Predicting octane numbers relying on principal component analysis and artificial neural network

S. Tipler,[*,†,‡,¶] G. D'Alessio,[†,§,¶] Q. Van Haute,[‖] A. Parente,[†,¶] F. Contino,[‡,¶] and A. Coussement[†,¶]

†*Université Libre de Bruxelles, Ecole Polytechnique de Bruxelles, Aero-Thermo-Mechanics Laboratory, Bruxelles, Belgium*

‡*Vrije Universiteit Brussel, Department of Mechanical Engineering, Bruxelles, Belgium*

¶*Université Libre de Bruxelles and Vrije Universiteit Brussel, Combustion and Robust Optimization Group (BURN), Bruxelles, Belgium*

§*CRECK Modeling Lab, Department of Chemistry, Materials and Chemical Engineering, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy*

‖*Comet Traitements, Rivage de Boubier, 25, B-6200 Châtelet, Belgium*

E-mail: steven.tipler@ulb.be

Abstract

Measuring the Research Octane Number (RON) and the Motor Octane Number (MON) at a low price is currently not feasible, thus making the use of predictive methods essential to accomplish this task. Nevertheless, the latter rely on expensive data and linear by volume models cannot be applied for complex fuels. In this work, we have investigated 41 parameters from inexpensive tests to find the inherent link between these fuel properties and the RON and the MON. To achieve this objective, we

first reduced the number of properties to only consider the principal ones relying on principal component analysis (PCA). Then, we applied artificial neural network (ANN) to identify the underlying links between the properties and the RON/MON. The measurement of the distillation curve, the atomic mass fraction and the specific gravity are the primary properties required for the current method. The achieved mean squared error (MSE) is equal to 0.7 [ON^2].

Introduction

Global warming is one of the most central issues of our century. It is a consequence of emissions from human activities, which depend mainly on our energy mix. To control emissions adequate energy policies must be implemented to limit global warming. To reach this objective a scenario and an ecological aim must be defined. One such energy policy is the Paris agreement which is an universal agreement made during the 21st conference of the parties (COP). Its purpose is to limit the global temperature from increasing below 2°C in 2100 compared with the pre-industrial levels. The world energy outlook (WEO) provides scenarios that assess the consequences of different energy policies.

According to the different scenarios, one important aspect to achieve sustainable development is the usage of alternative fuels. They could be energy carriers and electrofuels,[1,2] biofuels such as biodiesel,[3] fuels derived from wastes such as automotive shredder residues[4,5] and other types of molecules such as alcohols.[6]

The number of sources of alcohol molecules increases with the development of fermentation and gasification. These processes promoted the production of large and unconventional alcohol molecules such as propanols, butanols and pentanols.[6] Studying these molecules is important as they are produced from renewable sources. Biofuels have obtained growing interest as alternatives or supplements in conventional fuels. Biofuels are produced from biological raw materials, and mostly environment-friendly.[7] Additionally, these molecules have been proposed as blend components to reduce the petroleum consumption and the

greenhouse gas emissions.[6]

RON and MON are usually employed to characterize a mixture consisting of the aforementioned alcohol molecules and gasoline fuel, assessing the ignition delay of gasoline fuels under specific engine conditions.[8,9] Although the effectiveness of the two methods is not questioned here, they both result expensive (due to the costs of the products to be employed, the high level of qualification required and the cost of the engine to be employed) and they entail a very large amount of time to perform the analysis. Being able to characterize the octane numbers is capital. A too small octane number is characterized by a small autoignition delay time which creates undesirable end-gas auto-ignition. This autoignition leads to pressure waves that dramatically damages the piston. This phenomena is also called knock. Additionally, as far as alternative fuels are concerned, the fuel production plant is owned by a local company, so small-scale production plants play a role in the production process. To achieve profitability, small-scale production plants could require a high energy yield. Thus, being able to adapt the engine parameters according to the auto-ignition properties of the fuel (the octane numbers) is also a central challenge. As the fuel properties depend on the feedstock, they could vary over time which requires an instantaneous characterization of the fuel, which is given by predicting methods.

A valuable alternative to measurement methods is represented by linear by volume predicting models, as they are able to predict the octane numbers of simple fuels whose composition is known. In these methods, the octane number of the fuel is given by the linear combination of the octane numbers of each molecule weighted by the volume fraction of each molecule.[10] Nevertheless, when the composition cannot be precisely estimated due to its complexity or because of the cost of the measurement method, having a model based on properties is appropriate. Bayesian methods can accurately predict the RON and the MON of gasoline blendstocks mixed with an oxygenated molecule.[11] This class of methods has the advantage of being very effective and it can reproduce experimental data very well. On the other hand, the limitation of Bayesian methods is represented by their input quantities,

which are expensive. In fact, they completely rely on the saturates, olefins, aromatics and oxygenates (SOAOx) hydrocarbon class fractions. The latter are all expensive properties to be computed, especially if compared to the distillation curve or the specific gravity (SG).

Other predicting methods exist. Researchers have already applied ANN to petroleum products. Albahri[12] predicted the octane number of pure molecules based on their chemical composition. Nevertheless, this method requires the knowledge of the whole fuel composition. Pasadakis[13] applied ANNs to predict the RON of gasoline blends. To do so, the input data were representative for the volumetric concentration of the following refinery streams: streams from fluidized catalytic cracking, reforming, isomerization, alkylation, dimersol, butane and methyl tert-butyl ether (MTBE). This method performs well, nevertheless it is only applicable for the mentioned streams and not for gasoline mixed with large molecules. Similar work was performed by Doicin et al.[14] to predict the octane numbers of petroleum mixtures. Nevertheless, once again, this methodology does not account for large oxygenate molecules. Additionally, it does not identify broadly the input properties. More recently, Ibrahim et al.[15] relied on ANNS and PCA to predict the octane numbers based on several chemical features obtained from infrared spectroscopy. This methodology does not take large oxygenate molecules into account and rely on infrared spectroscopy, which is a method more expensive and complex that simple measurements of the specific gravity or the distillation curve. Abdul Jameel et al.[16] also predicted the octane numbers based on a chemical analysis, relying for instance on paraffinic CH3 groups, paraffinic CH2 groups, paraffinic CH groups, olefinic -CH=CH2 groups, naphthenic CH-CH2 groups, aromatic C-CH groups, and ethanolic OH groups. The methodology is again expensive and complex comparing to easier methods that measure the fuel properties and has not been developed for large oxygenates. Kubic et al.[17] estimated the octane numbers of hydrocarbons and oxygenated compounds with an ANN relying on the chemical composition with a group contribution method. This methodology requires the knowledge of the whole fuel composition, which is not always available. Vom Lehn et al.[18] also relied on a group contribution method to predict the ONs. This

methodology is based on the knowledge on the molecule structure and only estimates the octane numbers of pure molecules, and not fuel blends.

In the aforementioned studies from the literature, ANNs generally rely on chemical properties, such as volume fraction and group contributions, rather than physical properties which tends to be easier to measure. Moreover, no prediction models based on inexpensive properties exist for gasoline blendstock mixed with oxygenated molecules. Finally, the methods from the literature predict the octane number (ON) of pure molecules and of blends of several petroleum fractions, but, no method investigates single petroleum fractions which are not blended.

In the current paper, we focus on developing a new method to predict the octane number, which is based on input quantities which are easy to measure. We initially take into account as many properties as possible, to then identify and select only the ones which are well correlated with the ONs and which consequently have a key role in the prediction. Yet, we select a large number of properties (41 properties), with 32 out of 41 are not related. Consequently, it is mandatory to rely on a method that enables us to find any underlying relation between the octane numbers and the input properties. To this end, we rely on ANN to find the inherent link (if existing) between these properties and the octane numbers of a sample of gasoline blendstocks mixed with an oxygenated molecule. In light of the high number of properties which are taken into account in this work to predict the ONs, which is equal to 41, it is first appropriate to select a subset of properties that are composed of the ones that share the majority of the variance of all the original properties. In fact, by means of this operation, it is possible to remove the properties carrying a small amount of information, as well as to clean the data by removing the noise. To do that, we have applied the PCA.

