Optimization of an Artificial Neural Network for Pure Component Parameters based on a Group Contribution Method of PC-SAFT EoS

Masayuki Kitahara[1], Hiroaki Matsukawa[2], Yuya Murakami[2], Atsushi Shono[2], and Katsuto Otake[2]*

[1] Department of Industrial Chemistry, Graduate School of Engineering, Tokyo University of Science, 12-1 Ichigayafunagawara-machi, Shinjuku-ku, Tokyo 162-0826, Japan
[2] Department of Industrial Chemistry, Faculty of Engineering, Tokyo University of Science, 12-1 Ichigayafunagawara-machi, Shinjuku-ku, Tokyo 162-0826, Japan

* To whom correspondence should be addressed.
E-mail : k-otake@ci.kagu.tus.ac.jp Tel : +81-3-5228-8052

Abstract

In previous work, an artificial neural network (ANN) to estimate the pure component parameters of PC-SAFT EoS based on the group contribution method was developed. The model could not distinguish between structural isomers. In this study, the relative value to the main chain and the substituent pattern of aromatics was added to the input of that ANN. As a result of training, the present ANN showed the same estimation accuracy as the previous study. Furthermore, it was able to distinguish between different branching positions and substitution patterns of aromatics. However, the estimated values did not reflect the trend due to structural differences.

Keywords

PC-SAFT equation of state, Artificial neural network, Group contribution method

1. Introduction

With the tremendous efforts of researchers, new chemical substances have been developing one after another. Unfortunately, only a few of these chemicals are used in industry due to the lack of physical properties such as phase equilibrium, viscosity, thermal conductivity, and diffusion coefficient, which are crucial for the chemical process designs. Measurements of

these properties requires specialized equipment and considerable amount of time. Therefore, methods to estimate and predict the properties of substances are highly desired. Most of the current physical property prediction models estimate only a single physical property, such as vapor pressure, critical temperature, and critical pressure. For the estimation/prediction of these properties, the use of equation of state (EoS) is a method with a great promise. The EoS is an expression for the pressure–volume–temperature (PVT) relationship. From this equation, various thermodynamic properties as well as physical properties such as Gibbs free energy, entropy, vapor pressure, phase equilibrium, viscosity, and interfacial tension can be derived. In fact, calculations using the EoS are performed in the process simulator to ensure the model has high compatibility with the process simulator.

