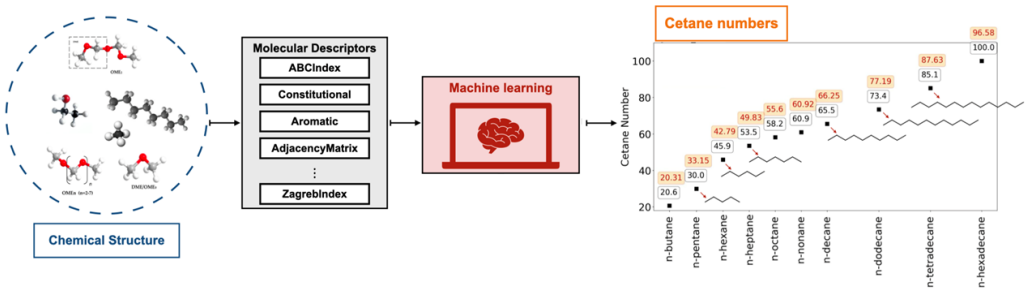
Graphical Abstract

Descriptors-based machine-learning prediction of cetane number using quantitative structure-property relationship

Rodolfo S. M. Freitas, Xi Jiang



Highlights

Descriptors-based machine-learning prediction of cetane number using quantitative structure-property relationship

Rodolfo S. M. Freitas, Xi Jiang

- Machine learning models were employed to predict the cetane number of hydrocarbon fuels from the chemical structures.
- A feature selection approach is used to select the most relevant molecular descriptors that describe the chemical structure of the fuel.
- Impact of the descriptors on cetane number predictions is evaluated using the SHAP method
- The models showed appreciable predictability and reproducibility, returning a mean absolute error of 5.07.

Descriptors-based machine-learning prediction of cetane number using quantitative structure-property relationship

Rodolfo S. M. Freitas[a], Xi Jiang[a]

[a]*School of Engineering and Materials Science, Queen Mary University of London, Mile End Road, London, E1 4NS, UK*

Abstract

The physicochemical properties of liquid alternative fuels are important but difficult to measure/predict, especially when complex surrogate fuels are concerned. In the present work, machine learning is used to develop quantitative structure-property relationship models. The fuel chemical structure is represented by molecular descriptors, allowing the linking of important features of the fuel composition and key properties of fuel utilization. Feature selection is employed to select the most relevant features that describe the chemical structure of the fuel and several machine learning algorithms are tested to construct interpretable models. The effectiveness of the methodology is demonstrated through the development of accurate and interpretable predictive models for cetane numbers, with a focus on understanding the link between molecular structure and fuel properties. In this context, matrix-based descriptors and descriptors related to the number of atoms in the molecule are directly linked with the cetane number of hydrocarbons. Furthermore, the results showed that molecular connectivity indices play a role in the cetane number for aromatic molecules. Also, the methodology is extended to predict the cetane number of ester and ether molecules, leveraging the design of alternative fuels toward fully sustainable fuel utilization.

Keywords: Chemical descriptors, Quantitative Structure-Property relationship, Machine Learning, Cetane number, Fuel design

1. Introduction

Low-carbon alternative fuels are becoming increasingly important, but fossil fuels still play a key role in energy supply, especially in difficult-to-decarbonize transport applications such as shipping, road freight, and aviation transport. Overall, they are responsible for emitting more than 50% CO_2 of the entire transport sector [1]. With the need to take a step towards net zero emissions and sustainable energy utilization, renewable fuels and biofuels derived from sources other than petroleum are becoming increasingly important [2]. The design of alternative fuels is often based on the life cycle assessment methodology, which is an all-encompassing evaluation method employed to estimate fuel viability and benefits, offering insights into their environmental impact, energy efficiency, economic feasibility, and sustainable decision-making through a holistic evaluation [3].

In light of the fuel design, one of the essential "fuel criteria" is that an alternative fuel must be compatible with the existing global infrastructure [4]. So, it can be integrated into the current transportation system using existing infrastructure and be burned in existing engines (such as diesel engines for optimal fuel economy) with minor adjustments as drop-in fuels. Among efforts on developing low-emission fuels for diesel engine combustion, liquid synthetic fuels have shown high potential for low-carbon transport applications [5]. Synthetic fuels are carbon-based liquid fuels chemically synthesized from a carbon source designed to mimic the physicochemical properties of fossil fuels [2, 4].

Synthetic fuels can be manufactured via chemical conversion processes from 'defossilised' carbon dioxide sources, such as point source capture from the exhausts of industrial processes and direct capture from air or biological sources. Bio-based fuels using carbon from biological sources are renewable and have been in use. Liquid synthetic fuels could be a direct alternative to liquid fossil fuels [2]. Also, gas-to-liquid fuels are synthetic fuels produced from natural gas and other hydrocarbon gases as feedstocks [6]. One example of synthetic fuels is oxymethylene ethers (OME_x). As a clean alternative fuel to compression-ignition diesel engines, OME_x is a class of dimethyl ether (DME) derivatives that can be produced from a range of waste feedstocks and biomass, thereby avoiding new fossil carbon from entering the supply chain [7, 8]. Also, a range of methodologies and processes have been used for the synthesis of OME_x [9]. In addition, OME_x can be produced from renewable energy stored via catalytic conversion of hydrogen (H_2) and carbon dioxide (CO_2), as an electrofuel (or e-fuel), thereby used as a sustainable energy carrier [4]. It has been demonstrated that OME_x can substantially reduce harmful emissions of nitrogen oxides (NO_x) and particulates (soot) avoiding the NO_x-soot trade-off [10, 11]. However, OME_x is currently used as a blending component instead of a complete replacement for diesel due to differences in their physicochemical properties [11].

For the rapid integration of liquid alternative fuels into current infrastructures for storage, transport, and direct injection in combustion engines, the physicochemical properties associated with fuel composition must be known.

This represents a significant challenge since alternative fuels are often composed of complex mixtures and the physicochemical properties depend on fuel composition variability linked with production source and process [2]. To address this challenge, accurate information on the physicochemical properties of complex mixtures over the engine operational ranges is mandatory to adapt the system operation to alternative fuels, but this is not readily available.

In terms of fuel utilization, important properties of liquid fuel include cetane number (CN), density, flash point, freezing point, autoignition temperature, energy content, and combustion emissions. Cetane number is a key physicochemical property to measure the ignition quality of diesel-like fuels. The ignition quality is directly linked to the molecular structure of the diesel fuel blends [12]. Although the chemical structures of liquid synthetic fuels are mostly known, a critical aspect faced in designing new diesel-like fuels is understanding the relationship between the molecular structure and fuel properties. Usually, the physicochemical properties are measured from costly experimental facilities [13, 14]. Also, no robust physicochemical models able to predict CN are available [15, 16].

Recently, machine-learning (ML) models have gained attention in the renewable energy sectors [17, 18]. This is the case with fuel design, where ML can be a powerful tool to predict the physicochemical properties of fuels from the chemical structures [19, 20]. Furthermore, an AI-assisted fuel design methodology was developed to predict the properties of pure compounds and

fuel blends [21]. Freitas et al. [22] proposed a methodology to explore the thermodynamic properties of practical fuels by combining molecular dynamics (MD) simulations and ML models. The results show that ML models can yield accurate predictions of fundamental fuel properties from the chemical compositions of the fuels.

Going further, research efforts have been dedicated to developing predictive CN models based on quantitative structure-property relationships [23, 24]. Guan et al. [25] proposed an active subspace method for descriptor selection to build a QSPR predictive model for cetane number. Saldana et al. [26] were pioneers in using ML techniques for CN prediction based on molecular structure. Also, an artificial neural network (ANN) to predict the cetane numbers of hydrocarbons and oxygenates, which are critical for compression-ignition engines was developed [27]. A systematic data quality analysis integrated with graph neural networks (GNNs) for a reliable CN prediction was also developed [28]. Also, Schweidtmann et al. [29] proposed a GNN for predicting fuel ignition quality indicators of hydrocarbons, combining multitask learning, transfer learning, and ensemble learning to address the limitation of experimental data to train the model.

Despite remarkable advances in the construction of QSPR machine learning models, these methods might suffer from a lack of interpretability, which limits the understanding of the underlying physics/chemistry of the predictive models, i.e., the relationship between the molecular structure and physicochemical property is not fully understood. Explainable artificial in-

telligence tools are being developed to leverage the interpretability of ML models so that the effects of the input quantities on the predictions can be better understood [30]. The present work aims to characterize the dependence of the cetane number of liquid synthetic fuels on chemical composition/structure using ML techniques to leverage data obtained through experiments. Moreover, the study provides insights into the interpretability of the ML models, highlighting the importance of feature selection and the correlation between molecular descriptors and physicochemical properties, with implications for the development of predictive models for renewable fuels. In particular, the novelty lies in the use of sophisticated ML techniques to design descriptor-based interpretable quantitative structure-property relationship (IQSPR) models. Such an approach allows an understanding of the underlying physics/chemistry which links the chemical structure with the physicochemical properties. To do so, the chemical structure of such complex fuels is described using molecular descriptors [31]. Descriptors allow us to correlate the chemical structure of the fuels with physicochemical properties. Furthermore, we use a recursive feature elimination (RFE) strategy to select the set of most suitable descriptors that best characterize the molecular structure to construct robust IQSPR predictive models. Moreover, we test different ML algorithms to evaluate their performance (like a model selection) to construct IQSPR predictive cetane number models. Finally, we use the SHapley Additive exPLanations (SHAP) [32] method to measure the impact of descriptors on CN predictions.