Comparing with the current literature, the novelty of the proposed methodology comes from the usage of properties rather that relying on the fuel composition. This enables the user of the method to rely on test methods that are inexpensive. The usage of PCA is a way

to select the principal properties from a large number rather than only using a small number of properties. This allow the usage of a very large number of candidate input properties even some that were unstudied until now. Coupling PCA with ANN is a powerful way to find the inherent link between the principal input properties and the octane numbers. Thanks to the develop methodology, the octane numbers can be calculated even if the fuel composition is unknown. Finally, most of the existing models were developed either for a small mixture of molecules or for mixtures of different types of fuel blends.

After describing the method to create the sample of fuels, we list the properties that were investigated. Finally, we detail the PCA and the ANN methods and we conclude with the precision of the developed method.

Composition of the investigated fuels

The investigated fuels are gasoline blendstock mixed with a single oxygenated molecule. Different molecules were used: 1-propanol, 2-propanol, 1-butanol, 2-butanol and 2-methyl-1-propanol. Such alcohol with a high number of carbon atoms can be produced via fermentation or gasification. This type of fuel was investigated by Christensen et al.[6] The investigated fuels were simulated, in a sense that their compositions were chosen numerically and their properties were calculated with Aspen Plus and with the available methods from the literature. Simulating the studied fuels enabled us to have a large number of points to study the data thanks to PCA, to benefit of a large number of points to carry the ANN regression and to have an homogeneous training set. The alternative would have been to experimentally study those fuels. It would have led to more accurate results but at an unbearable cost. Thereafter is explained how the simulated fuels were created.

First, molecules representative of gasoline were selected from the Aspen Plus database. 238 molecules were selected to represent the gasoline blendstock: 9 n-paraffins, 92 iso-paraffins, 83 olefins, 28 naphthenes, 26 aromatics.

Second, the volume fractions of each molecule in each fuel were calculated. To proceed, the fuel was divided into three layers: the hydrocarbon class (n-paraffin, iso-paraffin, olefin, naphthene, aromatic, and oxygenate (PIONAOx)), the isomer group, and the carbon distribution in an isomer group. This subdivision enabled us to get an accurate definition and to set an homogeneous distribution for the hydrocarbon class and isomer fractions. Additionally, it enabled us to use a mathematical law from the literature to define the molecular distribution for each isomer group.

The main hydrocarbon class volume fractions (saturate, olefin, aromatic) were sampled following a latin hypercube sampling (LHS) - to optimize the space filling and obtain an homogeneous distribution - within the bounds reported in Table 1. These bounds correspond to the range of applicability of the method developed in a previous work to predict the octane numbers of gasoline blendstock mixed with an oxygenated molecule.[11] Once the saturate, olefin and oxygenate class fractions were sampled, the oxygenate volume fraction was calculated, being 100% subtracted from the sum of the other class fractions. Thereafter, the samples whose oxygenate volume fraction did not match with the requested bounds (Table 1) were removed to respect the range of applicability of the law that define the target ON. The following approach allowed us to ensure a good distribution of the volume fractions with a small number of samples. We first evaluated by convergence the mean and the standard deviation of the six volume fractions (PIONAOx). From this study, we estimated the minimal number of samples to reach convergence. Finally, $N_{\text{class}}^{\text{Sampling}} = 25$ samples were generated in a while loop until reaching the predefined mean and standard deviation.

Table 1: Limits of the hydrocarbon class fractions of the simulated fuels.[6] For each sample fuel, the oxygenate class is composed of a single molecule among 1-propanol, 2-propanol, 1-butanol, 2-butanol and 2-methyl-propanol.

	S	O	A	Ox
Min	57.3	2.1	16.3	2.9
Max	74.0	7.8	28.9	15.3

Different isomer groups differentiate the molecules with the same molecular weight, but

with a different number of methyl substituent. In each hydrocarbon class, a distribution factor defines the percentage of each isomer group. After a convergence study similar to the one realized with the hydrocarbon class layer, $N_{molecule}^{Sampling} = 17$ samples were chosen.

A Gamma distribution drove the molecular weight in each hydrocarbon class, following the method presented by Riazi et al.[19] The distribution is given by the function:

$$f(x, \alpha, \beta, \eta) = \frac{(x - \eta)^{\alpha-1} e^{-\frac{x-\eta}{\beta}}}{\beta^{\alpha}\Gamma(\alpha)}. \quad (1)$$

Table 2: investigated shape and intensity parameters of the gamma function. η is the minimum molecular weight, M_{mean} is the investigated mean molecular weight range and α defines the probability density function shape.

Group	η	α		M_{mean}	
		Min	Max	Min	Max
n-Paraffin	58	1.5	20	85	115
iso-Paraffin	58	1.5	20	82	95
Olefin	70	2.8	10	80	116
Naphthene	70	2	10	80	95
Aromatic	78	2	20	113	128

This function requires four input parameters: x, is the molecular weight of a molecule in an hydrocarbon class, the three others, α, β and η, are design parameters that characterize the molecular distribution in each hydrocarbon class. The parameter β can be estimated with the following formula:[19]

$$\beta = \frac{M_{mean} - \eta}{\alpha}. \quad (2)$$

The parameter η is the molecular weight of the lightest molecule in the hydrocarbon class. β depends on the function shape, α, and on the mean molecular weight of the hydrocarbon class, M_{mean}. These two parameters were sampled. Their ranges were defined iteratively with the following procedure. Sets of 20 parameters were generated iteratively, giving raise to 20 fuels after each iteration. At each iteration corresponds different ranges of mean molecular weight, M_{mean}, and shape parameter, α. For each iteration, the distillation cut points of the

20 fuels were calculated. These cut points were compared with the experimental distillation cut points published by Christensen et al.[6] (Table 3) to obtain fuels for which the target ON can be calculated. The best correspondence gave the best range of parameters (Table 2). The final sample of fuels was generated with LHS within these ranges. As we did not study the co-variance of the parameters, part of the final fuels did not match with the experimental cut points. These fuels were removed, resulting in $N_{molecule}^{Sampling} = 14$ out of 20 fuels.

Table 3: Evaporation characteristics of the investigated gasoline blendstocks mixed with an oxygenated molecule.[6] TX refer to the distillation temperature (° C) to get X% evaporated, SL is the 10-90 slope (° C / %v) and MeABP is the mean average boiling point (° C) as defined by Riazi.[19]

	T5	T10	T30	T50	T70	T90	SL	MeABP
Min	35.85	41.1	58.6	69.9	91.8	163.3	1.21	79.4
Max	62.6	69.8	88.1	103.7	137.25	174.5	1.63	101.5

The Gamma distribution is continuous over the molecular weight, so, we discretized and integrated the Gamma function depending on the molecular weight of the molecules in the fuel. Moreover, it represents the probability density function of a molecule depending on its molecular weight. Thus, y_k, is given by:

$$y_k = \begin{cases} y_i \frac{r_k}{N_k} \int_{M_k}^{(M_k+M_k^{+1})/2} f(x,\alpha,\beta,\eta)dx \\ \qquad\qquad \text{if } M_k = M_k^{min} \\ y_i \frac{r_k}{N_k} \int_{(M_k+M_k^{-1})/2}^{(M_k+M_k^{+1})/2} f(x,\alpha,\beta,\eta)dx \\ \qquad\qquad \text{if } M_k \in]M_k^{min}; M_k^{max}[\\ y_i \frac{r_k}{N_k} \int_{(M_k+M_k^{-1})/2}^{M_k} f(x,\alpha,\beta,\eta)dx \\ \qquad\qquad \text{if } M_k = M_k^{max}. \end{cases} \quad (3)$$

M_k is the molecular weight of a molecule k. M_k^{-1} and M_k^{+1} are the molecular weights of molecules respectively lighter and heavier than the molecule k. M_k^{min} and M_k^{max} are the

minimal and maximal molecular weights in the hydrocarbon class of the molecule k. y_i is the hydrocarbon class volume fraction and r_k is the isomer distribution factor of the molecule k defined in the two previous sections. Several molecules have the same molecular weight in an isomer groups, such as 2-methyl-1-pentene and 4-methyl-1-pentene. Thus, the isomer factor was divided by N_k, which is the number of molecules that shares the same molecular weight with the molecule k in its isomer group.