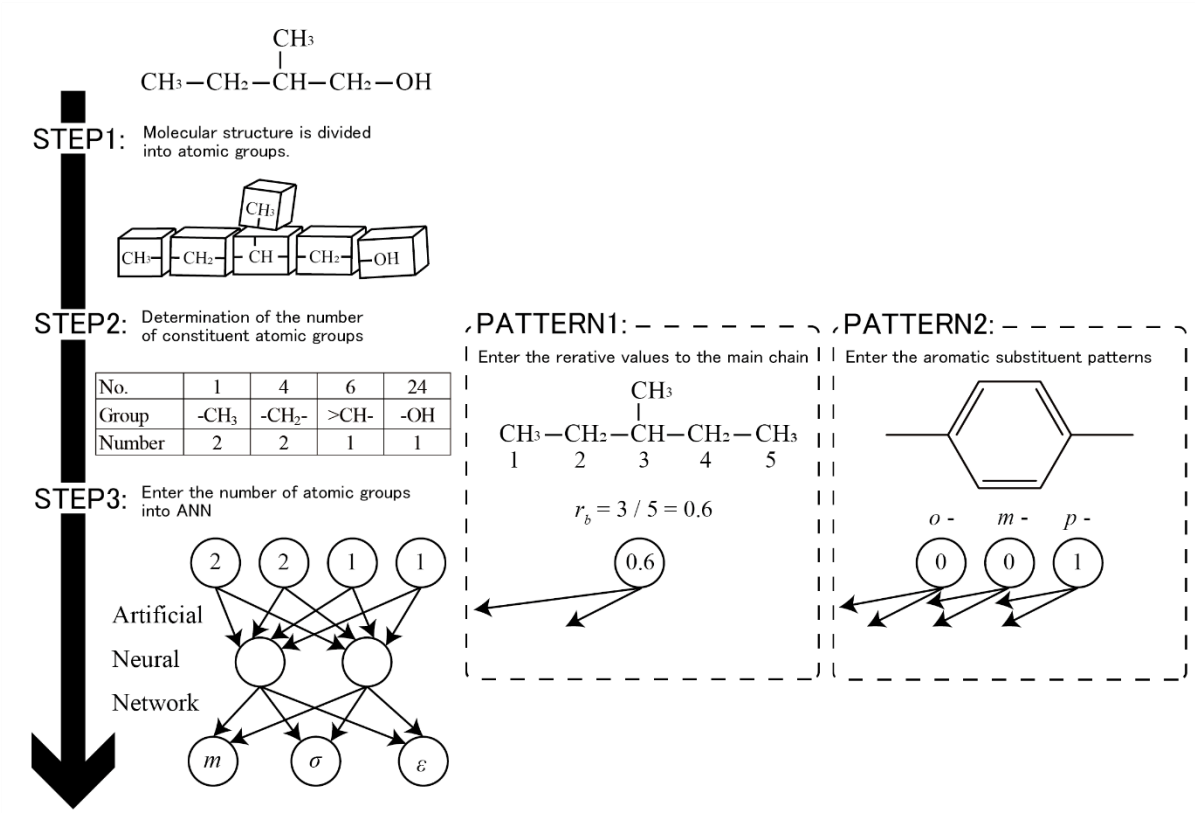
Pure component parameters are required for using the EoS, and these are determined from basic physical properties of individual pure substances. Unfortunately, as described above, many of these properties have not been reported. Therefore, as one of the methods to estimete/predict the pure component parameters, a group contribution method (GCM) is used. The GCM divides a substance molecule into specific atomic groups, determine the contribution of individual groups to the property, and estimates physical property of unknown substance by integrating the contributions of these groups.

In recent years, perturbation theory-type EoS, such as perturbed chain statical associating fluid theory (PC-SAFT) EoS [1-3], has been attracting attention owing to its applicability to a wide range of molecular families including polymers and ionic liquids. The pure component parameters of the PC-SAFT EoS are generally obtained from liquid density and saturated vapor pressure. Tamouza *et al*. [4] reported a GCM to estimate the pure component parameters of the original SAFT EoS [5] and SAFT variable range EoS [6], which was applied to vapor liquid equilibria of various hydrocarbon series. This method has further extended to binary systems [7], esters [8], aromatic hydrocarbons [9], and polymers [10]. Vijande *et al*. [11] reported a GCM to estimate the pure component parameters of the PC-SAFT EoS that is applicable to non-associated substances. Many other GCMs for perturbation theory-type EoS have been actively developed [12-16]. As discussed above, the GCM has been commonly used to determine the pure component parameters of the EoS, which are difficult to obtain the measured properties. Unfortunately, these GCMs are gradually expanding their scope of application though, the areas of the application are still limited. Further, as Nishiumi [17] suggested, there exists a limit of GCM. As the molecular weight and the number of functional groups increase, the estimation accuracy decreases. Therefore, attempts have been made to introduce correction terms [18, 19] though, modifying the formulas does not solve the problem satisfactorily.

In the previous work, the introduction of deep learning, especially artificial neural networks (ANNs), into GCM was proposed [20]. Deep learning is a form of machine learning

that allows computers to learn the tasks normally performed by humans. It is a technology that supports the rapid development of artificial intelligence (AI), and its progress leads to its practical application in various fields. The introduction of ANNs in place of contribution equations in the GCM was expected to solve the limitation of the GCMs and expand the range of target substances. One drawback of the ANNs is that no physicochemical theory is involved, making them a black box for estimation.