It is worth pointing out that the ML-assisted liquid synthetic fuel property modeling methodology depicted here has the potential to construct predictive models able to assist the design of alternative fuels with desired properties that can potentially lead to fully sustainable fuel utilization, i.e., the engines are 100% powered by sustainable/renewable synthetic fuels, which breaks a barrier in alternative fuel utilization as synthetic fuels are currently used as a 'drop-in' component instead of on their own. Furthermore, it is worth highlighting the present approach is used to build predictive models for cetane numbers. Still, such an approach can be extended to other physicochemical properties important for the design of liquid alternative fuels, as it is ubiquitous for data analytics of fuel properties.

The remainder of this paper is organized as follows. The idea of descriptors-based QSPR models is explained in Section 2. In Section 3, the CN dataset used to build the QSPR models is described. Finally, the results are shown in Section 4, and the conclusions are drawn in Section 5.

2. Methodology

2.1. Molecular Descriptors

A molecular descriptor is a mathematical characterization of a chemical structure. The main idea is transforming chemical information embedded within a symbolic representation of a molecule into a set of features useful to represent the chemical composition [33]. In particular, molecular descriptors reflect complex structural patterns, such as the number of atoms in the

molecule, molecular shape, size, as well as atomistic properties [34].

Molecular descriptors can be divided into classes concerning the information content about the chemical structure and the ease of computing them [35]. 1D descriptors are computed from the bulk properties of the molecules, such as the number of atoms, molecular weight, and so on. These descriptors are simple to compute but do not provide any information regarding the molecular structure. 2D molecular descriptors provide topological information based on the 2D structure representation of the molecule, such as the number of rings, the number of hydrogen bond acceptors and donors, matrix-based descriptors, etc. Finally, 3D descriptors reflect information about the spatial coordinates of atoms of the chemical structure. These descriptors may provide useful information about the molecule, such as the intramolecular hydrogen bonding. However, the determination of 3D descriptors might be a time-consuming task due to their complexity [36, 37]. In addition, molecular descriptors based on functional groups in the molecule structure have been proposed to build QSPR models [38, 39]. In particular, these descriptors are calculated by counting the number of functional groups, such as methyl ($-CH_3-$) and methylene ($-CH_2-$), in the chemical structure.

Molecular descriptors can be a powerful tool to convert molecular structures into values that can be used to construct predictive machine learning models [40, 41]. Quantitative structure-property relationship models to predict sooting tendencies were developed using molecular descriptors [42, 43]. Such a description may be used to connect important features of the fuel

composition with key properties of fuel utilization, allowing a step toward the development of IQSPR-ML models. Going further, revealing the dependence of physicochemical properties of liquid synthetic fuels on fuel mixture chemical composition/chemical structure may lead to new information about the physicochemical properties.

In the present work, we compute the molecular descriptors using the open-source descriptor-calculator Mordred software [34]. Mordred software is chosen due to the easiness of implementation, where the descriptors are generated from the simplified molecular input line entry specification (SMILES), and it has a lower computation time compared with other traditional descriptor calculators such as the PaDEL-Descriptor [44]. Mordred can compute more than 1800 molecular descriptors including 2D and 3D descriptors, such as the number of carbon atoms, the number of aromatic bonds, topological indices, matrix-based descriptors, and so on.

2.2. Feature Selection

As the descriptor calculator software can provide thousands of molecular descriptors, selecting the set of most suitable descriptors that best characterize the chemical structure is critical for developing interpretable and accurate predictive models. Moreover, over-parametrization may hamper the predictability of IQSPR-ML models. In this sense, feature selection techniques can be applied to remove irrelevant descriptors and select the most important features regarding the physicochemical properties [45].

In this regard, several strategies for selecting the descriptors that significantly contribute to constructing robust QSPR models have been developed [46, 47, 20]. St. John et al. [43] used principal component analysis (PCA) and RFE to select the best features for the prediction of sooting tendency from molecular structure. Furthermore, a QSPR model for the cetane number of hydrocarbons using the activate subspace method to remove irrelevant descriptors has been proposed [25]. Recently, Comesana et al. [45] developed a systematic methodology for selecting features to construct accurate and interpretable predictive models for several physicochemical properties of organic molecules.

Based on Comesana et al. [45], in the present work, we have developed a method using recursive feature elimination to develop an IQSPR model for the prediction of the cetane number of liquid synthetic fuels guided by the molecular structure. RFE is a supervised feature selection method that recursively removes features at each iteration based on feature importance. The procedure is repeated until a predefined desired number of features is achieved.

The main steps of the feature selection methodology proposed in the present work are given as follows. First, we generate the molecular descriptors from SMILES using the Mordred software for each molecule. Next, we remove non-numerical values, zeros, and descriptors with 95% of matching values across molecules, i.e., descriptors that have similar values for different molecules, since they do not provide reliable information about the molecule

structures. Autocorrelation descriptors may not be linked with the molecular structure and physicochemical properties of the molecule [45, 48], so they are also removed. We compute the correlation between the molecular descriptors and physicochemical properties to be predicted and rank the descriptors from the highest to lowest correlation. The correlation provides prior knowledge of the most relevant descriptors before employing the RFE method and avoids eliminating reliable information. Here, the Spearman coefficient is used [49]. Next, we selected the top-ranked descriptors and removed the descriptors with a Spearman correlation coefficient greater than 0.9 regarding the top-ranked descriptors to avoid redundant information. The remaining descriptors set is used as input for the RFE method. Furthermore, removing collinear features can help the QSPR model to generalize and improve the interpretability of the model [50].

Here, the Scikit-learn [51] ML library is used to implement the RFE method with a model estimator (e.g., a predictive model that assigns weights to features). Further details about the ML estimator models are presented in the next subsection.

2.3. Machine Learning Models

In this section, we present a brief description of ML models for a generic property γ function of the chemical structure/composition and state variables. In particular, the aim is to learn a mapping $\mathbf{f}$ characterizing the macroscopic thermodynamic relation between the physicochemical property

and the molecular structure:

$$\gamma = \mathbf{f}(\boldsymbol{\Phi}, \boldsymbol{\theta}, \boldsymbol{\xi}) \tag{1}$$

Here, $\mathbf{f}$ is a map that acts as an equation of state that may replace costly experimental measurements. $\boldsymbol{\Phi}$ is the vector of molecular descriptors that characterize the chemical structure of fuel. Also, $\boldsymbol{\theta}$ is the vector of state variables, such as temperature and pressure. The vector $\boldsymbol{\xi}$ denotes potential noise from the data and is often considered random.

Despite many works dedicated to constructing QSPR based on ML models [52, 53], there is no clear statement regarding the best ML algorithms to be chosen. The choice of the ML model is a challenge, as it might depend on the physicochemical properties to be learned and the relationship between them with the molecular descriptors. This becomes more critical when it faces a small data regime, where the ML models are prone to overfitting and difficult to generalize [54, 55].

In this context, we test eight different ML algorithms. The models used are the least absolute shrinkage and selection operator (LASSO), support vector regression (SVR), random forest (RF), gradient boosting regression (GBR), and Gaussian process regressor (GP). These models are implemented using the Scikit-learn [51]. We also implemented a light gradient-boosting machine (LGBM) [56], and extreme gradient boosting (XGB) [57]. Moreover, we train a multi-layer perceptron (MLP) using the PyTorch [58]. All ML

models are deterministic except the GP model. GP is a popular probabilistic Bayesian approach to building predictive models mainly due to its capacity to handle both parametric and epistemic uncertainties [59]. The optimization of model hyperparameters is detailed in the next section.

The aim of testing different ML algorithms is to evaluate the performance (like a model selection) of the ML algorithms to construct IQSPR predictive models for physicochemical properties guided by the molecular structure.

3. Dataset

In this paper, the cetane numbers dataset is acquired from the report by Murphy et al. [60]. The dataset consists of experimental CNs of hydrocarbons computed using four different methods, ASTM D613, ASTM D6890, blend method, and an unknown method. Constructing a dataset based on CNs of hydrocarbons is a natural choice since they are typical surrogates for diesel fuels [61], and the design of alternative fuel blends are based on them [2]. Here, we follow the same approach proposed by Guan et al. [25] to build the dataset, where the CNs from the first two methods are prioritized due to measurement repeatability, and the mean values are considered for fuels with multiple measurements. Thus, the final dataset consists of 110 CNs of hydrocarbons divided by n-alkanes, iso-alkanes, cycloalkanes, alkenes, and aromatics (see Table S1 in the *Supplementary Material* (SM) for a complete list of CNs).

