Numerical Approaches to Determine Cetane Number of Hydrocarbons and Oxygenated Compounds, Mixtures, and their Blends

Benoit Creton, Nathalie Brassart, Amandine Herbaut, Mickael Matrat

Numerical approaches to determine cetane number of hydrocarbons and oxygenated compounds, mixtures and their blends.

Benoit Creton[1,*], *Nathalie Brassart*[1], *Amandine Herbaut*[2], *Mickael Matrat*[1]

[1] IFP Energies nouvelles, 1 et 4 Avenue de Bois-Préau, 92852 Rueil-Malmaison, France.

[2] Service de l'Energie Opérationnelle, Case n°68, 60 Boulevard du Général Martial Valin, CS21623, 75509 Paris CEDEX 15, France.

[*] To whom the correspondence should be addressed. E-mail: benoit.creton@ifpen.fr

ABSTRACT. In the present work, we report the development and use of models to predict the cetane number of hydrocarbons and oxygenated compounds, mixtures and their blends. The study is divided in three steps: (i) the prediction of pure compounds CN using ML-based approaches, (ii) the development and the application of mixing rules, (iii) the external validation of models on a set of real fuels. Experimental CN values for 658 pure compounds are collected from the literature and merged to obtain a consistent and comprehensive database. ML-based models are then trained on the database. A second database is built from the collection of 572 experimental CN values for mixtures. Existing and proposed mixing rules powered either by experimental CN or CN predicted using the ML-based models are then assessed on the basis of the second database.

The new mixing rule involving the activity coefficients of mixtures' components shows the best performances. Finally, the application of our predictive numerical approach to 27 real fuels demonstrates its accuracy and relevance, and that it could be further used for testing large numbers of samples.

KEYWORDS. Cetane number, hydrocarbons, oxygenated compounds, mixtures, machine learning, mixing rules

1. Introduction

One of the most stringent properties related to combustion lies in the control of fuel ignition which can be expressed by the cetane number (CN). The cetane number reflects the tendency of a fluid to auto-ignite when exposed to pressure and heat, as it typically occurs in diesel or compression-ignition (CI) engines under working conditions reaching for instance in-cylinder pressures and temperatures up to 8 MPa and 1000 K, respectively.[1,2] Mainly due to their high efficiency and durability, CI motors are used worldwide in agricultural and industrial machinery, stationary power generators and obviously as sources of motion for many applications.

The use of such combustion systems has also been associated for decades with significant emissions, particularly in terms of nitrogen oxides, particles or even unburned hydrocarbons.[3,4] In addition, the use of fossil fuels contributes to increasing the effects of greenhouse gases e.g., carbon dioxide released into the atmosphere, resulting in climate changes currently observed. These issues impose various research challenges to be solved, including identifying new technologies and potential alternative energy sources to meet the demand and supply.[5] Emissions mitigation has led to the development of more efficient combustion systems operating at lower temperature. Gas turbines used in aircraft engines are also following this trend through the development of premixed lean combustion engines. However, operating in lean conditions contributes to increase combustion instabilities leading potentially to blow out. In this context, CN is considered as an important property for jet fuels to prevent lean blow out (LBO),[6] but also, more generally, for the development of alternative fuels. The CN of a fuel can be reduced by adding products issued from new energy pathways, notably Alcohol-to-Jet produced from isobutene or isobutanol,[7] contributing to affect LBO occurrence in gas turbines. Diesel fuels are also affected

by the use and/or incorporation of alternative fuels that include typical hydrocarbons present in fossil-based references or frequently oxygenated components with different reactivities.[8]

Evaluating alternative fuel properties is thus mandatory to ensure a safe and efficient use among the different applications. Their characterization is a challenge as low carbon fuels (LCF) encompass various options, including biogenic fuels from various feedstocks and synthetic fuels, e.g., hydrotreated vegetable oil, Fischer-Tropsch based processes, fermentation, and other thermochemical processes. These processes lead to different fuel compositions that can differ significantly compared to fossil references. Their impact on CN values has been demonstrated in the literature including frequently the use of additives to overcome the cetane number reduction. In a recent review, Benajes et al. indicated that blending different types of LCFs and including adequate additives can enhance combustion properties, emissions reductions, and the efficiency of CI engines.[9] Mohammed et al. investigated effects of different chemicals such as alcohols and ethers, in blends to improve CI engine performance and emission characteristics.[10] The use of relevant additives and cetane improvers,[10,11] in gasolines or jet fuels can lead to diesel-like fuels. The approach consisting in using a single fuel for all types of vehicles strongly simplified the storage and transport logistics, making this solution attractive in both a civil and military context. Military applications can rely on the single fuel policy where jet fuel is a unique reference for both aircraft and ground transports.[12]

Since first attempts in developing predictive approaches based on machine learning (ML),[13–15] many models have been proposed in the literature to estimate the CN of molecules.[16] Non-exhaustively focusing on recent years, models based on quantitative structure-property relationships (QSPR), including group contributions, have been developed to predict CN values.[17–24] In these latter works, different databases were used as support and made available via the

supporting information of the articles, the most extensive contains up to 630 CN values for hydrocarbons and oxygenated compounds.[20] Molecular structures were encoded either considering functional group counts or more abstract molecular descriptors extracted from 3-dimensional geometries. The use of algorithms such as neural networks or support vector machines (SVM) were reported in several of these works. From conclusions drawn by the authors, the use of ML-based models combined with a functional group encoding of molecules, obviously leads to interesting approaches to predict the CN for a molecule of which only its structure is known.[13–15,17–24]

CN for a fluid is derived considering a virtual volumetric mixing of n-hexadecane (CN = 100) and 2,2,4,4,6,8,8-heptamethylnonane (CN = 15) that produces a similar ignition delay. Therefore, CN for a mixture is by definition the sum of each constituent's CN weighted with its volumetric fraction.[25–27] However, non-ideal behaviours were observed when mixing blends, different chemical families or even with the introduction of oxygenated compounds, leading some authors to propose an additional adjustable parameter β related to the composition of blends.[27] Witkowski et al. proposed a multi-parameter group contribution working with mole fractions to predict ignition delay (τ) values which were then converted to CN values using the following relation: $CN = a + b/\tau$, with a = 4.46 and b = 186 ms. Li et al. proposed ML regressions based on a group contribution method to predict CN for pure compounds and mixtures, and showed a benefit of taking mixtures into account when training models.[22] Very recently, Sheyyab et al. represented fluids – hydrocarbon compounds and mixtures – with some of functional groups as defined within the UNIFAC (universal functional activity coefficient) method, then trained ML algorithms on CN values and obtained models with good predictive capabilities.

In this work, we propose to investigate the prediction of CN for mixtures only having information about their compositions. Two strategies were followed: (i) Pure compounds CN values were collected from the literature and merged to obtain a consistent and comprehensive database. ML-based models were then trained on the database. In parallel with these actions, a second database was built containing CN values for mixtures, including fuel surrogates. Existing and proposed mixing rules powered by experimental and predicted CN with the ML-based models were then assessed. (ii) The two databases were merged, and ML-based models were trained. The article is organized as follows: after presenting the data collections and curation methods and the strategy followed to develop new QSPR based models as well as new mixing rules, we expose the predictive performances of models and discuss their utilization for external data predictions for real fuels, before concluding.

2. Materials and methods

One of the key points for developing QSPR-based models is collecting data that will serve as a reference during the learning process. To this end, the database proposed in the Saldana et al. work – originally derived from the work of Creton et al.[15] and the compendium of Murphy et al.[28] – was used as a starting point including 329 CN values for hydrocarbons and oxygenated compounds.[23] Several more recent works have proposed CN databases,[17–22,29] and a merger of these latter databases has been carried out. The result was a collection of cetane number values for 274 hydrocarbons and 384 oxygen compounds. Figure 1 presents the distribution between hydrocarbons and oxygenated compounds in the database, as well as distributions of subfamilies within the two families. It shows that hydrocarbons and oxygenates represent roughly 42% and 58% of chemicals, respectively. A decomposition of hydrocarbons in terms of alkanes, alkenes,

alkynes, and cyclic molecules such as naphthenes and aromatics is proposed. It shows that 30% of hydrocarbons are saturated paraffins (n- and i-alkanes). In this latter class, n- alkanes and i-alkanes are distributed roughly at 22% and 78%, respectively. Unsaturated alkanes account for 15% of the hydrocarbons in the database, while only one CN value of alkyne has been reported in the literature. Cyclic compounds (naphthenes and aromatics) represent more than half of the hydrocarbons in the database. Regarding oxygenated compounds, the database includes, in decreasing order of occurrence: esters, ethers, alcohols, ketones, furans, aldehydes and carboxylic acids. It should be noted that 17 oxygenates are polyfunctionals, i.e. they are constituted of at least two of the latter chemical characteristics. For many compounds, several different values are reported from one source to another, the average value has been taken into account. In case of significant discrepancies between experimental values, outliers were simply discarded. The complete database containing CN values for 658 compounds is available as Supporting Information.

