Structure-Based Reaction Descriptors for Predicting Rate Constants by Machine Learning: Application to Hydrogen Abstraction from Alkanes by CH_3/H/O Radicals

Yu Zhang[1,2,#], Jinhui Yu[2,3,#], Hongwei Song[2,*] and Minghui Yang[2,4,*]

1. College of Physical Science and Technology, Huazhong Normal University, Wuhan 430079, China

2. Key Laboratory of Magnetic Resonance in Biological Systems, State Key Laboratory of Magnetic Resonance and Atomic and Molecular Physics, National Center for Magnetic Resonance in Wuhan, Wuhan Institute of Physics and Mathematics, Innovation Academy for Precision Measurement Science and Technology, Chinese Academy of Sciences, Wuhan 430071, China

3. Key Laboratory of Plant Germplasm Enhancement and Specialty Agriculture, Wuhan Botanical Garden, Chinese Academy of Sciences, Wuhan 430074, China.

4. Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan 430074, China

[#]Y. Zhang and J. Yu contributed equally to this work.

[*]Corresponding authors: hwsong@wipm.ac.cn and yangmh@wipm.ac.cn

ABSTRACT

Accurate determination of reaction rate constants in the combustion circumstance is very challenging both experimentally and theoretically. In this work, three supervised machine learning algorithms, including XGB, FNN and XGB-FNN, are used to develop quantitative structure−property relationship models for the estimation of the rate constants of hydrogen abstraction reactions from alkanes by the free radicals CH_3, H and O. The molecular similarity based on Morgan molecular fingerprints combined with the topological indices are proposed to represent chemical reactions in the machine learning models. Using the newly constructed descriptors, the performance of each algorithm in prediction was found to be comparable and even superior to the corresponding one using the activation energy as a descriptor. The use of activation energy as a descriptor has previously been shown to significantly improve prediction accuracy (Fuel, **2022**, 322, 124150) but typically requires cumbersome ab initio calculations. The hybrid XGB-FNN algorithm performed better than the other two algorithms, which could reasonably predict reaction rate constants of hydrogen abstractions from different sites of alkanes and their isomers, indicating a good generalization ability. It is expected that the reaction descriptors proposed in this work can be applied to build machine leaning models for other reactions.

Keywords: rate constant, machine learning, combustion reaction, reaction descriptor

1. INTRODUCTION

Developing reliable combustion models is one of the main objects in the field of combustion chemistry.[1-4] By utilizing combustion models, a comprehensive understanding of the combustion process can be achieved, leading to improved efficiency and better control over pollutant emissions. The reliability of combustion model is closely related to the accuracy of kinetic parameters of associated chemical reactions, such as rate constants.[3, 5, 6] However, it is generally very difficult to yield accurate reaction rate constants in the combustion circumstance both experimentally and theoretically.

Recently, machine learning, lying at the core of artificial intelligence and data science, has emerged as a promising method to yield highly reliable reaction rate constants.[7-9] The machine learning methods can be in principle clarified into three categories: supervised learning, unsupervised learning and reinforcement learning, in which the supervised machine learning is usually applied in predicting chemical reaction properties by using different molecular representations as inputs.[8, 10] In this regard, many pioneering works have been performed to learn activation energies and minimum energy paths of chemical reactions.[7, 11-18] Meanwhile, some efforts have been paid to directly predict rate constants.[8]

For gas-phase bimolecular reactions, Houston et al. employed Gaussian Process Regression to train thermal rate constants using a dataset of 13 reactions over a large

temperature range.[19-21] The reactions were described by the three parameters related to Eckart tunneling, the skew angle and the symmetric stretch vibrational frequency of the reactant. The predicted rate constants averaged over the 39 test reactions were within 80% of the accurate answer. A deep neural network was employed by Komp et al. to train ~1.5 million quantum reaction rate constants from one-dimensional model potentials, with the reactant mass and the structure of barriers as input features.[22] The predicted logarithm of the rate product had a relative error of 1.1%. In the meanwhile, chemical reactions in liquid phase have also attracted much attention. Borhani et al. applied multiple-linear regression and artificial neural network methods to model the hydroxyl radical rate constants of water contaminants on a dataset of 457 water contaminants, using three-dimensional geometries of the molecular structures and quantum-chemically calculated descriptors as molecular descriptors.[23] The absolute relative error of the predicted logarithmic rate constants was less than 4%. Zhong et al. combined a convolutional neural network with molecular image to model radical rate constants of water contaminants, whose predictive performance was comparable to the molecular fingerprint-deep neural network model developed by the same group.[9, 24-26] Greaves et al. predicted rate constants for organic processes in mixtures containing ionic liquids by applying multiple linear regression and artificial neural networks with descriptors taken mostly from the Dragon descriptor data base.[27] To make the established machine learning models broadly available, Sanches-Neto et al. developed a web application "pySiRC" that predicts reaction rate constants of radical-based

oxidation processes of aqueous organic contaminants by combining three machine learning algorithms with molecular fingerprints.[28, 29]

Alkane (C_nH_{2n+2}) is an important component of various fuels, such as natural gas (mainly composed of methane), lighter fuel (e.g., n-butane), motor gasoline (consisting of various compounds of alkane and aliphatic hydrocarbon) and aviation fuel. In the case of aviation fuel, the combustion is largely proceeded by abstracting hydrogen atoms from different sites of alkanes by free radicals such as O, H, OH, HO_2, CH_3 and so on.[6, 30, 31] Accurate measurement of thermal rate constants of combustion reactions is very challenging and even infeasible at high temperatures. Compared to the vast number of reactions associated with the combustion process, there are only a few simple reactions whose rate constants have been measured. Theoretically, transition state theory (TST), Rice-Ramsperger-Kassel-Marcus theory and master equation approaches, could provide powerful and useful supplements. However, they may not be practical as the number of atoms involved in the reaction increases.[3, 32] To address this challenge, our group proposed using multilayered neural network models to predict rate constants of hydroxyl radical reactions with alkanes, with the reactions initially represented by three topological indices for the molecular structure of alkane and the index for the broken C-H bond.[33] Furthermore, we developed a novel hybrid machine learning model by combing feedforward neural network with eXtreme gradient boosting (XGB-FNN) to predict rate constants of reactions between alkanes and CH_3 radical.[34] The model employed six descriptors, including temperature, activation energy,

and four Mordred generated descriptors that were selected through Pearson correlation analysis. The average deviation of the XGB-FNN model on the prediction set was about 40%. All these studies have shown that machine leaning models are capable of accurately predicting rate constants of combustion reactions with well-designed molecular descriptors.

