Energy and Fuels, Volume 32, 2018, pp. 5779-5788
DOI: 10.1021/acs.energyfuels.8b00470

Correlating asphaltene dimerization with its molecular structure by potential of mean force calculation and data mining

Xinzhe Zhu [a, b], Guozhong Wu [a, b], Frédéric Coulon [c], Lvwen Wu [d], Daoyi Chen [a, b, *]

[a] Division of Ocean Science and Technology, Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China

[b] School of Environment, Tsinghua University, Beijing 100084, China

[c] School of Water, Energy and Environment, Cranfield University, Cranfield MK43 0AL, UK

[d] School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

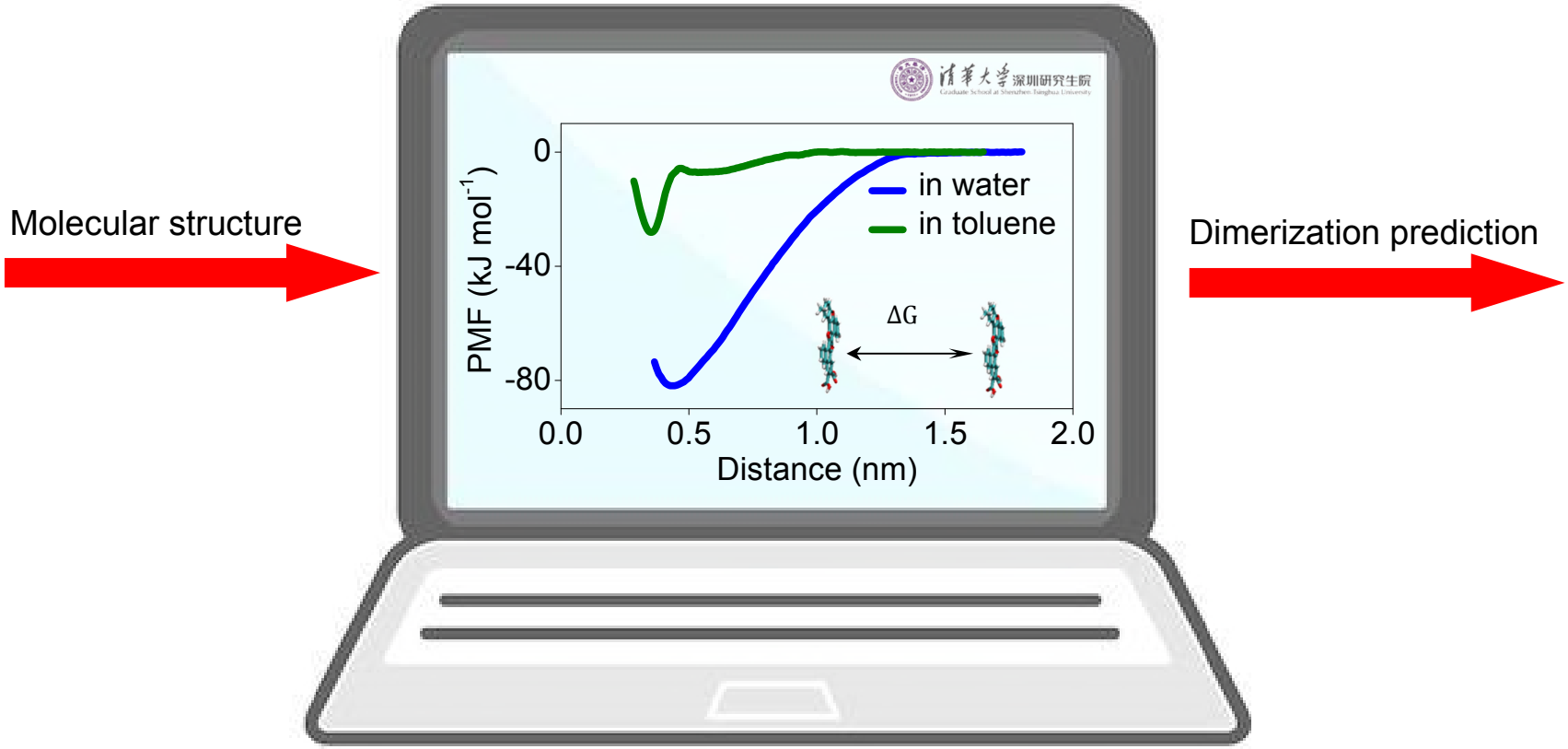
*Corresponding Author:

Phone: +86 0755 2603 0544

Fax: +86 0755 2603 0544

Email: chen.daoyi@sz.tsinghua.edu.cn

1



MACHINE LEARNING

Abstract: Asphaltene aggregation affects the entire production chain of petrochemical industry, which also poses environmental challenges for oil pollution remediation. The aggregation process has been investigated for decades, but it remains unclear how the free energy of asphaltene association in solvents is correlated with its molecular structure. In this study, dimerization energy of 28 types of asphaltenes in water and toluene were calculated using the umbrella sampling method. Structural parameters related to the atom types and functional groups were screened to identify the factors most influencing the dimerization energy using multiple linear regression, multi-layer perceptron and support vector regression. Results demonstrated that the influences of molecular structure on asphaltene association in water was nonlinear, while attempts to capture the relationship using linear regression had large error. The linkage per aromatic ring, number of aromatic carbons and aliphatic chains were the top three factors accounting for 52% of the dimerization energy variation in water. Asphaltene dimerization in toluene was dominated by the content of sulfur in aromatic rings and the number of aromatic carbons which contributed to 55% of the energy variation. To the best of our knowledge, this was the first study successfully predicting asphaltene dimerization using molecular structure ($R > 0.9$) and quantifying simultaneously the relative importance of each structural parameter. The proposed modelling approach supported the decision making on the number of structural parameters to investigate for predicting asphaltene aggregation.

Keywords: Asphaltene aggregation; PMF; Multiple linear regression; Multi-layer perceptron; Support vector regression

1. Introduction

Asphaltenes are the largest, heaviest and most polar fraction of crude oil, which usually contain poly-condensed aromatic ring linked by aliphatic chains and heteroatoms like nitrogen, sulfur and oxygen. Such structural properties make them easy to aggregate, flocculate and deposit even at low concentrations, which affects the entire production chain of petrochemical industry.[1] A large number of studies have been carried out to address the flow-assurance issues arose from asphaltene aggregation.[2] Our recent studies also demonstrated the environmental significance of asphaltene aggregation process on the fate and transport of oil pollutants in the contaminated soils. For example, asphaltene aggregates could form porous network in which the small oil molecules are sequestrated and become difficult to remove. [3] They also have the tendency to co-aggregate with the soil organic matter (SOM), a key component of soil responsible for the persistence of recalcitrant residual oils in the oil-soil matrix, at the oil-water interface and in the bulk water.[4] The formation of asphaltene-humic acids complex decrease the diffusion coefficient and increases the hydrophilicity of the asphaltenes in the clay pores.[5] These findings highlight the potential role of asphaltene aggregation on the extractability and bioavailability of the residual oil in the aged soil.

Successful prediction of asphaltene macro-scale behaviour requires an understanding of the aggregation in terms of micro-scale mechanism and aggregation strength. Without this fundamental knowledge, any attempt to model the environmental behaviour of asphaltenes will be incomplete or inaccurate. Previous studies indicated that asphaltene aggregation is influenced by many factors such as the molecular

structure and concentration of asphaltenes,[6-8] temperature and pressure,[9, 10] and solvent types.[11, 12] Among these factors, the molecular structure of asphaltenes are of particular interest due to the great diversity and complexity of asphaltene from various sources.[13] For instance, it was estimated that the number of rings in a single asphaltene fused ring system can range from a few rings to up to 20 rings.[13]

Several studies have been carried out to investigate the relationship between asphaltene aggregation and its molecular structure using molecular simulation techniques such as classical molecular mechanics, molecular dynamics (MD) simulation and density functional theory method in quantum mechanics. Generally, the π-π interaction between poly-aromatic rings was acknowledged as the main association force.[4, 14-16] The lower H/C ratio and the higher molecular weight and aromaticity were also proved to favour the asphaltene aggregation.[17, 18] However, the contribution of some other structural parameters to the asphaltene aggregation varied against studies with different simulation conditions. For example, hydrogen bonding has been proposed as one of the driving forces for asphaltene association in toluene or crude oil, [15, 19] but in other studies the intermolecular hydrogen bonds were not observed during the asphaltene dimerization in vacuum condition.[20] According to the modified Yen model,[21] the steric repulsion between aliphatic side chains was a force limiting asphaltene aggregation in toluene or heptane. In contrast other studies demonstrated that the asphaltene aggregates formation in water was favored by the very long or very short side chains in the asphaltene molecules.[22]

Another issue that was less taken into account in previous studies was the existence of

co-factors effects among the molecular structure parameters. For example, the same backbone structure was used by only changing the side-chain length,[22] but actually the varied side-chain length also led to the changes in the molecular weight and the ratio of hydrogen to carbon which might have synergistic or antagonistic effects on the asphaltene aggregation. If the side-chain length was investigated without changing the latter two parameters, then the number of chains had to be altered to keep mass balance. When more than one structural parameter varied simultaneously, the relative contribution of each parameter to the binding energy of two asphaltene molecules remained unelucidated. To the best of our knowledge, this was mainly due to the lack of statistical data. As a way forward it is suggested to develop multi-factor regression to evaluate the importance of concomitant molecular structure parameters on the strength of asphaltene dimerization.

