Journal of Molecular Liquids, Volume 370, January 2023, Article number 120952
DOI:10.1016/j.molliq.2022.120952

Machine learning models for fast selection of amino acids as green thermodynamic inhibitors for natural gas hydrate

Guozhong Wu[1, 2, *], Frederic Coulon[4], Jing-Chun Feng[1, 3], Zhifeng Yang[1], Yuelu Jiang[2], Ruifeng Zhang[5]

[1] School of Ecology, Environment and Resources, Guangdong University of Technology, Guangzhou 510006, China

[2] Institute for Ocean Engineering, Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China

[3] Research Centre of Ecology & Environment for Coastal Area and Deep Sea, Guangdong University of Technology & Southern Marine Science and Engineering Guangdong Laboratory (Guangzhou), Guangzhou 510006, China

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

[5]Shenzhen Qihay Academy Science and Technology Co., Ltd., Shenzhen 518055, China

* Corresponding Author

Email: wgz@gdut.edu.cn

Abstract

Natural amino acids are non-toxic thermodynamic hydrate inhibitors without negative environmental impact, but it is difficult to accurately select the appropriate amino acid as a quick response to the operational conditions changes in the natural gas pipeline. The objective of this study was to develop mathematical models to predict the hydrate formation temperature (HFT) in presence of amino acids, capture the relationship between amino acid structure properties and their hydrate inhibition strength, and determine the optimal type and concentration to use. The HFT prediction was evaluated using multiple linear regression (MLR) and three machine learning methods including random forest (RF), M5 Rule (M5R) and support vector machine (SVM). After parameter optimization using the trial-and-error method, the coefficient of determination (R^2) of the four models were 0.9328, 0.9793, 0.9795 and 0.9980, respectively. The SVM prediction of HFT outperformed other models as the root mean square error (RMSE) was 83%, 76% and 69% lower than that of the MLR, RF and M5R, respectively. Results also demonstrated that the relative importance of the amino acid concentration to the hydrate phase equilibrium was 5-fold higher than that of the intrinsic properties of the amino acid molecular. The SVM model proposed in this study served an easy-to-use tool for reliable prediction of HFT by just providing a new set of input data. This made it possible to accurately determine the minimum concentration of amino acids to be used during the gas pipeline transportation.

Keywords: natural gas hydrate; amino acid; phase transition prediction; support vector machine

1. Introduction

Natural gas is a carbon-neutral and environmentally friendly fuel which is widely used for energy production and consumption. It is estimated that the global demand for natural gas will continue to rise by 3% each year until 2030 [1]. Pipeline transportation plays a key role in the natural gas supply [2]. However, the pipeline transportation is not without challenge as often pipeline blockage happened due to the formation of natural gas hydrates. They are ice-like crystalline compounds formed by the trapping of gas molecules in hydrogen bonded water molecules [3]. Gas hydrate plugged pipelines can result in gas production shutdown from a few days to a month with significant economic loss [4]. Hydrate inhibitors are often injected to prevent hydrate formation and reduce the occurrence of pipeline plugging, which include thermodynamic hydrate inhibitors (THI) and low-dosage hydrate inhibitors (LDHI) [3]. Generally, LDHI can be divided to kinetic inhibitors (KHI) and anti-agglomerate inhibitors. Although LDHI are new and promising chemicals, they have serious limitations and their efficacy remain unstable. For example, they can only decrease the nucleation and growth rate of hydrate crystals or prevent to some extent the agglomeration of hydrate crystals, but cannot prevent hydrate formation. The anti-agglomerate inhibitors generally require a certain amount of continuous oil phase in order to be effective [5]. KHI may lose the capability of hydrate inhibition or even promote hydrate nucleation at high sub-cooling conditions [6-8]. To date, THI remain the predominant inhibitors in the industrial application especially for the long-distance gas pipeline in cold waters, because they can shift the phase boundary to lower temperature and higher pressure where gas hydrate cannot form. However, the THI extensively used in the industry are mainly alcohols such as methanol, ethanol, ethylene glycol, diethylene

glycol and glycerol, which are toxic compounds with negative environmental impacts [9]. These chemicals are often added in large dosages which require complex post-treatment process to recycle and reuse them. Therefore, there is a continuous search for novel THI with high efficiency and low toxicity.

There are growing interest to employ amino acids as green THI, because they are natural compounds without toxicity and biodegradability issues and they are less expensive [10, 11]. Previous studies demonstrated that some amino acids could thermodynamically inhibit hydrate information but some others could promote the kinetics of hydrate formation [10, 12]. It was observed that the inhibition or promotion effects depended on a variety of factors such as the composition of natural gas and the concentration and hydrophobicity of amino acid. There are over 20 types of amino acids and it remains confusing on how to fast select the suitable types of amino acid due to the knowledge gaps in the relationship between the molecular properties of amino acid and the degree of phase boundary shift after adding amino acid. Additionally, the minimum concentration of amino acids needs to be accurately determined at given operational conditions (e.g., temperature, pressure and natural gas composition in the pipeline) in order to reduce the amount of amino acids and the corresponding cost. In the current industrial application, it is a safe but conservative practice that THI are often injected at the maximum mass concentration (up to 60%) to prevent hydrate formation to the largest degree [5, 6]. If the hydrate formation temperature (HFT) can be accurately predicted, the dynamic relationship between the hydrate phase boundary and the amino acids concentration can be rapidly captured. The minimum concentration of amino acids can be intelligently selected as long as the HFT is kept below the temperature in the pipeline. It is very time-consuming to measure the HFT for each

amino acid using phase equilibrium experiments and it is unrealistic to completely replicate the natural gas composition and temperature-pressure conditions in the gas pipeline by bench-scale experiments. Therefore, mathematical model being able to predict HFT is highly demanded for the fast selection of the molecular type and optimal concentration of amino acids.

Statistical thermodynamic models have been developed to predict the gas hydrate phase equilibrium [13-18]. These models often require many parameters and need to find proper values for these parameters [19, 20]. Moreover, thermodynamic models often assume a specific form of mathematical equation and statistical regression is used to determine the unknown parameters, but it is hard to assume an empirical equation due to the lack knowledge on the relationship between amino acids and HFT. In contrast, machine learning (ML) is a data-based method which allows computers to "learn" and "recognize" the patterns of the empirical data without relying on prior knowledge on the amino acids. It can be used to distinguish the given data based on their different patterns, extract useful information from the data and output predictions with new input variables [21]. Recently, a number of ML algorithms such as artificial neural networks, decision trees, k-nearest neighbor, gradient boosting regression, adaptive neuro Fuzzy interference system, gene expression programming, random forest (RF), support vector machines (SVM) have been used to predict gas hydrate formation and dissociation [19, 20, 22-29]. These studies mainly focused on the hydrate phase transition in aqueous solutions with or without traditional hydrate inhibitor such as monoethylene glycol, Luvicap 55 W and ionic liquids which might have negative environmental impacts as aforementioned. To the best of our knowledge, mathematical model predicting gas hydrate phase equilibrium in the presence

of amino acids has not been reported so far. Knowledge gaps also exist on understanding if and how the physicochemical properties (e.g., molecular weight, isoelectric point and hydrophobicity) of the inhibitors would influence the hydrate phase boundary, making it difficult for amino acid selection.

To address these issues, three ML methods including RF, M5 Rule (M5R) and SVM were used in this study to predict the HFT in the presence of ten types of amino acids based on the experimental data. These methods were selected owing to their unique advantages. RF is a kind of decision tree algorithm with parameters easily to set and is relatively less sensitive to outlier data, which can automatically produce accuracy assessments and measure the variable importance [30]. The advantage of M5R over many other ML methods is that it is not black-box model, which creates "if-else-then" constraints that can be expressed by easy-to-understand rules. The SVM is very effective for sparse and high-dimensional data with better convergence guarantee and less overfitting problem, which can be used for the data that is not regularly distributed and have unknown distribution [31]. In the present study, these three ML models were trained and validated using the molecular properties and concentration of amino acids. After model optimization by parameter tuning, the accuracy of these models were compared and the best model could be chose for future prediction of HFT in presence of amino acids. The type and minimum concentration of amino acids could then be determined when providing a new set of temperature-pressure conditions in the gas pipeline. In summary, the principle novelty of this work was to propose an easy-to-use tool to (i) fast predict HFT in presence of amino acids and reveal the relationship between molecular properties of amino acids and their thermodynamic inhibition strength, and (ii) quickly determine which type and what

minimum concentration of amino acid were required to prevent gas hydrate formation as a quick response to the operational conditions changes in the pipeline.

