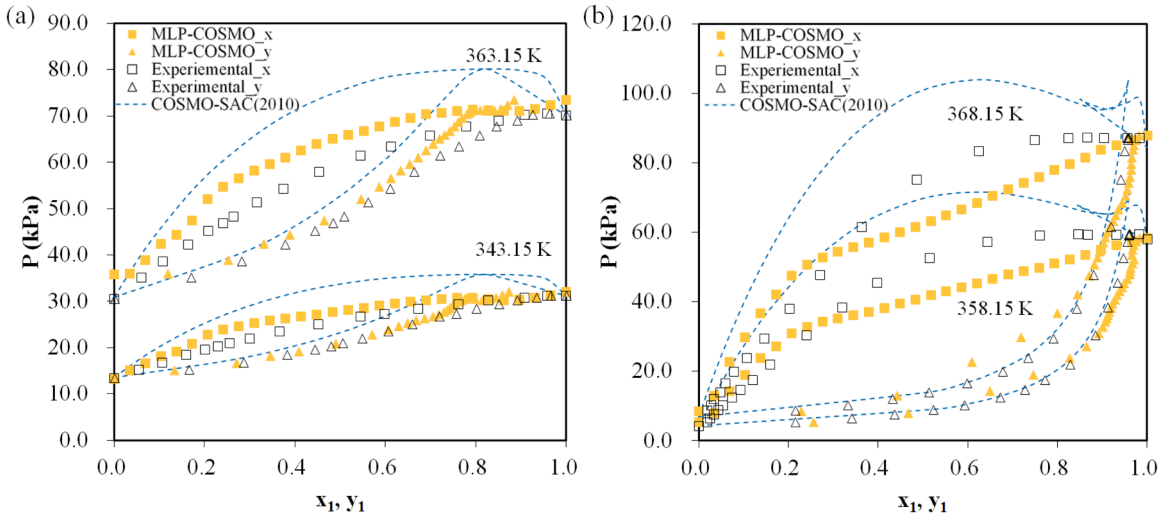


Fig. 9. Vapor-liquid equilibrium diagram of (a) water–2-methoxyethanol and (b) water - 2-butoxyethanol obtained from the best MLP-COSMO and COSMO-SAC (2010) modeling, compared to experimental data [57].

pressure within the training range, MLP-COSMO still demonstrates good predictive ability. In summary, we have successfully developed and investigated ML models based on MLP-ANN algorithms. Our models not only provided highly accurate predictions but also offered a satisfactory first-principle solution in VLE phase behavior [58].

CRediT authorship contribution statement

Hsiu-Min Hung: Writing – original draft, Validation, Methodology, Investigation, Data curation, Conceptualization. **Ying-Chieh Hung:** Writing – review & editing, Supervision, Resources, Project administration.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Ying-Chieh Hung reports financial support was provided by National Science and Technology Council. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper

Acknowledgements

This study is supported by the National Science and Technology Council (112–2222-E-027–001 -). We also acknowledge computational resources from the National Center for High-Performance Computing.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jtice.2025.106054.

Reference

[1] Renon H, Prausnitz JM. Estimation of parameters for the NRTL equation for excess gibbs energies of strongly nonideal liquid mixtures. Ind Eng Chem Process Des Devel 1969;8:413–9.
[2] Wilson G. Vapor-liquid equilibria, correlation by means of a modified Redlich-Kwong equation of state. In: Advances in Cryogenic Engineering: Proceedings of the 1963 Cryogenic Engineering Conference University of Colorado College of Engineering and National Bureau of Standards Boulder Laboratories Boulder, Colorado August 19–21. 1964. Springer; 1963. p. 168–76.
[3] Anderson TF, Prausnitz JM. Application of the UNIQUAC equation to calculation of multicomponent phase equilibria. 1. Vapor-liquid equilibria. Ind Eng Chem Process Des Devel 1978;17:552–61.

