Machine learning regression based group contribution method for cetane and octane numbers prediction of pure fuel compounds and mixtures

Li, Runzhao; Herreros, Martin; Tsolakis, Athanasios; Yang, Wenzhao

DOI:
10.1016/j.fuel.2020.118589

Document Version
Peer reviewed version

Citation for published version (Harvard):
Li, R, Herreros, M, Tsolakis, A & Yang, W 2020, 'Machine learning regression based group contribution method for cetane and octane numbers prediction of pure fuel compounds and mixtures', *Fuel*, vol. 280, 118589. https://doi.org/10.1016/j.fuel.2020.118589

Machine learning regression based group contribution method for cetane and octane numbers prediction of pure fuel compounds and mixtures

Runzhao Li,[†] Jose Martin Herreros,[†] Athanasios Tsolakis,[*,†] Wenzhao Yang [‡]

[†] *Department of Mechanical Engineering, School of Engineering, College of Engineering and Physical Sciences, University of Birmingham, Edgbaston, Birmingham B15 2TT, United Kingdom*

[‡] *Shenzhen Gas Corporation Ltd., Shenzhen 518049, China*

Abstract

Current methods to predict fuel ignition quality usually focus on either cetane numbers or research/motor octane numbers (CN, RON, MON) and most of them apply to pure compounds. A machine learning regression based group contribution method (GCM) is proposed to simultaneously predict CN, RON and MON of pure fuel compounds and mixtures. The GCM extracts the structural features of fuel molecules to build a molecular structure matrix. Then a mathematical model developed by machine learning correlates the molecular structure matrix with ignition quality (CN, RON, MON) matrix. A comprehensive fuel ignition quality database is built for model training which contains 603, 374, 371 compounds for CN, RON and MON respectively. High predictive precision is obtained for CN, RON, MON (R^2 equal to 0.9911, 0.9874, 0.9731) being superior to those obtained by neural

[*]Corresponding author.

E-mail address: a.tsolakis@bham.ac.uk

network. The method is successfully applied to a wide range of compounds including alkanes, alkenes, alkynes, cycloalkanes, cycloalkenes, aromatics, alcohols, aldehydes/ketones, ethers, esters, acids, furans and fuel mixtures. Three key factors contribute to the high predictive capacity: (i) GCM considers the structural features, functional group interaction and fuel reactivity of fuel molecules; (ii) the built-in machine learning algorithm automatically optimizes the model function and parameters and (iii) the fuel ignition quality database provides adequate model training data for different fuel types. This method provides an effective tool to obtain CN, RON and MON of pure compounds and mixtures and a fundamental understanding of the impact of fuel molecular structures on the ignition quality.

Highlights:

- CN/RON/MON prediction of pure compounds and mixtures by machine learning based group contribution method
- Group contribution method extracts the structural features and transforms into molecular structure matrix
- Machine learning regression model correlates the molecular structure matrix and ignition quality matrix
- A comprehensive fuel ignition quality database is developed for regression model training and validation

Keywords:

Fuel molecular structure; group contribution method; machine learning regression; CN/RON/MON prediction; pure fuel compounds & mixtures.

Nomenclature

Symbols

R^2	correlation coefficient

Abbreviations

ABFIS	adaptive network-based fuzzy inference system
AIChE	American Institute of Chemical Engineers
ASM	active subspace method
ASTM	American Society for Testing and Materials
CAS	chemical abstract service
CCDB	carbon-carbon double bond
CCTB	carbon-carbon triple bond
CFR	cooperative fuels research
CN	cetane number
$CN_{blending}$	cetane number of a specific fuel compound when it is blended with a base fuel in particular volume fractions
CVCC	constant volume combustion chamber
DCN	derived cetane number
FIT	fuel ignition tester
GCM	group contribution method

GPR	Gaussian process regression
IDT	ignition delay time
IUPAC	International Union of Pure and Applied Chemistry
IQT	ignition quality tester
KAUST	King Abdullah University of Science and Technology
LANL	Los Alamos National Laboratory
LLNL	Lawrence Livermore National Laboratory
MAE	mean absolute error
MLR	multiple linear regression
MSE	mean square error
MON	motor octane number
N/A	not applicable
NMR	nuclear magnetic resonance spectroscopy
NN	neural network
NREL	National Renewable Energy Laboratory
OS	octane sensitivity
PCR	principle component regression
QSPR	quantitative structure-property relationship
RMSE	root mean square error
RON	research octane number
SVM	support vector machines

SwRI	Southwest Research Institute
TI	topological indices
TPRF	toluene primary reference fuels (n-heptane-iso-octane-toluene mixture)
UOB	University of Birmingham

43

1. INTRODUCTION

Cetane number (CN) and research/motor octane number (RON/MON) are parameters to evaluate the fuel autoignition quality for compression ignition and spark ignition engine respectively. The CN and RON/MON are generally in an inverse proportion, the greater the CN, the more prone to autoignition while the greater the RON/MON, the more resistant to autoignition (anti-knock). Thence, the CN/RON/MON are the key fuel properties affecting engine combustion and emission performance.

Even though there are mature ASTM standards available for CN [1-4] /RON [5] /MON [6] measurement, there are still some significant challenges to be solved. First, it is expensive and time consuming to test the ignition quality by CFR engine or constant volume combustion chamber (CVCC) recommended by ASTM standards. The measurement of CN/RON/MON by CFR engine demand 500mL/sample 40min/sample (see Table S1 in supporting information) while the CVCC requires less quantity than CFR engine around 40~370 mL/sample. For those fuels not commercially available, it is unrealistic and unfeasible for researchers to produce such testable quantity and the higher the produced purity, the greater the cost. Second, most of the emerging fuels, including advanced biofuels derived from biomass, lack of measured ignition quality data (CN/RON/MON) and the only information available is the chemical formula. For example, the Fuel Properties Database [7] developed by Co-Optimization of Fuels & Engines project [8-10] contain 489 pure compounds, but only 291, 162, 110 compounds have measured CN, RON, MON. Third, there are no accurate method to calculate the CN/RON/MON of blending mixture. The mixture CN is usually estimated by a linear volume fraction weighted mixing rule from those pure compounds $CN_{mix} = \sum_{i=1}^{n} v_i \cdot CN_i$ because it gets more accurate estimations than linear relationship of molar fraction or mass fraction [11-14]. But this model does not consider non-linear interaction among fuel components which fails to reflect the synergistic or

antagonistic behavior with respect to its composition [15, 16]. Fourth, the typical testing ranges of CN/RON/MON by CFR engine are limited to 30~65 [1], 40~120.3 [5], 40~120 [6] as shown in Table S1 of supporting information. Even though the CFR engine can measure fuels outside the range, the precision has not been determined. Fifth, the advanced compression ignition engine requires low reactivity fuels (25<CN<40) [17, 18] but its ignition quality is difficult to be characterized by ASTM standards due to outside of the calibrated range.

The five challenges confronted by fuel ignition quality characterization lead to the development of alternative method to predict CN/RON/MON. The commonly used methods include: (1) group contribution method (GCM); (2) quantitative structure-property relationship (QSPR); (3) neural network (NN); (4) simulated ignition delay time (IDT); (5) active subspace method (ASM); (6) topological indices (TI); (7) adaptive network-based fuzzy inference system (ABFIS); (8) principle component regression (PCR); and (9) multiple linear regression (MLR) as summarized in Table 1. All these methods (except (4) simulated IDT method) correlate the ignition quality data with molecular structure being collectively called the QSPR method [19] . In theory, QSPR method can be applied to pure compounds and blending mixtures given a training database containing mixture data provided QSPR works on functional groups level instead of molecular level. The QSPR methods differ in terms of descriptor types and numbers of descriptors as shown in Table 1. The GCM method is one of the most commonly used QSPR methods that correlates the ignition quality data (CN/RON/MON) with types & numbers of functional group. NN method updates and adapts the regression model to new inputs and enables to capture the nonlinear relationship in the system. The neural network inputs can be molecular descriptors [20-22], functional groups [23, 24], NMR spectroscopy [25]. Topological indices are developed to deal with complex physicochemical properties which incorporate the branching

88 degree, shape and size of molecules [26-28] into consideration. The simulated IDT method is proposed
89 by Singh et al. [29] and the constant volume IDTs are correlated with RON, MON at equivalent RON
90 condition (750K, 25bar) and equivalent MON condition (825K, 25bar). There are some issues with this
91 simulated IDT method: (1) the predictive accuracy of regression model developed by correlating
92 CN/RON/MON with IDT is inferior to that by correlating CN/RON/MON with molecular structure.
93 This is because both CN/RON/MON and IDT are the secondary data of molecular structure, the
94 correlation function between CN/RON/MON and IDT contains uncertainty and measured error to some
95 extent. (2) The IDT is calculated by chemical kinetic mechanism, thus it is not applicable for fuels
96 without validated mechanisms. (3) It is unclear if it can be applied to CN prediction. (4) The validated
97 scope is limited to alkanes, alkenes, aromatics and their mixtures while the applicability to oxygenated
98 compounds is unclear.

99 There are three scales (CN, RON, MON) to characterize the fuel ignition quality, CN and
100 RON/MON apply to high reactivity fuels and low reactivity fuels respectively. The knowledge gap is
101 how to characterize ignition quality for specific fuel compound in these three scales respectively. The
102 conversion formula between CN and RON/MON is inaccurate and limited to limited fuel types (see
103 Table S2 in supporting information), it is necessary to predict theses three parameters simultaneously.
104 In addition, the ignition quality characterization for fuel mixtures is challenging and it is essential to
105 build a predictive model based on functional group level rather than molecular level. The scope of the
106 published predictive models is usually limited to a few compound groups (see Table 1), more chemical
107 classes especially oxygenates should be incorporated. This study develops a comprehensive
108 CN/RON/MON database, pure compounds and mixtures are included, to train and verify the predictive
109 model. A new group contribution method GCM-UOB 2.0 is proposed to extract the structural features

of different fuel types and their mixtures. This method is applicable to fuel compounds that the molecular structures are known. Machine learning algorithm is used to optimize the model functions and parameters to improve predictive accuracy.

113 **Table 1. Overview of CN and ON forecasting approaches**

Method	Target output	Model inputs	Optimal R^2	RMSE	No. of compounds	Scope	Ref.
GCM	CN/RON/MON	38 functional groups	0.90	N/A	449	Alkanes, alkenes, alkynes, cycloalkanes, cycloalkenes, aromatics, alcohols, aldehydes/ketones, ethers, esters, acids, furans	[30, 31]
	CN	13 functional groups, IQT ignition delay, vapor pressure	0.98	8.75	162	As above	[32]
QSPR	CN	28 functional groups	0.934	6.3	229	Hydrocarbons, alcohols, esters	[33]
	CN	150 molecular descriptors	0.978	N/A	147	Alkanes, alkenes, cycloalkanes, aromatics	[34]
	CN	^{13}C NMR spectroscopy and 7 group descriptors	0.64	N/A	127	34 pure alkanes, 93 hydrocarbon mixtures	[35]
	RON/MON	Molecular descriptors: 12 for RON, 23 for MON	0.92	N/A	552	279 for RON, 273 for MON, alkanes, alkenes, alkynes, cycloalkanes, cycloalkenes, aromatics, alcohols, esters, furans	[36]
	RON/MON	Molecular mass, hydration energy, boiling point, molar refractivity, octanol/water distribution coefficient, critical pressure, critical volume, critical temperature	0.9419	N/A	65	Alkanes, cycloalkanes	[37]
NN	CN	15 QSPR descriptors	0.963.	7.94	N/A	Alkanes, alkenes, alkynes, cycloalkanes, aromatics, alcohols, aldehydes/ketones, ethers, esters	[20]
	CN	12 hydrocarbon groups	0.97	N/A	69	Alkanes, cycloalkanes, aromatics	[23]
	CN	4 functional groups and boiling point for isoparaffins	0.97	N/A	141	iso-Paraffins and diesel fuels	[24]
	CN	10 molecular descriptors	0.934	N/A	349	Alkanes, alkenes, aromatics, alcohols, esters, others (3 ketones, 1 aldehyde, 8 ethers and 4 acids)	[21]
	CN	15 molecular descriptors	N/A	9.1	284	Alkanes, alkenes, alkynes, cycloalkanes, cycloalkenes, aromatics, alcohols, aldehydes/ketones, ethers, esters, acids, furans	[22]
	RON/MON	15 H types in H NMR spectroscopy	0.99	2.2	251	128 pure hydrocarbons: alkanes, alkenes, cycloalkanes, cycloalkenes, aromatics; 123 hydrocarbon blends: n-heptane, iso-octane toluene, trimethylbenzene, cyclopentane, 1-hexene, ethanol	[25]

TI	CN	Parameters of the general second degree equation of hyperbola	0.99998	N/A	71	Alkanes, cycloalkanes	[27]
	RON/MON	Fuel molecules	0.9643	N/A	27	Heptane isomers, octane isomers	[28]
	ON	Vector coefficients	0.9966	N/A	78	46 samples of alkanes, 32 samples of cycloalkanes	[26]
Simulated IDT	RON/MON	Computed ignition delay curve	N/A	N/A	N/A	Alkanes, alkenes, cycloalkane, aromatics, alcohols, ketones, esters, acids, furans	[38]
		RON: constant volume IDT at 750K, 25bar MON: constant volume IDT at 825K, 25bar	0.9932	N/A	N/A	Alkanes, alkenes, aromatics and their mixtures	[29]
		Compression ratio dependent variable volume IQT	0.9726	N/A	25	TPRF mixtures	[39]
ASM	CN	9 topological indices and 5 carbon-chain related descriptors	0.93	N/A	110	Alkanes, alkenes, cycloalkanes, aromatics	[40]
ABFIS	CN	4 evaporation relevant descriptors and 6 combustion relevant descriptors	0.986	3.38	496	204 hydrocarbons and 292 oxygenates, no further detail available	[41]
PCR	RON	Fourier-transform infrared absorption spectra	N/A	N/A	34	Alkanes, alkenes, cycloalkanes, aromatics	[42]
MLR	CN	H NMR spectroscopy	0.95	N/A	125	Alkanes, alkenes, alkynes, cycloalkanes, aromatics, hydrocarbon mixtures	[43]

114

2. Modeling approach

2.1 Methodological overview

A fuel ignition quality (CN/RON/MON) database is built based on open access database, project reports and published articles which are described in section 2.2. The fuel molecular structure and functional groups are the model input and CN/RON/MON is target output respectively. In the first step, the fuel molecular structure & functional groups information are accessed by chemical CAS (Chemical Abstract Service) number and IUPAC (International Union of Pure and Applied Chemistry) nomenclature (see section 2.2). In the second step, the GCM extracts structural features of fuel molecules to form molecular structure matrix (see section 2.3). Minimum number of feature descriptors are used to capture the fuel molecular structure characteristics to reduce model size and avoid overfitting. The molecular structure matrix contains n rows (number of samples) and 32 columns (number of structural feature descriptors). The ignition quality matrix contains n rows (number of samples) and 1 column (target fuel ignition quality of CN or RON or MON). It should be noted that the three predictive models of CN, RON and MON are trained separately using different datasets (see Table 3), therefore, only one column exists in the ignition quality matrix. In the third step, a machine learning regression model is developed to correlate the molecular structure matrix and ignition quality matrix. The ignition quality database and 23 machine learning algorithms (Linear regression algorithms: ① linear, ② interactions linear, ③ robust linear, ④ stepwise linear; Regression trees algorithms: ⑤ fine tree, ⑥ medium tree, ⑦ coarse tree, ⑧ optimizable tree; Support vector machines algorithms: ⑨ linear SVM, ⑩ quadratic SVM, ⑪ cubic SVM, ⑫ fine Gaussian SVM, ⑬ medium Gaussian SVM, ⑭ coarse Gaussian SVM, ⑮ optimizable SVM; Gaussian process regression algorithms: ⑯ rational quadratic, ⑰ squared exponential, ⑱ Matern 5/2, ⑲ exponential, ⑳ optimizable GPR; Ensembles of trees algorithms: ㉑ boosted trees, ㉒ bagged trees, ㉓ optimizable ensemble, see Table S3 in supporting information) are used to train the regression models in parallel,

then the algorithm with minimum root mean square error (RMSE) is selected. 10-fold cross validation is adopted to validate the model and prevent over-fitting. In the fourth step, the machine learning regression model developed is first deployed into a MATLAB APP with graphical interfaces and then into PC & Phone APP. Both the regression model (the model detail is discussed in detail in section 2.4) and the fuel ignition quality database are embed into a cloud database and a Web APP. The Web APP, Desktop APP and Phone APP are easy for users to characterize fuel ignition quality without programming knowledge requirement.