To train the descriptor-based IQSPR-ML models we preprocessed the

dataset. Here, we use the non-supervised K-means clustering algorithm to divide the dataset into six subdomains based on CN values. This allows data points of each subdomain in both training and testing sets. The dataset is scaled using the standard scaler, which means removing the mean and dividing by the standard deviation. Next, 75% data points of each subdomain are selected randomly to train the IQSPR-ML models. The remaining 25% are used to test them. Such a strategy allows the learning process to enhance the ability of the IQSPR models to generalize to unseen fuel molecules. Figure 1 shows an overview of training and testing data, including the percentage of the five categories of hydrocarbons used in the present work. In terms of numbers, 80 CNs are used for training, and 30 CNs are used for testing.

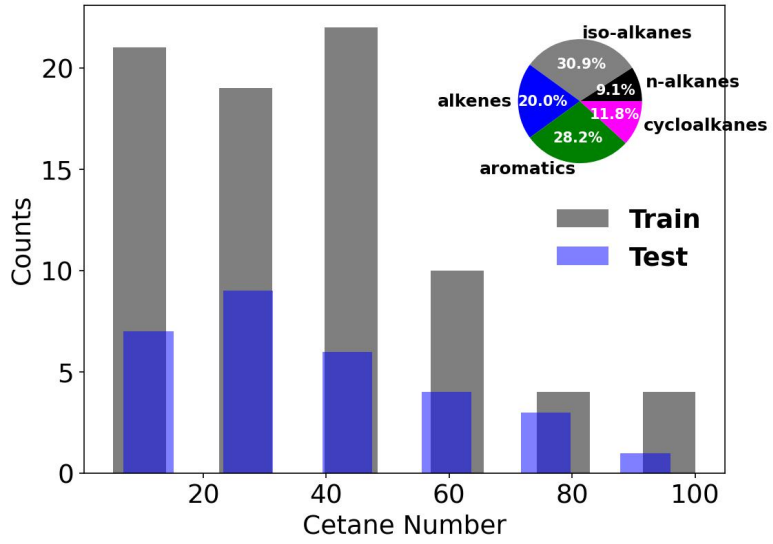


Figure 1: Distribution of the cetane number training and testing datasets.

4. Results and Discussions

In this section, we assess the performance of the ML-assisted liquid synthetic fuel design methodology to construct IQSPR predictive models of CNs for hydrocarbons. Such an assessment unfolds into two main goals: to evaluate the performance of descriptor-based IQSPR-ML models in predicting CNs and to investigate which molecular descriptors are most important to predicting CNs and the relationship between the molecular structure with the physicochemical property. Also, we describe how the methodology can be extended to the design of alternative fuels. The code and data accompanying this manuscript are made publicly available at `https://github.com/RodolfosmFreitas/AI-SyntheticFuel`.

4.1. Performance of descriptor-based IQSPR-ML models

To evaluate the performance of descriptor-based QSPR-ML models we compute the coefficient of determination (R^2-score), the mean absolute error (MAE), and the mean absolute percentage error (MAPE). These are commonly used metrics to measure the performance and accuracy of ML models. The R^2-score is given by:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\gamma_i - \hat{\gamma}_i)_2^2}{\sum_{i=1}^{n}(\gamma_i - \overline{\gamma})_2^2} \tag{2}$$

where n is the number of samples (molecules), γ_i is the measured physicochemical property, $\hat{\gamma}_i$ is predicted value, and $\overline{\gamma} = \frac{1}{N}\sum_{i=1}^{N}\gamma_i$ is the mean over the samples. The MAE is a commonly used metric due to being not very

sensitive to large errors and outliers, which is desirable if the dataset spans different scales. It is given as follows

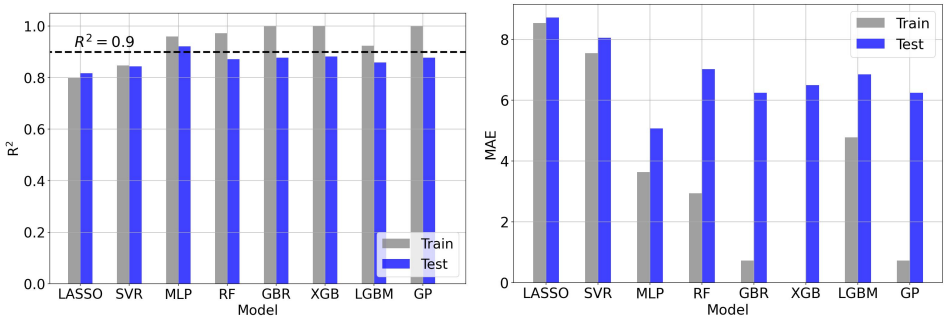
$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} \left|\gamma_i - \hat{\gamma}_i\right| . \tag{3}$$

Also, the MAPE metric is given by

$$\mathrm{MAPE} = 100 \times \frac{1}{n}\sum_{i=1}^{n} \left|\frac{\gamma_i - \hat{\gamma}_i}{\gamma_i}\right| . \tag{4}$$

Furthermore, to avoid overfitting we use k-fold cross-validation [54]. The models are trained using $k-1$ folds as the training data, and the remaining left-out fold is used as the validation set. Here, we split the training dataset into $k = 5$ folds. Also, the cross-validation is repeated 5 times with different random seeds to evaluate the robustness of this approach. Here, the ensemble models are optimized by grid-search over the number of estimators $\in [10, 20, 50, 100]$, i.e., the number of trees in the forest and the number of boosting stages to perform. For all models, the best number of estimators was 100. Moreover, we test different forms of the kernel for the GP model including the radial basis function (RBF) and the Matern family [59]. The RBF kernel returns the best predictions.

Figure 2(a) shows the train and test set R^2 score for the eight QSPR-ML models. As we can see, ensemble models such as RF, GBR, and XGB present high scores for the train set, with $R^2 \geq 0.95$. However, such models perform the worst in the test set, returning an R^2-score lower than 0.9. Here, we



(a) R^2-score (b) MAE

Figure 2: Accuracy metrics for the descriptor-based QSPR-ML models.

consider 0.9 a good target. Such behavior may suggest that these models are overfitting, which is typical when we are dealing with a small data regime. That is further confirmed in Fig. 2(b). We observe that the tree-based models present a lower MAE to the training set but the error increases significantly on the test set. The same behavior is depicted in the probabilistic GP model. Furthermore, linear models like LASSO and SVR (with a linear kernel) return a not satisfactory accuracy. That might suggest a non-linear link between the molecular descriptors and the cetane number.

Also, we can verify that the MLP model returns a satisfactory accuracy in both train and test sets. Thus, we top-ranked the MLP descriptor-based model for CN predictions since the goal here is to construct a QSPR model able to generalize patterns in training data so that it becomes feasible to predict CNs from new unseen data (molecules) correctly. Here, the MLP has 2 hidden layers and 32 neurons per layer. The MLP is constructed using Ray Tune [62] to search for the hyperparameters that give the lowest

mean squared error. The hidden layers use a hyperbolic tangent non-linearity activation function. The models are trained for 1,000 stochastic gradient descent steps using the Adam optimizer [63] with a learning rate of 5×10^{-3}, with a learning rate scheduler, that is used to drop 10 times on plateau during the training process. Indeed, MLPs are universal function approximators that can detect and decode intrinsic relations from data. MLP models are an effective tool for accurately predicting several physicochemical properties [40, 42].

Table 1 shows the test dataset accuracy metrics for the different models. As we can see, the MLP-QSPR model returns an MAE of 5.074, an acceptable value since the experimental error range of CN measurements is from 2.8 to 4.8 [64]. Also, by comparing the predictions of the present model with QSPR models available in the literature, we can verify our model presents competitive performance compared to the state-of-the-art QSPR models. In particular, the model by Guo et al. [27] used an ANN to construct predictive models for CNs of hydrocarbons and oxygenates with average absolute errors of 6.5 and 4.4 for compounds with and without rings. Finally, a QSPR model combining an active subspace method and linear regression returns an average absolute error of 5.0 [25]. Here, we do not claim that this represents a fair comparison as the models have been constructed on different train/test sets, but we use it to demonstrate the performance of the present methodology to build QSPR models.

Figure 3 shows the parity plots between the measured and the predicted

Table 1: Accuracy metrics of the ML models for the test set.