The activation energy of a chemical reaction looks like an "ideal" descriptor since it is closely related to the reaction rate. Previous work showed that using the activation energy as a descriptor can significantly improve the prediction accuracy of machine learning models.[34] However, there exist several difficulties in practical applications. Accurate determination of activation energy by quantum chemistry calculations has often been challenging for complex polyatomic reactions. In addition, if there exist multiple reaction pathways, it is difficult to choose the calculated activation energy used in the models. In this work, we intend to design new molecular descriptors to represent hydrogen abstraction reactions in combustion, in the condition that, on one hand, the descriptors don't require extra quantum chemistry calculations, and on the other hand, the prediction accuracy of machine learning methods based on the new descriptors archives the accuracy with the activation energy as a descriptor. Concretely speaking, the rate constants of hydrogen abstraction reactions between alkanes and radicals (CH_3, H and O) will be independently trained by combining different machine learning models with newly designed descriptors and then these models are applied to predict rate constants of combustion reactions with the number of involved carbon atoms in

alkanes up to 16. It is noteworthy that fuels with a carbon chain length of C8-C16 are the main constituent of aviation kerosene.[35]

2. METHODOLOGY

2.1 Data Collection

The prediction performance of machine learning models depends largely on the size and representativeness of the labeled dataset. In this study, the rate constants of site-specific H-abstraction reactions of alkanes by the radicals CH_3, H and O are collected from the National Institute of Standards and Technology (NIST) website (https://kinetics.nist.gov/kinetics/). Data cleaning follows the rules that experimental records take precedence over theoretical records and the recorded rate constants are averaged when there exist multiple experimental or theoretical records.

To enforce data self-consistency, the reserved rate constants for each reaction are fitted into a modified three-parameter Arrhenius equation:

$$k = AT^m exp(-E/RT), \quad (1)$$

in which *A* is the pre-exponential factor, *T* refers to the temperature, *m* is the additional temperature exponent coefficient, *E* is the exponential temperature coefficient and *R* is the universal gas constant. The fitted Arrhenius equation is then used to regenerate the dataset at intervals of 5 K in temperature, which increases the size of the dataset and helps models learn about the temperature dependence of rate constants. To prevent

over-fitting, the dataset is randomly divided into 90% as the training set and 10% as the validation set. The logarithms of rate constants are used in the training. Moreover, the reactions with only one or few recorded rate constants are devoted to assess the prediction accuracy.

2.2 Chemical Representation and Feature Selection

Molecular representation in machine-readable formats has becoming an important and active research field in computational chemistry, especially with the reviving of machine learning.[36, 37] The key to developing machine learning models is the designment of chemical representation, since no representation is perfect for every scenario. In this work, chemical reactions are represented by two types of descriptors: one describing different reactant alkanes and the other one characterizing different branches from the reaction.

To describe the reactant alkane, the open-source software Mordred[38] is applied to calculate the topological indices based on their SMILES representation as the input. The SMILES representations of alkanes are taken from the website of PubChem (https://pubchem.ncbi.nlm.nih.gov).

Chemical reactions definitely require more descriptors due to the possible existence of multiple reaction branches. The properties such as activation energy and reaction enthalpy may be appropriate, but are impractical due to time-consuming quantum chemical calculations. The reaction fingerprints, which are calculated by the difference

of the molecular fingerprints of reactants and products, have been successfully applied in reaction classification and similarity assessment.[39, 40] In the same spirit, the molecular similarity based on molecular fingerprints is used in this work as a reaction descriptor to represent different branches from a reaction. Molecular fingerprints are bit string representations of chemical structures originally designed for chemical database substructure searching and analysis.[41] The Morgan molecular fingerprints, which are widely utilized as binary features for the quantification of structural similarity of chemical compounds in two dimension,[41, 42] are employed to calculate the molecular similarity to describe H-abstraction reactions at different sites. The cosine similarity is used in this work and calculated by[43]

$$sim(A,B) = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}}, \quad (2)$$

where *A* and *B* are the values of the bits (0 or 1) for the reactant and product molecular fingerprints, respectively, and *i* refers to the bit.

Feature selection is usually necessary in machine learning, as it identifies the optimal relevant features and removes irrelevant ones.[44, 45] The imbalance between the dimensions of the feature space and the data points in the space can negatively affect the performance of machine learning models.[46] Ineffective descriptors with constant, reduplicative and/or missing values are first removed. The remaining descriptors are

then screened by the Pearson correlation analysis, implemented by the psych R package,[47] which guarantee a low correlation coefficient for the chosen descriptors.

2.3 Machine Learning Model Development and Evaluation

Three different machine learning algorithms are applied in this work, including feedforward neural networks (FNN), eXtreme gradient boosting (XGB) and hybrid FNN-XGB. Artificial neural network has been widely used in the field of nonlinear fitting. A simple type of acritical neural work is FNN, which are trained by error backpropagation. The information in FNN moves in only one direction, forward, from the input nodes to the output nodes. XGB is a tree-based ensemble machine learning algorithm,[48] which provides some advantages over the FNN, such as high tolerance, high efficiency and fast calculation speed.

The FNN model often converges slowly due to the error backpropagation algorithm, but a good weight initialization scheme can overcome this issue to some extent.[26, 49] To take advantages of both ANN and decision tree-based algorithms, a hybrid model, namely XGB-FNN, was designed in our previous work.[34] In the XGB-FNN model, a one-dimensional vector ($1 \times N$), representing the mean relative importance of descriptors, is generated by the trained XGB model. The sign ($\pm$) of the importance for each descriptor is determined by the correlation coefficient between the descriptor and the rate constants. Then, a new matrix ($M \times N$) is constructed by multiplying the one-dimensional vector of the mean relative importance by a randomly

generated Kaiming uniform distribution matrix ($M\times1$). The input weights of the FNN model are initialized by the newly constructed matrix. In the training, the XGB-FNN model uses the same hyperparameters as the FNN model.

The FNN model is constructed with an architecture of 1-3-1, *i. e.*, one input layer, three hidden layers, and one output layer, using the PyTorch framework (version 1.4.0).[50] The XGB model is built by XGBoost python package (version 0.80) under the scikit-learn framework (version 0.20.2).[51] The hyperparameters, such as the number of hidden neurons and the learning rate, have to be tested before training. Grid search is used to choose suitable hyperparameters.[22]

To assess the accuracy and robustness of the machine learning models, The Leave-One-Out (LOO) cross validations are firstly applied. The "one" left out in the cross validations denotes the rate constants of a reaction at different temperatures， which are taken as the test set. The rate constants of the remaining *n*-1 reactions are randomly divided into a training set and a validation set by a ratio of 9:1. Therefore, there are three datasets in the LOO cross validations: training, validation and test. For simplicity, the dataset group is labeled as "DG_{LOO}".

To build reliable prediction models, the generated data points are also divided into a training set and a validation set using the hierarchical random sampling by a ratio of 9:1, i.e., the ratio of the two datasets in each of the specified temperature interval keeps

unchanged. The reactions with only one or few raw rate constants are taken as the prediction set. The three datasets are labelled as "DG_{ALL}".

The mean square error loss function and Adam optimizer are used in the model training. The root mean square error (RMSE) is used to evaluate the performance of the developed models, which is defined as

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\log_{10}\frac{k_{pred}}{k_{obs}}\right)_i^2}, \quad (3)$$

where k_{obs} denotes the collected rate constant, k_{pred} stands for the predicted rate constant, i represents a data point in the dataset, N is the number of sampled rate constants in the dataset. Similar to Houston et al.'s definition,[19] the deviation δ is calculated by

$$\delta = 10^{\mathrm{RMSE}} - 1. \quad (4)$$

In this work, the predicted rate constant is thought to be "accurate" when the deviation is less than 100%, "reasonable" when the deviation is about 300%, and "inaccurate" when the deviation is larger than 500%. Houston et al. showed that the calculated rate constants by the TST methods were within 330% of the accurate answers on a set of 39 test reactions.[19] Therefore, the "reasonable" level is close to the accuracy of traditional TST methods.