In the ANN previously proposed, input parameter is the number of atomic groups that constitute the molecule. It is not possible to distinguish between substances with the same number of constituent atomic groups but different molecular structures. For example, compounds with different branching positions, such as 2-methylpentane and 3-methylpentane. Another example is compounds with different aromatic substituent patterns, such as *o*-xylene, *m*-xylene, and *p*-xylene. These structural isomeric compounds have different physical properties due to the differences in their structures. In this study, in order to distinguish between the two patterns of structural isomers mentioned above, the relative values to the main chain (Figure 1-pattern 1) and/or aromatic substituent patterns (Figure 1-pattern 2) as well as the number of atomic groups were incorporated into the ANN as an additional input. the molecular structure of a target substance is divided into predetermined atomic groups, and the identification number (id) of each constituent atomic group is determined. Then, the pure component parameters of unknown molecule are obtained by feeding the id and number of of atomic groups to the ANN constructed from the id, molecular structure and pure component parameters of known molecules. The previous GCM covers 57 kinds of atomic groups, which is much more applicable than other GCMs for pure component parameters of PC-SAFT EoS. And its estimation accuracy was comparable to other GCMs. To verify the accuracy of parameter estimation by increasing the number of inputs, PC-SAFT EoS was used to estimate the liquid density, vapor pressure, and critical properties with the obtained parameters. As for the critical properties, it was confirmed that the previous GCM can estimate them with the same or better accuracy compared to other GCMs that estimate the critical properties [18, 19, 21, 22].



No.	1	4	6	24
Group	$-CH_3$	$-CH_2-$	>CH-	-OH
Number	2	2	1	1

Fig. 1. Process of estimating the pure component parameters of PC-SAFT EoS in this work.

2. Computational methods

2.1. PC-SAFT EoS

The PC-SAFT EoS is a perturbation theory-type EoS proposed by Gross and Sadowski in 2001 [1], which differs from other perturbation theory EoSs in the treatment of dispersion forces. In other perturbation theory-based EoSs [23-26], dispersive forces are introduced to rigid spheres, and then chain formation is considered. Meanwhile, in the PC-SAFT EoS, dispersion force is introduced to the rigid sphere chain after the rigid sphere chain is formed. The residual Helmholtz free energy is expressed as follows:

$$\frac{A^{res}}{NkT} = \tilde{a}^{res} = \tilde{a}^{hc} + \tilde{a}^{disp} + \tilde{a}^{assoc} \qquad (1)$$

where $\tilde{a}^{hc}$, $\tilde{a}^{disp}$, and $\tilde{a}^{assoc}$ are contributions to the Helmholtz free energy by a chain of hard spheres by diffusion and association. In this equation, the three pure component parameters of the PC-SAFT EoS are m, σ, and ε, which describe the number of segments per chain, segment diameter, and depth of the pair potential, respectively. These parameters were obtained by fitting the experimental liquid density and vapor pressure of pure components to the PC-SAFT EoS. For substances such as alcohol where molecules associate, contribution of association

$\tilde{a}^{assoc}$ [5, 27, 28] should be determined. In the calculation of $\tilde{a}^{assoc}$, pure component parameters for association ($\kappa^{A_iB_i}$, the effective association volume, and $\varepsilon^{A_iB_i}$, the association energy) are introduced. Association parameters vary depending on the way of association of the molecules. In this study, pure component parameters, *m*, *σ*, and *ε*, which are common to both unassociated and associated substances, were the subjects of the estimation. In the calculation of associated substances, pure component association parameters were obtained from the literature.