2.2 Fuel ignition quality database development

A fuel ignition quality database is set up for predictive model training and validation and the data sources are summarized in Table 2. The CN of pure compounds mainly derive from: (i) Co-Optimization of Fuels & Engines: Fuel Properties Database [7] released by NREL (National Renewable Energy Laboratory); (ii) Octane and cetane number data tabulation [30, 31] provided by LANL (Los Alamos National Laboratory); (iii) Compendium of experimental cetane numbers [12] available at NREL; (iv) Cluster of Excellence "Tailor-Made Fuels from Biomass" [32] managed by RWTH Aachen University. The selection priority of experimental CN is as follow: ASTM D613 (CFR engine test) [1]>ASTM D6890 (IQT test) [2]=ASTM D7170 (FIT test) [44]. The CN of hydrocarbon mixtures are mainly derived from journal articles [45-47] and thesis [48]. The RON/MON of pure compounds are mainly obtained from: (i) Co-Optimization of Fuels & Engines: Fuel Properties Database [7] released by NREL; (ii) Octane and cetane number data tabulation [30, 31] provided by LANL; (iii) API Tech Data Book [49] published by AIChE (American institute of Chemical Engineers). The RON/MON of mixtures are mainly acquired from journal articles [25, 29, 38, 39, 50-53]. Unlike CN, the RON and MON are only measured by CFR engine according to ASTM D2699 [5] and ASTM D2700 [6], therefore, the data reproducibility is good in different data source. In summary, the CN, RON, MON datasets contain 603, 374, 371 samples and the numbers of different chemical classes are provided in Table 3. Particularly, the term "polyfunctionals" refers to multi-functional (aromatic bond, carbon-carbon double bond, carbon-carbon triple bond, hydroxyl group, carbonyl group, aldehyde group, ether group, ester group) compounds. For example, 2-methoxyethanol (CAS 109-86-4) belongs to "polyfunctionals" since it contains hydroxyl group and ether group.

163 **Table 2. Data source of measured CN/RON/MON for pure compounds and fuel mixtures**

Items	Fuel type	Institute	Ref.
CN/RON/MON	Alkanes, alkenes, alkynes, cycloalkanes, cycloalkenes, aromatics, alcohols, aldehydes/ketones, ethers, esters, acids, furans	NREL	[7]
CN/RON/MON	Alkanes, alkenes, alkynes, cycloalkanes, cycloalkenes, aromatics, alcohols, aldehydes/ketones, ethers, esters, acids, furans	LANL	[30, 31]
CN	Alkanes, alkenes, alkynes, cycloalkanes, cycloalkenes, aromatics, alcohols, aldehydes/ketones, ethers, esters, acids, furans	NREL	[12]
CN	Alkanes, alkenes, alkynes, cycloalkanes, cycloalkenes, aromatics, alcohols, aldehydes/ketones, ethers, esters, acids, furans	RWTH Aachen University	[32]
CN	Alkanes, alkenes, alkynes, cycloalkanes, aromatics, hydrocarbon mixtures	KAUST	[43]
CN	Alkanes, cycloalkanes	Russian Academy of Sciences	[54]
CN	Alkanes, cycloalkanes, aromatics	Hokkaido University	[55]
CN	Alkanes, aromatics, hydrocarbon mixtures	University of South Carolina	[45]
CN	Cycloalkanes, n-heptane-cycloalkane mixtures	University of South Carolina	[48]
CN	Hydrocarbon mixtures	Princeton University,	[46]
CN	Hydrocarbon mixtures	Stanford University	[47]
RON/MON	Alkanes, alkenes, alkynes, cycloalkanes, cycloalkenes, aromatics, alcohols, aldehydes/ketones, ethers, esters, acids, furans	AIChE	[49]
RON/MON	Alkanes, alkenes, alkynes, cycloalkanes, aromatics	ASTM	[56]
RON/MON	TPRF mixtures	Saudi Aramco	[50]
RON/MON	TPRF mixtures	University of Cambridge	[51]
RON/MON	TPRF mixtures	Saudi Aramco	[39]
RON/MON	TPRF mixtures	KAUST	[52]
RON/MON	TPRF-ethanol mixtures	University of Melbourne	[53]
RON/MON	Hydrocarbon mixtures	KAUST	[29]
RON/MON	Hydrocarbon-ethanol mixtures	LLNL	[38]
RON/MON	Alkanes, alkenes, cycloalkanes, cycloalkenes, aromatics, ethanol and their mixtures	KAUST	[25]

164

165 **Table 3. Number of compounds of different chemical classes in the ignition quality database for model**

166 **training**

	Number of compounds (measured data)		
Compound class	CN	RON	MON
Alkanes	74	46	46
Alkenes and alkynes	35	73	74
Naphthenes	52	40	35
Aromatics	56	35	37
Total oxygenates	266	24	23
Alcohol	52	13	12
Aldehydes/Ketones	19	2	2
Saturated esters	66	3	3
Unsaturated esters	19	N/A	N/A
Ethers	66	6	6
Carboxylic acids	5	N/A	N/A
Polyfunctionals	39	N/A	N/A
Fuel mixtures	120	156	156
Total	603	374	371

167

2.3 Structural features extraction by group contribution method (GCM)

The group contribution method GCM-UOB 2.0 is based on the authors' recently published paper (GCM-UOB 1.0) [57]. The GCM-UOB 2.0 adds 9 functional group position descriptors (functional group type 1.1~1.9 in Figure 1) and 1 fuel reactivity descriptor (functional group type 1.10 in Figure 1) to account for the substituent positions on the phenyl group (functional group type 1.1~1.6, 1.10), naphthyl group (functional group type 1.1~1.8, 1.10 in Figure 1) and anthranyl group (functional group 1.1~1.10 in Figure 1). The introduction of these 10 functional group descriptors in GCM-UOB 2.0 significantly enhances the distinguishability of aromatics compared to GCM-UOB 1.0 [57]. Other functional group identifiers and fuel reactivity descriptors are already existing in GCM-UOB 1.0, so they remain unchanged in GCM-UOB 2.0 and the detail explanation can refer to ref. [57].

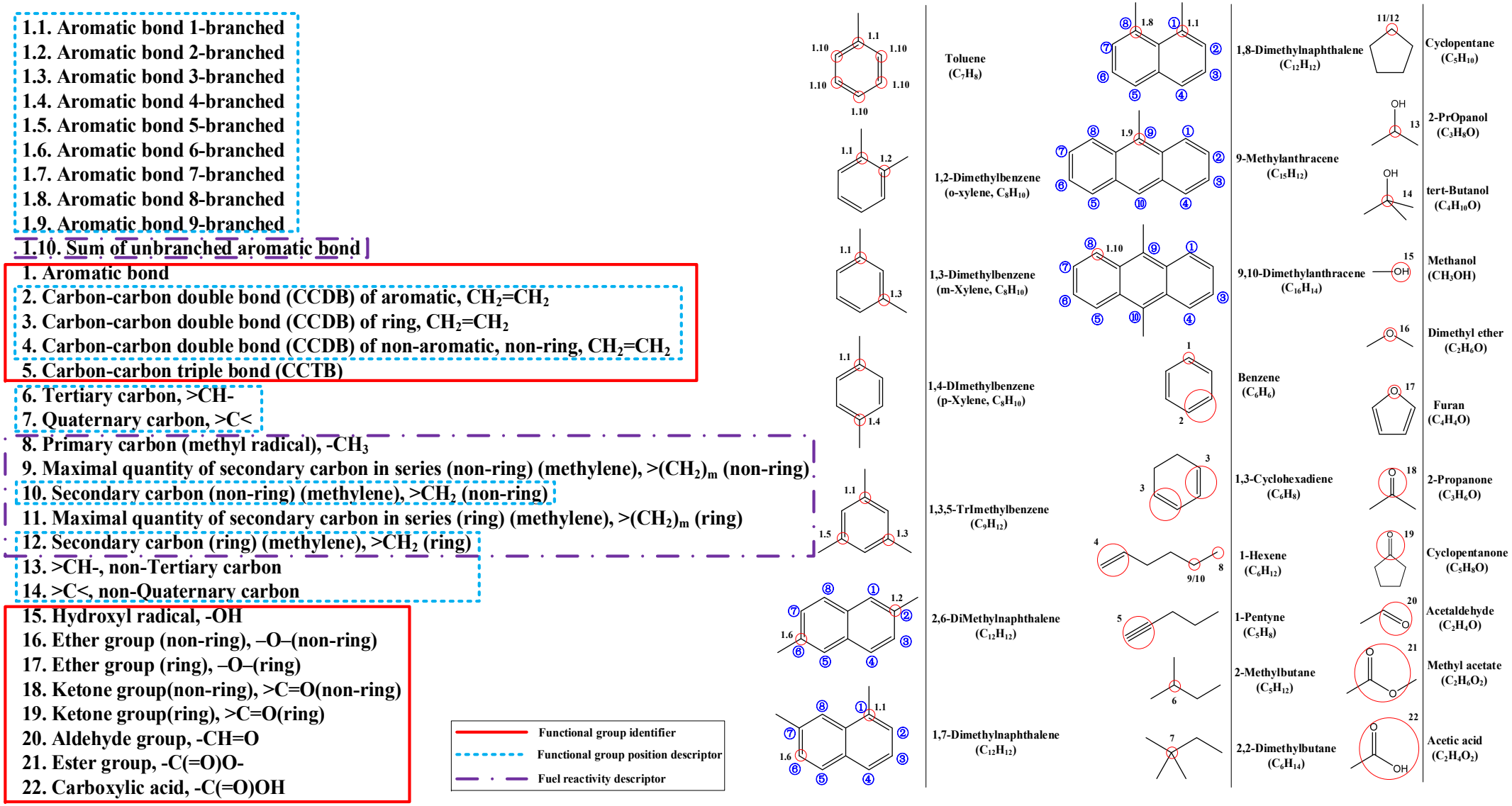

Figure 1. GCM-UOB 2.0 for structural features extraction. The functional groups are listed on the left, and an example of each functional group is circled in the molecular structure on the right.

2.4 Training and validation of machine learning regression model

The GCM-UOB 2.0 converts the fuel molecular structure into the molecular structure matrix and it is mapped to the ignition quality matrix by the machine learning regression model as shown in Figure 2. The functional relationship between molecular structure and ignition quality is challenging to quantify by a predetermined regression function, therefore, machine learning algorithms are needed to "learn" information directly from data. Machine learning regression has two implications: one is using regression model to describe the relationship between a response variable (output, in this work they are CN, RON, MON respectively) and predictor variables (input, in this work it the molecular structure matrix); and the other is using supervised learning (e.g. linear, generalized linear, nonlinear, and nonparametric techniques) to adaptively improve forecasting performance as increasing sample numbers for learning. The workflow for training regression models is as follow: (1) regression problem identification and data collection; (2) regression algorithm selection; (3) regression model training; (4) predictive performance assessment; (5) export regression model to predict new data. Pure compounds and full database (including both pure compounds and fuel mixtures, see Table 3) are used as model training dataset respectively. The regression model is developed by MATLAB regression learner module and 23 machine learning algorithms (see Table S3 in supporting information) are used to train the regression model in parallel. 10-fold cross validation is used to examine the predictive accuracy and prevent overfitting. The predictive accuracy is assessed by RMSE, mean absolute error (MAE), R-squared (R^2) and their equations can be found in section 4 of supporting information [58]. The Gaussian process regression algorithm obtains the minimal RMSE for CN/RON/MON and the corresponding models are selected as the optimal models in this work (the model detail is presented in Table S4 in supporting information).

The proposed method can predict CN, RON, MON of specific fuel compounds only if their molecular structures are known because it is based on functional group level instead of molecular level. For example, there

205 are no measured RON and MON for unsaturated esters in the training dataset (see Table 3), but they are
206 decomposed into type 4, 8, 9, 10 and 21 functional groups (see Figure 1) to obtain molecular structure matrix
207 (see Figure 2). The nonlinear relationship between these functional groups and ON is described by machine
208 learning regression model. Similarly, the proposed method can also apply to predict ON of carboxylic acids and
209 polyfunctionals even though experimental data is not available in the training dataset. The files of “Fuel ignition
210 quality database_Prediction” and“Fuel ignition quality database_ Training & Validation” in supplementary
211 material are the full database and the model training dataset (containing compounds with measured
212 CN/RON/MON only). They contain the predicted ignition quality for different types of fuel compounds.

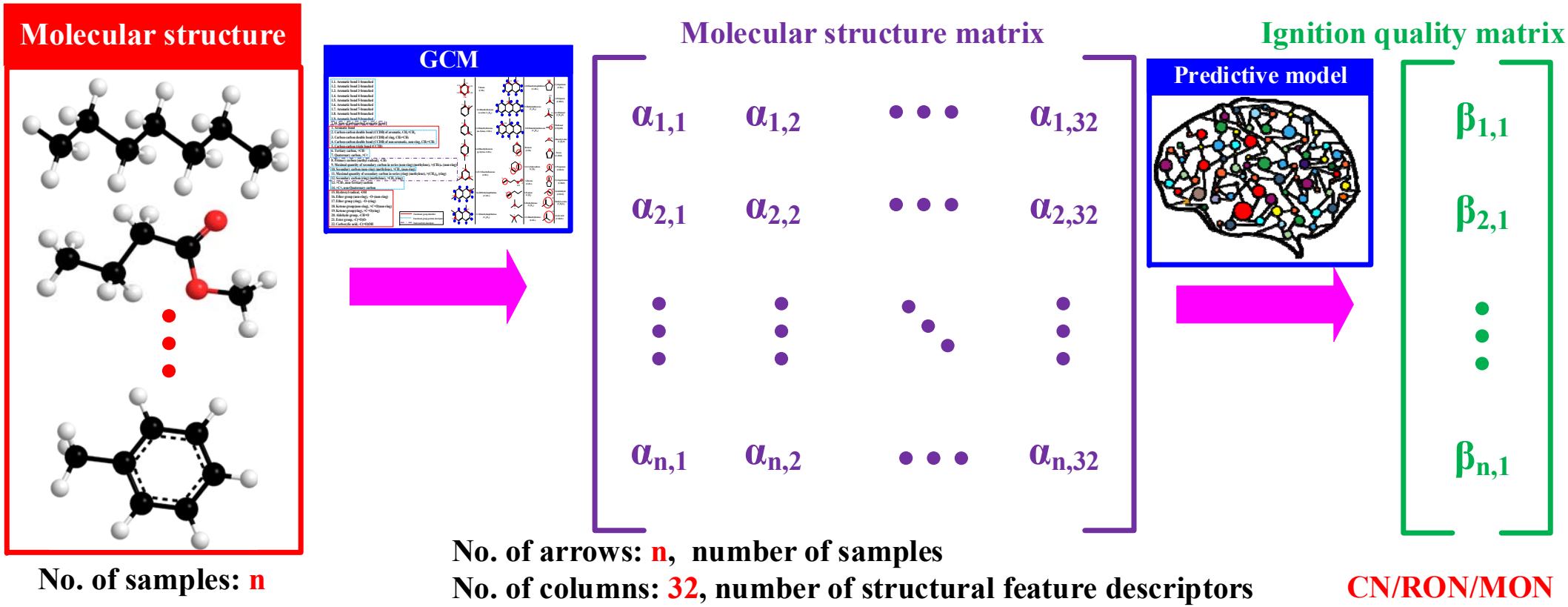

Figure 2. Flow chart of the CN/RON/MON prediction by coupling group contribution method and machine learning predictive model.

3. RESULTS AND DISCUSSION

This section discusses predictive accuracy of the proposed method and make a comparison with the published methods and neural network method (based on the same fuel ignition quality database, see section 5 in supporting information) since it is one of the most commonly used methods for ignition quality prediction.

3.1 Predictive accuracy of CN/RON/MON

3.1.1 Overall performance

The correlation coefficients of measured and predicted CN/RON/MON are shown in Figure 3 and the MAE, RMSE are demonstrated in Table 4. The regression models in the left and right columns of Figure 3 are trained by pure compounds dataset and full (pure compounds & mixtures) dataset respectively. The regression models trained by the full dataset reach higher predictive accuracy than those models trained by the pure compounds dataset because the former has additional 120, 156, 156 mixtures samples for CN/RON/MON (see Table 3). The availability of more data results in a better predictive model because the machine learning algorithms adaptively improve forecasting performance as increasing number of samples. Therefore, the machine learning regression models trained by full dataset are used in the following discussion by default. The regression models of CN/RON/MON obtain correlation coefficients of 0.9911, 0.9874, 0.9731 respectively which are superior to the published predictive methods (CN: 0.64~0.99, RON/MON: 0.92~0.99, see Table 1). Even though RON/MON predictive models proposed by Jameel et al. [25] can acquire R^2 up to 0.99, but oxygenates are excluded in their validated compounds. This study comprehensively validates against alkanes, alkenes, alkynes, cycloalkanes, cycloalkenes, aromatics, alcohols, aldehydes/ketones, ethers, esters, acids, furans and fuel mixtures. The trained regression models obtain low RMSE of 2.526, 2.454, 2.765 for CN, RON, MON respectively while the published methods have higher RMSE (CN: 3.38~9.1, RON/MON: 2.2~3.38, see Table 1). The RMSE results further support that the models trained by full dataset have higher predictive accuracy than those trained by pure compounds dataset as shown in Table 4. It should be noted that the measured CNs of propane (CAS 74-98-6)

239 [12], tert-butylbenzene (CAS 98-06-6) [30], 2,6-dimethylnaphthalene (CAS 581-42-0) [12],
240 1,3-diisopropylbenzene (CAS 99-62-7) [30], are -20, -1, -7, -7 (see Figure 3) and their ignition quality are out of
241 the calibrated range of ASTM D613 (CN: 30~65), ASTM D6890 (DCN: 31.5~75.1), ASTM D7668 (DCN:
242 30~70). Therefore, their CNs are indirectly measured by blending cetane number method (also known as
243 equivalent blending octane number). The $CN_{blending}$ parameter represents the autoignition quality of a specific
244 fuel compound when it is blended with a base fuel in particular volume fractions. The test fuel usually composes
245 of 10 vol.%, 20 vol.% or 30 vol.% of the binary mixture and the CN of test fuel is obtained by extrapolation [13,
246 15, 59, 60].

