

Towards Model-Based Identification of Biofuels for Compression Ignition Engines

Manuel Dahmen, Manuel Hechinger, Juan Victoria Villeda and Wolfgang Marquardt
RWTH Aachen University

ABSTRACT

Depleting fossil resources, a constantly rising energy demand and worries about greenhouse gas emissions force society to explore novel concepts for future mobile propulsion. In the context of biofuels, the identification of novel, sustainably producible, tailored molecules meeting the property specifications derived from advanced engine combustion concepts therefore constitutes a major objective. Due to the tremendous amount of possible molecular structures, solely experimental search strategies are infeasible and highlight the need for a computer-aided biofuel identification framework. To this end, a holistic approach for deriving truly predictive Quantitative-Structure-Property-Relationships (QSPRs) for engine-relevant fuel properties is presented. Such QSPRs are combined with a rigorous generation of molecular structures aiming at identification of single-compound fuel candidates for use in compression ignition (CI) engines. Besides addressing thermophysical fuel properties relevant to CI engines, an accurate QSPR model for the cetane number has been established and integrated into the fuel identification framework. A comprehensive list of potential CI fuel molecules is generated automatically and subsequently screened by sequentially applying property constraints to identify those compounds predicted to exhibit all desired physicochemical properties. In conjunction with the rigorous generation of structures, the outlined holistic QSPR methodology thus represents a feasible strategy for rapid computer-aided identification of fuel candidates tailored for optimal engine performance.

CITATION: Dahmen, M., Hechinger, M., Victoria Villeda, J. and Marquardt, W., "Towards Model-Based Identification of Biofuels for Compression Ignition Engines," *SAE Int. J. Fuels Lubr.* 5(3):2012, doi:10.4271/2012-01-1593.

INTRODUCTION

Limited availability of fossil oil, constantly growing demand in energy and transportation, and worries about climate gas emissions drive the exploration of novel concepts for future mobile propulsion. Liquid carbon-based energy carriers offer high volumetric energy densities and infrastructure for their storage and delivery exists in almost every part of the world. The inevitable transition to a renewable carbon source does not constitute a challenge only, but should be considered an opportunity [1]. While the use of biomass as a feedstock addresses the accumulation of carbon-dioxide in the earth's atmosphere, the design of tailored fuel molecules offers great potential in achieving optimal fuel properties important to manufacturing, combustion, storage and delivery.

Biodiesel, a mixture of fatty acid methyl esters, constitutes a renewable fuel suitable to compression ignition (CI) engines. However, the production of biodiesel from plant oils is feedstock limited at this time [2,3]. Alternatively, the gasification of biomass enables a subsequent synthesis of Fischer-Tropsch diesel from syngas. The composition of

Fischer-Tropsch diesel depends on the reaction conditions and the catalyst employed in the synthesis process [3,4]. In contrast, the current contribution focuses on the identification of novel single-compound fuels as they might result from direct refunctionalization of lignocellulosic biopolymers to fuel compounds by means of low-temperature production processes [1,5]. The Cluster of Excellence (CoE) "Tailor-Made Fuels from Biomass" (TMFB) at RWTH Aachen University aims at a holistic, sustainable approach from biomass supply to engine combustion employing the entire plant as a carbon resource and preserving the natural molecular structure in the raw materials to the extent possible. In the present contribution, the space of feasible product molecular structures exhibiting desired CI engine-relevant properties (cf. Figure 1) shall be explored by an approach based on generating and testing potential molecular structures in a computer-based framework.