2.2. Artificial Neural Network

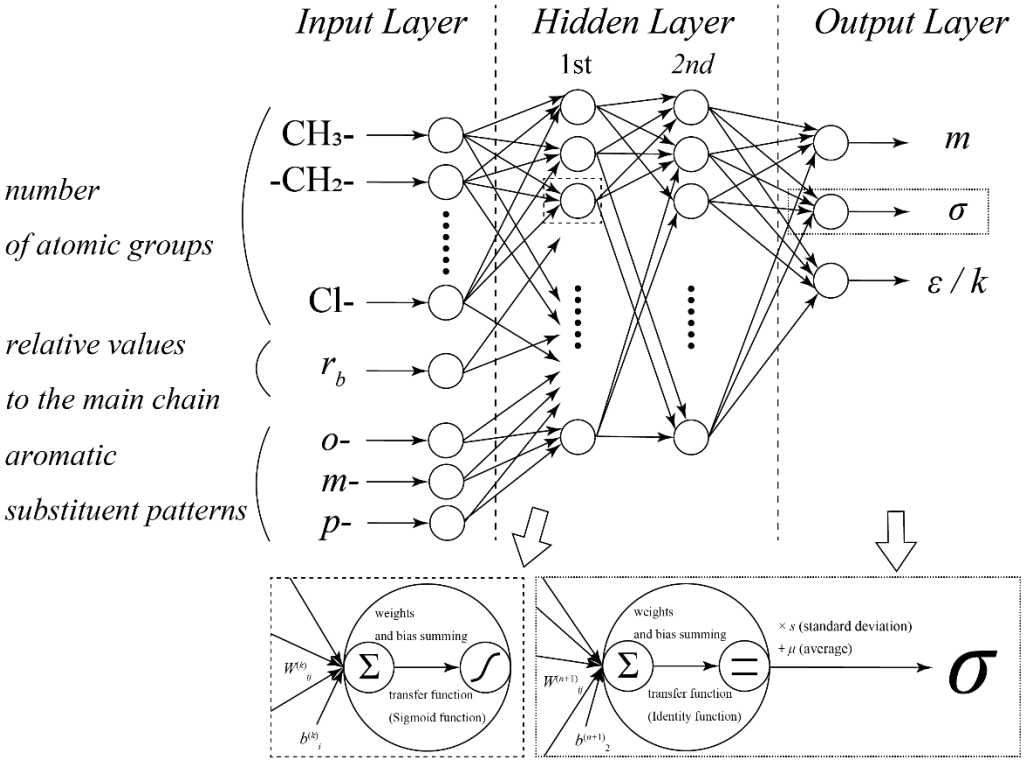


Fig. 2. Structure of the proposed neural network.

The theoretical description of the ANN can be found in the literatures [29-31]. Figure 2 shows the network structure of the ANN consisting of an input layer, hidden layers, and an output layer in this work.

Based on the GCM, the number of atomic groups (ex. CH_3-, >CH-) that constitute the molecule was entered in the input layer. The grouping was performed following the Nannoolal's group contribution method [18, 19]. The details of the atomic groups have been reported in the previous work [20]. The present ANN model covers 57 atomic groups. The relative value to the main chain and/or the aromatic substituent pattern are also the input parameter. The relative

value to the main chain r_b was the sum of the positions of the branched chains x_b divided by the number of carbon atoms in the main chain n_c, expressed as follows:

$$r_b = \sum x_b / n_c \qquad (2)$$

The main chain was selected as the longest series of the carbon atoms. For the aromatic substituent pattern, a neuron for each substitute position (*o*-, *m*-, and *p*-) was prepared. When a molecule has a substituent on the *o*- position but not on *m*- and *p*- position, "1" is entered for the *o*- positional neuron while “0” is entered for *m*- amd *p*- positional neurons.

In the hidden layer, the weights and biases from the previous layer are summed and transformed by the transfer function. Here, a sigmoid transfer function was used.

The outputs in the output layer are the pure component parameters m, σ, and ε of the PC-SAFT EoS. During the training, given that the dimensions of each parameter vary, the loss function will be biased if training is performed using the parameters’ original values without any transformation. Therefore, the training data were standardized for each parameter. After summing the weights and biases from the last layer of the hidden layer, the parameters are obtained through the identity function as the transfer function, followed by multiplying by the standard deviation obtained from the training data, and adding up the mean values.

Table 1 summarizes the pure component parameters database. The pure component parameters database consisted of values from the literature[1, 2, 32], including those of Gross and Sadowski [1,2]. The values of pure component parameters were those determined from saturated vapor pressure and liquid density. The total number of substances in the database was 282, of which 197 data sets were randomly selected as training data and the rest 85 data sets were used for test data. However, the training data were adjusted to ensure that each of the 57 atomic groups appeared in the input layer at least once during training. During training, the sum of squared errors, E, whose formula is shown below, was used as the loss function, and the weights and bias were optimized to minimize its value.

$$E = \frac{1}{2}\sum_k \left(y_k - t_k\right)^2 \qquad (3)$$

where y is the calculation result and t is the standardized training data. Optimization was performed by updating the parameters using the Adam method [33]. In particular, the number of parameter updates was set to 10,000, and the learning rate was set to 0.0005. The hyperparameters required to update the parameters in the Adam method were the recommended values reported in the literature [33]. In addition, the dropout method [34] was incorporated into the training to suppress overlearning.