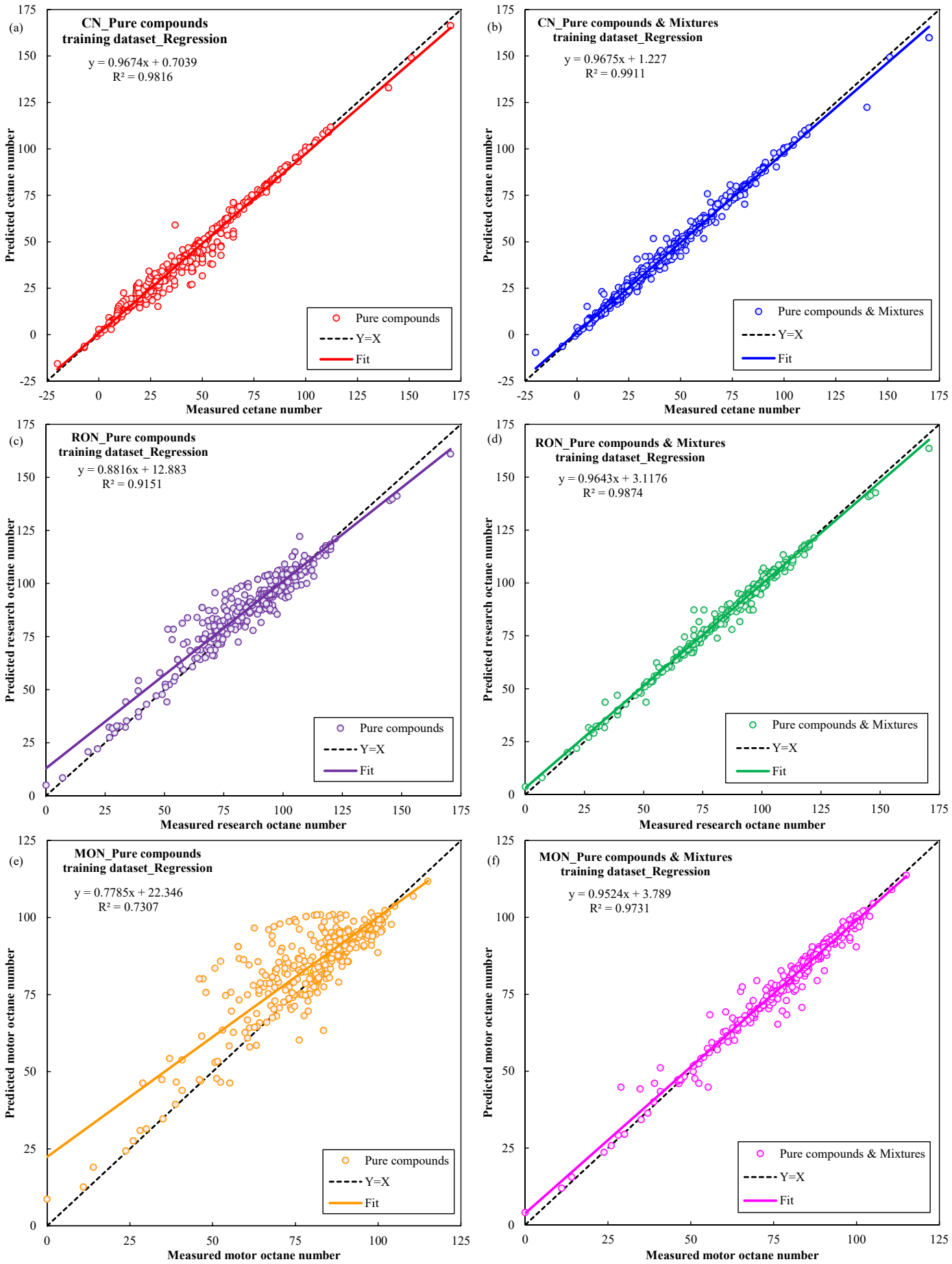

Figure 3. Parity plots for (a)-(b) CN, (c)-(d) RON, (e)-(f) MON between measured and predictive values by machine learning regression model. The regression models in the left and right columns are trained by pure compounds dataset and full (pure compounds & mixtures) dataset.

251 **Table 4. Statistical analysis of predictive performance for the machine learning regression models**

Property	Training dataset	R-squared	MAE	RMSE
CN	Pure compounds (483)	0.9816	1.891	3.580
	Pure compounds & Mixtures (603)	0.9911	1.460	2.526
RON	Pure compounds (217)	0.9151	4.543	6.795
	Pure compounds & Mixtures (373)	0.9874	1.386	2.454
MON	Pure compounds (215)	0.7307	6.500	9.922
	Pure compounds & Mixtures (371)	0.9731	1.567	2.765

252

3.1.2 Subgroup performance

Box-and-whisker plots in Figure 4 serve two functions: First, to assess the dispersion degree of forecast error for different compound groups. Second, to recognize the outliers that may contain great experimental error. The outliers are those more/less than 3/2 times of the upper/lower quartile. The maximum and minimum are the greatest and least values excluding outliers. The upper and lower quartiles denote 25% of data greater and less than the mean value. Median line and mean marker denote 50% of data greater than this value and mean of the selected data.

Figure 4 (a) ~ Figure 4(c) support that the predictive accuracy of CN mixture regression model is superior to those of RON and MON because the predictive residuals of CN are more concentrated around zero. The CN/RON/MON for different compound groups are successfully predicted by the machine learning regression models as most of correlation coefficients exceed 0.98 (see Table 5). The predicted performances of CN and MON of naphthenes are relatively low (R^2 of CN: 0.9599 and R^2 of MON: 0.9504) which indicate that additional functional group position descriptors for cycloalkanes should be added into the functional group classification system (see Figure 1) to increase the distinguishability. The R^2 of RON and MON for olefins and alkynes are low (R^2=0.9279 and 0.8839 for RON and MON respectively, see Table 5) because both have 11 outliers (see Figure 4). The gaps between upper and low quartiles of alkynes in RON and MON are abnormally disperse as shown in Figure 4 (b) and Figure 4 (c) because there are only 4 measured RONs and 2 measured MONs. It is necessary to further study the ignition quality of alkynes to enrich the training database and increase the model predictive capacity. The ketones, esters, ethers, furans also need more measured RON and MON as well. The proposed method has good predictive capacity on fuel mixtures that the R^2 of CN, RON, MON reach 0.98, 0.9982, 0.9908 respectively as shown in Table 5. Given that the ignition quality database contains 120, 156, 156 samples of CN, RON, MON, more mixture CN data is required to further improve its predictive accuracy.

All outliers recognized by the proposed method are provided in the supporting information, both the

measured and predictive ignition quality data are also enclosed. But the outliers probably come from measurement error rather than the predictive error and repeatability and reproducibility tests are needed for these compounds. For example, the CN value of 3,3-dimethylpentane reported by Lapidus et al. [54] is -10.3, then the CN predictive residual is abnormally high (29.59) as shown in Figure 4 (a) and Figure 4 (d). The predictive CN by machine learning regression model and the neural network model (see section 5 in supporting information) of 3,3-dimethylpentane are 19.29 and 20.36 respectively. Therefore, the nominal CN of 3,3-dimethylpentane should be around 19~20.5 instead of -10.3 reported by Lapidus et al. [54]. Another example, the RON of 6-methyl-2-heptene reported by Kubic et al. [30, 31] is 71.3, but the predictive values by machine learning regression model and neural network model (see section 5 in supporting information) are 87.27 and 87.19 respectively. The reported RON of 6-methyl-2-heptene may contain huge measurement error and uncertainty.

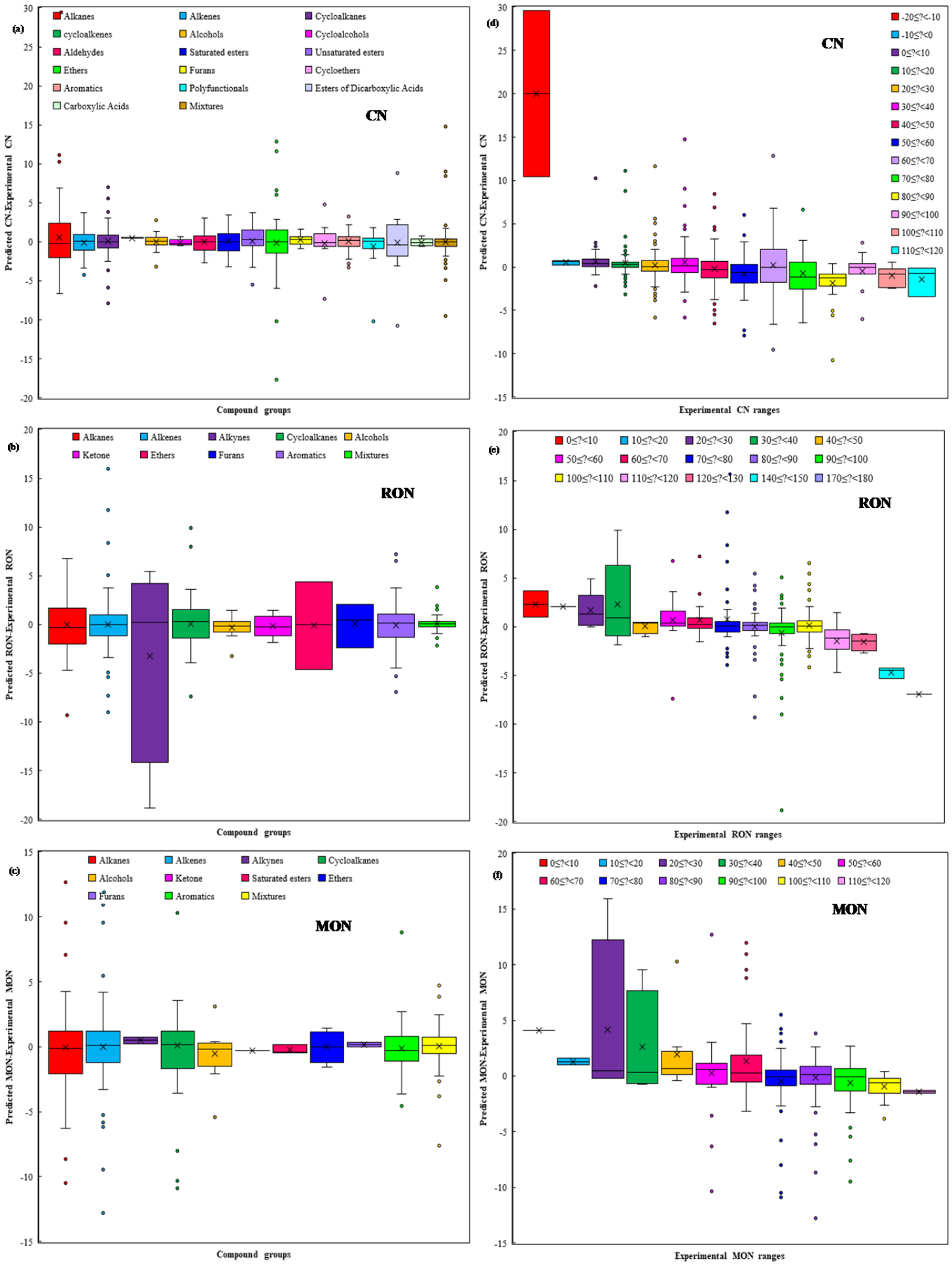

Figure 4. Predictive residuals for typical compound groups and within specified ranges for (a), (d) CN; (b), (e) RON; (c), (f) MON.

290 **Table 5. Comparison of correlation coefficients of different compound groups between current study and published methods**

R-squared	CN					RON			MON		
Compound class	Current	Saldana et al. [33]	Kubic et al. [31]	DeFries et al. [35]	Dahmen et al. [32]	Current	Kubic et al. [31]	Albahri [61]	Current	Kubic et al. [31]	Albahri [61]
Paraffins	0.9866	N/A	0.91	0.73	0.53	0.9902	0.94	0.86	0.9765	0.95	0.87
Olefins and alkynes	0.992	N/A	0.90	N/A	-0.48	0.9279	0.90	0.53	0.8839	0.65	-1.55
Naphthenes	0.9599	N/A	0.81	N/A	0.25	0.9839	0.85	0.75	0.9504	0.89	-0.40
Aromatics	0.9946	N/A	0.87	0.44	0.58	0.9896	0.76	-3.28	0.9722	N/A	N/A
Oxygenates	0.993	N/A	0.85	N/A	0.41	0.9821	0.62	N/A	0.9767	0.56	N/A
Alcohol	0.9977	N/A	N/A	N/A	N/A	0.9945	N/A	N/A	0.9433	N/A	N/A
Aldehydes/Ketones	0.9973	N/A	N/A	N/A	N/A	1	N/A	N/A	1	N/A	N/A
Saturated esters	0.9956	N/A	N/A	N/A	N/A	0.9977	N/A	N/A	0.9991	N/A	N/A
Unsaturated esters	0.9881	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Ethers	0.9905	N/A	N/A	N/A	N/A	0.9414	N/A	N/A	0.9943	N/A	N/A
Carboxylic acids	0.9996	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Polyfunctionals	0.9937	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Fuel mixtures	0.98	N/A	N/A	N/A	N/A	0.9982	N/A	N/A	0.9908	N/A	N/A
Overall	0.9911	0.934	0.90	0.64	0.53	0.9874	0.93	0.55	0.9731	0.91	-1.16

291

3.2 Method application: impact of fuel molecular structure on ignition quality

The machine learning regression based method enables to provide an insight into the impact of fuel molecular structure on ignition quality. Fuels with measured RON/MON are usually short carbon chain (C2~C10) molecules and the isomerization implements significant influence on the values. On the opposite, the fuels with measured CN are usually long carbon chain molecules (C4~C28), the impact of isomerization on CN weakens as increasing carbon chain length. RON metric is adopted to characterize the ignition quality for alkanes, alkenes, naphthenes, aromatics, alcohols, ethers, esters in section 3.2.1~3.2.7 because changing branching degree has more pronounced effect on ON than CN. CN is used to characterize the ignition quality of esters in section 3.2.8 since more experimental data are available.

3.2.1 Comparison of ignition quality for different fuel types

The straight chain alkane, alkene, alcohol, ether, aldehyde, ketone and ester with 5 carbon atoms are taken as an example to understand the ignition quality of different fuel types. Results in Figure 5 demonstrate that the predicted RONs of different group compounds rank from high to low as: ester (109.42) > ketone (106.35)> alkene (95.06) > ether (90.75) > alcohol (78.29) > alkane (60.21) > aldehyde (57.28). Only n-pentane (61.8 vs 60.21) and 1-pentanol (78 vs 78.29) exist measured RON and they are in good agreement to the predicted values. Therefore, the proposed method provide an effective tool to predict and compare the ignition quality of traditional and emerging fuels even for those no experimental data available. Results also indicate that all oxygenated chemical compounds, except aldehyde, have higher RON than the counterpart straight chain alkane. The RONs of unsaturated alkenes are greater than the corresponding straight chain alkanes as well [56, 62]. The order of RON discussed above may vary with carbon chain length, branching degree, functional group positions.

