Correa Gonzalez, Sandra; Kroyan, Yuri; Sarjovaara, Teemu; Kiiski, Ulla; Karvo, Anna; Toldy, Arpad I.; Larmi, Martti; Santasalo-Aarnio, Annukka

Prediction of Gasoline Blend Ignition Characteristics Using Machine Learning Models

Published in:
Energy and Fuels

DOI:
10.1021/acs.energyfuels.1c00749

Published: 03/06/2021

Document Version
Publisher's PDF, also known as Version of record

energy&fuels

pubs.acs.org/EF Article

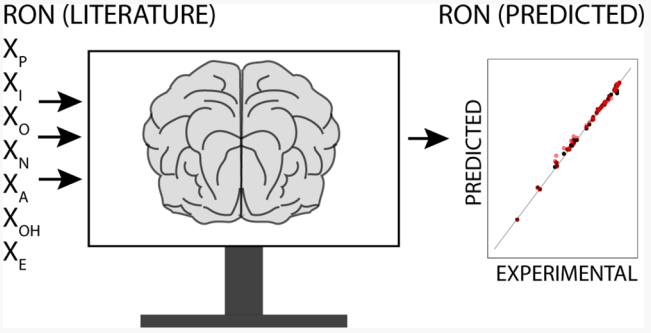
Prediction of Gasoline Blend Ignition Characteristics Using Machine Learning Models

Sandra Correa Gonzalez, Yuri Kroyan, Teemu Sarjovaara, Ulla Kiiski, Anna Karvo, Arpad I. Toldy, Martti Larmi, and Annukka Santasalo-Aarnio*

Cite This: *Energy Fuels* 2021, 35, 9332–9340 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Research Octane Number (RON), among other autoignition related properties, is a primary indicator of the grade of spark-ignition (SI) fuels. However, in many cases, the blending of various gasoline components affects the RON of the final fuel product in a nonlinear way. Currently, the lack of precise predictive models for RON challenges the accurate blending and production of commercial SI fuels. This study compares popular Machine Learning (ML) algorithms and evaluates their potential to develop state-of-the-art models able to predict key SI fuel properties. Typical gasoline composition was simplified and represented by a palette of seven characteristic molecules, including five hydrocarbons and two oxygenated species. Ordinary Least Squares (OLS), Nearest Neighbors (NN), Support Vector Machines (SVM), Decision Trees (DT), and Random Forest (RF) algorithms were trained, cross-validated, and tested using a database containing 243 gasoline-like fuel blends with known RON. Best results were obtained with nonlinear SVM algorithms able to reproduce synergistic and antagonistic molecular interactions. The Mean Absolute Error (MAE) on the test set was equal to 0.9, and the estimator maintained its accuracy when alterations were performed on the training data set. Linear methods performed better using molar compositions while predictions on a volumetric basis required nonlinear algorithms for satisfactory accuracy. Developed models allow one to quantify the nonlinear blending behavior of different hydrocarbons and oxygenates accounting for those effects during fuel blending and production. Moreover, these models contribute to a deeper understanding of the phenomena that will facilitate the introduction of alternative gasoline recipes and components.



1. INTRODUCTION

1.1. Prediction of Gasoline Blend Properties. The current transport sector is heavily dependent on fossil fuels and is responsible for 25% of the global CO_2 emissions.[1] Efforts to alleviate the environmental impact of internal combustion engines (ICEs) and current vehicle fleets focus on increasing overall combustion efficiency. Engine downsizing[2] or the introduction of high-octane fuels[3] helps increase fuel economy and reduce related emissions. Another approach that can reduce the environmental impact of ICEs is the use of alternative fuels obtained from sustainable feedstock. In part due to the diversity and higher complexity of spark-ignition (SI) fuels, alternative drop-in gasoline synthesis is still under development. The introduction of new raw materials such as municipal solid waste (MSW) or industrial waste could translate into changes in the chemistry of intermediate and final streams and pose challenges to manufacturing a consistent product. Gasoline blends are known to behave nonlinearly concerning several key properties, including antiknocking performance and volatility. For the most part, these properties are estimated using linear models due to the complexity of blends and the lack of more accurate tools. This uncertainty regarding gasoline blending behavior is one of the aspects preventing the increase of alternative feedstock in the production chain of SI engine fuels. Therefore, in order to be able to incorporate novel and/or renewable components into commercial fuel blends, a deeper understanding of the relation between composition and fuel properties is required.

Research Octane Number (RON) as well as Motor Octane Number (MON) measure the antiknock quality of fuels using a specially designed engine.[4] The tediousness of the standard test has led to an increased interest in alternative predictive tools both in the fuel design (i.e., prior to blending) and fuel analysis (i.e., tools utilizing data obtained post-blending) stages of production, as presented in Figure 1. In the fuel design stage, the simplest approach is to use linear tools. The first studies that included nonlinear interactions were based on

Received: March 9, 2021
Revised: May 6, 2021
Published: May 19, 2021

https://doi.org/10.1021/acs.energyfuels.1c00749
Energy Fuels 2021, 35, 9332–9340

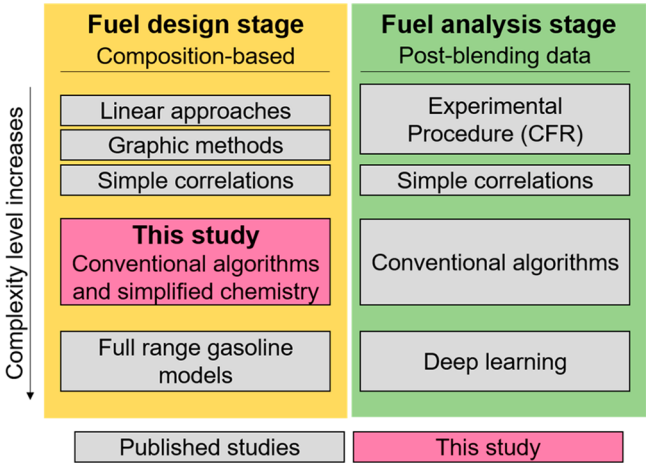


Figure 1. Summary of existing literature and the intended niche of this paper, as discussed in the Introduction. This study provides useful tools in the fuel design stage with an intermediate level of complexity.

graphical methods.[5,6] More recently, calculation methods based on the use of blending octane numbers have been proposed,[7] as well as correlations that include deviation terms to account for synergistic and antagonistic effects.[8] Multivariate statistical analysis based on near-infrared (NIR) spectral features has been used to perform online RON prediction of full range gasolines.[9−11]