2. Methodology

2.1 Datasets

The experimental hydrate phase equilibrium data used for ML model building were collected from previously published articles (Table S1 in the supplementary material). Overall 210 sets of data were collected for natural gas HFT in the presence of 10 types of amino acids including glycine, alanine, serine, proline, arginine, lysine, valine, threonine, asparagine and phenylalanine [32-36]. In order to predict the HFT, eight independent variables were selected including (i) physicochemical properties of the amino acids such as molecular weight (MW, g mol^{-1}), isoelectric point (pI), hydropathy index (hI) and mass concentration (conc, wt%), (ii) gas composition of the natural gas such as the percentage of CH_4 (C_1%), C_2H_6 (C_2%) and C_3H_8 (C_3%), and (iii) system pressure (P). The basic physicochemical properties of amino acids were compiled from Liu et al., [37] and Bavoh et al. [10], which were summarized in Table S1.

The dataset was normalized (or scaled) within a uniform range (i.e., 0-1) before model development. This pre-processing step could prevent larger numbers from overriding smaller ones, which was especially important when the input data had large values. Different methods for data normalization have been reported in literatures and there is not a standard procedure for data normalization. In this study, the minimum and maximum values of the variable (x_i) in the dataset were calculated, which were then used to calculate

the scaled value (x_i^*) using Equation 1. By this way, each recorded value was normalized to the closed interval [0, 1].

$$x_i^* = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)} \quad (1)$$

2.2 Machine learning models

Three ML models including RF, M5R and SVM were evaluated to predict the natural gas HFT using the collected dataset (Table S1). Multiple linear regression (MLR) was also performed for comparison. The widely used software WEKA was used for model building, which was available at https://www.cs.waikato.ac.nz/ml/weka/. The principles, parameters and applications of these models have been well elaborated in literatures [30, 38, 39], therefore, they are only briefly introduced in this study.

Random forest (RF): RF is an ensemble classifier proposed by Breiman [30], which randomly split the training data to build a set of decision trees and train these trees in parallel. The term "train" refers to the process to achieve pattern recognition of the given data by minimizing an error function [21]. Decision tree is a specific type of flow chart used to visualize the decision-making process by mapping out different courses of action and the potential outcomes. There are three typical elements in a decision tree including the root node that represents the ultimate decision to make, the branches that stem from the root and represent different options, and the leaf node that are attached at the end of the branches and represent possible outcomes for each action. Decision trees seek to find the best split to subset the data which are then trained through the classification and regression tree (CART) algorithm [40]. In the random forest, each tree is constructed independently and depends on a random vector sampled from the input data, with all the trees in the forest

having the same distribution. The predictions from the forests are averaged using bootstrap aggregation and random feature selection. The complete equations of RF algorithm has been detailed by Breiman [30]. In the present study, the RF model was constructed by several steps such as drawing bootstrap samples from the original data, selecting the splitting variable from a randomly selected subset of variables using the decrease of Gini impurity (DGI) criterion, growing a tree for each bootstrap data set, aggregating the predictions of all trees through averaging, computing the out-of-bat (OOB) error rate using the data not in the bootstrap sample, and repeating these steps until specified number of trees were obtained [41, 42].

M5 Rules (M5R): M5R is a rule-based ML method that generates a list of decisions for regression problems using separate-and-conquer algorithms [43]. The separate-and-conquer algorithm breakdowns a problem into multiple sub-problems of the same or related class, until these problems become simple enough to be solved directly. In this study, the M5R model was constructed as follows: a model tree was firstly build using the decision tree algorithm, which was then pruned to delete the branches that result error in the learning data using the CART algorithm [40, 44]. Subsequently, the partial regression tree (PART) algorithm [45] was applied to generate rules. In each rule, a heuristic was established based on the if-then rule and the best leaf was made in the rule [46]. These procedures were applied recursively until all the instances were included in the rules. An instance could be included by different rules at the same time. Finally, a full tree was built instead of partially explored tree and the best rule was selected at every pattern for each tree which was called M5 Rule [47].

Support vector machine (SVM): SVM is a ML method for both classification and regression based on the Vapnik-Chervonenkis theory, which is a computational learning theory related to the statistical learning theory [48]. The basic idea of support vector regression (SVR) is to map the input data to a high-dimensional space and then calculate the linear regression functions in high-dimensional feature space. By this way, the disorganized data points become linearly separable in high-dimensional space where the training error can be minimized in the linearly separable optimization process. The complete SVR equations have been detailed in published books [49, 50], therefore, the equation derivation process would not be repeated here and the final equation was directly wrote down as follows:

$$f(\mathrm{x}) = \sum_{i,j=1}^{n}(\alpha_i - \alpha_i^*)\langle\phi(x_i)\cdot\phi(x_j)\rangle + b \quad (2)$$

where b represents the offset of the regression function, α_i and α_i^* represent the Lagrange multipliers satisfying the constraint $0 \leq \alpha_i, \alpha_i^* \leq$ C. Only the non-zero α_i and α_i^* contribute to the final regression model, which are the so-called support vectors. C is the complexity parameter which is a constant determining the trade-off between the complexity of SVR model and the tolerance of errors. The ϕ(x) represents the mapping function which is often unknown in advance and is difficult to determine, therefore, the term $\langle\phi(x_i)\cdot\phi(x_j)\rangle$ is often transformed by a kernel function:

$$\mathrm{K}(x_i, x_j) = \langle\phi(x_i)\cdot\phi(x_j)\rangle \quad (3)$$

The kernel function returns the value of the dot product, which is represented using a symmetric square matrix termed as kernel matrix. The kernel function reorganizes the original non-linear input space to a higher dimensional feature space and then uses it as a new input set. By this way, the non-linear problem is transformed into a linear problem.

Different kernel functions such as linear, polynomial, radial basis function (RBF) and sigmoid can be used to include varying degrees of nonlinearity and flexibility in the model. So far, there is not uniform standard to determine which kernel will produce the most accurate SVR, which should be chosen based on the characteristics of the given data. In this study, the widely used polynomial kernels were used by adjusting the kernel parameters. During the iterative data training process, the sequential minimal optimization algorithm was used to minimize the error function [51].

2.3 Model validation

The overall dataset were randomly divided into two groups including training dataset (80%) and testing dataset (20%). The training dataset was used to develop the prediction model and the model were optimized using the 10-fold cross-validation method. The basic idea of cross-validation was to evenly split the training dataset into 10 folds, while the instances from 9 folds were used for training and the remaining 1 fold was used for validation. This process was repeated 10 times using a different fold for validation at each cycle [21]. The performance of the well-trained model was further validated using the testing dataset which was not used for model development. The effectiveness and predictability of the models were assessed using the coefficient of determination (R^2) and root mean squared error (RMSE).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(Y_i^{exp} - Y_i^{pred}\right)^2}{\sum_{i=1}^{n}\left(Y_i^{exp} - \overline{Y}^{exp}\right)^2} \tag{4}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i^{exp} - Y_i^{pred}\right)^2} \tag{5}$$

where Y_i^{exp} and Y_i^{pred} represent the hydrate formation temperature measured in the

experiments and predicted by the model, respectively. $\overline{Y}^{exp}$ represents the average value of the hydrate formation temperature measured in the experiments.

2.4 Relative importance of independent variables to the accuracy of models

It is interesting to reveal how each variable contributes to the accuracy of the model and which direction we need to focus on during the selection of amino acids. In order to get such information, each of the well-trained model was run for eight times again using the same model parameters. At each time, one independent variable was removed from the training dataset and then the RMSE of each running was quantified. The variable was identified as the most important one if the resulted RMSE was the largest after removing it from the model. The relative importance of all variables were characterized by dividing the corresponding RMSE with the largest RMSE and multiplying with 100. By this way, all the values were normalized to the range between 0 and 100. This method was employed as it could be consistently applied to all the models which allowed comparison of variables' contribution among different models.

