Progress in Reaction Kinetics and Mechanism

ARTICLE

Open Access

https://doi.org/10.48130/prkm-0025-0013

Progress in Reaction Kinetics and Mechanism **2025**, 50: e013

Prediction of the sooting tendency of fuel mixtures with quantitative structure-property relationship (QSPR) model based on artificial neural network approach

Haoyang Wang[1], Lei Zhu[2] and Liming Cai[1*]

[1] *School of Automotive Studies, Tongji University, Shanghai 201804, China*

[2] *Key Laboratory for Power Machinery and Engineering of M.O.E., Shanghai Jiao Tong University, Shanghai 200240, China*

* Corresponding author, E-mail: lcai@tongji.edu.cn

Abstract

The sooting characteristics of fuels are of high relevance for their application in advanced combustion devices. While practical petroleum fuels are often mixtures of various components, most literature studies have focused on the numerical and experimental investigation of the sooting behaviors of neat components, often in terms of the so-called yield sooting index (YSI). Recently, a YSI database was established for 21 neat fuel components and 151 mixtures of them. An artificial neural network (ANN) model was developed to predict the YSI of fuel mixtures with reasonable accuracy by only using the mass fractions of mixture components as input features. However, pioneering studies have demonstrated that if quantitative structure-property relationship (QSPR) is incorporated, the ANN models can match the measured physical and chemical properties of fuels very well and provide insights into the dependence of properties on fuel structures. Thus, in this study, in order to predict the YSI of fuel mixtures with improved accuracy and to understand the impact of functional groups on the sooting tendency of fuel blends, an artificial neural network (ANN)-based QSPR model is derived by taking functional group descriptors as input features, in conjunction with the mass fractions of mixture components. The model is cross-validated successfully and gives more satisfactory results than the literature ANN models, which consider only mass fractions as input features. In addition, the influence of functional groups on the sooting tendency of fuel mixtures is evaluated for various blending ratios and fuel classes by means of sensitivity analysis.

Citation: Wang H, Zhu L, Cai L. 2025. Prediction of the sooting tendency of fuel mixtures with quantitative structure-property relationship (QSPR) model based on artificial neural network approach. *Progress in Reaction Kinetics and Mechanism* 50: e013 https://doi.org/10.48130/prkm-0025-0013

Introduction

The design of fuels offers the possibility of simultaneously increasing thermal efficiency and decreasing pollutant emissions for practical combustion devices, which requires, nevertheless, an accurate knowledge of the physical and chemical properties of fuel candidates. In addition to the widely considered properties, such as octane number, cetane number, and heating value, the yield sooting index is of growing interest due to the increasingly stringent emission regulations and the importance of joint consideration of combustion and pollutant performance[1]. The yield sooting index (YSI), which was proposed originally by the Pfefferle group at Yale University, characterizes the fuel-sooting behavior[2] based on the maximum soot concentration of a methane/air co-flow flame doped by this fuel. Literature studies[2–4] determined the YSI values of 428 neat fuel species for various fuel classes, including alkanes, alkenes, aromatics, alcohols, ethers, and esters. Very recently, Cheng et al.[5] reported the YSI of 21 neat fuel components and 151 mixtures of them. This pioneering study and its valuable datasets open unprecedented possibilities for understanding the sooting characteristics of fuel blends.

Efforts were made in the literature to develop analytic and numerical models for the estimation of fuel properties, such as YSI, mostly for neat fuel components[3,6–13]. In previous work[5], artificial neural network (ANN) models were proposed to estimate the YSI of fuel mixtures by taking the reported dataset[5] as the training data and the mass fractions of mixture components as input features. However, literature studies[14–16] have demonstrated that ANN models can predict the physical and chemical properties of fuels very satisfactorily, including YSI, and provide insights into the influence of molecule structures on properties if structure information of fuel molecules can be taken into account in ANN models.

Therefore, in this study, to predict the YSI of fuel mixtures with improved accuracy and to understand the impact of molecular functional groups on the sooting tendency of fuel mixtures, an ANN-based quantitative structure-property relationship (QSPR) model is developed. In conjunction with the mass fractions of mixture components, functional group descriptors are taken as input features, following the previous works[14,15]. With this, the present model is supposed to provide increased prediction accuracy and enable functional analysis beyond previous models. The model is cross-validated against the datasets reported by Cheng et al.[5]. Sensitivity analyses are performed to evaluate the effects of functional groups on fuel's sooting tendency for various blending ratios and fuel classes, which is infeasible with previous models.