Model	**R^2-score**	**MAE**	**MAPE** (%)
LASSO	0.817	8.714	48.5
SVR	0.844	8.045	39.4
MLP	**0.921**	**5.074**	**19.4**
RF	0.871	7.018	38.5
GBR	0.877	6.237	26.1
XGB	0.882	6.498	34.9
LGBM	0.858	6.843	36.4
GP	0.887	6.411	32.1

CN for the five categories of hydrocarbons using the descriptor-based MLP model. Here, we use the whole dataset to show that the MLP model is not biased towards a specific category of hydrocarbons. The parity plot shows that the MLP model predictions lie near the diagonal $x = y$ line, which suggests that the model captures well the relationship between the molecular structure and CNs. Furthermore, we measure the structural deficiency of the MLP model (also called epistemic uncertainty [65]) for CN predictions. The model error follows a normal distribution with zero mean and constant variance, as shown in Fig. 4. This implies that such discrepancies are only due to uncorrelated noise/fluctuations in the data, i.e., not to structural deficiencies of the model.

In light of the design of alternative fuels, the knowledge of the physicochemical properties of such fuels raises the potential to leverage the efficiency of the existing global infrastructure and combustion engines, carbon neutrality (emissions), and financial viability. In this context, we extend the methodology to construct an IQSPR predictive model for the CN of esters and

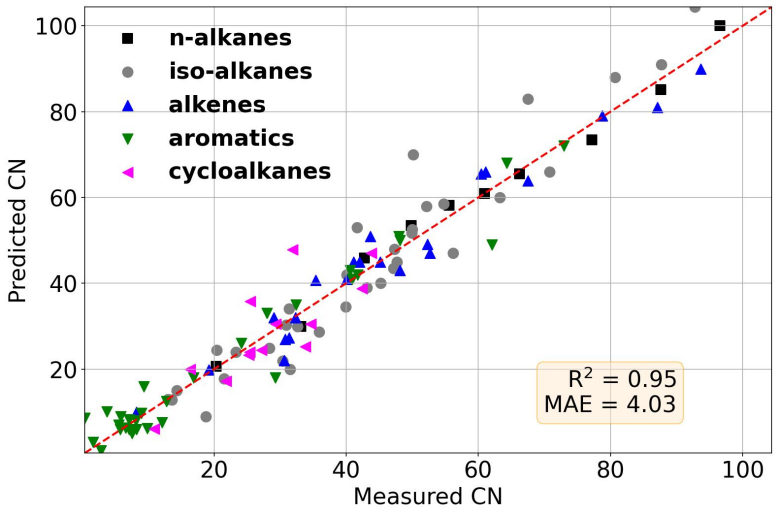


Figure 3: Parity plot between predicted and measured cetane number of hydrocarbons for the MLP model.

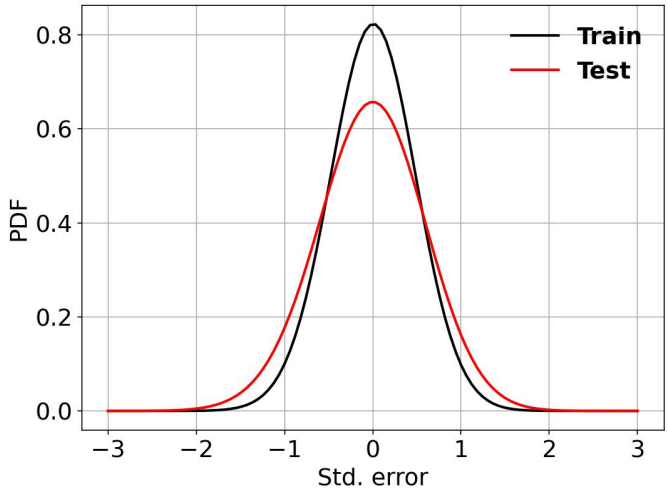
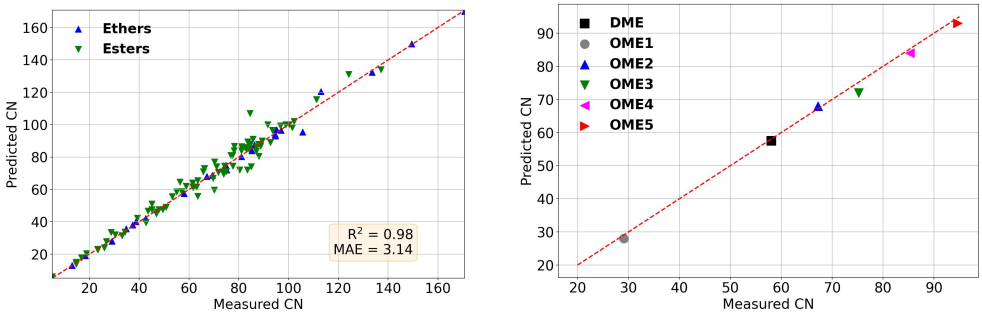


Figure 4: Probability distribution of the model error of the MLP model.

ethers. Biodiesel is a low-carbon alternative fuel to petrodiesel [66], which is primarily composed of fatty acid methyl esters (FAMEs). Furthermore, oxymethylene dimethyl ethers, including dimethyl ether (OME_0), have been



(a) (b) OME_x

Figure 5: Parity plot between predicted and measured cetane number of esters and ethers for the MLP model.

developed and deployed in heavy-duty transport. They have shown potential to be proxies of diesel-like fuels in difficult-to-decarbonize sectors. Therefore, constructing predictive models for esters and ethers' physicochemical properties is extremely important for the design of alternative liquid synthetic fuels. The results show that the approach can predict accurately the CNs of the main components of biodiesel, as shown in Fig. 5(a). Also, the parity plots between the measured and the predicted CN for OME_x using the descriptor-based IQSPR-MLP model are depicted in Fig. 5(b).

Here, the trained MLP-QSPR model for CNs of hydrocarbons is used to initialize the parameters of the MLP model for CNs of esters and ethers, which mimic a transfer learning approach [29], and the model is retrained using the ester and ether data. The results show that the model can predict the OME_x CNs very well, and the potential of the present approach to assist the design of new fuels. (see Table S2 in the *Supplementary Material* (SM) for a complete list of CNs of esters and ethers.)

4.2. Model Interpretability

In the current context, model interpretability means the ability to correctly and efficiently understand the relationship between the molecular structure and the physicochemical property, allowing the construction of IQSPR models that present the underlying chemistry behind predictions. Here, we do not claim to define the meaning of all molecular descriptors but instead, provide some physical/chemical basis for linking the top-ranked descriptors and physicochemical properties.

The key advantage of descriptors-based IQSPR ML models is the ability to reveal the descriptor's importance in predicting CNs. Here, feature importance is evaluated using the SHAP approach which is a game theoretic approach to explain the predictions of ML models. Fig. 6 shows the average impact on the cetane number of the top ten molecular descriptors for the MLP model. These descriptors suggest that molecular size, charge distribution, and atomistic properties affect the CN of hydrocarbons. In particular, we remark the spectral mean absolute deviation from Barysz matrix weight by vdw volume (SpMAD_Dzv) as the top-ranked descriptor. Indeed, matrix-based descriptors are topostructural, and topochemical indices are computed from a graph representation of the molecule [67]. More specifically, the molecule's structure is represented as a hydrogen-depleted graph, from which several graph-theoretical matrices are calculated and the chemical information is extracted through a set of basic algebraic operators. In particular, the SpMAD_Dzv is a weighted distance matrix-based descriptor

that encodes chemical and topological information derived from the presence of heteroatoms and multiple bonds in the molecule [68]. Such a descriptor accounts for molecular shape and branching, and it is computed by a sum of atomic properties of the carbon atom weighted based on bond order and atomic properties of the bonded atoms. So, it might be expected that the values tend to increase with the chain length, branching, and heteroatoms. The complementary information content (CIC) is an information-theoretic topological parameter calculated from the premise that each chemical structure contains a subset of elements. It is directly linked to the molecular structure and, specifically, it is a function of the number of atoms in the molecule [33]. Therefore, it is a natural choice since it is well known the CN for hydrocarbons is related to the number of carbons in the molecule [12]. Here, we also highlight the molecular connectivity indices (AXp-2dv) which encode different aspects of atom connectivity within a molecule, such as the amount of branching ring structures present, and flexibility. Molecular connectivity indices aim to measure the extent of branching of the carbon-atom skeleton of saturated hydrocarbons. These indices represent a good measure of the molecular area and are useful to describe the physicochemical properties of molecules as a measure of intermolecular interactions [69]. This descriptor plays a role in aromatic molecules as will be shown subsequently. Table 2 shows a complete definition of top-ranked descriptors.