[4] Weidlich U, Gmehling J. A modified UNIFAC model. 1. Prediction of VLE, hE, and. Gamma.. infin. Ind Eng Chem Res 1987;26:1372–81.
[5] Lohmann J, Joh R, Gmehling J. From UNIFAC to modified UNIFAC (Dortmund). Ind Eng Chem Res 2001;40:957–64.
[6] Klamt A, Jonas V, Bürger T, Lohrenz JCW. Refinement and parametrization of COSMO-RS. J Phys Chem A 1998;102:5074–85.
[7] Klamt A, Eckert F. COSMO-RS: a novel and efficient method for the a priori prediction of thermophysical data of liquids. Fluid Phase Equilib 2000;172:43–72.
[8] Klamt A. Conductor-like screening model for real solvents: A new approach to the quantitative calculation of solvation phenomena. J Phys Chem 1995;99:2224–35.
[9] Klamt A. The COSMO and COSMO-RS solvation models. Wiley Interd Rev Comput Mol Sci 2011;1:699–709.
[10] Klamt A, Eckert F, Arlt W. COSMO-RS: an alternative to simulation for calculating thermodynamic properties of liquid mixtures. Annu Rev Chem Biomol Eng 2010;1:101–22.
[11] Lin S-T, Sandler SI. A Priori phase equilibrium prediction from a segment contribution solvation model. Ind Eng Chem Res 2002;41:899–913.
[12] Lin S-T, Chang J, Wang S, Goddard WA, Sandler SI. Prediction of vapor pressures and enthalpies of vaporization using a COSMO solvation model. J Phys Chem A 2004;108:7429–39.
[13] Hsieh C-M, Lin S-T. Prediction of liquid–liquid equilibrium from the Peng–Robinson+COSMOSAC equation of state. Chem Eng Sci 2010;65:1955–63.
[14] Chang C-K, Chen W-L, Wu DT, Lin S-T. Improved directional hydrogen bonding interactions for the prediction of activity coefficients with COSMO-SAC. Ind Eng Chem Res 2018;57:11229–38.
[15] Fingerhut R, Chen W-L, Schedemann A, Cordes W, Rarey J, Hsieh C-M, Vrabec J, Lin S-T. Comprehensive assessment of COSMO-SAC models for predictions of fluid-phase equilibria. Ind Eng Chem Res 2017;56:9868–84.
[16] Hsieh C-M, Sandler SI, Lin S-T. Improvements of COSMO-SAC for vapor–liquid and liquid–liquid equilibrium predictions. Fluid Phase Equilib 2010;297:90–7.
[17] Tung H-H, Tabora J, Variankaval N, Bakken D, Chen C-C. Prediction of pharmaceutical solubility via NRTL-SAC and COSMO-SAC. J Pharm Sci 2008;97:1813–20.
[18] Hsieh C-M, Wang S, Lin S-T, Sandler SI. A predictive model for the solubility and octanol−water partition coefficient of pharmaceuticals. J Chem Eng Data 2011;56:936–45.
[19] Shu C-C, Lin S-T. Prediction of drug solubility in mixed solvent systems using the COSMO-SAC activity coefficient model. Ind Eng Chem Res 2011;50:142–7.
[20] Kataoka T, Hao Y, Hung YC, Orita Y, Shimoyama Y. Prediction of biphasic separation in CO2 absorption using a molecular surface information-based machine learning model. Environ Sci Processes Impacts 2022;24:2409–18.
[21] Hung Y-C, Hsieh C-M, Machida H, Lin S-T, Shimoyama Y. Unveiling the mechanism of CO2-driven phase change in amine + water + glycol ether ternary mixture. J Taiwan Inst Chem Eng 2022;131:104143.
[22] Hung Y-C, Hsieh C-M, Machida H, Lin S-T, Shimoyama Y. Towards design of phase separation solvent for CO2 capture using COSMO-SAC model. J Mol Liq 2021;336:116229.
[23] Lee B-S, Lin S-T. Screening of ionic liquids for CO2 capture using the COSMO-SAC model. Chem Eng Sci 2015;121:157–68.
[24] Lin S-T, Sandler SI. Prediction of octanol−water partition coefficients using a group contribution solvation model. Ind Eng Chem Res 1999;38:4081–91.