Quantitative structure-property relationship (QSPR) modeling is a useful tool to predict the relationships between molecular structures and the properties.[23] One challenge is the selection of a number of molecular models for asphaltenes with different structures, which is then used to calculate the free energy for the binding of each pair of asphaltenes and to establish mathematical models correlating these energies with the structural parameters. Although the asphaltenes are complex heterogeneous fractions, a variety of asphaltene model molecules have been proposed by many research groups that can mimic the properties of asphaltenes,[24] making it possible to address the above issue. In the present study, 28 asphaltene models from the literatures were selected to explore the relationship between asphaltene dimerization and its molecular structure using

conventional statistical methods (e.g. multiple linear regression) and machine learning methods (e.g. multi-layer perceptron and support vector regression) that have been successfully used in our previous studies for predicting environmental phenomenon.[25-27] Specific objectives of this study were to screen structural parameters that were significant to the asphaltene dimerization, predict asphaltene dimerization by structural parameter, and quantify the relative importance of each structural parameter to the asphaltene dimerization.

2. Methodology

2.1 Molecular dynamics simulation

MD simulations were performed using the GROMACS 5.1 software,[28] while the molecular interactions was calculated by the CHARMM36 force field.[29, 30] As shown in Fig. 1, 28 types of asphaltene molecules with different structures were selected from literatures.[6, 11, 22, 31-34] Each molecule contained a single polyaromatic core with side chains, because previous experiments evidenced that such "continental model" was the dominant asphaltene architecture, providing the main aspects proposed by the Yen-Mullins model.[35] For example, using the Laser desorption laser ionization mass spectra, Sabbah et al.[36] demonstrated that all the 23 types of model compounds having one aromatic core showed little or no fragmentation, which well-mimicked the behavior observed for the 2 petroleum asphaltene samples. However, all model compounds with more than one aromatic core showed energy-dependent fragmentation, suggesting that the "archipelago models" were not dominant in asphaltenes. Recently, using the atomic

force microscopy and scanning tunnelling microscopy, Schuler et al.[37] measured the structure of more than 100 asphaltene molecules, which also concluded that a single aromatic core was the dominant asphaltene architecture although in some cases the "archipelago structure" was also present.

The free energy of dimerization of asphaltene molecules in water and toluene was obtained using the umbrella sampling method.[6, 38] A step-by-step guide for the umbrella sampling was available from the Gromacs tutorials website (http://www.bevanlab.biochem.vt.edu/Pages/Personal/justin/gmx-tutorials/umbrella/index.html). Before umbrella sampling, the configuration of each asphaltene dimer was obtained from a 2 ns MD simulation, where each two asphaltene molecules were randomly placed in a simulation box (4 nm × 4 nm × 4 nm) filled with water or toluene. The coordinates of asphaltene dimer at the end of MD simulation was extracted and located in the edge of a new simulation box (4 nm × 4 nm × 7.5 nm) with the aromatic rings parallel to the x-y plane. The simulation box was filled with water and toluene, respectively. It was then equilibrated by running MD simulation at NVT and NPT ensemble for 1 ns, respectively, where the asphaltenes were kept restrained. The center of mass (COM) pulling was employed to generate a series of initial configurations for each umbrella window. The COM of one asphaltene molecules was pulled along the Z-axis at the rate of 2.6 nm ns^{-1} for 1 ns, while the other one was restrained with a harmonic force constant of 1000 kJ mol^{-1} nm^{-2}. Approximately 35 windows were selected with a space of 0.05 nm between two consecutive distances from the resulted pulling simulation. For each window, separate MD simulation (10 ns)

was performed using the biased umbrella potentials to restrain the molecules within the window. The potential of mean force (PMF) was calculated from unbiased probability distributions of the systems with the weighted histogram analysis method.[39] The free energy for asphaltene dimerization was obtained from the PMF curves, which was defined as the free energy when two asphaltene monomers moved from very far away (i.e., distance larger than 1.8 nm when the interaction between two monomers was negligible) to the most stable state with the lowest energy. Accordingly to previous studies, the most stable structure of asphaltenes dimers were formed when the two monomers were parallel stacked and the distance between each other was about 0.4 nm.[4, 40, 41] The absolute values of the dimerization energy were used for the following data processing and results discussions. A larger value suggested an easier aggregation of asphaltenes. Throughout simulations, the temperature (298 K) and pressure (1 bar) were controlled using the V-rescale thermostat[42] and Parrinello-Rahman method[43], respectively.

2.2 Data preprocessing

Structural parameters definition: In order to gain insights into the influence of molecular structures on the binding energy of asphaltene dimer, eleven molecular structural parameters were selected as follows: molecular weight (MW), number of aromatic carbons (N_{Aro-C}), number of aliphatic chains (N_{chains}), number of aliphatic carbons (N_{Ali-C}), number of carbons in the longest chains (N_{Ali-CL}), mass percentage of heteroatoms in the aliphatic chains (HA%) , mass percentage of N in the aromatic ring

(N%), mass percentage of S in the aromatic ring (S%), the number of hydroxyl groups (N_{OH}), the ratio of hydrogen to carbon (H/C), number of carboxyl groups (N_{COOH}), and the linkage per aromatic ring where a linkage was counted when each two aromatic rings shared one edge (N_{link}/ N_{aro}).

Normalization: To eliminate the effects from the units of different variables during regression, the data was standardized to the same scale with Z-score normalization according to the following equation:

$$x_i^* = \frac{x_i - \bar{x}}{\sigma}$$

where x_i^* and x_i represent the standardized value and initial value of the variable, $\bar{x}$ is the mean value of the variable, σ is the standard error of the variable. The standardized value of each parameter obeyed a normal distribution.

Cluster analysis: Cluster analysis was performed to split the dataset into several classes using the WEKA program package[44] with the expectation maximization algorithm.[45] The molecular structures were similar in one class and different from that in other classes. Molecules (80%) were randomly selected from each class for regression, while the remaining 20% molecules were used for external validation.

Structural parameters screening: when multiple influence factors coexisted, there might be multicollinearity among them, which would impact the accuracy of regression model and should be removed before model development. The first step was to carry out bivariate correlation between each two parameters and between each parameter and the dimerization free energy. If the correlation coefficient between two parameters exceed 0.95, the one with poorer correlation with the dimerization free energy was

removed and the other was retained.[46] Next step was to perform stepwise multiple linear regression (SMLR) to identify the significant parameters to the dimerization free energy among the above screened parameters.[47, 48] Briefly, partial F-statistics were computed for each parameter and the one with the highest F-value was inserted into the model. The partial F-statistics were computed again for all the remaining parameters and the one with the highest F-value was added to the model if the corresponding F-value exceeded a specified threshold value. These two parameters in the model were evaluated with partial F-test to see if each one was still significant. The parameter was removed from the model if it was no longer significant. This procedure was repeated for the remaining parameters until no more parameters yielded a partial F-value greater than the threshold and all parameters in the model remained significant.[49] The parameters screening process was carried out with the SPSS (version 22.0), from which the selected parameters were used for regression models building in the next subsection.

2. 3 Regression and validation

Multiple linear regression (MLR): MLR was employed to predict the relationship between the molecular structure parameters and the binding energy, which could be expressed by the following equation: $y = a_1x_1 + a_2x_2 + \cdots + a_kx_k + a_0$, where the a_1, a_2,…, a_k were the regression coefficients, and a_0 was the intercede of the regression line.

Multi-layer perceptron (MLP): MLP is one of the most commonly used artificial neural network (ANN) models, consisting of an input layer, hidden layers and an output

layer. Each layer consists of nodes that are connected with a certain weight to every node in the following layer. Except for the input nodes, each node is a nonlinear activation function that enables the network to compute complex nonlinear problems. The connection weights are changed after each data passing through the nodes in the network using the 'back propagation' technique. By this way, the error in the output compared to the expected result is minimized. This process is termed as training, which stops automatically when no more decrease in the errors of cross-validation samples. To seek for the model with high precision, two parameters are optimized by trial and error during the training process including (i) the learning rate, which controls the degree of changes in the connection weights during each iteration,[50] and (ii) the number of nodes in the hidden layer, which ranges from $(2n^{1/2} + m)$ to $(2n + 1)$ where *n* and *m* represented the number of input node and output node, respectively.[51]

Support vector regression (SVR): SVR is a regression technique with excellent performances in regression and time series prediction application, which is capable of rearranging the input data from nonlinear to linear using mathematical functions known as kernels. In this study, the SMOreg with a RBF kernel or poly-kernel function was also used to develop regression model, which could transform nominal attributes into binary ones and had been successfully applied to predict the QSPR models.[52, 53] The complexity parameter, which determines the trade-off between the model complexity and the tolerance of errors, was optimized by a number of trials.