In this study, based on the results of previous study [20], number of hidden layers which consisted from 40 neurons is set to 2. The optimization results were evaluated by the mean squared deviation (*RMSD*) and the average absolute relative deviation (*AARD*) described as

follows.

$$RMSD = \sqrt{\frac{1}{N_{data}} \sum_{k} \left(Y_k - T_k \right)^2} \qquad (4)$$

$$AARD = \frac{100}{N_{data}} \sum_{k} \left| \frac{Y_k - T_k}{T_k} \right| \qquad (5)$$

where N_{data} is the number of the total data, and Y and T are the estimated parameter and the parameter from the literature, respectively. The more details of the network structure and the training method have been described elsewhere [20].

Table 1. Database of the pure component parameters of PC-SAFT EoS.

Number of data	Example	Ref.
71	methane, isobutane, ethylene	Gross and Sadowski [1]
16	methanol, methylamine	Gross and Sadowski [2]
195	2,2,3-Triethylpentane 1,1-Dimthylcyclohexane Propadiene, Dibutyl ether	Tihic *et al.* [32]
Total number of data = 282 (training data; 197, test data; 85)		

3. Results and Discussion

Figure 3 and Table 2 show the estimation results of the parameters of the PC-SAFT EoS. The table also summarizes the estimation results without additional neurons to distinguish the structural isomers taken from the previous work. From the table, it could be seen that the accuracy of both training and test data remained almost unchanged when the relative value to the main chain and the aromatic substituent pattern were added as inputs.

Table 2 Estimation results of the pure component parameters of PC-SAFT EoS.

test data	Previous work (without structural parameter)			This work (with structural parameter)		
	m / -	σ / Å	εk^{-1} / K	m / -	σ / Å	εk^{-1} / K
AARD %	11.9	3.04	5.93	11.9	3.03	5.54
RMSD	0.563	0.191	26.9	0.569	0.185	25.8