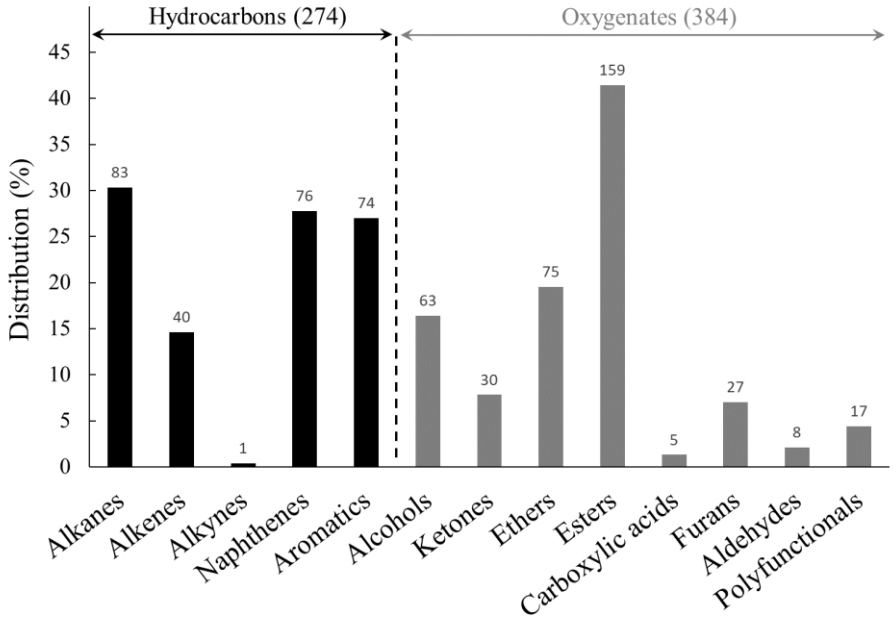

Figure 1. Percentage distributions of hydrocarbons (black) and oxygenates (dark grey) in the pure compounds database. Values in brackets and values above bars stand for the number of compounds in the chemical families and subclasses, respectively.

An additional collection of data was performed to build a database including a series of mixtures for which experimental CN values are known. To this end, works reporting experimental CN values for mixtures were identified in the literature, and mixtures compositions as well as their CN values were extracted.[22,30–44] A merger of these latter data has been carried out resulting in a collection of cetane number values for 572 mixtures after removing duplicates. Mixtures individually involve in-between 2 and 9 components, and a total of 43 hydrocarbons and oxygenates are represented in the database. Furthermore, among the 572 mixtures, only 28 involve at least one oxygenated compound. Figure 2 illustrates mole percentage distributions of subclasses of hydrocarbons and oxygenates in the database. It reveals that 96% of mixture components are hydrocarbons, belonging in decreasing order of occurrence: n-alkanes (nC7 to nC20), i-alkanes (C8 to C18), aromatics (C6 to C15), naphthenes (C5 to C15), alkenes (C6 to C8), and alkynes (C6). The 4% oxygenates are distributed in alcohols (C1 to C6) and ethers (C8 to C12). Considering all collected mixtures, CN values range from 7.3 to 101.5, with an average value of 37. The complete database containing CN values for the 572 mixtures is available as Supporting Information.

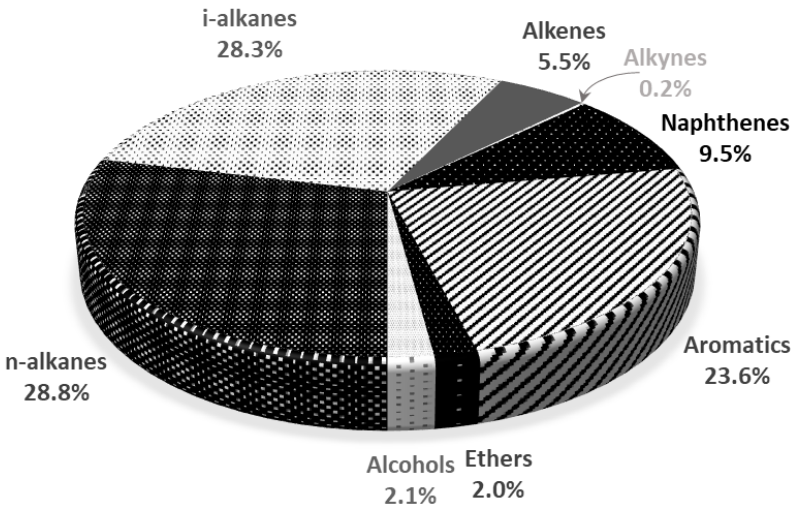

Figure 2. Percentage distributions of hydrocarbons and oxygenates subclasses in the mixture database. The value under each subclass represents its mole percentage representation in the entire database.

From comparisons performed in previous studies,[45–48] each molecule in the pure compound database was encoded using descriptors – labelled as functional group count descriptors (FGCD) – calculated on the basis of the chemical and structural formulae. In the FGCD family of molecular descriptors are included counts of atoms and groups of atoms identified as relevant from chemical aspects. It was demonstrated that such a simple representation of compounds provides relevant descriptors usable for the development of QSPR.[49,50] Simplified molecular input line entry specification (SMILES) codes were assigned to each molecule within the databases. FGCD were automatically generated using the RDKit Python package[51] and SMILES arbitrary target specification (SMARTS) matching functionalities.[52] Table 1 reports SMARTS codes for FGCD – labelled from X1 to X60 – under consideration in this study. For example, FGCD labelled X1 to X3 denote the number of hydrogen, carbon, and oxygen atoms, respectively. FGCD labelled X4

to X7 count the numbers of CH_3, CH_2, and CH groups in the molecule, respectively. Some descriptors were designed to consider effects of chemical function position, e.g., X31 to X34 transcoding the position of the alcohol function, with X31 and X32 standing for primary and secondary alcohols, respectively. The descriptor X60 (MM) denotes the molar mass of neat compounds. Descriptor values for mixtures were calculated as linear combinations of corresponding descriptors for individual compounds weighted by associated volumetric fractions v_i. For instance, for a given descriptor X1, the mixture descriptor $\mathrm{X1}_{mix}$, is defined as follows:

$$\mathrm{X1}_{mix} = \sum_{i=1}^{N} v_i \,.\, \mathrm{X1}_i \qquad (1)$$

where *N* is the number of mixture components. Equation (1) was already used to compute mixtures descriptors in the modeling of properties such as flash points,[49] or even surfactants' properties.[53]

Table 1. SMARTS codes defined for molecular descriptors (FGCD) selected for the development of models.