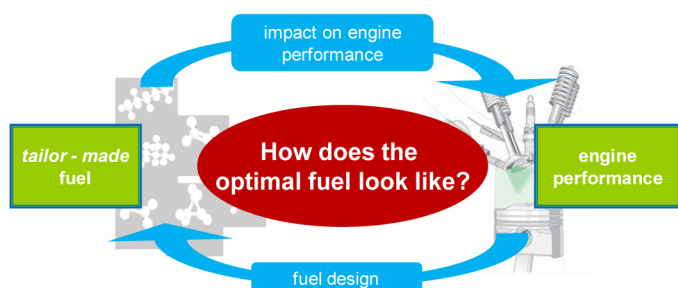


Figure 1. TMFB approach to fuel design: engine performance requirements define constraints for a tailored molecular structure to be used as a future biofuel

Properties important to CI engines specified in the CoE [6] and considered in rating potential fuel structures in the present contribution include cetane number, enthalpy of combustion, normal boiling point, liquid density, enthalpy of vaporization, melting point, liquid viscosity and surface tension. In addition, the flash point is considered in order to cover safety aspects in fuel handling and storage. The cetane number is utilized for rating a fuel's tendency to auto-ignite. A low cetane number implies a long ignition delay, which is measured from fuel injection into hot air until the event of auto-ignition takes place [7]. A long ignition delay allows for good premixing associated with reduced soot formation [8]. However, shorter ignition delays (higher cetane numbers) are expected to be beneficial for reduction of carbon-monoxide, unburned hydrocarbon and noise emissions [8]. The enthalpy of combustion and the liquid density determine the volumetric energy density of a fuel. Hence, both property values should be maximized. Normal boiling point and enthalpy of vaporization should be limited in order to achieve full and rapid vaporization of the injected fuel. Since the overall ignition delay is comprised of physical ignition delay (due to vaporization) and chemical ignition delay (due to combustion kinetics), fast vaporization is expected to enable improved homogeneity of combustion [8]. Viscosity and surface tension are important to in-cylinder fuel atomization [9] and the melting point is used in the assessment of low temperature behavior of the single-compound fuel candidates.

The present contribution outlines a framework for the identification of promising biofuel molecules including algorithmic generation of candidate molecular structures and subsequent testing of these structures by means of predictive property models. A holistic property modeling methodology is employed to describe all eight above-mentioned thermophysical properties and the cetane number as a largely chemistry controlled-property, thereby highlighting its feasibility and broad applicability range. The considered properties are not intended to represent a final list of all relevant fuel properties. Instead, the outlined methodology can be applied directly to further properties if sufficient measurement data is available. This is the first time a rigorous generation of structures is combined with a holistic

property modeling approach to rapidly narrow down the set of potential CI fuel candidates. Subsequently, experimental investigations can be directed to the most promising molecular structures. This idea constitutes the core of the inverse fuel design problem (cf. Figure 1), which aims at tailored molecular structures exhibiting optimal engine-relevant properties.

FUEL IDENTIFICATION FRAMEWORK

The product design task of identifying novel single-compound fuel candidates (cf. Figure 1) cannot be approached solely by experiments as the molecular search space under consideration consists of a tremendous amount of largely unexplored molecular structures. Facing this challenge, a model-based identification framework directing experimental efforts to the most promising structures has been proposed by our group recently [10]. The inverse problem of finding tailored fuel molecules exhibiting optimal engine-relevant properties is addressed by a generate-and-test-approach. To this end, the process is initiated by a rigorous generation of molecules in a certain carbon range, e.g., by defining sum formulas as inputs to a mathematically rigorous structure generator [11] (cf. Figure 2). Unstable bonds and sub-structures which are expected to be unstable are excluded from generation. An empirical correlation between enthalpy of combustion and two structural descriptors (namely the number of carbon atoms (nC) and the sum of conventional bond orders (SCBO) [12]) is translated into a structural constraint for the molecule generation. This way, only molecules with an enthalpy of combustion of at least around 38 MJ/kg are generated [10].