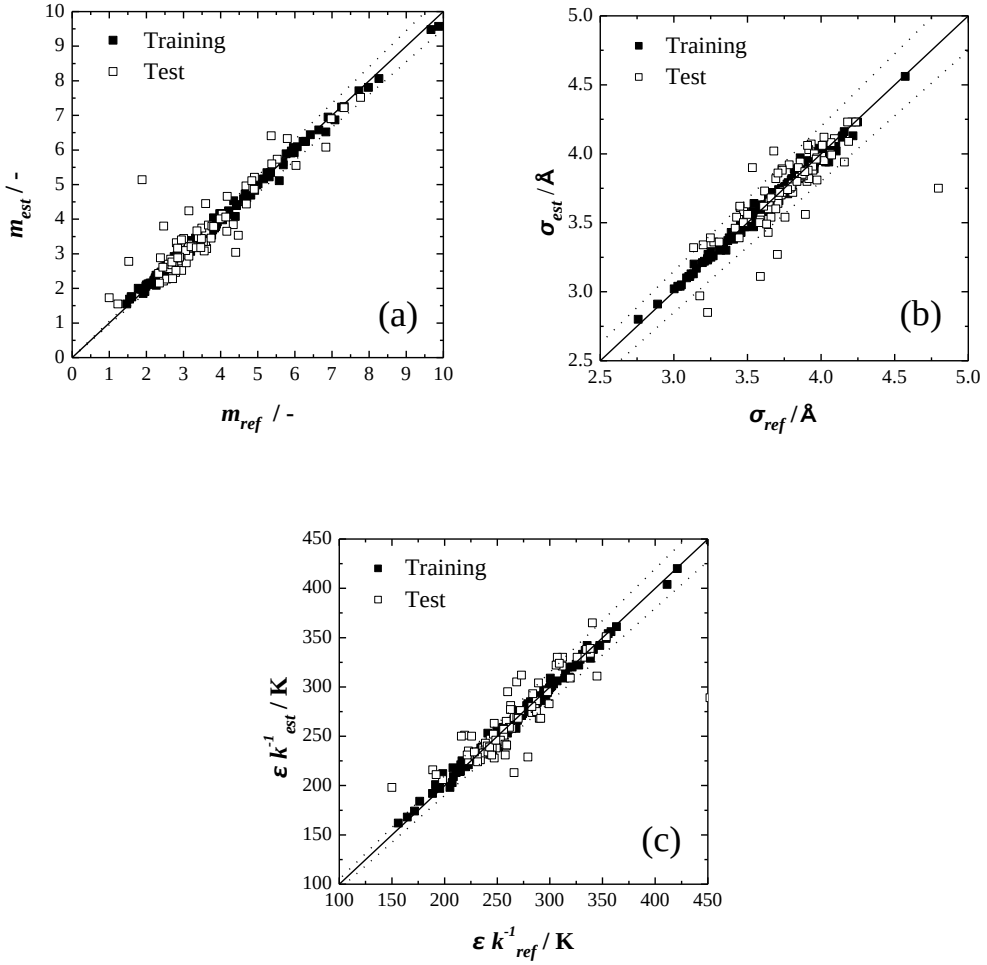


Fig. 3. Modeling ability of the optimized ANN to predict pure parameters of PC-SAFT EoS ((a) segment number, *m*; (b) segment diameter, *σ*; and (c) depth of pair potential, *ε* k^{-1})

Table 3 shows the estimation results of PC-SAFT EoS parameters for 2-methylpentane and 3-methylpentane. The table also shows the estimation results by ANN of the previous study [20]. From the table, it could be seen that in the previous study, the same values were estimated for 2-methylpentane and 3-methylpentane, while in this work, different values were estimated. It was possible to distinguish the difference in the branching position relative to the main chain. Unfortunately, when compared to the literature values, the large and small relationships due to structural differences were not reflected in the estimated values.

Table 3 Estimation results of the pure component parameters of PC-SAFT EoS for 2-methylpentane and 3-methylpentane.

	literature value			previous work			this work		
	m / -	σ / Å	εk^{-1} / K	m / -	σ / Å	εk^{-1} / K	m / -	σ / Å	εk^{-1} / K
2-methylpentane	2.93	3.85	236	2.96	3.84	235	2.94	3.83	235
3-methylpentane	2.89	3.86	240	2.96	3.84	235	2.99	3.83	236

Table 4 shows the estimation results of the parameters of PC-SAFT EoS for *o*-xylene, *m*-xylene, and *p*-xylene. The table also shows the estimation results by ANN of the previous study [20]. Similar to the case described above, different values were estimated though, the large and small relationships due to the different placement positions were not reflected in the estimated values.

Table 4 Estimation results of the pure component parameters of PC-SAFT EoS for *o*-xylene, *m*-xylene, and *p*-xylene.

	literature value			previous work			this work		
	m / -	σ / Å	εk^{-1} / K	m / -	σ / Å	εk^{-1} / K	m / -	σ / Å	εk^{-1} / K
o-xylene	3.14	3.76	291	3.14	3.76	286	3.16	3.76	291
m-xylene	3.19	3.76	284	3.14	3.76	286	3.15	3.76	286
p-xylene	3.17	3.78	284	3.14	3.76	286	3.18	3.77	284

It could be concluded that both the branching positions and the aromatic substituent patterns could be distinguished by the approach proposed in this study though, the estimated values were not consistent with the large and small relationships due to structural differences. One of the reasons for this could be the lack of training data. Unfortunately, as mentioned in Introduction, replenishing the training data is difficult and limited. Therefore, it is necessary to change the structure of the ANN to achieve better learning with less training data.