The hyperparameters of these machine learning models are re-optimized by the grid search method. Finally, the so-called ensemble approach is used to diminish

random errors. In other words, a batch of NN models are trained under the same hyperparameters and the average of the three optimal models are employed to predict rate constants.

3. APPLICATIONS AND DISCUSSION

3.1 Description of Datasets, Input Features and Hyperparameters

In the dataset, 12 reactions are collected for the hydrogen abstraction of alkanes by CH_3 (labeled as CH_3Rs and listed in Table S1), in which two of them are taken as the prediction set in the DG_{ALL}. 12 reactions are included for the hydrogen abstraction by H (labeled as HRs and listed in Table S2) with three of them as the prediction set in the $DG_{ALL.}$ In addition, 20 reactions are collected for the hydrogen abstraction by O (labeled as ORs and listed in Table S3) with one of them as the prediction set in the DG_{ALL}. By sampling rate constants at every 5 K interval of temperature using the fitted Arrhenius equation, there are 3026 data points in the DG_{LOO} (2 more in the DG_{ALL} for prediction) of CH_3Rs, 3424 date points in the DG_{LOO} (18 more in the DG_{ALL}) of HRs, and 7079 data points in the DG_{LOO} (17 more in the DG_{ALL}) of ORs. The compositions of the dataset groups DG_{LOO} and DG_{ALL} for CH_3Rs, HRs and ORs are listed in Tables S4, S5 and S6, respectively.

Feature selection is necessary since the molecular descriptors generated by the open-source software Mordred are often highly correlated. Our previous work demonstrated that the number of six descriptors performed well in training the rate

constants of CH_3Rs.[34] In this work, the descriptors are selected using the pair-wise linear correlation analysis and the hierarchical clustering heat-map analysis.[52] Four descriptors are screened out from the Mordred generated descriptors to describe the reactants. Different from our previous work,[34] the molecular similarity, instead of the activation energy, is taken as a descriptor to represent different reaction branches. Together with temperature, there are totally six descriptors. Table 1 provides the selected descriptors for the three kinds of reactions, CH_3Rs, HRs and ORs, and their physical meanings.

Molecular similarity is calculated using the reactant and product molecular fingerprints. However, the value of each bit in fingerprint depends on the length of the bit string and the cut-off radius, resulting in different values of molecular similarity. It is thus necessary to choose a suitable bit string length and the cut-off radius to yield one-to-one molecular similarity for each branch reaction. The reactant alkanes with the number of involving carbon atoms up to 16 are interested in this work due to their importance in combustion reactions.[35] Therefore, all the alkanes and their isomers (less than 17 carbon atoms) with their SMILESs available from PubChem are collected and used to test the length and the cut-off radius. As a result, the length of the bit string is taken as 433 and the radius is taken as 8. Table 2 shows the calculated similarities for the reactions with the reactant alkanes involving less than 6 carbon atoms. Clearly, the different branches for the reactions can be well distinguished. A full list of generated similarities is given in Tables S7-S17.

The hyperparameters of the XGB and FNN models are selected using the grid search method on the DG_{LOO} and DG_{ALL}, respectively. Table S18 shows the optimal hyperparameters on the DG_{LOO} and Table S19 gives the optimal hyperparameters on the DG_{ALL}. Note that the hybrid XGB-FNN model uses the same hyperparameters as the FNN model for both the DG_{LOO} and DG_{ALL}.

3.2 Evaluation of Machine Learning Models

Figure 1 compares the deviations of the XGB, FNN and XGB-FNN models on the DG_{LOO} test sets with either the similarity or the activation energy as a descriptor. The activation energy-based results are taken from our previous work.[34, 53] The boxplot consists of the most extreme values in the data set (the green lines), the lower and upper quartiles (the black box), the median (the red line), the mean (the blue dashed line) and potential outliers (discrete red dots). The top three panels show the deviations with the similarity as the descriptor while the bottom three panels refer to previous results from the activation energy. Note that the reactions involving the reactant CH_4 are excluded from the LOO test dataset since the similarity between the reactant CH_4 and the product CH_3 is 0. For the reactions between alkanes and CH_3 (the first column), the deviations with the similarity as a descriptor are mostly distributed in between 75% and 225% for the three models, slightly smaller than those obtained using the activation energy as a descriptor, which are largely located in between 75% and 300%. For the reactions between alkanes and H (the middle column), the deviations with the similarity as a descriptor are mainly distributed in between 75% (75%, 75%) and 150% (225%, 200%)

for the XGB (FNN, XGB-FNN) model, which is larger than the deviations with the activation energy as a descriptor, which are mostly lower than 100%. For the reactions between alkanes and O (the third column), the deviations for the XGB model are largely smaller than 100% with the similarity as a descriptor while they are mostly in between 50% and 150% when using the activation energy as one of the descriptors. For the FNN and XGB-FNN models, most of the deviations with the similarity as a descriptor are slightly larger than those with the activation energy as a descriptor. Overall, all three machine learning models provide reasonable descriptions of the thermal rate constants and the performance of the activation energy is slightly better than that of the similarity.

The LOO cross validations presented above show that the machine learning combined with the similarity input provide a good alternative to predict thermal rate constants of combustion reactions. To improve the robustness of the models, the models are re-trained on the DG_{ALL} generated by the hierarchical random sampling. Three best models are selected and averaged for each algorithm to reduce random errors. The average deviations of the models on the DG_{ALL} are given in Figure 2. Note that the models using the activation energy as one of the descriptors are re-trained as well to make the comparison on an equal footing. Clearly, the three machine learning models using the activation energy as a descriptor all yield small deviations on the training and validation sets while produce relatively large deviations on the prediction set. In contrast, the FNN and XGB-FNN models with similarity as a descriptor generally provide comparable deviations across the three data sets. For the reactions between

alkanes and CH_3, the averaged deviations on the prediction set for the three models with the similarity as a descriptor are less than 60%, which is visibly smaller than those using the activation energy as a descriptor (about 90%). The FNN and XGB-FNN models with similarity as a descriptor also perform better than those with activation energy as a descriptor for the reaction between alkanes and H/O. However, the XGB model using the activation energy as a descriptor behaves better than that of similarity for the reactions between alkanes and O. In conclusion, the models developed in this work are "accurate" or close to "accurate".

The above assessments demonstrate that the performance of the three models using similarity as one of the descriptors is comparable to the corresponding activation energy-based models. However, in sharp contrast to the activation energy, the similarity can be easily generated from the reactant and product fingerprints and thus doesn't require extra quantum chemistry calculations, endowing great potential in practical applications. By comparing the three machine learning models that use similarity as a descriptor, the XGB-FNN model achieves a good balance between the accuracy and the stability in prediction. In addition, the predicted rate constant by the XGB model may exhibit discontinuous behavior with temperature, as shown in Fig. S1 for the reaction between n-C_8H_{18} and CH_3 as an example. By contrast, the predicted rate constant by the XGB-FNN model has a relatively smooth temperature dependence. Therefore, the rate constants will be predicted by the XGB-FNN model with similarity as a descriptor hereafter.