The use of Machine Learning (ML) algorithms has also been demonstrated in previous studies to predict gasoline autoignition behavior both on the individual component and the blend level. For individual components, quantitative structure−property relationship models have been effective. These models (often employing Artificial Neural Networks, ANNs) use features extracted from the molecular structure, such as the presence and position of functional groups, level of oxygenates, etc., to predict properties.[12,13] More recently, Li et al. developed a new group contribution model based on ML that is able to predict octane and cetane numbers for a wide range of individual compounds and mixtures of known composition.[14] These models have enabled the creation of extensive and accurate databases of fuel-relevant molecules and their properties.[15]

For finished gasoline blends, studies have typically focused on the fuel analysis stage, involving data from noninvasive optical techniques. NIR spectra have been correlated with RON using the Random Forest (RF) algorithm[16] and the performance of Support Vector Machine (SVM) algorithms has been compared to Partial Least Squares (PLS) on Nuclear Magnetic Resonance (NMR) data.[17] ANNs have been deployed on NMR spectral results[18] and gas chromatography.[19,20] Composition-based models for the fuel design stage are also available at various levels of complexity: especially ethanol-containing gasoline (and surrogate) blend properties have been accurately modeled using multivariate linear regression by Anderson and Wallington[21] and using blending rules coupled with measurements on a new fuel surrogate model by AlRamadan et al.[22] Albahri[23] predicted RON among other gasoline properties using ANN. Input features included distillation curve points, Reid vapor pressure (RVP), and content of olefins, saturates, and aromatics. Pasadakis et al.[24] tuned an ANN to predict gasoline RON based on refinery stream blending ratios. Similar studies include additional gasoline properties as input features such as distillation curves and density.[25]

1.2. Gasoline composition simplification. A simplistic representation of gasoline can be achieved using a single molecule, isooctane. With both RON and MON equal to 100 and similar carbon to hydrogen ratio to real gasoline, several studies have focused on isooctane oxidation kinetics.[26] Isooctane is used in binary mixtures of primary reference fuels (PRF) together with *n*-heptane. Using a linear-by-volume approach, PRFs are used to predict the knocking resistance of fuels in standard test procedures. Ternary blends of PRFs and toluene give the possibility of accounting for the octane sensitivity of the fuels (S), defined as the difference between RON and MON.[27] More complex surrogate fuels have also been proposed in the literature and are widely used in research.[28−31] Fuels for Advanced Combustion Engine (FACE) gasolines are formulated to emulate several gasoline recipes.[32] These surrogates include a wider range of molecules and extend ternary blends with the addition of olefins and naphthenes. Surrogate fuels are often used in research to help handle complex gasoline composition and its computational representation and study. Their exact formulation highly depends on the phenomena of interest. However, as kinetic modeling develops and computational capacity increases, multipurpose surrogates able to match several target properties become more relevant and applicable. This study focuses on RON prediction using a simple but comprehensive chemistry to represent gasoline fuels. Despite the complex chemistry of fossil-based gasoline, hydrocarbons present in it can be categorized into five different groups, namely, linear paraffins (P), branched or isoparaffins (I), olefins (O), naphthenes (N), and aromatics (A). In addition to that, the blending of oxygenated species such as alcohols (OH) and ethers (E) is a common practice to enhance fuel octane rating. Moreover, the oxygen content in commercial fuels is expected to prevail with the integration of larger shares of biobased feedstock. For these reasons, and for the sake of data availability in the literature, blends of these seven components were chosen as the subject of this study.

1.3. Objectives of this study. The present study positions itself in the fuel design space and compares the suitability of 8 ML algorithms to build predictive models that can be used to define gasoline recipes for a given RON using a simplified gasoline representation. Literature data on blends consisting of up to seven common ingredients representing paraffins, isoparaffins, olefins, naphtenes, aromatics, alcohols, and ethers is used as an input. The models are also intended to accommodate new approaches to fuel chemistry and improve the estimation accuracy of existing tools while still relying on simplified gasoline representations. In addition to the predictive capability, these models are presented as a tool to investigate the behavior of different chemical species in multicomponent blends, as their moderate complexity enables the interpretability of the results. Furthermore, a better understanding of the underlying phenomena behind nonlinear behaviors is expected to facilitate the refinery blending operations with more complex input streams without requiring the extensive implementation of spectroscopic analysis throughout the production process. Models for RON prediction are presented in this paper, but the same methodology could be applied for any other properties showing nonlinear behavior and based on any other data set of interest.[33]

https://doi.org/10.1021/acs.energyfuels.1c00749
Energy Fuels 2021, 35, 9332−9340

2. DATA AND METHODOLOGY

2.1. Definition of Compositions. Motivated by the abundance of the molecular groups described in the Introduction in commercial gasoline as well as by data availability, this study proposes a palette of seven species to represent PIONA+OH+E compounds, as shown in Table 1.

Table 1. Proposed Palette of Species for Gasoline Representation in This Study and Their Corresponding RON Values[7,34,35]

group		molecule	RON
hydrocarbons	linear paraffins	*n*-heptane	0
	isoparaffins	isooctane	100
	olefin	1-hexene	73.6
	naphthene	cyclopentane	100
	aromatic	toluene	118
oxygenated species	alcohol	ethanol	108
	ether	ethyl *tert*-butyl-ether (ETBE)	117

2.2. Data Set. The data set used in this study was collected from existing literature and is presented in the Supporting Information. During the data acquisition stage, points were included in the set if they satisfied two main criteria: the RON value had been measured using the standard experimental tests, and the blend only includes the species listed in Table 1. As for the number of components or the compositional ratios, no restrictions were applied. The resulting data set consists of 243 samples with unique volume-based composition and RON values in the range 0−118 corresponding to neat *n*-heptane and toluene, respectively. Because of the limited size of the data set and the need to collect it from several sources, data cleaning was performed manually. Duplicates (i.e., data points with identical composition) were kept as long as their RON values were within the experimental error. All features had the same scale and ranges, and therefore normalization was not applied. A second data set was derived from the first one where the composition of each sample is expressed on a molar basis instead, as previous studies[36−38] have found that using molar concentrations improve the accuracy of linear octane blending models, especially for gasoline−ethanol blends. However, the performance difference of nonlinear models is not well-known.

