Screening of phase-separation CO_2 absorbent using machine learning combined with molecular information

Taishi Kataoka, Yingquan Hao, Ying-Chieh Hung, Yusuke Shimoyama*

Department of Chemical Science and Engineering, Tokyo Institute of Technology, 2-12-1 S1-33, Ookayama, Meguro-ku, Tokyo 152-8550, Japan

* To whom correspondence should be addressed.
E-mail : yshimo@chemeng.titech.ac.jp Tel : +81-3-5734-3285

Abstract

Carbon dioxide capture technologies have been focused to overcome global warming. Although a chemical absorption process using aqueous alkanolamine solvents is one of the most established technologies for CO_2 capture, it consumes a large amount of energy due to its high desorption temperature. Recently, to reduce energy consumption, phase-separation absorbents have been reported. When the absorbents react with CO_2, a single liquid phase transforms into two liquid phases. This unique phase behavior assists the desorption of CO_2, decreasing the desorption temperature from 120 ˚C of the conventional chemical absorption process to 80 ˚C. However, to find the phase-separation absorbents, screening experiments are needed because of its unique phase behaviors. In this work, we developed a screening method of the phase-separation absorbents using supervised machine learning models: random forest, logistic regression, and support vector machine. There are 61 mixed-solvent absorbents containing alkanolamine/glycol ether or alcohol in a dataset. Extended-connectivity fingerprints, physical molecular properties, and molecular surface charge distributions are used as molecular descriptors in the machine learning models. The machine learning models successfully predicted the phase behaviors of the mixed-solvent absorbents with accuracy of approximately 90 %. Furthermore, we analyzed contributions of explanatory variables to predict the phase states in the model.

Keywords

CO_2 capture, phase-separation absorbent, machine learning

1. Introduction

Carbon dioxide capture technologies which can reduce CO_2 emissions attract much attention because CO_2 concentration in the atmosphere has been increasing. One of the most established CO_2 capture technologies is a chemical absorption process using aqueous alkanolamine solvents. This process can absorb CO_2 with low partial pressure from power plants [1, 2]. However, amine species in this process react with CO_2 to form carbamate species which need high temperatures and large energy to desorb CO_2 from the solvents [3].

A phase-separation CO_2 absorbent is one of the most attractive CO_2 capture technologies which are expected to reduce CO_2 desorption energy [4]. When this solvent reacts with CO_2, one liquid phase of the absorbent transforms into two liquid phases, CO_2-rich and CO_2-lean phases. In the absorption process, the phase-separated absorbent accelerates absorption [5]. Moreover, in the desorption process, the phase-separated solvent is mixing and getting to a homogeneous phase, which promotes CO_2 desorption. Machida et al. reported that the phase-separation capture process reduces desorption temperature from 120 ˚C of the conventional chemical absorption process to 80 ˚C, resulting in the reduction of desorption energy by approximately 50 % [6]. However, it is difficult to know phase behaviors of the mixed-solvent absorbents. Therefore, screening experiments for long time are required for finding candidates as phase-separation absorbents.

In this work, we developed a screening method of the phase-separation absorbents using supervised machine learning models: random forest (RF), logistic regression (LGR), and support vector machine (SVM). The mixed-solvent absorbents containing alkanolamine/glycol ether or alcohol are candidates for the phase-separation absorbents in this work because little amount of water in the solvent reduces latent and sensitive heat during CO_2 desorption [7]. Phase behaviors of the mixed-solvent absorbents in a dataset are obtained from the literatures. Molecular descriptors in this model include molecular fingerprints, two-dimensional RDkit descriptors, and molecular surface charge distributions. Furthermore, we analyzed the contributions of explanatory variables to predict the phase states.

2. Computational method

2.1. Dataset

A dataset contains 61 phase behaviors of three classes of mixed-solvent absorbents: 34 amine/glycol ether [7], 6 amine/alcohol [8], and 21 amine/glycol ether/water solvents [9]. Three types of phase behaviors exist in a dataset: phase separation type which becomes phase-separated by CO_2 absorption, miscible type which is homogeneous before and after absorption, and immiscible type which is phase-separated before and after absorption. Weight fractions of amine species in the solvents were 30 wt%. Besides, all the experiment data in the literature were obtained at approximately 313 K. Table 1 summarizes all solvents included in a dataset.

Table 1
Chemicals in a dataset.