Materials and methods

ANN method

The ANN model was developed with the Tensorflow Package provided by Keras[17]. Following previous works[14,15,18,19], the molecular group descriptors were employed as input features. For a specific fuel *j*, the quantity of group i contained in it is denoted by $x_{i,j}$ and its YSI value is indicated by y_j. To ensure a good training convergence, the input data $x_{i,j}$ and the target data y_j were normalized as:

$$x_{i,j,normal} = \frac{x_{i,j} - \overline{x}}{\sigma_{x,i}} \quad \text{and} \quad y_{j,normal} = \log_{10} y_j$$

here, $\bar{x}$, and $\sigma_{x,i}$ are the mean values and standard deviations of the numbers of feature *i* on the entire fuel dataset. The quantity of group *i* in the mixture is estimated as the quantity of this group in component *j* multiplied by the mass fraction of this component in the mixture.

In the ANN model, the input data is transferred to the hidden layers, where the input of each neuron consists of the sum of the outputs of neurons in the prior layer and additional biases. Following our previous works[14,15,20], K-fold cross-validation with a variety of random seeds was considered in the model development, which rigorously evaluates the model prediction accuracy by testing its performance across diverse and statistically representative subsets. The dataset was partitioned randomly into K subsets of identical size for the training of K models. Each model was trained with a distinct subset as the validation set, while the remainder was used for training. The model performance was evaluated based on the values, including the correlation coefficient (R^2), mean absolute percentage error (MAPE), and root mean squared error (RMSE) across the K models. Due to the influence of varying distributions of the K subsets on prediction accuracy, a number of random seeds were utilized, with the selection of the seed that yielded the optimal model performance. The implementation of a 5-fold cross-validation was adopted in this study.

QSPR method

As the physical and chemical properties of fuels are highly sensitive to their molecular structures, QSPR models are often developed to estimate the property values[14,15,20–22]. In this study, an ANN-based QSPR model was proposed by providing molecule functional groups as input features. To determine molecular groups and their values, the library descriptors within the open-source cheminformatics software RDKit[23] were used to extract required information from molecules, which are represented by simple molecular-input line-entry system (SMILES) codes[24]. Eight functional groups, which are involved in the species available in the dataset[5] and shown in Table 1, were selected as initial input features. Note that more functional group descriptors and model retaining will be required if additional fuel classes, such as ester and ether, are added to the training dataset.

As the prediction of YSI may be insensitive to certain features, an iterative backward feature elimination process[25] was carried out to obtain an optimal set of input features. The feature combination with the highest correlation coefficient (R^2) was selected to ensure the highest model prediction accuracy. Features with minimal impact on generalization capability were eliminated. It was found that the highest R^2 was achieved by removing the group Tr. Thus, seven features were selected as input features for the final model.

Computational details

The definition of hyper-parameters is of high importance for the performance of the ANN model[14,20]. To find the optimal combination of hyper-parameters with low deviation and computational cost, the random search method was employed in this study to assess the model performance for different sets of hyperparameters. The adaptive moment estimation algorithm (Adam)[26] provided by TensorFlow[27] was employed as an optimizer due to its high convergence performance. The method used in this work incorporates two loops with cross-validation. The first inner loop identifies the best model for different hyper-parameter combinations in terms of mean RMSE over all cross-validation splits. It is then considered in the second loop, where its performance is evaluated based on held-out test data. The output consists of a list of parameter sets, along with the scores of the first (RMSE_inner) and second loop (RMSE_test, r2_test), facilitating the identification of the optimal set of hyper-parameters.

It was found that the optimal numbers of neurons for the two hidden layers are 128 and 1,024, respectively, and the rates for the two dropout layers are both 10%. The Adam hyper-parameters α, β_1, and β_2 are 0.008, 0.9, and 0.9995, respectively. The activation function is ReLU (Rectified Linear Unit), and the optimal training epoch number is 700.