The next step involves the construction of mathematical models to predict fuel properties solely based on molecular structure. Subsequently, such models are applied to predict property data for the generated molecular structures. The application of constraints on these properties obtained from engine research gradually reduces the number of molecules for further investigation. In case of fuels for use in CI engines, constraints [6] are applied to cetane number, boiling point, enthalpy of combustion, liquid density, melting point, liquid viscosity and surface tension with the aim of stepwise compound exclusion (cf. Figure 2). At the present state, enthalpy of vaporization and flash point, which are calculated for the resulting set of compounds, constitute additional information for manual selection of candidates to be forwarded to experimental validation or synthetic accessibility determination. It should be noted, that a molecule exhibiting desired values for the above mentioned properties is still denoted a *candidate* fuel compound since additional properties and phenomena, e.g., ignition behavior beyond cetane number, have to be studied in order to draw final conclusions on the feasibility of the structure as a fuel compound. Moreover, the term *candidate fuel compound* refers to a molecule which is *predicted* to fulfill the property

constraints. Obviously, any property model might fail in adequately characterizing the properties of a molecular structure. Thus, experimental validation of the considered properties should be performed before continuing with further types of experiments, e.g., engine experiments or detailed studies on combustion kinetics.

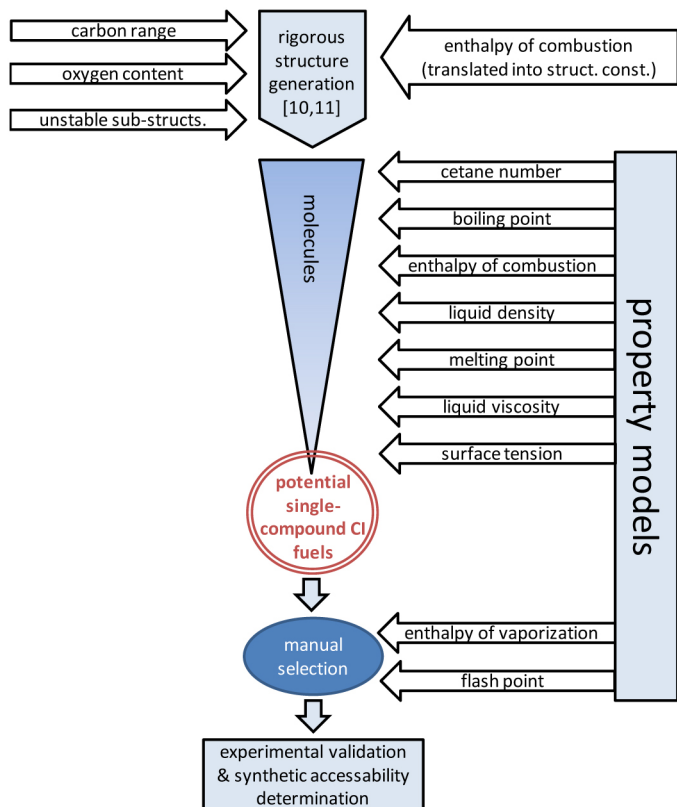


Figure 2. TMFB fuel identification framework adapted for CI engines comprising rigorous structure generation and property-constrained reduction of a set of potential fuel molecules

FUEL PROPERTY MODELING

Since the presented fuel identification process is initiated by an algorithmic generation of a large number of potential fuel structures, fast property prediction techniques are required, which do not need to be of the highest possible accuracy in a first step. Sophisticated, high-level property models (e.g., quantum mechanics or detailed kinetic mechanisms) are infeasible for screening large amounts of molecules and their use is therefore restricted to the most promising structures only. Instead, Quantitative-Structure-Property-Relationships (QSPRs) [13] can link molecular structure to specific fuel properties to facilitate the rapid screening of molecular structures. QSPR methods employ molecular descriptors [14], which are calculated from the molecules' two- or three-dimensional structure and might offer more information for model building than frequently used group contribution methods [15]. However, the huge

number of descriptors derived by today's software packages [12,14] must be taken with caution as multi-collinearity may be present and the number of measured fuel compounds is often small. Therefore, an adequate choice of data pretreatment methods and model-building techniques is required and is addressed in detail in the present contribution.