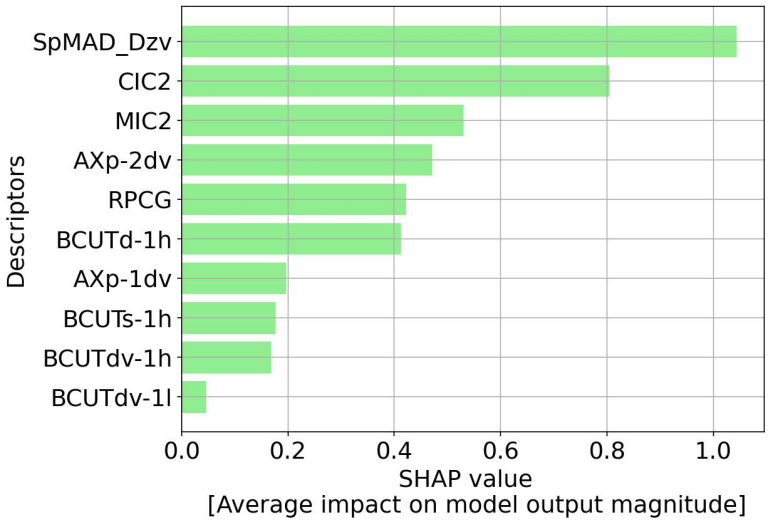


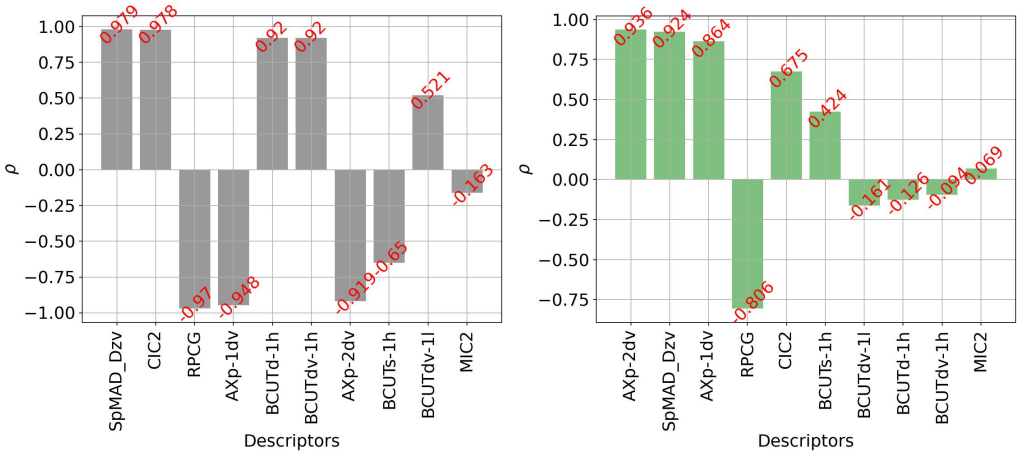
Figure 6: Feature importance of descriptor-based MLP model. The top 10 important features are measured as the average magnitude of SHapley Additive exPLanations (SHAP) values.

Table 2: List of top-ranked descriptors for hydrocarbons, esters, and ethers molecules.

Description	**Name**	**Dim**
Spectral mean absolute deviation from Barysz matrix weight by vdw volume	SpMAD_Dzv	2D
Spectral mean absolute deviation from Barysz matrix weight by polarizability	SpMAD_Dzp	2D
2-ordered complementary information content	CIC2	2D
2-ordered modified information content	MIC2	2D
2-ordered averaged Chi path weighted by valence electrons	AXp-2dv	2D
Relative positive charge	RPCG	2D
1-ordered averaged Chi path weighted by valence electrons	AXp-1dv	2D
Averaged ETA composite index for reference graph	AETA_eta	2D
Averaged ETA functionality index	AETA_eta_F	2D
Max of sCH3	MAXsCH3	2D
Sum of sCH3	SsCH3	2D
First highest eigenvalue of the Burden matrix weighted by sigma electrons	BCUTd-1h	2D
First highest eigenvalue of Burden matrix weighted by intrinsic state	BCUTs-1h	2D
First lowest eigenvalue of Burden matrix weighted by intrinsic state	BCUTs-1l	2D
First highest eigenvalue of Burden matrix weighted by valence electrons	BCUTdv-1h	2D
First lowest eigenvalue of Burden matrix weighted by valence electrons	BCUTdv-1l	2D
First lowest eigenvalue of Burden matrix weighted by vdw volume	BCUTv-1l	2D
First highest eigenvalue of Burden matrix weighted by Pauling EN	BCUTpe-1h	2D

Going further, we evaluate the correlation of these ten descriptors with different categories of hydrocarbons. Here, the aim is to evaluate the relationship between the molecular structures and descriptors. This allows leveraging the predictability of IQSPR models by selecting the most relevant descriptors regarding the categories of hydrocarbons and presenting the underlying basis behind predictions and decision-making in a physically explainable way. To do so, we compute Pearson's correlation between the molecular descriptors and the CNs for different categories of hydrocarbons, as shown in Fig. 7. The results show that CNs of alkanes have a high positive correlation with the SpMAD_Dzv and CIC as these descriptors are directly connected with the molecular size. They also present a negative correlation with the relative positive charge (RPCG) which is a charged partial surface area molecular descriptor to encode information about intermolecular hydrogen-bonding interactions [70]. This result suggests that polar interactions between molecules can have a strong influence on CNs and intermolecular hydrogen bonding may affect negatively the CNs of hydrocarbons. Regarding aromatics, the results show a high correlation of CNs with molecular connectivity indices (AXp-1dv and AXp-2dv) since they are related to the amount of branching ring structures in the molecule. It is well known that aromatic structures with multiple rings have lower CNs compared to single-ring structures [12]. (Fig. S1 in the SM illustrates the correlation analysis for iso-alkanes, cycloalkanes, and alkenes.)

In fact, such correlation analysis allows us to construct IQSPR predictive



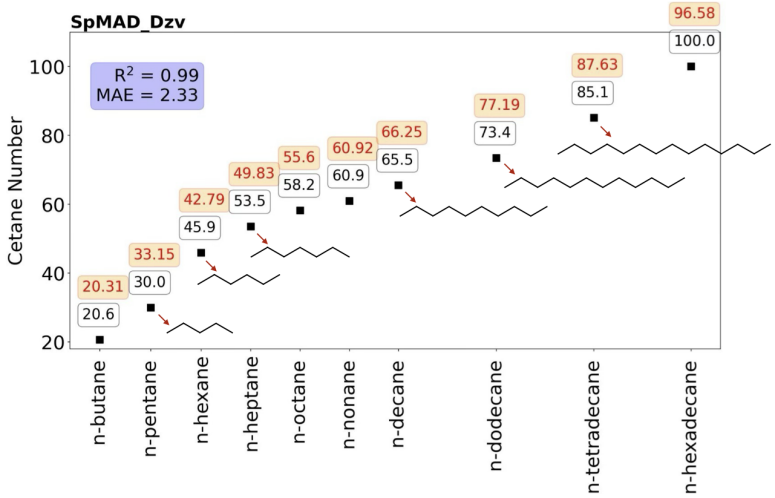
(a) alkanes (b) aromatics

Figure 7: Pearson's correlation between the molecular descriptors and cetane number for different molecular structures.

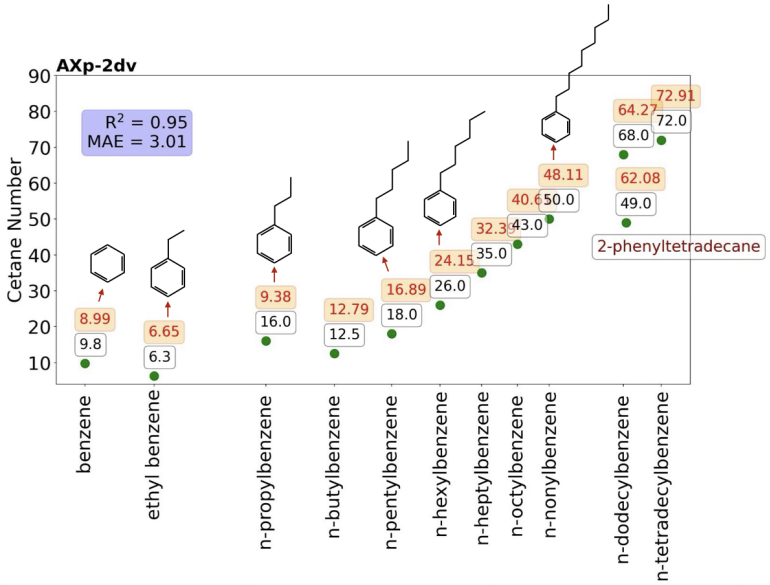
models by removing less important descriptors but keeping a satisfactory accuracy for different categories of hydrocarbons. This saves computational resources without loss of prediction performance. Figures 8(a) and 8(b) show the CNs as a function of the top correlated descriptor for alkanes and aromatics. As we can note, the SpMAD_Dzv descriptor has a direct correlation with CNs for alkanes. The x-axis represents the name of the molecules weighed by the value of the molecular descriptor. This result suggests that matrix-based descriptors may be a good choice to construct IQSPR predictive models for CNs prediction of alkanes since they encode topological information of the molecule such as molecular shape and size. Also, the influence of molecular size on CNs is confirmed, as it has a high positive correlation with SpMAD_Dzv, as shown in Fig. 7. The same tendency is confirmed for alkenes and iso-alkanes, as we can see in Fig. S2 in the SM. For iso-alkanes,

if the branch is concentrated at one end of the molecular structure, leaving a long chain at the other end, the CNs tend to increase with the size of the long chain. Also, the size of the branch and the position play a role but are secondary compared with the size of the long chain. Similar to alkanes, CNs of alkenes tend to increase with molecular size, aligning with the choice of SpMAD_Dzv as the most important descriptor. On the other hand, aromatics have presented a non-monotonic behavior with the top correlated descriptor (AXp-2dv). Also, as we can see, the CNs of aromatics tend to increase with the size of the n-alkyl group attached to the aromatic ring, aligning with the results found in the literature. This explains the choice of SpMAD_Dzv as the second most correlated descriptor since it is directly linked with molecular size [12]. Furthermore, these results suggested that the number of descriptors required to describe the molecule correctly might increase proportionally with the complexity of the structure. That is the case of n-dodecylbenzene and 2-phenyltetradecane, where both have very similar structures, as can be seen in Fig. 9, and more than one descriptor is required to distinguish the molecules. Thus, the selection of the most relevant descriptors is crucial, the wrong choice may hamper the predictability of QSPR-ML models, especially in a small data regime.