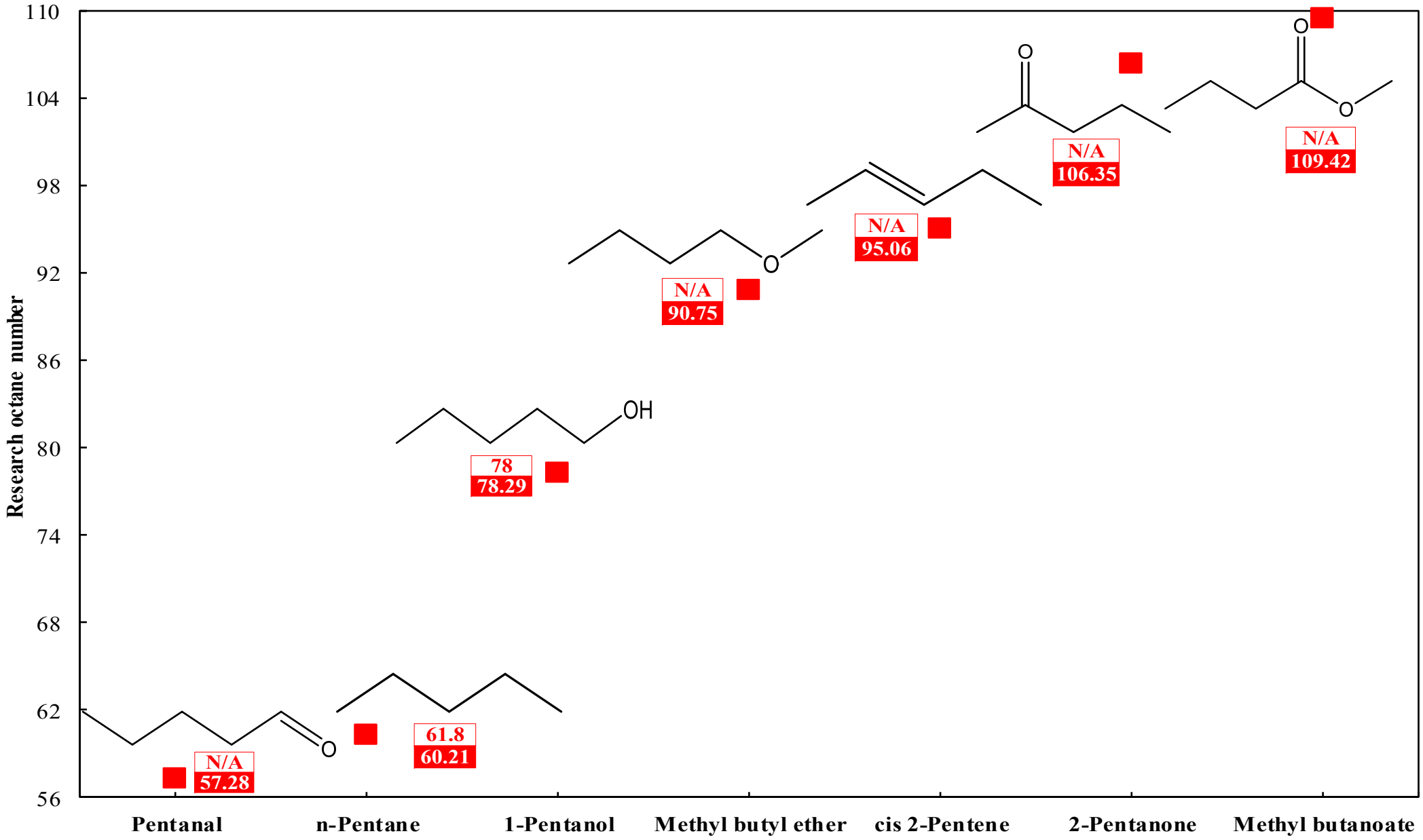

Figure 5. RON of different fuel types, numbers with red frames and red backgrounds are measured values and predictive values.

315 **3.2.2 Impact of alkanes structural features on ignition quality**

316 Alkanes are the main constituents of commercial diesel and gasoline and it the best studied class of
317 compounds [63-65]. The complete RON picture of C1~C8 alkanes as a function of carbon atom numbers is
318 shown in Figure 6. The pink lines direct toward bottom right corresponding to increase carbon chain length. The
319 blue lines direct toward upper right indicate adding methyl group into fuel molecule other than terminal position.
320 The green lines direct vertical upwards corresponds to centralization of fuel molecules. The ignition quality of
321 alkanes can be affected in three ways by modifying the molecular structure: First, increasing carbon chain length
322 decreases RON. For example, the predicted RONs of 2-methylpropane (102.67), 2-methylbutane (92.5),
323 2-methylpentane (72.36), 2-methylhexane (42.7), 2-methylheptane (21.83) decrease as increasing carbon chain
324 length. Second, increasing branching degree raise the RON. The impact of increasing branching/centralization
325 degree on RON for octane isomers is visualized in Figure 7. For example, the predicted RONs of n-octane
326 (8.53), 2, 5-dimethylhexane (68.56), 2, 3, 4-trimethylpentane (103.27), tetramethylbutane (118.54) increase
327 progressively as increasing branching degree. Third, increasing centralization degree (moving methyl group
328 toward the center of the molecule) increases RON. From example, the predictive RON of n-octane (8.53),
329 2-methylheptane (21.83), 3-methylheptane (27.06), 4-methylheptane (31.6) as the methyl group moving toward
330 center position. The impact of increasing branching/centralization degree on ignition quality is opposite to
331 lengthening carbon chain length [56, 62].

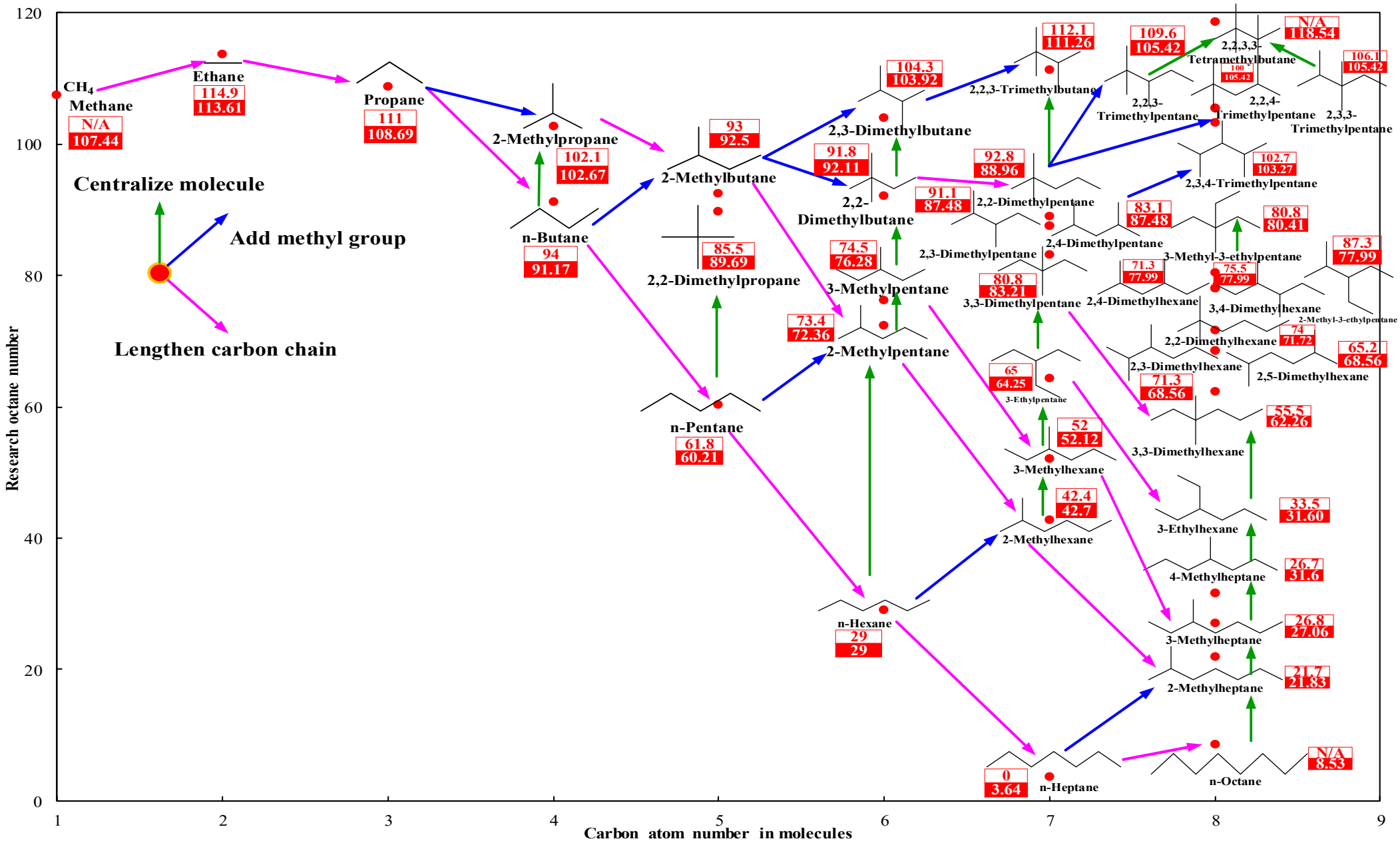

Figure 6. RON of C1~C8 alkanes, numbers with red frames and red backgrounds are measured values and predictive values.

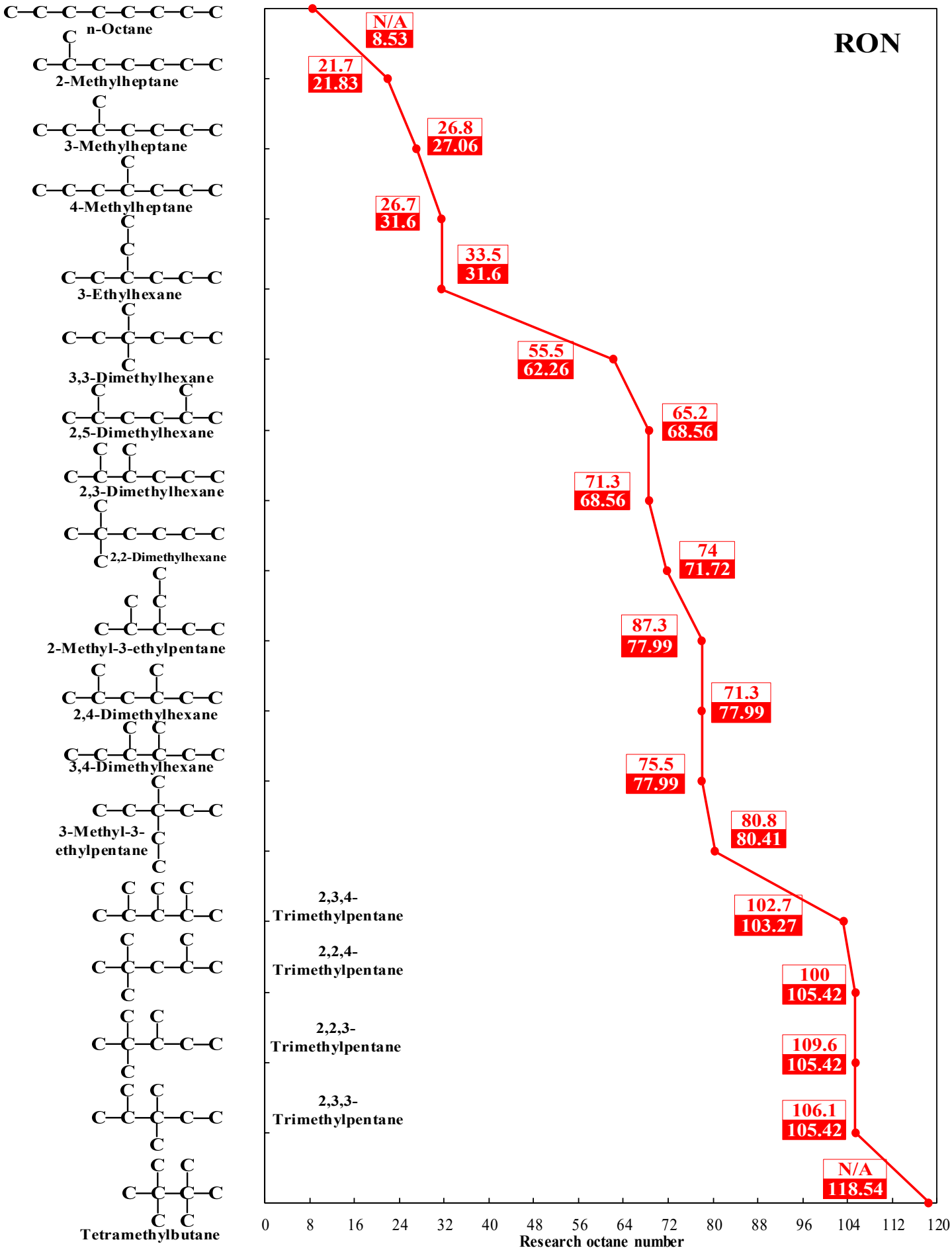

Figure 7. Impact of branching/centralization on RON of C8 alkanes, numbers with red frames and red backgrounds are measured values and predictive values.

3.2.3 Impact of alkenes structural features on ignition quality

Alkenes are unsaturated hydrocarbons and good octane components of gasoline [66]. The influence of the position of carbon-carbon double bond on the straight chain alkenes is examined by comparing with the corresponding alkanes as shown in Figure 8. The results are summarized below: First, the RON decreases consistently as increasing carbon chain length similar to n-alkanes but the decline is less severe. As a consequence, the alkenes lighter than 1-butene have lower RON than the corresponding alkanes while those heavier than 1-butene have higher RON. Therefore, replacing C4 or above straight chain alkanes with corresponding straight chain alkenes can increase the RON which is useful for advanced gasoline compositions design [56]. Second, the RON increasing magnitudes vary with the double bond position in the straight chain alkenes and the centralization of double bond increases the RON. In other words, the closer the carbon-carbon double bond to the center of molecule, the greater the RON. For example, the predicted RONs of 1-nonene (30.83), 2-nonene (42.39), 3-nonene (50.85) and 4-nonene (56.96) increase progressively as the double bond moving toward molecular center as shown in Figure 8.

The RONs of branched chain aliphatic alkenes and the counterpart alkanes are plotted in Figure 9. The green arrows direct from alkanes to alkenes. The introduction of double bond to form a branched chain alkenes do not always lead to a higher RON than the corresponding branched alkanes which depends on the position and centralization degree of the double bond. For example, the predicted RONs of 2-methyl-2-hexene (92.36) is higher than 2-methylhexane (42.71) while the 2,3-dimethyl-2-butene (98.1) is lower than 2,3-dimethylbutane (103.92) as shown in Figure 9. In general, the RON of branched chain alkene and the corresponding alkane does not exist a clear relationship.

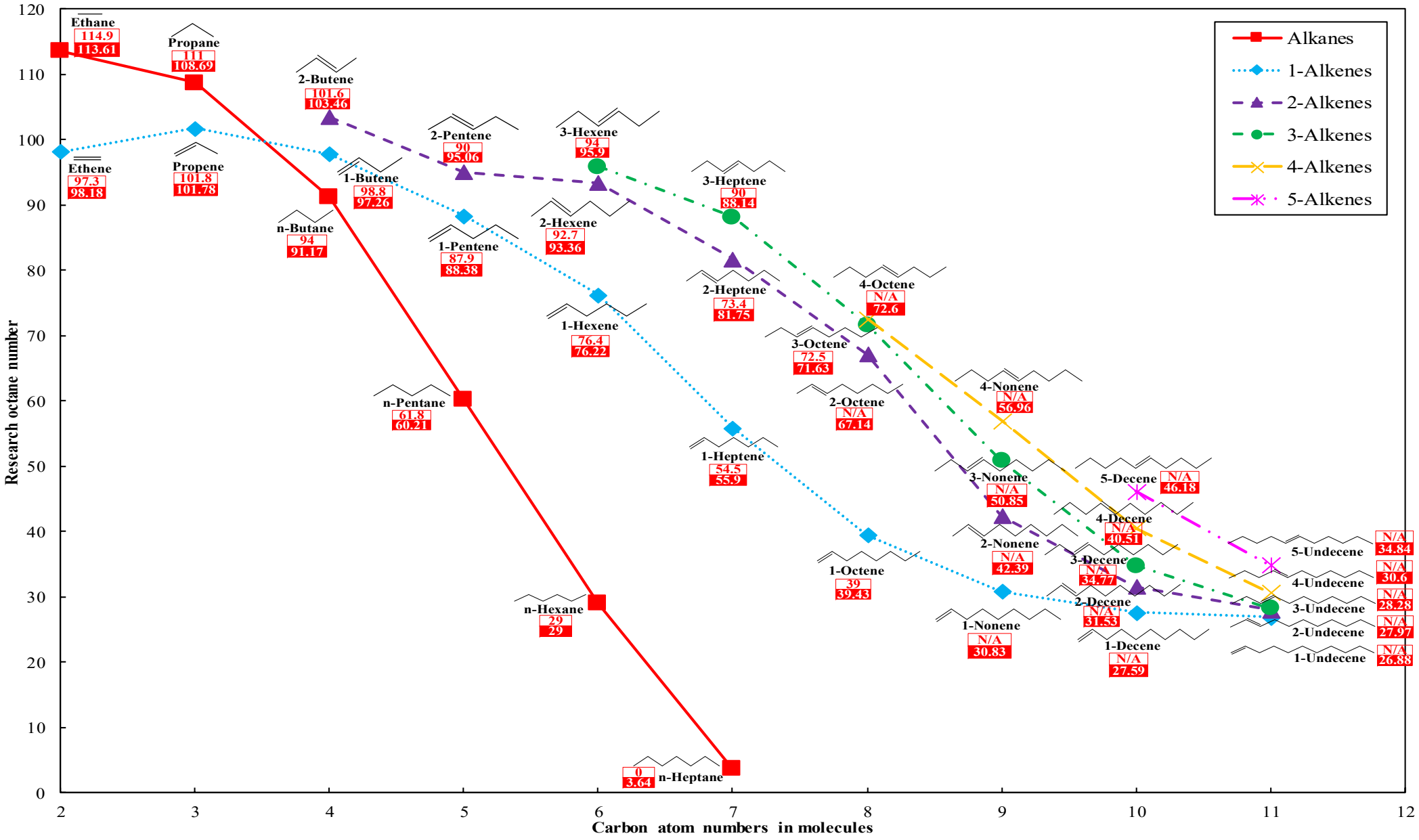

Figure 8. RON of C2~C11 straight chain alkenes, numbers with red frames and red backgrounds are measured values and predictive values.