2.3. Machine Learning Algorithms. ML methods can be divided into two broad categories, supervised and unsupervised learning.[39] Supervised learning requires labeled data for the training stage, that is, data that contains both input features and desired outputs. Unsupervised learning infers natural structures within unlabeled data instead. All algorithms applied in this study fall into the first category, suitable for regression tasks, which can also be considered traditional methods. Traditional algorithms are often advised for small data sets in contraposition to deep learning approaches.[40] In this study, the need for experimental data is the limiting factor that constrains the size of the data sets. Under those premises, five popular and well-defined types of supervised algorithms were selected and compared.

Ordinary Least Squares (OLS) is a widely used regression algorithm both for statistical analysis and machine learning. The comparison of OLS with the simple weighted average of the molecules' neat RON aims to help in the detection of prevailing synergistic or antagonistic behaviors of the components among the majority of the samples.

Nearest Neighbors (NN) are simple nonparametric methods based on the concept of similarity among objects coexisting in close proximity in the feature space. Two versions of the NN algorithm are tested in this study: in the k-Nearest Neighbors (k-NN) approach, the number of neighbors set through the hyperparameter (i.e., a parameter that is fixed *a priori*) k are averaged, while in the radius-based approach (r-NN) objects within a tuned radius r to the query point are averaged.

Support Vector Machines are powerful algorithms suitable for small complex sets. The models are built on support vectors or data points that carry more critical information than others, reducing the impact of outliers. Nonlinear regression models using kernel functions (SVR and NuSVR) are tested against the linear version of the algorithm (LinSVR).

Decision Trees (DT) are white-box algorithms that infer simple decision rules from training data. Unlike most ML algorithms, DT results can be visually interpreted and provide information on the contribution of the different variables to the predictions.

The Random Forest (RF) algorithm is an ensemble method that constructs and averages multiple decision trees. This method was developed to overcome some of the weaknesses of simple DT, such as overfitting. RF is less sensitive to noisy data and provides more stable results. Wrong predictions made by individual trees are counteracted by other trees.

2.4. Modeling Approach. The eight selected algorithms were implemented using the Scikit-learn library for the Python programming language.[41] Volume and mole-based compositions are used as input features and RON is defined as the target variable. Figure 2 shows an overview of the modeling approach. The collected data set was randomly divided into two groups for training and testing purposes. An 80/20 ratio was used, with 80% of the data points used in the training phase and the remaining 20% of the data reserved to test the performance of the models. The test set serves the purpose of external validation as well, as it was not shown to the algorithms during the training phase. Figure 3 shows the distribution of the collected data set according to RON values and the data splitting for training and testing phases in the

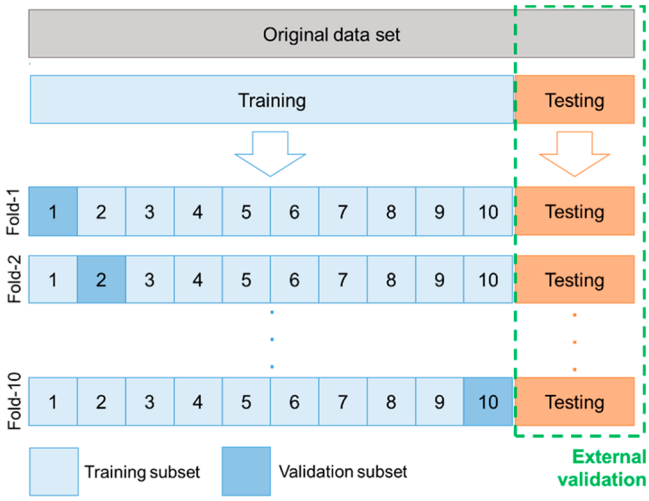


Figure 2. Internal validation of the models is carried out using a 10-fold cross-validation approach, while 20% of the data set was set aside for external validation (testing).

https://doi.org/10.1021/acs.energyfuels.1c00749
Energy Fuels 2021, 35, 9332−9340

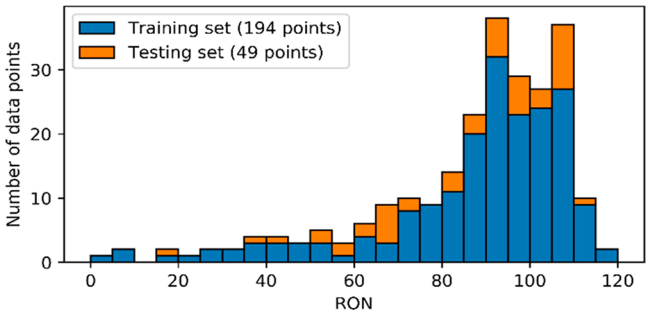


Figure 3. Data distribution according to RON values and training/test splitting.

different ranges. Data was mainly concentrated in the range between 80 and 110, which corresponds to the most common RON values in refinery streams and commercial gasolines.

The relatively small size of the data set motivates the use of *k*-fold cross-validation during the training to reduce overfitting risk with *k* equal to 10. It also allows for hyperparameter tuning for each algorithm and provides an estimation of the generalization performances of the final estimators for preliminary model comparison. Cross-validation uses dynamic validation subsets to identify optimal algorithm settings in an unbiased way. The training set is initially divided into *k* subsets and the algorithm is trained *k* times as shown in Figure 2. Over those *k*-folds, the best values for the hyperparameters are identified by retaining one of the subsets for validation and using the remaining $k - 1$ for training.

Grid search was the approach used for hyperparameter tuning of the algorithms in the learning stage. The suitable combination of hyperparameters for each algorithm was found by an exhaustive search over a manually defined search subspace, guided by cross-validation metrics and the estimator's score method. Table 2 presents the explored hyperparameter space for each algorithm and the optimal values for the two data sets in this study.

Only hyperparameters with the highest potential to affect model performance according to literature have been considered for tuning. Optimal hyperparameters were retained and used in the training phase, and the trained models were used to predict the RON of the test set.

3. RESULTS AND DISCUSSION

This study aims to evaluate the implementation of ML as a predictive tool by assessing popular algorithms and identifying the best candidates. The performance of the models both in the training and testing phases was compared using two different metrics. The coefficient of determination (R^2) reflects the proportion of the RON variance that is captured by the models. The Mean Absolute Error (MAE) was carefully examined since it can help identify potential algorithms with the capacity to predict RON with a deviation within the error range of the standard experimental testing procedure.