For the reactions listed in the DG_{ALL} prediction set, the rate constants are predicted in the temperature range from 300 to 2500 K for the reactions between alkanes and CH_3/H and from 300 to 2000 K for the reaction between 1-C_6H_{14} and O, as shown in Fig. 3. They are very close to the available observed values. Since the measured rate constants for prediction are mostly distributed in a very narrow temperature range, it is infeasible to assess the performance of the models over the temperature range of combustion. In this respect, we use the three-parameter Arrhenius equation to assess the performance based on the following considerations. On one hand, the training dataset is generated by the fitted three-parameter Arrhenius equation, which should be learned by the XGB-FNN model. On the other hand, the three reactions interested in this work are believed to follow the Arrhenius behavior in the temperature range of combustion due to the existence of well-defined barriers. Thus, the predicted rate constants for each reaction are fitted by the three-parameter Arrhenius equation. As expected, the predicted values closely align with the fitted curve, implying the good performance of the models over the temperature range of combustion.

3.3 Generalization Ability of the XGB-FNN Models

The alkanes containing up to 16 carbon atoms are regarded to be important in the combustion circumstance.[35] It is thus of great significance for developing accurate combustion models that could predict thermal rate constants of hydrogen abstraction reactions involving these reactants. Although Machine learning is widely perceived to have very limited extrapolation capabilities, this work built a machine learning model

for each kind of combustion reactions and the reactions in each kind should have similar reaction characteristics. If the designed descriptors take hold of the inherent features, we can expect good generalizability for machine learning models.

The rate constants of hydrogen abstraction reactions between alkanes and CH_3/H/O at different sites are predicted by the XGB-FNN models. The results are given in the supporting material (Figs. S2-S67). We will provide some predicted rate constants for the reactions between alkanes and CH_3 as examples to show the generalization abilities of the XGB-FNN models. Figure 4 displays the predicted rate constants of the primary hydrogen abstraction reaction with the normal alkanes involving 5-16 carbon atoms, namely n-C_5H_{12}, n-C_6H_{14}, n-C_7H_{16}, n-C_9H_{20}, n-$C_{10}H_{22}$, n-$C_{11}H_{24}$, n-$C_{12}H_{26}$, n-$C_{13}H_{28}$, n-$C_{14}H_{30}$, n-$C_{15}H_{32}$ and n-$C_{16}H_{34}$. The predicted rate constants of each reaction are further fitted by the three-parameter Arrhenius equation to assess the reliability of the prediction. It can be seen that the temperature dependence of rate constants is reasonably predicted by the XGB-FNN model. The predicted rate constants for each reaction follow well the Arrhenius equation. However, the accuracy of the predicted values cannot be definitely assessed without corresponding observed values for comparison.

The hydrogen atom can be abstracted from various sites of alkanes in the reactions. The reaction mechanism can be in-depth understood if one can determine the rate constants from different sites. Figure 5 shows the predicted rate constants of hydrogen

abstraction reactions at different sites of n-$C_{12}H_{26}$. The different sites of alkane are labeled as C_nH_{2n+2}-i, in which *i* is a positive integer representing the position of abstracted hydrogen atom from the end to the middle of C_nH_{2n+2}. The predicted rate constants from different sites of n-$C_{12}H_{26}$ are obviously distinct at low temperatures and the difference diminishes as the temperature rises up. In the temperature range from 300 to 1000 K, the predicted rate constants follow the order of n-$C_{12}H_{26}$-1 (the end) < n-$C_{12}H_{26}$-2 < n-$C_{12}H_{26}$-3 < n-$C_{12}H_{26}$-4 < n-$C_{12}H_{26}$-5 < n-$C_{12}H_{26}$-6 (the middle). The predicted thermal rate constants for other hydrogen abstraction reactions of alkanes with CH_3/H/O are carefully checked as well. We observed that the primary alkane site always has the lowest reactivity. The reactivity of the secondary site gradually diminishes from the last but one to the middle and becomes close to each other for large alkanes. It is widely recognized that the abstraction of secondary hydrogen atoms is easier than primary ones due to their lower bond dissociation energy.[6] Diego Troya investigated the barriers of hydrogen abstractions from primary, secondary, and tertiary sites of acyclic alkanes by ground-state oxygen atoms by high-level ab initio calculations.[54] The calculated thermal rate constants via the transition-state theory indicate that the room-temperature relative reactivities of primary, secondary, and tertiary alkane sites are 1, 29 and 422, following the order of primary < secondary < tertiary, in agreement with our findings. This gives us more confidence on the capability of generalization.

Figure 6 shows the predicted rate constants of the hydrogen abstraction reactions between CH_3 and several isomers of $C_{12}H_{26}$. The predicted rate constants increase monotonically with temperature. The reactivities for different $C_{12}H_{26}$ isomers are visibly different, indicating the influence of the molecular structure on the reactivity.

A web application is available at the following link - mlrate.apm.ac.cn. This application predicts the thermal rate constants of combustion reactions between alkanes and free radicals CH_3, H, and O by the constructed XGB-FNN models. Detailed instructions on how to use the web application to generate rate constants can be found on the homepage of the website.

4. CONCLUSION

In this work, three supervised machine learning algorithms, XGB, FNN and XGB-FNN, are employed to train and predict thermal rate constants of important combustion reactions between alkanes and free radicals (CH_3, H, O). The cosine similarity between the reactant alkane and the resulting alkyl radical combined with the screened topological indices of alkanes are taken as the reaction descriptors to feed the machine learning models. The hybrid XGB-FNN algorithm is found to perform better than the XGB and FNN algorithms on predicting thermal rate constants. The prediction accuracy of the similarity-based XGB-FNN model is comparable to that using the activation energy as a descriptor. Different from the activation energy, the similarity can be directly generated from the reactant and product fingerprints and doesn't require

extra quantum chemistry calculations. The XGB-FNN models have ability to predict thermal rate constants of hydrogen abstraction reactions from different sites of alkanes and their isomers. The newly designed reaction representations endow machine learning with great potential in developing reliable models to predict thermal rate constants. It's our hope that these reaction descriptors can be applied to develop other quantitative structure−property relationship models relevant to gas-phase chemical reactions.

Supporting Information

Collected reactions in the datasets, composition of different datasets, generated similarities, optimal hyperparameters and predicted rate constants. Data are provided with this paper and can be accessed from the file named "data".

Author contributions

M.Y. and H.S. conceived and supervised the research. All calculations were performed and analyzed by Y.Z. with assistance from J. Y. The manuscript was written by H. S. with input from Y.Z.

Conflicts of interest

There are no conflicts of interest to declare.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant Nos. 21973109 to H.S., 21973108 and 21921004 to M.Y.).

REFERENCES

(1) Smoot, L. D. A decade of combustion research. *Prog. Energy Combust. Sci.* **1997**, *23*, 203-232.

(2) Eaton, A. M.; Smoot, L. D.; Hill, S. C.; Eatough, C. N. Components, formulations, solutions, evaluation, and application of comprehensive combustion models. *Prog. Energy Combust. Sci.* **1999**, *25*, 387-436.

(3) Yuan, W.; Li, Y.; Qi, F. Challenges and perspectives of combustion chemistry research. *Sci. China Chem.* **2017**, *60*, 1391-1401.