Table 1. Eight molecular groups initially considered as input features.

Group	Abbreviation	Type
$-CH_3$	P	Primary carbon (non-ring, saturated)
$-CH_2-$	S	Secondary carbon (non-ring, saturated)
>CH–	T	Tertiary carbon (non-ring, saturated)
>C<	Q	Quaternary carbon (non-ring, saturated)
$-CH_2-$(ring)	Sr	Secondary carbon (ring, saturated)
>CH–(ring)	Tr	Tertiary carbon (ring, saturated)
=CH–ring)	Sr*	Secondary carbon (ring, unsaturated)
=C<(ring)	Tr*	Tertiary carbon (ring, unsaturated)

Training dataset

The YSI dataset of 21 neat fuel components and 151 mixtures of them reported by Cheng et al.[5] was considered in this work for the model training. Three fuel classes, alkanes, cycloalkanes, and aromatics, are involved. Isomers are considered by distinguishing them with functional groups. The YSI of each fuel was employed as a benchmark metric to evaluate model prediction efficacy. Functional groups and the mass fractions were integrated as input features for model training. The blending ratios of mixtures are between 0 to 94%.

Sensitivity analysis

To gain a deeper insight into the impacts of different molecular groups on the sooting tendency of the fuel mixtures, local sensitivity analysis on the input features is performed by using the perturbation method[28], following our previous works[14,15,20]. Each input feature was perturbed over the whole dataset with 20% of its standard deviation over the dataset. The YSI's sensitivity $S_{i,j}$ on the feature x_i for a particular fuel species *j* is calculated as:

$$S_{i,j} = \left.\frac{\partial y}{\partial x_i}\right|_{x_j} = \left.\frac{\log_{10}[YSI(X_{i,j}+0.2\sigma_{x,i})]-\log_{10}[YSI(X_{i,j}-0.2\sigma_{x,i})]}{0.4\sigma_{x,i}}\right|_{x_j}$$

where, x_j is the vector of feature values for fuel species *j*.

Results and discussion

An ANN-based QSPR model was developed using the approaches described in the previous section. Its performance is demonstrated in this section. The results of sensitivity analysis are shown here as well for a deep understanding of the impacts of functional groups on the sooting tendency of fuel blends.

The performance of the derived ANN-based QSPR model is illustrated in Fig. 1. It is seen that the model predicts the YSI of neat components and their mixtures with high accuracy. An R^2 value of 0.990 is obtained in conjunction with an mean absolute error (MAE) value of 2.015.

The present model is compared with those ANN models proposed by Cheng et al.[5] in terms of error measures, including correlation coefficient, mean deviation, mean relative error, root mean square error, and mean absolute error. The results are shown in Table 2. It is seen that, regarding all error measures, the QSPR model developed in this study gives more satisfactory results than those models developed by taking only mixture compositions as input features, while some models are proposed with ANN approaches as well. Obviously, the incorporation of chemical molecule information in

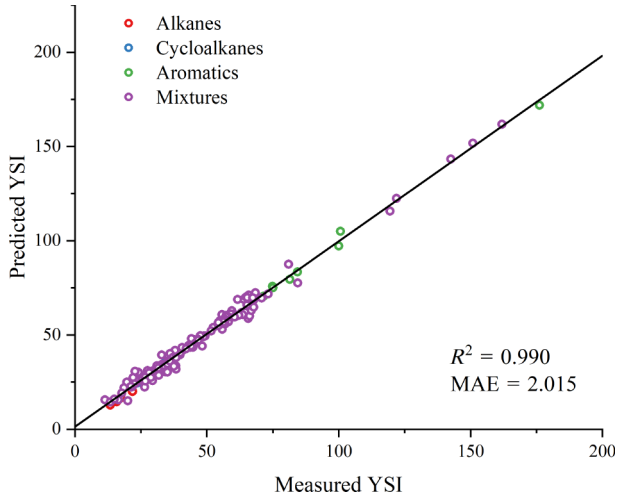

Fig. 1 Comparison of measured and predicted YSI values.

Table 2. Prediction performance of the ANN model developed in the present work compared with literature models.