3.1. Training. In the training stage, cross-validation results showed a low standard deviation for R^2 for all models, which indicates adequacy and consistency of the data set and suggests a low risk of overfitting. Regression metrics from the training stage are shown in Figure 4 and reflect the performance of the validation subset over the 10 cross-validation folds. Training results suggest nonlinear SVR algorithms as the best candidate to predict RON. On both a molar and volumetric basis, SVR and NuSVR models achieved an average R^2 of 0.99 and a

Table 2. Hyperparameters Tuned Using Grid Search and Optimal Combinations for the Two Data Sets in This Study

algorithm	hyperparameter	best (volume-based data set)	best (mole-based data set)
OLS	N/A	N/A	N/A
k-NN	number of neighbors	7	5
	weights	distance	distance
	metric	Manhattan	Euclidean
r-NN	radius	1.9	1.9
	weight	distance	distance
	metric	Manhattan	Manhattan
LinSVR	epsilon	2	2
	C	10^4	10^4
SVR	kernel function	RBF	RBF
	epsilon	0.5	1
	gamma	0.1	0.1
	C	10^4	10^3
NuSVR	kernel function	RBF	RBF
	gamma	0.1	0.1
	C	10^4	10^4
	nu	0.5	0.3
DT	maximum depth	17	15
	minimum sample split	2	3
	splitter	random	random
	criterion	MSE	MSE
RF	maximum depth	12	13
	minimum sample split	2	2
	criterion	MSE	MAE
	number of trees	24	128

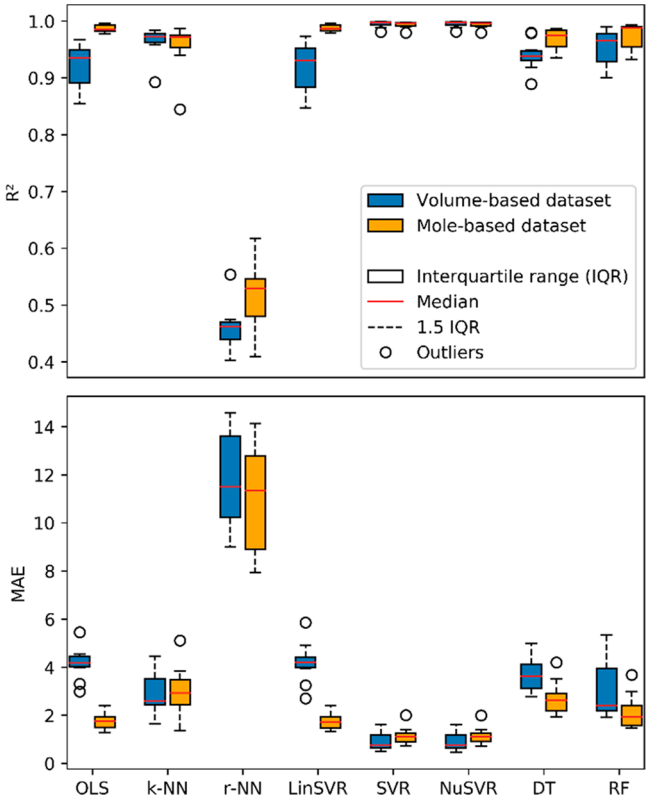


Figure 4. Comparison of the response of the training subset in the 10-fold cross-validation.

https://doi.org/10.1021/acs.energyfuels.1c00749
Energy Fuels 2021, 35, 9332−9340

standard deviation of 0.006. Conversely, the r-NN algorithm showed poor accuracy and MAE that exceeded 10 octane numbers. Cross-validation results also showed that linear models, OLS and LinSVR, reduce their predictive error when trained with molar composition data by 57% and 76%, respectively.

3.2. Testing. The models were tested on the 49 data points included in the test set. The results from the testing phase were in agreement with the cross-validation performance of the models. R^2 and MAE for the eight algorithms and two data sets are reported in Table 3, while Figure 5 shows the error distribution and outliers behavior.

Table 3. Performance of the Eight Algorithms Used in This Study on the Test Set

	volume-based data set		mole-based data set	
model	R^2	MAE	R^2	MAE
OLS	0.9255	3.9644	0.9884	1.7313
k-NN	0.9571	3.2107	0.9606	3.0234
r-NN	0.4891	12.1186	0.4766	12.3332
LinSVR	0.918	4.1832	0.988	1.7567
SVR	0.9962	0.9224	0.9903	1.412
NuSVR	0.9964	0.9072	0.9903	1.418
DT	0.8714	4.9806	0.9393	3.5622
RF	0.9701	2.636	0.9852	2.0741

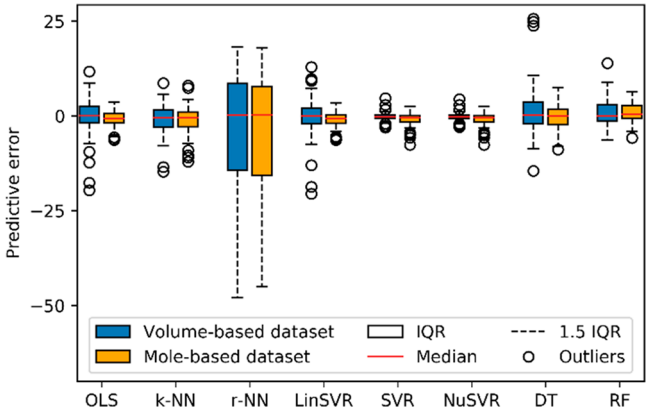


Figure 5. Prediction error distribution for points in the test set.

As shown in Figure 5, using molar data improved the accuracy of the estimators for most models (and no significant impact was observed for the rest). The remarkable improvement in the case of linear algorithms for the molar-based approach can be explained by the fact that fuel interaction in the engine cylinder during a RON test occurs in the gas phase,[36] where various chemical kinetic interactions between fuel compounds can occur. An example could be a competition between various chemical compounds for small radicals, which would lead to different reaction pathways.[42] Such interactions affect the global reactivity of the combustible mixture and the autoignition properties of the fuel. This, in turn, could lead to the deviation from the RON additivity based estimations toward synergistic or antagonistic blending effects. As the reaction rates for chemical compounds are better explained by molar concentrations rather than volumetric, models developed based on the molar rule turned out to have superior performance. Nevertheless, the developed models apply to hydrocarbons (linear paraffins, isoparaffins, olefins, naphthenes, aromatic compounds = PIONA) and oxygenated species including ethanol and ETBE. The molar rule for novel synthetic or biobased fuel compounds such as esters, ketones, etc. could be used as well. However, that would require the development of new models based on the experimental database of RON measurements for gasoline blends that include those new blendstocks. The current models could be applied only to RON estimations for various molar concentrations of compound classes that were taken into the training of the model (PIONA + OH and E). Moreover, the limited availability of data for certain components in the blends used to build this model (ethers and naphtenes in particular have only been present in 10 and 6 blends, respectively) may lead to lower prediction accuracies in some cases.