Validation: The 10-fold cross-validation was used to test the performance of the models.[54] Briefly, the datasets used for training were evenly split into 10 folds. The

instances from 9 folds were used for training while the remaining one fold was used for testing. This process was repeated 10 times using a different fold for testing at each cycle. The effectiveness of the model training was assessed by the correlation coefficient (R) and root mean squared error (RMSE). To validate the predictability of the model, these two parameters were also calculated for those validation datasets not used in the training process.

3. Results and discussion

3.1 Free energy for asphaltene dimerization

Fig. 2 shows the changes in the PMF during the dimerization process of 28 asphaltene molecules in water and toluene, respectively. Despite the variances in the shape of the PMF curves, the deepest well of all the PMF curves were observed at 0.3 ~ 0.5 nm, which corresponded to the separation distance between the centers of mass of each two asphaltene molecules parallel stacking due to the π-π interactions in the aromatic regions.[4, 55, 56] It suggested that such type of stacking required low energy for asphaltene dimerization and therefore offered the stable configuration for the asphaltene dimers in both water and toluene. By a closer examination of the structure of asphaltene dimers at different positions on the PMF curves (two examples are shown in Fig. 3), we noted that the stacking manner between the two monomers changed by rotation during pulling process. In other words, if two monomers were initially located as the T-shape stacking, they would spontaneously transform to the parallel stacking when they moved closer to each other.

The free energy of dimerization was determined by the difference between the minimum value of PMF (r = 0.3 ~ 0.5 nm) and the value of PMF at higher distances when it reached a plateau (r > 1.8 nm). The calculated binding energy of asphaltenes with different molecular structures, which ranged from 22.3 to 81.5 kJ mol^{-1} in water and from 2.5 to 39.0 kJ mol^{-1} in toluene (Fig. 2), respectively. Particularly, the dimerization free energy for each specific type of asphaltene was 48 - 91% lower in toluene than in water (Fig. 2). This might be attributed to the interactions between the

benzene ring of toluene molecules with the aromatic sheets of asphaltenes through π - π interaction.[57] It suggested a greater tendency for asphaltenes aggregation in water than in toluene, which was consistent with previous studies.[4, 56]

3.2 Bivariate correlation and structural parameters screening

A preliminary examination of the bivariate correlation indicated that the absolute value of the Pearson correlation coefficient between the free energy of asphaltene dimerization in water and each individual structural parameter ranged from 0.068 to 0.459 (Table 1). The mass percentage of nitrogen in the aromatic rings was characterized as the most significant parameter ($p < 0.01$), while relatively less significance was found for the mass percentage of sulfur in the aromatic rings, molecular weight, number of aromatic carbons, linkage per aromatic ring, and number of aliphatic chains ($p < 0.05$). The remaining five parameters showed insignificant influence on the asphaltene dimerization in water ($p > 0.05$). By contrast, only four parameters were identified as significant factors influencing asphaltene dimerization in toluene including the mass percentage of sulfur in the aromatic rings, number of aromatic carbons, ratio of hydrogen to carbon and number of carboxyl groups. It indicated that some insignificant parameters became significant after changing the solvent. For example, the Pearson correlation coefficient between the dimerization free energy in toluene and the number of carboxyl groups was 0.518, which decreased to 0.053 in water (Table 1). It might be attributed to the fact that the hydrogen bonding between the carboxylic groups of two asphaltene molecules was a driving force for its

dimerization in toluene, but dimerization driven by this way was less in water because the carboxylic groups in asphaltene would also form hydrogen bonds with water.

Results also demonstrated that there were significant correlation among some independent variables. For example, the molecular weight was significantly correlated with all the other parameters except the linkage per aromatic ring, the mass percentage of sulfur in the aromatic rings and the number of carboxyl groups (Table 1). The number of aliphatic carbons was highly correlated with the molecular weight, the ratio of hydrogen to carbon and the number of carbon in the longest chains with the corresponding Pearson correlation coefficient of 0.797, 0.883 and 0.824, respectively. This meant that the changes in one structural parameter would result in the changes in some other structural parameter. Therefore, when we tried to interpret the effects of one specific structural parameter on the asphaltene aggregation, we should take into account the multi-factor interactions between this parameter and other parameters. Moreover, the Pearson correlation coefficients among independent variables were not large enough to allow removing any one before input to the regression models. According to the parameter screening protocol (see section 2.2), it remained impossible at this stage to remove any of these structural parameters because the Pearson correlation coefficients between each two structural parameters were all less than 0.95 (Table 1).

To reduce the parameters to a suitable number without losing any important information, we further performed SMLR analysis using the dimerization free energy data in the water and toluene, respectively. After removal of the non-significant parameters, seven and five parameters were selected to correlate with dimerization energy in water and in

toluene, respectively (Table 2). The variance inflation factor (VIF) was less than the threshold value of 10 suggesting that the remaining parameters were of no multicollinearity.[58]

According to the similarity in the molecular structure, the 28 asphaltene molecules were divided into three representative classes by cluster analysis, in which the number of asphaltene molecules were 8, 8 and 12, respectively. From each class, 80% molecules were randomly selected for regression model development and the remaining molecules were used for model validation.

3.3 Prediction of asphaltene dimerization in water

MLR: The free energy of asphaltene dimerization in water was correlated with the screened structural parameters using the MLR model as follows: E_{water} = 0.6168 $N_{Aro\text{-}C}$ + 0.5819 N_{link}/N_{aro} - 0.3979 $N_{ali\text{-}C}$ - 0.2488 HA% - 0.4006 N% + 0.4042 S% + 0.695 N_{chains} - 0.0012. The R and RMSE for the training dataset calculated by 10-fold cross validation was 0.9231 and 0.3624, respectively (Table 3). These two parameters were about 4% lower and 80% higher for the validation dataset, respectively, compared with that during model training. It should be noted that the energy predicted from this equation was dimensionless due to the normalization treatment before modelling. It could be transformed to the energy with units (kJ mol^{-1}) using the mean value and standard error as indicated in Section 2.2. The input and predicted free energies are shown in Fig. 4.

MLP: The well trained MLP network consisted of 7, 6 and 1 nodes in the input layer,

hidden layer and output layer, respectively. The optimized learning rate was 0.02. Higher or lower learning rate both resulted in larger error in the training and validation datasets. For example, when the learning rate increased from 0.02 to 0.05, the RMSE for the validation data almost doubled (Fig. 5). After training the data with these optimal parameters, the R and RMSE was 0.9247 and 0.3608, respectively (Fig. 4b). Little difference was observed in the R between training and validation data, but the RMSE was about 14% higher in the latter.

SVR: The optimized SVR model was obtained using the complexity parameter of 1200 and the RBF Kernel with gamma of 0.008. The R for the training and validation datasets was 0.9462 and 0.9104, respectively (Table 3). The RMSE for the training and validation datasets was 0.3055 and 0.5173, respectively.

Model comparison: The performance of the three models for predicting the dimerization free energy in water was compared by the RMSE for the validation dataset (Table 3). It demonstrated that the MLP model had the best performance with the lowest RMSE. The highest RMSE was observed in the MLR model, which was 13 - 32% higher than that in the other two models. This finding suggested that the influences of asphaltene molecular structure on its aggregation in water was more likely nonlinear, while attempts to capture the relationship using a linear regression method might resulted in larger error. Accordingly, the relative contribution of the structure descriptors to the predictive dimerization free energy was calculated using the hidden-input and hidden-output connection weights resulted from the MLP modelling.[25] As shown in Fig. 6a, the linkage per aromatic ring was characterized as the most important

factor which accounted for 18% of the changes in the dimerization free energy. Similar importance was found between the number of aromatic carbons and the number of aliphatic chains. The least important factor was the number of aliphatic carbons which contributed to less than 10% of the asphaltene dimerization.

3.4 Prediction of asphaltene dimerization in toluene

MLR: The free energy of asphaltene dimerization in toluene was correlated with the five selected parameters using the MLR model as follows: $E_{toluene}$ = 0.5192 $N_{Aro\text{-}C}$ + 0.1745 N_{link}/N_{aro} + 0.5757 S% + 0.4413 N_{COOH} + 0.1817 N_{chains} - 0.0203. The R and RMSE for the training dataset was 0.8931 and 0.4436, respectively (Table 3). For the validation dataset, little change was found in the R while the RMSE increased by about 6%. The input and predicted free energies are shown in Fig. 7.

MLP: The well trained MLP network consisted of 5, 6 and 1 nodes in the input layer, hidden layer and output layer, respectively. The optimized learning rate was 0.04. The R and RMSE for the training dataset was 0.8934 and 0.4217, respectively (Table 3). Little difference was observed in the R between training and validation data, but the RMSE was about 10% higher in the latter.

SVR: The optimized SVR model was obtained using the polynomial kernel with exponent of 0.72. The R for the training and validation datasets was 0.9114 and 0.8989, respectively (Table 3). The RMSE for the training datasets was 0.3663 which increased by 30% for the validation dataset.