No.	Solvent	Abbreviation	Type
1	2-(2-Aminoethoxy)ethanol	AEE	Alkanolamine
2	2-Amino-1-butanol	AM2B	Alkanolamine
3	1-Amino-2-propanol	AM2P	Alkanolamine
4	2-Amino-1-methoxybutane	AMB	Alkanolamine
5	2-(Butylamino)ethanol	BAE	Alkanolamine
6	2-(Benzylamino)ethanol	BZMEA	Alkanolamine
7	Diethanolamine	DEA	Alkanolamine
8	2-(Ethylamino)ethanol	EAE	Alkanolamine
9	2-(isopropylamino)ethanol	IPMEA	Alkanolamine
10	2-(Methylamino)ethanol	MAE	Alkanolamine
11	Monoethanolamine	MEA	Alkanolamine
12	3-Amino-1-propanol	MPA	Alkanolamine
13	Diethylene glycol diethyl ether	DEGDEE	Glycol ether
14	Diethylene glycol dimethyl ether	DEGDME	Glycol ether
15	Diethylene glycol ethyl methyl ether	DEGEME	Glycol ether
16	Diethylene glycol monoethyl ether	DEGMEE	Glycol ether
17	Diethylene glycol methyl ether	DEGMME	Glycol ether
18	Ethylene glycol butyl ether	EGBE	Glycol ether
19	1-Heptanol	Heptanol	Alcohol
20	Isooctanol	Isooctanol	Alcohol
21	1-Octanol	Octanol	Alcohol

2.2. Molecular descriptor

The molecular descriptors in this work were calculated using RDkit [10] and TURBOMOLE 6.5 [11] software. RDkit software was used to obtain extended-connectivity fingerprint (ECFP) [12], topological polar surface area (TPSA) [13], octanol-water partition coefficient (logP) [14], Labute's approximate surface area (Labute ASA), and 12 SlogP_VSA descriptors which combine logP and Labute ASA [15]. ECFP is a class of topological fingerprints for molecular characterization. TPSA and logP are two-dimensional RDkit descriptors calculated by functional group contributions. Labute ASA is a conformation independent three-dimensional property that requires only two-dimensional connection information. SlogP_VSA descriptors intend to capture hydrophobic and hydrophilic effects.

The large value of SlogP_VSA1 means the large molecular surface area which is hydrophilic, whereas the large value of SlogP_VSA12 means the large molecular surface area which is hydrophobic. TURBOMOLE 6.5 software was used to conduct a quantum chemical calculation based on a conductor-like screening model (COSMO) [16] and to obtain molecular surface charge profiles (σ-profile), molecular surface area, and molecular volume. The σ-profile of each solvent has 25 segments whose range is from -0.03 to 0.03 e Å^{-2}. Table 2 summarizes four sets of molecular descriptors.

Table 2
Molecular descriptors in this work.

Descriptor	Content	Calculation
COSMO	σ-profile and molecular volume	COSMO
ECFP	ECFP2	RDkit
RDkit	TPSA, logP, and logP_VSA, Labute ASA	RDkit
ECFP + RDkit	ECFP2, TPSA, logP, Labute ASA, and logP_VSA	RDkit

2.3. Model development

In this work, the phase behaviors of the mixed-solvent absorbents are predicted by combining the predicted results for the phase states before and after CO_2 absorption. First, the model predicts whether the solvent containing the unreacted amine forms a homogeneous phase or not. Then, the model predicts whether the solvent containing the amine that reacted with CO_2 forms a homogeneous phase or not. Here, the mixed-solvent absorbent that forms a homogeneous phase before CO_2 absorption and separated phases after absorption is classified as a phase-separation type absorbent. To obtain molecular information of amine after CO_2 absorption, we assumed all amine molecules react with CO_2 to form carbamate and protonated amine as shown in the following equation [9].

$$2R_1R_2\mathrm{NH} + \mathrm{CO_2} \rightarrow R_1R_2\mathrm{NH_2^+} + R_1R_2\mathrm{NCOO^-} \quad (1)$$

where R_1 and R_2 represent substituents, and R_1R_2NH, $R_1R_2NH_2^+$, and $R_1R_2NCOO^-$ neutral amine, protonated amine, and carbamate, respectively. When the mixed-solvent absorbents contain three components including water, σ-profiles of amine are converted as the following equation using σ-profile of water.

$$p_{mix}(\sigma) = \frac{x_{\mathrm{Amine}}A_{\mathrm{Amine}}p_{\mathrm{Amine}}(\sigma) + x_{\mathrm{Water}}A_{\mathrm{Water}}p_{\mathrm{Water}}(\sigma)}{x_{\mathrm{Amine}}A_{\mathrm{Amine}} + x_{\mathrm{Water}}A_{\mathrm{Water}}} \quad (2)$$

where x_C, A_C, and $p_C(\sigma)$ are the mole fraction, molecular surface area, and σ-profile for each solvent C.