3. Results and Discussion

3.1 Statistical distribution of variables

Fig. 1 shows the statistical distribution of the variables used in this study. The interquartile range (IQR) was used to characterize the data variability by dividing the dataset into the first quartile (Q_1), the median and the third quartile (Q_3). Results indicated that the molecular weight of the amino acids used as THI ranged from 75.1 to 174.2 g mol^{-1} while 75% of them was less than 120 g mol^{-1}. No outlier was observed, suggesting that all the

data was distributed between Q_1 - 1.5 IQR and Q_3 + 1.5 IQR. Majority of the isoelectric point ranged from 6.0 to 6.3, which meant that these amino acids solutions were electrically neutral at around pH 6. Two of the three outliers were identified at the isoelectric point of 9.5 and 10.8, respectively, which exceeded the range of Q_3 + 10 IQR. It demonstrated that these two samples would carry electrical charge unless the solution became strong alkaline. The hydropathy index is a number representing the hydrophobic or hydrophilic properties of the amino acid's sidechain. Higher hydropathy index represent strong hydrophobicity while lower values represent strong hydrophilicity. Previous studies had showed significant effect of this variable on the gas hydrate inhibition strength [52, 53], but a weak linear correlation was observed between the amino acids' hydropathy and the methane hydrate inhibition. Bavoh et al. [10] suggested that the corresponding correlation coefficient (R^2) highly depended on the hydropathy scale. For example, the R^2 equaled 0.4567 and 0.6651, respectively, when the hydropathy scale ranged from -1.5 to 2.0 and from -4.5 to 2.0. In the present study, a much wider scale (-4.5 ~ 4.2) of hydropathy was used, which should be able to better interpret the influence of this variable on the hydrate phase equilibrium (Fig. 1). Another attributable factor for the above observation might be that the intrinsic relationship between the hydropathy index and the hydrate inhibition strength of amino acids was highly non-linear. A simple linear correlation was insufficient to capture such relationship while the ML methods were suitable for addressing such non-linear issues.

Another variable used for ML model building was the concentration of amino acids. The lowest concentration was 1.0 wt% and the majority of the data was in the range between 5.0 wt% and 11.4 wt% (Fig. 1). It should be noted that both mol% and wt% were used to

quantify the concentration of amino acid in previous studies, but opposing inhibition impact might be observed considering both units [54]. In this study, the data from different literature was converted to the equivalent concentration in wt% as it was more frequently used in industrial applications [55]. According to Yousif [55], the normal concentration needed to avoid hydrate formation in the industry was about 24 wt%. From the data collected in this study, only two examples had the concentration higher than 20 wt%.

The fraction of CH_4 in the natural gas ranged from 93% to 100% while the fraction of C_2H_8 and C_3H_8 did not exceed 5% in the data collected (Fig. 1). The system pressure ranged from 1.2 to 10.4 MPa and most of them was distributed between 3.3 and 7.4 MPa. The hydrate formation temperature in the collected data was predominantly distributed in the region between 277.2 and 283.3 K with a median value of 280.6 K.

3.2 Parameter optimization for each model

Each model had its own parameters to be specified, which were optimized by empirical attempts and the trial-and-error method. For the M5R model, only one parameter (i.e., the minimum number of instances allowing a leaf node) need to be specified. Cross-validation results indicated that the best performance was obtained when this parameter was set as 6. For the RF model, five parameters may affect the estimation ability of the model including the maximum depth of the tree, the random number of seed to be used, the number of execution slots for constructing the ensemble, the number of features at each node, and the number of trees in the random forest. In this study, the first three parameters was set as default (i.e., unlimited, 1, and 1, respectively), while the other two parameters varied from 1 to 9 and from 20 to 2000, respectively, during the parameter tuning process. It was found

that a larger number of trees in the RF resulted in a higher R^2 when a small number of features ($N_{features} \leq 4$) were used, but this tendency was not observed when the number of trees exceeded 1000 or when the number of features exceeded 4 (Fig. 2a). For example, R^2 increased from 0.9291 to 0.9592 when only one feature was used if the number of trees increased from 20 to 1000. When 6 features were used, the relative deviation between the highest and lowest R^2 was only 0.6% when the number of trees varied from 20 to 2000. A more obvious tendency was observed that the changes in the number of features resulted in initial decrease followed by increase in the RMSE (Fig. 2b). For instance, the RMSE decreased by 44% when the number of features increased from 1 to 6, which subsequently increased by 12% when the number of features increased from 6 to 9. Overall, the results indicated that the optimal number of features and the number of trees were 6 and 100, respectively.

For the SVM model, two key parameters needed to be carefully tuned. One was the aforementioned complexity parameter C. The other one was the polynomial order of the kernel function. The cross-validation results during model training demonstrated that the polynomial kernel had a higher prediction ability than the linear kernel and RBF kernel in this study. The polynomial order ranging from 2 to 4 was set in this study and tuned according to the variance in the R^2 and RMSE. Results clearly showed the significant increase in the R^2 when the polynomial order increased from 2 to 3 (Fig. 3a), while the opposite trend was observed for the RMSE (Fig. 3b). For example, the highest R^2 and the lowest RMSE was 0.9833 and 0.6306, respectively, when the polynomial order equaled 2. By contrast, the lowest R^2 and the highest RMSE was 0.9964 and 0.2966, respectively, when the polynomial order increased to 3. Moreover, there was a “V-shape” evolution

tendency of the RMSE against the complexity parameters. However, the model performance in terms of R^2 and RMSE did not improve much when the polynomial order increased from 3 to 4. Accordingly, the optimal value of the polynomial order and the complexity parameter was 3 and 60, respectively.

3.3 Comparison among different models

Fig. 4 shows the experimental HFT and the HFT predicted by the ML models using their optimized parameters. Before discussing the performance of ML models, the MLR results were also plotted for comparison. As shown in Fig. 4a, the HFT was linearly correlated with the eight independent variables using the equation as follows: T = -0.2098 mw + 0.811 pI + 0.1903 hI - 8.5936 conc - 2.5202 CH_4% +2.5204 C_2H_6% + 2.5204 C_3H_8% + 16.127 P + 275.0311. It can be seen that the R^2 and RMSE in the MLR was 0.9328 and 1.2821, respectively. Although the R^2 was relatively good, the relative deviation between the predicted HFT and experimental HFT was large. As shown in Fig. 5a, the distribution of the relative deviation was highly scattered. About 38% of the predicted HFT had the 0.5 - 1% relative deviation against the experimental HFT. Majority of the HFT deviation ranged from -1.4 K to 0.7 K while the maximum deviation was as high as 2.8 K (Fig. 5b). Using such linear regression model to predict HFT would be misleading and might result in wrong decision-making for amino acid selection, because the degree of HFT decrease after adding low concentration of many amino acids such as glycine, proline, serine, alanine and arginine was less than 1 K [10]. Although the application of high concentration amino acids might result in larger degree of HFT decrease, it was confusion if the difference was originated from the role of amino acid or from the model error. Therefore, the

thermodynamic hydrate inhibition characteristics of amino acids could not be correctly identified by MLR model due to the large error.

Fig. 4b demonstrated that RF model was better than MLR, which resulted in about 5% increase in the R^2 and 29% decrease in the RMSE compared with MLR. However, the model error remained unsatisfied especially in the low temperature zone ($T < 280$ K). About 17% of the predicted HFT had 0.5 - 1% relative deviation against the experimental HFT (Fig. 5a), while the largest deviation reached -1.8 K (Fig. 5b). The application of M5R model further improved the accuracy of prediction. The R^2 increased to 0.9795 while the RMSE decreased to 0.6951 (Fig. 4c). The relative deviation of all the predicted results were less than 0.5% except one sample where the relative deviation was 0.54% (Fig. 5a). The main deviation of the predicted HFT ranged from -0.5 to 0.3 K (Fig. 5b).

The best performance was observed by using the SVM model. The R^2 reached 0.9980 which was 7%, 2% and 1% higher than that of the MLR, RF and M5R, respectively (Fig. 4d). More importantly, the RMSE was reduced to 0.2188, which was about 83%, 76% and 69% smaller that of the other three models, respectively. The relative deviation of all the predicted HFT were less than 0.2% except one sample where the relative deviation was 0.23% (Fig. 5a). The absolute deviations of the predicted HFT were predominantly distributed within the range of 0.1 - 0.2 K (Fig. 5b). This made it possible to accurately predict HFT with good generalization ability, because the above results of R^2 and RMSE were based on the independently test dataset which was not used for model training. Compared with the other models, one possible reason for the better performance of the SVM model was that this model was based on the structural risk minimization principle whereas other models were based on the empirical risk minimization principle [31]. Instead

of minimizing the absolute error in classical adaptation algorithm, the SVM minimized the Vapnik-Chervonenkis bounds which was proven to have lower probability of error even without knowing the underlying probability distribution [56].