Model comparison: As we did in the water, the predictability of the three models for

dimerization free energy in toluene was compared by the RMSE for the validation dataset (Table 3). The highest RMSE was found in the SVR model, while very little difference (~ 1%) was observed between MLR and MLP. Same tendency was observed in the relative contribution of each structural parameter to the predicted dimerization free energy between MLR and MLP (Fig. 6b). As suggested by both models, the mass percentage of sulfur in the aromatic ring and the number of aromatic carbons were recognized as the two most important factor which contributed to more than 55% for the asphaltene aggregation (Fig. 6b).

It was interesting to find that (i) the number of aromatic carbons was strongly correlated with the asphaltene dimerization both in water and toluene, suggesting the attraction between poly-aromatic cores was the main driving force in both cases; (ii) the number of carboxyl groups had ignorable influence on the asphaltene aggregation in water, but it significantly influence the asphaltene aggregation in toluene with about 20% contributions. This was attributed to the hydrogen bonding phenomenon as aforementioned in the bivariate correlation; (iii) the top two most important factors (linkage per aromatic ring and number of aliphatic chains) for asphaltene aggregation in water became the least important in toluene. It suggested that the interactions between aliphatic chains contributed little to the strength of aggregation in toluene, but played very important roles for aggregation in water. This finding was consistent with previous study which demonstrated that the contacts between aliphatic chains in water was 2.6 - 5.7 folds of that in toluene. [22, 59]

Overall results indicated that the combination of MD simulation with machine learning

provided a potential way to predict asphaltene dimerization using the molecular structure parameters. It is well-acknowledged that asphaltene aggregation highly depend upon its molecular structure and efforts have been made to identify the sensitivity of asphaltene aggregation toward the molecular architecture. However, the correlation between asphaltene binding energy and multiple structure parameters with statistical significance has not been well determined. This study is one of the first attempts to do so. Unlike conventional statistical regression methods, machine learning approaches do not rely on any assumption of thermodynamic equation before modelling. For example, asphaltene aggregation was often hypothesized as a first or pseudo-first order reaction in some reported models.[60, 61] Even though such approximation was simple enough, the value of the rate constant was uncertain or even lack for asphaltenes with various molecular structures. By contrast, there is not a parameter like rate constant during machine learning prediction. Instead, it is able to distinguish the given data based on their different patterns and make useful decisions in new data. Therefore, machine learning methods do not have the issue of rate constant encountered in statistical methods. Such methods are particularly useful to handle problems with high non-linearity that will increase the difficulty in determining the rate constant or the form of thermodynamic equations for asphaltene aggregation. While not all parameters associated with the molecular architecture are necessary for understanding the asphaltene aggregation phenomenon, its prediction still requires detailed informative data in order to assess to what extent each parameter contribute to the aggregation. In the approach proposed in this study, the concomitant effects of parameters were taken

into account instead of only focusing one or two parameters by fixing the others. The study helps to identify and reduce the number of parameters that need to focus for predicting asphaltene aggregation.

4. Conclusions

Results demonstrated that the dimerization strength for each type of asphaltene was 48 - 91% lower in toluene than in water. Asphaltenes association in water was significantly influenced by seven structural parameters including the linkage per aromatic ring, number of aliphatic chains, number of aromatic carbons, number of aliphatic carbons, mass percentage of heteroatoms in the aliphatic chains, mass percentage of N in the aromatic ring and mass percentage of S in the aromatic ring. The first three were characterized as the most important parameters by the MLP model ($R = 0.9243$) which outperformed the MLR and SVR models. Five structural parameters were screened for predicting asphaltene association in toluene including the mass percentage of S in the aromatic ring, number of aromatic carbons, number of carboxyl groups, linkage per aromatic ring and number of aliphatic chains, while the first two parameters contributed to 55% of the energy variation for dimerization. Comparison between the two solvents highlighted the strong contribution of the aliphatic side chains to the asphaltene aggregation in water, which was insensitive in toluene. Overall results from this study provided a sound basis for improving the asphaltene phase behavior prediction. Further works are needed to investigate how microscopic predictions can be integrated into macroscopic deposition models to allow easy integration with commercially available

multiphase flow prediction tools.

Acknowledgements

This study was financially supported by the Shenzhen Peacock Plan Research Grant (KQJSCX20170330151956264) and the Fundamental Research Project of Shenzhen, China (JCYJ20160513103756736).

References

(1) Adams, J. J., Asphaltene Adsorption, a Literature Review. *Energy Fuels* **2014**, 28, 2831-2856.

(2) Zhang, X.; Pedrosa, N.; Moorwood, T., Modeling asphaltene phase behavior: comparison of methods for flow assurance studies. *Energy Fuels* **2012,** 26, 2611-2620.

(3) Li, X.; He, L.; Wu, G.; Sun, W.; Li, H.; Sui, H., Operational Parameters, Evaluation Methods, And Fundamental Mechanisms: Aspects of Nonaqueous Extraction of Bitumen from Oil Sands. *Energy Fuels* **2012,** 26, 3553-3563.

(4) Zhu, X.; Chen, D.; Wu, G., Molecular dynamic simulation of asphaltene co-aggregation with humic acid during oil spill. *Chemosphere* **2015,** 138, 412-421.

(5) Zhu, X.; Chen, D.; Wu, G., Insights into asphaltene aggregation in the Na-montmorillonite interlayer. *Chemosphere* **2016,** 160, 62-70.

(6) Sedghi, M.; Goual, L.; Welch, W.; Kubelka, J., Effect of asphaltene structure on association and aggregation using molecular dynamics. *J. Phys. Chem.* B **2013,** 117, 5765-5776.

(7) Mullins, O. C., Review of the molecular structure and aggregation of asphaltenes and petroleomics. *Spe. J.* **2008,** 13, 48-57.

(8) Roux, J. N.; Broseta, D.; Demé, B., SANS study of asphaltene aggregation: concentration and solvent quality effects. *Langmuir* **2001,** 17, 5085-5092.

(9) Maqbool, T.; Srikiratiwong, P.; Fogler, H. S., Effect of temperature on the precipitation kinetics of asphaltenes. *Energy Fuels* **2011,** 25, 694-700.

(10) Espinat, D.; Fenistein, D.; Barre, L.; Frot, D.; Briolant, Y., Effects of temperature and pressure on asphaltenes agglomeration in toluene. A light, X-ray, and neutron scattering investigation. *Energy fuels* **2004,** 18, 1243-1249.

(11) Kuznicki, T.; Masliyah, J. H.; Bhattacharjee, S., Aggregation and partitioning of model asphaltenes at toluene-water interfaces: Molecular dynamics simulations. *Energy Fuels* **2009,** 23, 5027-5035.

(12) Rane, J. P.; Harbottle, D.; Pauchard, V.; Couzis, A.; Banerjee, S., Adsorption kinetics of asphaltenes at the oil–water interface and nanoaggregation in the bulk. *Langmuir* **2012,** 28, 9986-9995.

(13) Groenzin, H.; Mullins, O. C., Molecular size and structure of asphaltenes from various sources. *Energy Fuels* **2000,** 14, 677-684.

(14) da Costa, L. M.; Stoyanov, S. R.; Gusarov, S.; Tan, X.; Gray, M. R.; Stryker, J. M.; Tykwinski, R.; de M. Carneiro, J. W.; Seidl, P. R.; Kovalenko, A., Density Functional Theory Investigation of the Contributions of π–π Stacking and Hydrogen-Bonding Interactions to the Aggregation of Model Asphaltene Compounds. *Energy Fuels* **2012,** 26, 2727-2735.

(15) Murgich, J., Intermolecular Forces in Aggregates of Asphaltenes and Resins. *Petrol. Sci. Technol.* **2002,** 20, 983-997.

(16) Tan, X.; Fenniri, H.; Gray, M. R., Pyrene derivatives of 2, 2-bipyridine as models for asphaltenes: synthesis, characterization, and supramolecular organization. *Energy Fuels* **2008,** 22, 715-720.

(17) Rogel, E., Simulation of interactions in asphaltene aggregates. *Energy Fuels* **2000,** 14, 566-574..

(18) Ungerer, P.; Rigby, D.; Leblanc, B.; Yiannourakou, M., Sensitivity of the aggregation behaviour of asphaltenes to molecular weight and structure using molecular dynamics. *Mol. Simulat.* **2014,** 40, 115-122.

(19) Liao, Z.; Zhou, H.; Graciaa, A.; Chrostowska, A.; Creux, P.; Geng, A., Adsorption/occlusion characteristics of asphaltenes: Some implication for asphaltene structural features. *Energy fuels* **2005,** 19, 180-186.