(4) Kohse-Höinghaus, K. Combustion in the future: The importance of chemistry. *Proc. Combust. Inst.* **2021**, *38*, 1-56.

(5) Pilling, M. J. Interactions between theory and experiment in the investigation of elementary reactions of importance in combustion. *Chem. Soc. Rev.* **2008**, *37*, 676-685.

(6) Curran, H. J. Developing detailed chemical kinetic mechanisms for fuel combustion. *Proc. Combust. Inst.* **2019**, *37*, 57-81.

(7) Madzhidov, T. I.; Rakhimbekova, A.; Afonina, V. A.; Gimadiev, T. R.; Mukhametgaleev, R. N.; Nugmanov, R. I.; Baskin, I. I.; Varnek, A. Machine learning modelling of chemical reaction characteristics: yesterday, today, tomorrow. *Mendeleev Commun.* **2021**, *31*, 769-780.

(8) Komp, E.; Janulaitis, N.; Valleau, S. Progress towards machine learning reaction rate constants. *Phys. Chem. Chem. Phys.* **2022**, *24*, 2692-2705.

(9) Zhang, K.; Zhang, H. Machine Learning Modeling of Environmentally Relevant Chemical Reactions for Organic Compounds. *ACS ES&T Water* **2022**.

(10) Ihme, M.; Chung, W. T.; Mishra, A. A. Combustion machine learning: Principles, progress and prospects. *Prog. Energy Combust. Sci.* **2022**, *91*, 101010.

(11) Jorner, K.; Brinck, T.; Norrby, P.-O.; Buttar, D. Machine learning meets mechanistic modelling for accurate prediction of experimental activation energies. *Chem. Sci.* **2021**, *12*, 1163-1175.

(12) Heid, E.; Green, W. H. Machine Learning of Reaction Properties via Learned Representations of the Condensed Graph of Reaction. *J. Chem. Inf. Model.* **2022**, *62*, 2101-2110.

(13) Farrar, E. H. E.; Grayson, M. N. Machine learning and semi-empirical calculations: a synergistic approach to rapid, accurate, and mechanism-based reaction barrier prediction. *Chem. Sci.* **2022**, *13*, 7594-7603.

(14) Komp, E.; Valleau, S. Low-cost prediction of molecular and transition state partition functions via machine learning. *Chem. Sci.* **2022**, *13*, 7900-7906.

(15) Spiekermann, K. A.; Pattanaik, L.; Green, W. H. Fast Predictions of Reaction Barrier Heights: Toward Coupled-Cluster Accuracy. *J. Phys. Chem. A* **2022**, *126*, 3976-3986.

(16) Stuyver, T.; Coley, C. W. Quantum chemistry-augmented neural networks for reactivity prediction: Performance, generalizability, and explainability. *J. Chem. Phys.* **2022**, *156*, 084104.

(17) Garcia-Andrade, X.; Garcia Tahoces, P.; Perez-Rios, J.; Martinez Nunez, E. Barrier Height Prediction by Machine Learning Correction of Semiempirical Calculations. *J. Phys. Chem. A* **2023**, *127*, 2274-2283.

(18) Choi, S. Prediction of transition state structures of gas-phase chemical reactions via machine learning. *Nat. Commun.* **2023**, *14*, 1168.

(19) Houston, P. L.; Nandi, A.; Bowman, J. M. A Machine Learning Approach for Prediction of Rate Constants. *J. Phys. Chem. Lett.* **2019**, *10*, 5250-5258.

(20) Nandi, A.; Bowman, J. M.; Houston, P. A Machine Learning Approach for Rate Constants. II. Clustering, Training, and Predictions for the $O(^3P) + HCl \rightarrow OH + Cl$ Reaction. *J. Phys. Chem. A* **2020**, *124*, 5746-5755.

(21) Houston, P. L.; Nandi, A.; Bowman, J. M. A Machine Learning Approach for Rate Constants. III. Application to the $Cl(^2P) + CH_4 \rightarrow CH_3 + HCl$ Reaction. *J. Phys. Chem. A* **2022**, *126*, 5672-5679.

(22) Komp, E.; Valleau, S. Machine Learning Quantum Reaction Rate Constants. *J. Phys. Chem. A* **2020**, *124*, 8607-8613.

(23) Borhani, T. N. G.; Saniedanesh, M.; Bagheri, M.; Lim, J. S. QSPR prediction of the hydroxyl radical rate constant of water contaminants. *Water Res.* **2016**, *98*, 344-353.

(24) Zhong, S.; Hu, J.; Fan, X.; Yu, X.; Zhang, H. A deep neural network combined with molecular fingerprints (DNN-MF) to develop predictive models for hydroxyl radical rate constants of water contaminants. *J. Hazard. Mater.* **2020**, *383*, 121141.

(25) Zhong, S.; Hu, J.; Yu, X.; Zhang, H. Molecular image-convolutional neural network (CNN) assisted QSAR models for predicting contaminant reactivity toward OH radicals: Transfer learning, data augmentation and model interpretation. *Chem. Eng. J.* **2021**, *408*, 127998.

(26) Zhong, S.; Zhang, K.; Wang, D.; Zhang, H. Shedding light on "Black Box" machine learning models for predicting the reactivity of HO radicals toward organic compounds. *Chem. Eng. J.* **2021**, *405*, 126627.

(27) Greaves, T. L.; Schaffarczyk McHale, K. S.; Burkart-Radke, R. F.; Harper, J. B.; Le, T. C. Machine learning approaches to understand and predict rate constants for organic processes in mixtures containing ionic liquids. *Phys. Chem. Chem. Phys.* **2021**, *23*, 2742-2752.

(28) Sanches-Neto, F. O.; Dias-Silva, J. R.; Keng Queiroz Junior, L. H.; Carvalho-Silva, V. H. "pySiRC": Machine Learning Combined with Molecular Fingerprints to Predict the Reaction Rate Constant of the Radical-Based Oxidation Processes of Aqueous Organic Contaminants. *Environ. Sci. Technol.* **2021**, *55*, 12437-12448.

(29) Sanches-Neto, F. O.; Dias-Silva, J. R.; de Oliveira, V. M.; Aquilanti, V.; Carvalho-Silva, V. H. Evaluating and elucidating the reactivity of OH radicals with atmospheric organic pollutants: Reaction kinetics and mechanisms by machine learning. *Atmos. Environ.* **2022**, *275*, 119019.

(30) Simmie, J. M. Detailed chemical kinetic models for the combustion of hydrocarbon fuels. *Prog. Energy Combust. Sci.* **2003**, *29*, 599-634.

(31) Griffiths, J. F. Reduced kinetic models and their application to practical combustion systems. *Prog. Energy Combust. Sci.* **1995**, *21*, 25-107.

(32) Golden, D. M.; Barker, J. R. Pressure- and temperature-dependent combustion reactions. *Combust. Flame* **2011**, *158*, 602-617.

(33) Lu, J.; Zhang, H.; Yu, J.; Shan, D.; Qi, J.; Chen, J.; Song, H.; Yang, M. Predicting Rate Constants of Hydroxyl Radical Reactions with Alkanes Using Machine Learning. *J. Chem. Inf. Model.* **2021**, *61*, 4259-4265.

(34) Yu, J.; Shan, D.; Song, H.; Yang, M. A novel hybrid machine learning model for predicting rate constants of the reactions between alkane and CH_3 radical. *Fuel* **2022**, *322*, 124150.