Label	SMARTS[a]	Label	SMARTS[a]
X1	[#1]	X31	[CX4H2][OH]
X2	[#6]	X32	[#6][#6]([OH])[CX4H3]
X3	[#8]	X33	[#6][#6][#6]([OH])[#6][CX4H3]
X4	[CX4H3]	X34	[#6][#6][#6][#6]([OH])[#6][#6][CX4H3]
X5	[CX4H2]	X35	[OX2H1][cX3]:[c]
X6	[CX4H1]	X36	[CX3H0](=[O])[OX2H1]
X7	[CX4H0]	X37	[CX3H0](=[O])[OX2H0]
X8	[CX3H2]	X38	[CX4H3][CX3H0](=[O])[OX2H0]
X9	[CX3H1]	X39	[CX4H2][CX3H0](=[O])[OX2H0]
X10	[CX3H0]	X40	[CX4H1][CX3H0](=[O])[OX2H0]
X11	[CX2H0]	X41	[CX3H0](=[O])[OX2H0][CX4H3]
X12	[CX4H2R]	X42	[CX3H0](=[O])[OX2H0][CX4H2]
X13	[CX4H1R]	X43	[CX3H0](=[O])[OX2H0][CX4H1]
X14	[CX4H0R]	X44	[CX3H0](=[O])
X15	[CX3H1R]	X45	[CX3H1](=[O])
X16	[CX3H0R]	X46	[#6][#6](=[O])[CX4H3]
X17	[cX3H1](:*):*	X47	[#6][#6][#6](=[O])[#6][CX4H3]

X18	[cX3H0](:*)(:*)*	X48	[#6][#6][#6][#6](=[O])[#6][#6][CX4H3]
X19	[c][CX4H3]	X49	[CX3H0R](=[O])
X20	[c][CX4H2]	X50	[OX2H0]
X21	[c][CX4H1]	X51	[#6][OX2H0][CX4H3]
X22	[CX4H3][CX4H1]	X52	[#6][OX2H0][CX4H2]
X23	[CX4H3][CX4H0]	X53	[#6][OX2H0][CX4H1]
X24	[CX4H3][#6][CX4H1]	X54	[#6][OX2H0][CX4H0]
X25	[CX4H3][#6][CX4H0]	X55	[OX2H0R]
X26	[CX4H3][#6][#6][CX4H1]	X56	[oX2H0](:*):*
X27	[CX4H3][#6][#6][CX4H0]	X57	[O;R]
X28	[CX4H3][#6][#6][#6][CX4H1]	X58	[C;R]
X29	[CX4H3][#6][#6][#6][CX4H0]	X59	[c;R]
X30	[OX2H1]	X60	MM

[a] With the exception of X60 described in the text, interpreters can be used to visualize SMARTS, e.g., https://smarts.plus/

The multidimensional space formed by descriptors was used to define the applicability domain of developed QSPR models. Among existing approaches,[54] the chemical information contained implicitly within our databases was pre-processed by applying a principal component analysis (PCA) on X1 to X60 descriptor values taken by both pure compounds and mixtures. The space formed by PC1 and PC2 – the first two principal components resulting from the PCA – was used as an approximated graphical representation of the chemical space of our databases. Figure 3 presents projections of compounds within the chemical space, and it reveals that the domain occupied by mixtures is encompassed in the space of pure compounds. Additionally, empty circles symbolized compounds on the edges of the populated area – with possible quite different structures as compared to others – which can pose certain problems during the learning process as discussed below.

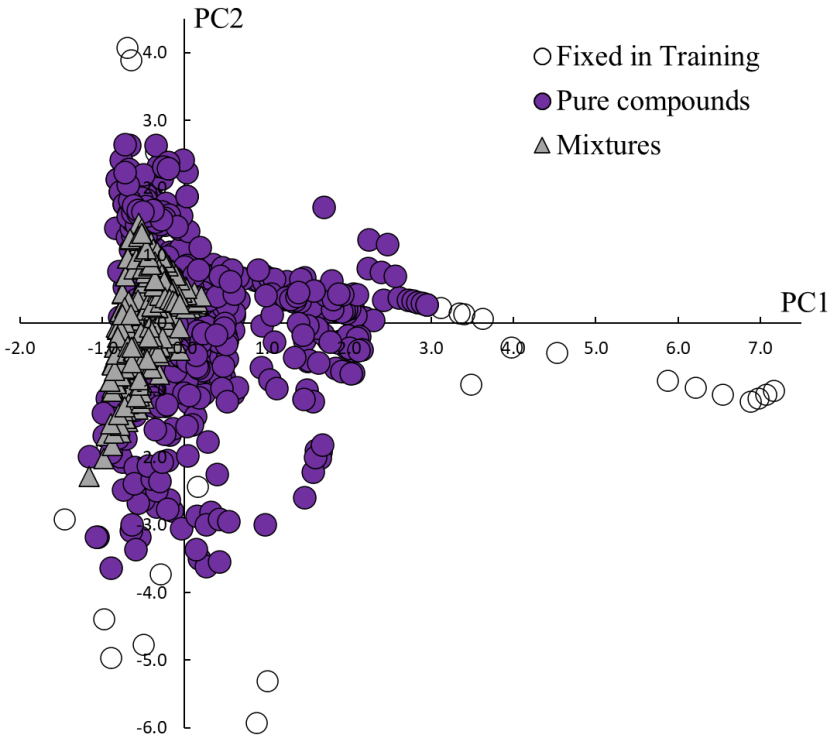

Figure 3. Projections of compounds into the approximated chemical space formed by PC1 and PC2, the two first principal components resulting from the PCA applied to descriptor values. Pure compounds are represented with black circles and mixtures with grey triangles. Empty circles stand for compounds fixed in the Training set during the learning procedure. The percentages of variance explained by PC1 and PC2 are 15% and 9%, respectively.

Since our first developments of QSPR based models for fuels property specifications[55,56] to very recent works on per- and polyfluoroalkyl substances,[45] several machine learning algorithms have been tested in conjunction with different molecular structure encodings. It was widely demonstrated that the combination of Support Vector Machines (SVM) and FGCD provides accurate solutions in terms of property modelling.[23,49,50] The LibSVM library[57] was used for Support Vector Regression (SVR) with both linear and radial basis function kernels, and with an epsilon insensitive zone. This led us to optimize three SVR hyperparameters: cost, epsilon, and

gamma, and the approach previously proposed by Gantzer et al. was used within a n-fold cross-validation (n-CV) procedure.[58] In n-CV, the data set is randomly divided in approximately equal *n* portions. An aggregate of *n-1* portions forms a Training set – subsequently used to train models – and the remaining portion constitutes an external set labelled as Test set used to assess model performances. As a result, no data points belonging to the external sets were used to derive models. The procedure is repeated *n* times, choosing at each iteration a new portion of data as an external set, leading to *n* different models. To avoid any strong violation of the applicability domain of the *n* models during the cross-validation procedure, we fixed 39 of the compounds located on the edges of the populated area (as illustrated in Figure 3) in a specific fold (labelled fold-00) which will always be used to form Training sets. For instance, the series of empty circles located on the right part of the diagram denote triglycerides, and the three in the upper left part of the diagram represent furans. It should be noted that other structural outlier compounds were included in this set, for instance, for example hept-1-yne which is the unique alkyne in the pure compounds database. In this work, a 5-CV was applied and details about fold assignment for each compound or mixture are provided in the supporting information. Finally, a SVR model was developed using the set of optimized hyperparameters (i.e., cost, epsilon, and gamma) and considering data points belonging to all folds.[58]

Obtained SVR models were evaluated on to their capability to reproduce CN values for hydrocarbons and oxygenated compounds and their blends. Predicted values were compared to reference experimental data, and performances of models are evaluated by means of metrics such as Mean Absolute Error (MAE, equation (2)), Root Mean Squared Error (RMSE, equation (3)), or the coefficient of determination (R^2, equation (4)), defined respectively as:

$$MAE = \frac{1}{N}\sum_{1}^{N}|y_i - x_i| \quad (2)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{1}^{N}(y_i - x_i)^2} \quad (3)$$

$$R^2 = 1 - {\sum_{1}^{N}(y_i - x_i)^2}\Big/{\sum_{1}^{N}(x_i - \bar{x})^2} \quad (4)$$

with y_i the predicted value, x_i the experimental value, $\bar{x}$ the average of experimental CMC values, and *N* is the number of data points in the considered set. As the endpoint of models is unitless, all considered metrics MAE, RMSE and R^2 are unitless.

3. Results

In this section, the results are organized as follows: first, we report the development of ML-based models to predict CN for hydrocarbons and oxygenated compounds, ones are SVR trained on CN for pure compounds, and the others are SVR trained on CN for pure compounds and mixtures. Then, the prediction of CN for mixtures is addressed, and the development and use of mixing rules is investigated. Finally, as an external validation, the proposed numerical approaches are applied to predict the CN for real fuels. Very recently, Flora et al. reported the use of a somewhat similar approach to determine CN for hydrocarbon mixtures.[59] The compilation of the databases used as support in our work provides a more comprehensive database and a variety in terms of chemical families such as oxygenated compounds. Moreover, new mixing rules are proposed and validated on a large dataset of mixtures containing well-detailed composition of fuel surrogates.