[25] Xiong R, Sandler SI, Burnett RI. An improvement to COSMO-SAC for predicting thermodynamic properties. Ind Eng Chem Res 2014;53:8265–78.
[26] Boobier S, Hose DRJ, Blacker AJ, Nguyen BN. Machine learning with physicochemical relationships: solubility prediction in organic solvents and water. Nat Commun 2020;11.
[27] Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T. The rise of deep learning in drug discovery. Drug Discov Today 2018;23:1241–50.
[28] Goldsmith BR, Esterhuizen J, Liu J-X, Bartel CJ, Sutton C. Machine learning for heterogeneous catalyst design and discovery. AICHE J 2018;64:2311–23.
[29] Ma Y, Gao Z, Shi P, Chen M, Wu S, Yang C, Wang J, Cheng J, Gong J. Machine learning-based solubility prediction and methodology evaluation of active pharmaceutical ingredients in industrial crystallization. Front Chem Sci Eng 2022;16:523–35.
[30] Mu Y, Dai T, Fan J, Cheng Y. Prediction of acetylene solubility by a mechanism-data hybrid-driven machine learning model constructed based on COSMO-RS theory. J Mol Liq 2024:126194.
[31] Aniceto JP, Zêzere B, Silva CM. Prediction of diffusion coefficients in aqueous systems by machine learning models. J Mol Liq 2024;405:125009.
[32] Roosta A, Haghbakhsh R, Rita C. Duarte A, Raeissi S. Deep eutectic solvent viscosity prediction by hybrid machine learning and group contribution. J Mol Liq 2023;388:122747.
[33] Cheng Y-H, Sung IT, Hsieh C-M, Lin L-C. Module-based machine learning models using sigma profiles of organic linkers to predict gaseous adsorption in metal-organic frameworks. J Taiwan Inst Chem Eng 2024;165:105728.
[34] Zain NS, Mahmoud MHH, Khan MI, Zafar F, Manzoor S, Akhtar N, Khan MA, El Azab IH, El-Bahy ZM. Machine learning-assisted optimization and evaluation of methylene blue adsorption kinetics on citrus aurantifolia leaves: insights from isotherm and thermodynamic studies. J Taiwan Inst Chem Eng 2024;164:105696.
[35] Mohanty S. Estimation of vapour liquid equilibria of binary systems, carbon dioxide–ethyl caproate, ethyl caprylate and ethyl caprate using artificial neural networks. Fluid Phase Equilib 2005;235:92–8.
[36] Vaferi B, Lashkarbolooki M, Esmaeili H, Shariati A. Toward artificial intelligence-based modeling of vapor liquid equilibria of carbon dioxide and refrigerant binary systems. J Serb Chem Soc 2018;83:199–211.
[37] Sun G, Zhao Z, Sun S, Ma Y, Li H, Gao X. Vapor-liquid phase equilibria behavior prediction of binary mixtures using machine learning. Chem Eng Sci 2023;282:119358.
[38] Mesbah M, Soroush E, Azari V, Lee M, Bahadori A, Habibnia S. Vapor liquid equilibrium prediction of carbon dioxide and hydrocarbon systems using LSSVM algorithm. J Supercrit Fluids 2015;97:256–67.
[39] Hao Y, Hung YC, Shimoyama Y. Investigating spatial charge descriptors for prediction of cocrystal formation using machine learning algorithms. Cryst Growth Des 2022;22:6608–15.
[40] Abranches DO, Zhang Y, Maginn EJ, Colón YJ. Sigma profiles in deep learning: towards a universal molecular descriptor. Chem Commun 2022;58:5630–3.
[41] Li J-E, Chien S-C, Hsieh C-M. Modeling solid solute solubility in supercritical carbon dioxide by machine learning algorithms using molecular sigma profiles. J Mol Liq 2024;395:123884.
[42] Mohan M, Demerdash O, Simmons BA, Smith JC, Kidder MK, Singh S. Accurate prediction of carbon dioxide capture by deep eutectic solvents using quantum chemistry and a neural network. Green Chem 2023;25:3475–92.