Accordingly, the well trained SVM model could be used to predict the HFT once a new set of data of the input parameters were given. To evaluate the relationship between the molecular properties of amino acids and their inhibition strength on hydrate formation, sensitivity analysis was carried out by running the trained SVM model again but increasing each of the four physicochemical properties (i.e., MW, pI, hI and conc) of amino acids, one at a time, at a rate of about 10%. When one property was varied, the other three properties were kept constant at the average value of the dataset (Fig. 1). The hydrate inhibition strength of amino acids was estimated by the average depression temperature (ΔT) as follows:

$$\text{average } \Delta T = \frac{1}{m} \sum_{i=1}^{m} (HFT_{water} - HFT_{AA}) \quad (6)$$

where m was the number of data points in the hydrate phase equilibrium curve, HFT_{water} was the HFT in pure water which was obtained from experimental data [35], HFT_{AA} was the HFT in presence of amino acids predicted by SVM model. It should be noted that the HFT_{water} and HFT_{AA} should be obtained under the same pressure when calculating their difference. A positive ΔT suggested that the amino acids could effectively inhibit hydrate formation by shifting the HFT to a lower value. The larger the ΔT, the stronger inhibition ability of the amino acids. Otherwise the amino acids had no effects or even promoted hydrate formation if ΔT was zero or negative.

Results indicated that the thermodynamic inhibition ability of amino acids monotonically increased with the isoelectric point and the concentration but decreased with the molecular

weight (Fig. 6). A parabolic shape was found between the inhibition strength and the hydropathy index where the maximum average depression temperature was about 1.8 K. Overall, the hydropathy index was the least sensitive property as the change from -4.5 to 4.2 in this parameter only resulted in about 0.4 ~ 0.6 K difference in the average depression temperature. Concentration was the most significant property especially when it exceeded 15 wt% and it was able to promote the average depression temperature by up to 8 K within the concentration range of study. As suggested by Mannar et al. [34], the increase of hydrate inhibition strength with the amino acids' concentration might be associated with the solution density and refractive index changes at high concentration. Another attributable factor was that the H-bonding interaction between amino acids and water increased with the concentration which enhanced the disruption of water structure during hydrate formation and therefore increased the hydrate inhibition strength.

3.4 Relative importance of independent variables to the accuracy of models

The above results demonstrated that the combination of the eight selected independent variables were sufficient to correctly capture the evolution patterns of natural gas HFT in the presence of amino acids. Fig. 7 shows the relative importance of each independent variable on the accuracy of the HFT predicted by different models. It can be seen that the system pressure and the concentration of amino acids ranked the first two important variables. All the four models successfully captured such overall tendency. Compared with the system pressure, the relative importance of the concentration of amino acid was 67%, 39%, 54% and 47%, respectively, in the MLR, RF, M5R and SVM prediction. According to the best model (i.e., SVM) in this study, the third important variable was the gas

composition in the natural gas. For instance, the relative importance of the fraction of methane in the natural gas was about 20% of that of the system pressure (Fig. 7d). The contribution of the physicochemical properties of amino acid (i.e., hI, pI and MW) to the accuracy of the predicted HFT was relatively slight (Fig. 7d). The relative importance of these three variables were similar but ranged from 20% to 40% in the other three models, suggesting that these models overestimated the role of these intrinsic properties of the amino acids. The above results demonstrated that a slight variance in the system pressure and gas composition would result in obvious changes in the HFT. In order to effectively shift the hydrate phase boundary to the unstable regions of natural gas hydrate, it might be the priority of choice to consider increasing the amino acid concentration instead of changing the type of amino acid (e.g., change an amino acid with higher hydropathy index).

4. Conclusions

This study demonstrated that the equilibrium temperature of natural gas hydrate formation could be accurately predicted using the SVM model with the selected independent variables. The relative deviation of the predicted HFT with the experimental HFT was less than 0.2%, while the corresponding R^2 and RMSE was 0.9980 and 0.2188, respectively. Compared with conventional MLR, the application of SVM effectively reduced the RMSE by 83%. The SVM model proposed in this study served a useful tool to help decide which type of amino acid to use and what minimum concentration was required to assure the HFT could be shifted to below the pipeline temperature. For instance, once the gas composition in the pipeline changed, a new series of hydrate phase boundary curves under a series of amino acid concentration could be quickly predicted using the proposed model. The

minimum concentration could be characterized as the concentration corresponding to the hydrate phase boundary curve nearest to the pipeline temperature. If the pipeline temperature increased, the nearest hydrate phase boundary would change accordingly and a smaller minimum concentration could be easily determined. By this way, the minimum concentration could be updated in time as a quick response to the operational conditions changes in the pipeline. An intelligent gas hydrate inhibitor injection system can be developed in future works which can not only monitor the real-time changes of the pressure, temperature and gas composition in the gas pipeline but also automatically select and inject the most appropriate amino acids using the machine learning algorithms.

Acknowledgements

This study was financially supported by Shenzhen Science and Technology Program (No. GJHZ20200731095600002, No. JCYJ20210324140810027), Guangdong Basic and Applied Basic Research Foundation (No. 2022A1515011834), National Key Research and Development Program of China for Young Scientist (No. 2021YFF0502300), Key Special Project for Introduced Talents Team of Southern Marine Science and Engineering Guangdong Laboratory (Guangzhou) (No. GML2019ZD0403, No. GML2019ZD0401), and Stable Support Funding from the Science, Technology and Innovation Commission of Shenzhen Municipality (No. WDZC20200818183253001).

References

[1] A.K. Arya, R. Jain, S. Yadav, S. Bisht, S. Gautam, Recent trends in gas pipeline optimization, Materials Today: Proceedings 57 (2021) 1455.

[2] Q. Chen, C. Wu, L. Zuo, M. Mehrtash, Y. Wang, Y. Bu, R. Sadiq, Y. Cao, Multi-objective transient peak shaving optimization of a gas pipeline system under demand uncertainty, Comput Chem Eng 147 (2021) 107260.

[3] D. Sloan, J. Creek, A.K. Sum, in: D. Sloan, C. Koh, A. K. Sum, A. L. Ballard, J. Creek, M. Eaton, J. Lachance, N. McMullen, T. Palermo, G. Shoup, L. Talley (Eds.), Natural Gas Hydrates in Flow Assurance, Gulf Professional Publishing, Boston, 2011, p. 13-36.

[4] S.-W. Zhang, L.-Y. Shang, L. Zhou, Z.-B. Lv, Hydrate deposition model and flow assurance technology in gas-dominant pipeline transportation systems: a review, Energ Fuel 36 (2022) 1747.

[5] S. Brustad, K.-P. Løken, J.G. Waalmann, Offshore technology conference, OnePetro, 2005.

[6] S.-P. Kang, J.-Y. Shin, J.-S. Lim, S. Lee, Experimental measurement of the induction time of natural gas Hydrate and its prediction with polymeric kinetic inhibitor, Chem Eng Sci 116 (2014) 817.

[7] L. Cheng, K. Liao, Z. Li, J. Cui, B. Liu, F. Li, G. Chen, C. Sun, The invalidation mechanism of kinetic hydrate inhibitors under high subcooling conditions, Chem Eng Sci 207 (2019) 305.

[8] W. Ke, T.M. Svartaas, J.T. Kvaløy, B.R. Kosberg, Inhibition–Promotion: Dual Effects of Polyvinylpyrrolidone (PVP) on Structure-II Hydrate Nucleation, Energ Fuel 30 (2016) 7646.

[9] G.A. Tabaaza, I.U. Haq, D.B. Zain, B. Lal, Toxicological issues of conventional gas hydrate inhibitors, Process Saf Prog 41 (2021) 5135.

[10] C.B. Bavoh, B. Lal, H. Osei, K.M. Sabil, H. Mukhtar, A review on the role of amino acids in gas hydrate inhibition, CO_2 capture and sequestration, and natural gas storage, J Nat Gas Sci Eng 64 (2019) 52.

[11] Q. Nasir, H. Suleman, Y.A. Elsheikh, A review on the role and impact of various additives as promoters/ inhibitors for gas hydrate formation, J Nat Gas Sci Eng 76 (2020) 103211.

[12] G. Bhattacharjee, P. Linga, Amino acids as kinetic promoters for gas hydrate applications: a mini review, Energ Fuel 35 (2021) 7553.

[13] G.-J. Chen, T.-M. Guo, Thermodynamic modeling of hydrate formation based on new concepts, Fluid Phase Equilibr 122 (1996) 43.

[14] G.-J. Chen, T.-M. Guo, A new approach to gas hydrate modelling, Chem Eng J 71 (1998) 145.