Pure compounds cetane number. As a preamble, the first objective of this work was to develop a CN prediction model for hydrocarbons and oxygenated compounds. Since the ML-based models proposed by our group in the early 2010s,[15,23] methodologies have evolved and numerous works have been published on the subject, and as mentioned above, new CN data have been made available. The CN prediction for pure compounds was performed by applying our now well-proven methodology for SVR-based model training,[45,48,58] to the data collections described above. Two cases were investigated: in the first, only the dataset including pure compounds is taken into account resulting in type (I) models, while in the second, both pure compounds and mixtures are used to train models resulting in type (II) models. During the learning procedure, it was noted that models always strongly (error greater than 30 points of CN) failed in reproducing some of experimental CN values, and thus, ten compounds were ultimately removed from the training procedure. Four of these latter structures are alkyl octadec-9-enoates, with numbers of carbon atoms in the alkyl chain of 6, 8, 10, and 12, noting that Kim et al. also reported such structures as outliers because of CN trends inconsistencies.[20] Propane also belongs to the ten discarded structures, as well as five compounds holding one or more ether functions.

A 5-CV was applied resulting in a splitting of the database into five folds (each containing about 120 and 235 entries for type (I) and type (II) models, respectively), plus one additional – Fold-00 – containing the 39 compounds fixed in the Training set to avoid any violation of the applicability domain. It should be emphasized that the folds' assignment defined for type (I) models is reused for type (II) models – supplemented with a random distribution of mixtures – which enable fair comparisons. The SVR hyperparameters were optimized considering the six folds, and a final SVR model was developed using optimized cost, epsilon, and gamma. Considering pure compounds, performances of type (I) and type (II) models are roughly similar (with RMSE of 5 points of CN),

thus only performances of type (II) models are discussed in detail. Table 2 presents for type (II) model RMSE and R^2 values calculated on the Training and Test sets for the five ephemeral models generated during the 5-CV. It shows that values of metrics are roughly stable from one decomposition to another with for instance, mean RMSE and R^2 values of 4.3 and 0.968 on Training sets, respectively. Regarding Test sets, mean RMSE and R^2 values are 8.4 and 0.874, respectively. The set of optimized hyperparameters values are 458.032, 2.747, and 1.306 for cost, epsilon, and gamma, respectively.

Table 2. RMSE and R^2 values calculated on the Training and Test sets for the ephemeral SVR based models' type (II) generated during the 5-CV.

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Training					
RMSE	4.3	4.2	4.2	4.4	4.5
R^2	0.967	0.970	0.969	0.965	0.967
Test					
RMSE	7.6	9.5	8.2	8.7	8.2
R^2	0.906	0.829	0.897	0.883	0.856

Figure 4 presents the parity plot of experimental vs. predicted CN values using the type (II) SVR model over the entire database. All data points are not too scattered on both sides of the bisector, indicating that the predicted values are in good agreement with reference data. Figure 5 presents the distribution of errors when predicting CN values using the final SVR models of type (II). It shows that 89% of compounds and mixtures are predicted with an absolute error less than 5 CN points, and it reaches 96% if an interval [0;10[is considered. The largest error is obtained for 1-ethoxybutane for which the model predicts a CN value of 77 while an experimental value of 110.2 was reported by Kim et al.[20] Noting that the application of the final SVR model of type (I) a value of 91 was obtained. Figure 6 presents the evolution of CN values with the number of carbon atoms

in each hydrocarbon chain in the dialkyl ether family, and it proposes a comparison between experimental CN values in the database and trends obtained using the final SVR models of type (I) and type (II). This first highlights the aberrant nature of some experimental values for this family, more particularly the CN value for diethyl ether which does not follow the trend of the other compounds belonging to this family. This justifies our choice to discard the value for diethyl ether from the training process. Then, Figure 6 shows different trends returned by type (I) and type (II) models. Although type (I) predictions are reasonably in agreement with the trend of the experimental data, those of type (II) deviate slightly. It is interesting to note that some ethers are involved in mixtures as shown in Figure 2, more precisely dibutyl ether and dihexyl ether. As mixtures involving oxygenates are poorly represented in the database, the non-linear effects they imply are just as important. Thus, it can be assumed that the type (II) model counterbalances such effects underestimated individual CN values for most of dialkyl ethers.

For most mixtures, Figure 4 shows that predictions fairly agree with reference experimental values. However, the largest deviation is observed for a 4-component mixture containing cyclohexane, 2,2,4,4,6,8,8-heptamethylnonane, n-hexadecane, oct-1-ene in molar proportions of 0.1071:0.4962:0.2973:0.0994, for which the Type (II) model predicts a CN value of 55.6 while an experimental value of 78.5 is reported.[44]

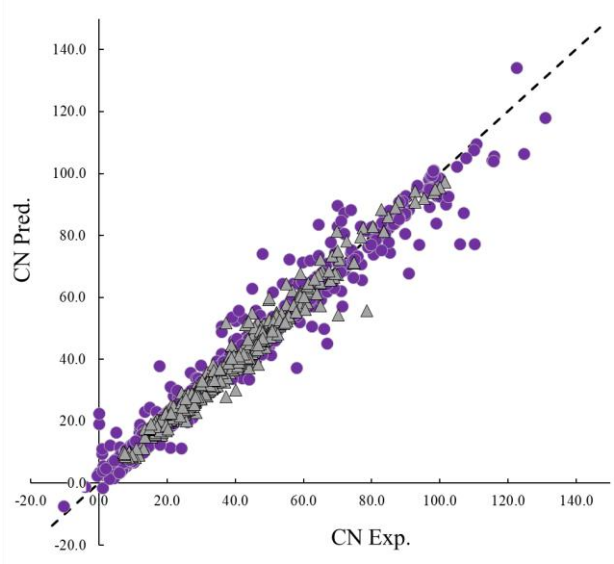

Figure 4. Scatterplots of experimental vs. predicted CN values using the final SVR models of type (II). The pure compounds are represented with purple circles, and the mixtures with grey triangles. The dashed line stands for the bisector of the diagram.

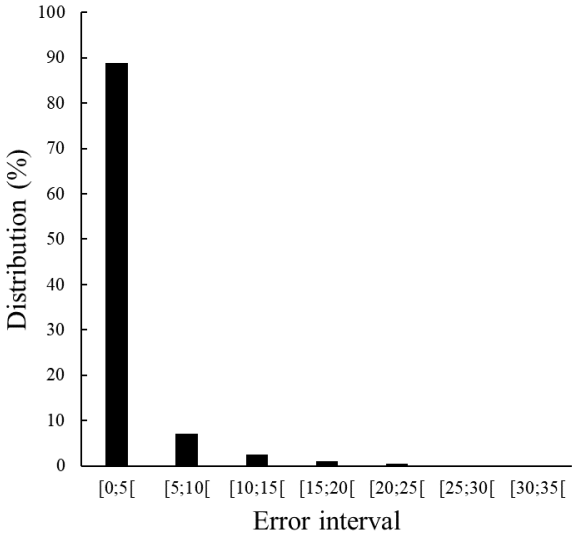

Figure 5. Distribution of errors when predicting CN values using the final SVR models of type (II).

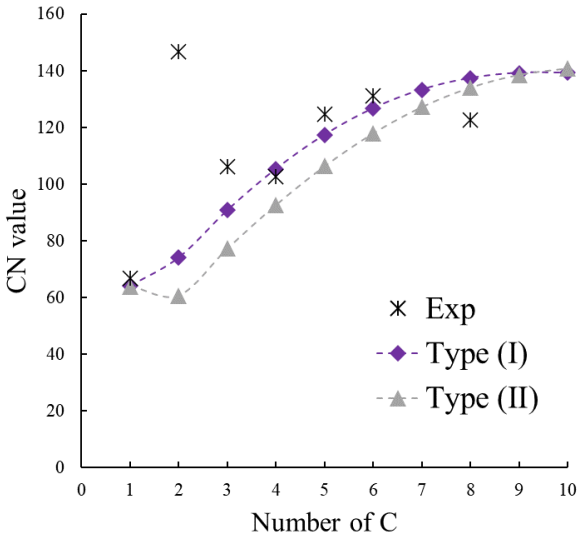

Figure 6. Evolution of CN values with the number of carbon atoms in each branch (R) in the dialkyl ether (R-O-R) family. Predictions obtained using the final SVR models of type (I) and type (II) are compared with reference experimental CN values in the database. The dashed lines for the predicted values are intended to guide the eye.