(20) Carauta, A. N.; Correia, J. C.; Seidl, P. R.; Silva, D. M., Conformational search and dimerization study

of average structures of asphaltenes. *J. Mol. Struc. THEOCHEM* **2005,** 755, 1-8.
(21) Mullins, O. C., The modified Yen model. *Energy Fuels* **2010,** 24, 2179-2207.
(22) Jian, C.; Tang, T.; Bhattacharjee, S., Probing the Effect of Side-Chain Length on the Aggregation of a Model Asphaltene Using Molecular Dynamics Simulations. *Energy Fuels* **2013,** 27, 2057-2067.
(23) Katritzky, A. R.; Maran, U.; Lobanov, V. S.; Karelson, M., Structurally diverse quantitative structure−property relationship correlations of technologically relevant physical properties. J. Chem. Inf. Comp. Sci. **2000,** 40, 1-18.
(24) Sjöblom, J.; Simon, S.; Xu, Z., Model molecules mimicking asphaltenes. *Adv. Colloid. Interfac.* **2015,** 218, 1-16.
(25) Wu, G.; Kechavarzi, C.; Li, X.; Wu, S.; Pollard, S. J. T.; Sui, H.; Coulon, F., Machine learning models for predicting PAHs bioavailability in compost amended soils. *Chem. Eng. J.* **2013,** 223, 747-754.
(26) Sui, H.; Li, L.; Zhu, X.; Chen, D.; Wu, G., Modeling the adsorption of PAH mixture in silica nanopores by molecular dynamic simulation combined with machine learning. *Chemosphere* **2016,** 144, 1950-1959.
(27) Wu, G.; Coulon, F., Modelling the Environmental Fate of Petroleum Hydrocarbons During Bioremediation. In: McGenity TJ, Timmis KN, Nogales B, editors. Hydrocarbon and Lipid Microbiology Protocols: Statistics, Data Analysis, Bioinformatics and Modelling, Springer Berlin Heidelberg; 2016, pp 165-180.
(28) Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J., GROMACS: fast, flexible, and free. *J. Comput. Chem.* **2005,** 26, 1701-1718.
(29) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; Mackerell, A. D., Jr., CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.* **2010,** 31, 671-90.
(30) Rane, K. S.; Kumar, V.; Wierzchowski, S.; Shaik, M.; Errington, J. R., Liquid–Vapor Phase Behavior of Asphaltene-like Molecules. *Ind. Eng. Chem. Res.* **2014,** 53, 17833-17842.
(31) Gao, F.; Xu, Z.; Liu, G.; Yuan, S., Molecular Dynamics Simulation: The Behavior of Asphaltene in Crude Oil and at the Oil/Water Interface. *Energy Fuels* **2014,** 28, 7368-7376.
(32) Mullins, O. C., The asphaltenes. *Ann. Rev. Anal. Chem.* **2011,** 4, 393-418.
(33) Aray, Y.; Hernández-Bravo, R.; Parra, J. G.; Rodríguez, J.; Coll, D. S., Exploring the structure–solubility relationship of asphaltene models in toluene, heptane, and amphiphiles using a molecular dynamic atomistic methodology. *J. Phys. Chem. A* **2011,** 115, 11495-11507.
(34) Silva, H. S.; Sodero, A. C.; Bouyssiere, B.; Carrier, H.; Korb, J.-P.; Alfarra, A.; Vallverdu, G.; Bégué, D.; Baraille, I., Molecular Dynamics Study of Nanoaggregation in Asphaltene Mixtures: Effects of the N, O, and S Heteroatoms. *Energy Fuels* **2016,** 30, 5656-5664.
(35) Mullins, O. C.; Sabbah, H.; Eyssautier, J.; Pomerantz, A. E.; Barré, L.; Andrews, A. B.; Ruiz-Morales, Y.; Mostowfi, F.; McFarlane, R.; Goual, L., Advances in asphaltene science and the Yen–Mullins model. *Energy Fuels* **2012,** 26, 3986-4003.
(36) Sabbah, H.; Morrow, A. L.; Pomerantz, A. E.; Zare, R. N., Evidence for island structures as the dominant architecture of asphaltenes. *Energy Fuels* **2011,** 25, 1597-1604.
(37) Schuler, B.; Meyer, G.; Pena, D.; Mullins, O. C.; Gross, L., Unraveling the Molecular Structures of Asphaltenes by Atomic Force Microscopy. *J. Am. Chem. Soc.* **2015,** 137, 9870-6.
(38) Kästner, J., Umbrella sampling. *WIRES Comput. Mol. Sci.* **2011,** 1, 932-942.
(39) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A., The weighted histogram analysis method for free‐energy calculations on biomolecules. I. The method. *J. Comput. Chem.* **1992,**

13, 1011-1021.
(40) Alvarez-Ramirez, F.; Ramirez-Jaramillo, E.; Ruiz-Morales, Y., Calculation of the interaction potential curve between asphaltene-asphaltene, asphaltene-resin, and resin-resin systems using density functional theory. *Energy Fuels* **2006,** 20, 195-204.
(41) Pahlavan, F.; Mousavi, M.; Hung, A.; Fini, E. H., Investigating molecular interactions and surface morphology of wax-doped asphaltenes. *Phys. Chem. Chem. Phys.* **2016,** 18, 8840-8854.
(42) Bussi, G.; Donadio, D.; Parrinello, M., Canonical sampling through velocity rescaling. *J. Chem. Phys.* **2007,** 126, 014101.
(43) Parrinello, M.; Rahman, A., Strain fluctuations and elastic constants. J. Chem. Phys. **1982,** 76, 2662-2666.
(44) Frank, E.; Hall, M.; Holmes, G.; Kirkby, R.; Pfahringer, B.; Witten, I. H.; Trigg, L., Weka-a machine learning workbench for data mining. In: Maimon O, Rokach L, editors. Data mining and knowledge discovery handbook. Springer, Boston,MA; 2009, pp 1269-1277.
(45) Do, C. B.; Batzoglou, S., What is the expectation maximization algorithm. Nat. Biotechnol. **2008,** 26, 897-899.
(46) Samari, F.; Yousefinejad, S., Quantitative structural modeling on the wavelength interval (Δλ) in synchronous fluorescence spectroscopy. *J. Mol. Struct.* **2017,** 1148, 101-110.
(47) Riahi, S.; Mousavi, M.; Shamsipur, M., Prediction of selectivity coefficients of a theophylline-selective electrode using MLR and ANN. *Talanta* **2006,** 69, 736-740.
(48) Bordbar, M.; Ghasemi, J.; Faal, A. Y.; Fazaeli, R., Chemometric modeling to predict aquatic toxicity of benzene derivatives using stepwise-multi linear regression and partial least square. *Asian J. Chem.* **2013,** 25, 331.
(49) Fatemi, M. H.; Gharaghani, S.; Mohammadkhani, S.; Rezaie, Z., Prediction of selectivity coefficients of univalent anions for anion-selective electrode using support vector machine. *Electrochim. Acta* **2008,** 53, 4276-4282.
(50) Khandelia, H.; Mouritsen, O. G., Lipid gymnastics: evidence of complete acyl chain reversal in oxidized phospholipids from molecular simulations. *Biophys. J.* **2009,** 96, 2734-2743.
(51) Fletcher, D.; Goss, E., Forecasting with neural networks: an application using bankruptcy data. *Inform. Manag.* **1993,** 24, 159-167.
(52) Chauhan, J. S.; Dhanda, S. K.; Singla, D.; Agarwal, S. M.; Raghava, G. P.; Consortium, O. S. D. D., QSAR-based models for designing quinazoline/imidazothiazoles/pyrazolopyrimidines based inhibitors against wild and mutant EGFR. *PLoS One* **2014,** 9, e101079.
(53) Singla, D.; Anurag, M.; Dash, D.; Raghava, G. P., A web server for predicting inhibitors against bacterial target GlmU protein. *BMC pharmacol.* **2011,** 11, 5.
(54) Refaeilzadeh, P.; Tang, L.; Liu, H., Cross-validation. In *Encyclopedia of database systems*, Springer: 2009; pp 532-538.
(55) Pisula, W.; Tomovic, Z.; Simpson, C.; Kastler, M.; Pakula, T.; Müllen, K., Relationship between core size, side chain length, and the supramolecular organization of polycyclic aromatic hydrocarbons. *Chem. Mater.* **2005,** 17, 4296-4303.
(56) Kuznicki, T.; Masliyah, J. H.; Bhattacharjee, S., Molecular dynamics study of model molecules resembling asphaltene-like structures in aqueous organic solvent systems. *Energy Fuels* **2008,** 22, 2379-2389.
(57) Headen TF, Boek ES, Skipper NT. Evidence for asphaltene nanoaggregation in toluene and heptane from molecular dynamics simulations. *Energy Fuels* **2009**, 23, 1220-1229.

(58) Dormann, C. F.; Elith, J.; Bacher, S.; Buchmann, C.; Carl, G.; Carré, G.; Marquéz, J. R. G.; Gruber, B.; Lafourcade, B.; Leitão, P. J., Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* **2013,** 36, 27-46.
(59) Jian, C.; Tang, T.; Bhattacharjee, S., Molecular Dynamics Investigation on the Aggregation of Violanthrone78-based Model Asphaltenes in Toluene. *Energy Fuels* **2014**, 28, 3604-3613.
(60) Kurup, A. S.; Vargas, F. M.; Wang, J.; Buckley, J.; Creek, J. L.; Subramani, H., J; Chapman, W. G., Development and application of an asphaltene deposition tool (ADEPT) for well bores. *Energy Fuels* **2011,** 25, 4506-4516.
(61) Vargas, F. M.; Creek, J. L.; Chapman, W. G., On the development of an asphaltene deposition simulator. *Energy Fuels* **2010,** 24, 2294-2299.