The number of samples that were simulated is equal to $N_{class}^{Sampling} \times N_{molecule}^{Sampling} \times N_{molecule}^{Sampling} = 5950$. Thanks to the procedure described before, this high number of fuels defines a homogeneous training dataset generate a large training dataset for PCA and ANN.

Properties of the investigated fuels

The properties of the simulated fuels and their calculation methods were carefully chosen. Fuel properties have been characterized for many years. A lot of calculation methods thus exist. Among all the available references, we selected as many properties as possible in order to benefit of a large choice of properties to later identify which are the ones that contains a lot of information (maximal variance). We selected property methods from the *Characterization and Properties of Petroleum Fractions* by Riazi,[19] the API *Technical Data Book - Petroleum Refining*,[20] and the Peng-Robinson property package from Aspen HYSYS (HYSPR) especially developed for hydrocarbon systems.[21] We investigated the two octane numbers, and from the above references we were able to select 13 thermodynamic, 9 chemical and 7 transport petroleum properties. Some of these 29 properties were calculated with several methods, which results in 41 candidate inputs properties. Relying on several methods is a way to increase the number of candidate properties and dig into the highest number of properties as possible. It is noteworthy that the selected methods are based on a combination of simple properties such as the specific gravity, which measures the density. Relying on these simple properties is a way to perform the smaller number of measurements as possible and

estimate the other properties based on these simple properties. The selected properties are listed thereafter.

Thermodynamic properties

The distillation curve cut points of ASTM D86 at 5, 10, 30, 50, 70 and 90% are the first properties. They were calculated with the HYSYS Peng-Robinson EOS package from Aspen Plus. The mean average boiling point (MeABP) was also calculated and included in the study. The MeABP depends on the distillation cut points:

$$\mathrm{MeABP} = \mathrm{VABP} - \Delta \mathrm{T_{Me}} \tag{4}$$

with the volume average boiling point (VABP):

$$\mathrm{VABP} = \frac{\mathrm{T_{10}} + \mathrm{T_{30}} + \mathrm{T_{50}} + \mathrm{T_{70}} + \mathrm{T_{90}}}{5}, \tag{5}$$

$\mathrm{T_{XX}}$ is the temperature at which xx% of the fuel is evaporated and the correction temperature is defined by:[19]

$$\ln(\Delta \mathrm{T_{Me}}) = -1.53181 - 0.0128(\mathrm{VABP} - 273.15)^{0.6667} + 3.646064\mathrm{SL}^{0.333} \tag{6}$$

with SL is the 10–90 slope:

$$\mathrm{SL} = \frac{\mathrm{T_{90}} - \mathrm{T_{10}}}{80}. \tag{7}$$

The critical volume, pressure, temperature, density and compressibility factors were also considered. The following relations were adopted to calculate them from properties that are easily measurable.

The critical volume was calculated with the following formula:[19]

$$\mathrm{V_c} = 1.7842 \times 10^{-4} \mathrm{T_b^{2.3829}} \mathrm{SG^{-1.683}} \tag{8}$$

where V_c is in cm^3/mol and T_b is the MeABP in kelvin. This formula is accurate for lower molecular weights such as gasoline blends.

The critical pressure was calculated with two methods. The first is given by the Riazi-Daubert method:

$$\begin{aligned} \mathrm{Pc}= 3.1958 \times 10^5 [\exp(&-8.505 \times 10^{-3}\mathrm{Tb} - 4.8014\mathrm{SG} \\ &+5.749 \times 10^{-3}\mathrm{TbSG})]\mathrm{Tb}^{4.0844}\mathrm{SG}^{4.0846}, \end{aligned} \tag{9}$$

and the second method to calculate the critical pressure is known as the Cavett method:

$$\begin{aligned} \log(\mathrm{P_c}) = {} & 1.6675956 + (9.412011 \times 10^{-4})(1.8\mathrm{T_b} - 459.67) \\ & -(3.047475 \times 10^{-6})(1.8\mathrm{T_b} - 459.67)^2 \\ & -(2.087611 \times 10^{-5})(\mathrm{API})(1.8\mathrm{T_b} - 459.67) \\ & +(1.5184103 \times 10^{-9})(1.8\mathrm{T_b} - 459.67)^3 \\ & +(1.1047899 \times 10^{-8})(\mathrm{API})(1.8\mathrm{T_b} - 459.67)^2 \\ & -(4.8271599 \times 10^{-8})(\mathrm{API}^2)(1.8\mathrm{T_b} - 459.67) \\ & +(1.3949619 \times 10^{-10})(\mathrm{API}^2)(1.8\mathrm{T_b} - 459.67)^2. \end{aligned} \tag{10}$$

The critical temperature was calculated also with two methods. The first is the API method:[20]

$$\begin{aligned} \mathrm{T_c}= 10.6443 &[\exp(-5.1747 \times 10^{-4}\mathrm{T_b} - 0.54444\mathrm{S} + 3.5995 \times 10^{-4}\mathrm{T_b S})] \\ &\times \mathrm{T_b}^{0.81067}\mathrm{S}^{0.53691}, \end{aligned} \tag{11}$$

and the second is the Cavett method:[19]

$$\begin{aligned} \mathrm{T_c}= & 426.7062278 + (9.5187183\mathrm{x}10^{-1})(1.8\mathrm{Tb} - 459.67) \\ & -(6.01889 \times 10^{-4})(1.8\mathrm{T_b} - 459.67)^2 \\ & -(4.95625 \times 10^{-3})(\mathrm{API})(1.8\mathrm{Tb} - 459.67) \\ & +(2.160588 \times 10^{-7})(1.8\mathrm{Tb} - 459.67)^3 \\ & +(2.949718 \times 10^{-6})(\mathrm{API})(1.8\mathrm{T_b} - 459.67)^2 \\ & +(1.817311 \times 10^{-8})(\mathrm{API}^2)(1.8\mathrm{T_b} - 459.67)^2. \end{aligned} \tag{12}$$

The critical density was calculated with:

$$\mathrm{d_c} = \frac{\mathrm{M}}{\mathrm{V_c}} \tag{13}$$

where the critical volume is given by Equation 8. The molecular weight is given by three methods described in the following section dedicated to chemical properties.

The critical compressibility factor is given by the following relation:[19]

$$\mathrm{Z_c} = \frac{\mathrm{P_c V_c}}{\mathrm{R T_c}} \tag{14}$$

with the critical properties calculated with Equations 8, 9 and 11

The acentric factor given by the Lee-Kesler method[19] is also included in the study:

$$\omega = \frac{-\ln(\mathrm{P_c}/1.01325) - 5.92714 + 6.09648/\mathrm{T_{br}} + 1.28862\ln(\mathrm{T_{br}}) - 0.169347\mathrm{T_{br}^6}}{15.2518 - 15.6875/\mathrm{T_{br}} - 13.4721\ln(\mathrm{T_{br}}) + 0.43577\mathrm{T_{br}^6}}, \tag{15}$$

with

$$\mathrm{T_{br}} = \mathrm{T_b}/\mathrm{T_c}. \tag{16}$$

The last property of the current section is the Watson K factor:

$$\mathrm{K_w} = \frac{(1.8\mathrm{T_b})^{1/3}}{\mathrm{SG}} \tag{17}$$

Chemical properties

On top of these thermodynamic properties, chemical properties were included. The oxygen-to-carbon and the carbon-to-oxygen ratios, the carbon, hydrogen and oxygen weight ratios were obtained from Aspen Plus by counting the amounts of atoms.