Internal and external validation techniques avoid overfitting of a model to the training data and thus are considered essential in building truly predictive structure-property models [16]. Unfortunately, application of state-of-the-art validation techniques is still lacking in many QSPR studies [17]. In order to ensure transferability of derived structure-property-relationships from training data to novel, computer-generated compounds, similarity assessment to training data is crucial to prevent risky extrapolation. To this end, the concept of a model's applicability domain [18,19] has been established.

QSPR modeling is commonly applied to thermophysical properties. Consequently, there are several studies in literature on some of the properties considered here [13,20,21,22,23]. However, most of these models were derived for small numbers of compounds belonging to specific structural classes. Moreover, the range of model applicability is not explicitly stated in most QSPR studies. The cetane number represents a property largely controlled by complex combustion chemistry. Even in case of a single-compound fuel like n-heptane, several hundreds of species interact in several thousands of elementary reactions [24]. Despite this complexity on the level of chemical kinetics, QSPR approaches have been applied to cetane number modeling recently [25,26]. In contrast to these studies, the present contribution not only involves data pretreatment for outlier exclusion and the concept of a model's domain of applicability, but also adds ether structures to the modeling dataset since some ethers exhibit promising effects on cetane number. Such extension of the training database affects property predictions for oxygenated compounds which are derived from a biomass feedstock and are associated with great potential in reduction of polycyclic aromatic hydrocarbon (PAH) emissions [8].

Deriving and applying predictive QSPRs in the present contribution is accomplished by the procedure outlined in Figure 3. The process of establishing QSPR models is discussed comprehensively, because of the aforementioned shortcomings in a number of QSPR studies. The same modeling approach is applied to all of the nine investigated fuel properties and can readily be extended towards different properties.

Initially, a comprehensive property database is collected and structural information is obtained for the investigated molecules. To this end, a set of molecular descriptors is calculated from the two-dimensional molecular structure. The resulting structure-property database is inspected and pretreated to obtain the final modeling dataset. Subsequently, this dataset is split into a training dataset used for model building and a validation dataset utilized to check whether a

model is able to predict unknown compounds. Finally, a validated QSPR model is employed in predicting property values for computer-generated compounds within the model's domain of applicability.

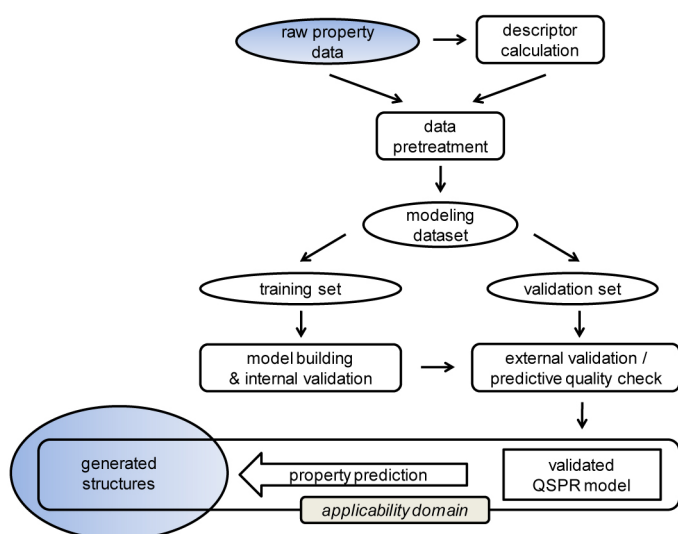


Figure 3. Workflow employed to build and apply predictive QSPR models for engine-relevant fuel properties