Table 1 Pearson correlation matrix of different variables (** represents $P < 0.01$, * represents $0.01 < P < 0.05$)

	MW	N_{Aro-C}	N_{link}/N_{aro}	N_{Ali-C}	HA%	N%	S%	N_{OH}	H/C	N_{COOH}	N_{chains}	N_{Ali-CL}
MW	1.000											
N_{Aro-C}	0.428*	1.000										
N_{link}/N_{aro}	0.183	-0.070	1.000									
N_{Ali-C}	0.797**	-0.054	0.329*	1.000								
HA%	0.461**	0.200	0.039	0.286	1.000							
N%	-0.475*	-0.120	-0.178	-0.423*	-0.309	1.000						
S%	-0.081	0.096	-0.277	-0.259	-0.283	-0.212	1.000					
N_{OH}	-0.166	-0.508**	-0.152	0.083	-0.194	-0.182	0.115	1.000				
H/C	0.621**	-0.357*	0.144	0.883**	0.244	-0.287*	-0.250	0.287	1.000			
N_{COOH}	-0.189	0.095	-0.199	-0.430*	-0.176	0.216	0.147	-0.138	-0.515**	1.000		
N_{chains}	0.418*	-0.271	-0.128	0.511**	-0.093	-0.099	0.033	0.263	0.673**	-0.154	1.000	
N_{Ali-CL}	0.760**	0.057	0.346*	0.824**	0.403*	-0.346*	-0.287	-0.089	0.731**	-0.351*	0.207	1.000
Energy in water	0.387*	0.350*	0.318*	0.170	-0.232	-0.459**	0.425*	-0.139	0.002	0.053	0.346*	0.068
Energy in toluene	0.228	0.561**	-0.053	-0.176	-0.134	-0.263	0.608**	-0.212	-0.404*	0.518**	0.009	-0.234

Table 2 Screening of the structural parameters by stepwise multiple linear regression

(√ represents the parameters selected for the regression; X represents the parameters deleted before input to the regression model)

Solvent	MW	N_{Aro-C}	N_{link}/N_{aro}	N_{Ali-C}	HA%	N%	S%	H/C	N_{OH}	N_{COOH}	N_{chains}	N_{Ali-CL}	VIF
Water	X	√	√	√	√	√	√	X	X	X	√	X	2.554
Toluene	X	√	√	X	X	X	√	X	X	√	√	X	1.148

Table 3 Performance comparison between MLR, MLP and SVR models

Solvent	Model	Parameters	Training dataset	Validation dataset
Water	MLR	R	0.9231	0.8907
		RMSE	0.3624	0.6506
	MLP	R	0.9247	0.9243
		RMSE	0.3608	0.4975
	SVR	R	0.9462	0.9104
		RMSE	0.3055	0.5173
Toluene	MLR	R	0.8931	0.9061
		RMSE	0.4436	0.4687
	MLP	R	0.8934	0.9048
		RMSE	0.4217	0.4640
	SVR	R	0.9114	0.8989
		RMSE	0.3663	0.4769

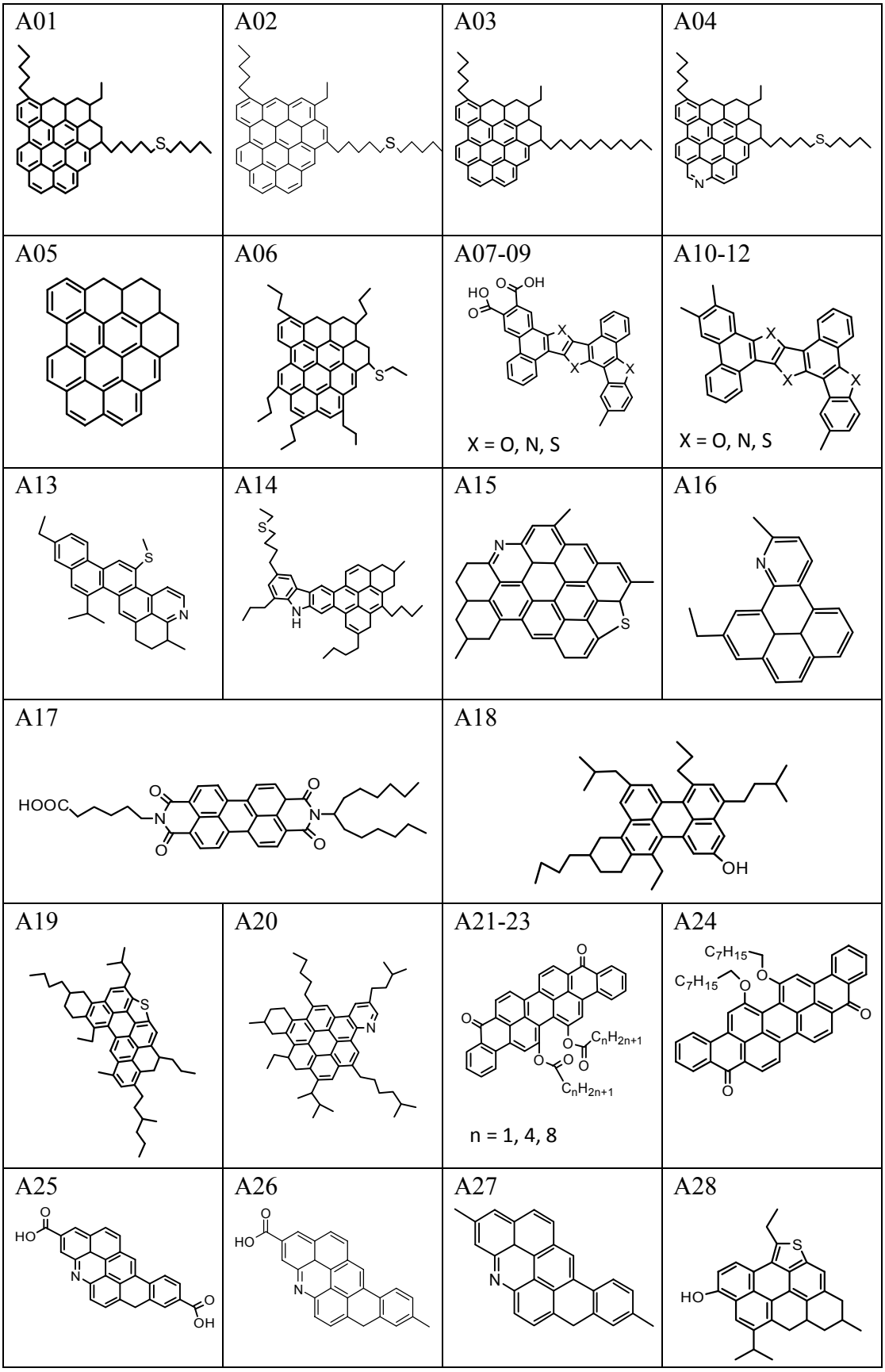
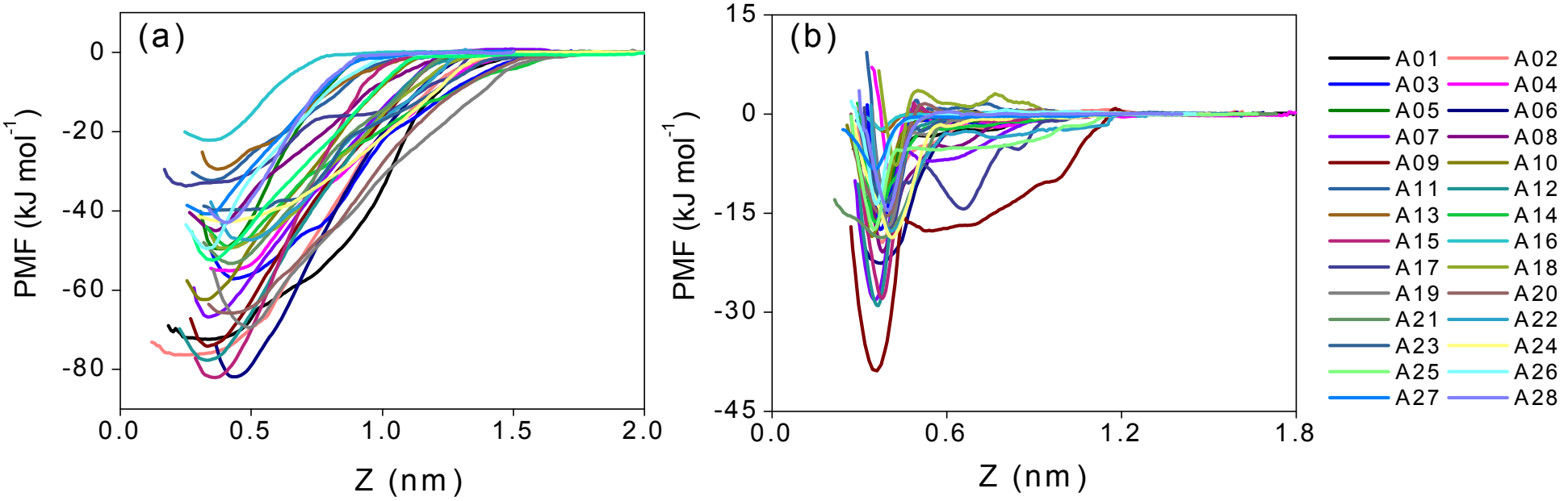


Fig. 1 Molecular structure of the 28 types of asphaltene models used in the study


(a)
PMF (kJ mol-1)
0
-20
-40
-60
-80
0.0
0.5
1.0
1.5
2.0
Z (nm)
(b)
PMF (kJ mol-1)
15
0
-15
-30
-45
0.0
0.6
1.2
1.8
Z (nm)
A01
A02
A03
A04
A05
A06
A07
A08
A09
A10
A11
A12
A13
A14
A15
A16
A17
A18
A19
A20
A21
A22
A23
A24
A25
A26
A27
A28

	A01	A02	A03	A04	A05	A06	A07	A08	A09	A10	A11	A12	A13	A14
E_{water}	72.3	76.3	57.4	55.0	49.7	81.5	67.6	45.1	74.3	62.4	32.1	77.1	29.9	49.5
$E_{toluene}$	14.9	19.4	17.2	8.8	13.3	22.6	28.0	21.5	39.0	18.6	13.7	28.5	2.7	16.4
	A15	A16	A17	A18	A19	A20	A21	A22	A23	A24	A25	A26	A27	A28
E_{water}	81.4	22.3	34.7	49.9	72.6	65.5	52.6	47.4	39.7	43.0	51.9	48.8	40.6	42.8
$E_{toluene}$	28.0	2.5	17.6	7.9	16.8	15.8	18.6	17.3	13.4	18.7	21.3	13.8	8.1	14.6

Fig. 2 PMF curves of asphaltene dimers in (a) water and (b) toluene.
Data in the table shows the free energy of dimerization (kJ mol^{-1}) for each type of asphaltene.

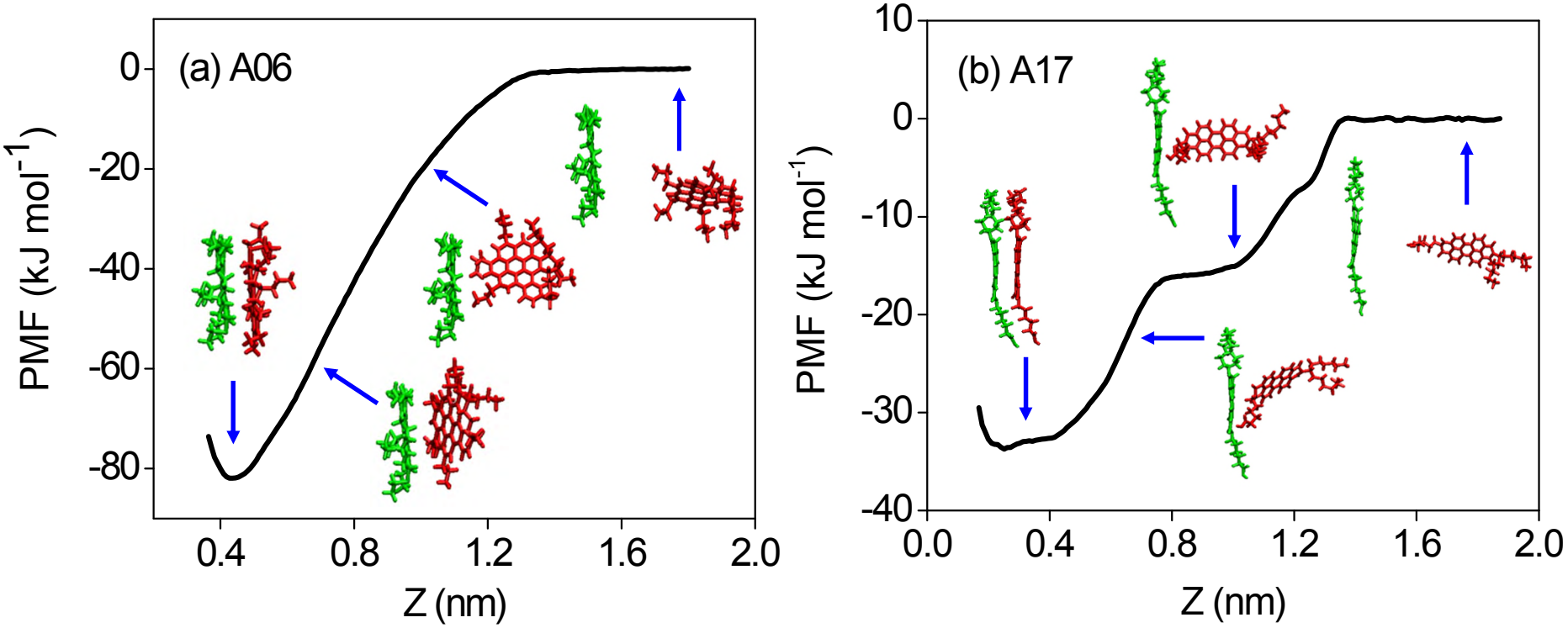


Fig. 3 PMF and structure of asphaltene dimers in water (A06 and A17 represent two examples selected from the 28 types of asphaltenes)

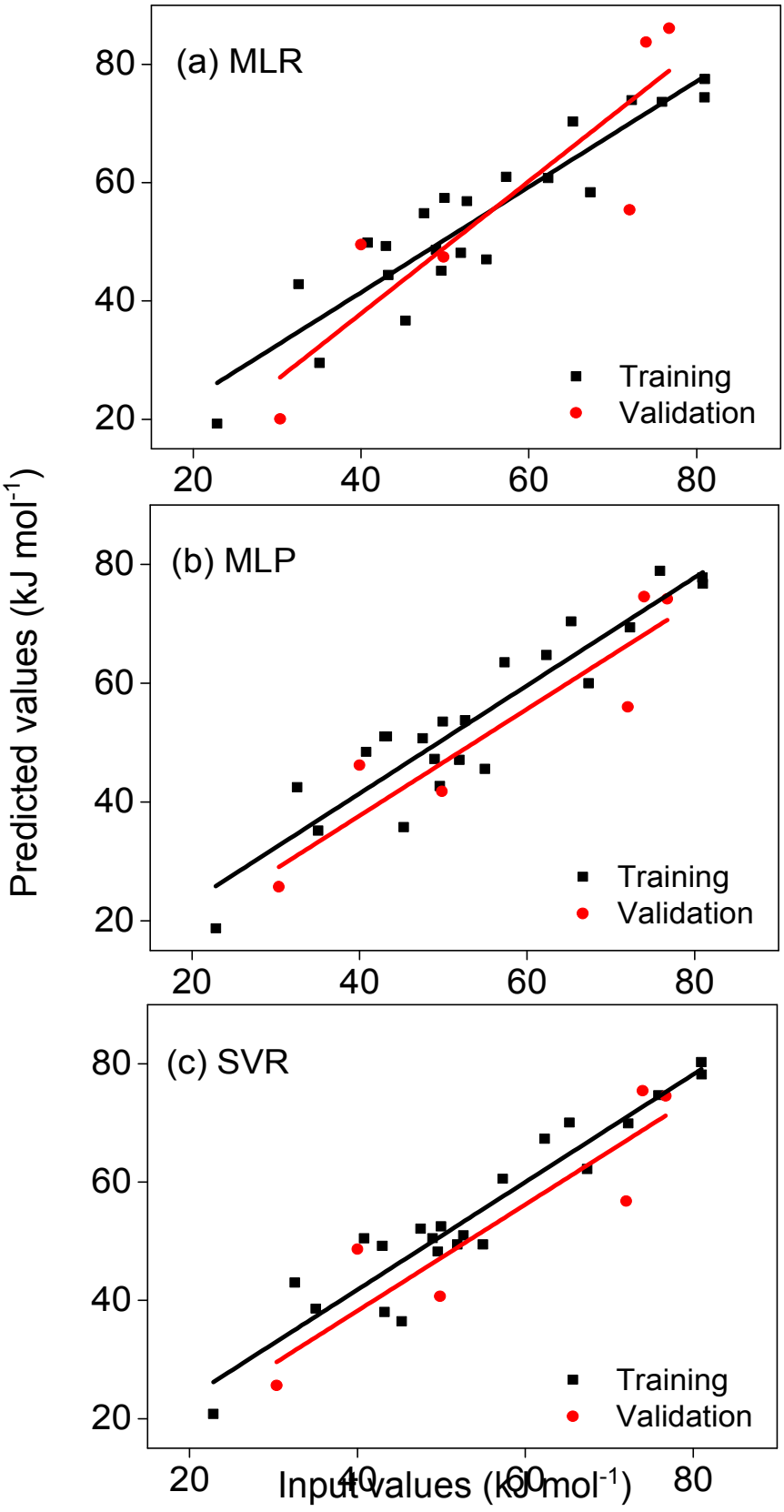


Fig. 4 Input and predicted values of asphaltene dimerization energy in water

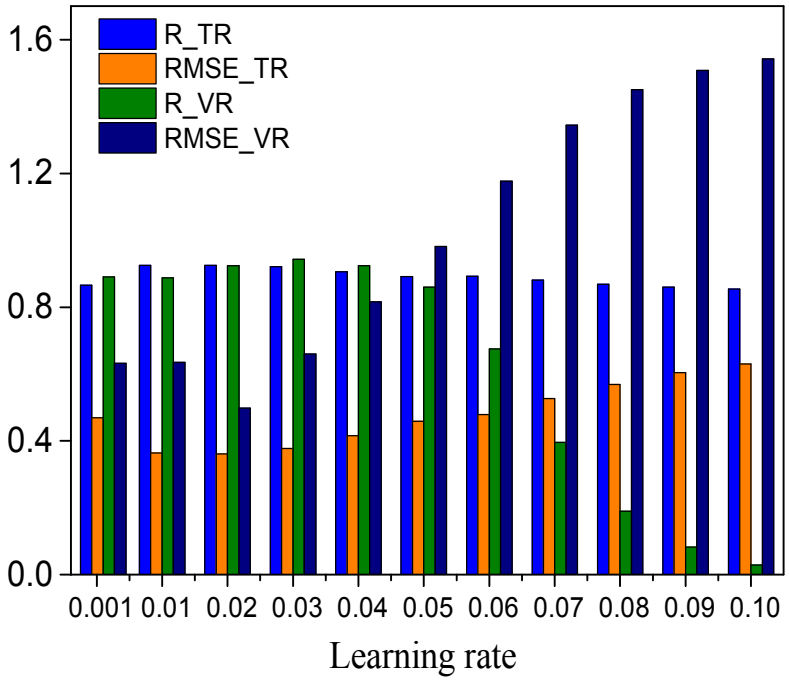


Fig. 5 Variation of R and RMSE for the training (TR) and validation (VR) datasets with different learning rates during MLP training process

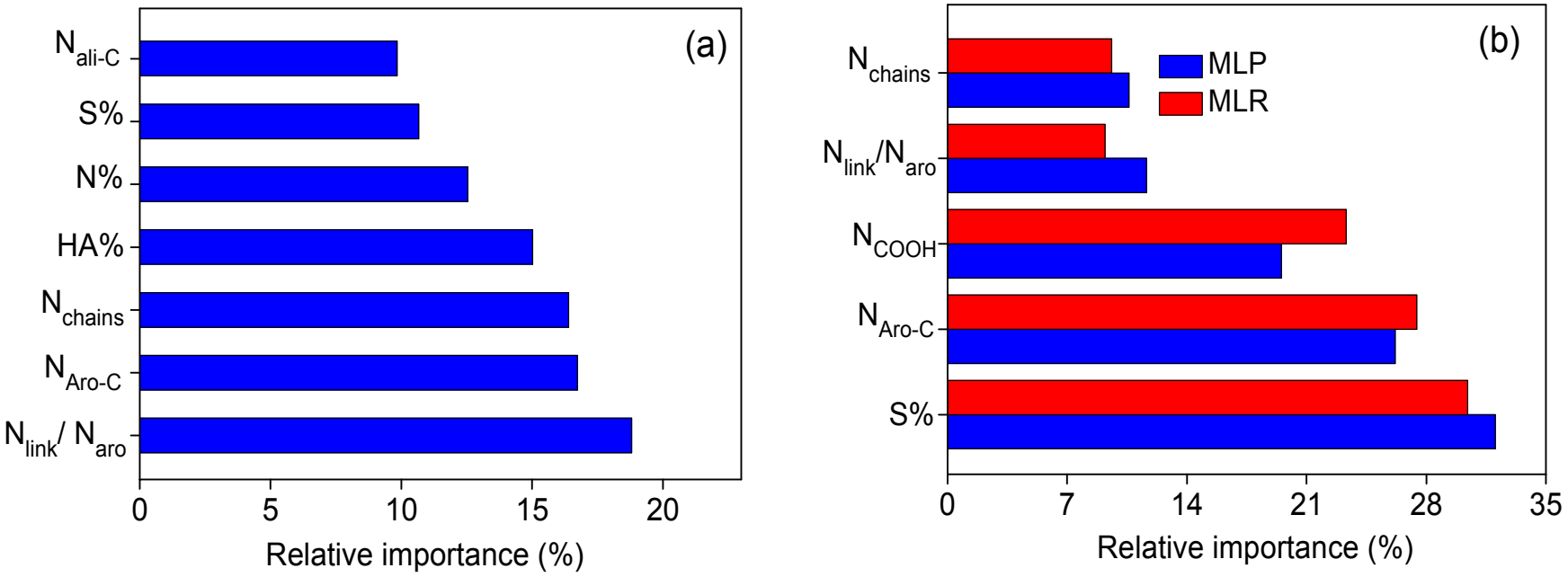


Fig. 6 Relative importance of structural parameters to the asphaltene dimerization in (a) water and (b) toluene. In panel A, only the results calculated by MLP model are shown because this model outperforms the other models.

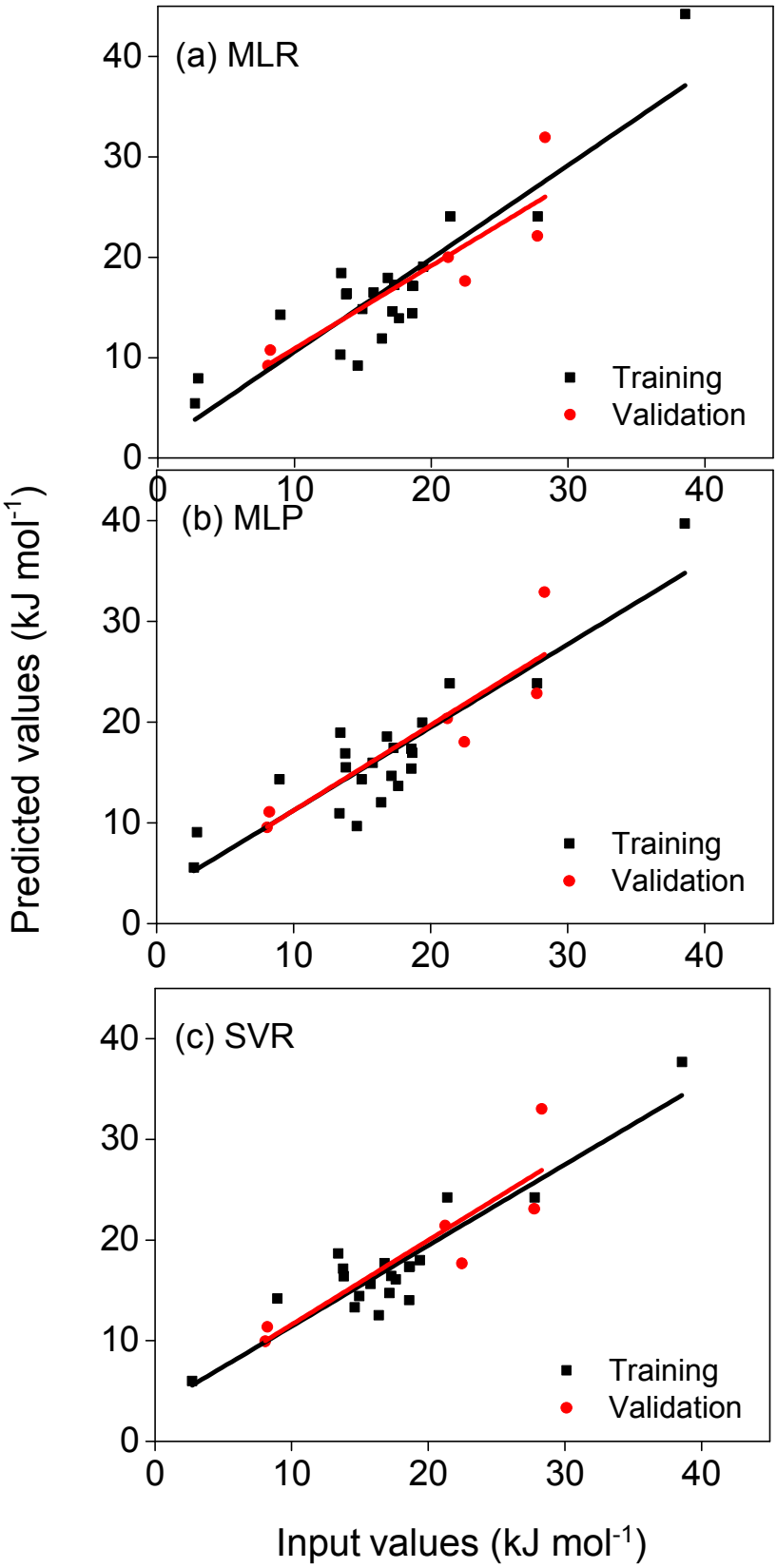


Fig. 7 Input and predicted values of asphaltene dimerization energy in toluene