The molecular weight was estimated with three methods. First, the Lee-Kelser method:[19]

$$\begin{aligned}\mathrm{M}= &-12272.6 + 9486.4\mathrm{SG} + (8.3741 - 5.9917\mathrm{SG})\mathrm{T_b} \\ &+(1 - 0.77084\mathrm{SG} - 0.02058\mathrm{SG}^2) \\ &\times(0.7465 - 222.466/\mathrm{Tb})10^7/\mathrm{Tb} \\ &+(1 - 0.80882\mathrm{SG} + 0.02226\mathrm{SG}^2) \\ &\times(0.3228 - 17.335/\mathrm{T_b})10^{12}/\mathrm{T_b}^3,\end{aligned} \tag{18}$$

second, the Riazi-Daubert method:[19]

$$\mathrm{M} = 1.6607 \times 10^{-4}\mathrm{T_b}^{2.1962}\mathrm{SG}^{-1.0164}, \tag{19}$$

third, the API method:[20]

$$\begin{aligned}\mathrm{M}= 20.486[&\exp(1.165 \times 10^{-4}\mathrm{T_b} - 7.78712\mathrm{SG} \\ &+1.1582 \times 10^{-3}\mathrm{T_bSG})]\mathrm{T_b}^{1.26007}\mathrm{SG}^{4.98308},\end{aligned} \tag{20}$$

Other chemical properties were estimated: the SG, the stoichiometric air-to-fuel ratio,

and the aniline point calculated with the API method:[20]

$$\mathrm{AP} = -1253.7 - 0.139\mathrm{MeABP} + 107.8\mathrm{K_w} + 868.7\mathrm{SG}, \tag{21}$$

and with the Albahri et al. method:[19]

$$\mathrm{AP} = -9805.269\mathrm{R_i} + 711.85761\mathrm{SG} + 9778.7069, \tag{22}$$

with $\mathrm{R_i}$ the refractivity intercept defined by:

$$\mathrm{R_i} = \mathrm{n} - \frac{\mathrm{d}}{2} \tag{23}$$

where d and n are respectively the density and the refractive index at 20°C and at 1 atm.

Transport properties

Finally, transport properties were included in the study.

The refractive index was calculated with:

$$\mathrm{n} = (\frac{1 + 2\mathrm{I}}{1 - \mathrm{I}})^{1/2} \tag{24}$$

where I was calculated with three methods. Firstly with the API method:[20]

$$\begin{aligned} \mathrm{I} =& 2.266 \times 10^{-2} \exp(3.905 \times 10^{-4}\mathrm{MeABP} + 2.468\mathrm{SG} \\ & -5.704 \times 10^{-4}\mathrm{MeABP\,SG})\mathrm{MeABP}^{0.0572}\mathrm{SG}^{-0.720}, \end{aligned} \tag{25}$$

secondly with the Riazi-Daubert method:[19]

$$\mathrm{I} = 0.3773\mathrm{T_b}^{-0.02269}\mathrm{SG}^{0.9182} \tag{26}$$

and last, by the method developed by Riazi and Daubert and included in the API-TDB method:[19]

$$\begin{aligned} \mathrm{I} =& 2.34348 \times 10^{-2}[\exp(7.029 \times 10^{-4}\mathrm{T_b} + 2.468\mathrm{SG} \\ & -1.0267 \times 10^{-3}\mathrm{T_b SG})]\mathrm{T_b}^{0.0572}\mathrm{SG}^{-0.720}. \end{aligned} \tag{27}$$

The refractivity intercept, previous defined by Equation 23 was also included.

The m parameter was also considered:

$$\mathrm{m} = \mathrm{M}(\mathrm{n} - 1.475). \tag{28}$$

The liquid and gas thermal conductivity were investigated. The liquid thermal conductivity at 20°C was calculated with the API method:[20]

$$\mathrm{k} = \mathrm{MeABP}^{0.2904} \times (9.961 \times 10^{-3} - 5.364 \times 10^{-6} \times \mathrm{T}) \tag{29}$$

and with another method developed by the API group:[19]

$$\mathrm{k} = {\mathrm{T_b}}^{0.2904} \times (2.551 \times 10^{-2} - 1.982 \times 10^{-5}\mathrm{T}) \tag{30}$$

where T and $\mathrm{T_b}$ are both in Kelvin.

The gas thermal conductivity was calculated with:[19]

$$\mathrm{k} = 1.7307\mathrm{E}(1.8\mathrm{T_b})^{\mathrm{B}}\mathrm{SG}^{\mathrm{C}} \tag{31}$$

with the constants A, B and C defined as:

$$\mathrm{A} = \exp(21.87 - 8.07986\mathrm{t} + 1.1298\mathrm{t}^2 - 0.05309\mathrm{t}^3), \tag{32}$$

$$\mathrm{B} = -4.13948 + 1.29924\mathrm{t} - 0.17813\mathrm{t}^2 + 0.00833\mathrm{t}^3, \tag{33}$$

$$C = 0.19876 - 0.0312t - 0.00567t^2, \tag{34}$$

and t is a variable such as

$$t = \frac{1.8T - 460}{100}. \tag{35}$$

k is given in W/m.K, and the temperatures T and T_b are in kelvin.

Finally the kinematic viscosity and the viscosity gravity function (VGF) were included in the study. The kinematic viscosity is given by the API method:[19]

$$\begin{aligned}\log(\nu_{38}) =& 4.9371 - 1.94733K_w + 0.12769{K_w}^2 \\ &+ 3.2629 \times 10^{-4}API^2 - 1.18246 \times 10^{-2}K_wAPI \\ &+ \frac{0.171617{K_w}^2 + 10.9943(API) + 9.50663 \times 10^{-2}(API)^2 - 0.860218Kw(API)}{API + 50.3642 - 4.78231K_w}\end{aligned} \tag{36}$$

where API is the API gravity.

The VGF was included, calculated with the following relation:

$$VGF = -1.816 + 3.484SG - 0.1156ln\nu_{38}, \tag{37}$$

with ν_{38} in mm^2/s.

Octane numbers

For what concern the octane numbers, the Bayesian pseudocomponent (PC) method was adopted. This method was developed in a previous work and allowed us to precisely estimate the octane number with an uncertainty lower than 2%.[11] This model was developed for fuels mixed with large alcohols studied by Christensen et al.,[6] so it cannot be used with ethanol, methanol, methyl tert-butyl ether (MTBE) and ethyl tert-butyl ether (ETBE).

The predicting law is given by the following formulation:

$$\mathrm{ON}^{*} = \left[\mathbf{y}\right]^{\intercal} \times \left(\left[\mathrm{E}\left(\mathbf{K}\right) \circ \mathbf{ON_{pc}} \right] \right) + \sigma^{*}, \tag{38}$$

with σ^{*} the random error, i.e. an unpredictable error due to the measurement method, "$\circ$" the Hadamard product and "$\times$" the Cartesian product.

y is the volume fraction of each hydrocarbon class (saturate, olefin, aromatic, oxygenate), $\mathrm{E}\left(\mathbf{K}\right)$ represents a correcting factor which corrects the initial PC method and $\mathrm{ON_{pc}}$ are the octane numbers of the PC. More details about these different terms and the calculation method is described in a previous work.[11]

Variable selection and structure of the ANN

In the previous section, we selected as many properties as possible to create the model. Nevertheless, some of these properties might be useless if they do not have a large variance when compared to the other properties. For this reason, we applied PCA to reduce the number of properties to be taken into account.