Three machine learning models, RF, SVM, and LGR, are used to predict the phase behaviors of the mixed-solvent absorbents. A scikit-learn library in python is used for machine learning [17]. RF is an ensemble learning model based on decision trees [18]. SVM is a flexible model to solve regression and classification problems [19]. LGR is represented as a linear classification model using a log-odds unit, logit, as follows [20].

$$p = \frac{1}{1 + \exp(-\boldsymbol{w}^{\top}\boldsymbol{x})} \tag{3}$$

where p, $\boldsymbol{w} = \{w_0, w_1, w_2, \ldots, w_m\}$, $\boldsymbol{x} = \{1, x_1, x_2, \ldots, x_m\}$ represents probabilities that objective variable is positive ($y = 1$), regression coefficient, and explanatory variables, respectively. Here, m denotes the number of explanatory variables. The value of the loss function is minimized to obtain the optimal value of $\boldsymbol{w}$.

Explanatory variables $\boldsymbol{x}$ are composed of the molecular descriptors of amine and glycol ether species, water content, and mole fraction of amine species. All explanatory variables are rescaled to ensure the mean and the standard deviation to be 0 and 1, respectively. Objective variable y represents the phase state, homogeneous or not. When the phase state is homogenous, the objective variable y is represented as 0.

The model was evaluated using the following procedure: (A) one absorbent is picked as a test set; (B) a training set is composed of all absorbents except the ones containing the solvents included in the absorbent picked as a test set, and the hyperparameters of the model are optimized by grid search with 5-fold cross-validation in the training set; (C) the model developed by the training set is used to predict the phase behaviors of the absorbent picked as the test set; (D) another absorbent is picked as a test set, and the procedures of (A) to (C) are repeated until all absorbents are predicted.

3. Results and discussion

3.1. Predicted results

Table 3 summarizes the predicted results for phase behaviors of the mixed-solvent absorbents. Accuracy is the fraction of the total samples which are correctly classified by the model as the following equation.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \tag{4}$$

The accuracy of the model using the COSMO descriptor is higher than the other models. On the other hand, the model using only ECFP has the lowest accuracy of the four descriptors. Herein, the COSMO and RDkit descriptors contain information on polarity and the ECFP contains information on chemical structure. From the accuracy of each model, information on polarity is important to predict the phase behaviors of the mixed-solvent absorbents.

Furthermore, the RDkit and EFCP descriptors of all molecules in this work were calculated within 0.1 seconds. Compared to the COSMO calculation which takes longer computational times, these descriptors are more appropriate for a large-scale screening of the phase-separation absorbents than the COSMO descriptor.

Table 3
Accuracies of predicted results.

Descriptor	Accuracy		
	LGR	SVC	RF
COSMO	90.2 %	88.5 %	82.0 %
ECFP	65.6 %	72.1 %	57.4 %
RDkit	83.6 %	78.7 %	73.8 %
ECFP + RDkit	88.5 %	88.5 %	62.3 %

3.2. Regression coefficient in the model

Figure 1 shows the average values of the regression coefficients corresponding to the molecular descriptors in the LGR model using the RDkit descriptors. From Eq. (3), the explanatory variable which has a large absolute value of regression coefficient is important to determine phase behaviors. Furthermore, when the product of the variable x_i and the regression coefficient w_i is positive or negative, it contributes to immiscibility or miscibility, respectively. Figure 1 indicates that amine species that have small logP and large TPSA are likely to be phase-separated. On the other hand, amine species that have high logP and small TPSA are likely to form a homogeneous phase. Therefore, amine species whose logP and TPSA are significantly changed during CO_2 absorption seem to be appropriate for the phase-separation absorbent because the solvent containing these amine species can be easily phase-separated.