(35) Blakey, S.; Rye, L.; Wilson, C. W. Aviation gas turbine alternative fuels: A review. *Proc. Combust. Inst.* **2011**, *33*, 2863-2885.

(36) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; 2000. DOI: 10.1002/9783527613106.

(37) Wigh, D. S.; Goodman, J. M.; Lapkin, A. A. A review of molecular representation in the age of machine learning. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2022**, *12*, e1603.

(38) Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. Mordred: a molecular descriptor calculator. *J. Cheminformatics* **2018**, *10*, 4.

(39) Patel, H.; Bodkin, M. J.; Chen, B.; Gillet, V. J. Knowledge-Based Approach to de Novo Design Using Reaction Vectors. *J. Chem. Inf. Model.* **2009**, *49*, 1163-1184.

(40) Schneider, N.; Lowe, D. M.; Sayle, R. A.; Landrum, G. A. Development of a Novel Fingerprint for Chemical Reactions and Its Application to Large-Scale Reaction Classification and Similarity. *J. Chem. Inf. Model.* **2015**, *55*, 39-53.

(41) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742-754.

(42) Kuwahara, H.; Gao, X. Analysis of the effects of related fingerprints on molecular similarity using an eigenvalue entropy approach. *J. Cheminformatics* **2021**, *13*, 27.

(43) Ye, J. Cosine similarity measures for intuitionistic fuzzy sets and their applications. *Math. Comput. Modell.* **2011**, *53*, 91-97.

(44) Cunningham, P. Dimension Reduction. In *Machine Learning Techniques for Multimedia*, Cord, M., Cunningham, P. Eds.; Cognitive Technologies, Springer Berlin Heidelberg, 2008; pp 91-112.

(45) Coelho, D.; Madureira, A.; Pereira, I.; Gonçalves, R. A Review on Dimensionality Reduction for Machine Learning. In *Innovations in Bio-Inspired Computing and Applications*, Cham, 2023; Abraham, A., Bajaj, A., Gandhi, N., Madureira, A. M., Kahraman, C., Eds.; Springer Nature Switzerland: Vol. 649, pp 287-296. DOI: https://doi.org/10.1007/978-3-031-27499-2_27.

(46) Jia, W.; Sun, M.; Lian, J.; Hou, S. Feature dimensionality reduction: a review. *Complex Intell. Syst.* **2022**, *8*, 2663-2693.

(47) Revelle, W.; Revelle, M. W. Package 'Psych'. The Comprehensive R Archive Network. 2015.

(48) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA; 2016.

(49) Wang, Y.; Liao, Z.; Mathieu, S.; Bin, F.; Tu, X. Prediction and evaluation of plasma arc reforming of naphthalene using a hybrid machine learning model. *J. Hazard. Mater.* **2021**, *404*, 123965.

(50) Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: an imperative style, high-performance deep learning library. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Curran Associates Inc., 2019; p Article 721.

(51) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(52) Kolde, R. Pheatmap: pretty heatmaps. **2012**.

(53) Yu, J. Machine Learning Research on Rate Constants of Reactions bewteen Alkane and Free Radial. University of Chinese Academy of Sciences (Innovation Academy for Precision Measurement Science and Technology), Wuhan, 2022.

(54) Troya, D. Barriers of Hydrogen Abstraction from Primary, Secondary, and Tertiary Alkane Sites by O(^{3}P). *J. Phys. Chem. A* **2007**, *111*, 10745-10753.

Table 1: Selected descriptors for the reactions CH_3Rs, HRs and ORs, and their physical meanings.

Descriptor	Physical meaning	CH_3Rs	HRs	ORs
T	Temperature	√	√	√
Similarity	Structural similarity between alkane and alkyl radical	√	√	√
SIC1	1-ordered structural information content	√		
IC1	1-ordered neighborhood information content	√	√	
AATSC2c	Averaged and centered moreau-broto autocorrelation of lag 2 weighted by gasteiger charge	√		
AATS2i	Averaged moreau-broto autocorrelation of lag 2 weighted by ionization potential	√		
BIC1	1-ordered bonding information content		√	
CIC2	2-ordered complementary information content		√	
ATSC1c	Centered moreau-broto autocorrelation of lag 1 weighted by gasteiger charge		√	
AATS2Z	Averaged moreau-broto autocorrelation of lag 2 weighted by atomic number			√
AATSC1c	Averaged and centered moreau-broto autocorrelation of lag 1 weighted by gasteiger charge			√
SIC3	3-ordered structural information content			√
VE2_A	VE2 of adjacency matrix			√

Table 2: Calculated similarities for the reactions with the reactant alkanes involving less than 6 carbon atoms.

Reactant chemical formula	Reactant SMILES	Product SMILES	Similarity
CH_4	C	[CH3]	0
C_2H_6	CC	C[CH2]	0.40825
C_3H_8	CCC	C[CH]C	0.28868
		CC[CH2]	0.51640
C_4H_{10}	CCCC	CCC[CH2]	0.56695
		CC[CH]C	0.37796
	CC(C)C	CC(C)[CH2]	0.61237
		C[C](C)C	0.5
	CCCCC	CC[CH]CC	0.33333
		CCCC[CH2]	0.68041
C_5H_{12}	CCC(C)C	CCC[CH]C	0.38730
		CC[C](C)C	0.28571
		CC(C)C[CH2]	0.53452
		CCC(C)[CH2]	0.62994
	CC(C)(C)C	C[CH]C(C)C	0.40089
		CC(C)(C)[CH2]	0.61237

Figure captions

Figure 1: Comparison of deviations of the XGB, FNN and XGB-FNN models with either the activation energy or the similarity as reaction descriptor on the DG_{LOO} test sets.

Figure 2: Comparison of average deviations of the XGB, FNN and XGB-FNN models with either the activation energy or the similarity as a descriptor on the DG_{ALL}.

Figure 3: Comparison of predicted thermal rate constants with available experimental values for the hydrogen abstraction reactions from alkanes by CH_3, H and O Radicals.

Figure 4: Predicted and fitted thermal rate constants for the primary hydrogen abstraction reactions between the CH_3 radical and (a) n-C_5H_{12}, (b) n-C_6H_{14}, (c) n-C_7H_{16}, (d) n-C_8H_{18}, (e) n-C_9H_{20}, (f) n-$C_{10}H_{22}$, (g) n-$C_{11}H_{24}$, (h) n-$C_{12}H_{26}$, (i) n-$C_{13}H_{28}$, (j) n-$C_{14}H_{30}$, (k) n-$C_{15}H_{32}$ and (l) n-$C_{16}H_{34}$.

Figure 5: Predicted thermal rate constants for hydrogen abstractions at different sites between n-$C_{12}H_{26}$ and CH_3. The number *i* near each circle denotes the position of abstracted hydrogen atom from the end to the middle of n-$C_{12}H_{26}$.

Figure 6: Predicted thermal rate constants for hydrogen abstraction reactions from different isomers of $C_{12}H_{26}$ by CH_3. The circles represent different abstraction sites for different isomers of $C_{12}H_{26}$.