In addition to a dimensionality reduction via feature extraction, i.e., by means of a projection of a reduced basis of eigenvectors, PCA can also be used to perform feature selection, i.e., to select a subset of variables from the original number of properties. Thus, among the 41 input variables which were taken into account (i.e., the chemical properties), by means of PCA coherent subsets of variables were selected (the Principal Variables). Additional information regarding the PCA reduction via feature extraction and feature selection is available in literature.[22–29] Mathematically speaking, by mean of PCA the distribution of the data is analysed in order to define a new coordinate system where the new directions correspond to the maximal variance. In addition, in light of the multivariate nature of the input data considered in the current study (i.e., since the variables have different units and ranges), the operation of centering and scaling are necessary prior to the variable selection process.[27]

Centering is accomplished by subtracting the mean value, therefore each matrix row can

be seen as a fluctuation around the mean. Scaling, on the other hand, is carried out by dividing the centered observation by a factor $\mathbf{d_j}$:

$$\tilde{\mathbf{x}}_{\boldsymbol{j}} = \frac{(\mathbf{x}_{\boldsymbol{j}} - \bar{\boldsymbol{x}}_{\mathbf{j}})}{\mathbf{d}_{\boldsymbol{j}}}, \tag{39}$$

with $\mathbf{d_j}$ the scaling parameter. This scaling parameter depends on the scaling method (Auto scaling, Range scaling, Vast scaling): these scaling methods are accurately described in the literature and there are only briefly discussed here.[22]

The Auto scaling relies on the standard deviation as the scaling factor, the Range scaling employs instead the difference between the maximum and the minimum value for each variable as scaling factor. Finally, Vast scaling relies on the the product between the standard deviation and the so-called coefficient of variation, defined as the standard deviation divided by the mean value.

In the current study, we consider an input matrix $\mathbf{X}$ of Q variables (Q=41) and n observations (n=5950). By means of PCA, the input matrix $\mathbf{X}$ is projected on a basis of orthogonal eigenvectors, represented by the matrix $\mathbf{A}$ in (40), which is obtained by computing a spectral decomposition on the covariance matrix $\mathbf{S}$ associated to $\mathbf{X}$:

$$\mathbf{S} = \frac{1}{\mathrm{n}-1}\mathbf{X}^{\intercal}\mathbf{X} = \mathbf{A}\mathbf{L}\mathbf{A}^{\intercal}, \tag{40}$$

From the basis defined by the eigenvectors $\mathbf{A}$, a submatrix $\mathbf{A_q}$ can be extracted. This submatrix is the one whose columns are associated with the largest eigenvalues of $\mathbf{L}$, while the number of the selected eigenvalues is a value obtained by a convergence study of the explained variance. When the original data are projected on the reduced basis, $\mathbf{A_q}$, the scores matrix $\mathbf{Z_q}$ is obtained as follows:

$$\mathbf{Z_q} = \mathbf{X}\mathbf{A_q}. \tag{41}$$

An approximate reconstruction of the original dimensional sample can be obtained after-

wards by inverting Equation 41:

$$\mathbf{X_q} = \mathbf{Z_q}\mathbf{A_q}^{\intercal}. \tag{42}$$

To assess the amount of input data variance being explained by the original dataset, the eigenvalues matrix can be analyzed. In fact, the eigenvalues which are found on the diagonal of the matrix L are in descending order of magnitude. Thus, since they represent the amount of information carried by each PC, the first ones ($\lambda_{i=1\ldots q}$) are retained and the last ones ($\lambda_{j=q+1\ldots Q}$) can be discarded. Finally, the explained variance with q eigen vectors can be defined as:[30]

$$v_q = \frac{\sum_{i=1}^{q} \lambda_i}{\sum_{j=1}^{Q} \lambda_j} \tag{43}$$

To compare the three scaling methods, 8 eigenvectors were retained, as in this way more than the 99% of the original data variance was explained by the lower-dimensional basis. In addition, the number of variables was reduced to 10, 15 and 20 to compare the performance of the ANN with each of these number of variables. The Vast method is particularly interesting compared to the other methods because it allows to reduce the dimensionality even more, with only 4 eigenvectors. Vast scaling method will be investigated apart from the others to test the ANN with the lower dimensions as possible, i.e. 5 properties.

Two PCA-based variable selection methods were compared: the Procrustes and the B2 methods[24,25,28] which are summarized thereafter but more information can be found in Jolliffe et al., Krzanowski et al., and D'Alessio et al.[24,25,28]

The variable selection via B2 method is accomplished by means of a backward elimination. As the PCs are obtained as a linear combination of the original variable, and each variable is represented by a certain weight on a given principal component (PC), it is possible to delete the ones which are most correlated to the last eigenvectors (i.e, which have the highest weight). In fact, since the last eigenvectors are associated with the smallest eigenvalues, they carry a small amount of information, and are often associated with numerical noise

For the Procrustes method, let us consider $\mathbf{X}$ that is to be reduced to a smaller size. First each variable is iteratively deleted from this matrix, and p matrices $\tilde{\mathbf{X}}$ are thus obtained. PCA then is applied to project both matrices (the full and the reduced, respectively) on a lower dimensional manifold, and the two corresponding score matrices, $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$, are consequently obtained. A Procrustes analysis is applied to each of these matrices. This means that p coefficients M^2 are calculated to evaluate the differences between $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$. These coefficients are given by:

$$M^2 = \mathrm{Tr}(\mathbf{Z'Z} - \tilde{\mathbf{Z}}'\tilde{\mathbf{Z}} - 2\mathbf{\Sigma}) \tag{44}$$

where $\mathbf{\Sigma}$ is the matrix of the values obtained from the singular value decomposition of the square matrix $\tilde{\mathbf{Z}}'\mathbf{Z}$.

One additional aspect that was necessary to take into consideration in this study is represented by the ANN architecture, as no defined rule is available to set the number of neurons, as well as the number of layers, a priori. Thus, a convergence study a convergence study was realized to determine the size of the hidden layer, starting with 10 and with up to 300 artificial neurons. The final architecture consisted of one input layer, one hidden layer and one output layer, each being composed of a variable number of nodes, 42, 20, 15, 10 (up to 5 for the Vast method) for the input layer and from 10 to 300 for the hidden layer. The output layer is only consisted of two neurons, which are representative for the two ON. The software Matlab R2018a was used and over the 5950 samples available from the original input data, 70% of the latter were used to train the model, 15% were used for validation and 15% for testing.

Results

In this section, the different methods are compared. First, the variable selection methods are compared: Procrustes vs B2. Second, the scaling methods are compared: Auto vs Range vs