[15] C.-Y. Sun, G.-J. Chen, Modelling the hydrate formation condition for sour gas and mixtures, Chem Eng Sci 60 (2005) 4879.

[16] A. Eslamimanesh, A.H. Mohammadi, D. Richon, Thermodynamic model for predicting phase equilibria of simple clathrate hydrates of refrigerants, Chem Eng Sci 66 (2011) 5439.

[17] G. Moradi, E. Khosravani, Application of PRSV2 equation of state to predict hydrate formation temperature in the presence of inhibitors, Fluid Phase Equilibr 333 (2012) 18.

[18] K. Nasrifar, M. Moshfeghian, A model for prediction of gas hydrate formation conditions in aqueous solutions containing electrolytes and/or alcohol, J Chem Thermodyn 33 (2001) 999.

[19] C. Li, Twin support vector regression for prediction of natural gas hydrate formation conditions, Ind Eng Chem Res 60 (2021) 18519.

[20] M. Mesbah, E. Soroush, M. Rezakazemi, Development of a least squares support vector machine model for prediction of natural gas hydrate formation temperature, Chinese J Chem Eng 25 (2017) 1238.

[21] G. Wu, C. Kechavarzi, X. Li, S. Wu, S.J. Pollard, H. Sui, F. Coulon, Machine learning models for predicting PAHs bioavailability in compost amended soils, Chem Eng J 223 (2013) 747.

[22] M. Zare, S. Zendehboudi, M.A. Abdi, Deterministic tools to estimate induction time for methane hydrate formation in the presence of Luvicap 55 W solutions, J Mol Liq 348 (2022) 118374.

[23] S. Kim, M. Lee, K. Lee, T. Ahn, J. Lee, Data-driven estimation of three-phase saturation during gas hydrate depressurization using CT images, J Petrol Sci Eng 205 (2021) 108916.

[24] S. Kim, K. Lee, M. Lee, J. Lee, T. Ahn, J.-T. Lim, Evaluation of saturation changes during gas hydrate dissociation core experiment using deep learning with data augmentation, J Petrol Sci Eng 209 (2022) 109820.

[25] H. Xu, Z. Jiao, Z. Zhang, M. Huffman, Q. Wang, Prediction of methane hydrate formation conditions in salt water using machine learning algorithms, Comput Chem Eng 151 (2021) 107358.

[26] M. Mehrizadeh, Prediction of gas hydrate formation using empirical equations and data-driven models, Materials Today: Proceedings 42 (2021) 1592.

[27] H. Yarveicy, M.M. Ghiasi, Modeling of gas hydrate phase equilibria: Extremely randomized trees and LSSVM approaches, J Mol Liq 243 (2017) 533.

[28] A. Kamari, A. Bahadori, A.H. Mohammadi, S. Zendehboudi, New tools predict monoethylene glycol injection rate for natural gas hydrate inhibition, J Loss Prevent Proc 33 (2015) 222.

[29] M.M. Ghiasi, A.H. Mohammadi, S. Zendehboudi, Modeling stability conditions of methane Clathrate hydrate in ionic liquid aqueous solutions, 325 (2021) 114804.

[30] L. Breiman, Random forests, Mach Learn 45 (2001) 5.

[31] A. Ukil, Intelligent Systems and Signal Processing in Power Engineering. Springer, 2007.

[32] J.H. Sa, G.H. Kwak, K. Han, D. Ahn, S.J. Cho, J.D. Lee, K.H. Lee, Inhibition of methane and natural gas hydrate formation by altering the structure of water with amino acids, Sci Rep 6 (2016) 31582.

[33] C.B. Bavoh, B. Partoon, B. Lal, L. Kok Keong, Methane hydrate-liquid-vapour-equilibrium phase condition measurements in the presence of natural amino acids, J Nat Gas Sci Eng 37 (2017) 425.

[34] N. Mannar, C.B. Bavoh, A.H. Baharudin, B. Lal, N.B. Mellon, Thermophysical properties of aqueous lysine and its inhibition influence on methane and carbon dioxide hydrate phase boundary condition, Fluid Phase Equilibr 454 (2017) 57.

[35] C.B. Bavoh, M.S. Khan, B. Lal, N.I. Bt Abdul Ghaniri, K.M. Sabil, New methane hydrate phase boundary data in the presence of aqueous amino acids, Fluid Phase Equilibr 478 (2018) 129.

[36] C.B. Bavoh, O. Nashed, M.S. Khan, B. Partoon, B. Lal, A.M. Sharif, The impact of amino acids on methane hydrate phase boundary and formation kinetics, J Chem Thermodyn 117 (2018) 48.

[37] H.X. Liu, R.S. Zhang, X.J. Yao, M.C. Liu, Z.D. Hu, B.T. Fan, Prediction of the isoelectric point of an amino acid based on GA-PLS and SVMs, J Chem Inf Comput Sci 44 (2004) 161.

[38] R. Gholami, N. Fakhari, in: P. Samui, S. Sekhar, V.E. Balas (Eds.), Handbook of Neural Computation, Academic Press, 2017, p. 515-535.

[39] I.A. Basheer, M. Hajmeer, Artificial neural networks: fundamentals, computing, design, and application, J Microbiol Meth 43 (2000) 3.

[40] W.Y. Loh, Classification and regression trees, WIREs Data Mining Knowl Discov 1 (2011) 14.

[41] X. Chen, H. Ishwaran, Random forests for genomic data analysis, Genomics 99 (2012) 323.

[42] A.L. Boulesteix, S. Janitza, J. Kruppa, I.R. König, Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics, WIREs Data Mining Knowl Discov 2 (2012) 493.

[43] J. Fürnkranz, Separate-and-conquer rule learning, Artif Intell Rev 13 (1999) 3.

[44] J.R. Quinlan, Simplifying decision trees, International Journal of Man-Machine Studies 27 (1987) 221.

[45] E. Frank, I.H. Witten, Proceeding of the Fifteenth International Conference on Machine Learning, Morgan Kaufmann, 1998, p. 144-151.

[46] Q. Fang, H. Nguyen, X.-N. Bui, T. Nguyen-Thoi, Prediction of blast-induced ground vibration in open-pit mines using a new technique based on imperialist competitive algorithm and M5Rules, Nat Resour Res 29 (2020) 791.

[47] Y. Ayaz, A.F. Kocamaz, M.B. Karakoç, Modeling of compressive strength and UPV of high-volume mineral-admixtured concrete using rule-based M5 rule and tree model M5P classifiers, Constr Build Mater 94 (2015) 235.

[48] V.N. Vapnik, The nature of statistical learning theory. Springer, 1995.

[49] N. Cristianini, J. Shawe-Taylor, An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press, 2000.

[50] B. Schölkopf, A.J. Smola, F. Bach, Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press, 2002.

[51] G.W. Flake, S. Lawrence, Efficient SVM regression training with SMO, Mach Learn 46 (2002) 271.

[52] J.H. Sa, G.H. Kwak, K. Han, D. Ahn, K.H. Lee, Gas hydrate inhibition by perturbation of liquid water structure, Sci Rep 5 (2015) 11526.

[53] J.H. Sa, B.R. Lee, D.H. Park, K. Han, H.D. Chun, K.H. Lee, Amino acids as natural inhibitors for hydrate formation in CO_2 sequestration, Environ Sci Technol 45 (2011) 5885.

[54] D. Mech, G. Pandey, J.S. Sangwai, Effect of Molecular Weight of Polyethylene Glycol on the Equilibrium Dissociation Pressures of Methane Hydrate System, J Chem Eng Data 60 (2015) 1878.

[55] M.H. Yousif, Effect of underinhibition with methanol and ethylene glycol on the hydrate-control process, SPE Prod Facil 13 (1998) 184.

[56] V.N. Vapnik, A.Y. Chervonenkis, in: V. Vovk, H. Papadopoulos, A. Gammerman (Ed.) Measures of Complexity, Springer, 2015, p. 11-30.