Model	Correlation coefficient (R^2)	Mean deviation (MD)	Mean relative error (MRE)	Root mean square error (RMSE)	Mean absolute error (MAE)
Linear[5]	0.9640	4.24	13.72%	5.41	4.24
ANN[5]	0.8349	5.43	15.39%	8.30	7.15
S-ANN[5]	0.9796	2.27	6.55%	3.12	2.22
QSPR-ANN	0.9900	1.85	2.29%	2.67	2.01

terms of functional groups improves the performance of the ANN model.

The results of sensitivity analyses are presented next. The averaged sensitivities of YSI of neat components and mixtures on different functional groups are shown in Fig. 2. It should be noted that some groups are missing in certain fuel classes. For instance, the group Q is only available in alkane fuels. Thus, even though it is numerically possible to estimate the sensitivity of YSI of cycloalkanes and aromatics in group Q, this meaningless value is not presented and considered for further analysis.

As shown in Fig. 2, while the impacts of some groups, such as S, Q, Sr, and Tr*, increase by blending, the groups T and Sr* have a smaller influence in the mixtures than in the neat components. Moreover, it is interesting that the sensitivities of YSI of blends on groups can be significantly different from those of the neat components, for instance, for the P group. This can be attributed to the relatively high composition of alkanes in the mixtures with cycloalkanes or aromatics. The mixture composition plays an important role in the sooting tendency of fuel blends in conjunction with the functional groups.

The average sensitivities of YSI of mixtures with different compositions on functional groups are shown in Fig. 3. For most groups, mixtures only consisting of alkanes show higher sensitivities than other mixtures. However, this should be further verified and analyzed in more detail, as the input datasets contain only one fuel mixture that is solely composed of alkanes. More data and analysis are required to confirm this very high sensitivity. For blends of alkanes and cycloalkanes, the functional groups T and Sr* show opposite effects on the YSI compared to other mixtures. With the exception of blends composed only of alkanes, the mixtures containing alkanes, cycloalkanes, and aromatics in the dataset exhibit the highest sensitivities in most functional groups.

Group Q exhibits consistently large negative sensitivity across all fuel mixtures and also for neat components. Data from Pfefferle

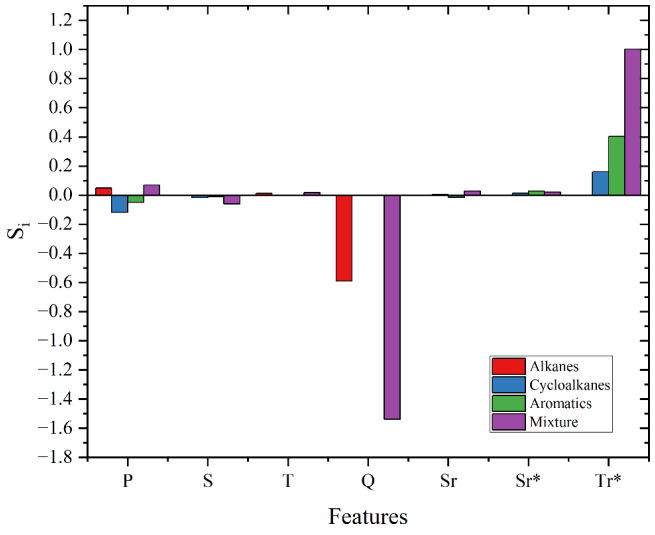

Fig. 2 Averaged sensitivities of YSI of neat components and mixtures on different functional features.