[43] Abdi-Khanghah M, Bemani A, Naserzadeh Z, Zhang Z. Prediction of solubility of N-alkanes in supercritical CO2 using RBF-ANN and MLP-ANN. J CO2 Utilizat 2018;25:108–19.
[44] Amani M, Saadati Ardestani N. Investigation of the solubility of anticancer drugs in the supercritical solvent for development of innovative drug delivery systems; artificial intelligence paradigms (MLP-ANN) and thermodynamic correlations. J Mol Liq 2024;394:123701.
[45] Hekayati J, Rahimpour MR. Estimation of the saturation pressure of pure ionic liquids using MLP artificial neural networks and the revised isofugacity criterion. J Mol Liq 2017;230:85–95.
[46] Ghasemi M, Qazani MRC, Lennon CW, Sedighi M. Analysis and prediction of microbial fuel cell behaviour using MLP and SVR. J Taiwan Inst Chem Eng 2023;151:105101.
[47] Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M. arXiv preprint. 2016.
[48] Chawla NV, Japkowicz N, Kotcz A. Special issue on learning from imbalanced data sets. ACM SIGKDD Explorat Newsletter 2004;6:1–6.
[49] Liu Y, An A, Huang X. Boosting prediction accuracy on imbalanced datasets with SVM ensembles. In: Advances in knowledge discovery and data mining: 10th Pacific-Asia conference, PAKDD 2006. Singapore: Springer; 2006. p. 107–18. April 9-12, 2006. Proceedings 10.
[50] Sen S, Singh KP, Chakraborty P. Dealing with imbalanced regression problem for large dataset using scalable Artificial Neural Network. New Astronomy 2023;99:101959.
[51] Gmehling J, Onken U, Arlt W. Vapor-liquid equilibrium data collection. DECHEMA. Chem Data Ser 1977;1.
[52] M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, G. Scalmani, V. Barone, G.A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A.V. Marenich, J. Bloino, B.G. Janesko, R. Gomperts, B. Mennucci, H.P. Hratchian, J.V. Ortiz, A.F. Izmaylov, J.L. Sonnenberg, Williams, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V.G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J.A. Montgomery Jr., J.E. Peralta, F. Ogliaro, M.J. Bearpark, J.J. Heyd, E.N. Brothers, K.N. Kudin, V.N. Staroverov, T.A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A.P. Rendell, J.C. Burant, S.S. Iyengar, J. Tomasi, M. Cossi, J.M. Millam, M. Klene, C. Adamo, R. Cammi, J.W. Ochterski, R.L. Martin, K. Morokuma, O. Farkas, J.B. Foresman, D.J. Fox, Gaussian 16 rev. A.03, in, Wallingford, CT, 2016.
[53] Chen W-L, Hsieh C-M, Yang L, Hsu C-C, Lin S-T. A critical evaluation on the performance of COSMO-SAC models for vapor–Liquid and liquid–Liquid equilibrium predictions based on different quantum chemical calculations. Ind Eng Chem Res 2016;55:9312–22.
[54] Bento AP, Hersey A, Félix E, Landrum G, Gaulton A, Atkinson F, Bellis LJ, De Veij M, Leach AR. An open source chemical structure curation pipeline using RDKit. J Cheminform 2020;12:1–16.
[55] Tsai C-C, Lin S-T. Improved vapor pressure prediction from PR + COSMOSAC EOS using normal boiling temperature. AICHE J 2023;69:e17997.
[56] Lin Y-H, Liang H-H, Lin S-T, Li Y-P. Advancing vapor pressure prediction: A machine learning approach with directed message passing neural networks. J Taiwan Inst Chem Eng 2024:105926.
[57] Chiavone-Filho O, Proust P, Rasmussen P. Vapor-liquid equilibria for glycol ether+water systems. J Chem Eng Data 1993;38:128–31.
[58] Gmehling J, Lohmann J, Jakob A, Li J, Joh R. A modified UNIFAC (Dortmund) model. 3. Revision and extension. Ind Eng Chem Res 1998;37:4876–82.