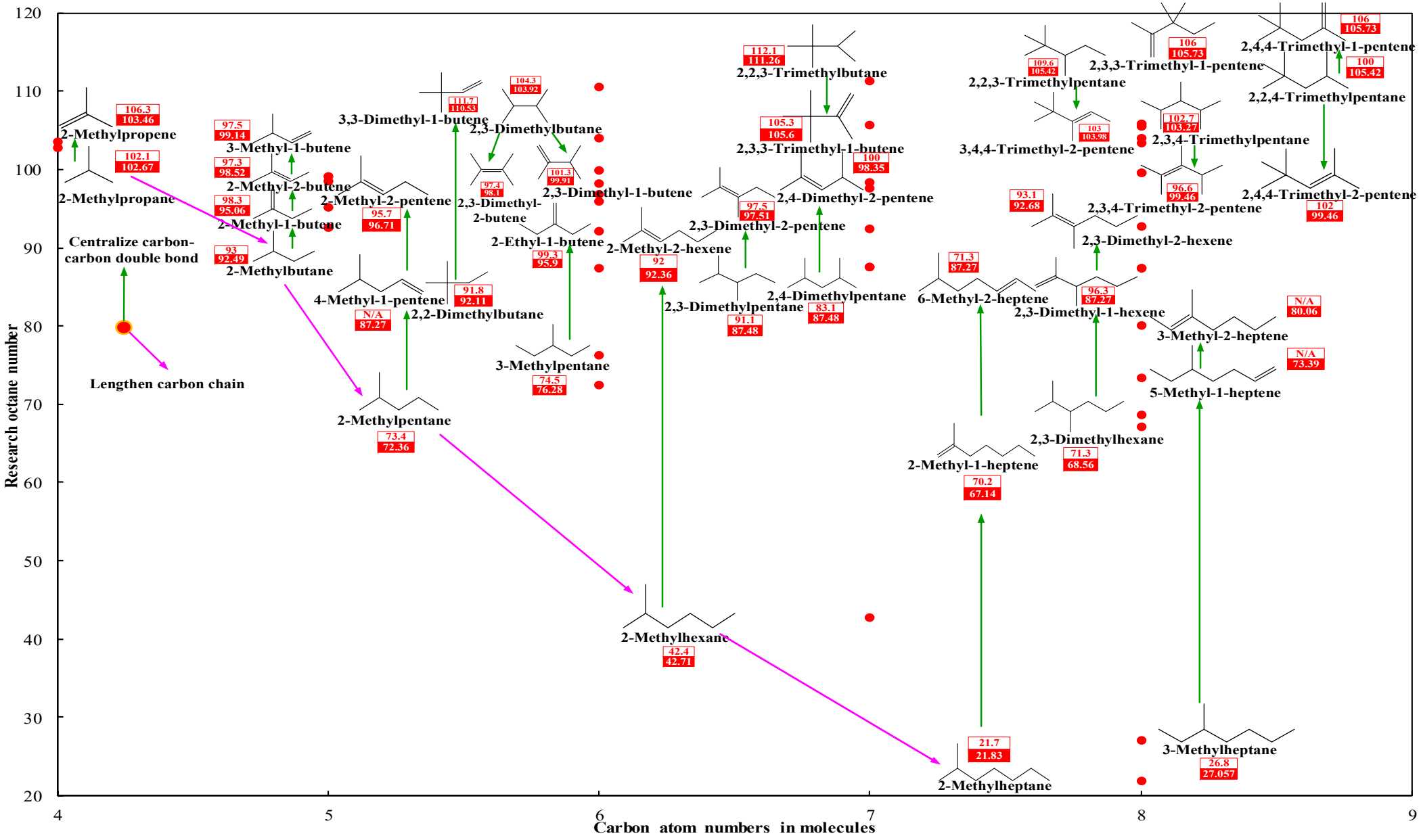

Figure 9. RON of C4~C8 branched chain alkenes, numbers with red frames and red backgrounds are measured values and predictive values.

3.2.4 Impact of naphthenes structural features on ignition quality

Only few measured RON/MON available for alkyl cycloalkanes [56], thus it is not possible to summarize the regularity due to the lack of experimental data and the generation of predicted data is required. Ignition quality studies on cyclic hydrocarbons are also presented as critical compression ratio [62] and aniline equivalent [67]. The predicted RONs of alkyl cyclopropane, alkyl cyclopentane, alkyl cyclohexane are plotted for various carbon atom numbers in Figure S2, Figure S3 in supporting information and Figure 10 and the generalizations are summarized. First, the RON increases as decreasing the ring size of cycloalkanes provided that they have identical substituted groups. For example, the predicted RON of methylcyclohexane (74.05), methylcyclopentane (88.45), methylcyclopropane (95.09) increases progressively as the ring size decrease as 6, 5, 3. Second, the addition of straight chain side group into the cycloalkane ring decreases the RON. The RON reduces as increasing side chain length and increases as branching side chain. For example, the predicted RON drops from 74.05 of methylcyclohexane to 13.13 of butylcyclohexane as increasing side chain length while the RON of tert-butylcyclohexane increases to 95.84 as branching side chain. Third, the distribution of one single side chain into several separated side chains increases the RON and the side chain position affects the RON. The more closer and compact of the substituents, the greater the RON of the molecule. The fuel molecule reaches the greatest RON as the two side chains connect on the same ring carbon. For example, the predicted RONs of ethylcyclohexane (46.93), 1,4-dimethylcyclohexane (68.43), 1,3-dimethylcyclohexane (71.36), 1,2-dimethylcyclohexane (80.16), 1,1-dimethylcyclohexane (87.64) increase progressively as two side chains become closer. Polysubstituted cycloalkanes of 1,2,3,5-tetramethylcyclohexane (85.63), 1,2,3,4-tetramethylcyclohexane (86.05), 1,1,3,5-tetramethylcyclohexane (89.64), 1,1,4,4-tetramethylcyclohexane (100.14) follow the same rule. In summary, reducing the ring size of cycloalkane, separating single substituent into multiple side chains and further compacting the substituted groups on the ring are the main methods to increase ON of naphthenes.

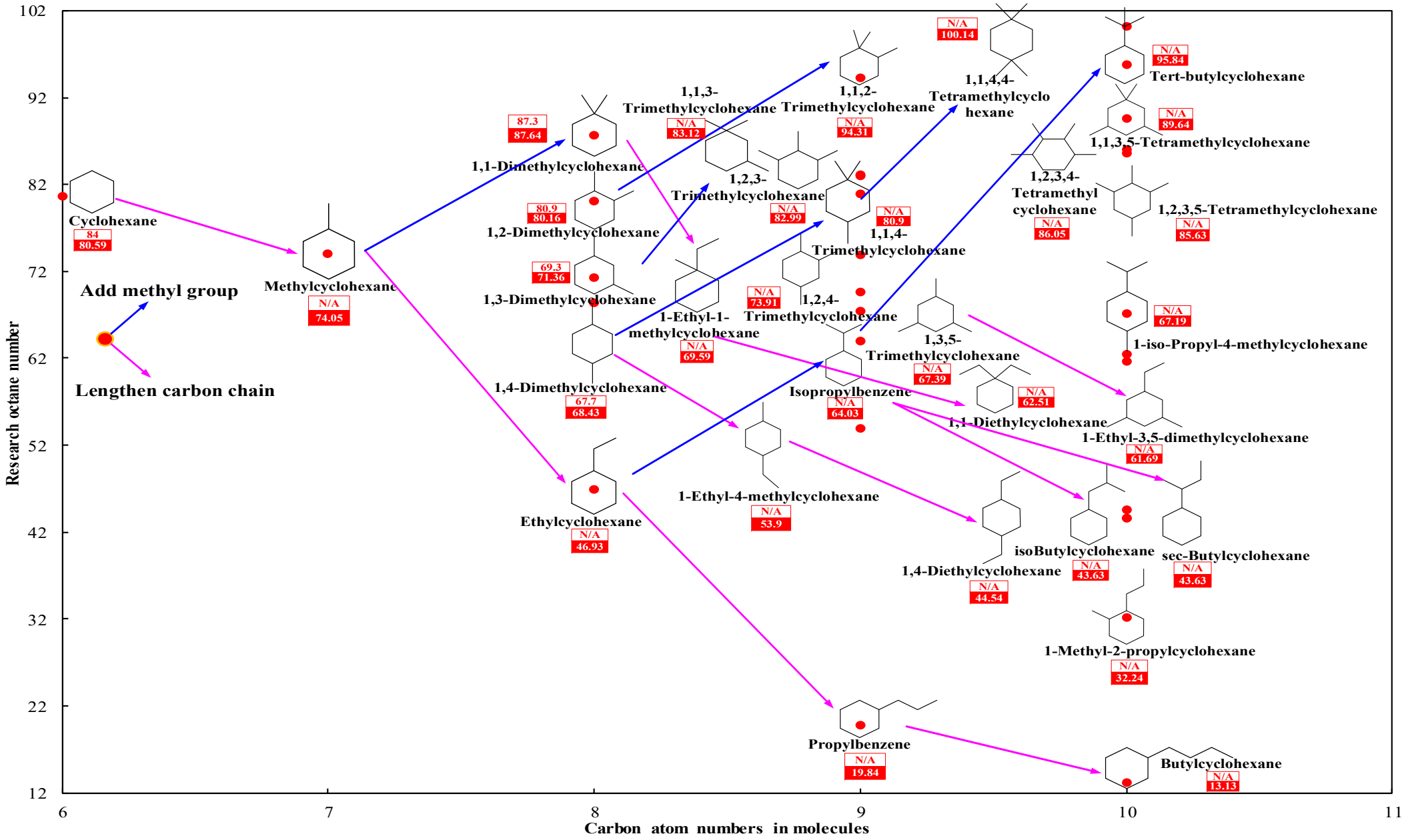

Figure 10. RON of C6~C10 alkyl cyclohexanes, numbers with red frames and red backgrounds are measured values and predictive values.

3.2.5 Impact of aromatics structural features on ignition quality

Aromatic hydrocarbons are the second most constituents in commercial gasoline (comprising up to 35vol.%) [66], but it is challenging to rate its ON because they are resistant to autoignition and some of them exceed the ON calibrated range (RON: 120.3, MON: 120, see Table S1 in supporting information) [56]. As a consequence, the measured RONs of aromatic hydrocarbons may be subjected to greater experimental error and uncertainty. Generating predicted data help to extrapolate the ON scale for high anti-knock fuel compounds like aromatics. The impact of the substituents of phenyl group on the RON is shown in Figure 11 and four important generalizations is obtained. First, for phenyl group with one substituent, increasing side group carbon chain length reduces the RON while the branching side chain plays an opposite role. For example, the predicted RONs of toluene (117.33), ethylbenzene (107.62), propylbenzene (101.17), butylbenzene (97.91) decrease progressively as increasing substituent carbon chain length while the RONs of butylbenzen (97.91), isobutylbenzene (99.34), tert-butylbenzene (104.43) increase as substituent branching. Second, for phenyl group with two substituents, the RON of compounds ranks in order from highest to lowest: para-isomers ≈ meta-isomers > ortho-isomers [56, 62]. For example, the predicted RONs of p-xylene, m-xylene and o-xylene are 141.49, 140.79 and 118.2 respectively. Third, for phenyl group with three substituents, the better the symmetry of the substituents, the greater the RON of the fuel molecule. As shown in Figure 11, the predicted RONs of 1, 2, 3-trimethylbenzene (117.15), 1, 2, 4-trimethylbenzene (142.68), 1, 3, 5-trimethylbenzene (163.65) increase consistently as substituents becoming more symmetric. Fourth, the distribution of the carbon atoms of a single substituted group into multiple substituted groups increases the RON. For example, the predicted RONs of propylbenzene (101.17), 1-ethyl-3-methylbenzene (105.54), 1, 3, 5-trimethylbenzene (163.65) increase progressively as increasing number of side groups. From the perspective of developing high anti-knock gasoline, short side chain aromatics (such as toluene) or multiple short side chains aromatics (dimethyl benzene, trimethyl benzene and tetramethyl benzene) significantly increases gasoline anti-knock quality.

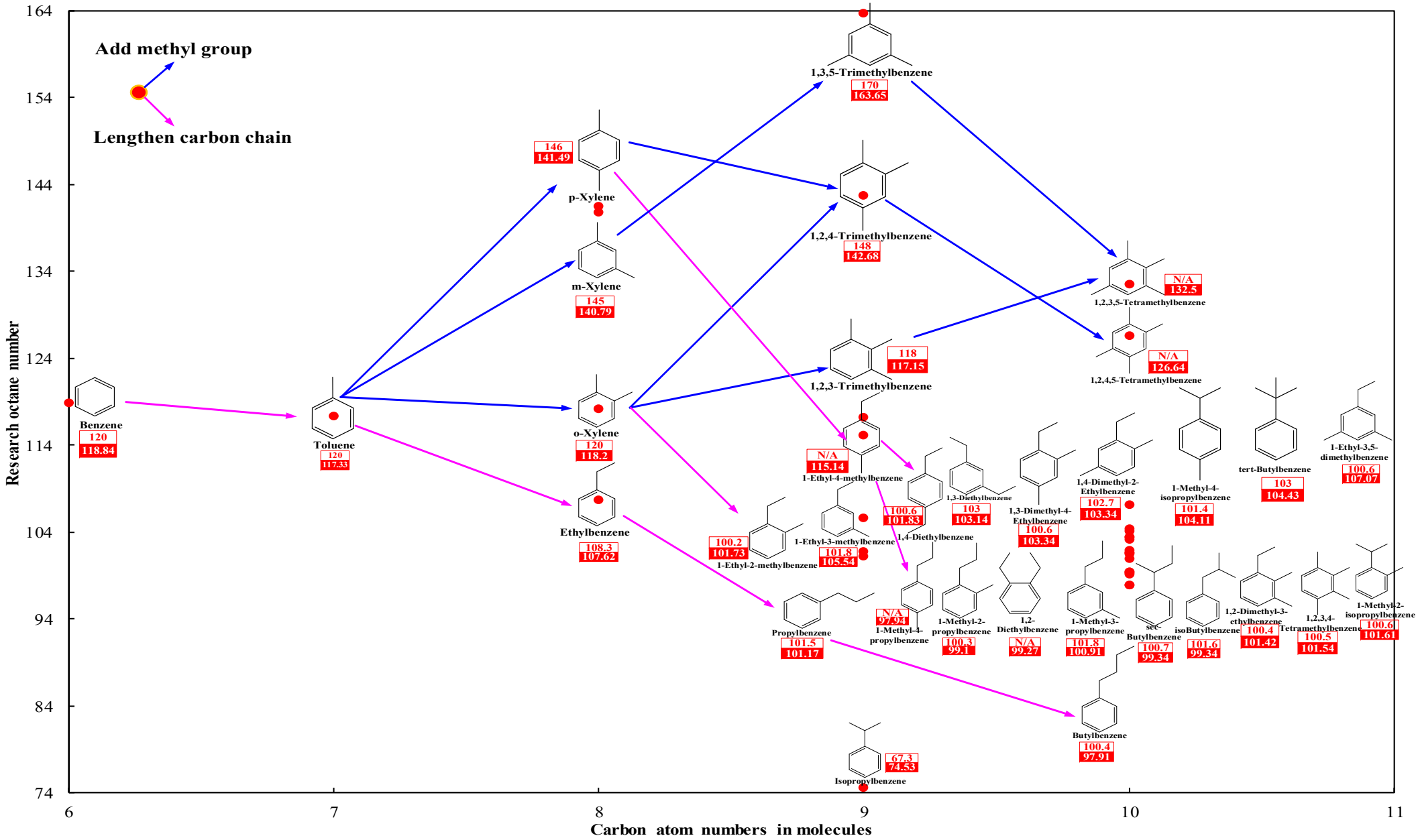
Add methyl group
Lengthen carbon chain
Research octane number
Carbon atom numbers in molecules
164
154
144
134
124
114
104
94
84
74
6
7
8
9
10
11
Benzene
120
118.84
Toluene
120
117.33
Ethylbenzene
108.3
107.62
p-Xylene
146
141.49
m-Xylene
145
140.79
o-Xylene
120
118.2
1,3,5-Trimethylbenzene
170
163.65
1,2,4-Trimethylbenzene
148
142.68
1,2,3-Trimethylbenzene
118
117.15
1,2,3,5-Tetramethylbenzene
N/A
132.5
1,2,4,5-Tetramethylbenzene
N/A
126.64
Isopropylbenzene
67.3
74.53
Propylbenzene
101.5
101.17
Butylbenzene
100.4
97.91
1-Ethyl-4-methylbenzene
N/A
115.14
1-Ethyl-3-methylbenzene
101.8
105.54
1-Ethyl-2-methylbenzene
100.2
101.73
1,4-Diethylbenzene
100.6
101.83
1,3-Diethylbenzene
103
103.14
1,3-Dimethyl-4-Ethylbenzene
100.6
103.34
1,4-Dimethyl-2-Ethylbenzene
102.7
103.34
1-Methyl-4-isopropylbenzene
101.4
104.11
tert-Butylbenzene
103
104.43
1-Ethyl-3,5-dimethylbenzene
100.6
107.07
1-Methyl-4-propylbenzene
N/A
97.94
1-Methyl-2-propylbenzene
100.3
99.1
1,2-Diethylbenzene
N/A
99.27
1-Methyl-3-propylbenzene
101.8
100.91
sec-Butylbenzene
100.7
99.34
isoButylbenzene
101.6
99.34
1,2-Dimethyl-3-ethylbenzene
100.4
101.42
1,2,3,4-Tetramethylbenzene
100.5
101.54
1-Methyl-2-isopropylbenzene
100.6
101.61

Figure 11. RON of C6~C11 aromatic hydrocarbons, numbers with red frames and red backgrounds are measured values and predictive values.