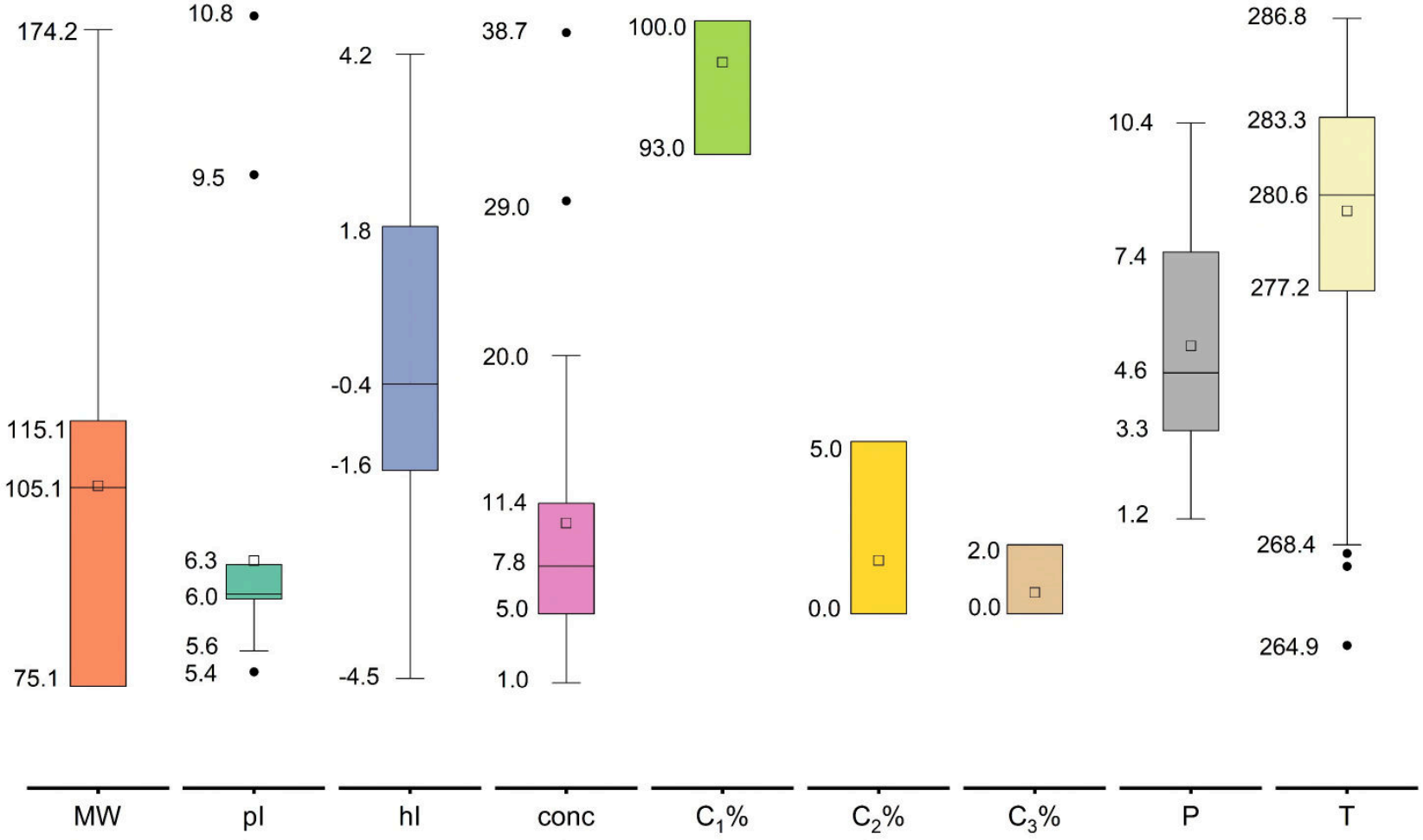

Fig. 1 Boxplot depicting the statistical distribution of the variables used in this study

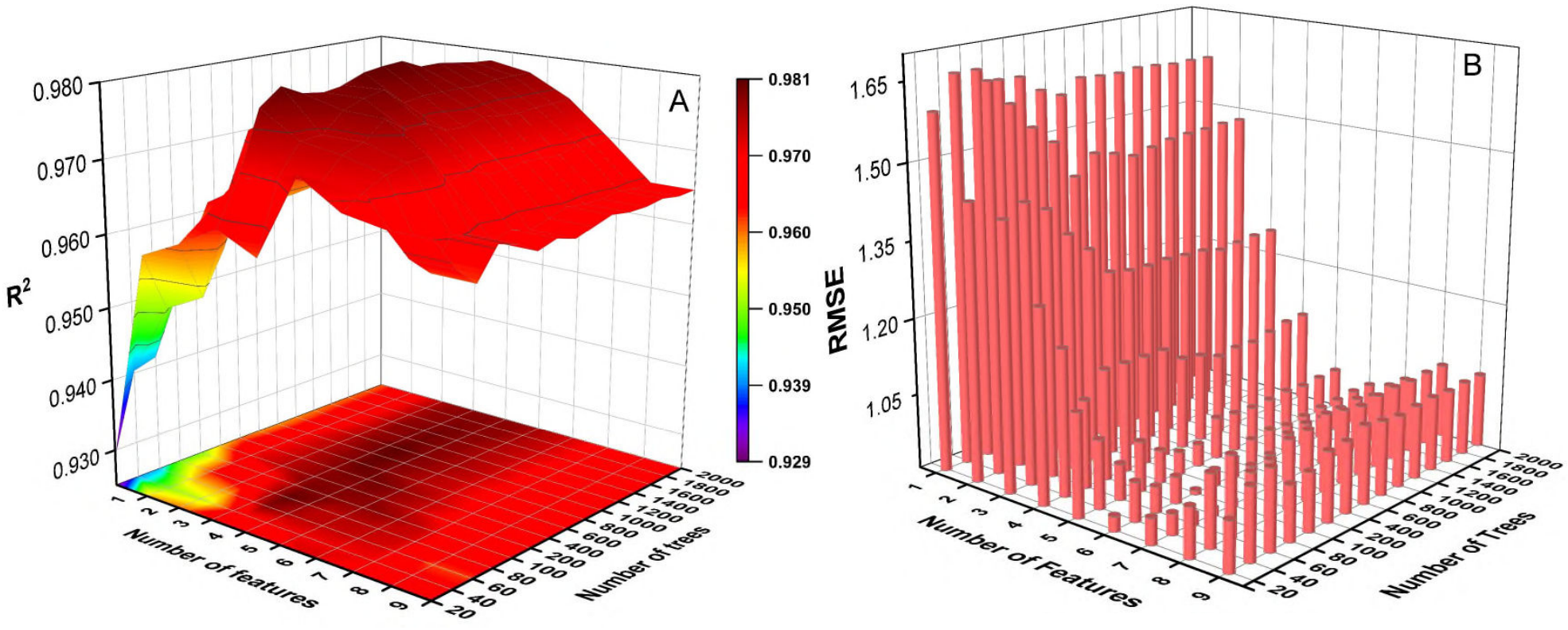

Fig. 2 Variance of R^2 and RMSE with the RF model parameters (optimal value: number of features = 6, number of trees = 100)