4. Conclusion

In this study, a model for estimating the pure component parameters of the PC-SAFT EoS m, σ, and ε by an ANN based on GCM was developed. In order to distinguish between structural isomers, the relative value to the main chain and the substitution pattern of aromatics were input in addition to the number of atomic groups. As a result of the training, the accuracy of the PC-SAFT parameter estimation was equivalent to that of the number of atomic groups only. In addition, it was possible to distinguish the branching positions and the substitution patterns of aromatics. However, the obtained estimates did not reflect the trend due to structural

differences. It is necessary to further improve the estimation accuracy by optimizing the structure of the ANN. Furthermore, the parameters introduced in this study are not enough to distinguish all the structural isomers. Therefore, we have to consider introducing more parameters.

References

[1] J. Gross, G. Sadowski, Perturbed-Chain SAFT: An Equation of State Based on a Perturbation Theory for Chain Molecules, Ind. Eng. Chem. Res., 40 (2001) 1244-1260.

[2] J. Gross, G. Sadowski, Application of the Pertuebed-Chain SAFT Equation of State to Associating Systems, Ind. Eng. Chem. Res., 41 (2002) 5510-5515.

[3] J. Gross, G. Sadowski, Modeling Polymer Systems Using the Perturbed-Chain Statistical Associating Fluid Theory Equation of State, Ind. Eng. Chem. Res., 41 (2002) 1084-1093.

[4] S. Tamouza, J.P. Passarello, P. Tobaly, J.C. de Hemptinne, Group contribution method with SAFT EOS applied to vapor liquid equilibria of various hydrocarbon series, Fluid Phase Equilibria, 222-223 (2004) 67-76.

[5] W.G. Chapman, K.E. Gubbins, G. Jackson, M. Radosz, New Reference Equation of State for Associating Liquids, Ind. Eng. Chem. Res., 29 (1990) 1709-1721.

[6] A. Gil-Villegas, A. Galindo, P.J. Whitehead, S.J. Mills, G. Jackson, A.N. Burgess, Statistical associating fluid theory for chain molecules with attractive potentials of variable range, The Journal of Chemical Physics, 106 (1997) 4168-4186.

[7] S. Tamouza, J.-P. Passarello, P. Tobaly, J.-C.d. Hemptinne, Application to binary mixtures of a group contribution SAFT EOS (GC-SAFT), Fluid Phase Equilibria, 228-229 (2005) 409-419.

[8] T.X.N. Thi, S. Tamouza, P. Tobaly, J.-P. Passarello, J.-C.d. Hemptinne, Application of group contribution SAFT equation of state (GC-SAFT) to model phase behaviour of light and heavy esters, Fluid Phase Equilibria, 238 (2005) 254-261.

[9] D.N. Huynh, M. Benamira, J.-P. Passarello, P. Tobaly, J.-C.d. Hemptinne, Application of GC-SAFT EOS to polycyclic aromatic hydrocarbons, Fluid Phase Equilibria, 254 (2007) 60-66.

[10] A. Tihic, G.M. Kontogeorgis, N.v. Solms, M.L. Michelsen, A Predictive Group-Contribution Simplified PC-SAFT Equation of State: Application to Polymer Systems, Ind. Eng. Chem. Res., 47 (2008) 5092-5101.

[11] J. Vijande, M.M. Pineiro, J.L. Legido, Group-Contribution Method for the Molecular Parameters of the PC-SAFT Equation of State Taking into Account the Proximity Effect. Apprication to Nonassociated Compounds, Ind. Eng. Chem. Res., 49 (2010) 9394-9406.

[12] Y. Peng, K.D. Goff, M.C.d. Ramos, C. McCabe, Developing a predictive group-contribution-based SAFT-VR equation of state, Fluid Phase Equilibria, 277 (2009) 131-

144.
[13] A. Lymperiadis, C.S. Adjiman, A. Galindo, G. Jackson, A group contribution method for associating chain molecules based on the statistical associating fluid theory (SAFT-gamma), J Chem Phys, 127 (2007) 234903.
[14] W.A. Burgess, D. Tapriyal, B.D. Morreale, Y. Wu, M.A. McHugh, H. Baled, R.M. Enick, Prediction of fluid density at extreme conditions using the perturbed-chain SAFT equation correlated to high temperature, high pressure density data, Fluid Phase Equilibria, 319 (2012) 55-66.
[15] W.A. Burgess, D. Tapriyal, B.D. Morreale, Y. Soong, H.O. Baled, R.M. Enick, Y. Wu, B.A. Bamgbade, M.A. McHugh, Volume-translated cubic EoS and PC-SAFT density models and a free volume-based viscosity model for hydrocarbons at extreme temperature and pressure conditions, Fluid Phase Equilibria, 359 (2013) 38-44.
[16] W.A. Burgess, D. Tapriyal, I.K. Gamwo, Y. Wu, M.A. McHugh, R.M. Enick, New Group-Contribution Parameters for the Calculation of PC-SAFT Parameters for Use at Pressures to 276 MPa and Temperatures to 533 K, Industrial & Engineering Chemistry Research, 53 (2014) 2520-2528.
[17] H. Nishiumi, Thermodynamic property prediction for high molecular weight molecules based on their constituent family, Fluid Phase Equilibria, 420 (2016) 1-6.
[18] Y. Nannoolal, J. Rarey, D. Ramjugernath, W. Cordes, Estimation of pure component properties, Fluid Phase Equilibria, 226 (2004) 45-63.
[19] Y. Nannoolal, J. Rarey, D. Ramjugernath, Estimation of pure component properties, Fluid Phase Equilibria, 252 (2007) 1-27.
[20] H. Matsukawa, M. Kitahara, K. Otake, Estimation of pure component parameters of PC-SAFT EoS by an artificial neural network based on a group contribution method, Fluid Phase Equilibria, 548 (2021) 113179.
[21] K.G. Joback, R.C. Reid, ESTIMATION OF PURE-COMPONENT PROPERTIES FROM GROUP-CONTRIBUTIONS, Chemical Engineering Communications, 57 (1987) 233-243.
[22] L. Constantinou, R. Gani, New Group Contribution Method for Estimating Properties of Pure Compounds, AIChE Journal, 40 (1994) 1697-1710.
[23] M.S. Wertheim, Fluids with Highly Directional Attractive Forces. I. Statistical Thermodynamics, Journal of Statistical Physics, 35 (1984) 19-34.
[24] M.S. Wertheim, Fluid with Highly Directional Attractive Forces. II. Thermodynamics Perturbation Theory and Integral Equations, Journal of Statistical Physics, 35 (1984) 35-47.
[25] M.S. Wertheim, Fluids with Highly Directional Attractive Forces. III. Multiple Attraction Sites, Journal of Statistical Physics, 42 (1986) 459-476.
[26] M.S. Wertheim, Fluids with Highly Directional Attractive Foreces. IV. Equilibrium

Polymerization, Journal of Statistical Physics, 42 (1986) 477-492.
[27] S.H. Huang, M. Radosz, Equation of State for Small, Large, Polydisperse, and Associating Molecules, Ind. Eng. Chem. Res., 29 (1990) 2284-2294.
[28] S.H. Huang, M. Radosz, Equation of State for Small, Large, Polydisperse, and Associating Molecules: Extension to Fluid Mixtures, Ind. Eng. Chem. Res., 30 (1991) 1994-2005.
[29] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, The MIT Press, 2016.
[30] C.M. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, USA, 1996.
[31] C.M. Bishop, Pattern Recognition and Machine Learning 1st ed., Springer, 2006.
[32] A. Tihic, G.M. Kontogeorgis, N. von Solms, M.L. Michelsen, Applications of the simplified perturbed-chain SAFT equation of state using an extended parameter table, Fluid Phase Equilibria, 248 (2006) 29-43.
[33] D.P. Kingma, J.L. Ba, Adam: A Method for Stochastic Optimization, in: the 3rd International Conference for Learning Representations, San Diego, 2014, pp. 1-15.
[34] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Drop out: A Simple Way to Prevent Neural Networks from Overfitting, Journal of Machine Learning Research, 15 (2014) 1929-1958.