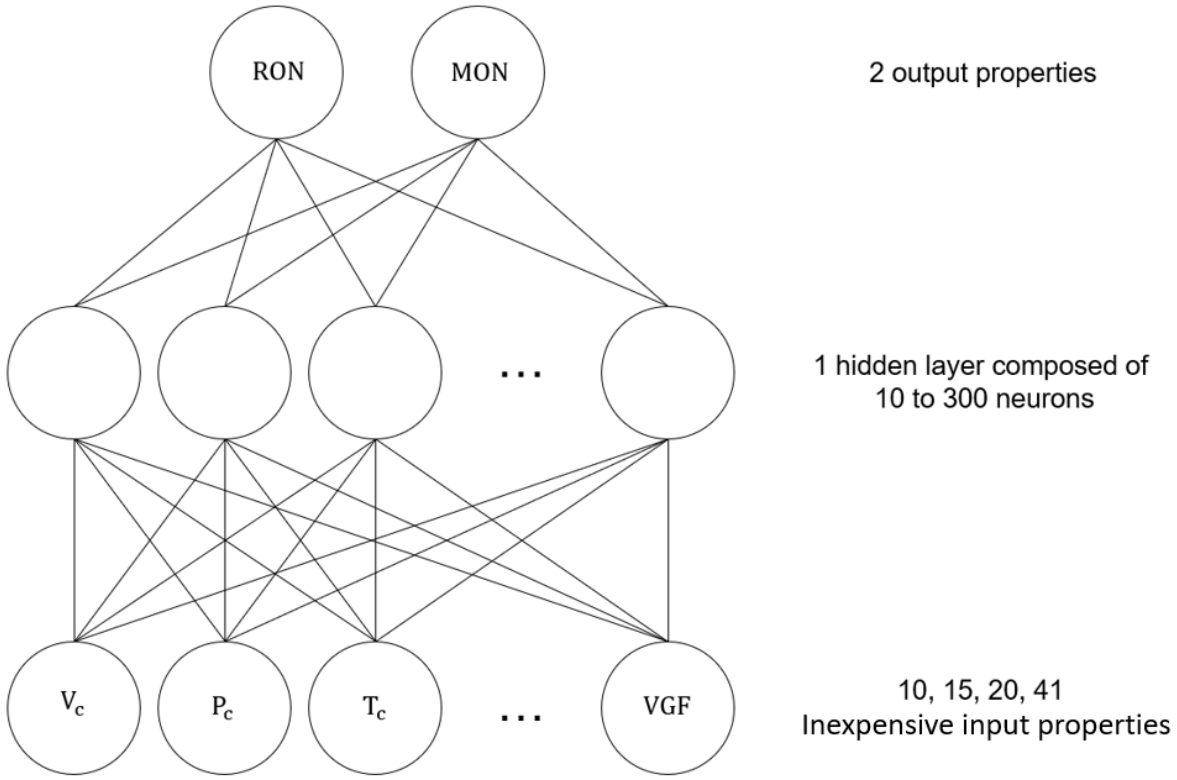

Figure 1: Scheme of the artificial neural network.

Vast (Figure 4). Third, the size of the hidden layer is discussed. Then, the selected variables are compared with the best methods. Finally, the MSE obtained with the Vast method and 4 eigen vectors is discussed.

First, the Procrustes and the B2 methods are compared by looking at the MSE obtained with a different size of the hidden layer. This MSE is obtained by computing the error between the true octane number and the estimated octane number. We can see on figures 2 and 3 that increasing the size of the hidden layer enables us to reduce the MSE because additional correlations between the input properties and ON can be found. A variable extraction has to be selected to realize this comparison. We select the Auto method for two reasons showed by the results below: 1- this method allowed us to reach a low MSE, 2- conversely with other methods, the MSE does not fluctuate a lot when the size of the hidden layer increases. The Procrustes and the B2 methods are compared on Figures 2 and

3. The B2 and the Procrustes methods give similar performance. Nevertheless, the selection via Procrustes analysis was considered to be more robust with respect to the B2 algorithm because of how the latter are defined. In fact, the selection with Procrustes is accomplished by means of a geometric comparison of the two scores matrices obtained from the full and the reduced matrix, respectively, while the selection via B2 is achieved via an elimination after the inspection of the weights distribution on the last (Q – m) PCs, with Q being the original dimensionality of the input matrix, and m being the total number of features to be selected. Thus, the studies in the current section were realized relying on the Procrustes method. As showed on Figures 2, 3 and 4, the convergence is reached with at least 100 neurons in the hidden layers. Nevertheless, a high variation is still present with up to 180. Thus, 180 neurons is appropriate.

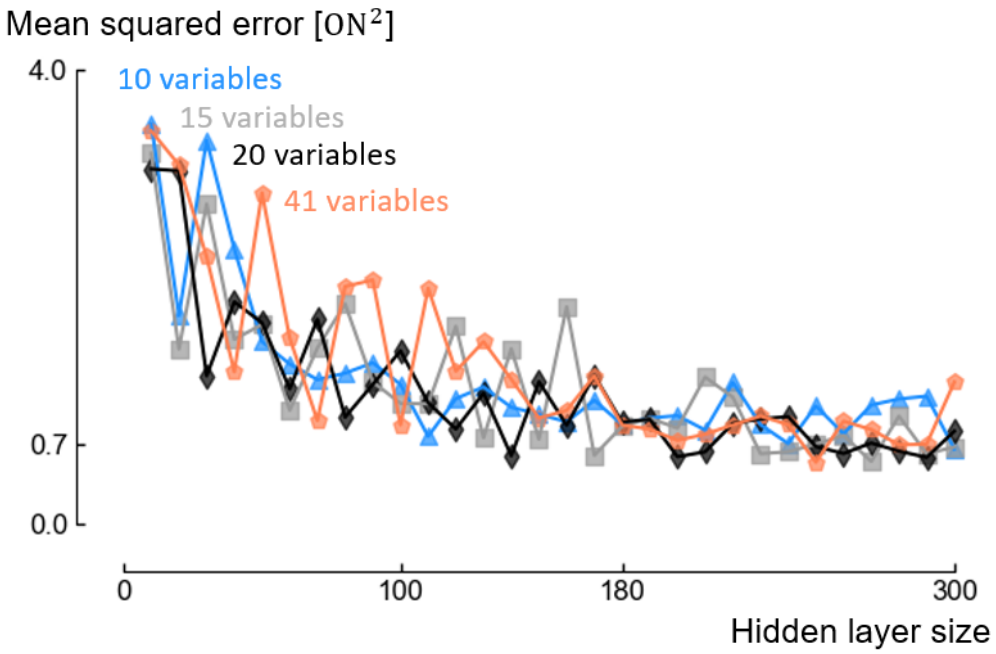

Figure 2: MSE with the Procrustes and the Auto methods. The MSE decreases with the size of the hidden layer. Moreover, if the number of neurons is higher than 180, the MSE is not impacted when the number of selected variables decreases. Therefore, the Procrustes selection method is appropriate.

Second, the scaling methods are compared while the Procrustes method is active as it provides the best results. As the variable selection method was previously selected, this

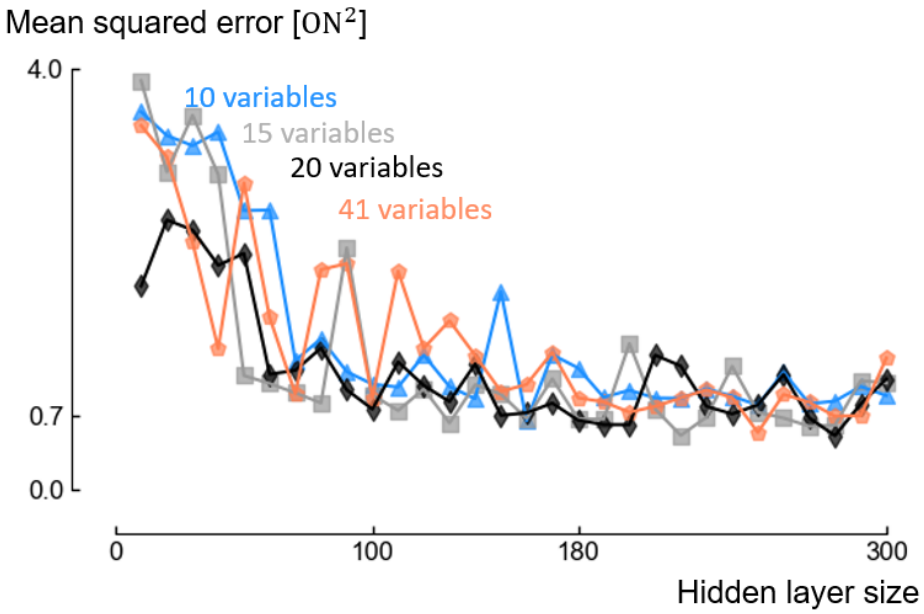

Figure 3: MSE with the B2 and the Auto methods. The MSE decreases with the size of the hidden layer. Moreover, the MSE increases when the number of selected variable decreases. Therefore, the variable selection method B2 is not appropriate.

comparison is realized with the lower number of variables, i.e. 10. The three methods are similar in terms of achieved mean squared error. It is noteworthy that the mean squared error fluctuates with the Vast method. Therefore, the Auto and the Range method are the better to predict the ON.