Data Pretreatment

In QSPR modeling, structural outliers are of major concern and should be addressed carefully. Structural outliers differ strongly from the remaining portion of the dataset in terms of their molecular descriptors. The basic assumption of QSPR modeling is that structurally similar objects behave in a similar way, i.e., they exhibit similar properties [16]. As a consequence, if a structural outlier is assigned to the training data set, it might significantly alter the general model performance at its core region of applicability in the molecular space, because the model has to account for a compound differing strongly from the majority of molecules considered. On the other hand, if a structural outlier is assigned to the external validation set, the predictive capability of the general model might be underestimated as the model cannot be assumed to extrapolate correctly beyond its region of applicability. Thus, structural outliers are removed from the modeling dataset prior to model building by means of one of the several detection techniques available [16]. For instance, the nearest neighbor for all compounds in the molecular space is determined to find isolated compounds. Yet, a crucial question is how to define the molecular space and the distance measure applied. Many descriptors cover substantially different orders of magnitude requiring range- or auto-scaling [16] to equalize the influence of all descriptors. However, multi-collinearity might be present among descriptors, i.e., descriptors may be linear combinations of other descriptors. Thus, principal component analysis (PCA) [27] is performed in the current contribution

to extract a set of latent variables from the high-dimensional descriptor data. These principal components represent a new, orthogonal coordinate system reflecting truly independent variation in the data. Similar to [28], for the present study, the first three principal components are considered for manual inspection of a compound's location in the set of structural data.

Model Building

QSPRs belong to the class of empirical modeling techniques and are closely related to response surface modeling (RSM) in statistical experimental design [29]. The basic principle is straightforward: The property under investigation (y) is described by a linear combination of the descriptor variables D_i which are derived from molecular structure:

$$y = \alpha_1 D_1 + \alpha_2 D_2 + \alpha_3 D_3 + \dots \quad (1)$$

The task of QSPR model building constitutes (i) finding a suitable subset of descriptors (D_i) and (ii) estimating their respective weights α_i from given experimental property data y^* . To this end, multiple linear regression (MLR) algorithms have been implemented [13] in combination with statistical hypothesis testing to decide whether an additional descriptor presents a significant contribution to the quality of fit [29]. These techniques try to achieve the strongest reduction in the residual ($\|y - y^*\|_2$) by a minimum number of descriptors and also take care of experimental measurement error. For instance, in the stepwise regression algorithm SROV [30], inclusion of further descriptors is stopped immediately after the residual falls below the noise level defined by experimental uncertainty. The theory behind multiple/stepwise linear regression allows for confidence interval prediction [29]. However, the MLR framework requires the number of observations (measured compounds) to be a multiple of the number of variables (molecular descriptors) available for selection and assumes independent decision variables [27]. Today, several software packages offer convenient calculation of hundreds or even thousands of molecular descriptors from a two- or three-dimensional molecular structure [12,14]. Applying multiple/stepwise linear regression algorithms despite an unfavorable descriptors/observations-ratio bears the risk of experiencing chance correlations [31,32]. Instead of excluding descriptors (and maybe essential information), multivariate data analysis techniques should be preferred. Indeed, partial least squares (PLS) [33] analysis is suited in cases where the number of descriptors is several times larger than the number of observations. The capability of dealing with (multi-)collinearity in the data [27] is an important strength of PLS analysis rendering the technique a method of choice for the present modeling task. However, those advantages attributed to projections computed by PLS come at a cost: A (nearly) perfect correlation between very few or even one descriptor and the response variable might be overlooked by

the PLS algorithm in the presence of a huge amount of descriptor data containing no useful information [31,34]. The practical relevance of these findings, however, is considered less important since a true, unique correlation between one descriptor and a property is not expected to emerge in a real QSPR study and large descriptor sets are usually interrelated [31].