Despite the simplicity of the OLS algorithm, its results are comparable to more complex estimators like LinSVR. In both cases, the largest deviations from the experimental results occurred for high RON values close to 120 as shown in Figure 6. The reason for those errors is a low data availability in

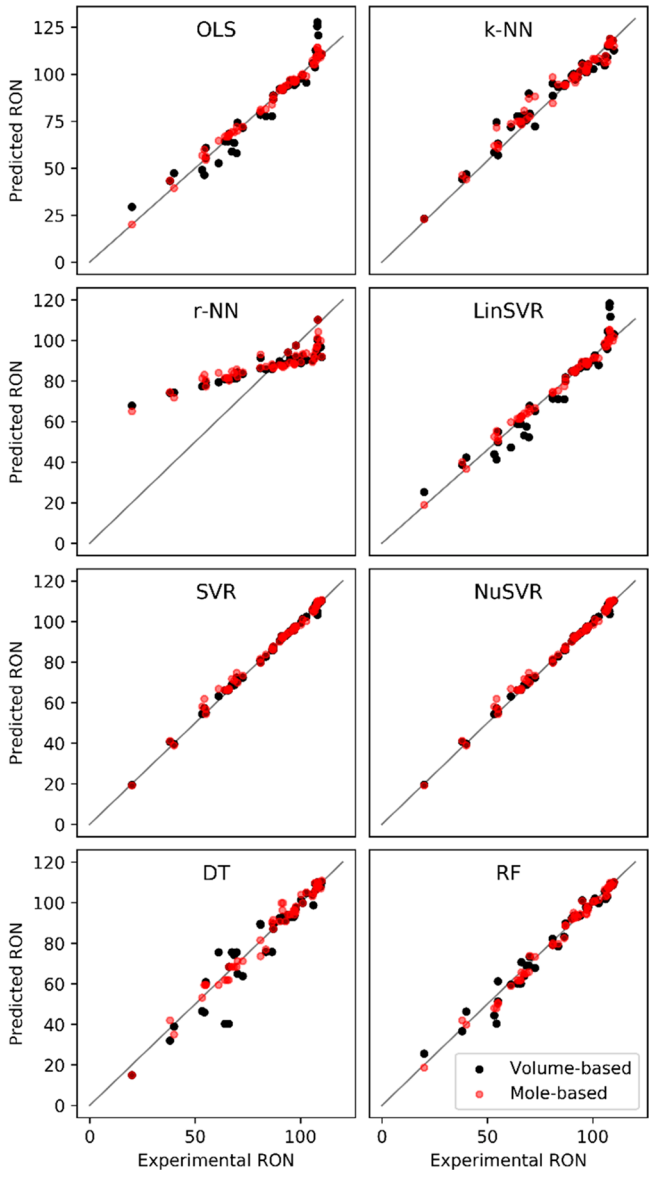


Figure 6. Comparison of experimental and predicted RON values for the eight algorithms in this study.

https://doi.org/10.1021/acs.energyfuels.1c00749
Energy Fuels 2021, 35, 9332−9340

literature for such high RON value fuels, as presented in Figure 3. In practice, high RON value refinery streams and compounds serve as octane boosters that are blended with gasoline to increase RON and therefore meet fuel standards. k-NN models outperformed r-NN estimators and the difference is extremely notable for low and high RON values. The distribution of the data set and the mechanism used by r-NN explain this difference. Predictions for query points located in regions with low data density are computed using fewer neighbors than in rich areas, where the data points are not necessarily homogeneously distributed. The k-NN model showed acceptable results but its predictive accuracy was far from outstanding.

Figure 6 also suggests that the use of ensemble methods like RF potentially improves the models compared to the use of simple DTs, especially for low RON values. Furthermore, it can be seen from Table 2 that the volume-based model was trained using 24 trees while the mole-based included 128. Despite those numbers of estimators being optimal for each data set, this difference was further analyzed and increasing the number of trees beyond 24 had almost no impact on the results. Therefore, the difference in the performance can be attributed to intrinsic characteristics on the data sets. An advantage of both DT and RF algorithms is their white-box nature. This allows for the interpretability of the results, although a large number of trees in the RF might make the analysis process more cumbersome.

Hyperparameter tuning of the NuSVR algorithm yielded an equivalent model to the SVR algorithm; thus, similar results were obtained. The accuracy of these two models was the highest among all methods compared in the present study. These results confirm the suitability of these algorithms for small and complex data sets. By projecting the features into a higher dimensional space through a radial-based kernel function, the SVR algorithm is able to retain nonlinearities within the training data. Figure 7 shows the effectiveness of the kernel trick by comparing the predictions made by the LinSVR and the SVR algorithms in a binary mixture of three hydrocarbons (*n*-heptane, isooctane, and toluene from top to bottom) and ethanol.

The same comparison for more complex blends included in the test set is shown in Figure 8. The use of a nonlinear approach reduced the inaccuracy of the prediction for the great majority of the samples in the test set. R^2 improved from 0.918 to 0.996 and the MAE dropped from 4.18 to 0.92. Prevailing errors corresponded in most cases to samples with a high content of ethanol, in part due to the absence of the pure molecule in the training set. In light of these results, including neat components manually in the training set could be a possible strategy to improve the models when the objective is to optimize the model performance for the component in question (e.g., a preliminary trial with ethanol resulted in a slight improvement of accuracy on ethanol-containing blends, but no change in the overall performance of the model).

To ratify the robustness of SVM methods, a sensitivity analysis of the models was carried out. The three algorithms with the highest accuracy (i.e., SVR, k-NN, and RF) were fitted using 16 alternative data sets derived from the original ones by applying restrictions to the RON or to the number of components present in the blends. To specify two examples: (1) the data set was limited to data points that fall into the RON range of 80−100; (2) the data set was limited to ternary blends (i.e., only 3 out of the 7 components were present).

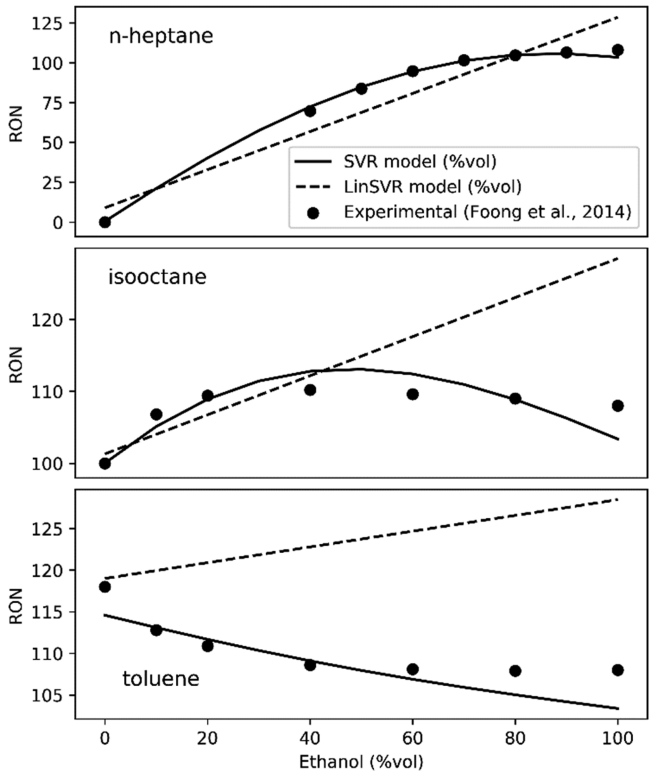


Figure 7. Comparison of RON predictions for linear and nonlinear SVR models in binary mixtures of hydrocarbons and ethanol using volume-based compositions.

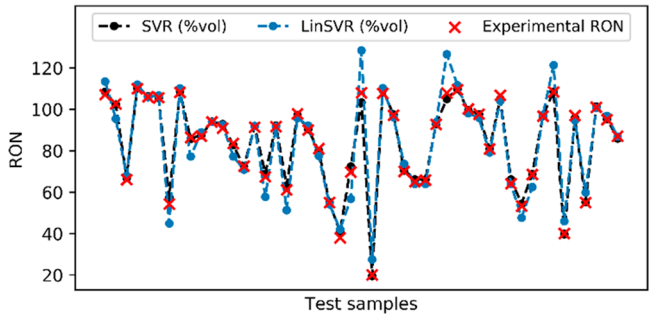


Figure 8. Predictive performance of linear and nonlinear SVR models to predict RON on the test set points, as compared to the corresponding experimentally determined RON values obtained from the literature. The horizontal axis represents individual points in the test set.

Unlike the other algorithms, under changing circumstances, SVR models maintained their predictive accuracy, with the MAE kept below 1.2 octane numbers in all cases. These results highlight the robustness of the method and its suitability for the studied application and other similar tasks. Ultimately, achieving consistent predictions within the reproducibility limit[43] will rely on the availability of more data.

4. CONCLUSIONS

RON is used as a quality indicator for gasoline fuels; however, its prediction in complex blends is not trivial. The increase of oxygenated species like alcohol or ethers in gasoline blends highlights the importance of new models able to capture nonlinear interactions and provide more accurate estimations for this property. This study investigated the suitability of eight

https://doi.org/10.1021/acs.energyfuels.1c00749
Energy Fuels 2021, 35, 9332−9340

traditional ML algorithms in the prediction of RON using simplified representations of gasoline and gasoline fractions. A palette of seven species was defined to provide an accurate representation of the fuels. Moreover, volume and mole-based data sets were used to evaluate the response and applicability of the estimators.

SVR proved to be the best algorithm for RON predictions, confirming its capacity to handle small and complex data sets. Especially when the model is trained and used with volumetric concentrations, it shows a predictive accuracy comparable to the traditional RON testing procedure in a Cooperative Fuel Research (CFR) engine, although consistent prediction within the reproducibility limit will most likely require more data. Moreover, the resulting model captures the nonlinear interactions between blending components and is able to reflect both synergistic and antagonistic effects between molecules. These findings show that modestly sized data sets do not impede the development of accurate predictive tools insofar consistent data is available. The comparison of volume and mole-based models also showed an improvement in the performance of linear algorithms when the molar composition was used, reinforcing the idea of a possible linear-by-mole blending rule.

The methodology proposed in this paper and the models derived from it are expected to foster future gasoline production and accelerate the transition toward alternative and more sustainable fuels. The obtained results also suggest that the developed approach could be applied to predict other key gasoline properties with nonlinear behavior, such as distillation curves and other volatility-related properties. Nevertheless, the extension of the applicability of these models is subjected to an increment in the number of species used for the fuel's representation.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.energyfuels.1c00749.

> Database used for this study, compiled from refs 4, 7, 27, 29−31, 34, 37, and 44−50, including the compositions and RON values of the individual samples along with a reference to the source of each sample (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Annukka Santasalo-Aarnio − *Research Group of Energy Conversion, Department of Mechanical Engineering, School of Engineering, Aalto University, FI-00076 Aalto, Finland;* orcid.org/0000-0003-0077-966X; Email: annukka.santasalo@aalto.fi

Authors

Sandra Correa Gonzalez − *Research Group of Energy Conversion, Department of Mechanical Engineering, School of Engineering, Aalto University, FI-00076 Aalto, Finland*

Yuri Kroyan − *Research Group of Energy Conversion, Department of Mechanical Engineering, School of Engineering, Aalto University, FI-00076 Aalto, Finland*

Teemu Sarjovaara − *Technology Centre, Neste Corporation, 06101 Porvoo, Finland*

Ulla Kiiski − *Technology Centre, Neste Corporation, 06101 Porvoo, Finland*

Anna Karvo − *Technology Centre, Neste Corporation, 06101 Porvoo, Finland*

Arpad I. Toldy − *Research Group of Energy Conversion, Department of Mechanical Engineering, School of Engineering, Aalto University, FI-00076 Aalto, Finland;* orcid.org/0000-0001-9027-4854

Martti Larmi − *Research Group of Energy Conversion, Department of Mechanical Engineering, School of Engineering, Aalto University, FI-00076 Aalto, Finland*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.energyfuels.1c00749

Author Contributions

S.C.G.: conceptualization, methodology, software, formal analysis, writing−original draft. Y.K.: methodology, writing−review and editing. T.S.: funding acquisition, writing−review and editing; U.K.: writing−review and editing. A.K.: writing−review and editing. A.I.T.: writing−review and editing, visualization. M.L.: supervision, funding acquisition. A.S.-A.: funding acquisition, resources, supervision, writing−review and editing.

Funding

This work was funded by Neste Corporation.

Notes

The authors declare no competing financial interest.

■ ABBREVIATIONS

A = Aromatics
ANN = Artificial Neural Network
CFR = Cooperative Fuel Research
DT = Decision Trees
E = Ethers
ETBE = Ethyl *tert*-butyl-ether
FACE = Fuels for Advanced Combustion Engines
I = Isoparaffins
ICE = Internal Combustion Engine
IQR = Interquartile Range
k-NN = k-Nearest Neighbors
LinSVR = Linear Support Vector Regression
MAE = Mean Absolute Error
ML = Machine Learning
MON = Motor Octane Number
MSW = Municipal solid waste
N = Naphthenes
NIR = Near-infrared
NMR = Nuclear Magnetic Resonance
NN = Nearest Neighbors
NuSVR = Nu Support Vector Regression
O = Olefins
OH = Alcohols
OLS = Ordinary Least Squares
P = Linear paraffins
PLS = Partial Least Squares
PRF = Primary Reference Fuels
R^2 = Coefficient of determination
RF = Random Forest
r-NN = Radius-Nearest Neighbors
RON = Research Octane Number
RVP = Reid Vapor Pressure
S = Octane Sensitivity
SI = Spark-Ignition
SVM = Support Vector Machines

SVR = Support Vector Regression

REFERENCES

(1) *Data & Statistics*; International Energy Agency. https://www.iea.org/data-and-statistics (accessed 2020-03-05).

(2) Fraser, N.; Blaxill, H.; Lumsden, G.; Bassett, M. Challenges for Increased Efficiency through Gasoline Engine Downsizing. *SAE Int. J. Engines* **2009**, *2*, 991−1008.

(3) Leone, T. G.; Anderson, J. E.; Davis, R. S.; Iqbal, A.; Reese, R. A.; Shelby, M. H.; et al. The Effect of Compression Ratio, Fuel Octane Rating, and Ethanol Content on Spark-Ignition Engine Efficiency. *Environ. Sci. Technol.* **2015**, *49*, 10778−89.

(4) D02 Committee. *Test Method for Research Octane Number of Spark-Ignition Engine Fuel*; ASTM International. DOI: 10.1520/D2699-15.

(5) Schoen, W. F.; Mrstik, A. V. Calculating Gasoline Blend Octane Ratings. *Ind. Eng. Chem.* **1955**, *47*, 1740−2.

(6) Metwally, M. M. Approach to Accurate Octane Number Calculation for Gasoline Blending. *Acad. Res. Community Publ* **2019**, *2*, 506−506.

(7) Ghosh, P.; Hickey, K. J.; Jaffe, S. B. Development of a Detailed Gasoline Composition-Based Octane Model. *Ind. Eng. Chem. Res.* **2006**, *45*, 337−45.

(8) Kirgina, M. V.; Ivanchina, E. D.; Dolganov, I. M.; Chekantsev, N. V.; Kravtsov, A. V.; Fu, F. Computer Program for Optimizing Compounding of High-Octane Gasoline. *Chem. Technol. Fuels Oils* **2014**, *50*, 17−27.

(9) Kelly, J. J.; Barlow, C. H.; Jinguji, T. M.; Callis, J. B. Prediction of gasoline octane numbers from near-infrared spectral features in the range 660−1215 nm. *Anal. Chem.* **1989**, *61*, 313−20.

(10) Bohács, G.; Ovádi, Z.; Salgó, A. Prediction of Gasoline Properties with near Infrared Spectroscopy. *J. Near Infrared Spectrosc.* **1998**, *6*, 341−8.

(11) Litani-Barzilai, I.; Sela, I.; Bulatov, V.; Zilberman, I.; Schechter, I. On-line remote prediction of gasoline properties by combined optical methods. *Anal. Chim. Acta* **1997**, *339*, 193−9.

(12) Schweidtmann, A. M.; Rittig, J. G.; König, A.; Grohe, M.; Mitsos, A.; Dahmen, M. Graph Neural Networks for Prediction of Fuel Ignition Quality. *Energy Fuels* **2020**, *34*, 11395−407.

(13) Kubic, W. L.; Jenkins, R. W.; Moore, C. M.; Semelsberger, T. A.; Sutton, A. D. 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*, 12236−45.

(14) Li, R.; Herreros, J. M.; Tsolakis, A.; Yang, W. Machine learning regression based group contribution method for cetane and octane numbers prediction of pure fuel compounds and mixtures. *Fuel* **2020**, *280*, 118589.

(15) vom Lehn, F.; Cai, L.; Tripathi, R.; Broda, R.; Pitsch, H. A property database of fuel compounds with emphasis on spark-ignition engine applications. *Appl. Energy Combust Sci.* **2021**, *5*, 100018.

(16) Lee, S.; Choi, H.; Cha, K.; Chung, H. Random forest as a potential multivariate method for near-infrared (NIR) spectroscopic analysis of complex mixture samples: Gasoline and naphtha. *Microchem. J.* **2013**, *110*, 739−48.

(17) Voigt, M.; Legner, R.; Haefner, S.; Friesen, A.; Wirtz, A.; Jaeger, M. Using fieldable spectrometers and chemometric methods to determine RON of gasoline from petrol stations: A comparison of low-field 1H NMR@80 MHz, handheld RAMAN and benchtop NIR. *Fuel* **2019**, *236*, 829−35.

(18) Abdul Jameel, A. G.; Van Oudenhoven, V.; Emwas, A.-H.; Sarathy, S. M. Predicting Octane Number Using Nuclear Magnetic Resonance Spectroscopy and Artificial Neural Networks. *Energy Fuels* **2018**, *32*, 6309−29.

(19) van Leeuwen, J. A.; Jonker, R. J.; Gill, R. Octane number prediction based on gas chromatographic analysis with non-linear regression techniques. *Chemom. Intell. Lab. Syst.* **1994**, *25*, 325−40.

(20) Liu, M.; Zhou, P.; Kong, P.; Yang, C.; Li, G.; Mu, M. Determination of octane number of gasoline by double ANN algorithm combined with multidimensional gas chromatography. *2010 Sixth Int. Conf. Nat. Comput.* **2010**, *3*, 1640−2.

(21) Anderson, J. E.; Wallington, T. J. Novel Method to Estimate the Octane Ratings of Ethanol-Gasoline Mixtures Using Base Fuel Properties. *Energy Fuels* **2020**, *34*, 4632−42.

(22) AlRamadan, A. S.; Sarathy, S. M.; Badra, J. Unraveling the octane response of gasoline/ethanol blends: Paving the way to formulating gasoline surrogates. *Fuel* **2021**, *299*, 120882.

(23) Albahri, T. A. Specific Gravity, RVP, Octane Number, and Saturates, Olefins, and Aromatics Fractional Composition of Gasoline and Petroleum Fractions by Neural Network Algorithms. *Pet. Sci. Technol.* **2014**, *32*, 1219−26.

(24) Pasadakis, N.; Gaganis, V.; Foteinopoulos, C. Octane number prediction for gasoline blends. *Fuel Process. Technol.* **2006**, *87*, 505−9.

(25) Murty, B. S. N.; Rao, R. N. Global optimization for prediction of blend composition of gasolines of desired octane number and properties. *Fuel Process. Technol.* **2004**, *85*, 1595−602.

(26) Curran, H. J.; Gaffuri, P.; Pitz, W. J.; Westbrook, C. K. A comprehensive modeling study of iso-octane oxidation. *Combust. Flame* **2002**, *129*, 253−80.

(27) Morgan, N.; Smallbone, A.; Bhave, A.; Kraft, M.; Cracknell, R.; Kalghatgi, G. Mapping surrogate gasoline compositions into RON/MON space. *Combust. Flame* **2010**, *157*, 1122−31.

(28) Cancino, L. R.; Fikri, M.; Oliveira, A. A. M.; Schulz, C. Ignition delay times of ethanol-containing multi-component gasoline surrogates: Shock-tube experiments and detailed modeling. *Fuel* **2011**, *90*, 1238−44.

(29) Singh, E.; Badra, J.; Mehl, M.; Sarathy, S. M. Chemical Kinetic Insights into the Octane Number and Octane Sensitivity of Gasoline Surrogate Mixtures. *Energy Fuels* **2017**, *31*, 1945−60.

(30) Wolk, B.; Ekoto, I.; Northrop, W. Investigation of Fuel Effects on In-Cylinder Reforming Chemistry Using Gas Chromatography. *SAE Int. J. Engines* **2016**, *9*, 964−78.

(31) Perez, P. L.; Boehman, A. L. Experimental Investigation of the Autoignition Behavior of Surrogate Gasoline Fuels in a Constant-Volume Combustion Bomb Apparatus and Its Relevance to HCCI Combustion. *Energy Fuels* **2012**, *26*, 6106−17.

(32) Daly, S. R.; Niemeyer, K. E.; Cannella, W. J.; Hagen, C. L. FACE Gasoline Surrogates Formulated by an Enhanced Multivariate Optimization Framework. *Energy Fuels* **2018**, *32*, 7916−32.

(33) Correa Gonzalez, S. Modeling the effect of blending multiple components on gasoline properties. Masters Thesis, Aalto University, 2019.

(34) Badra, J.; AlRamadan, A. S.; Sarathy, S. M. Optimization of the octane response of gasoline/ethanol blends. *Appl. Energy* **2017**, *203*, 778−93.

(35) Hu, E.; Ku, J.; Yin, G.; Li, C.; Lu, X.; Huang, Z. Laminar Flame Characteristics and Kinetic Modeling Study of Ethyl Tertiary Butyl Ether Compared with Methyl Tertiary Butyl Ether, Ethanol, iso-Octane, and Gasoline. *Energy Fuels* **2018**, *32*, 3935−49.

(36) Anderson, J. E.; Leone, T. G.; Shelby, M. H.; Wallington, T. J.; Bizub, J. J.; Foster, M. et al. *Octane Numbers of Ethanol-Gasoline Blends: Measurements and Novel Estimation Method from Molar Composition*; SAE International: Warrendale, PA, 2012. DOI: 10.4271/2012-01-1274.

(37) Foong, T. M.; Morganti, K. J.; Brear, M. J.; da Silva, G.; Yang, Y.; Dryer, F. L. The octane numbers of ethanol blended with gasoline and its surrogates. *Fuel* **2014**, *115*, 727−39.

(38) AlRamadan, A. S.; Sarathy, S. M.; Khurshid, M.; Badra, J. A blending rule for octane numbers of PRFs and TPRFs with ethanol. *Fuel* **2016**, *180*, 175−86.

(39) James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An introduction to statistical learning*; Springer, 2013; Vol. *112*.

(40) Ng, A. What data scientists should know about deep learning, **2015**.

(41) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; et al. Scikit-learn: Machine Learning in Python. *J. Mach Learn Res.* **2011**, *12*, 2825−2830.

https://doi.org/10.1021/acs.energyfuels.1c00749
Energy Fuels 2021, 35, 9332−9340

(42) Song, H.; Dauphin, R.; Vanhove, G. A kinetic investigation on the synergistic low-temperature reactivity, antagonistic RON blending of high-octane fuels: Diisobutylene and cyclopentane. *Combust. Flame* **2020**, *220*, 23−33.
(43) Knop, V.; Loos, M.; Pera, C.; Jeuland, N. A linear-by-mole blending rule for octane numbers of n-heptane/iso-octane/toluene mixtures. *Fuel* **2014**, *115*, 666−73.
(44) Kalghatgi, G.; Babiker, H.; Badra, J. A Simple Method to Predict Knock Using Toluene, N-Heptane and Iso-Octane Blends (TPRF) as Gasoline Surrogates. *SAE Int. J. Engines* **2015**, *8*, 505−19.
(45) Badra, J. A.; Bokhumseen, N.; Mulla, N.; Sarathy, S. M.; 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.
(46) Sarathy, S. M.; Kukkadapu, G.; Mehl, M.; Javed, T.; Ahmed, A.; Naser, N.; et al. Compositional effects on the ignition of FACE gasolines. *Combust. Flame* **2016**, *169*, 171−93.
(47) Li, B.; Jiang, Y. Chemical Kinetic Model of a Multicomponent Gasoline Surrogate with Cross Reactions. *Energy Fuels* **2018**, *32*, 9859−71.
(48) Cameron, D. M. Autoignition Studies of Gasoline Surrogate Fuels in the Advanced Fuel Ignition Delay Analyzer. Masters Thesis, University of Colorado Boulder, 2017.
(49) Ogura, T.; Sakai, Y.; Miyoshi, A.; Koshi, M.; Dagaut, P. Modeling of the Oxidation of Primary Reference Fuel in the Presence of Oxygenated Octane Improvers: Ethyl Tert-Butyl Ether and Ethanol. *Energy Fuels* **2007**, *21*, 3233−9.
(50) Cancino, L. R.; Fikri, M.; Oliveira, A. A. M.; Schulz, C. Autoignition of gasoline surrogate mixtures at intermediate temperatures and high pressures: Experimental and numerical approaches. *Proc. Combust. Inst.* **2009**, *32*, 501−8.

https://doi.org/10.1021/acs.energyfuels.1c00749
Energy Fuels 2021, 35, 9332−9340