3.2.6 Impact of alcohols structural features on ignition quality

Ignition propensity of alcohols vary significantly with different molecular structures (i.e. carbon number and substitutions), however, the regularity is not well understood [68]. C1~C5 alcohols usually blend with commercial gasoline due to low reactivity [69] while C8 and larger alcohols can blend with diesel due to less resistant to autoignition. The RONs of C1-C5 alcohols are shown in Figure 12 and the generalized regularities are summarized. First, the RONs of straight chain alcohols reduce as increasing carbon chain length similar to alkanes. For example, the predicted RONs of methanol (121.3), ethanol (110.83), 1-propanol (102.82), 1-butanol (94.81), 1-pentanol (78.3) reduce progressively as increasing carbon chain length. Second, the isomerization of straight chain alcohols boosts the RON but high branching degree does not always result in high RON. For example, the predicted RONs of 1-butanol (94.81), 2-methyl-1-propanol (106.53), 2-butanol (108.19) increase progressively as increasing branching degree, but the RON of tert-butanol drops down to 106.84 as increasing chain branching degree.

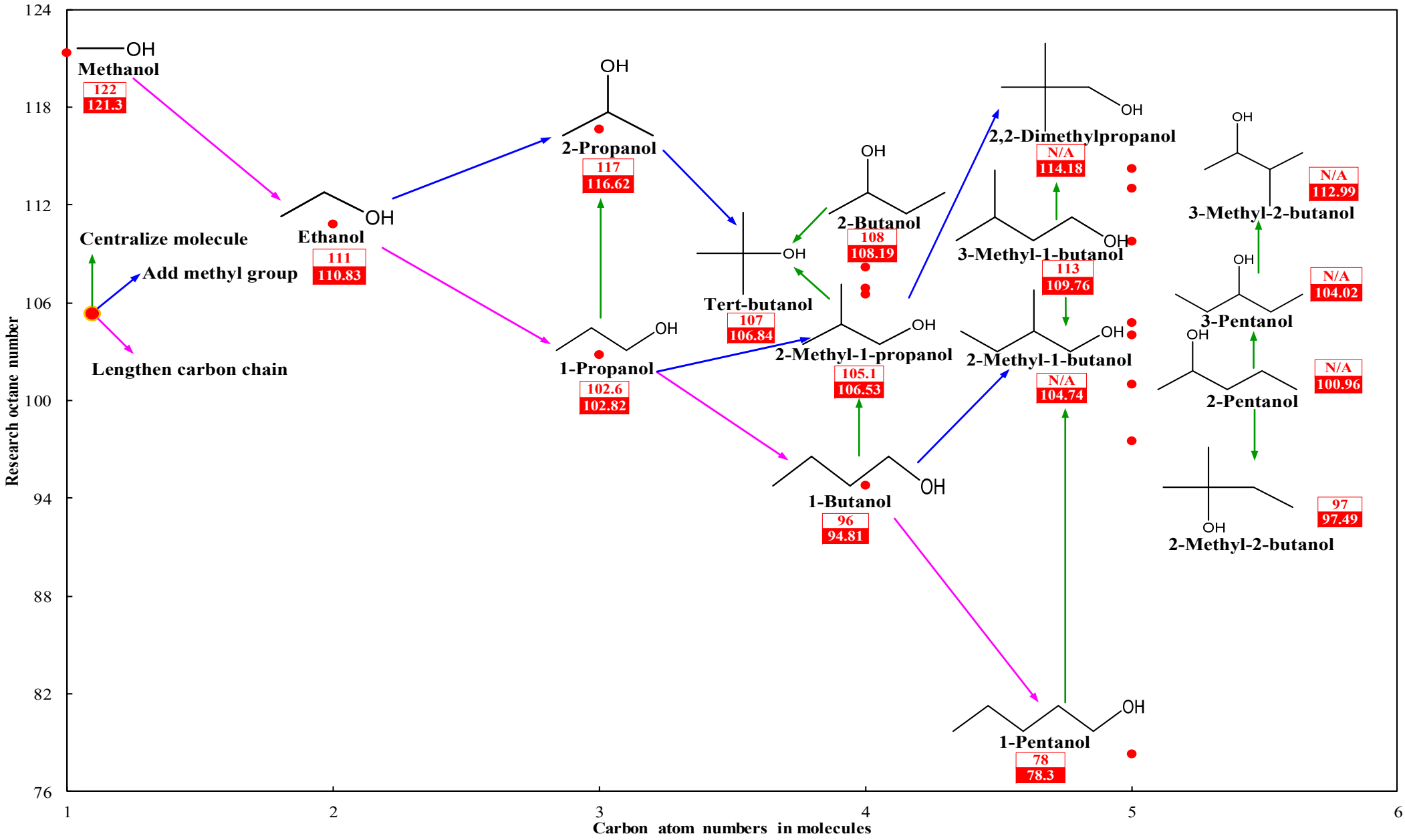

Figure 12. RON of C1~C5 alcohols, numbers with red frames and red backgrounds are measured values and predictive values.

3.2.7 Impact of ethers structural features on ignition quality

The oxygen atom in an ether (R1-O-R2) interrupts and divides the carbon chain into two alkyl chains (R1 and R2), therefore, the impacts of ethers structural features (increasing one side/both sides carbon chain, changing branching degree or forming ring) on RON are demonstrated in a different manner as shown in Figure 13. The data implications are summarized below: First, the RON decreases progressively with increasing carbon chain length for both symmetric and asymmetric ethers, but the former have greater RON than corresponding asymmetric ethers with the same number of carbon atoms. In other words, the RONs of straight chain ethers increase as oxygen atom moving toward the center of fuel molecule. For example, the predicted RONs of dimethyl ether (111.25), diethyl ether (100.57), dipropyl ether (89.69), dibutyl ether (79.45), dipentyl ether (70.54) decrease progressively as prolonging chain length at both sides. These symmetric ethers have higher RONs than the corresponding asymmetric ethers of methyl propyl ether (97.67), methyl pentyl ether (84.03), methylheptyl ether (72.05) and 1-methoxydecane (59.23). Second, increasing the carbon chain branching degree raises the RON. For example, the predicted RONs of methyl pentyl ether (84.03), 2-methoxypentane (95.53), 1-methoxy-2-methylbutane (98.42), 2-methoxy-3-methylbutane (107.06), tert-amyl methyl ether (109.79) increase progressively as increasing branching degree. However, the GCM-UOB 2.0 cannot distinguish the butane, 1-methoxy-3-methyl-, 2-methoxypentane, 3-methoxypentane because they have the same number of functional group types of 6, 8, 9, 10, 16. The distinguishability of functional group position descriptor and fuel reactivity descriptor in current GCM should be further improved.

The symmetric long carbon chain (C8 or above) ethers and the short carbon chain (C4 or less) ethers can be used as CN improver and ON booster respectively. All studied ethers in Figure 13 do not have measured CN/RON/MON data except dimethyl ether, therefore, more experimental data is needed to enrich the ignition quality database and refine the machine learning regression model.

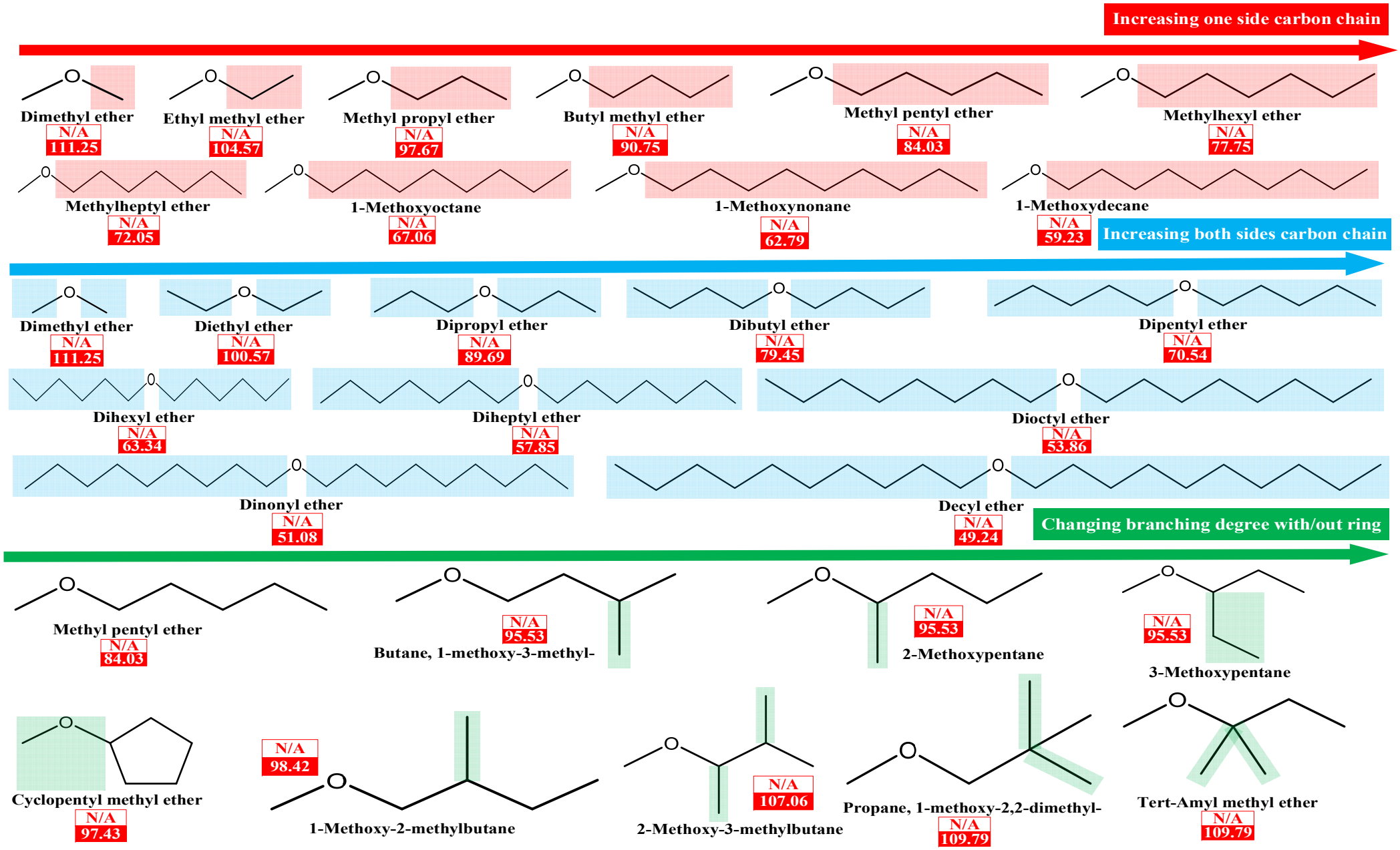

Figure 13. RON of typical ethers, numbers with red frames and red backgrounds are measured values and predictive values.

3.2.8 Impact of esters structural features on ignition quality

The long chained fatty acid methyl esters (FAME) are the major constituents of biodiesel and its ignition quality is different from straight chain alkanes due to the presence of ester group and asymmetric structure. The ignition quality of C4~C18 saturated esters, unsaturated esters-2, unsaturated esters-5 and other unsaturated esters is characterized by CN (see Figure 14) since no measured RON/MON data is available. Three important generalizations are obtained from this illustration: First, the CN increases with increasing carbon chain length which applies to both saturated and unsaturated esters. For example, the predicted CNs of methyl butanoate (8.80), methyl pentanoate (13.72), methyl hexanoate (21.47), methyl heptanoate (32.77), methyl octanoate (36.93), methyl nonanoate (43), methyl decanoate (48.84), methyl undecanoate (59.64), methyl laurate (64.13), methyl palmitate (83.91), methyl stearate (85.51) increase progressively as increasing chain length. Unsaturated esters have similar behavior. The predicted CNs of methyl propanoate (12.34) vs methyl butanoate (8.80), methyl acrylate (7.45) vs methyl-2-butenoate (6.55), methyl-5-hexenoate (18.91) vs methyl-5-heptenoate (17.63) do not exactly follow this rule. These abnormal predicted CNs mainly cause by insufficient model training due to lack of measured data and more experimental data is needed. Second, the addition of carbon-carbon double bond into ester molecules decreases CN compared to the corresponding saturated esters. Both moving double bond toward center of molecule and increasing the number of double bonds reduce CN. For example, the predicted CNs of methyl laurate (64.13), methyl-2-dodecenoate (46.59), methyl-5-dodecenoate (35.51) decrease as introduction and centralization of double bond. The predicted CNs of methyl oleate (57.2), methyl linoleate (42.5), methyl linolenate (40.44) reduce progressively as the number of double bond increases from one to three. Third, the short chained methyl esters have lower reactivity than corresponding alkanes which are the preferred octane booster for high performance gasoline. For example, methyl propanoate, methyl acrylate have particular high RON as 114.58, 119.54 as shown in Table S7 of supporting information.

Algae and waste fish oil have been identified as a feedstock to produce large amount of FAME which has the

potential to relieve the food vs fuel issue for biodiesel production [70]. These materials contain significant amounts of highly polyunsaturated FAME (more than two double bond) and the proposed method is used to predict the CN. Methyl 5(Z),8(Z),11(Z),14(Z)-eicosatetraenoate (CAS 2566-89-4) and methyl 4(Z),7(Z),10(Z),13(Z),16(Z),19(Z)-docosahexaenoate (CAS 2566-90-7) are typical algal oils compositions which contain 4 and 6 unsaturated double bonds respectively. The predicted CNs by this method are 29.675 and 26.9932 respectively which approach the measured DCN of 29.57 and 24.35 [70, 71]. It proves that the proposed method has good extrapolation capacity.

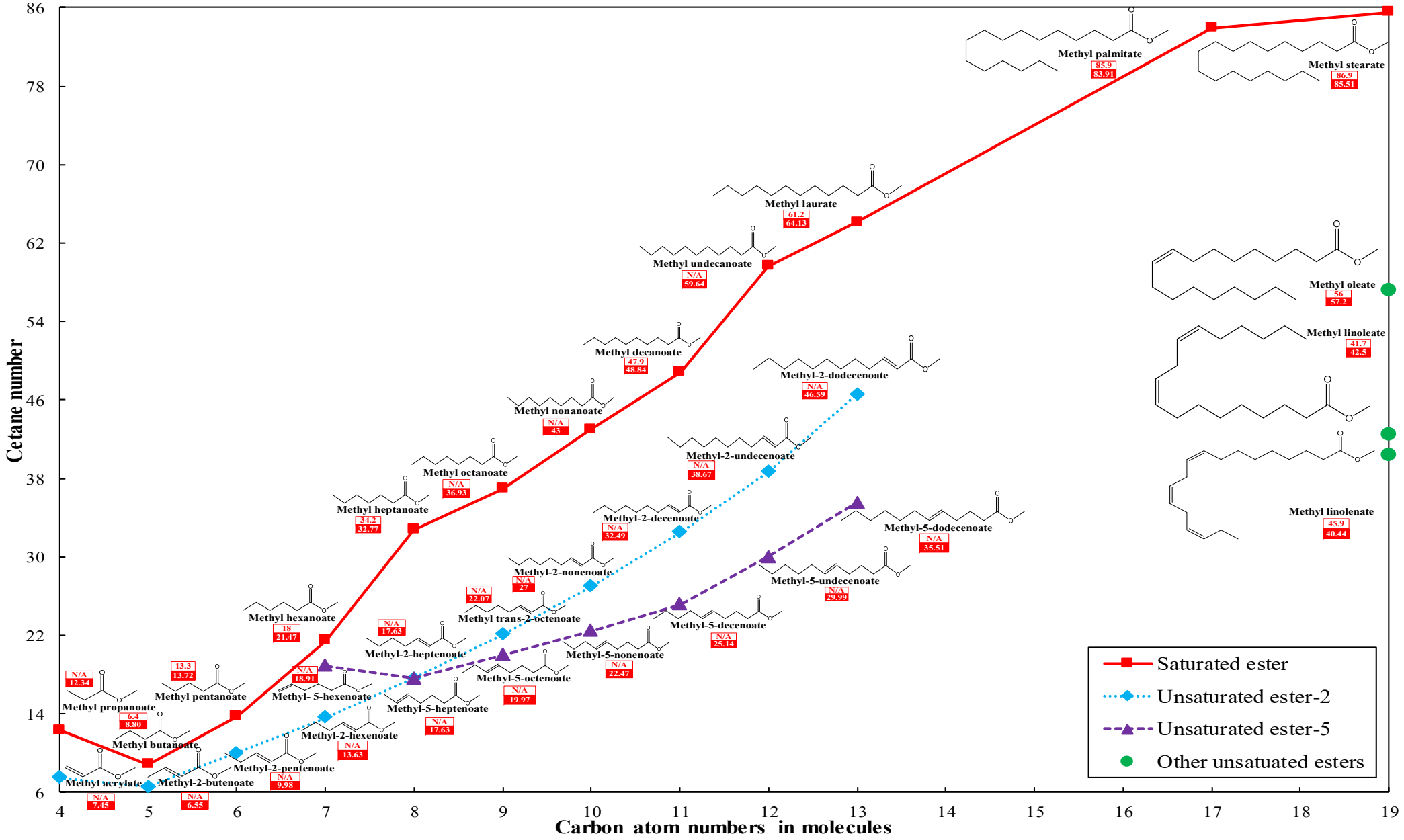

Figure 14. CN of C4~C19 esters, numbers with red frames and red backgrounds are measured values and predictive values.