Fig. 1

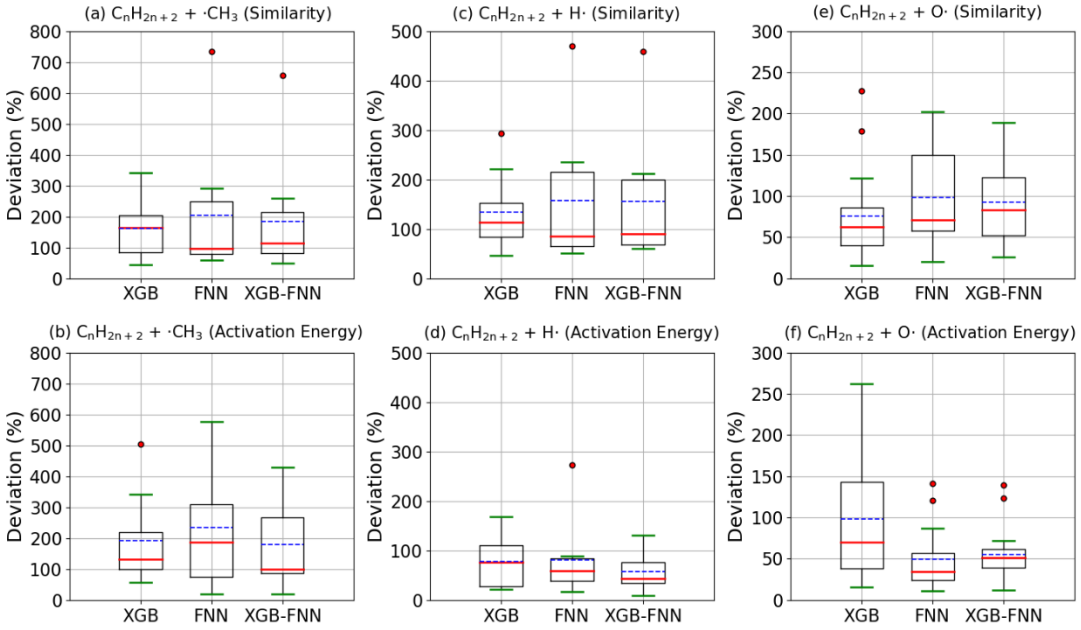
(a) C_nH_{2n+2} + ·CH_3 (Similarity)
(c) C_nH_{2n+2} + H· (Similarity)
(e) C_nH_{2n+2} + O· (Similarity)
(b) C_nH_{2n+2} + ·CH_3 (Activation Energy)
(d) C_nH_{2n+2} + H· (Activation Energy)
(f) C_nH_{2n+2} + O· (Activation Energy)
Deviation (%)
XGB
FNN
XGB-FNN

Fig. 2

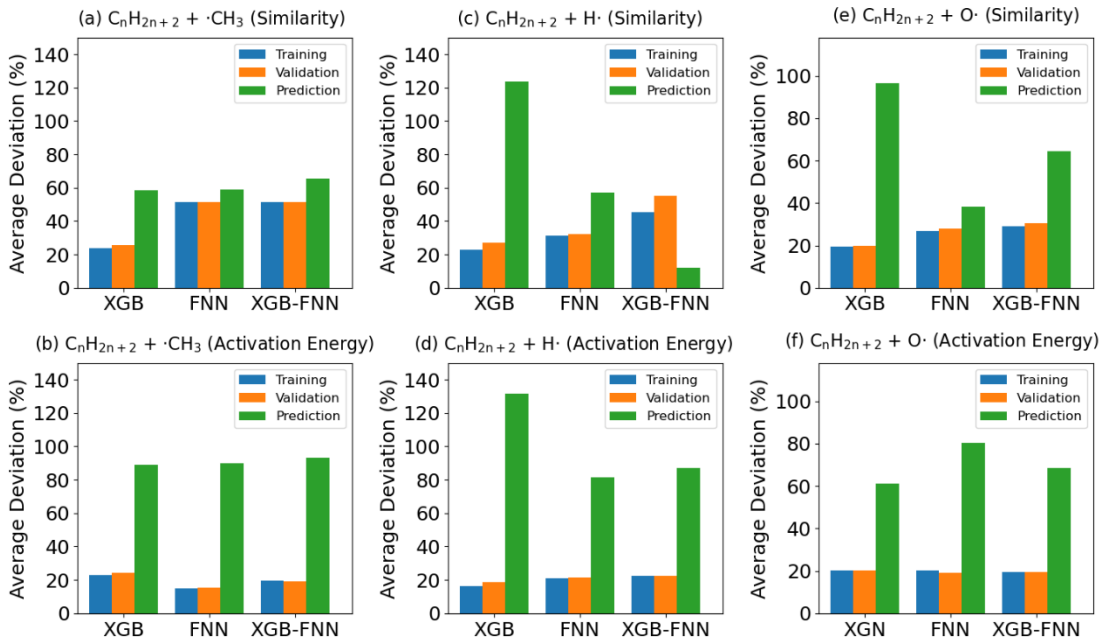
(a) C_nH_{2n+2} + ·CH_3 (Similarity)
(c) C_nH_{2n+2} + H· (Similarity)
(e) C_nH_{2n+2} + O· (Similarity)
(b) C_nH_{2n+2} + ·CH_3 (Activation Energy)
(d) C_nH_{2n+2} + H· (Activation Energy)
(f) C_nH_{2n+2} + O· (Activation Energy)
Average Deviation (%)
Training
Validation
Prediction
XGB
FNN
XGB-FNN
XGN

Fig. 3

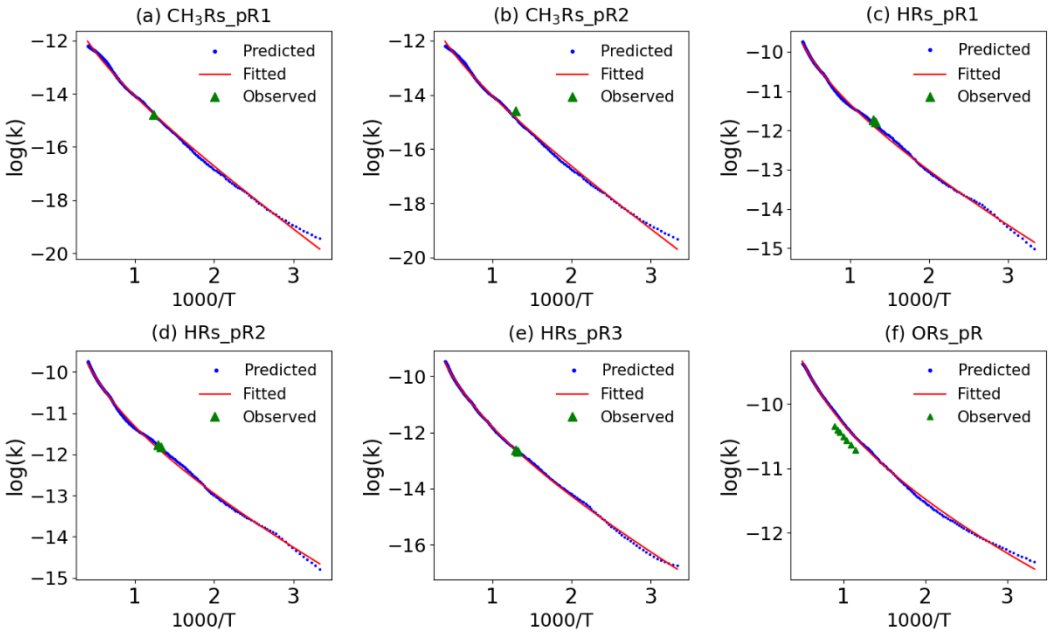
(a) CH3Rs_pR1
(b) CH3Rs_pR2
(c) HRs_pR1
(d) HRs_pR2
(e) HRs_pR3
(f) ORs_pR
Predicted
Fitted
Observed
log(k)
1000/T

