Molecular Reconstruction of Naphtha based on Physical Information Neural Network

Fangyuan Ma*, **, Xin Zheng*, Chengyu Han, Jingde Wang*, Wei Sun*

** College of Chemical Engineering, Beijing University of Chemical Technology, 100029 Beijing, China.*
*** Center of process monitoring and data analysis, Wuxi Research Institute of Applied Technologies, Tsinghua University, 214072 Wuxi, China.*
Corresponding Author: Jingde Wang, jingdewang@mail.buct.edu.cn, Wei Sun, sunwei@mail.buct.edu.cn

Abstract: A molecular reconstruction method based on physical information neural network is proposed for predicting the molecular composition of naphtha. By embedding physical information utilized in typical molecular reconstruction methods, such as mixing rules, into the loss function of the neural network, the model tends to converge to the state conforming to physical rules in training stage. The neural network model obtained by the method contains certain physical information, which can improve the generalization ability of the model. The results show that the prediction performance and application range of the proposed method are better than those of the typical ANN-based molecular reconstruction method.

Keywords: Artificial Neural Network, Mixing Rules, Generalization Ability.

1. INTRODUCTION

In industrial practice, the composition of feedstock will significantly affect the product yields and quality. A detailed molecular composition information will aid in determining whether the feed conforms to the design and environmental requirements of the unit in question, as well as identifying potential bottlenecks. Naphtha is a complex mixture of hydrocarbons, including *n*-paraffins(P), *iso*-paraffins(I), naphthenes(N), aromatics(A), etc, which is an important feedstock in the chemical industry. Determining a detailed molecular composition of naphtha is essential for refineries to improve product quality and increase profitability (Ren, et al., 2019a).

In order to obtain the detailed molecular composition of naphtha, several instrumental analysis techniques can be applied, such as gas-chromatography (GC), GC×GC, and GC-mass spectrometry (GC-MS). However, it is difficult to apply these methods widely in industry, because they are in general very time-consuming and expensive (Bi, et al., 2019a). In an attempt to avoid the usage of instrumental analysis techniques, a method known as “molecular reconstruction” has been proposed and widely studied by researchers in recent years. The molecular composition of naphtha can be determined based on a number of average properties or so-called commercial indices. These commercial indices are usually relatively easy-to-obtain analytical data, e.g., the average molecular weight of the mixture, the specific density, the global PINA weight fractions, hydrogen-carbon molar ratio, some points of a boiling point distillation curve and so on (Riazi., 2005).

Typical molecular reconstruction methods mainly include stochastic reconstruction (SR) method, structure-oriented lumping (SOL) method, molecular type-homologous series (MTHS) matrix method, entropy maximization method and etc (Stratiev, et al., 2019). These methods can be summarized as five steps for molecular reconstruction: the construction of molecular library, the acquisition of the properties of pure components, the calculation of average properties of the mixture, the establishment of the objective function, and the adjustment of the mole fractions (or mass fraction) of molecules (Ren, et al., 2019b). The key to these methods is to optimize a specific objective function to determine a detailed molecule composition (Wang, et al., 2017). The objective function can be generated from theoretical concepts like Shannon entropy, or it can be a cost function, such as calculation error of commercial indices. However, it is difficult to accurately determine the corresponding molecular composition of naphtha only from the commercial indices, because there is no unique correlation between the average properties and detailed composition. The molecular composition obtained by these molecular reconstruction methods is the most likely of all the possible molecular compositions that theoretically conform to the specified commercial indices. Previous research has also found that typical molecular reconstruction methods are more accurate in predicting commercial indices than molecular composition, because the commercial indices is a direct optimization object (Bi, et al., 2019b).

Artificial neural network (ANN) is a deep learning method with powerful learning ability, which can be used to fit complex nonlinear relationships among variables. Steven P. et al. applied artificial neural network in the field of molecular reconstruction and compared it with typical molecular reconstruction methods (Steven, et al., 2010). The results shown that the composition of naphtha can be reconstructed with great accuracy, provided the considered naphtha has similar characteristics compared to the large number of training data used to develop the ANN. However,

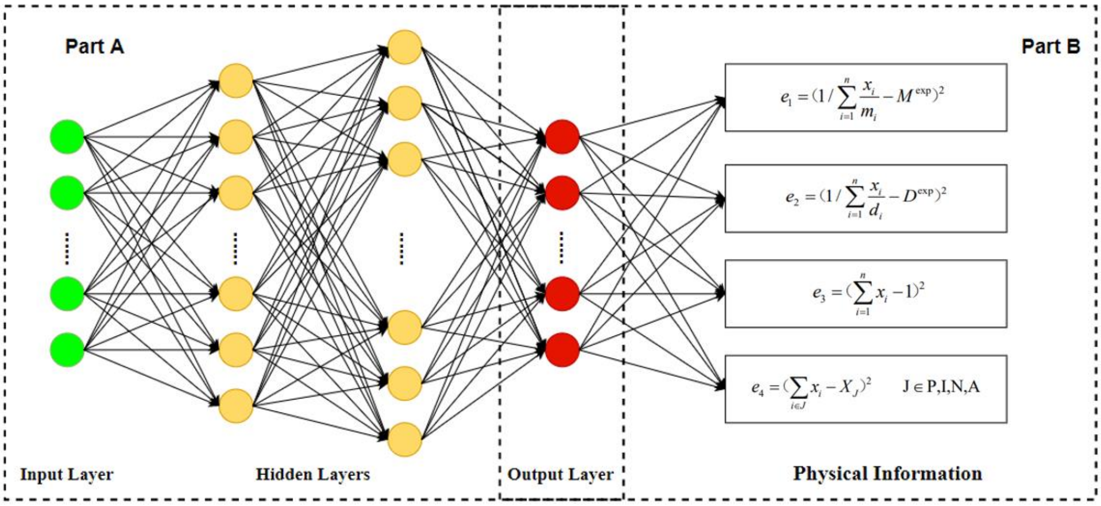
Part A
Part B
$e_1 = (1/\sum_{i=1}^{n} \frac{x_i}{m_i} - M^{\exp})^2$
$e_2 = (1/\sum_{i=1}^{n} \frac{x_i}{d_i} - D^{\exp})^2$
$e_3 = (\sum_{i=1}^{n} x_i - 1)^2$
$e_4 = (\sum_{i \in J} x_i - X_J)^2$ J ∈ P,I,N,A
Input Layer
Hidden Layers
Output Layer
Physical Information

Fig. 1. The structural layout of the molecular reconstruction method based on PINN.

outside this range, the performance of ANN declines dramatically, while the performance of typical molecular reconstruction methods is unaffected by the characteristics of naphtha considered. Since the ANN is hard to associate with containing physical meaning, its application range will be determined by employed training set, which is evidently finite.

Physical Information Neural Network (PINN) is a type of artificial neural network proposed by a Brown University research team, which employs physical equations as constraints and is frequently applied to solve partial differential equations (PDEs) (Raissi. et al., 2018). By embedding the PDEs containing physical information into the loss function of the neural network, the model tends to converge to the state conforming to the underlying physical rules in training stage (Karniadakis. et al., 2021). It can be seen as a method of using the physical information as a regularization agent to enhance the generalization ability of the model. Inspired by PINN, if the physical information utilized in typical molecular reconstruction methods can be embedded into the loss function of the neural network, the generalization ability of the model may be further improved.

In this work, a molecular reconstruction method based on PINN is proposed. In order to make the neural network contain some physical information, the mixing rules are embedded into the loss function. The neural network parameters are optimized by minimizing molecular composition prediction errors and the commercial indices errors to further improve the generalization ability of the network. Data generated based on the characteristics of naphtha samples in the literature are investigated to validate the proposed method (Mei. et al., 2017). The verification results show that the application range of molecular reconstruction model established by the proposed method is expanded and the predicted values are in good agreement with the experimental values.

The remaining sections are arranged as follows. The procedure of PINN-based molecular reconstruction method is detailed in Section 2. In Section 3, the proposed method is applied to the molecular reconstruction of naphtha. The molecular reconstruction results are shown and discussed. Conclusion is drawn at the end.