3.3 Method application to fuel mixtures

The proposed method can not only predict the CN, RON, MON of pure compounds, but can also apply to fuel mixtures. In this section, TPRF mixture (n-heptane-iso-octane-toluene) is taken as an example to demonstrate the predictive capacity since sufficient experimental data are available. The measured and predicted CN/RON/MON/OS and predicted error (measured value-predictive value) of TPRF are provided in Figure 15. The proposed method accurately reproduces CN/RON/MON of TPRF and the correlation coefficients reach 0.9933, 0.9984, 0.991 respectively as shown in Figure 16. The extremely low MAE and RMSE in Table 6 also indicate that the proposed method successfully captures the non-linear relationship between mixing proportion and ignition quality as well as molecule interaction. The predicted OS is the difference between predicted RON and predicted MON and its R^2 is relatively low as 0.8849 because there is an outlier of 20.80mol.%n-heptane-79.20mol.%toluene (measured OS=2, predicted OS=10.072). These RON and MON may contain significant measurement error and uncertainty. This is an example of using the proposed method to diagnose the abnormal experimental data. Another ternary mixture of n-heptane-dibutyl ether-ethanol is tested by the proposed method to obtain the CN/RON/MON/OS information. n-Heptane and dibutyl ether are chose because the former is a typical diesel surrogate and the latter is an alternative biofuel with high cetnae number and soot suppression performance [72, 73]. Ignition quality is varied by adjusting the ethanol proportion. The predicted CN/RON/MON/OS of n-heptane-dibutyl ether-ethanol mixture is shown in Figure 17. There are no published experimental and predicted data available for this mixture and it is the first time to report its ignition quality. Further ignition quality test of this ternary mixture by CFR engine test is necessary to verify the predicted values.

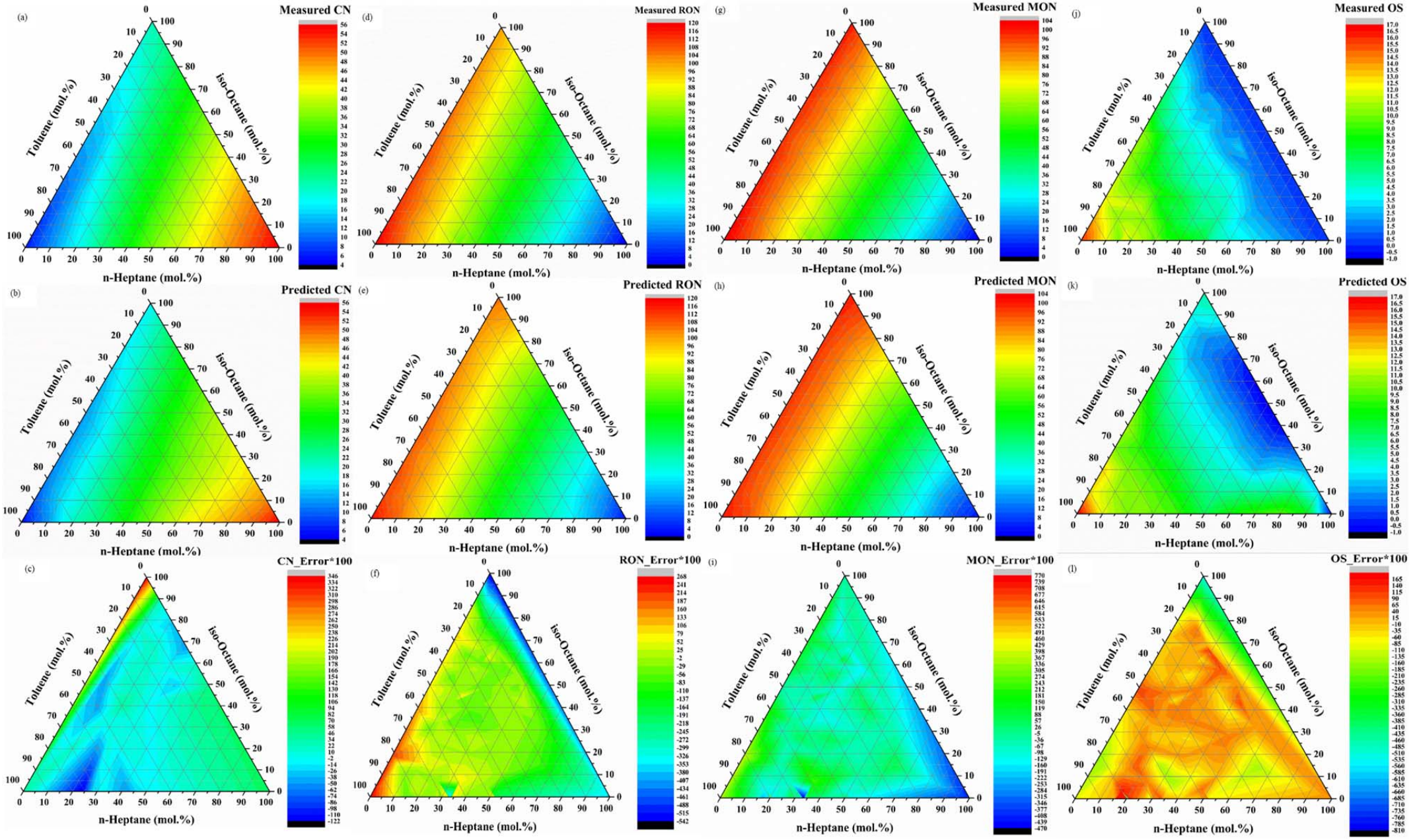

Figure 15. Comparison between measured, predicted values and errors of (a)~(c) CN, (d)~(f) RON, (g)~(i) MON, (j)~(l) OS of TPRF mixtures.

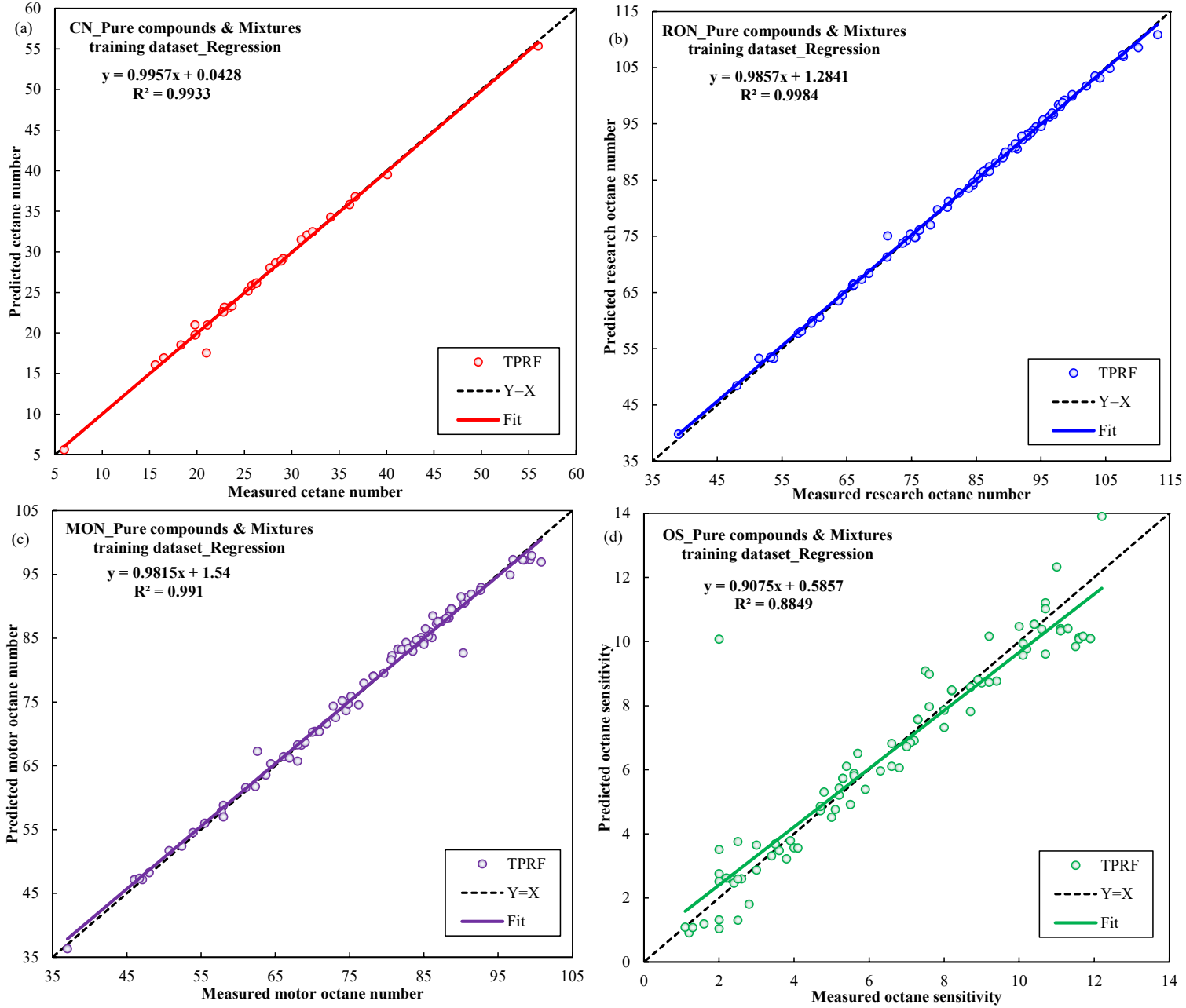

Figure 16. Parity plots for (a) CN, (b) RON, (c) MON, (d) OS of TPRF mixtures between measured and predictive values by machine learning regression model.

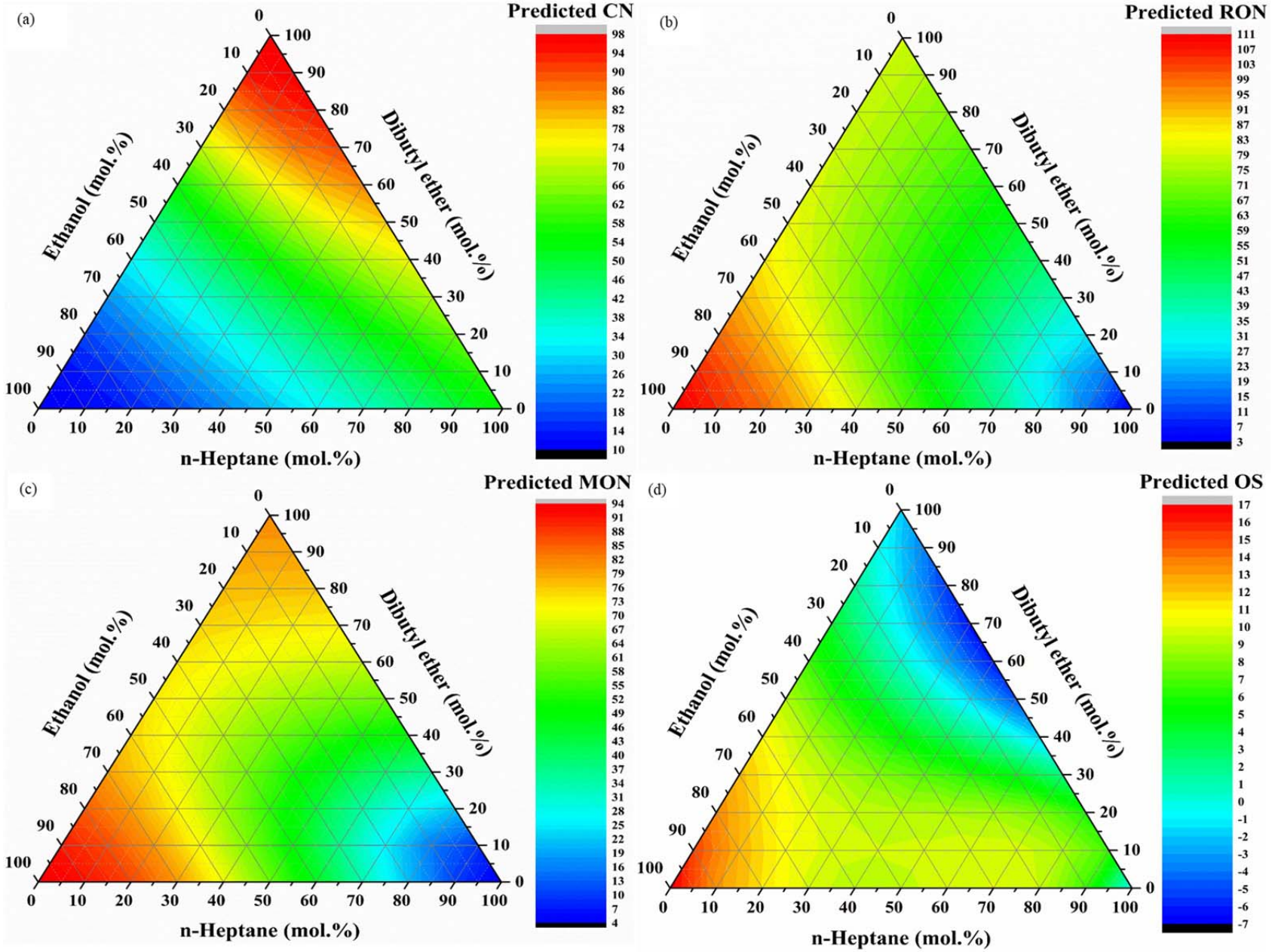

Figure 17. Predicted (a) CN, (b) RON, (c) MON, (d) OS of n-heptane-dibutyl ether-ethanol mixtures by machine learning regression model.

Table 6. Statistical analysis of predictive performance of TPRF mixtures for machine learning regression models

Property	Fuel mixtures (No. of measured data)	R-squared	MAE	RMSE
CN	TPRF (30)	0.9933	0.395	0.732
RON	TPRF (87)	0.9984	0.4	0.655
MON	TPRF (87)	0.991	0.869	1.368
OS	TPRF (87)	0.8849	0.661	1.139

4. CONCLUSIONS

A machine learning regression based group contribution method is proposed to simultaneously predict CN/RON/MON of pure fuel compounds and mixtures. The proposed method is applicable to a wide range of group compounds including alkanes, alkenes, alkynes, cycloalkanes, cycloalkenes, aromatics, alcohols, aldehydes/ketones, ethers, esters, acids, furans and fuel mixtures. High predictive precision is achieved and the overall R-squared for CN/RON/MON are 0.9911, 0.9874, 0.9731 respectively which are superior to traditional neural network method of 0.934, 0.9627, 0.9634.

The method has the following advantages over the published methods: (i) It can predict the CN/RON/MON simultaneously and avoid using the inaccurate conversion formulas between CN and RON/MON. (ii) GCM is applicable to both pure compounds and fuel mixtures and the latter is always a great challenge for ignition quality characterization. (iii) Abnormal experimental CN/RON/MON can be easily discovered by box-and-whisker chart analysis. (iv) It can generate insight into the impact of fuel molecular structure on the ignition quality even lack of experimental data.