Finally, we compare the selected properties with the best set of methods: Procrustes with the Auto method, and Procrustes with the Range method. The properties shared by the two sets of properties are important. They are the following ones: the temperature at which 10% of the fuel is evaporated, the oxygen weight fraction, the Watson K factor. The carbon ratio is also an important properties as it appears in the CH weight fraction and in the OC weight fraction.

The correlations between the 10 principal properties obtained with the two best methods are discussed thanks to the analysis of the covariance matrix of the scaled variables. These covariance matrices are represented in Figures 5 and 6. From these figures, we can see that some of the properties listed in Table 4 are correlated within the methods. It shows that

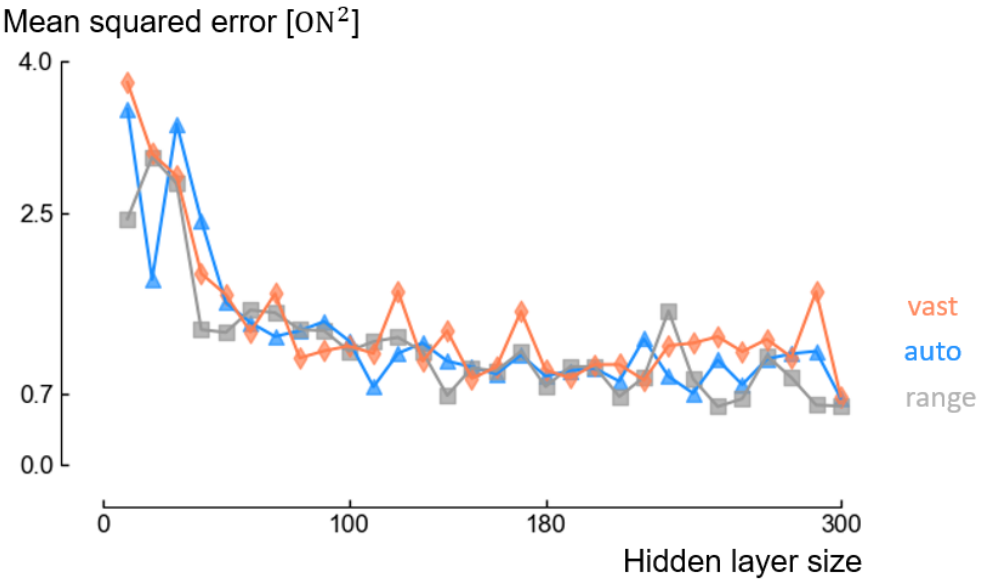

Figure 4: Comparison of the scaling methods.

Table 4: 10 principal properties with the combination of the best methods.

Variable selection	Procrustes	Procrustes
Scaling	Auto	Range
Property	T10%	T10%
	O%	T70%
	n (Eq. 24 and 26)	H%
	k (Eq. 30)	O%
	CH%	k (Eq. 31)
	MW (Eq. 20)	OC%
	SG	AP (Eq. 21)
	K_w	K_w
	P_c (Eq. 9)	T_c (Eq. 12)
	Z_c	d_c (Eq. 13 and 20)

similar information are carried by some properties which are not shared between the method.

This is for instance the case for the following properties when the scaling is done with the auto method:

- the CH ratio (auto method) with the hydrogen fraction (range method).
- the oxygen fraction (auto method) with the OC ratio (range method).
- the refractive index (auto method) with d_c (range method).

- k (auto method) with T_c (range method).

- Z_c (auto method) with T_c (range method).

	T10%	T70%	H%	O%	n (Eq. 149 and 151)	k (Eq. 155)	k (Eq. 156)	OC%	CH%	MW (Eq. 145)	SG	AP (Eq. 146)	Kw	Pc (Eq. 134)	Tc (Eq. 137)	dc (Eq. 138 and 145)	Zc
T10%	1.00	0.42	-0.27	0.10	0.52	0.85	-0.77	0.09	-0.21	0.85	0.54	0.12	-0.18	-0.58	0.84	0.19	-0.83
T70%	0.42	1.00	-0.38	-0.16	0.47	0.78	-0.69	-0.16	-0.42	0.78	0.49	0.12	-0.15	-0.54	0.77	0.16	-0.76
H%	-0.27	-0.38	1.00	-0.06	-0.86	-0.37	0.26	-0.05	0.91	-0.30	-0.86	0.74	0.86	-0.33	-0.51	-0.87	0.17
O%	0.10	-0.16	-0.06	1.00	0.17	0.02	0.02	1.00	0.37	0.00	0.17	-0.19	-0.20	0.13	0.06	0.20	0.02
n (Eq. 149 and 151)	0.52	0.47	-0.86	0.17	1.00	0.60	-0.47	0.17	-0.73	0.54	1.00	-0.70	-0.90	0.16	0.74	0.91	-0.41
k (Eq. 155)	0.85	0.78	-0.37	0.02	0.60	1.00	-0.90	0.02	-0.33	1.00	0.63	0.15	-0.20	-0.69	0.98	0.22	-0.97
k (Eq. 156)	-0.77	-0.69	0.26	0.02	-0.47	-0.90	1.00	0.02	0.25	-0.89	-0.49	-0.21	0.10	0.68	-0.86	-0.11	0.88
OC%	0.09	-0.16	-0.05	1.00	0.17	0.02	0.02	1.00	0.37	0.00	0.16	-0.19	-0.19	0.13	0.06	0.19	0.02
CH%	-0.21	-0.42	0.91	0.37	-0.73	-0.33	0.25	0.37	1.00	-0.28	-0.73	0.61	0.72	-0.25	-0.45	-0.72	0.17
MW (Eq. 145)	0.85	0.78	-0.30	0.00	0.54	1.00	-0.89	0.00	-0.28	1.00	0.56	0.23	-0.12	-0.74	0.96	0.14	-0.99
SG	0.54	0.49	-0.86	0.17	1.00	0.63	-0.49	0.16	-0.73	0.56	1.00	-0.68	-0.89	0.14	0.76	0.90	-0.43
AP (Eq. 146)	0.12	0.12	0.74	-0.19	-0.70	-0.15	0.21	-0.19	0.61	0.23	-0.68	1.00	0.94	-0.82	-0.04	-0.93	-0.37
Kw	-0.18	-0.15	0.86	-0.20	-0.90	-0.20	0.10	-0.19	0.72	-0.12	-0.89	0.94	1.00	-0.57	-0.38	-1.00	-0.03
Pc (Eq. 134)	-0.58	-0.54	-0.33	0.13	0.16	-0.69	0.68	0.13	-0.25	-0.74	0.14	-0.82	-0.57	1.00	-0.54	0.56	0.83
Tc (Eq. 137)	0.84	0.77	-0.51	0.06	0.74	0.98	-0.86	0.06	-0.45	0.96	0.76	-0.04	-0.38	-0.54	1.00	0.40	-0.91
dc (Eq. 138 and 145)	0.19	0.16	-0.87	0.20	0.91	0.22	-0.11	0.19	-0.72	0.14	0.90	-0.93	-1.00	0.56	0.40	1.00	0.01
Zc	-0.83	-0.76	0.17	0.02	-0.41	-0.97	0.88	0.02	0.17	-0.99	-0.43	-0.37	-0.03	0.83	-0.91	0.01	1.00

Figure 5: Covariance matrices obtained with all the scaled variables (with the auto method). Only the properties listed from Table 1 are reported. Green refer to high positive correlation (higher than the threshold 0.90. Red refer to high negative correlation (lower than the threshold -0.9), yellow refer to correlations equal to 1 or -1.

A similar study was done for the range method. The higher correlations are find between:

- the hydrogen fraction (range method) with the CH ratio (auto method), the aniline point (range method), the Watson factor (auto and range methods) and d_c (range method).

- the oxygen fraction (auto method) with the OC ratio (range method).

- the aniline point (range method) with the Watson factor (auto and range methods) and d_c (range method).

The 10 principal properties listed in Table 4 are correlated between each of the scaling method. This shows that the octane numbers are correlated with similar properties.