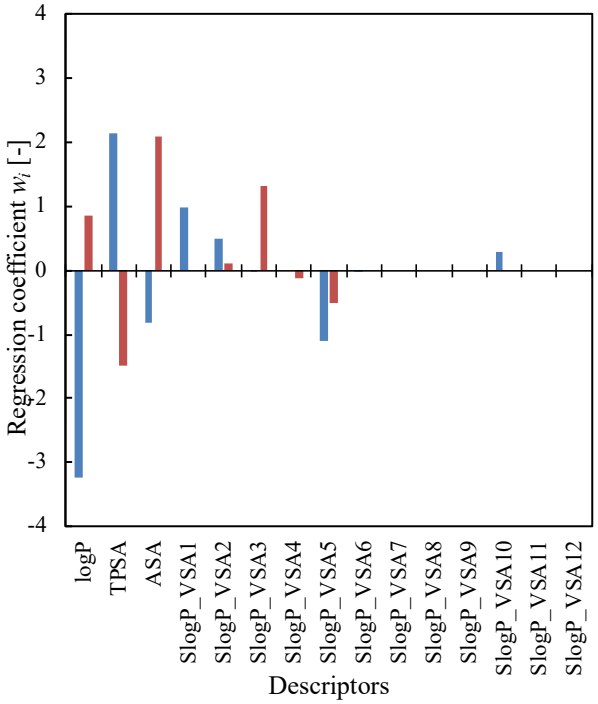

Fig. 1. Regression coefficients in the model using RDkit descriptors. Blue bar: coefficient related to amine species, red bar: coefficient related to glycol ether species.

4. Conclusion

We developed a machine learning model for the screening of the phase-separation CO_2 absorbents. The ECFP, RDkit descriptor, and COSMO descriptors are used as molecular descriptors in this work. The model using the COSMO and ECFP + RDkit descriptors achieved the prediction accuracy of the phase behavior 90.2 % and 88.5 %, respectively. Considering that the COSMO calculation takes longer computational times, the model using the RDkit and ECFP descriptors seems to be a promising approach because of comparable accuracy to the COSMO descriptor and short computational time. Moreover, polarity information of the solvent seems to be important to determine the phase behaviors of the mixed-solvent absorbents.

References

[1] G.T. Rochelle, Science. 325 (2009) 1652–1654.

[2] Y. Tan, W. Nookuea, H. Li, E. Thorin, J. Yan, Energy Convers. Manag. 118 (2016) 204–222.

[3] I. Sreedhar, T. Nahar, A. Venugopal, B. Srinivas, Sustain. Energy Rev. 76 (2017) 1080–1107.

[4] Q. Zhuang, B. Clements, J. Dai, L. Carrigan, Int. J. Greenh. Gas Control. 52 (2016) 449–460.

[5] T. Esaki, H. Machida, K. Norinaga, J. Adv. Manuf. Process. 2 (2020) 1‐7.

[6] H. Machida, R. Ando, T. Esaki, T. Yamaguchi, H. Horizoe, A. Kishimoto, K. Akiyama, M. Nishimura, Int. J. Greenh. Gas Control. 75 (2018) 1–7.
[7] F. Barzagli, F. Mani, M. Peruzzini, Int. J. Greenh. Gas Control. 60 (2017) 100–109.
[8] Y.E. Kim, J.H. Park, S.H. Yun, S.C. Nam, S.K. Jeong, Y. Il Yoon, J. Ind. Eng. Chem. 20 (2014) 1486–1492.
[9] M. Nakaoka, K.V.B. Tran, K. Yanase, H. Machida, K. Norinaga, Ind. Eng. Chem. Res. 59 (2020) 19020–19029.
[10] G. Landrum, RDKit: Open-Source Cheminformatics Software, (2016).
[11] TURBOMOLE 6.5; COSMO logic GmbH & Co.; K.G., Leverkusen, Germany, (2013)
[12] D. Rogers, M. Hahn, J. Chem. Inf. Model. 50 (2010) 742–754
[13] P. Ertl, B. Rohde, P. Selzer, J. Med. Chem. 43 (2000) 3714–3717.
[14] S.A. Wildman, G.M. Crippen, J. Chem. Inf. Comput. Sci. 39 (1999) 868–873.
[15] P. Labute, J. Mol. Graph. Model. 3263 (2000) 464–477.
[16] A. Klamt, G. Schüürmann, J. Chem. SOC., Perkin Trans. 2., (1993) 799-805.
[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, J. Mach. Learn. Res. 12 (2011) 2825–830.
[18] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[19] I. Guyon, A. Elisseeff, J. Mach. Learn. Res. 3 (2003) 1157–1182.
[20] C.Y.J. Peng, K.L. Lee, G.M. Ingersoll, J. Educ. Res. 96 (2002) 3–14.