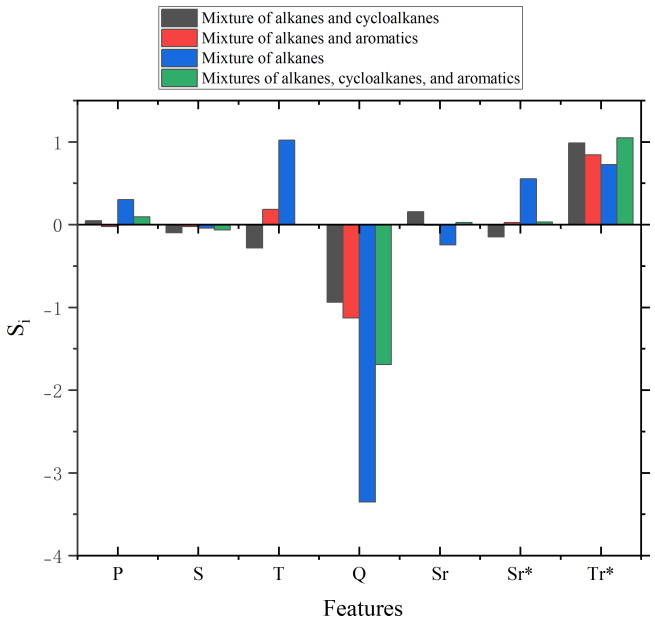

Fig. 3 Averaged sensitivities of YSI of mixtures on functional features.

et al.[29] also show that Q-rich neat fuels exhibit significantly lower YSI. This suggests that group Q has a positive effect in reducing the sooting tendency of both neat and blended fuels. The reason for this may lie in the inhibited production of soot precursors, such as acetylene and propyne, during the oxidation and combustion of Q-rich fuels.

For a deeper insight, the sensitivities of YSI on functional groups are also determined for different blending ratios of fuel classes. The results are shown in Fig. 4. For alkanes, notable sensitivities on functional groups are observed for blending ratios of 60%–70%, while impacts of functional groups are more pronounced at blending ratios of < 40% for cycloalkanes and aromatics. Interestingly, for alkanes at blending ratios of roughly 60%, the functional groups Sr and Sr* can have both positive and negative influences on the sooting behaviors of different blends. This can be observed for cycloalkanes and aromatics at small blending ratios as well. The impact of the secondary carbon group on the YSI is obviously ambiguous and

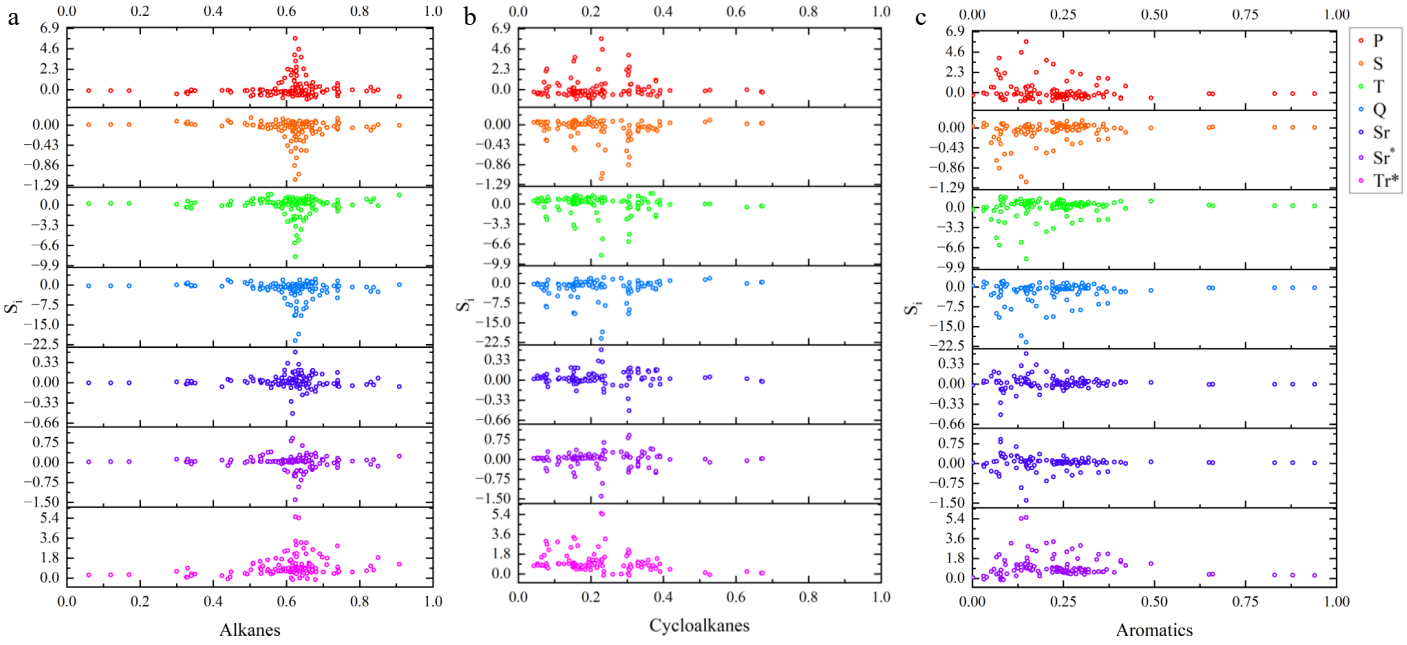

Fig. 4 Sensitivities of YSI on functional features over blending ratios of fuel classes.

depends on mixtures. For the other five groups, their impacts for different fuel classes are, in general, consistent.

In summary, group Q suppresses the soot formation of fuel mixtures significantly, while group Tr* strongly enhances the sooting tendency. Sr and Sr* can have both positive and negative influences on the sooting behaviors, depending on the fuel blends. These observations provide valuable information for the future design of fuel mixtures with low soot emissions.

Conclusions

An ANN-based QSPR model is developed in this study to estimate the YSI of fuel components and their mixtures. In addition to mass fractions of mixture components, functional groups of molecules are taken as additional input features, establishing a more accurate ANN model. As expected, the present model shows prediction advantages over ANN models proposed in the literature, which only consider mixture composition as input information. Moreover, it facilitates the functional group analysis, and thus, the impacts of functional groups on the sooting tendency of fuel mixtures are explored. It was found that the impact of functional groups is not necessarily larger in mixtures and can be different in mixtures than in neat components. Regardless of the mixture components, the quaternary carbon group in the molecule structure can contribute to a soot reduction. Except for secondary carbon groups, the impacts of functional groups are, in general, consistent for different fuel classes but depend on blending ratios.

Author contributions

The authors confirm their contributions to the paper as follows: investigation: Wang H; writing: Wang H, Zhu L, Cai L; supervision: Zhu L, Cai L; conceptualization: Cai L. All authors reviewed the results and approved the final version of the manuscript.

Data availability

All data generated or analyzed during this study are included in this published article.

Acknowledgments

The work was supported by the National Natural Science Foundation of China (Grant No. 52276133) and the Shanghai Science and Technology Program (Grant No. 23160711900).

Conflict of interest

The authors declare that they have no conflict of interest.

Dates

Received 4 April 2025; Revised 27 May 2025; Accepted 16 June 2025; Published online 8 July 2025

References

1. Bond TC, Doherty SJ, Fahey DW, Forster PM, Berntsen T, et al. 2013. Bounding the role of black carbon in the climate system: a scientific assessment. *Journal of Geophysical Research: Atmospheres* 118:5380–552
2. McEnally, CS, Das DD, Pfefferle LD. 2017. *Yield Sooting Index Database. Volume 2: Sooting Tendencies of a Wide Range of Fuel Compounds on a Unified Scale*. Cambridge, MA, USA: Harvard Dataverse, Harvard University
3. Das DD, St John PC, McEnally CS, Kim S, Pfefferle LD. 2018. Measuring and predicting sooting tendencies of oxygenates, alkanes, alkenes, cycloalkanes, and aromatics on a unified scale. *Combustion and Flame* 190:349–64
4. McEnally CS, Pfefferle LD. 2007. Improved sooting tendency measurements for aromatic hydrocarbons and their implications for naphthalene formation pathways. *Combustion and Flame* 148:210–22
5. Cheng X, Ren F, Gao Z, Zhu L, Huang Z. 2022. Synergistic effect analysis on sooting tendency based on soot-specialized artificial neural network algorithm with experimental and numerical validation. *Fuel* 315:122538
6. Han J. 2022. *CFD Modeling of Ignition and Soot Formation for Advanced Compression-Ignition Engines*. Ph. D. Thesis. The Pennsylvania State University, University Park, PA, USA
7. St John PC, Kairys P, Das DD, McEnally CS, Pfefferle LD, et al. 2017. A quantitative model for the prediction of sooting tendency from molecular structure. *Energy & Fuels* 31:9983–90

8. Pepiot-Desjardins P, Pitsch H, Malhotra R, Kirby SR, Boehman AL. 2008. Structural group analysis for soot reduction tendency of oxygenated fuels. *Combustion and Flame* 154:191–205
9. Barrientos EJ, Lapuerta M, Boehman AL. 2013. Group additivity in soot formation for the example of C-5 oxygenated hydrocarbon fuels. *Combustion and Flame* 160:1484–98
10. Gao Z, Zou X, Huang Z, Zhu L. 2019. Predicting sooting tendencies of oxygenated hydrocarbon fuels with machine learning algorithms. *Fuel* 242:438–46
11. Abdul Jameel AG. 2021. Predicting sooting propensity of oxygenated fuels using artificial neural networks. *Processes* 9:1070
12. Cai G, Liu Z, Zhang L, Shi Q, Zhao S, et al. 2021. Systematic performance evaluation of gasoline molecules based on quantitative structure-property relationship models. *Chemical Engineering Science* 229:116077
13. Li R, Herreros JM, Tsolakis A, Yang W. 2021. Machine learning-quantitative structure property relationship (ML-QSPR) method for fuel physicochemical properties prediction of multiple fuel types. *Fuel* 304:121437
14. Chen Z, Vom Lehn F, Pitsch H, Cai L. 2023. Prediction of sooting index of fuel compounds for spark-ignition engine applications based on a machine learning approach. *Journal of Thermal Science* 32:521–30
15. vom Lehn F, Cai L, Copa Cáceres B, Pitsch H. 2021. Exploring the fuel structure dependence of laminar burning velocity: a machine learning based group contribution approach. *Combustion and Flame* 232:111525
16. Kessler T, St John PC, Zhu J, McEnally CS, Pfefferle LD, et al. 2021. A comparison of computational models for predicting yield sooting index. *Proceedings of the Combustion Institute* 38(1):1385–93
17. Gulli A, Pal S. 2017. *Deep learning with Keras*. vol. 1. Birmingham, UK: Packt Publishing Ltd. 296 pp. www.packtpub.com/product/deep-learning-with-keras/9781787128422
18. vom Lehn F, Cai L, Tripathi R, Broda R, Pitsch H. 2021. A property database of fuel compounds with emphasis on spark-ignition engine applications. *Applications in Energy and Combustion Science* 5:100018
19. Joback KG, Reid RC. 1987. Estimation of pure-component properties from group-contributions. *Chemical Engineering Communications* 57:233–43
20. vom Lehn F, Brosius B, Broda R, Cai L, Pitsch H. 2020. Using machine learning with target-specific feature sets for structure-property relationship modeling of octane numbers and octane sensitivity. *Fuel* 281:118772
21. Nagaraja SS, Sarathy SM, Mohan B, Chang J. 2024. Machine learning-driven screening of fuel additives for increased spark-ignition engine efficiency. *Proceedings of the Combustion Institute* 40:105658
22. Katritzky AR, Lobanov VS, Karelson M. 1995. QSPR: the correlation and quantitative prediction of chemical and physical properties from structure. *Chemical Society Reviews* 24:279–87
23. Landrum GI. 2013. *Rdkit: a software suite for cheminformatics, computational chemistry, and predictive modeling*. www.rdkit.org/RDKit_Overview.pdf
24. Weininger D. 1988. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences* 28:31–36
25. Kohavi R, Sommerfield D. 1995. Feature subset selection using the wrapper method: overfltting and dynamic search space topology. *Proc. First International Conference on Knowledge Discovery and Data Mining, Montreal, Canada, 1995*. Menlo Park, CA: AAAI Press. pp. 192–97. https://cdn.aaai.org/KDD/1995/KDD95-049.pdf
26. Kingma DP, Ba JL. 2015. Adam: a method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, San Diego, USA*. Ithaca, NY: arXiv.org. pp. 1–10. doi: 10.48550/arXiv.1412.6980
27. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, et al. 2016. Tensorflow: large-scale machine learning on heterogeneous distributed systems. *arXiv* Preprint:1–10
28. Gevrey M, Dimopoulos I, Lek S. 2003. Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecological Modelling* 160:249–64
29. Das DD, McEnally CS, Kwan TA, Zimmerman JB, Cannella WJ, et al. 2017. Sooting tendencies of diesel fuels, jet fuels, and their surrogates in diffusion flames. *Fuel* 197:445–58