2. METHODOLOGY

Artificial neural network (ANN) is a deep learning method, which is widely applied to capture complex and nonlinear relationships between multiple input variables and output variables. As shown in Part A of Fig. 1, the structural layout of a typical ANN consists of an input layer, hidden layers and an output layer. Among them, hidden layers are the core structure of ANN. By utilizing a nonlinear activation function, the nonlinear mapping of the model can be realized. The mathematical expression can be expressed as follows:

$$H_i = f(H_{i-1} \otimes W_i + b_i) \tag{1}$$

where H_i indicates the output of *i*-th layer, W_i and b_i are the transformation parameters, also known as weights and biases, and $f()$ is the activation function. Threshold function, Sigmoid function and Tanh function are the most commonly applied activation functions for neural networks.

Input data is fed into the network in a forward direction. Each hidden layer accepts data, processes it according to the activation function and passes to the next layer. The predicted value of the ANN is produced from the output layer. A loss function is applied to evaluate the performance of the ANN. Then, the back-propagation algorithm is utilized to determine parameters of the neural network, such as weights and biases of all layers (Rumelhart. et al., 1986). In regression algorithms, the mean square error function (*MSE*) as a regression evaluation index is frequently the first choice of loss function.

$$Loss = MSE = \sum_{i=1}^{n} \left(x_i^{\exp} - x_i^{cal} \right)^2 \tag{2}$$

Table 1. Components of simulated naphtha samples

Carbon number	*n*-Paraffins	*iso*-Paraffins	Naphthenes	Aromatics
C4	*n*-butane	C_4 *iso*-paraffins	-	-
C5	*n*-pentane	C_5 *iso*-paraffins	Cyclopentane	-
C6	*n*-hexane	C_6 *iso*-paraffins	C_6 naphthenes	Benzene
C7	*n*-heptane	C_7 *iso*-paraffins	C_7 naphthenes	Ethylbenzene
C8	*n*-octane	C_8 *iso*-paraffins	C_8 naphthenes	C_8 aromatics
C9	*n*-nonane	C_9 *iso*-paraffins	C_9 naphthenes	C_9 aromatics
C10	*n*-decane	C_{10} *iso*-paraffins	C_{10} naphthenes	C_{10} aromatics
C11	*n*-undecane	C_{11} *iso*-paraffins	C_{11} naphthenes	C_{11} aromatics
C12	*n*-dodecane	C_{12} *iso*-paraffins	-	-

where $x_i^{\exp}$ indicates the experimental value of mass fraction of pure component *i* in naphtha, x_i^{cal} indicates the calculated mass fraction of pure component *i* in naphtha and *n* is the number of pure components.

It is well known that the application range of typical ANN is limited by the size of the employed training set, because it does not contain physical information. In this work, in order to improve the generalization ability of the model, mixing rules are embedded into the loss function to train the model. The mixing rules obtained from the literature are shown as follows (Riazi., 2005):

$$M = \frac{1}{\sum_{i=1}^{n} \frac{x_i}{m_i}} \tag{3}$$

$$D = \frac{1}{\sum_{i=1}^{n} \frac{x_i}{d_i}} \tag{4}$$

$$X_J = \sum_{i \in J} x_i \qquad \mathrm{J} \in \mathrm{P, I, N, A} \tag{5}$$

where x_i is mass fraction of pure component *i* in naphtha; m_i and d_i are molecular weight and density of pure component *i* in naphtha, respectively; *M* and *D* are calculated molecular weight and calculated density of naphtha, respectively; *J* is homologous series and X_J is mass fraction of the homologous series *J*.

According to the mixing rules, commercial indices of naphtha, e.g., the molecular weight *M*, the density *D* and mass fraction of the homologous series *J*, can be calculated. Meanwhile, the sum of the mass fractions of the pure components should equal 1. Theoretically, the calculated value should be equal to the experimental value. Therefore, Eq. (6) - (9) are added to Eq. (2) to obtain a new loss function, as shown in Eq. (10).

$$e_1 = \left(\frac{1}{\sum_{i=1}^{n} \frac{x_i^{cal}}{m_i}} - M^{\exp}\right)^2 \tag{6}$$

$$e_2 = \left(\frac{1}{\sum_{i=1}^{n} \frac{x_i^{cal}}{d_i}} - D^{\exp}\right)^2 \tag{7}$$

$$e_3 = \left(\sum_{i \in J} x_i^{cal} - X_J^{\exp}\right)^2 \qquad \mathrm{J} \in \mathrm{P, I, N, A} \tag{8}$$

$$e_4 = \left(\sum_{i=1}^{n} x_i^{cal} - 1\right)^2 \tag{9}$$

$$Loss_{new} = MSE + e_1 + e_2 + e_3 + e_4 \tag{10}$$

At this point, the molecular reconstruction method based on PINN has been constructed, and its structural layout is shown in Fig. 1.

3. CASE STUDIES

3.1 Data Details

Table 2. Commercial indices of naphtha

Group-type analysis (wt %)	
Total amount of *n*-paraffins	Total amount of *iso*-paraffins
Total amount of naphthenes	Total amount of aromatics
Simulated distillation (°C)	
Initial boiling point (IBP)	5 vol% boiling point ($T_{5\%}$)
10 vol% boiling point ($T_{10\%}$)	30 vol% boiling point ($T_{30\%}$)
50 vol% boiling point ($T_{50\%}$)	70 vol% boiling point ($T_{70\%}$)
90 vol% boiling point ($T_{90\%}$)	95 vol% boiling point ($T_{95\%}$)
Final boiling point (IBP)	
Molecular weight	
Density	

The PINN-based molecular reconstruction model needs to be trained with a large set of experimental data. However, gathering enough experimental data to train the neural networks is extremely challenging in practice. In this paper, based on the components distributions of naphtha samples given in the literature (Mei. et al., 2017), 1000 sets of simulated samples of naphtha are generated. As illustrated in Table 1, a total number of 31 components are considered. Commercial indicators of these simulated naphtha samples can be obtained utilizing simulation software. The detailed information of the commercial indicators is shown in Table 2.

In this work, 900 samples of simulated naphtha are utilized to established the PINN-based molecular reconstruction. Among them, 800 samples are training data and 100 samples are validation data. In addition, 100 samples are used as test data to validate the performance of the model, with 54 samples falling outside of the training data range.

3.2 Modelling

As mentioned in the methodology part, a PINN-based molecular reconstruction model is established, including an input layer, a hidden layer and an output layer. According to the number of commercial indicators, the number of neurons in the input layer is fixed to 15. Based on the theory of Kolmogorov, the number of neurons in the hidden layer is fixed to 31 (Kolmogorov., 1957). Because there are 31 naphtha components to be predicted, the number of neurons in the output layer is determined to be 31. In addition, some hyperparameters of the neural network are also very important, such as iteration epoch and activation functions, etc., which can be determined through grid search algorithm. In this work, ReLU function is determined to be the activation function, and the number of iteration epoch is 400.

3.2 Results and Discussion

Apart from the proposed method, an ANN-based molecular reconstruction model is established for comparison. To indicate these differences quantitatively, Table 3 shows the mean difference (MD), the mean absolute difference (MAD) and the root-mean-square difference (RMSD) of the calculated commercial indices of test data obtained ANN and proposed method. From the table, the MD, MAD, and RMSD of molecular weight and density of test data obtained by proposed method are -0.8601, 1.6891, 2.1053 and -0.0015, 0.0095, 0.0117, respectively, which are more accurate than those obtained by ANN-based molecular reconstruction model. Meanwhile, it can be seen that the prediction accuracy of some points of a boiling point distillation curve has not been significantly improved. Because the data of boiling range are not embedded into the loss function of the neural network, which can be researched in future.

$$MD = \frac{1}{n}\sum_{i=1}^{n}(x_i^{cal} - x_i^{\exp}) \tag{11}$$

$$MAD = \frac{1}{n}\sum_{i=1}^{n}\left|x_i^{cal} - x_i^{\exp}\right| \tag{12}$$

$$RMSD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i^{cal} - x_i^{\exp})^2} \tag{13}$$

The absolute errors of the calculated mass fractions of 20 samples, which are outside of training data range, obtained by proposed method is illustrated in Fig. 2, and that obtained by ANN is shown in Fig. 3. It can be found that the absolute error of most component mass fractions calculated by proposed method is smaller than that calculated by ANN. This indicates that a molecular composition closer to the true value can be obtained by PINN-based molecular reconstruction method. By embedding the mixing rules into the loss function of the neural network, the model tends to converge to the state conforming to the mixing rules in the training stage. The generalization ability of the model is improved since the neural network obtained using this method incorporates some physical information.

The one sample (naphtha A) out of the above 20 samples is applied to provide a detailed introduction to the molecular reconstruction performance of the proposed method. In order to visually distinguish the difference between the training data and the naphtha A, Principal Components Analysis (PCA) is applied to data visualization (Steven, et al., 2010). The visualization result is illustrated in Fig. 4. It can be seen

Table 3. Comparison of the Commercial Indices of test data obtained ANN and proposed method.

	ANN			Proposed Method		
Bulk properties	MD	MAD	RMSD	MD	MAD	RMSD
Molecular weight	-1.5131	2.4669	3.2012	-0.8601	1.6891	2.1053
Density (g/cm^3)	-0.0074	0.0151	0.0198	-0.0015	0.0095	0.0117
n-paraffins (wt %)	0.5935	2.1672	2.6021	-0.0300	1.9367	2.3801
iso-paraffins (wt %)	0.8170	2.2699	2.9049	0.8231	2.3992	2.9476
Naphthenes (wt %)	-0.4201	1.0499	1.3102	-0.7033	1.3543	1.7525
Aromatics (wt%)	-0.3189	0.4341	0.6360	-0.3516	0.4879	0.6315
IBP (°C)	3.7778	9.6979	9.6980	4.7727	6.6368	8.8253
$T_{5\%}$ (°C)	-1.2321	2.5461	2.5461	-1.7279	2.4008	2.9653
$T_{10\%}$ (°C)	-2.3768	3.7305	3.7305	-3.0684	3.4981	4.2172
$T_{30\%}$ (°C)	-3.1971	5.0134	5.0134	-4.6039	4.8095	6.6891
$T_{50\%}$ (°C)	-5.0142	8.0077	8.0077	-5.3822	5.6813	8.4843
$T_{70\%}$ (°C)	-4.6900	6.9265	6.9265	-4.7468	4.8196	6.7692
$T_{90\%}$ (°C)	-0.6089	2.5430	2.5430	-1.4055	1.6479	2.2033
$T_{95\%}$ (°C)	-0.9568	2.0265	2.0265	-1.2655	1.7176	2.5525
FBP (°C)	-0.8201	2.4769	2.4769	-2.3351	2.3516	3.3229

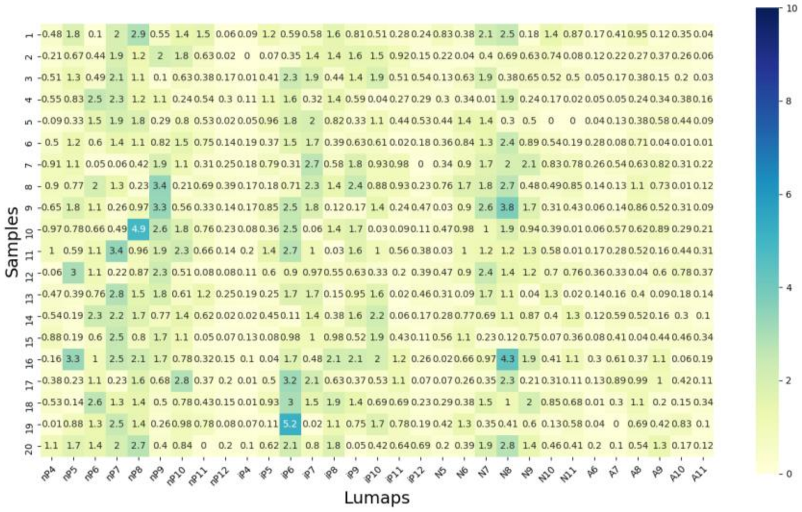

Fig. 2. Absolute errors of calculated mass fractions of 20 samples obtained by proposed method.

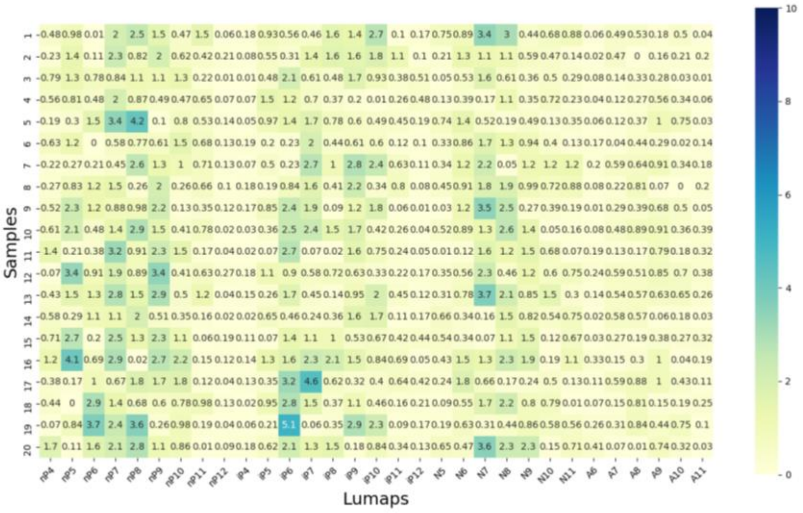

Fig. 3. Absolute errors of calculated mass fractions of 20 samples obtained by ANN.

that naphtha A is far from the training data range. The experimental and calculated values of molecular composition of the naphtha A are illustrated in Fig. 5. As shown in Fig. 5, the prediction performance of PINN-based molecular reconstruction method is better than ANN-based molecular reconstruction method, especially for *n*-pentane, C_{11} *iso*-paraffins, C_6 naphthenes, C_9 naphthenes, C_{11} naphthenes, ethylbenzene, C_9 aromatics and etc.

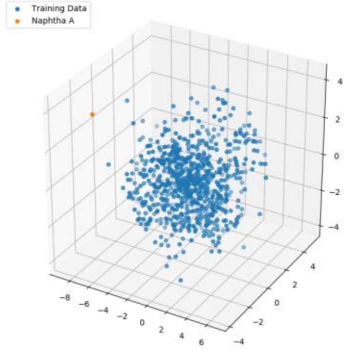

Fig. 4. Visualization result for training data and naphtha A.

4. CONCLUSIONS

In this work, a PINN-based molecular reconstruction method is proposed. By embedding mixing rules into the loss function of neural network, part of the network parameters could be explained with certain physical meanings, which can improve the generalization ability of the neural network model. Naphtha samples are investigated to validate the proposed method. The results show that compared with the