Mixing rules. The use of mixing rules was then investigated to predict CN of mixtures (CN_{mix}). The model initially proposed by Ghosh and Jaffe to estimate CN_{mix} of lumps – defined as a compositional abstraction of fractions with significant contributions to CN –[27] was used as a starting point, and is summarized as follows:

$$\mathrm{CN}_{mix} = \left. \sum_{i=1}^{N} \upsilon_i \beta_i \mathrm{CN}_i \right/ \sum_{i=1}^{N} \upsilon_i \beta_i \qquad (5)$$

where i runs over the N components in the mixture, and CN_i and υ_i are the cetane number and the volumetric fraction of component i, respectively. β_i is an adjustable parameter to represent a molecule's contribution to the CN of the fuel. As discussed, several variations of equation (5) have been investigated mainly by modifying the parameter β_i expression and/or by replacing υ_i with w_i or x_i the mass and mole fractions, respectively. To account for an individual contribution of each

molecule to the CN value, the parameter β_i was thought as a function of the activity of components in the mixture. From comparisons performed on our collection of experimental CN values for mixtures, the following expression has been shown to minimize errors:

$$\beta_i = {}^{1}/_{exp(\gamma_i)} = exp(-\gamma_i) \quad (6)$$

where γ_i is the activity coefficient of the component *i* in the mixture. The activity coefficients were calculated using the original UNIFAC, a functional groups based activity coefficients model[60,61] together with a recently reoptimized set of parameters.[62] It is noteworthy that the calculation of γ_i requires the knowledge of mole fractions, and that γ_i is a temperature-dependent factor. On this latter point, calculations performed at temperatures typically encountered in CI engines under working conditions (up to 1000 K) did not improve predictions as compared to the use of T = 298 K. Equation (5) was powered by CN_i values both extracted from experimental values collected in the literature and generated using type (I) and type (II) models. Table 3 summarizes results obtained when applying equation (5) on the database containing 572 mixtures and considering two cases, i.e., the parameter β_i as defined in equation (6) or screened with $\beta_i = 1$. From RMSE and MAE values reported in Table 3, it is obvious that considering non-linear interactions via the parameter β_i improves predictions as compared to a simple linear volumetric mixing rule ($\beta_i = 1$). Additionally, discrepancies are observed when equation (5) is either fed with experimental CN_i or predicted using type (I) and type (II) models. With type (I) model values predictions are slightly less accurate than when using experimental CN_i. On the contrary, the use of CN_i generated with the type (II) model leads to a noticeable improvement as compared to the two other sources of CN_i. An in-depth analysis shows discrepancies between the 544 mixtures involving hydrocarbons only and the 28 with at least one oxygenates. For mixtures involving hydrocarbons only, similar RMSE values (4.0 points of CN) are obtained when using equation (6)

or $\beta_i = 1$. In contrast, when considering the 28 mixtures containing at least one oxygenates, RMSE values of 10.9 and 15.3 are obtained using equation (6) or $\beta_i = 1$, respectively. Finally, in equation (5) replacing v_i with the mass or mole fractions led to already known conclusions, i.e., the use of w_i leads to quite similar results than when v_i are considered, and the use of x_i strongly deteriorates predictions (RMSE = 7.1 and MAE = 4.6, calculated on the 572 mixtures, and feeding equation (5) with CN_i predicted using the type (II) model).

Table 3. RMSE and MAE values obtained when applying equation (5) on the database containing 572 mixtures. CN_i values are extracted from experimental values collected in the literature and generated using type (I) and type (II) models.

	Metrics	CN_i from Exp.	CN_i from Type (I)	CN_i from Type (II)
β_i= 1	RMSE	6.3	6.1	5.1
	MAE	3.9	4.2	3.3
$\beta_i = 1/exp(\gamma_i)$	RMSE	5.4	5.5	4.6
	MAE	3.6	3.9	3.2

Figure 7 presents the distribution of errors when predicting CN values for the 572 mixtures using equation (5), equation (6), and CN_i predicted with the SVR model of type (II). It shows that 80% of mixtures are predicted with an absolute error less than 5 CN points, and it reaches 97% if an interval [0;10[is considered. However, the largest deviation (32.9 CN points) is observed for a 4-component mixture containing cyclohexane, 2,2,4,4,6,8,8-heptamethylnonane, n-hexadecane, oct-1-ene in molar proportions of 0.1071:0.4962:0.2973:0.0994. It should be noted that the second largest deviation (28.0 CN points) is observed for the same 4-component mixture in molar proportions 0.1063:0.3971:0.1990:0.2976, for which the model predicts a CN value of 42.2 while an experimental value of 70.2 is reported.[44] Interestingly, the CNs of these two mixtures are also poorly reproduced using the type (II) SVR model alone. In the range [20;25[is a 2-component

mixture, dihexyl ether and hexan-1-ol, in two molar proportions: 0.48:0.52 and 0.67:0.33, for which Witkowski et al. reported experimental CN values of 64.3 and 82.9, respectively.[43] Our full-predictive approach (equation (5) powered by type (II) model predictions) leads to CN values 86.5 and 104.8, and using individual experimental CN in combination with equation (5) leads to CN values of 94.0 and 115.5.

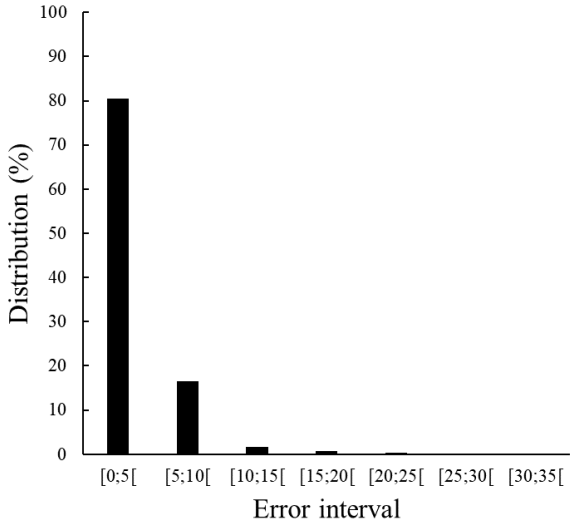

Figure 7. Distribution of errors when predicting CN values for the 572 mixtures using equation (5), equation (6), and CN_i predicted with the SVR model of type (II).

Application to the prediction of CN for real fuels. An external validation of numerical approaches described above was performed on a total of 27 fuel candidates, i.e., 9 blends containing Jet-A1 and catalytic hydrothermolysis jet, 9 blends containing Jet-A1 and hydroprocessed esters and fatty acids (HEFA), and 9 blends containing Jet-A1, aromatics, and HEFA. For all fuels, CN measurements were performed by running each sample in a single-cylinder cooperative fuel research (CFR) engine, as specified in the American Society for Testing and Materials (ASTM) D613. It should be highlighted that this method is time-consuming and

requires a large volume of sample (about 1 L). A fuel is a mixture of thousands of hydrocarbons, each molecule contributing to the CN of the mixture. To apply the predictive approaches mentioned above, it is necessary to characterize and simplify the fuels to surrogates. In previous works,[46,47,63] fuels were analysed with gas chromatography techniques such as the two-dimensional gas chromatography (GCxGC) which is able to provide a detailed characterization of a fuel chemical composition, with only few millilitres of the fluid.[64] The 27 fuels were analysed by means of GCxGC, and compositions were expressed as distributions of mass fractions as a function of the number of carbon atoms for hydrocarbon families such as n-paraffins, i-paraffins, naphthenes, and aromatics. Table 4 reports ranges of numbers of C atoms considered to represent the fuels. A representative molecular structure is attributed to each family/number of carbon atom bin, resulting in a fuel representation involving a maximum of 246 compounds. Each of the 27 fuels was analysed accordingly and was represented by a surrogate containing in-between 77 and 94 components. Compositions and experimental CN values for the 27 surrogates are available as Supporting Information.

Table 4. Ranges of number of carbon atoms to represent the 27 fuel candidates.