The success of the method can be majorly attributed to three key factors:

1. The improved group contribution method GCM-UOB 2.0 takes into account structural features, functional group interaction and fuel reactivity by using functional group identifier, functional group position descriptor and fuel reactivity descriptor. It significantly improves the distinguishability of aromatics containing phenyl group, naphthyl group, and anthranyl group.
2. The embedded machine learning algorithm automatically optimizes the model functions and parameters which results in higher predictive accuracy than neural network method.
3. The comprehensive fuel ignition quality database developed by this work provides a good foundation for model training and validation. This is based on the requirement of machine

543 learning theory of the bigger the data, the better the model.

544 This method provides an effective tool to obtain CN/RON/MON of both pure compounds and fuel
545 mixtures and a fundamental understanding of the influence of fuel molecular structure on ignition
546 quality. The latter provides a design rules of future higher performance diesel and gasoline, ultimately
547 leads to improved engine efficiency, reduction of carbon footprint and pollutant emissions.

548

549 # ACKNOWLEDGEMENT

550 This work is supported by Innovate UK (The Technology Strategy Board, TSB, No. 400176/149)
551 and Engineering & Physical Sciences Research Council (EPSRC, No. EP/P03117X/1). Runzhao Li
552 also thanks to University of Birmingham for the award of a Ph.D. research scholarship (No. 1871018).
553 The work is conducted in Future Engines & Fuels Lab, University of Birmingham. The authors thank
554 the laboratory managers and staff workers for their hospitability, time, and opinions. Special thanks go
555 to Shenzhen Gas Corporation Ltd. for providing us the technical guidance. The authors are indebted to
556 the reviewers of this article for their invaluable suggestions.
557

REFERENCES

[1] ASTM D613-18a Standard Test Method for Cetane Number of Diesel Fuel Oil. American Society for Testing and Materials (ASTM) international; 2018.

[2] ASTM D6890-18 Standard Test Method for Determination of Ignition Delay and Derived Cetane Number (DCN) of Diesel Fuel Oils by Combustion in a Constant Volume Chamber. American Society for Testing and Materials (ASTM) international; 2018.

[3] ASTM D7668-17 Standard Test Method for Determination of Derived Cetane Number (DCN) of Diesel Fuel Oils—Ignition Delay and Combustion Delay Using a Constant Volume Combustion Chamber Method. American Society for Testing and Materials (ASTM) international; 2017.

[4] ASTM D8183-18 Standard Test Method for Determination of Indicated Cetane Number (ICN) of Diesel Fuel Oils using a Constant Volume Combustion Chamber—Reference Fuels Calibration Method. American Society for Testing and Materials (ASTM) international; 2018.

[5] ASTM D2699-19 Standard Test Method for Research Octane Number of Spark-Ignition Engine Fuel. American Society for Testing and Materials (ASTM) international; 2019.

[6] ASTM D2700-19 Standard Test Method for Motor Octane Number of Spark-Ignition Engine Fuel. American Society for Testing and Materials (ASTM) international; 2019.

[7] Laboratory NRE. Co-Optimization of Fuels & Engines: Fuel Properties Database. https://wwwnrelgov/transportation/fuels-properties-database/.

[8] Farrell J, Holladay J, Wagner R. Co-Optimization of Fuels & Engines-FY16 Year in Review. 2016.

[9] Farrell J, Wagner R, Holladay J, Moen C. Co-Optimization of Fuels-Engines FY17 Year in Review. 2017.

[10] Farrell J, Wagner R, Gaspar D, Moen C. Co-Optimization of Fuels-Engines FY18 Year in Review. 2018.

[11] Narayanaswamy K, Pitsch H, Pepiot P. A component library framework for deriving kinetic mechanisms for multi-component fuel surrogates: Application for jet fuel surrogates. Combustion and Flame 2016;165:288-309.

[12] Yanowitz J, Ratcliff MA, McCormick RL, Taylor JD, Murphy MJ. Compendium of Experimental Cetane Numbers. 2017.

[13] Combustion characteristics of compression ignition engine fuel components. SAE Technical Papers 600112 1960.

[14] Naik CV, Puduppakkam K, Wang C, Kottalam J, Liang L, Hodgson D, et al. Applying Detailed Kinetics to Realistic Engine Simulation: the Surrogate Blend Optimizer and Mechanism Reduction Strategies. SAE International Journal of Engines 2010;3(1):241-59.

[15] Agosta A. Development of a chemical surrogate for JP-8 aviation fuel using a pressurized flow reactor. Master thesis of Drexel University 2002.

[16] Ghosh P, Jaffe SB. Detailed composition-based model for predicting the cetane number of diesel fuels. Ind Eng Chem Res 2006;45(1):346-51.

[17] Manente V, Johansson B, Cannella W. Gasoline partially premixed combustion, the future of internal combustion engines? International Journal of Engine Research 2011;12(3):194-208.

[18] ASTM D975-19a Standard Specification for diesel fuel oils. American Society for Testing and Materials (ASTM) international; 2019.

602 [19] Katritzky AR, Kuanar M, Slavov S, Hall CD, Karelson M, Kahn I, et al. Quantitative
603 Correlation of Physical and Chemical Properties with Chemical Structure: Utility for Prediction.
604 Chem Rev 2010;110(10):5714-89.
605 [20] Kessler T, Dorian G, Mack JH. Application of a Rectified Linear Unit (Relu) Based Artificial
606 Neural Network to Cetane Number Predictions. Proceedings of the Asme Internal Combustion
607 Engine Fall Technical Conference, 2017, Vol 1 2017.
608 [21] Guo Z, Lim KH, Chen M, Thio BJR, Loo BLW. Predicting cetane numbers of hydrocarbons and
609 oxygenates from highly accessible descriptors by using artificial neural networks. Fuel
610 2017;207:344-51.
611 [22] Kessler T, Sacia ER, Bell AT, Mack JH. Predicting the Cetane Number of Furanic Biofuel
612 Candidates Using an Improved Artificial Neural Network Based on Molecular Structure.
613 Proceedings of the Asme Internal Combustion Engine Fall Technical Conference, 2016 2016.
614 [23] Yang H, Ring Z, Briker Y, McLean N, Friesen W, Fairbridge C. Neural network prediction of
615 cetane number and density of diesel fuel from its chemical composition determined by LC and
616 GC-MS. Fuel 2002;81(1):65-74.
617 [24] Yang H, Fairbridge C, Ring Z. Neural Network Prediction of Cetane Numbers for Isoparaffins
618 and Diesel Fuel. Petroleum Science and Technology 2001;19(5-6):573-86.
619 [25] Abdul Jameel AG, Van Oudenhoven V, Emwas A-H, Sarathy SM. Predicting Octane Number
620 Using Nuclear Magnetic Resonance Spectroscopy and Artificial Neural Networks. Energy &
621 Fuels 2018;32(5):6309-29.
622 [26] Smolenskii EA, Ryzhov AN, Bavykin VM, Myshenkova TN, Lapidus AL. Octane numbers
623 (ONs) of hydrocarbons: a QSPR study using optimal topological indices for the topological

equivalents of the ONs. Russ Chem B+ 2007;56(9):1681-93.

[27] Smolenskii EA, Bavykin VM, Ryzhov AN, Slovokhotova OL, Chuvaeva IV, Lapidus AL. Cetane numbers of hydrocarbons: calculations using optimal topological indices. Russ Chem B+ 2008;57(3):461-7.

[28] Hosoya H. Chemical meaning of octane number analyzed by topological indices. Croat Chem Acta 2002;75(2):433-45.

[29] Singh E, Badra J, Mehl M, Sarathy SM. Chemical Kinetic Insights into the Octane Number and Octane Sensitivity of Gasoline Surrogate Mixtures. Energy & Fuels 2017;31(2):1945-60.

[30] Kubic WL. A Group Contribution Method for Estimating Cetane and Octane Numbers. Los Alamos National Laboratory Report No LA-UR-16-25529 2016.

[31] Kubic WL, Jenkins RW, Moore CM, Semelsberger TA, Sutton AD. Artificial Neural Network Based Group Contribution Method for Estimating Cetane and Octane Numbers of Hydrocarbons and Oxygenated Organic Compounds. Ind Eng Chem Res 2017;56(42):12236-45.

[32] Dahmen M, Marquardt W. A Novel Group Contribution Method for the Prediction of the Derived Cetane Number of Oxygenated Hydrocarbons. Energy & Fuels 2015;29(9):5781-801.

[33] Saldana DA, Starck L, Mougin P, Rousseau B, Pidol L, Jeuland N, et al. Flash Point and Cetane Number Predictions for Fuel Compounds Using Quantitative Structure Property Relationship (QSPR) Methods. Energy & Fuels 2011;25(9):3900-8.

[34] Creton B, Dartiguelongue C, de Bruin T, Toulhoat H. Prediction of the Cetane Number of Diesel Compounds Using the Quantitative Structure Property Relationship. Energy & Fuels 2010;24(10):5396-403.

[35] DeFries TH, Kastrup RV, Indritz D. Prediction of cetane number by group additivity and

carbon-13 Nuclear Magnetic Resonance. Ind Eng Chem Res 1987;26(2):188-93.

[36] Liu Z, Zhang L, Elkamel A, Liang D, Zhao S, Xu C, et al. Multiobjective Feature Selection Approach to Quantitative Structure Property Relationship Models for Predicting the Octane Number of Compounds Found in Gasoline. Energy & Fuels 2017;31(6):5828-39.

[37] Al-Fahemi JH, Albis NA, Gad EAM. QSPR Models for Octane Number Prediction. Journal of Theoretical Chemistry 2014;2014:1-6.

[38] Westbrook CK, Sjöberg M, Cernansky NP. A new chemical kinetic method of determining RON and MON values for single component and multicomponent mixtures of engine fuels. Combustion and Flame 2018;195:50-62.

[39] Badra JA, Bokhumseen N, Mulla N, Sarathy SM, Farooq A, Kalghatgi G, et al. A methodology to relate octane numbers of binary and ternary n-heptane, iso-octane and toluene mixtures with simulated ignition delay times. Fuel 2015;160:458-69.

[40] Guan C, Zhai J, Han D. Cetane number prediction for hydrocarbons from molecular structural descriptors based on active subspace methodology. Fuel 2019;249:1-7.

[41] Baghban A, Adelizadeh M. On the determination of cetane number of hydrocarbons and oxygenates using Adaptive Neuro Fuzzy Inference System optimized with evolutionary algorithms. Fuel 2018;230:344-54.

[42] Daly SR, Niemeyer KE, Cannella WJ, Hagen CL. Predicting fuel research octane number using Fourier-transform infrared absorption spectra of neat hydrocarbons. Fuel 2016;183:359-65.

[43] Abdul Jameel AG, Naser N, Emwas A-H, Dooley S, Sarathy SM. Predicting Fuel Ignition Quality Using 1H NMR Spectroscopy and Multiple Linear Regression. Energy & Fuels 2016;30(11):9819-35.

668 [44] ASTM D7170-16 Standard Test Method for Determination of Derived Cetane Number (DCN)
669 of Diesel Fuel Oils—Fixed Range Injection Period, Constant Volume Combustion Chamber
670 Method. American Society for Testing and Materials (ASTM) international; 2016.
671 [45] Won SH, Haas FM, Dooley S, Dryer FL. Chemical functional group descriptor for jet fuel
672 surrogate. 10th US National Combustion Meeting 2017.
673 [46] Won SH, Dooley S, Veloo PS, Wang H, Oehlschlaeger MA, Dryer FL, et al. The combustion
674 properties of 2,6,10-trimethyl dodecane and a chemical functional group analysis. Combustion
675 and Flame 2014;161(3):826-34.
676 [47] Wang Y, Cao Y, Wei W, Davidson DF, Hanson RK. A new method of estimating derived cetane
677 number for hydrocarbon fuels. Fuel 2019;241:319-26.
678 [48] Carpenter DO. Impact of Cycloalkanes on Ignition Propensity Measured as Derived Cetane
679 Number in Multi-Component Surrogate Mixtures. Master's thesis of University of South
680 Carolina 2019.
681 [49] API Tech Data Book. http://wwwepconcom/api-data-bookhtml.
682 [50] Kalghatgi G, Babiker H, Badra J. A Simple Method to Predict Knock Using Toluene, N-Heptane
683 and Iso-Octane Blends (TPRF) as Gasoline Surrogates. SAE International Journal of Engines
684 2015;8(2):505-19.
685 [51] Morgan N, Smallbone A, Bhave A, Kraft M, Cracknell R, Kalghatgi G. Mapping surrogate
686 gasoline compositions into RON/MON space. Combustion and Flame 2010;157(6):1122-31.
687 [52] Naser N, Yang SY, Kalghatgi G, Chung SH. Relating the octane numbers of fuels to ignition
688 delay times measured in an ignition quality tester (IQT). Fuel 2017;187:117-27.
689 [53] Foong TM, Morganti KJ, Brear MJ, da Silva G, Yang Y, Dryer FL. The octane numbers of

690 ethanol blended with gasoline and its surrogates. Fuel 2014;115:727-39.
691 [54] Lapidus AL, Bavykin VM, Smolenskii EA, Chuvaeva IV. Cetane numbers of hydrocarbons as a
692 function of their molecular structure. Doklady Chemistry 2008;420(2):150-5.
693 [55] Ogawa H, Nishimoto H, Morita A, Shibata G. Predicted diesel ignitability index based on the
694 molecular structures of hydrocarbons. International Journal of Engine Research
695 2016;17(7):766-75.
696 [56] Knocking Characteristics of Pure Hydrocarbons. American Petroleum Institute Research Project
697 45 1958.
698 [57] Li R, Herreros JM, Tsolakis A, Yang W. Novel Functional Group Contribution Method for
699 Surrogate Formulation with Accurate Fuel Compositions. Energy & Fuels
700 2020;34(3):2989-3012.
701 [58] Montgomery DC. Design and Analysis of Experiments, 8th Edition. John Wiley & Sons; 2012.
702 [59] Puckett AD, Caudle BH. Ignition qualities of hydrocarbons in the diesel-fuel boiling range. US
703 Dept of the Interior, Bureau of Mines 1948.
704 [60] Abou-Rachid H, Bonneviot L, Xu G, Kaliaguine S. On the correlation between kinetic rate
705 constants in the auto-ignition process of some oxygenates and their cetane number: a quantum
706 chemical study. Journal of Molecular Structure: THEOCHEM 2003;621(3):293-304.
707 [61] Albahri TA. Structural Group Contribution Method for Predicting the Octane Number of Pure
708 Hydrocarbon Liquids. Ind Eng Chem Res 2003;42(3):657-62.
709 [62] Lovell WG. Knocking Characteristics of Hydrocarbons. Industrial & Engineering Chemistry
710 1948;40(12):2388-438.
711 [63] Boot MD, Tian M, Hensen EJM, Mani Sarathy S. Impact of fuel molecular structure on

712 auto-ignition behavior – Design rules for future high performance gasolines. Progress in Energy
713 and Combustion Science 2017;60:1-25.
714 [64] Pitz WJ, Mueller CJ. Recent progress in the development of diesel surrogate fuels. Progress in
715 Energy and Combustion Science 2011;37(3):330-50.
716 [65] Sarathy SM, Farooq A, Kalghatgi GT. Recent progress in gasoline surrogate fuels. Progress in
717 Energy and Combustion Science 2018;65:67-108.
718 [66] Worldwide Fuel Charter, 5th Edition. Organisation Internationale des Constructeurs
719 d'Automobiles (OICA); 2013.
720 [67] Lovell WG, Campbell JM, Boyd TA. Knocking characteristics of naphthene hydrocarbons.
721 Industrial And Engineering Chemistry 1933;25(10):1107-10.
722 [68] Sarathy SM, Oßwald P, Hansen N, Kohse-Höinghaus K. Alcohol combustion chemistry.
723 Progress in Energy and Combustion Science 2014;44:40-102.
724 [69] Sarathy SM, Vranckx S, Yasunaga K, Mehl M, Oßwald P, Metcalfe WK, et al. A comprehensive
725 chemical kinetic combustion model for the four butanol isomers. Combustion and Flame
726 2012;159(6):2028-55.
727 [70] Knothe G. Fuel Properties of Highly Polyunsaturated Fatty Acid Methyl Esters. Prediction of
728 Fuel Properties of Algal Biodiesel. Energy & Fuels 2012;26(8):5265-73.
729 [71] Knothe G. A comprehensive evaluation of the cetane numbers of fatty acid methyl esters. Fuel
730 2014;119:6-13.
731 [72] Gao Z, Zhu L, Zou X, Liu C, Tian B, Huang Z. Soot reduction effects of dibutyl ether (DBE)
732 addition to a biodiesel surrogate in laminar coflow diffusion flames. Proceedings of the
733 Combustion Institute 2019;37(1):1265-72.

734 [73] Gao Z, Zou X, Huang Z, Zhu L. Predicting sooting tendencies of oxygenated hydrocarbon fuels
735 with machine learning algorithms. Fuel 2019;242:438-46.

AUTHOR INFORMATION

Corresponding Author

*E-mail: a.tsolakis@bham.ac.uk

ORCID

Runzhao Li: 0000-0001-5120-9849

Jose Martin Herreros: 0000-0002-6939-121X

Athanasios Tsolakis: 0000-0002-6222-6393

Notes

The authors declare no competing financial interest.