Model Validation

Adding descriptors to an existing linear combination doesn't improve the model in cases where the residual ($\|y - y^*\|_2$) is already located below the noise level. Instead, the model will fit the measurement error in the training data and therefore lose predictive quality. However, in model building, additional uncertainty concerns the use of a linear relationship between descriptor-variables and the investigated property, since the degree of accuracy of this assumption cannot be known in advance. Consequently, adding descriptors to the model might have to stop even before the residual reaches the noise level. To address both problems, uncertainty due to measurement error and due to the model's inappropriateness, validation techniques are employed. There exist two classes of validation methods, namely internal and external validation [17]. Internal validation techniques like cross-validation [27,35] work on the data set, which is used for model development. In contrast, external validation requires a distinct set of compounds, which is used exclusively to assess the predictive capability of a model and thus does not influence model building [17]. In the present contribution, a combination of internal and external validation is employed. Thus, a fraction of approximately 10% of the compounds is removed from the training dataset and assigned to an external validation dataset (cf. Figure 3). Both sets should span approximately the same portion of the molecular space. Furthermore, they should represent a diverse sampling of this space, e.g., compounds should be well distributed to avoid enclosed, but empty spheres [16]. To this end, stochastic cluster analysis (SCA) [36] is conducted for diversity-based partitioning of the modeling dataset. In the present contribution, SCA is applied to principal components extracted from the high-dimensional descriptor space to account for (multi-)collinearity among original descriptors.

Applicability Domain

There is a strong need for determining whether a novel compound is located inside a model's applicability domain [18,19]. Even in face of a comprehensive collection of training data, a model should never be assumed to cover the whole molecular space. For instance, data for critical temperature has been extracted from the DIPPR database [37] at two different error levels (<5% and <10%). Figure 4 illustrates that the two datasets differ not only in the number of compounds. The larger dataset also includes significantly different molecular structures. These provide additional

information allowing for the construction of models, which cover larger portions of the molecular space.

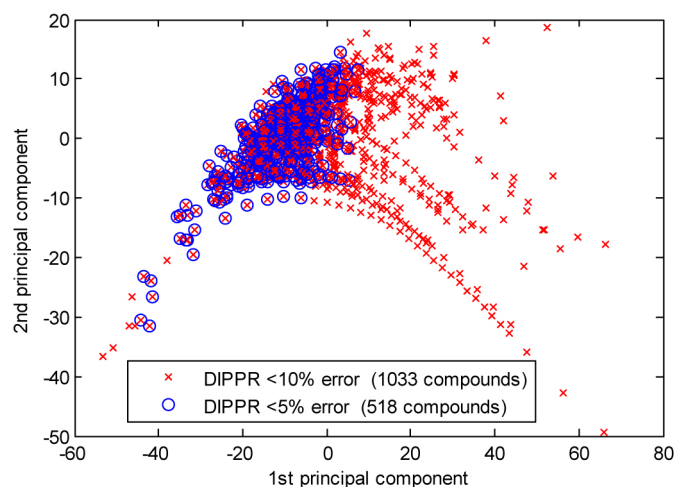


Figure 4. Critical temperature data extracted from DIPPR database [37]: principal component illustration of all compounds comprising <5% and <10% measurement error, respectively

In PLS analysis, the DModX criterion [27] is commonly employed to assess whether new data belongs to the same class of data used for training. In the present contribution, a novel compound is characterized by a descriptor-vector, which contains information calculated solely from the structure of the molecule. Thus, DModX is employed to determine, whether the structural data of a novel compound is similar to the structural data used for model building. Values above a defined critical value derived from the F-distribution are considered moderate outliers [27,38]. Strong outliers (two times above the critical value) are associated with severe prediction error [38].

A second measure of similarity [18] can be used to estimate the model's domain of applicability. Here, similarity to training compounds is assessed based on Euclidean distances to training compounds in the so-called X-scores. Since the X-scores represent latent variables being extracted from the descriptor-data in order to model the response (y) [27], they are interpreted as specific molecular characteristics being important to describe the property. Consequently, similarity is measured in this low-dimensional space. Even after removal of structural outliers, few compounds may be located relatively isolated. Thus, instead of the distance to the closest neighbor, the averaged distance to the three nearest neighbors in the training data is calculated for each compound. A threshold value for outlier detection is derived from the largest distance observed on the training data. Strong outliers identified by either DModX or Euclidean distances in X-scores are excluded from further consideration in the automatic fuel identification algorithm (cf. Figure 2).

Applicability domain concepts have demonstrated strong capability in classifying the reliability of predictions for a set of compounds [18]. Usually, a compound's location relative