Family	Formulae	Number of C atoms
n-paraffins	n-C_nH_{2n+2}	3 to 30
i-paraffins	i-C_nH_{2n+2}	4 to 30
mono-naphthenes	C_nH_{2n}	6 to 30
di-naphthenes	C_nH_{2n-2}	9 to 30
tri-naphthenes	C_nH_{2n-4}	13 to 30
mono-aromatics	C_nH_{2n-6}	6 to 30
naphtheno-mono-aromatics	C_nH_{2n-8}	9 to 30
naphtheno-mono-aromatics	C_nH_{2n-10}	10 to 30
di-aromatics	C_nH_{2n-12}	10 to 30
Naphtheno-di-aromatics	C_nH_{2n-14}	12 to 30
Naphtheno-di-aromatics	C_nH_{2n-16}	13 to 30

The type (I) and type (II) SVR models were used to predict CN for each of the 246 representative molecular structures, and then equation (5) was applied to predict the CN for the 27 fuel surrogates. Figure 8 presents the parity plot of experimental vs. predicted CN values for the 27 fuel surrogates. The combination of equation (5) with type (I) or with type (II) SVR models leads to similar results, and an overall good agreement with experimental CN values. The largest deviations are observed for both combinations for the same surrogate with deviations of 6.1 and 6.7 with type (I) and type (II), respectively. For the 27 surrogates, the obtained RMSE and MAE are 2.7 and 2.4, and 3.0 and 2.6, for equation (5) powered by type (I) and type (II), respectively. Noting that similar RMSE and MAE values are obtained when using equation (6) or $\beta_i = 1$, in line with observations made for the 544 mixtures involving hydrocarbons only. The RMSE values calculated for the 27 surrogates are about 1 CN point lower than those calculated on the subset of 544 mixtures involving only hydrocarbons. This discrepancy could be explained by (i) the large number of hydrocarbons, each with a low fraction in the surrogates, contributing to a possible balance of errors; (ii) or the fact that the data for the 544 mixtures originates from various sources. These values are greatly inferior to the experimental uncertainty on the CN measurement (estimated to 5 points of CN), which demonstrates the relevance of the proposed numerical approach. It is noteworthy that applying the type (II) SVR model alone on the 27 surrogates did not lead to more accurate prediction as compared to the use of equation (5).

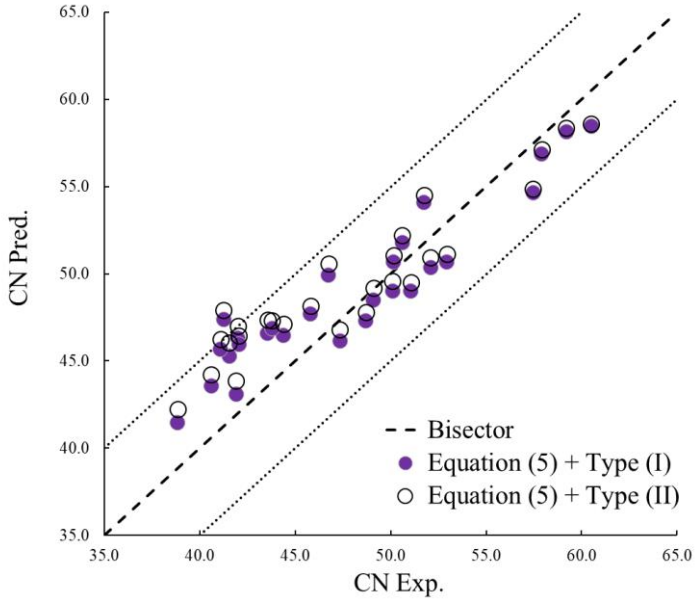

Figure 8. Scatterplots of experimental vs. predicted CN values for the 27 fuel surrogates, using equation (5) powered by either type (I) or type (II) models. The dashed line stands for the bisector of the diagram. The dotted lines denote deviations of 5 CN points to represent the uncertainty on measurements.

4. Conclusions

Numerical approaches were proposed to determine cetane number of hydrocarbons and oxygenated compounds and their mixtures, either in a fully predictive manner or half-predictive manner by powering a new mixing rule with predictive or available experimental data, respectively. The study was divided in three steps: (i) the prediction of pure compounds CN with ML-based approaches, (ii) the development and the application of mixing rules, (iii) the external validation of models on a set of real fuels. For each of these steps, comparisons performed with respect to available experimental data demonstrate the accuracy of the proposed models. When applying our numerical approach to real fuels, values obtained for some metrics are greatly inferior to the experimental uncertainty on the CN measurement (estimated to 5 points of CN), which

demonstrates the relevance of the proposed numerical approach. Furthermore, volumes of products required for the GCxGC analysis make the numerical approach very interesting for testing large numbers of samples as compared to CN measurements performed on CFR engines.

In the context of low carbon fuels, the fuel reactivity through parameters like the CN is of utmost importance to fulfil usage requirements. This is not only relevant for internal combustion engine applications but also for gas turbines. Indeed, any change in the fuel reactivity can affect the engine operation contributing to either reconsider the fuel use or the engine characteristics. The use of additives can also be considered to adapt the fuel reactivity. However, the identification or even the development of additives that contribute to increase/decrease fuels CN remains challenging especially for very low reactivity fuels. The application of numerical approaches as proposed in this work could be considered to explore such effects. To reach this objective, efforts are still required such as supplementing the databases and adapting the methods to the targeted chemicals, for example alkyl nitrates.

ASSOCIATED CONTENT

Supporting Information. Pure compounds and mixtures databases, as well as compositions of the fuel surrogates are available in a spreadsheet file format.

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

References

(1) Förster, F.; Crua, C.; Davy, M.; Ewart, P. Temperature measurements under diesel engine conditions using laser induced grating spectroscopy. *Combustion and Flame* **2019**, *199*, 249–257. DOI: 10.1016/j.combustflame.2018.10.017.

(2) Liang, Y.; Zhou, L.; Xu, M.; Yao, G. Analysis on Combustion Process of a 396 Series Diesel Engine with Flat-top Convex Basin Combustion Chamber. *IOP Conf. Ser.: Mater. Sci. Eng.* **2020**, *782* (4), 42050. DOI: 10.1088/1757-899X/782/4/042050.

(3) Simsek, S. Effects of biodiesel obtained from Canola, sefflower oils and waste oils on the engine performance and exhaust emissions. *Fuel* **2020**, *265*, 117026. DOI: 10.1016/j.fuel.2020.117026.

(4) Zulqarnain; Yusoff, M. H. M.; Ayoub, M.; Hamza Nazir, M.; Zahid, I.; Ameen, M.; Abbas, W.; Shoparwe, N. F.; Abbas, N. Comprehensive Review on Biodiesel Production from Palm Oil Mill Effluent. *ChemBioEng Reviews* **2021**, *8* (5), 439–462. DOI: 10.1002/cben.202100007.

(5) Das, A. K.; Sahu, S. K.; Panda, A. K. Current status and prospects of alternate liquid transportation fuels in compression ignition engines: A critical review. *Renewable and Sustainable Energy Reviews* **2022**, *161*, 112358. DOI: 10.1016/j.rser.2022.112358.

(6) Zheng, L.; Boylu, R.; Cronly, J.; Ahmed, I.; Ubogu, E.; Khandelwal, B. Experimental study on the impact of alternative jet fuel properties and derived cetane number on lean blowout limit. *Aeronaut. j.* **2022**, *126* (1306), 1997–2016. DOI: 10.1017/aer.2022.33.

(7) Abanteriba, S.; Yildirim, U.; Webster, R.; Evans, D.; Rawson, P. Derived Cetane Number, Distillation and Ignition Delay Properties of Diesel and Jet Fuels Containing Blended Synthetic Paraffinic Mixtures. *SAE Int. J. Fuels Lubr.* **2016**, *9* (3), 703–711. DOI: 10.4271/2016-01-9076.

(8) Busch, S. Diesel-like Fuels, Combustion, and Emissions, SAND2021-14170R, Sandia National Laboratories, **2021.** DOI: 10.2172/1832084.

(9) Benajes, J.; García, A.; Monsalve-Serrano, J.; Guzmán-Mendoza, M. A review on low carbon fuels for road vehicles: The good, the bad and the energy potential for the transport sector. *Fuel* **2024**, *361*, 130647. DOI: 10.1016/j.fuel.2023.130647.