Finally, we assess the correlation analysis of the ten top descriptors with CNs of esters and ethers. Note that CNs of esters have correlations with topological indices such as the extended topochemical atom (ETA) index and the electrotopological state indices (E-state) for the -CH3- groups in the



(a) alkanes



(b) aromatics

Figure 8: Predictions of the descriptor-based IQSPR-ML model for cetane number of hydrocarbons. Measured CNs (black) and predicted CNs (red).

molecule (MAXsCH3 and SsCH3). The ETA descriptors are directly correlated with molecular shape and size [71]. The E-state indices combine the

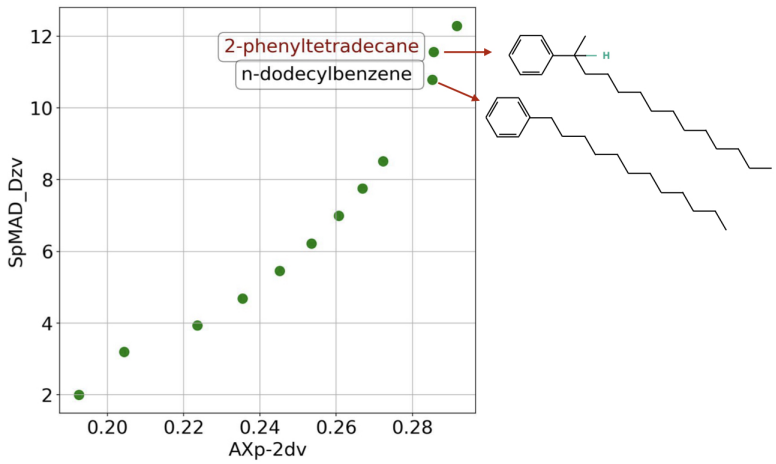
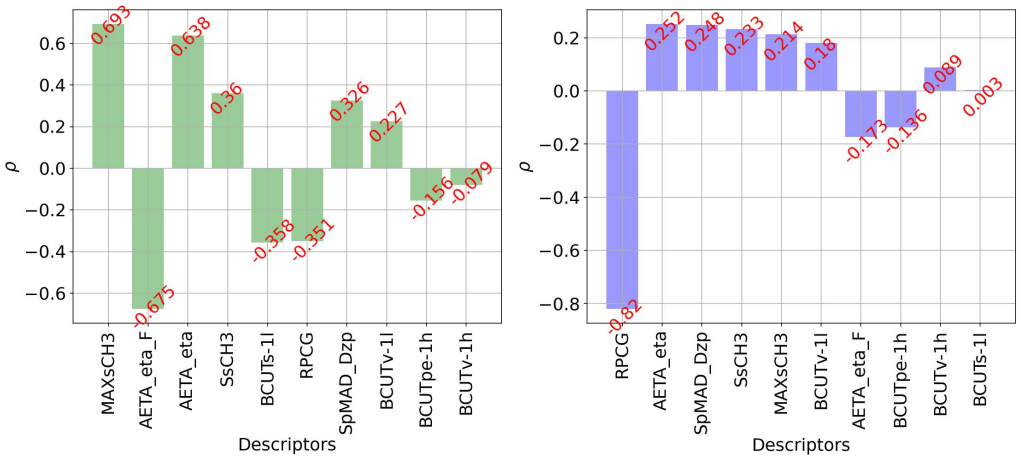


Figure 9: Comparison of molecular descriptors of n-dodecylbenzene and 2-phenyltetradecane.

electronic state of the bonded hydrogen atoms with the topological structure of the molecule. These molecular descriptors have been demonstrated very useful to construct predictive QSPR models [72]. Moreover, we can verify that ether molecules present a high negative correlation with the RPCG descriptor. This result might suggest that descriptors related to intermolecular bonding interactions can be a good choice to construct IQSPR predictive models for CNs prediction of renewable fuels like oxymethylene dimethyl ethers, as shown in Fig. 11.

5. Conclusions

We have presented a novel ML-assisted liquid synthetic fuel property prediction methodology that can potentially assist in fully sustainable fuel utilization, i.e., the engines are 100% powered by sustainable/renewable syn-



(a) esters (b) ethers

Figure 10: Pearson's correlation between the molecular descriptors and cetane number for different molecular structures.

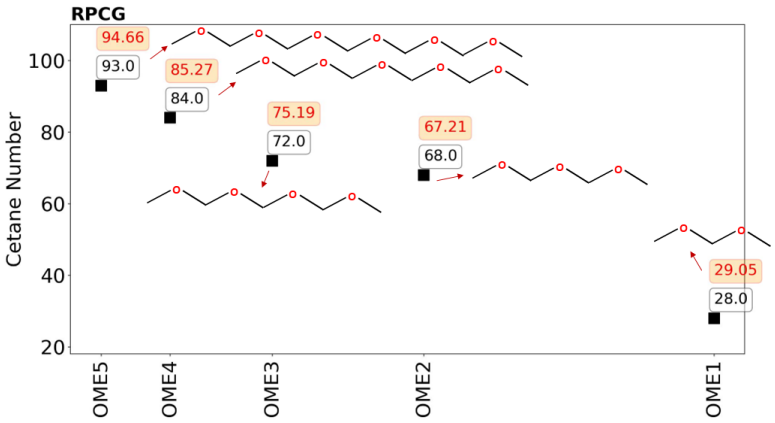


Figure 11: Predictions of the descriptor-based IQSPR-ML model for cetane number of OME_x. Measured CNs (black) and predicted CNs (red).

thetic fuels. The methodology developed can be summarized as follows. Molecular descriptors have proven to be a useful tool to characterize the molecular structure of fuels. The use of feature selection methods may reduce the number of features by minimizing the correlations between the molecular

descriptors to develop high-performing models. The novelty of the presented approach lies in the use of ML models to construct robust interpretable QSPR predictive models of fuel properties based on molecular structure. Such an approach provides tools to understand the underlying physics/chemistry of the link between the molecular structure and the fuel properties.

The effectiveness of the approach was tested by constructing IQSPR machine-learning models for CN predictions. CN is a key physicochemical property to measure the quality of diesel fuels during the ignition process. Therefore, developing accurate IQSPR predictive models for CN predictions is essential for the design of alternative diesel fuels. The results showed that the methodology predicts well the CN based on different fuel molecules. Moreover, the key advantage of the resulting models is the ability to correctly and efficiently reveal the relationship between the molecular structure and CNs.

It is worth mentioning that the present approach was used to predict the CN, but such a methodology can be extended to any important physicochemical property for the design of liquid alternative fuels because the ML approach is based on an interpretable quantitative structure-property relationship that is ubiquitous for data analytics of fuel properties. Indeed, the prediction of reliable physicochemical properties of alternative fuels is an important step forward towards the generation of digital tools that can assist in decarbonization by the use of renewable fuels. In this context, future studies will be focused on extending the methodology to inverse fuel design

by exploring the chemical space to learn fuel blends matching the desired properties, leveraging the design of alternative fuels toward fully sustainable fuel utilization.

Acknowledgement

This study is supported by the UK Physical Sciences Research Council under Grant No. EP/X019551/1.

References

[1] IEA, Global CO_2 emissions from transport by sub-sector in the Net Zero Scenario, IEA, Paris https://www.iea.org/data-and-statistics/charts/global-co2-emissions-from-transport-by-sub-sector-in-the-net-zero-scenario-2000-2030, IEA. Licence: CC BY 4.0 (2000).

[2] G. Hutchings, M. Davidson, P. Atkins, P. Collier, N. Jackson, A. Morton, M. Muskett, M. Rosseinsky, P. Styring, P. Thornley, C. Williams, Sustainable synthetic carbon based fuels for transport: Policy Briefing, The Royal Society, United Kingdom, 2019.

[3] F. Liu, M. Shafique, X. Luo, Literature review on life cycle assessment of transportation alternative fuels, Environmental Technology & Innovation 32 (2023) 103343. doi:https://doi.org/10.1016/j.eti.2023.103343.

URL https://www.sciencedirect.com/science/article/pii/S2352186423003395

[4] P. Styring, G. R. M. Dowson, I. O. Tozer, Synthetic fuels based on dimethyl ether as a future non-fossil fuel for road transport from sustainable feedstocks, Frontiers in Energy Research 9 (2021). doi:10.3389/fenrg.2021.663331.
URL https://www.frontiersin.org/articles/10.3389/fenrg.2021.663331

[5] V. Ram, S. R. Salkuti, An overview of major synthetic fuels, Energies 16 (6) (2023).
URL https://www.mdpi.com/1996-1073/16/6/2834

[6] J. C. Bart, N. Palmeri, S. Cavallaro, 15 - evolution of biodiesel and alternative diesel fuels, in: J. C. Bart, N. Palmeri, S. Cavallaro (Eds.), Biodiesel Science and Technology, Woodhead Publishing Series in Energy, Woodhead Publishing, 2010, pp. 713–782. doi:https://doi.org/10.1533/9781845697761.713.
URL https://www.sciencedirect.com/science/article/pii/B9781845695910500152

[7] A. Omari, B. Heuser, S. Pischinger, Potential of oxymethylenether-diesel blends for ultra-low emission engines, Fuel 209 (2017) 232–237. doi:https://doi.org/10.1016/j.fuel.2017.07.107.

URL https://www.sciencedirect.com/science/article/pii/S0016236117309699

[8] D. Pélerin, K. Gaukel, M. Härtl, E. Jacob, G. Wachtmeister, Potentials to simplify the engine system using the alternative diesel fuels oxymethylene ether ome1 and ome3-6 on a heavy-duty engine, Fuel 259 (2020) 116231. doi:https://doi.org/10.1016/j.fuel.2019.116231.
URL https://www.sciencedirect.com/science/article/pii/S0016236119315856

[9] J. Burger, M. Siegert, E. Ströfer, H. Hasse, Poly(oxymethylene) dimethyl ethers as components of tailored diesel fuel: Properties, synthesis and purification concepts, Fuel 89 (11) (2010) 3315–3319. doi:https://doi.org/10.1016/j.fuel.2010.05.014.
URL https://www.sciencedirect.com/science/article/pii/S0016236110002383

[10] Y. Putrasari, O. Lim, Dimethyl ether as the next generation fuel to control nitrogen oxides and particulate matter emissions from internal combustion engines: A review, ACS Omega 7 (1) (2022) 32–37.

[11] J. Liu, L. Wang, P. Wang, P. Sun, H. Liu, Z. Meng, L. Zhang, H. Ma, An overview of polyoxymethylene dimethyl ethers as alternative fuel for compression ignition engines, Fuel 318 (2022) 123582. doi:https://doi.org/10.1016/j.fuel.2022.123582.

URL https://www.sciencedirect.com/science/article/pii/S0016236122004471

[12] P. Ghosh, S. B. Jaffe, Detailed composition-based model for predicting the cetane number of diesel fuels, Industrial & Engineering Chemistry Research 45 (1) (2006) 346–351. doi:10.1021/ie0508132.
URL https://doi.org/10.1021/ie0508132

[13] M. Dahmen, W. Marquardt, A novel group contribution method for the prediction of the derived cetane number of oxygenated hydrocarbons, Energy & Fuels 29 (9) (2015) 5781–5801.

[14] H. Liu, K. H. Yoo, A. L. Boehman, Z. Zheng, Experimental study of autoignition characteristics of the ethanol effect on biodiesel/n-heptane blend in a motored engine and a constant-volume combustion chamber, Energy & Fuels 32 (2) (2018) 1884–1892.

[15] M. C. Ingham, J. A. Bert, L. J. Painter, Improved predictive equations for cetane number, SAE Transactions 95 (1986) 72–85.

[16] N. Ladommatos, J. Goacher, Equations for predicting the cetane number of diesel fuels from their physical properties, Fuel 74 (7) (1995) 1083–1093.

[17] Y. Zhou, Artificial intelligence in renewable systems for transformation towards intelligent buildings, Energy and AI 10 (2022) 100182. doi:https://doi.org/10.1016/j.egyai.2022.100182.

URL https://www.sciencedirect.com/science/article/pii/S2666546822000313

[18] Z. Liu, Y. Sun, C. Xing, J. Liu, Y. He, Y. Zhou, G. Zhang, Artificial intelligence powered large-scale renewable integrations in multi-energy systems for carbon neutrality transition: Challenges and future perspectives, Energy and AI 10 (2022) 100195. doi:https://doi.org/10.1016/j.egyai.2022.100195.
URL https://www.sciencedirect.com/science/article/pii/S2666546822000428

[19] R. Li, J. M. Herreros, A. Tsolakis, W. Yang, Machine learning-quantitative structure property relationship (ml-qspr) method for fuel physicochemical properties prediction of multiple fuel types, Fuel 304 (2021) 121437.

[20] F. vom Lehn, B. Brosius, R. Broda, L. Cai, H. Pitsch, Using machine learning with target-specific feature sets for structure-property relationship modeling of octane numbers and octane sensitivity, Fuel 281 (2020) 118772.

[21] N. Kuzhagaliyeva, S. Horváth, J. Williams, et al., Artificial intelligence-driven design of fuel mixtures, Commun Chem 111 (1) (2022).

[22] R. S. Freitas, Ágatha P.F. Lima, C. Chen, F. A. Rochinha, D. Mira, X. Jiang, Towards predicting liquid fuel physicochemical properties us-

ing molecular dynamics guided machine learning models, Fuel 329 (2022) 125415.

[23] A. R. Katritzky, M. Kuanar, S. Slavov, C. D. Hall, M. Karelson, I. Kahn, D. A. Dobchev, Quantitative correlation of physical and chemical properties with chemical structure: Utility for prediction, Chemical Reviews 110 (10) (2010) 5714–5789, pMID: 20731377. arXiv:https://doi.org/10.1021/cr900238d, doi:10.1021/cr900238d.
URL https://doi.org/10.1021/cr900238d

[24] D. V. Nadai, J. B. Simões, C. E. Gatts, P. C. Miranda, Inference of the biodiesel cetane number by multivariate techniques, Fuel 105 (2013) 325–330. doi:https://doi.org/10.1016/j.fuel.2012.06.018.
URL https://www.sciencedirect.com/science/article/pii/S0016236112004383

[25] C. Guan, J. Zhai, D. Han, Cetane number prediction for hydrocarbons from molecular structural descriptors based on active subspace methodology, Fuel 249 (2019) 1–7.

[26] D. A. Saldana, L. Starck, P. Mougin, B. Rousseau, L. Pidol, N. Jeuland, B. Creton, Flash point and cetane number predictions for fuel compounds using quantitative structure property relationship (qspr) methods, Energy & Fuels 25 (9) (2011) 3900–3908.

[27] Z. Guo, K. H. Lim, M. Chen, B. J. R. Thio, B. L. W. Loo, Predicting

cetane numbers of hydrocarbons and oxygenates from highly accessible descriptors by using artificial neural networks, Fuel 207 (2017) 344–351.

[28] Y. Kim, J. Cho, N. Naser, S. Kumar, K. Jeong, R. L. McCormick, P. C. St. John, S. Kim, Physics-informed graph neural networks for predicting cetane number with systematic data quality analysis, Proceedings of the Combustion Institute 39 (4) (2023) 4969–4978.

[29] A. M. Schweidtmann, J. G. Rittig, A. König, M. Grohe, A. Mitsos, M. Dahmen, Graph neural networks for prediction of fuel ignition quality, Energy & Fuels 34 (9) (2020) 11395–11407. arXiv:https://doi.org/10.1021/acs.energyfuels.0c01533, doi:10.1021/acs.energyfuels.0c01533.
URL https://doi.org/10.1021/acs.energyfuels.0c01533

[30] R. Machlev, L. Heistrene, M. Perl, K. Levy, J. Belikov, S. Mannor, Y. Levron, Explainable artificial intelligence (xai) techniques for energy and power systems: Review, challenges and opportunities, Energy and AI 9 (2022) 100169. doi:https://doi.org/10.1016/j.egyai.2022.100169.
URL https://www.sciencedirect.com/science/article/pii/S2666546822000246

[31] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley, 2000.

[32] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Vol. 30, Curran Associates, Inc., 2017.

[33] R. Todeschini, V. Consonni, Handbook of molecular descriptors. WileyVCH, Weinheim, Vol. 11, 2000.

[34] H. Moriwaki, Y.-S. Tian, N. Kawashita, T. Takagi, Mordred: a molecular descriptor calculator, Journal of Cheminformatics 10 (1) (2018) 4.

[35] F. Grisoni, D. Ballabio, R. Todeschini, V. Consonni, Molecular descriptors for structure-activity applications: A hands-on approach, Computational Toxicology: Methods and Protocols, Springer New York (2018) 3–53.

[36] F. R. Burden, M. J. Polley, D. A. Winkler, Toward novel universal descriptors: Charge fingerprints, Journal of Chemical Information and Modeling 49 (3) (2009) 710–715, pMID: 19434903.

[37] V. H. Masand, V. Rastija, Pydescriptor: A new pymol plugin for calculating thousands of easily understandable molecular descriptors, Chemometrics and Intelligent Laboratory Systems 169 (2017) 12–18.

[38] K. JOBACK, R. REID, Estimation of pure-component properties from group-contributions, Chemical Engineering Communications 57 (1-6) (1987) 233–243. `arXiv:https://doi.org/10.1080/`

00986448708960487, doi:10.1080/00986448708960487.
URL https://doi.org/10.1080/00986448708960487

[39] R. Gani, B. Nielsen, A. Fredenslund, A group contribution approach to computer-aided molecular design, AIChE Journal 37 (9) (1991) 1318–1332. arXiv:https://aiche.onlinelibrary.wiley.com/doi/pdf/10.1002/aic.690370905, doi:https://doi.org/10.1002/aic.690370905.
URL https://aiche.onlinelibrary.wiley.com/doi/abs/10.1002/aic.690370905

[40] P. Santak, G. Conduit, Predicting physical properties of alkanes with neural networks, Fluid Phase Equilibria 501 (2019) 112259.

[41] P. C. St. John, P. Kairys, D. D. Das, C. S. McEnally, L. D. Pfefferle, D. J. Robichaud, M. R. Nimlos, B. T. Zigler, R. L. McCormick, T. D. Foust, Y. J. Bomble, S. Kim, A quantitative model for the prediction of sooting tendency from molecular structure, Energy & Fuels 31 (9) (2017) 9983–9990.

[42] T. Kessler, P. C. St. John, J. Zhu, C. S. McEnally, L. D. Pfefferle, J. H. Mack, A comparison of computational models for predicting yield sooting index, Proceedings of the Combustion Institute 38 (1) (2021) 1385–1393.

[43] P. C. St. John, P. Kairys, D. D. Das, C. S. McEnally, L. D. Pfefferle,

D. J. Robichaud, M. R. Nimlos, B. T. Zigler, R. L. McCormick, T. D. Foust, Y. J. Bomble, S. Kim, A quantitative model for the prediction of sooting tendency from molecular structure, Energy & Fuels 31 (9) (2017) 9983–9990.

[44] C. W. Yap, Padel-descriptor: An open source software to calculate molecular descriptors and fingerprints, Journal of Computational Chemistry 32 (7) (2011) 1466–1474.

[45] A. E. Comesana, T. T. Huntington, C. D. Scown, K. E. Niemeyer, V. H. Rapp, A systematic method for selecting molecular descriptors as features when training models for predicting physiochemical properties, Fuel 321 (2022) 123836.

[46] B. Hemmateenejad, R. Miri, M. Akhond, M. Shamsipur, Qsar study of the calcium channel antagonist activity of some recently synthesized dihydropyridine derivatives. an application of genetic algorithm for variable selection in mlr and pls methods, Chemometrics and Intelligent Laboratory Systems 64 (1) (2002) 91–99.

[47] J. Li, D. Luo, T. Wen, Q. Liu, Z. Mo, Representative feature selection of molecular descriptors in qsar modeling, Journal of Molecular Structure 1244 (2021) 131249.

[48] B. Hollas, An analysis of the autocorrelation descriptor for molecules, Journal of Mathematical Chemistry 33 (2) (2003) 91–101.

[49] Spearman Rank Correlation Coefficient, Springer New York, New York, NY, 2008, pp. 502–505.

[50] M. C. J.C. Dearden, K. Kaiser, How not to develop a quantitative structure–activity or structure–property relationship (qsar/qspr), SAR and QSAR in Environmental Research 20 (3-4) (2009) 241–266.

[51] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[52] N. V. S. Avula, S. K. Veesam, S. Behera, S. Balasubramanian, Building robust machine learning models for small chemical science data: the case of shear viscosity of fluids, Machine Learning: Science and Technology 3 (4) (2022) 045032.

[53] A. K. Chew, M. Sender, Z. Kaplan, A. Chandrasekaran, A. R. Browning, H. S. Kwak, M. D. Halls, M. A. F. Afzal, Advancing material property prediction: Using physics-informed machine learning models for viscosity (jun 2023). doi:10.26434/chemrxiv-2023-1qfw8.
URL https://doi.org/10.26434%2Fchemrxiv-2023-1qfw8

[54] S. L. Brunton, J. N. Kutz, Data-Driven Science and Engineering: Ma-

chine Learning, Dynamical Systems, and Control, Cambridge University Press, 2019. doi:10.1017/9781108380690.

[55] G. Varoquaux, Cross-validation failure: Small sample sizes lead to large error bars, NeuroImage 180 (2018) 68–77, new advances in encoding and decoding of brain signals.

[56] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, Lightgbm: A highly efficient gradient boosting decision tree, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Vol. 30, Curran Associates, Inc., 2017.

[57] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 785–794.

[58] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Vol. 32, Curran Associates, Inc., 2019.

[59] C. E. Rasmussen, Gaussian Processes in Machine Learning, Springer Berlin Heidelberg, Berlin, Heidelberg, 2004, pp. 63–71. doi:10.1007/978-3-540-28650-9_4.
URL https://doi.org/10.1007/978-3-540-28650-9_4

[60] M. Murphy, J. Taylor, R. McCormick, Compendium of experimental cetane number data, National Renewable Energy Lab. (NREL), Golden, CO (United States) (01 2004).

[61] P. G. Szymkowicz, J. Benajes, Development of a diesel surrogate fuel library, Fuel 222 (2018) 21–34.

[62] R. Liaw, E. Liang, R. Nishihara, P. Moritz, J. E. Gonzalez, I. Stoica, Tune: A research platform for distributed model selection and training, arXiv preprint arXiv:1807.05118 (2018).

[63] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization (2017). arXiv:1412.6980.

[64] R. C. Santana, P. T. Do, M. Santikunaporn, W. E. Alvarez, J. D. Taylor, E. L. Sughrue, D. E. Resasco, Evaluation of different reaction strategies for the improvement of cetane number in diesel fuels, Fuel 85 (5) (2006) 643–656. doi:https://doi.org/10.1016/j.fuel.2005.08.028.
URL https://www.sciencedirect.com/science/article/pii/S0016236105003133

[65] E. Hüllermeier, W. Waegeman, Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods, Machine Learning 110 (2021) 457–506.

[66] H. K. Jeswani, A. Chilvers, A. Azapagic, Environmental sustainability of biofuels: a review, Proc. Math. Phys. Eng. Sci. 476 (2243) (2020) 20200351.

[67] F. Grisoni, D. Reker, P. Schneider, L. Friedrich, V. Consonni, R. Todeschini, A. Koeberle, O. Werz, G. Schneider, Matrix-based molecular descriptors for prospective virtual compound screening, Molecular Informatics 36 (1-2) (2017) 1600091. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/minf.201600091, doi:https://doi.org/10.1002/minf.201600091.
URL https://onlinelibrary.wiley.com/doi/abs/10.1002/minf.201600091

[68] O. Ivanciuc, T. Ivanciuc, D. Cabrol-Bass, A. T. Balaban, Comparison of weighting schemes for molecular graph descriptors: Application in quantitative structure-retention relationship models for alkylphenols in gas-liquid chromatography, Journal of Chemical Information and Computer Sciences 40 (3) (2000) 732–743, pMID: 10850777. arXiv:https://doi.org/10.1021/ci990129w, doi:10.1021/ci990129w.
URL https://doi.org/10.1021/ci990129w

[69] E. Estrada, Physicochemical interpretation of molecular connectivity

indices, The Journal of Physical Chemistry A 106 (39) (2002) 9085–9091. arXiv:https://doi.org/10.1021/jp026238m, doi:10.1021/jp026238m.
URL https://doi.org/10.1021/jp026238m

[70] D. T. Stanton, L. M. Egolf, P. C. Jurs, M. G. Hicks, Computer-assisted prediction of normal boiling points of pyrans and pyrroles, Journal of Chemical Information and Computer Sciences 32 (4) (1992) 306–316.

[71] K. Roy, G. Ghosh, Qstr with extended topochemical atom indices. 2. fish toxicity of substituted benzenes, Journal of Chemical Information and Computer Sciences 44 (2) (2004) 559–567, pMID: 15032536. arXiv:https://doi.org/10.1021/ci0342066, doi:10.1021/ci0342066.
URL https://doi.org/10.1021/ci0342066

[72] L. H. Hall, L. B. Kier, Electrotopological state indices for atom types: A novel combination of electronic, topological, and valence state information, Journal of Chemical Information and Computer Sciences 35 (6) (1995) 1039–1045. arXiv:https://doi.org/10.1021/ci00028a014, doi:10.1021/ci00028a014.
URL https://doi.org/10.1021/ci00028a014