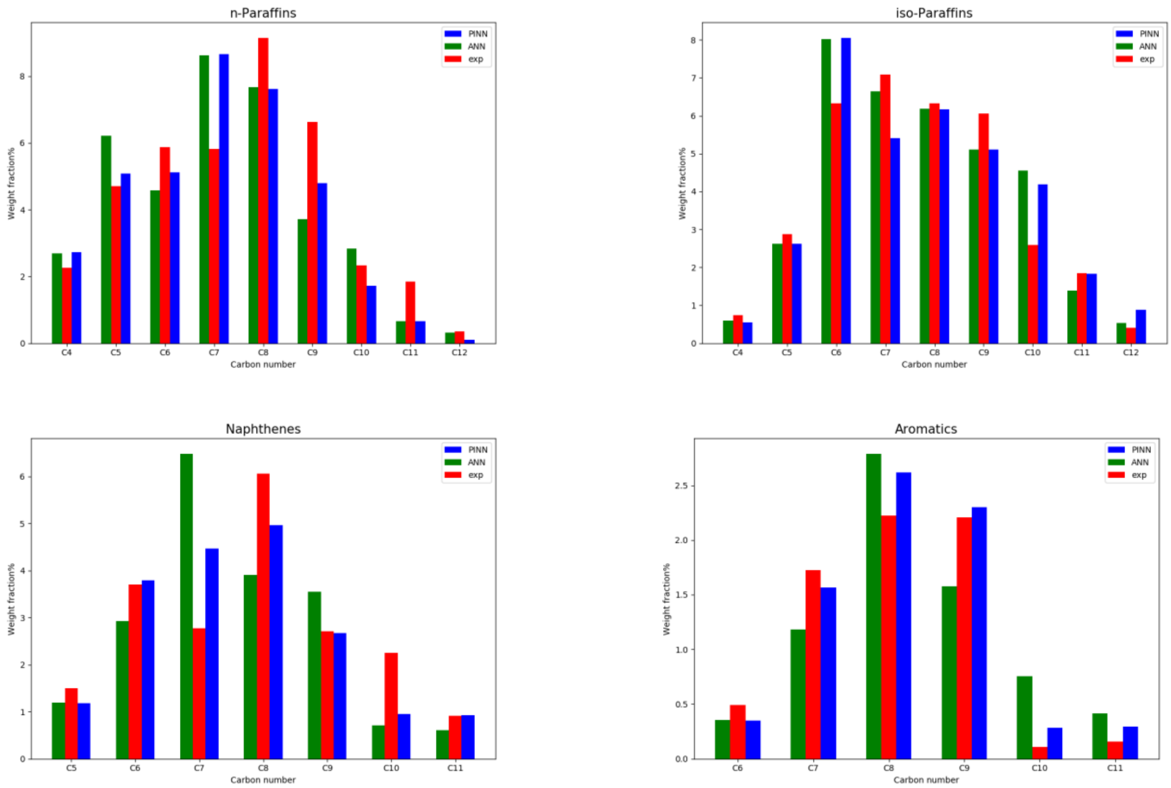

Fig. 5. Experimental values and calculated values of the molecular composition of naphtha A.

typical ANN-based molecular reconstruction method, the prediction performance and application range of the proposed method are improved.

REFERENCES

Bi K., Qiu T., (2019a). A Novel Naphtha Molecular Reconstruction Process Using Self-Adaptive Cloud Model and Hybrid GA-PSO Algorithm[J]. Industrial & Engineering Chemistry Research, 58(36):16753-16760.

Bi, K., & Qiu, T. (2019b). An intelligent SVM modeling process for crude oil properties prediction based on a hybrid GA-PSO method[J]. Chinese Journal of Chemical Engineering, 27(8), 1888-1894.

Karniadakis G. E., Kevrekidis I. G., Lu L., et al., (2021). Physics-informed machine learning[J]. Nature Reviews Physics.

Kolmogorov A. N., (1957). On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition[J]. Dokl.akad.nauk Sssr, 2(2):953-956.

Mei, H., Wang, Z., Huang, B. (2017)., Molecular-based Bayesian regression model of petroleum fractions[J]. Industrial & Engineering Chemistry Research. 56 (50): 14865-14872.

Ren Y., Liao Z., Sun J., et al., (2019a). Molecular Reconstruction of Naphtha via Limited Bulk Properties: Methods and Comparisons[J]. Industrial & Engineering Chemistry Research, 58:18742-18755.

Ren Y., Liao Z., Sun J., et al., (2019b). Molecular reconstruction: Recent progress toward composition modeling of petroleum fractions[J]. Chemical Engineering Journal, 357: 761-775.

Raissi M., Perdikaris P., Karniadakis G. E., (2018). Physics-Informed Neural Networks: A Deep Learning Framework for Solving Forward and Inverse Problems Involving Nonlinear Partial Differential Equations[J]. Journal of Computational Physics, 378:686-707.

Riazi M. (2005). Characterization and Properties of Petroleum Fractions[M]. West Conshohocken: ASTM International.

Rumelhart D. E., Hinton G. E., Williams R.J., (1986). Learning representations by back-propagating errors, Nature 323 (6088): 533–536.

Wang, K., & Li, S. (2017). Modified molecular matrix model for predicting molecular composition of naphtha[J]. Chinese Journal of Chemical Engineering, 25(12), 1856-1862.

Stratiev, D., Shishkova, I., Tankov, I., & Pavlova, A. (2019). Challenges in characterization of residual oils[J]. A review. Journal of Petroleum Science and Engineering, 178, 227-250.

Steven P., Kevin M., Marie-Franoise R., et al., (2010). Molecular reconstruction of complex hydrocarbon mixtures: An application of principal component analysis[J]. Aiche Journal, 2010, 56: 3174-3188.