(10) Mohammed, A. S.; Atnaw, S. M.; Ramaya, A. V.; Alemayehu, G. A comprehensive review on the effect of ethers, antioxidants, and cetane improver additives on biodiesel-diesel blend in CI engine performance and emission characteristics. *Journal of the Energy Institute* **2023**, *108*, 101227. DOI: 10.1016/j.joei.2023.101227.

(11) Zhang, Y.; Gao, S.; Zhang, Z.; Li, W.; Yuan, T.; Tan, D.; Duan, L.; Yang, G. A comprehensive review on combustion, performance and emission aspects of higher alcohols and its additive effect on the diesel engine. *Fuel* **2023**, *335*, 127011. DOI: 10.1016/j.fuel.2022.127011.

(12) Spudić, R.; Somek, K.; Kovačević, V. Single Fuel Concept for Croatian Army Ground Vehicles. *Promet- Traffic & Transportation* **2008**, *20* (3), 181–187. DOI: 10.7307/ptt.v21i3.1000.

(13) Yang, H.; Fairbridge, C.; Ring, Z. Neural network prediction of cetane numbers for isoparaffins and diesel fuel. *Petroleum Science and Technology* **2001**, *19* (5-6), 573–586. DOI: 10.1081/LFT-100105275.

(14) Smolenskii, E. A.; Bavykin, V. M.; Ryzhov, A. N.; Slovokhotova, O. L.; Chuvaeva, I. V.; Lapidus, A. L. Cetane numbers of hydrocarbons: calculations using optimal topological indices. *Russ Chem Bull* **2008**, *57* (3), 461–467. DOI: 10.1007/s11172-008-0073-0.

(15) Creton, B.; Dartiguelongue, C.; Bruin, T. de; Toulhoat, H. Prediction of the Cetane Number of Diesel Compounds Using the Quantitative Structure Property Relationship. *Energy Fuels* **2010**, *24* (10), 5396–5403. DOI: 10.1021/ef1008456.
(16) Freitas, R. S.; Jiang, X. Descriptors-based machine-learning prediction of cetane number using quantitative structure–property relationship. *Energy and AI* **2024**, *17*, 100385. DOI: 10.1016/j.egyai.2024.100385.
(17) 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), 5781–5801. DOI: 10.1021/acs.energyfuels.5b01032.
(18) 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. DOI: 10.1016/j.fuel.2019.03.092.
(19) Guo, Z.; Lim, K. H.; Chen, M.; Thio, B. J. R.; Loo, B. L. W. Predicting cetane numbers of hydrocarbons and oxygenates from highly accessible descriptors by using artificial neural networks. *Fuel* **2017**, *207*, 344–351. DOI: 10.1016/j.fuel.2017.06.104.
(20) Kim, Y.; Cho, J.; Naser, N.; Kumar, S.; Jeong, K.; McCormick, R. L.; St. John, P. C.; Kim, S. Physics-informed graph neural networks for predicting cetane number with systematic data quality analysis. *Proceedings of the Combustion Institute* **2023**, *39* (4), 4969–4978. DOI: 10.1016/j.proci.2022.09.059.
(21) 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* (42), 12236–12245. DOI: 10.1021/acs.iecr.7b02753.
(22) 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. DOI: 10.1016/j.fuel.2020.118589.
(23) Saldana, D. A.; Starck, L.; Mougin, P.; Rousseau, B.; Pidol, L.; Jeuland, N.; Creton, B. Flash Point and Cetane Number Predictions for Fuel Compounds Using Quantitative Structure Property Relationship (QSPR) Methods. *Energy Fuels* **2011**, *25* (9), 3900–3908. DOI: 10.1021/ef200795j.
(24) Chen, Y.; Zheng, Z.; Lu, Z.; Wang, H.; Wang, C.; Sun, X.; Xu, L.; Yao, M. Machine learning-based screening of fuel properties for SI and CI engines using a hybrid group extraction method. *Applied Energy* **2024**, *366*, 123257. DOI: 10.1016/j.apenergy.2024.123257.
(25) Naik, C. V.; Puduppakkam, K.; Wang, C.; Kottalam, J.; Liang, L.; Hodgson, D.; Meeks, E. Applying Detailed Kinetics to Realistic Engine Simulation: the Surrogate Blend Optimizer and Mechanism Reduction Strategies. *SAE Int. J. Engines* **2010**, *3* (1), 241–259. DOI: 10.4271/2010-01-0541.
(26) Mueller, C. J.; Cannella, W. J.; Bruno, T. J.; Bunting, B.; Dettman, H. D.; Franz, J. A.; Huber, M. L.; Natarajan, M.; Pitz, W. J.; Ratcliff, M. A.; Wright, K. Methodology for Formulating Diesel Surrogate Fuels with Accurate Compositional, Ignition-Quality, and Volatility Characteristics. *Energy Fuels* **2012**, *26* (6), 3284–3303. DOI: 10.1021/ef300303e.
(27) Ghosh, P.; Jaffe, S. B. Detailed Composition-Based Model for Predicting the Cetane Number of Diesel Fuels. *Ind. Eng. Chem. Res.* **2006**, *45* (1), 346–351. DOI: 10.1021/ie0508132.

(28) Murphy, M. J.; Taylor, J. D.; McCormick, R. L. *Compendium of Experimental Cetane Number Data*, NREL/SR-540-36805, National Renewable Energy Laboratory (NREL), **2004**. https://digital.library.unt.edu/ark:/67531/metadc835154/.
(29) J. Yanowitz; M.A. Ratcliff; R.L. McCormick; J.D. Taylor; M.J. Murphy. *Compendium of Experimental Cetane Numbers*, NREL/TP-5400-67585, National Renewable Energy Laboratory (NREL), **2017**. https://www.nrel.gov/docs/fy17osti/67585.pdf.
(30) Luo, J.; Yao, M.; Liu, H.; Yang, B. Experimental and numerical study on suitable diesel fuel surrogates in low temperature combustion conditions. *Fuel* **2012**, *97*, 621–629. DOI: 10.1016/j.fuel.2012.02.057.
(31) Dooley, S.; Won, S. H.; Heyne, J.; Farouk, T. I.; Ju, Y.; Dryer, F. L.; Kumar, K.; Hui, X.; Sung, C.-J.; Wang, H.; Oehlschlaeger, M. A.; Iyer, V.; Iyer, S.; Litzinger, T. A.; Santoro, R. J.; Malewicki, T.; Brezinsky, K. The experimental evaluation of a methodology for surrogate fuel formulation to emulate gas phase combustion kinetic phenomena. *Combustion and Flame* **2012**, *159* (4), 1444–1466. DOI: 10.1016/j.combustflame.2011.11.002.
(32) Xiao, G.; Zhang, Y.; Lang, J. Kinetic Modeling Study of the Ignition Process of Homogeneous Charge Compression Ignition Engine Fueled with Three-Component Diesel Surrogate. *Ind. Eng. Chem. Res.* **2013**, *52* (10), 3732–3741. DOI: 10.1021/ie303406k.
(33) Dooley, S.; Heyne, J.; Won, S. H.; Dievart, P.; Ju, Y.; Dryer, F. L. Importance of a Cycloalkane Functionality in the Oxidation of a Real Fuel. *Energy Fuels* **2014**, *28* (12), 7649–7661. DOI: 10.1021/ef5008962.
(34) Dryer, F. L.; Jahangirian, S.; Dooley, S.; Won, S. H.; Heyne, J.; Iyer, V. R.; Litzinger, T. A.; Santoro, R. J. Emulating the Combustion Behavior of Real Jet Aviation Fuels by Surrogate Mixtures of Hydrocarbon Fluid Blends: Implications for Science and Engineering. *Energy Fuels* **2014**, *28* (5), 3474–3485. DOI: 10.1021/ef500284x.
(35) Kim, D.; Martz, J.; Violi, A. A surrogate for emulating the physical and chemical properties of conventional jet fuel. *Combustion and Flame* **2014**, *161* (6), 1489–1498. DOI: 10.1016/j.combustflame.2013.12.015.
(36) Liu, X.; Wang, H.; Wang, X.; Zheng, Z.; Yao, M. Experimental and modelling investigations of the diesel surrogate fuels in direct injection compression ignition combustion. *Applied Energy* **2017**, *189*, 187–200. DOI: 10.1016/j.apenergy.2016.12.054.
(37) Kim, D.; Martz, J.; Abdul-Nour, A.; Yu, X.; Jansons, M.; Violi, A. A six-component surrogate for emulating the physical and chemical characteristics of conventional and alternative jet fuels and their blends. *Combustion and Flame* **2017**, *179*, 86–94. DOI: 10.1016/j.combustflame.2017.01.025.
(38) Luning Prak, D. J.; Romanczyk, M.; Wehde, K. E.; Ye, S.; McLaughlin, M.; Luning Prak, P. J.; Foley, M. P.; Kenttämaa, H. I.; Trulove, P. C.; Kilaz, G.; Xu, L.; Cowart, J. S. Analysis of Catalytic Hydrothermal Conversion Jet Fuel and Surrogate Mixture Formulation: Components, Properties, and Combustion. *Energy Fuels* **2017**, *31* (12), 13802–13814. DOI: 10.1021/acs.energyfuels.7b02960.
(39) Abdul Jameel, A. G.; Naser, N.; Emwas, A.-H.; Dooley, S.; Sarathy, S. M. Predicting Fuel Ignition Quality Using 1 H NMR Spectroscopy and Multiple Linear Regression. *Energy Fuels* **2016**, *30* (11), 9819–9835. DOI: 10.1021/acs.energyfuels.6b01690.
(40) Mueller, C. J.; Cannella, W. J.; Bays, J. T.; Bruno, T. J.; DeFabio, K.; Dettman, H. D.; Gieleciak, R. M.; Huber, M. L.; Kweon, C.-B.; McConnell, S. S.; Pitz, W. J.; Ratcliff, M. A. Diesel Surrogate Fuels for Engine Testing and Chemical-Kinetic Modeling: Compositions and Properties. *Energy Fuels* **2016**, *30* (2), 1445–1461. DOI: 10.1021/acs.energyfuels.5b02879.