Fig. 4

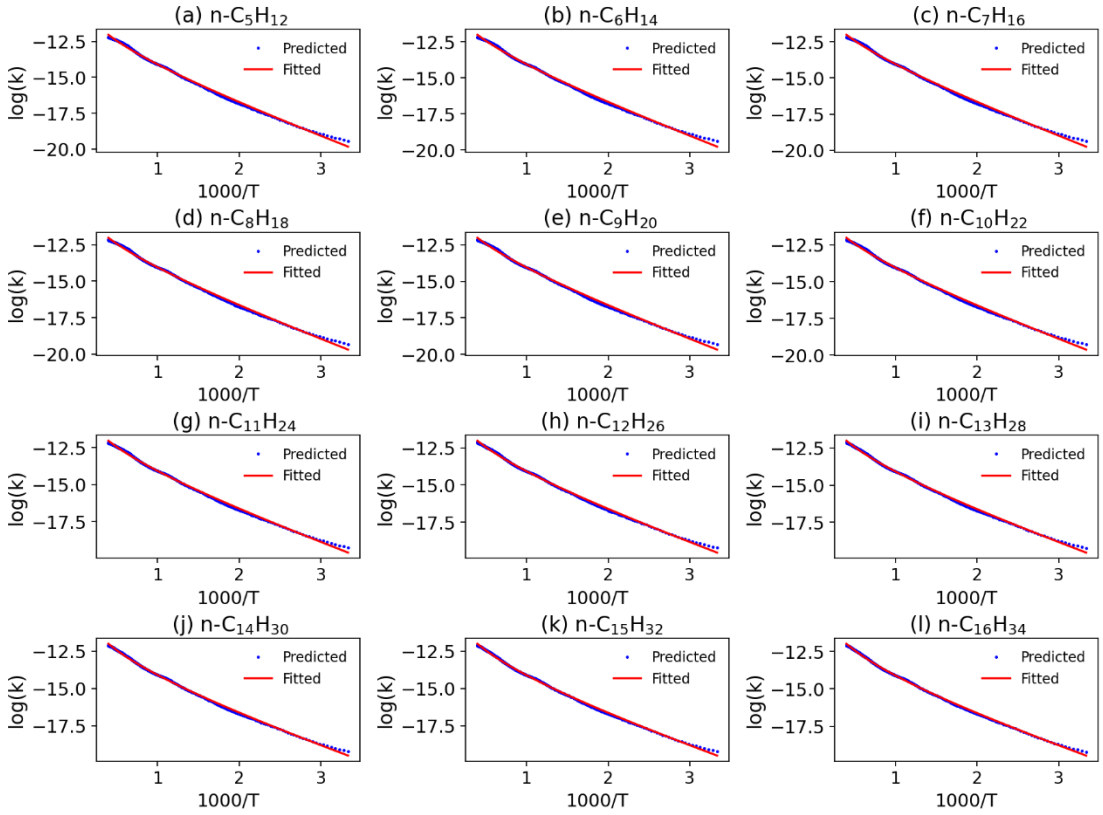
(a) n-C5H12
(b) n-C6H14
(c) n-C7H16
(d) n-C8H18
(e) n-C9H20
(f) n-C10H22
(g) n-C11H24
(h) n-C12H26
(i) n-C13H28
(j) n-C14H30
(k) n-C15H32
(l) n-C16H34
Predicted
Fitted
log(k)
1000/T

Fig. 5

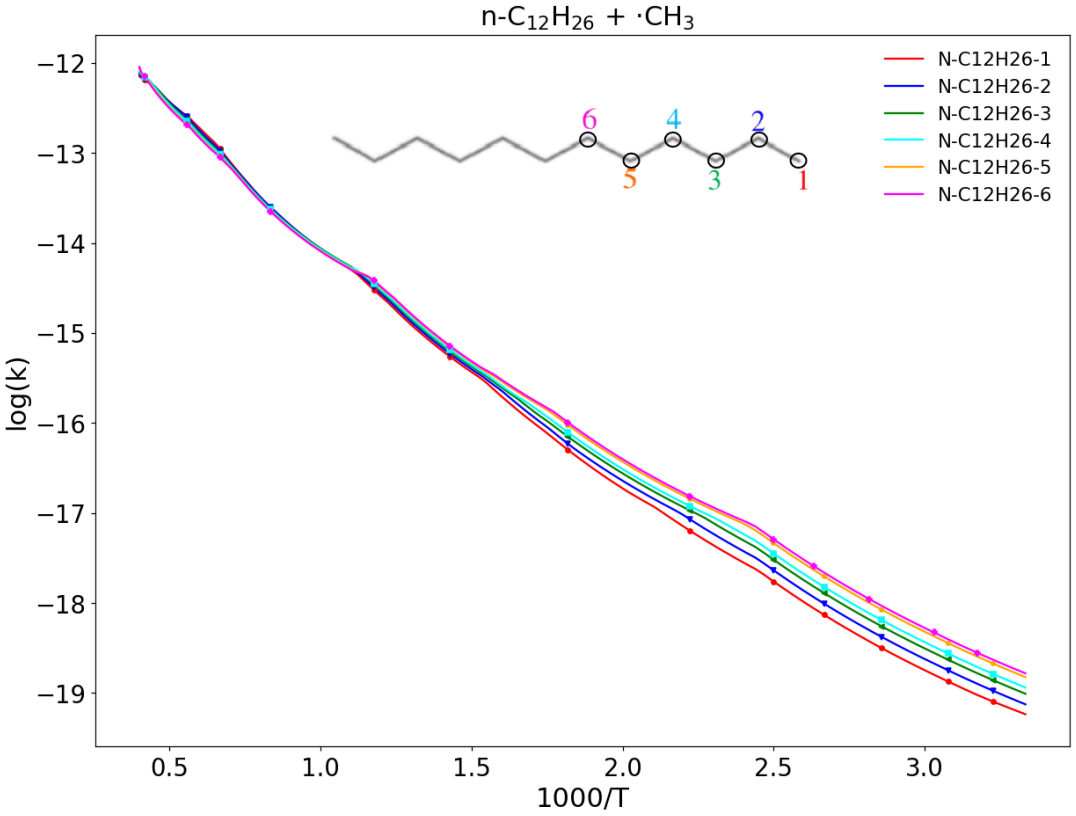
n-C12H26 + ·CH3
log(k)
1000/T
−12
−13
−14
−15
−16
−17
−18
−19
0.5
1.0
1.5
2.0
2.5
3.0
N-C12H26-1
N-C12H26-2
N-C12H26-3
N-C12H26-4
N-C12H26-5
N-C12H26-6
6
4
2
5
3
1

Fig. 6: