



Original Paper

Systematic diesel molecular performance evaluation based on quantitative structure-property relationship model

Guang-Qing Cai, Lin-Zhou Zhang*

State Key Laboratory of Heavy Oil Processing, Petroleum Molecular Engineering Center (PMEC), China University of Petroleum, Beijing 102249, China



ARTICLE INFO

Article history:

Received 20 January 2021

Accepted 28 June 2021

Available online 30 October 2021

Edited by Xiu-Qiu Peng

Keywords:

Diesel

Optimal molecules

QSPR

Molecular screening

ABSTRACT

The production and separation of optimal molecules in diesel fuels require a systematic property evaluation for the containing molecules. This paper evaluates the diesel molecules based on four key quality indicators: low-temperature fluidity, cleanliness, ignition, and power performance. We established the corresponding quantitative structure-property relationship models for corresponding properties, which are freezing point, yield sooting index, cetane number, and combustion heat. The models were applied for the screening of the high-performance molecules that is suitable for diesel. The molecular performance distribution of the conventional diesel and biodiesel were also compared. Moreover, we analyzed the effect of different transformation paths on molecular properties, giving guidance on the conversion process design.

© 2021 The Authors. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Diesel is still one of the primary motor fuels all over the world. The vehicle's emission has been considered to contribute significantly to the air pollution. A lot of attentions have been paid to the environmental protection, leading to a stricter emission regulation on diesel fuel. There is a growing demand for refineries to produce high-performance and clean diesel to meet the quality standards (Liu H et al., 2008). While the traditional diesel processing technology mainly focused on the bulk properties of feedstocks and products, the emerging diesel processing technology requires molecular-level information to achieve better product quality with high selectivity.

In recent years, petroleum molecular engineering technology has attracted widespread attentions. One of the critical problems is to find the proper separation and transformation routines to convert feedstock molecules into high value-added product molecules (Zhu, 2018; Hu et al., 2002; Hu and Zhu, 2004). The determination of high value-added molecules is critically important for the process design, which requires a systematic evaluation of the molecular property. Our previous work has built a comprehensive evaluation model for gasoline molecules, predicting five quality indicators: anti-knock, evaporation, power, stability, and

cleanliness performance (Cai et al., 2021). A systematic evaluation for diesel molecules has not yet been reported.

The property prediction for the pure component can be mainly classified into three categories based on the input features: group contribution, chemical descriptor, and deep learning methods. Many group contribution models have been developed for decades to calculate commonly used properties, such as boiling point, critical properties, freezing point (FP), and flash point (Joback and Reid, 1987; Constantinou and Gani, 1994; Marrero and Gani, 2001; Hukkerikar et al., 2012; Tsibanogiannis et al., 1995; White, 1986; Cordes and Rarey, 2002; Benson et al., 1969; Marrero and Gani 2001, 2001; Albahri and Tareq, 2003; Wen and Qiang, 2001; Li et al., 1994; Wang et al., 2020). Kubic et al. used group contribution and machine-learning algorithms to predict the cetane number (CN) of hydrocarbons and oxygenated compounds, showing good applicability on biofuels (Kubic et al., 2017). Saldana et al. developed a variety of machine-learning based QSPR models to estimate properties such as cetane number and flash point (Saldana et al. 2011, 2012). The advantage of the group contribution method is that the calculation procedure is simple and straightforward, but the determination of proper functional groups and the mathematic expression is time-consuming. Comparing to group contribution methods, chemical descriptors provide molecular features with a higher diversity. Karthikeyan et al. used chemical descriptors and artificial neural network (ANN) to predict the freezing point (Karthikeyan et al., 2005). Pan et al. used support vector machine (SVM) based QSPR models to predict properties

* Corresponding author.

E-mail address: lzz@cup.edu.cn (L.-Z. Zhang).

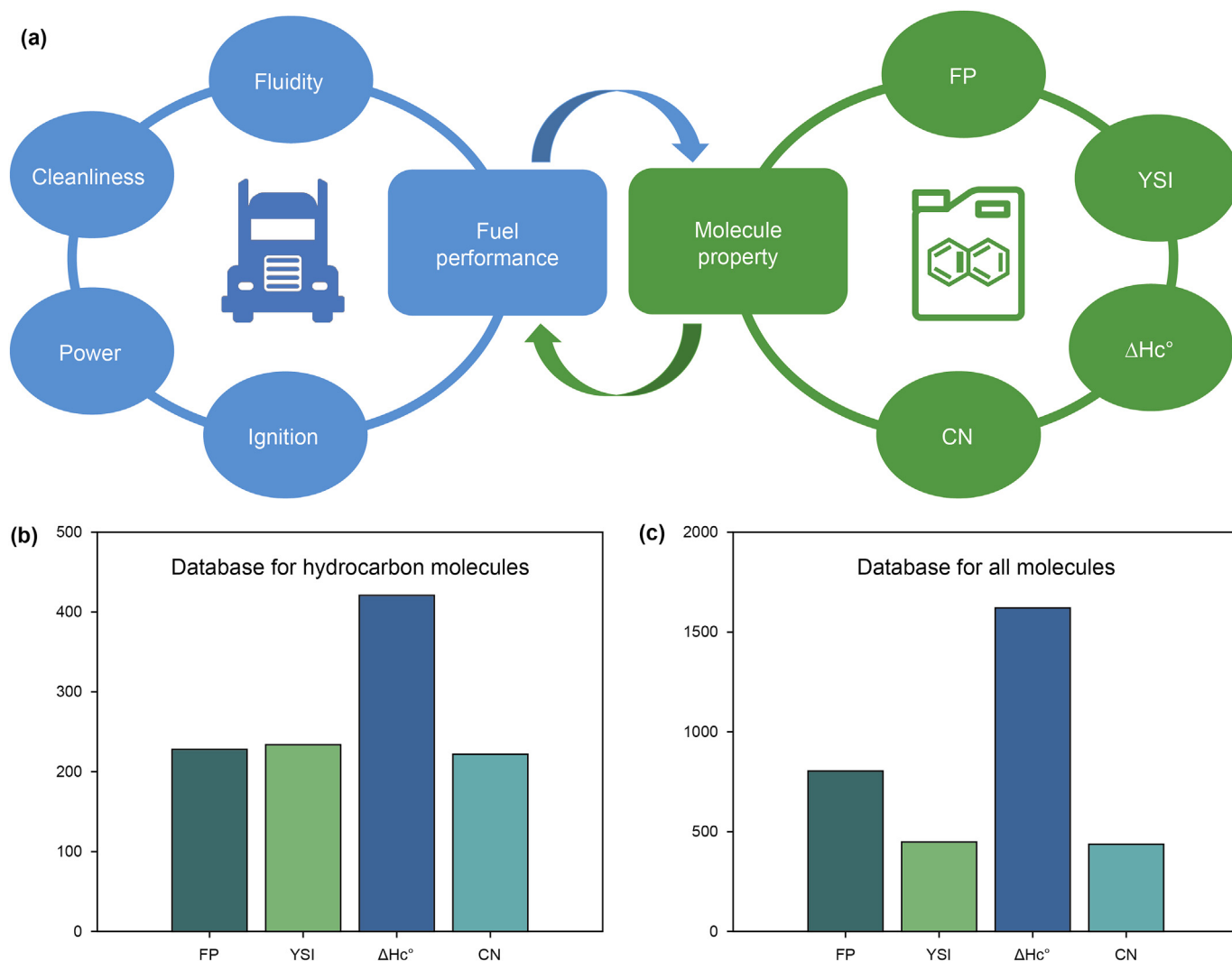


Fig. 1. a–c Main diesel quality indicators and the corresponding molecular properties; Database of key properties for hydrocarbon molecules and all molecules.

such as flash point and combustion heat (ΔH_c°) (Pan et al. 2008, 2011).

Recently, the deep learning method has been widely applied to property estimation since it has higher performance on the feature engineering and the regression process (Su et al., 2019; Goh et al., 2017; Han et al., 2017; Lashkenari et al., 2013; Dong et al., 2010). Schweidtmann et al. used the graph neural network to estimate the octane number and cetane number of pure compounds, in which the chemical bonds between atoms and atoms constitute the edges and nodes of the graph (Schweidtmann et al., 2020). Due to the lack of sufficient experimental data of octane number and cetane number, they combined multi-task learning, transfer learning, and integrated learning to overcome the small sample problem. Coley et al. applied deep learning to predict the products of a reaction (Coley et al., 2019). The model incorporated the possible sites of reactivity and evaluated their relative likelihoods. The accuracy of the main product predicted was higher than 85%. Su et al. combined tree-structured long short-term memory network and back-propagation neural network to predict critical properties (Goh et al., 2017). Different from traditional property prediction methods, deep learning method achieves a higher degree of automation during feature engineering and model training. However, a large amount of data was required to perform an effective network

structure parameters training.

So far, many property prediction methods for diesel molecules have been reported in the literature. However, a systematic performance evaluation model of diesel molecules is still necessary for its usage. In this paper, a systematic evaluation model was developed, which was composed of QSPR models for freezing point, YSI, cetane number, and combustion heat. The high value-added molecules of diesel can be distinguished by the model. Then, we discussed the distribution characteristics of the molecular performance of conventional diesel and biodiesel. The effects of different reaction pathways on the molecular property were also analyzed.

2. Materials

2.1. Data preparation

To establish a systematic evaluation model, firstly we must determine the key quality indicators for diesel usage. The diesel quality standard specifies solidifying point, cloud point, and cetane number as main quality indicators. The solidifying point and cloud point represent low-temperature fluidity. To avoid diesel solidification at low temperature, a lower solidifying point and cloud point

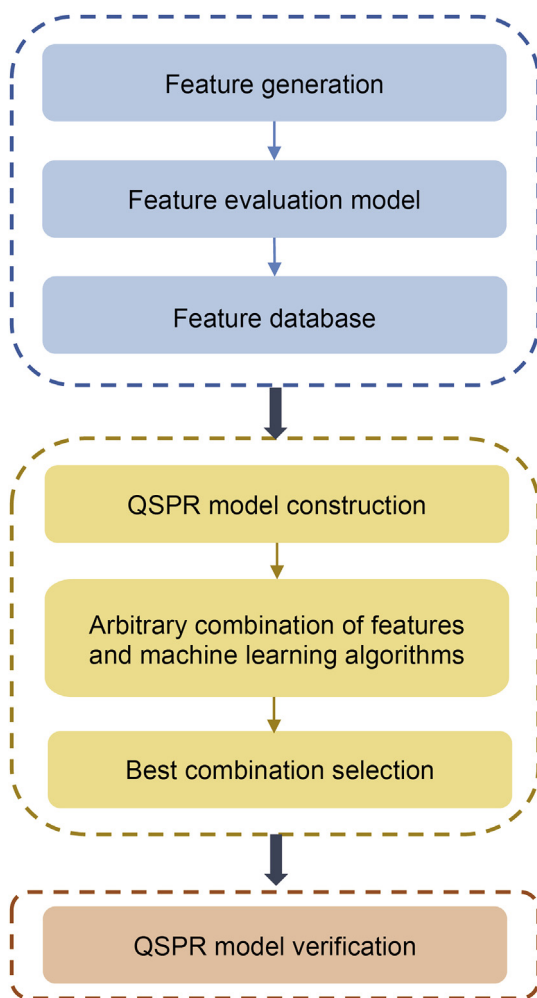


Fig. 2. Flow chart of diesel molecular performance evaluation model construction.

is preferred. The cetane number represents ignition performance. A higher cetane number donate to a smooth engine work. In the common diesel standards, the cetane number is required to be maintained in an appropriate range. Hydrocarbon provides sufficient power for the engine through combustion. The combustion heat is an indicator of diesel mileage. Vehicle emission is an important source of environmental pollution. The pollutants emitted should be reduced to a lower threshold. In summary, high-quality diesel should have good low-temperature fluidity, ignition stability, combustion heat, and cleanliness performance.

According to diesel quality standards and related literature, the low-temperature fluidity of diesel is generally represented by cloud point, pour point, and solidifying point. Nevertheless, there is no available data related to the low-temperature fluidity of diesel molecules in the literature. The freezing point is an indicator of the low-temperature fluidity of aviation fuels. Freezing points of hydrocarbons and oxygenates have been extensively reported. Therefore, we used the freezing point to represent the low-temperature fluidity of diesel. YSI represents the emissions emitted from diesel vehicles (Das et al. 2015, 2017, 2018; St John et al., 2017). Combustion heat represents the power performance. Cetane number represents the ignition performance. Fig. 1 (a) shows a schematic diagram of the four main indicators and the corresponding molecular properties. After determining the key properties, the corresponding dataset should be collected. This

work collected two sets of databases. For conventional diesel, the hydrocarbon database composed of only C and H elements was built with a total of 1105 data points. For biodiesel, the universal database including C, H, S, O, and N elements was established with 3311 data points (Kubic et al., 2017; Pan et al., 2011; Das et al., 2018; Gharagheizi et al., 2014; American Petroleum Institute, 1981). Fig. 1(b and c) displays the specific distribution of the main properties.

2.2. Molecular feature

In this paper, diesel molecules are expressed by a chemical language of simplified molecular input line entry specification (SMILES), which can encode chemical structure as simple text strings. Based on SMILES language, we calculated structural groups and chemical descriptors of diesel molecules. Structural groups were manually determined and automatically calculated for the diesel molecules, which can directly reflect the structural variation of molecules. Chemical descriptors contain molecular composition, topology, geometry, and quantum chemistry information, which have more rich structural information and can be automatically generated.

3. QSPR model development

The construction of the evaluation model was divided into three steps. Fig. 2 shows a flow chart of property prediction modeling. The first step was to build a feature evaluation model. The features that contribute significantly to the properties were selected, so the corresponding feature database was implemented. The second step was to integrate the selected features with machine learning algorithms to screen out the best combination of key properties. The third was to validate the extrapolation ability of QSPR models. It can prevent the model from over-fitting and ensure that the model has strong practicality.

3.1. Feature evaluation model construction

A good QSPR model can accurately predict the target properties with the fewest features. In this paper, two feature engineering methods were used for structural groups and chemical descriptors, namely the neural network feature screening method and principal component analysis. Fig. 3 is a schematic diagram of two feature engineering methods. Conventional diesel is mainly composed of hydrocarbon compounds and a small number of heteroatomic compounds (oxygen, nitrogen, sulfur). Since the combinatorial complexity of the structural group is small, limited groups can be used to precisely describe the structural variation of diesel molecules. To select structural groups that mainly contribute to their properties, a feature evaluation model should be established. The first stage was to input the basic groups into the neural network, where the training set accounts for 100% of the dataset. Meanwhile, the network parameters are constantly adjusted to train the network. The second was to add the united groups manually. The role of the united groups was to identify and distinguish the isomers in the database. In this way united groups were continuously added to the feature evaluation model until an acceptable error was obtained; Finally, the basic groups and united groups constituted the feature database.

Moreover, biodiesel consists of more oxygen elements besides hydrocarbons. The limited structural groups cannot fully describe the diverse structures of biodiesel. Therefore, we used chemical descriptors as features that include abundant chemical information. Due to a large number of descriptors, the principal component analysis was applied to screen effective descriptors and the

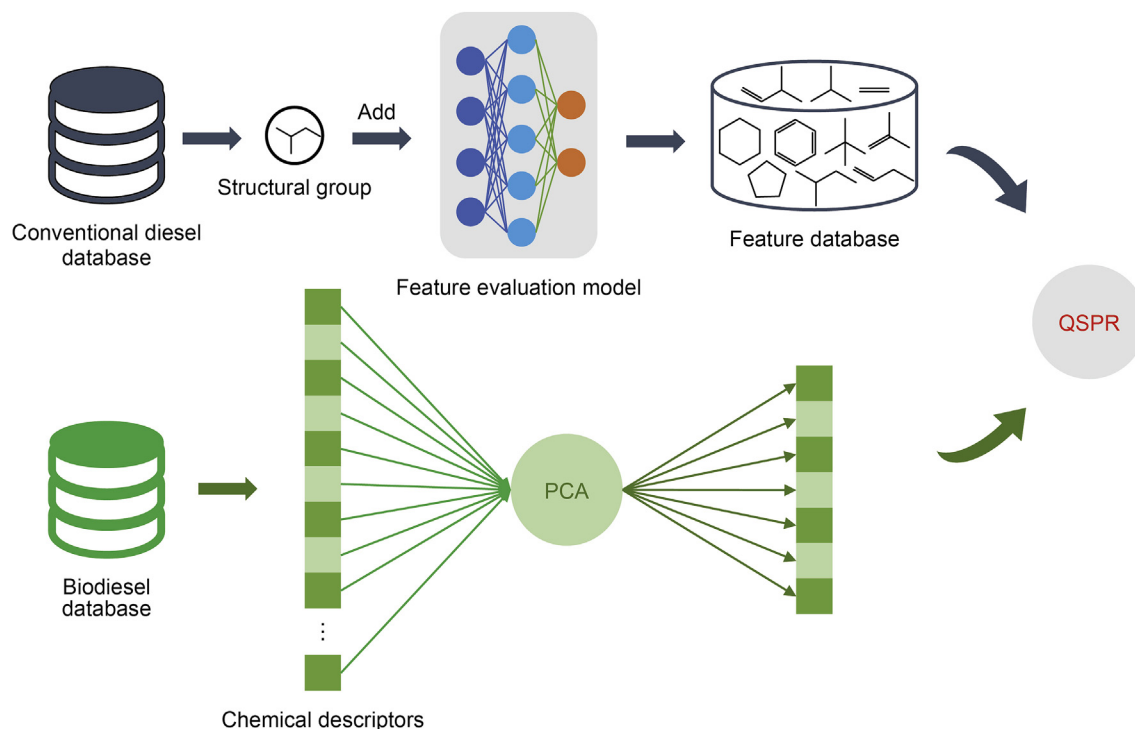


Fig. 3. Schematic diagram of feature engineering modeling: neural network feature screening method and principal component analysis.

database of chemical descriptors was determined. Conventional diesel could employ the neural network feature screening method and principal component analysis, while biodiesel only used principal component analysis.

3.2. Combination of QSPR models

Two types of features including structural groups and chemical descriptors were used in this work. The regression methods were prevalent machine learning approaches currently, containing support vector machine, random forest, and artificial neural network. The algorithms had a better capacity to regress nonlinearity and were very adapted to the complex and diverse molecular structures and properties of diesel. For the sake of gaining the optimal QSPR model for the key properties, we combined features and mathematical models arbitrarily to screen out the best combination for each property. For conventional diesel, YSI, cetane number, and combustion heat employed the combination of structural groups and ANN, and freezing point used chemical descriptors and RF through repeated trials and parameters adjustment. For biodiesel, it was found that the combination of chemical descriptors and ANN had higher performance for each property.

It is necessary to balance the fitness and model extrapolation while training QSPR models. Fig. 4 (a) takes the cetane number as a case study to compare the prediction results of ANN, SVM, and RF. It can be seen that the correlation coefficient of RF in the training process is 0.9825, which indicates the best performance for the training set. The SVM performs the best in the test set. However, The QSPR models were validated in two aspects, namely test set and extrapolation performance test. The extrapolation performance of a model is especially important. The extrapolation performance of the cetane number in Fig. 4 (b) corresponds to the methods in Fig. 4 (a). In Fig. 4 (b), the extrapolation curve of RF is not smooth enough, indicating the overfitting of model. Secondly, the extrapolation effect of RF on hydrocarbons with multiple rings and

oxygenated compounds is not good. Moreover, the extrapolation curve of SVM on hydrocarbons with multiple rings in the graph tends to be flat. The cetane number does not vary with increasing carbon number, which is not under common knowledge. The ANN has good extrapolation performance on various hydrocarbon homologous series. The error between ANN and SVM in the training set and test set is very small. Therefore, the ANN model was selected for cetane number prediction finally.

3.3. QSPR model validation

Taking the freezing point as an example, we investigated the construction and verification process of the QSPR model. The low-temperature fluidity of diesel is not only related to whether diesel engines can supply fuel normally, but also has a close relationship with storage and transportation of diesel at low temperatures. The freezing points of straight-chain paraffins, branched paraffins and naphthenes are quite different, indicating that the freezing point is very sensitive to the structure. Hence, the prediction of the freezing point for diesel molecules is a challenge. In this paper, the freezing point prediction model applied the group of chemical descriptors and RF. Fig. 5 (a) shows the parity plot of the experimental and estimated values. The correlation coefficients were 0.9766 and 0.9736, respectively. The predicted value was in good agreement with the experimental value.

After training the model, it was indispensable to test the homologous extrapolation ability of the model. Fig. 5 (b) shows the homologous series extrapolation test of different types of hydrocarbons. The predicted values were consistent with the experimental data. A growing trend of freezing point with carbon number can be observed from the figure for the same homologous series. The only exception was that the freezing points of some naphthenes without branched chains were usually higher. The variation tendency is in line with experimental data. Moreover, there is no experimental data on the freezing point of hydrocarbons with

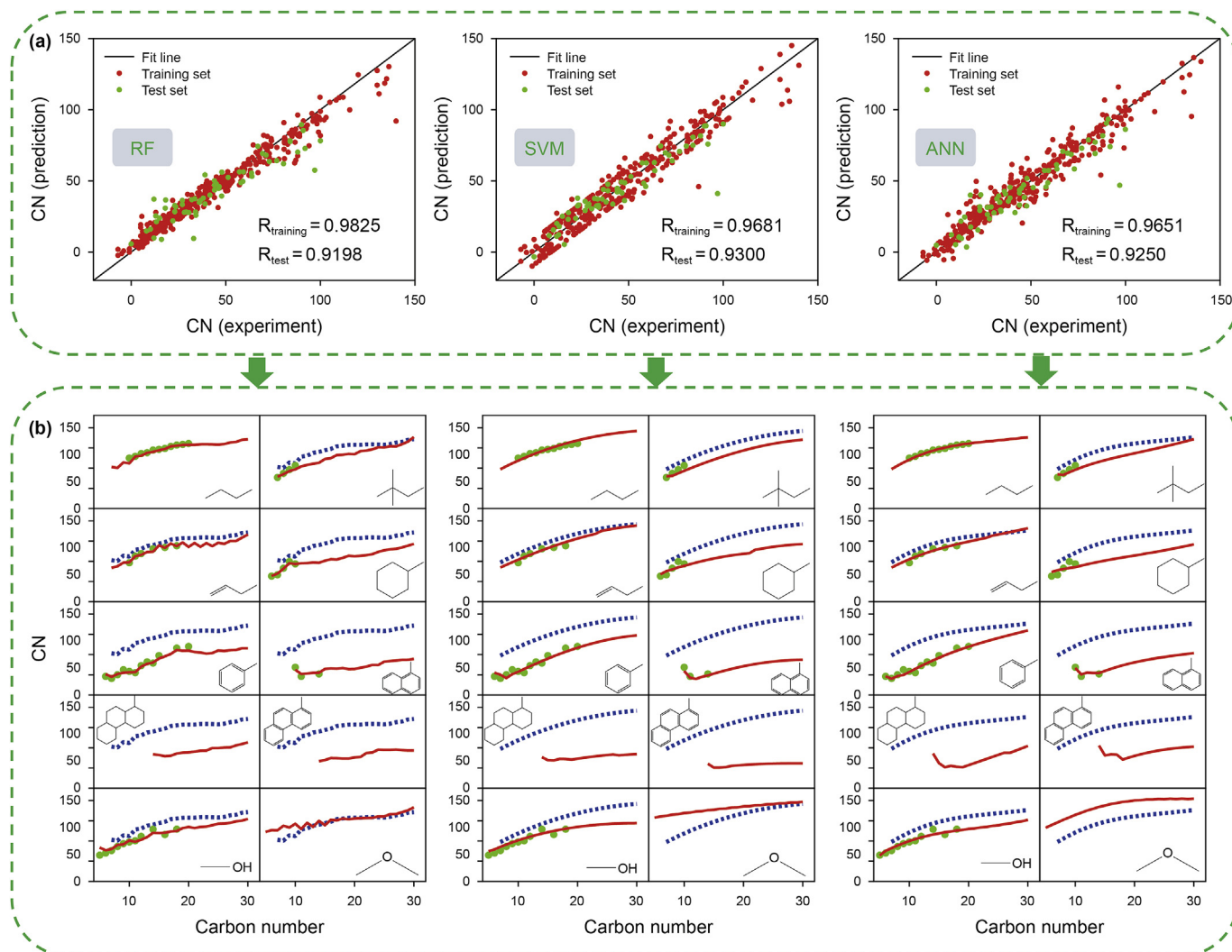


Fig. 4. Comparison of results of the neural network, support vector machine, and random forest for cetane number prediction.

multiple rings in the database. But the QSPR model we established can still precisely predict its variation trend, demonstrating that the model has a strong extrapolation ability. From Fig. 5(b), it can also be found that in addition to hydrocarbons with multiple rings, branched chains, double bonds, and rings are all structural features that have a positive contribution to the freezing point. This revealed that the conversion of normal paraffins into these structures is helpful to reduce the freezing point and increase the low-temperature fluidity. The model construction of other properties was similar to the freezing point. The prediction results of YSI and combustion heat can be checked in our previous work for gasoline molecules (Hu and Zhu, 2004). For the results of other properties, please see Fig. S1–S8 in the supporting information.

In summary, we established two groups of evaluation models, which are applied to two scenarios of conventional diesel and biodiesel. Each model has undergone repeated training and strict extrapolation, thereby obtaining optimum models with high accuracy and strong generalization. Then, we applied QSPR models to the molecular screening and conversion process. Therefore, we will elaborate on the application of evaluation models in the selection of optimal molecules and the optimal path.

4. Model application

4.1. Performance evaluation of pure component

The QSPR models of the key properties were combined to form a comprehensive evaluation system. We took a radar chart to show the pros and cons of diesel molecules intuitively. Several representative compounds were selected, including paraffins, iso-paraffins, olefins, naphthenes, aromatics, and oxygenated compounds. In Fig. 6, the coordinates are cetane number, combustion heat, YSI, and freezing point. The molecules with a higher cetane number and combustion heat represent better molecules. Similarly, the molecules with lower YSI and freezing points are better. Correspondingly, the shaded area denotes high-performance molecules in the radar chart. From the figure, we can readily find high value-added molecules, such as hexyl-cyclohexane and 2,6,10-trimethylundecane, which have good low-temperature fluidity, emission, and combustion performance. In consequence, iso-paraffins and naphthenes are ideal blending components of high-quality clean diesel. Refineries should purposefully convert feedstocks into these two types of molecules to improve diesel quality in practical production.

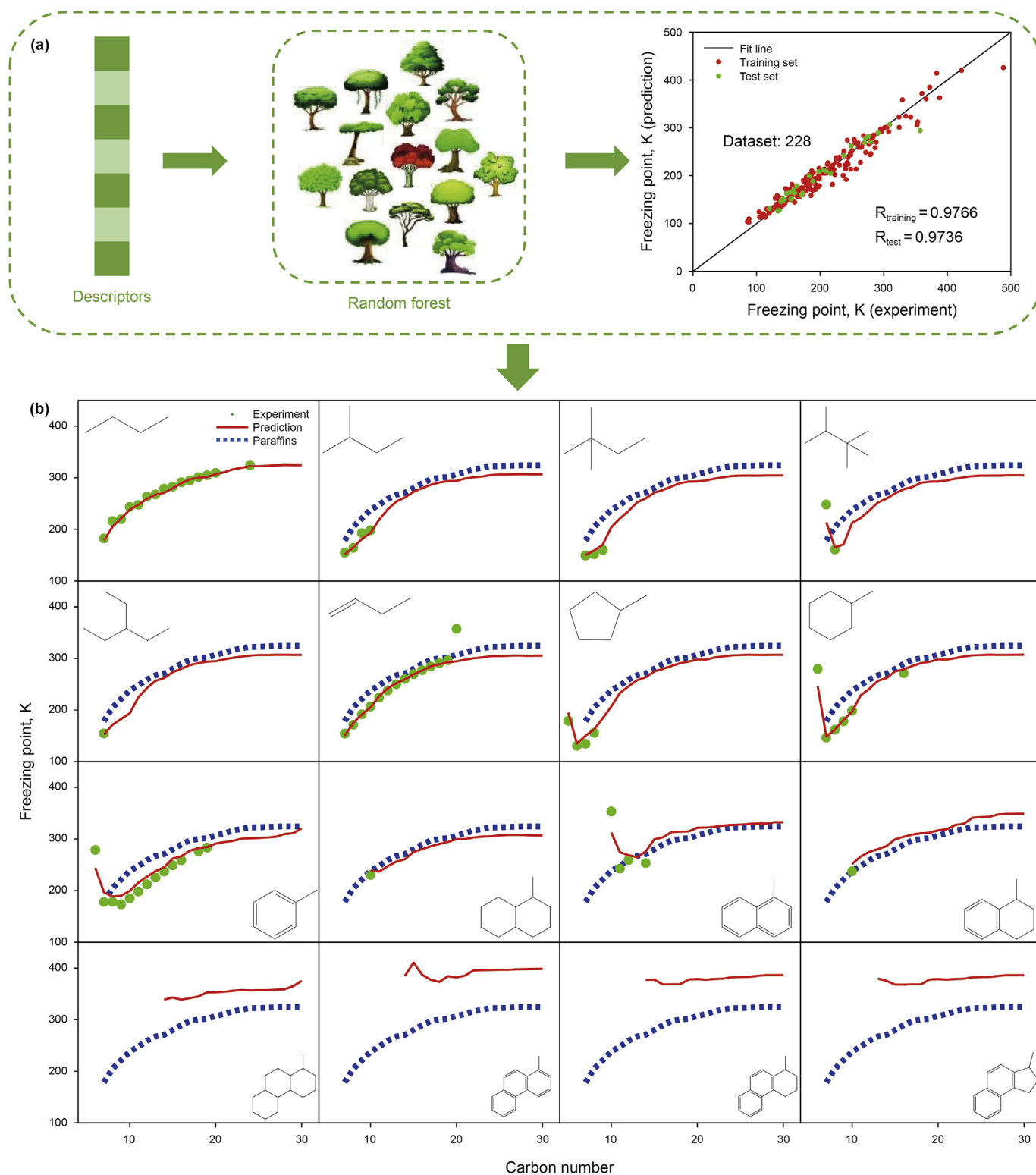


Fig. 5. a–b Schematic diagram of freezing point prediction model construction using chemical descriptors and random forest combination methods; Comparison of the experimental and predicted freezing point of different kinds of hydrocarbons. The blue dashed line represents paraffin data used as a reference.

4.2. Molecular performance distribution characteristic

The performance distribution of molecules in diesel mixture is useful for the separation and conversion process design. Therefore, we investigate the molecular performance distributions for typical

diesels. Fig. 7 shows the molecular performance distribution maps of hydrotreating diesel, catalytic cracking diesel, biodiesel produced from waste tires and palm oil (Aguing et al., 2012; Ayanoglu and Yumrutaş, 2016; Mota et al., 2014). The distribution map was partitioned into four parts, in which the part of low YSI and low

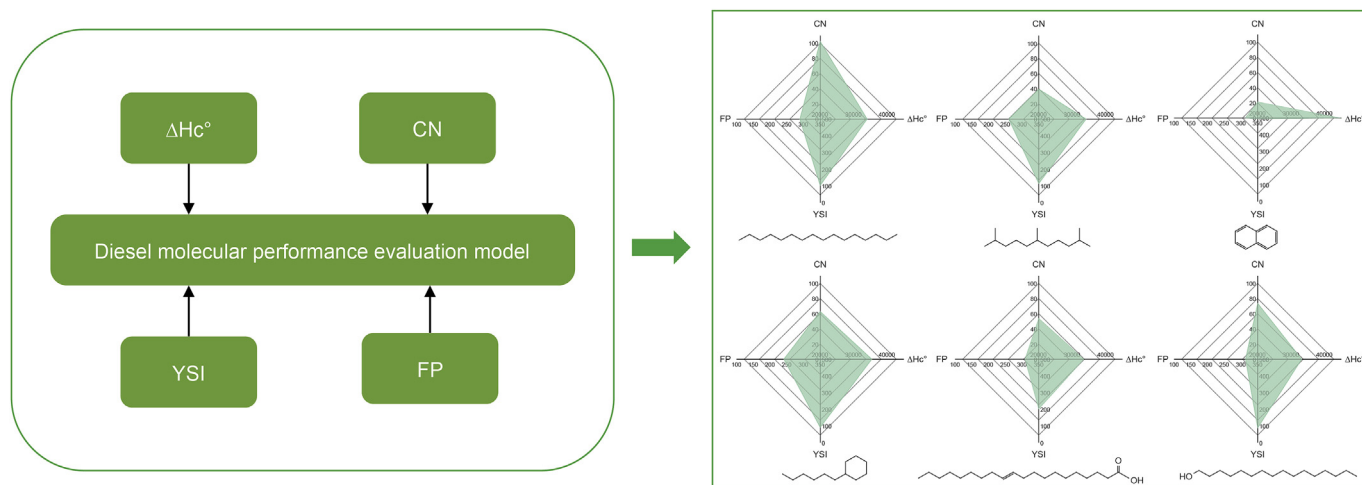


Fig. 6. Radar chart of comprehensive performance of representative molecules using the combination of QSPR models.

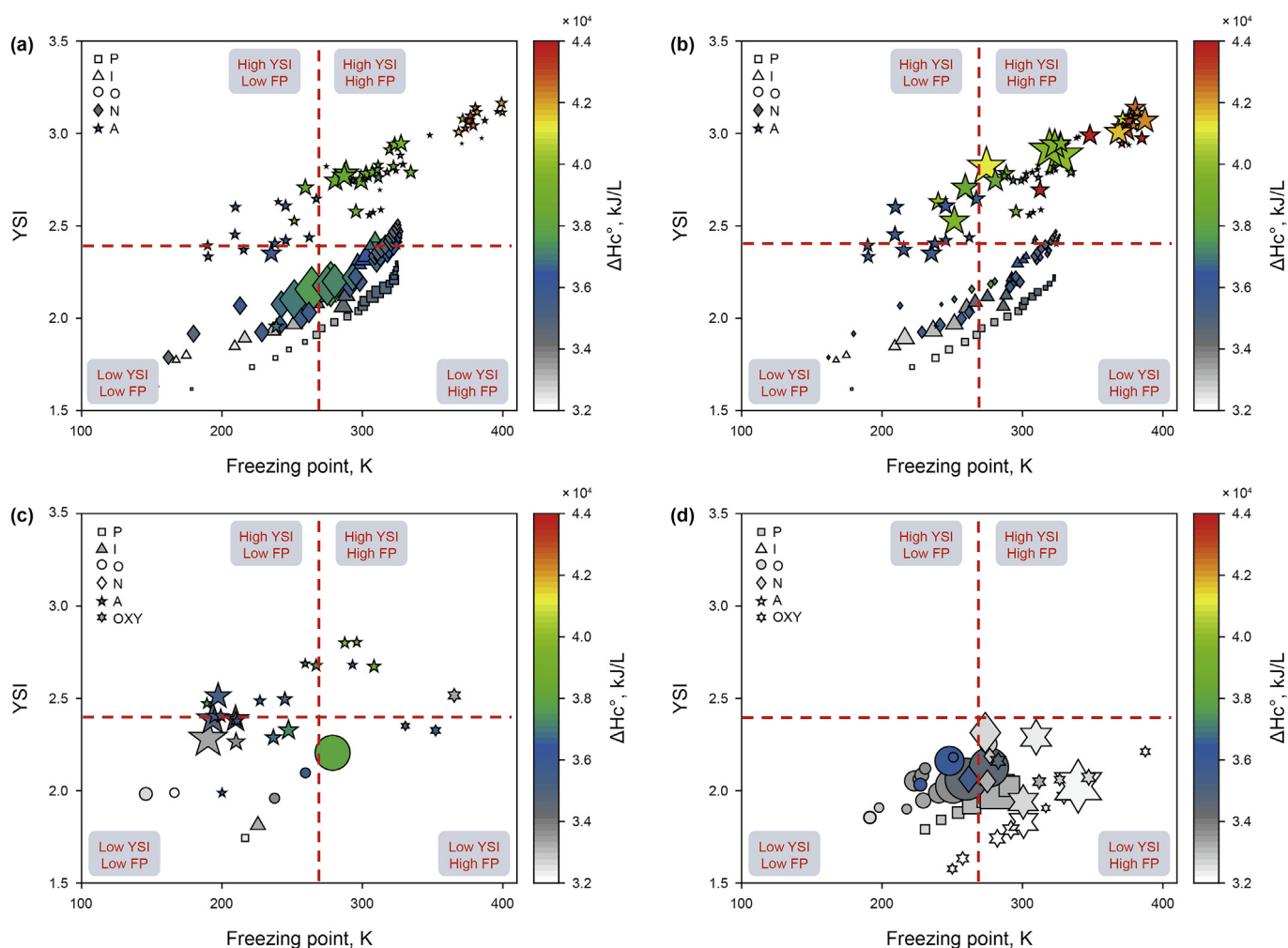


Fig. 7. a–d The distribution characteristics of various diesels. High naphthenic diesel after hydrotreatment; Olefinic and aromatic diesel after fluid catalytic cracking; Biodiesel from waste tire oil using catalytic cracking; Biodiesel from crude palm oil using thermal catalytic cracking. Dot size represents the concentration.

freezing point are better blending components. It can be observed that there are more compounds with low freezing point and YSI in the hydrotreating diesel, where the content of naphthenes

accounts for the majority. When used as a blending component, it will improve the low-temperature fluidity and reduce emissions of blended diesel. The catalytic cracking diesel contains more

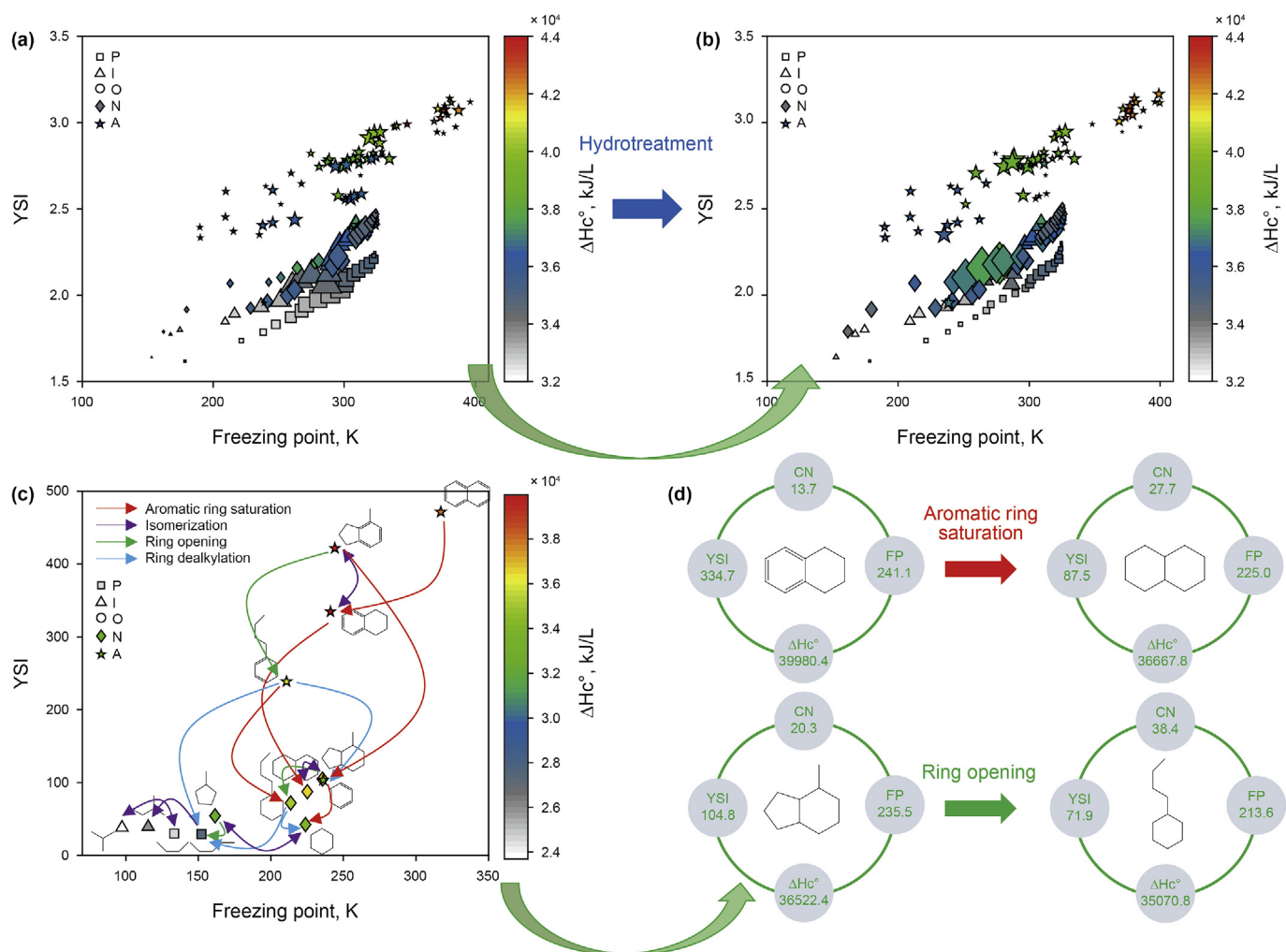


Fig. 8. a–d The conversion pathways of hydrotreatment for naphthalene. The distribution characteristics of raw material and product; Transformation pathways of hydrotreatment for naphthalene compound; Specific numerical variation of the key properties.

components with high YSI and freezing point, which will emit more pollutants and have poor low-temperature fluidity. Biodiesel, as a substitute fuel for traditional petrochemical diesel, is relatively clean and renewable energy. It has the advantages of wide sources, low cost, and environmental protection. Biodiesel from catalytic pyrolysis of waste tires has higher content of low molecular weight aromatics, which can be used as raw materials for chemicals production. Biodiesel produced by thermal catalytic cracking of palm oil contains more olefins and oxygenated compounds. The components are mainly concentrated in the low YSI area, indicating that it is a low-emission clean fuel. With the increasing importance of environmental protection today, we need to develop more related processes of biodiesel production to reduce environmental pollution. In summary, both hydrotreating diesel and biodiesel produced by catalytic cracking of palm oil can be used as better components for diesel blending. Their common characteristics are low YSI and low freezing point, which can improve the quality of blended diesel.

4.3. Effect of conversion pathways on molecular properties

The optimal transformation pathways were investigated in the hydrotreating process. We took the hydrotreating reaction as an example to illustrate the process. Fig. 8 exhibit the distribution

characteristics of raw materials and products through hydrotreating. The raw material involves more alkanes and fewer aromatics. The product contains more naphthenes through the hydrotreating reaction. The paraffins and iso-paraffins were greatly reduced, thereby lowering the freezing point of hydrotreated diesel.

The naphthalene compound was taken to illustrate the selection of optimal paths in the hydrotreating reaction. The hydrotreating reaction network was constructed at the molecular level. The aromatic ring saturation, isomerization, ring-opening, ring dealkylation mainly occurred in the hydrotreating reaction. In Fig. 8 (c), we have drawn the hydrotreating pathways diagram of naphthalene. The molecules moving left signify a reducing freezing point and moving down denotes a decreasing YSI. It can be found that all pathways in the hydrotreating reaction of naphthalene are all moving down or left, which indicates that hydrotreating is an efficient process for producing high-quality clean diesel. Fig. 8 (d) shows the specific numerical variation of the key properties of diesel before and after hydrotreating. When the aromatic ring is saturated, the cetane number increases, the freezing point, YSI, and combustion heat all decrease. The comprehensive performance of the product has been significantly improved compared with the reactant, especially the YSI value. The ring-opening reaction is similar to aromatic ring saturation. The key properties of the

product are all improved.

Hence, the reaction paths of all colors are ideal routes of moving down or left through detailed analysis. This demonstrates that hydrotreating is an effective means to improve the quality of diesel. It can improve the ignition performance and low-temperature fluidity while decreasing the emissions of motor vehicles. The reaction pathways on the map can help us understand the effect of various conversion paths on the properties deeply and direct refineries to select and control the optimal route.

5. Conclusion

In this work, two groups of systematic evaluation models have been constructed for the main quality indicators of diesel, including low-temperature fluidity, cleanliness, power, and ignition performance. One was the hydrocarbon database composed of only C and H elements for conventional diesel. The QSPR models were built by structural group + ANN and chemical descriptor + RF. Its accuracy was high, but the application range was small. The other was the universal database containing C, H, S, O, and N elements for biodiesel. The models were established by chemical Descriptor + principal component analysis + ANN, which have a broader application and good generalization. Then, we used the evaluation model to guide the optimal structure selection and the identification of the optimal pathways. First, we adopted the radar chart to visually display the pros and cons of representative molecules. Second, the distribution characteristics of various diesels were compared, including conventional diesel and biodiesel. Also, we investigated the ideal component for diesel blending. Finally, a detailed analysis of the hydrotreating reaction process of naphthalene was carried out. The reaction routes that improve the quality of diesel can be clearly observed. The systematic evaluation model we developed has important guiding significance for high-quality diesel production.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 22021004 and U19B2002). The authors declare no competing financial interests.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.petsci.2021.10.022>.

References

- A.P.I.R.P. and K.R. Hall, 1981. *Selected Values of Properties of Hydrocarbons and Related Compounds*. Thermodynamics Research Center.
- Albahri, Tareq A., 2003. Structural group contribution method for predicting the octane number of pure hydrocarbon liquids. *Ind. Eng. Chem. Res.* 43 (3), 657–662. <https://doi.org/10.1021/ie020306+>.
- Altae-Tran, H., Ramsundar, B., Pappu, A.S., et al., 2017. Low data drug discovery with one-shot learning. *ACS Cent. Sci.* 3 (4), 283–293. <https://doi.org/10.1021/acscentsci.6b00367>.
- Aquino, M., et al., 2012. Composition analysis and viscosity prediction of complex fuel mixtures using a molecular-based approach. *Energy Fuels* 26 (4), 2220–2230. <https://doi.org/10.1021/ef300106z>.
- Ayanoglu, A., Yumrutaş, R., 2016. Production of gasoline and diesel like fuels from waste tire oil by using catalytic pyrolysis. *Energy* 103, 456–468. <https://doi.org/10.1016/j.energy.2016.02.155>.
- Benson, S.W., Cruickshank, F.R., Golden, D.M., et al., 1969. Additivity rules for the estimation of thermochemical properties. *Chem. Rev.* 69 (3), 279–324. <https://doi.org/10.1021/cr60259a002>.
- Cai, G., Liu, Z., Zhang, L., et al., 2021. Systematic performance evaluation of gasoline molecules based on quantitative structure-property relationship models. *Chem. Eng. Sci.* 229, 116077. <https://doi.org/10.1016/j.ces.2020.116077>.
- Coley, C.W., Jin, W., Rogers, L., et al., 2019. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chem. Sci.* 10 (2), 370–377. <https://doi.org/10.1039/c8sc04228d>.
- Constantinou, L., Gani, R., 1994. New group contribution method for estimating properties of pure compounds. *AIChE J.* 40 (10), 1697–1710. <https://doi.org/10.1002/aic.690401011>.
- Cordes, W., Rarey, J., 2002. A new method for the estimation of the normal boiling point of non-electrolyte organic compounds. *Fluid Phase Equil.* 201 (2), 409–433. [https://doi.org/10.1016/S0378-3812\(02\)00050-X](https://doi.org/10.1016/S0378-3812(02)00050-X).
- Da Mota, S.A.P., et al., 2014. Production of green diesel by thermal catalytic cracking of crude palm oil (*Elaeis guineensis* Jacq) in a pilot plant. *J. Anal. Appl. Pyrol.* 110, 1–11. <https://doi.org/10.1016/j.jaap.2014.06.011>.
- Das, D.D., McEnally, C.S., Pfefferle, L.D., 2015. Sooting tendencies of unsaturated esters in nonpremixed flames. *Combust. Flame* 162 (4), 1489–1497. <https://doi.org/10.1016/j.combustflame.2014.11.012>.
- Das, D.D., Cannella, W.J., McEnally, C.S., et al., 2017a. Two-dimensional soot volume fraction measurements in flames doped with large hydrocarbons. *Proc. Combust. Inst.* 36 (1), 871–879. <https://doi.org/10.1016/j.proci.2016.06.047> doi: 10.1016/j.proci.2016.06.047.
- Das, D.D., McEnally, C.S., Kwan, T.A., et al., 2017b. Sooting tendencies of diesel fuels, jet fuels, and their surrogates in diffusion flames. *Fuel* 197, 445–458. <https://doi.org/10.1016/j.fuel.2017.01.099>.
- Das, D.D., et al., 2018. Measuring and predicting sooting tendencies of oxygenates, alkanes, alkenes, cycloalkanes, and aromatics on a unified scale. *Combust. Flame* 190, 349–364. <https://doi.org/10.1016/j.combustflame.2017.12.005>.
- Dong, X., Wang, S., Sun, R., et al., 2010. Design of artificial neural networks using a genetic algorithm to predict saturates of vacuum gas oil. *Petrol. Sci.* 7 (1), 118–122. <https://doi.org/10.1007/s12182-010-0015-y>.
- Gharagheizi, F., et al., 2014. A group contribution model for the prediction of the freezing point of organic compounds. *Fluid Phase Equil.* 382, 21–30. <https://doi.org/10.1016/j.fluid.2014.08.025>.
- Goh, G.B., Hodas, N.O., Siegel, C., et al., 2017a. Smiles2vec: an interpretable general-purpose deep neural network for predicting chemical properties. *arXiv preprint arXiv:1712.02034*.
- Goh, G.B., Siegel, C., Vishnu, A., et al., 2017b. Chemception: a deep neural network with minimal chemistry knowledge matches the performance of expert-developed QSAR/QSPR models. *arXiv preprint arXiv:1706.06689*.
- Hu, S., Zhu, X.X., 2004. Molecular modeling for plant-wide optimization. *Chem. Eng. Commun.* 191 (4), 513–530. <https://doi.org/10.1080/00986440490277965>.
- Hu, S., Towler, G., et al., 2002. Combine molecular modeling with optimization to stretch refinery operation. *Ind. Eng. Chem. Res.* 41 (4), 825–841. <https://doi.org/10.1021/ie0010215>.
- Hukkerikar, A.S., et al., 2012. Group-contribution+ (GC+) based estimation of properties of pure components: improved property estimation and uncertainty analysis. *Fluid Phase Equil.* 321, 25–43. <https://doi.org/10.1016/j.fluid.2012.02.010>.
- Joback, K.G., Reid, R.C., 1987. Estimation of pure-component properties from group-contributions. *Chem. Eng. Commun.* 57 (1–6), 233–243. <https://doi.org/10.1080/00986448708960487>.
- Karthikeyan, Muthukumarasamy, Robert, C., 2005. Glen, and Andreas Bender. General melting point prediction based on a diverse compound data set and artificial neural networks. *J. Chem. Inf. Model.* 45 (3), 581–590. <https://doi.org/10.1021/ci0500132>.
- Kubic Jr., W.L., et al., 2017. Artificial neural network based group contribution method for estimating cetane and octane numbers of hydrocarbons and oxygenated organic compounds. *Ind. Eng. Chem. Res.* 56 (42), 12236–12245. <https://doi.org/10.1021/acs.iecr.7b02753>.
- Lashkenari, M.S., Taghizadeh, M., Mehdiadeh, B., 2013. Viscosity prediction in selected Iranian light oil reservoirs: artificial neural network versus empirical correlations. *Petrol. Sci.* 10 (1), 126–133. <https://doi.org/10.1007/s12182-013-0259-4>.
- Li, P., Ma, P.S., Yi, S.Z., et al., 1994. A new Corresponding-States Group-Contribution method (CSGC) for estimating vapor pressures of pure compounds. *Fluid Phase Equil.* 101, 101–119. [https://doi.org/10.1016/0378-3812\(94\)02607-6](https://doi.org/10.1016/0378-3812(94)02607-6).
- Liu, H., Yu, J., Fan, Y., et al., 2008. A scenario-based clean gasoline production strategy for China National Petroleum Corporation. *Petrol. Sci.* 5 (3), 285–294. <https://doi.org/10.1007/s12182-008-0048-7>.
- Marrero, J., Gani, R., 2001. Group-contribution based estimation of pure component properties. *Fluid Phase Equil.* 183, 183–208. [https://doi.org/10.1016/S0378-3812\(01\)00431-9](https://doi.org/10.1016/S0378-3812(01)00431-9).
- Pan, Y., et al., 2011. Predicting the net heat of combustion of organic compounds from molecular structures based on ant colony optimization. *J. Loss Prev. Process. Ind.* 24 (1), 85–89. <https://doi.org/10.1016/j.jlp.2010.11.001>.
- Pan, Y., et al., 2008. Quantitative structure–property relationship studies for predicting flash points of organic compounds using support vector machines. *QSAR Comb. Sci.* 27 (8), 1013–1019. <https://doi.org/10.1002/qsar.200810009>.
- Saldana, D.A., et al., 2011. Flash point and cetane number predictions for fuel compounds using quantitative structure property relationship (QSPR) methods. *Energy Fuels* 25 (9), 3900–3908. <https://doi.org/10.1021/ef200795j>.
- Saldana, D.A., et al., 2012. Prediction of density and viscosity of biofuel compounds using machine learning methods. *Energy Fuels* 26 (4), 2416–2426. <https://doi.org/10.1021/ef3001339>.
- Schweidtmann, A.M., et al., 2020. Graph neural networks for prediction of fuel ignition quality. *Energy Fuels* 34 (9), 11395–11407. <https://doi.org/10.1021/acs.energyfuels.0c01533>.
- St John, P.C., Kairys, P., Das, D.D., et al., 2017. A quantitative model for the prediction of sooting tendency from molecular structure. *Energy Fuels* 31 (9), 9983–9990.

- <https://doi.org/10.1021/acs.energyfuels.7b00616>.
- Su, Y., Wang, Z., Jin, S., et al., 2019. An architecture of deep learning in QSPR modeling for the prediction of critical properties using molecular signatures. *AIChE J.* 65 (9), e16678. <https://doi.org/10.1002/aic.16678>.
- Tsibanogiannis, I.N., Kalospiros, N.S., Tassios, D.P., 1995. Prediction of normal boiling point temperature of medium/high molecular weight compounds. *Ind. Eng. Chem. Res.* 34 (3), 997–1002. <https://doi.org/10.1021/ie00042a039>.
- Wang, Z., Su, Y., Jin, S., et al., 2020. A novel unambiguous strategy of molecular feature extraction in machine learning assisted predictive models for environmental properties. *Green Chem.* <https://doi.org/10.1039/d0gc01122c>.
- Wen, X., Qiang, Y., 2001. A new group contribution method for estimating critical properties of organic compounds. *Ind. Eng. Chem. Res.* 40 (26), 6245–6250. <https://doi.org/10.1021/ie010374g>.
- White, C.M., 1986. Prediction of the boiling point, heat of vaporization, and vapor pressure at various temperatures for polycyclic aromatic hydrocarbons. *J. Chem. Eng. Data* 31 (2), 198–203. <https://doi.org/10.1021/je00044a020>.
- Zhu, F.X.X., 2018. Molecular analysis for process synthesis. *Ind. Eng. Chem. Res.* 57 (43), 14583–14595. <https://doi.org/10.1021/acs.iecr.8b01919>.