(41) Al Ibrahim, E.; Farooq, A. Prediction of the Derived Cetane Number and Carbon/Hydrogen Ratio from Infrared Spectroscopic Data. *Energy Fuels* **2021**, *35* (9), 8141–8152. DOI: 10.1021/acs.energyfuels.0c03899.
(42) D. F. Chuahy, F.; Sluder, C. S.; Curran, S. J.; Kukkadapu, G.; Wagnon, S. W.; Whitesides, R. Numerical assessment of fuel physical properties on high-dilution diesel advanced compression ignition combustion. *Applications in Energy and Combustion Science* **2023**, *13*, 100102. DOI: 10.1016/j.jaecs.2022.100102.
(43) Witkowski, D.; Groendyk, M.; Rothamer, D. A. Kinetic model-based group contribution method for derived cetane number prediction of oxygenated fuel components and blends. *Combustion and Flame* **2023**, *255*, 112883. DOI: 10.1016/j.combustflame.2023.112883.
(44) Sheyyab, M.; Lynch, P. T.; Mayhew, E. K.; Brezinsky, K. Optimized synthetic data and semi-supervised learning for Derived Cetane Number prediction. *Combustion and Flame* **2024**, *259*, 113184. DOI: 10.1016/j.combustflame.2023.113184.
(45) Creton, B.; Barraud, E.; Nieto-Draghi, C. Prediction of critical micelle concentration for per- and polyfluoroalkyl substances. *SAR and QSAR in environmental research* **2024**, 1–16. DOI: 10.1080/1062936X.2024.2337011.
(46) Creton, B.; Veyrat, B.; Klopffer, M.-H. Fuel sorption into polymers: Experimental and machine learning studies. *Fluid Phase Equilibria* **2022**, *556*, 113403. DOI: 10.1016/j.fluid.2022.113403.
(47) Hall, C.; Creton, B.; Rauch, B.; Bauder, U.; Aigner, M. Probabilistic Mean Quantitative Structure–Property Relationship Modeling of Jet Fuel Properties. *Energy Fuels* **2022**, *36* (1), 463–479. DOI: 10.1021/acs.energyfuels.1c03334.
(48) Moreno Jimenez, R.; Creton, B.; Marre, S. Machine learning-based models for accessing thermal conductivity of liquids at different temperature conditions. *SAR and QSAR in environmental research* **2023**, *34* (8), 605–617. DOI: 10.1080/1062936X.2023.2244410.
(49) Saldana, D. A.; Starck, L.; Mougin, P.; Rousseau, B.; Creton, B. On the rational formulation of alternative fuels: melting point and net heat of combustion predictions for fuel compounds using machine learning methods. *SAR and QSAR in environmental research* **2013**, *24* (4), 259–277. DOI: 10.1080/1062936X.2013.766634.
(50) Saldana, D. A.; Starck, L.; Mougin, P.; Rousseau, B.; Ferrando, N.; Creton, B. Prediction of Density and Viscosity of Biofuel Compounds Using Machine Learning Methods. *Energy Fuels* **2012**, *26* (4), 2416–2426. DOI: 10.1021/ef3001339.
(51) *RDKit: Open-Source Cheminformatics Software*. http://www.rdkit.org/ (accessed 2024).
(52) *SMARTS: a language for describing molecular patterns; daylight chemical information systems inc.: Laguna niguel, ca*. https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (accessed 2023).
(53) Muller, C.; Maldonado, A. G.; Varnek, A.; Creton, B. Prediction of Optimal Salinities for Surfactant Formulations Using a Quantitative Structure–Property Relationships Approach. *Energy Fuels* **2015**, *29* (7), 4281–4288. DOI: 10.1021/acs.energyfuels.5b00825.
(54) Roy, K.; Kar, S.; Ambure, P. On a simple approach for determining applicability domain of QSAR models. *Chemometrics and Intelligent Laboratory Systems* **2015**, *145*, 22–29. DOI: 10.1016/j.chemolab.2015.04.013.
(55) Creton, B. Chemoinformatics at IFP Energies Nouvelles: Applications in the Fields of Energy, Transport, and Environment. *Molecular informatics* **2017**, *36* (10). DOI: 10.1002/minf.201700028.

(56) Saldana, D. A.; Creton, B.; Mougin, P.; Jeuland, N.; Rousseau, B.; Starck, L. Rational Formulation of Alternative Fuels using QSPR Methods: Application to Jet Fuels. *Oil Gas Sci. Technol. – Rev. IFP Energies nouvelles* **2013**, *68* (4), 651–662. DOI: 10.2516/ogst/2012034.
(57) Chang, C.-C.; Lin, C.-J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2* (3), 1–27. DOI: 10.1145/1961189.1961199.
(58) Gantzer, P.; Creton, B.; Nieto-Draghi, C. Comparisons of Molecular Structure Generation Methods Based on Fragment Assemblies and Genetic Graphs. *JCIM* **2021**, *61* (9), 4245–4258. DOI: 10.1021/acs.jcim.1c00803.
(59) Flora, G.; Karimzadeh, F.; Kahandawala, M. S.; DeWitt, M. J.; Corporan, E. Prediction of hydrocarbons ignition performances using machine learning modeling. *Fuel* **2024**, *368*, 131619. DOI: 10.1016/j.fuel.2024.131619.
(60) Fredenslund, A.; Jones, R. L.; Prausnitz, J. M. Group-contribution estimation of activity coefficients in nonideal liquid mixtures. *AIChE Journal* **1975**, *21* (6), 1086–1099. DOI: 10.1002/aic.690210607.
(61) Vidal, J. *Thermodynamics*:*Applications in chemical engineering and the petroleum industry*; Institut français du pétrole publications; Editions Technip, 2003.
(62) *The 20th joint UNIFAC consortium sponsor and DDB user meeting*. https:// unifac.ddbst.com/unifac_.html (accessed 2023).
(63) Villanueva, N.; Flaconnèche, B.; Creton, B. Prediction of Alternative Gasoline Sorption in a Semicrystalline Poly(ethylene). *ACS combinatorial science* **2015**, *17* (10), 631–640. DOI: 10.1021/acscombsci.5b00094.
(64) Vendeuvre, C.; Ruiz-Guerrero, R.; Bertoncini, F.; Duval, L.; Thiébaut, D. Comprehensive Two-Dimensional Gas Chromatography for Detailed Characterisation of Petroleum Products. *Oil Gas Sci. Technol. – Rev. IFP Energies nouvelles* **2007**, *62* (1), 43–55. DOI: 10.2516/ogst:2007004.