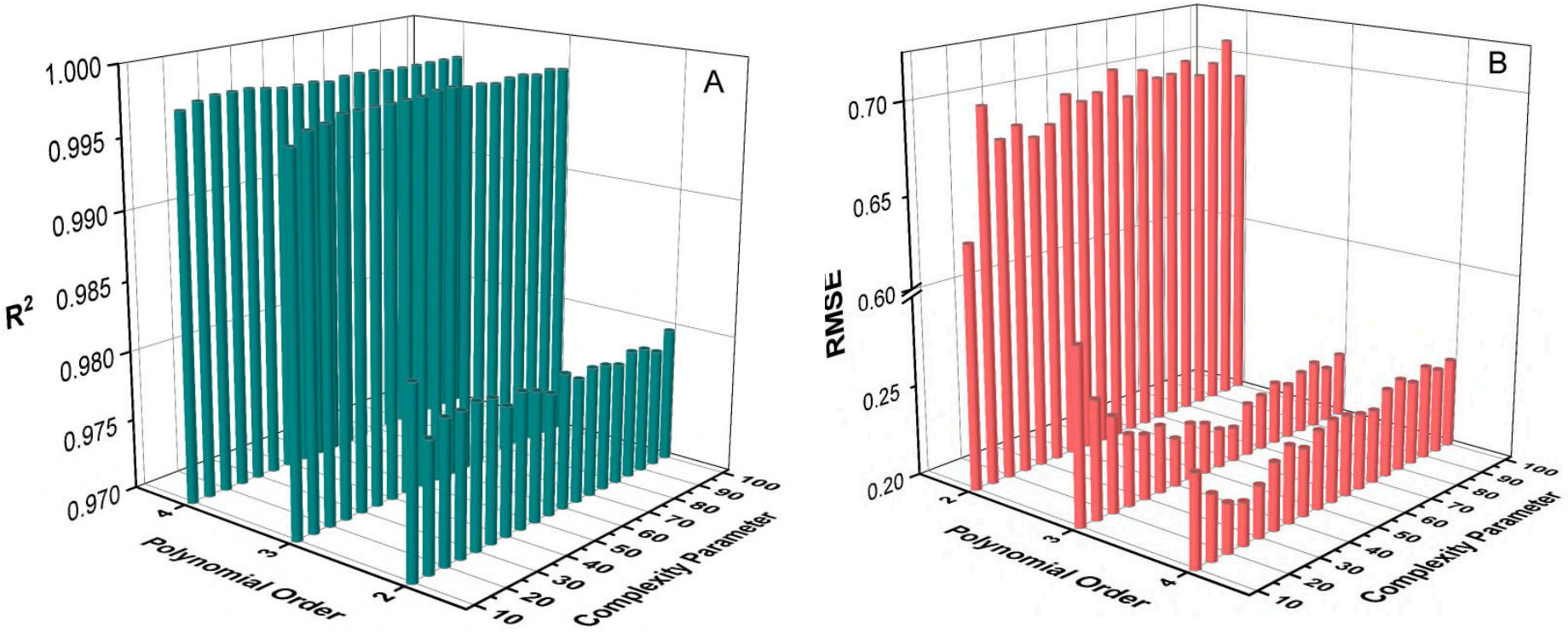

Fig. 3 Variance of R^2 and RMSE with the SVM model parameters (optimal value: polynomial order = 3, complexity parameter = 60)

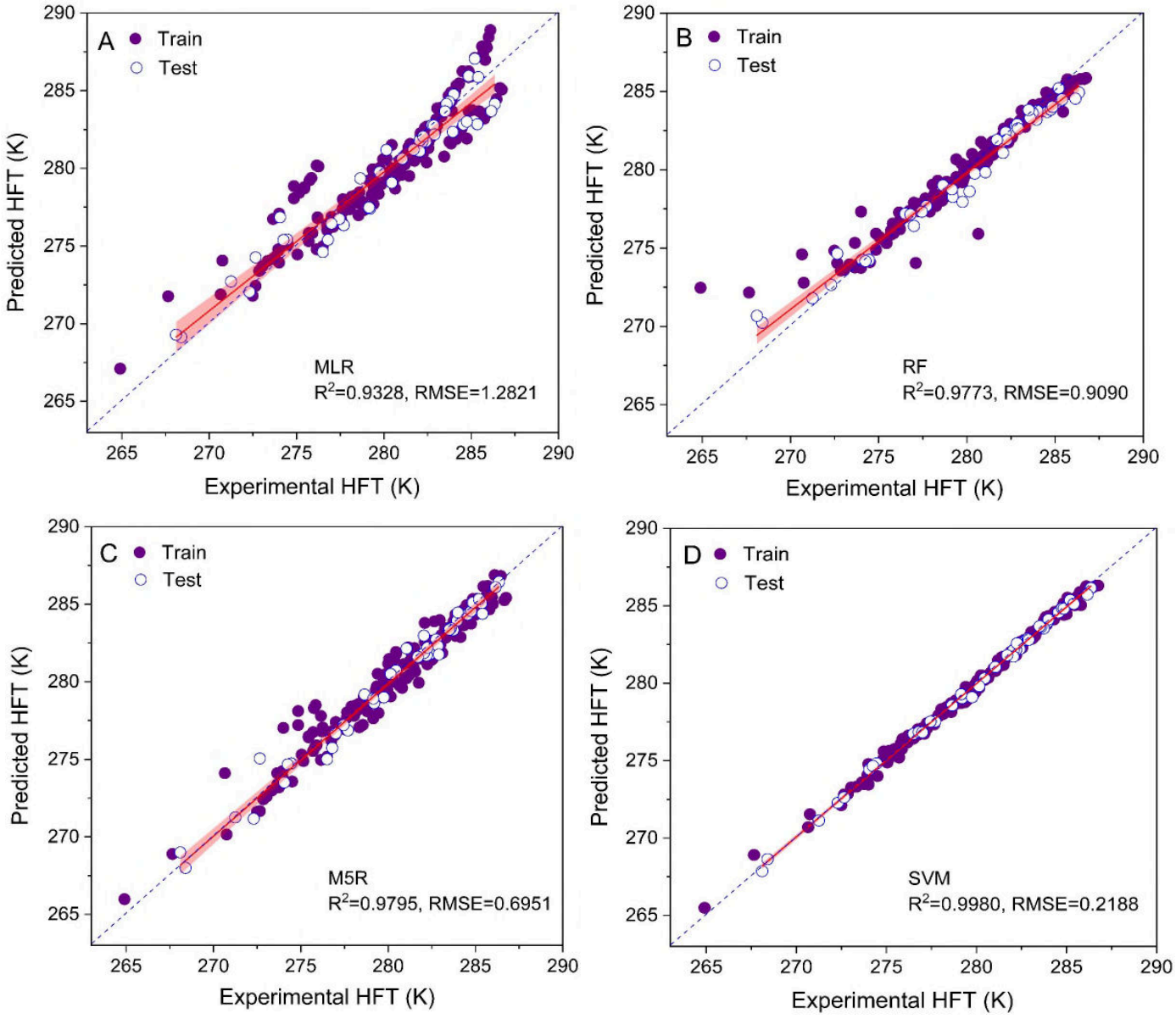

Fig. 4 Comparison between the predicted and experimental hydrate formation temperature. The red line represents the fit line of the test data and the shaded area indicates 95% confidence interval. The R and RMSE were calculated based on the test dataset which was not used in model training.

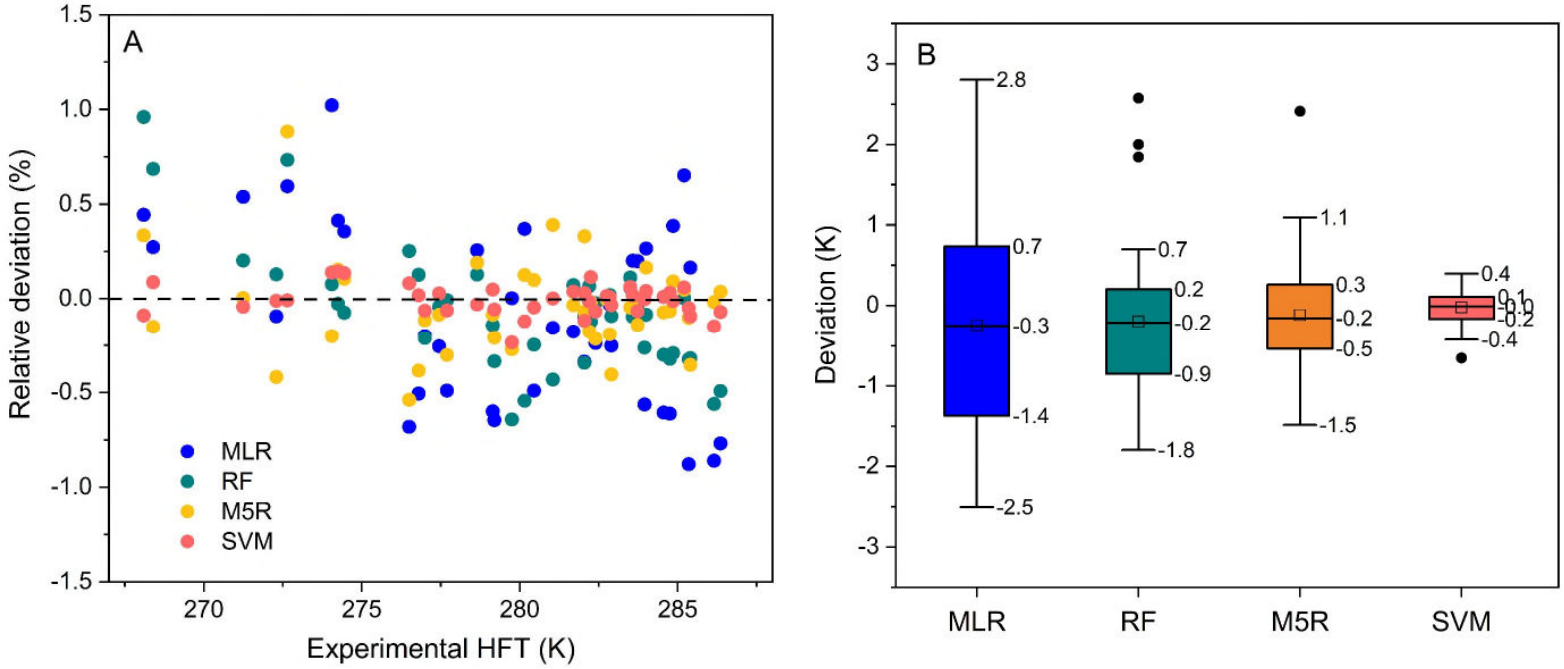

Fig. 5 Deviation and relative deviations of the predicted HFT with experimental HFT. The values were calculated based on the test dataset which was not used in model training.

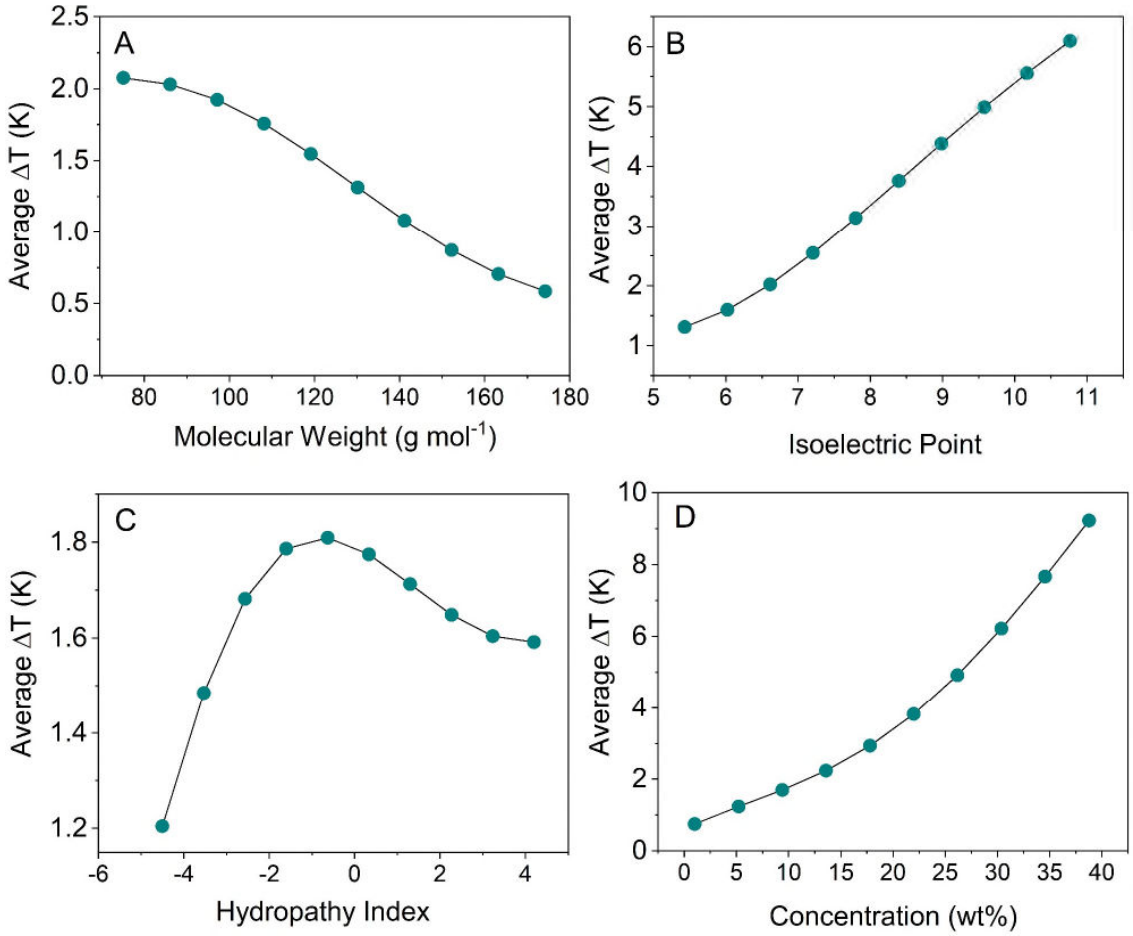

Fig. 6 Relationship between the molecular properties of amino acids and their inhibition strength on methane hydrate formation

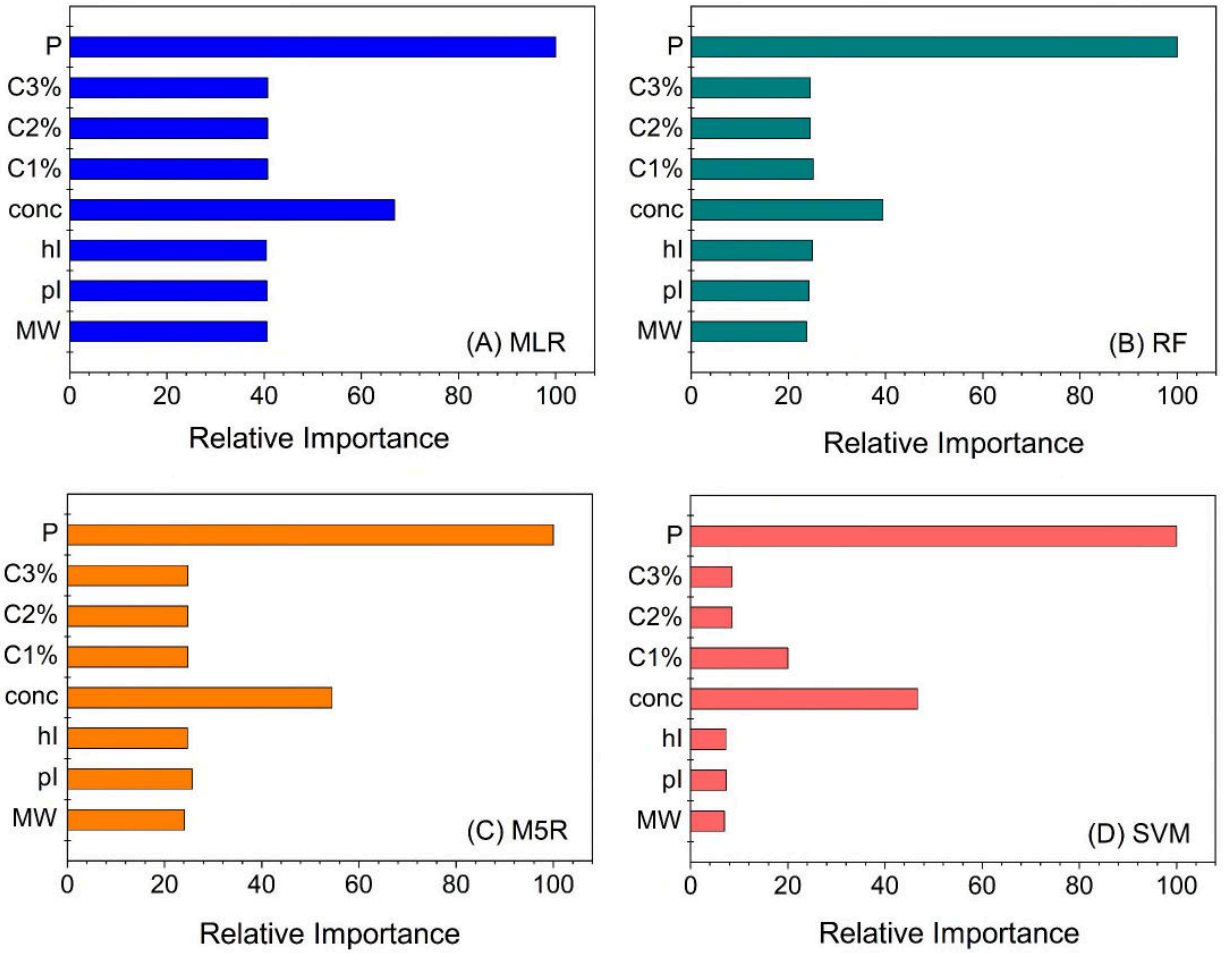

Fig. 7 Relative importance of independent variables to the accuracy of MLR and ML models

Machine learning models for fast selection of amino acids as green thermodynamic inhibitors for natural gas hydrate

Wu, Guozhong

2022-12-13

Wu G, Coulon F, Feng JC, et al., (2023) Machine learning models for fast selection of amino acids as green thermodynamic inhibitors for natural gas hydrate, Journal of Molecular Liquids. Volume 370, January 2023, Article number 120952

https://doi.org/10.1016/j.molliq.2022.120952