The Vast method is discussed alone as it is the only scaling method that enables to explain 99% of the variance with only 4 eigen vectors. Unfortunately, the Vast method has a low prediction capability when the number of properties is decreased up to the maximum, i.e. 5 properties (Figure 7).

	T10%	T70%	H%	O%	n (Eq. 149 and 151)	k (Eq. 155)	k (Eq. 156)	OC%	CH%	MW (Eq. 145)	SG	AP (Eq. 146)	Kw	Pc (Eq. 134)	Tc (Eq. 137)	dc (Eq. 138 and 145)	Zc
T10%	0.04	0.02	-0.01	0.01	0.02	0.03	0.03	0.01	-0.01	0.03	0.02	0.00	-0.01	0.02	0.03	0.01	-0.03
T70%	0.02	0.04	-0.02	-0.01	0.02	0.03	0.03	-0.01	-0.02	0.03	0.02	0.01	-0.01	0.02	0.03	0.01	-0.03
H%	-0.01	-0.02	0.04	-0.00	0.03	0.01	0.01	-0.00	0.04	0.01	-0.03	0.03	0.03	0.01	0.02	0.03	0.01
O%	0.01	-0.01	-0.00	0.08	0.01	0.00	0.00	0.08	0.02	0.00	0.01	0.01	-0.01	0.01	0.00	0.01	0.00
n (Eq. 149 and 151)	0.02	0.02	-0.03	0.01	0.03	0.02	0.01	0.01	-0.03	0.01	0.03	0.02	-0.03	0.00	0.02	0.03	-0.01
k (Eq. 155)	0.03	0.03	-0.01	0.00	0.02	0.03	0.03	0.00	-0.01	0.03	0.02	0.00	-0.01	0.02	0.03	0.01	-0.03
k (Eq. 156)	-0.03	-0.03	0.01	0.00	0.01	0.03	0.03	0.00	0.01	0.02	-0.01	0.01	0.00	0.02	0.02	0.00	0.02
OC%	0.01	-0.01	-0.00	0.08	0.01	0.00	0.00	0.08	0.02	0.00	0.01	0.01	-0.01	0.01	0.00	0.01	0.00
CH%	-0.01	-0.02	0.04	0.02	0.03	0.01	0.01	0.02	0.04	0.01	-0.03	0.03	0.03	0.01	0.02	0.03	0.01
MW (Eq. 145)	0.03	0.03	-0.01	0.00	0.01	0.03	0.02	0.00	-0.01	0.03	0.01	0.01	-0.00	0.02	0.03	0.00	-0.03
SG	0.02	0.02	-0.03	0.01	0.03	0.02	0.01	0.01	-0.03	0.01	0.03	0.02	-0.03	0.00	0.02	0.03	-0.01
AP (Eq. 146)	0.00	0.01	0.03	-0.01	0.02	0.00	0.01	-0.01	0.03	0.01	-0.02	0.04	0.03	0.03	0.00	0.03	-0.01
Kw	-0.01	-0.01	0.03	-0.01	0.03	0.01	0.00	-0.01	0.03	0.00	-0.03	0.03	0.03	0.02	0.01	0.03	-0.00
Pc (Eq. 134)	-0.02	-0.02	-0.01	0.01	0.00	0.02	0.02	0.01	-0.01	0.02	0.00	0.03	-0.02	0.03	0.01	0.02	0.02
Tc (Eq. 137)	0.03	0.03	-0.02	0.00	0.02	0.03	0.02	0.00	-0.02	0.03	0.02	0.00	-0.01	0.01	0.03	0.01	-0.02
dc (Eq. 138 and 145)	0.01	0.01	-0.03	0.01	0.03	0.01	0.00	0.01	-0.03	0.00	0.03	0.03	-0.03	0.02	0.01	0.03	0.00
Zc	-0.03	-0.03	0.01	0.00	0.01	0.03	0.02	0.00	0.01	0.03	-0.01	0.01	-0.00	0.02	0.02	0.00	0.03

Figure 6: Covariance matrices obtained with all the scaled variables (with the range method). Only the properties listed from Table 1 are reported. Green refer to high positive correlation (higher than the threshold 0.028. Red refer to high negative correlation (lower than the threshold -0.028).

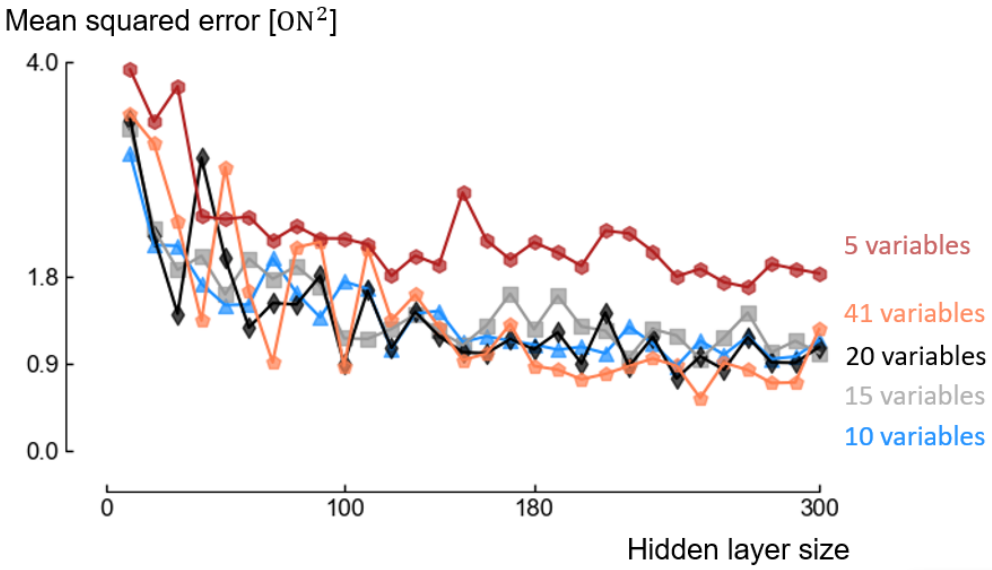

Figure 7: MSE with the Procrustes and the Vast methods with the principal variables selected based on 4 eigen vectors. The MSE decreases with the size of the hidden layer. Moreover, the MSE decreases when only 5 properties are considered. Therefore, the Procrustes selection method is not able to reduce the number of variables to 5.

Finally, the estimated octane numbers and the sensitivity are plotted against the target data for the combination of the best methods on Figures 8, 9 and 10.

The octane numbers are first discussed followed by the sensitivity. The prediction tends to be over-estimated for the low octane numbers and slightly under-estimated for the high octane numbers. To obtain a more precise view of the produced octane numbers comparing with the target octane numbers, the best and the worst prediction are reported in Tables 5

and 6. It is noteworthy that the worst values are not the most representative predictions as majority of the predicted data points fall within a 2% confident interval, as shown on Figures 8 and 9. The prediction ability of 2% is not better than linear by volume blending models, nevertheless, these predicting models can only be applied for simple fuels which composition is known. The power of the new model come from its ability to predict the octane numbers of complex fuels even if the composition is unknown. All the octane numbers are also reported in the Supporting Information.

The accuracy of the predicted sensitivity increases when the number of input variables decreases. The best prediction corresponds to a maximal error of $\pm$ 1 sensitivity point. Analysing the sensitivity is also a way to evaluate the propensity of the model to correctly estimate the kinetic behaviour of the fuel. More specifically, it is well known that paraffinic fuels show negative temperature coefficient (NTC) behaviour with a pronounced low temperature reactivity. Addition of aromatics or alcohols tends to suppress the NTC region. This observation impacts the sensitivity which increases as the aromatic or the alcohol fraction is added in the studied fuel.[31] Thus, the sensitivity for the best predicting method, i.e. Procustes coupled with the Auto method and 10 properties, was plotted depending on the different hydrocarbon class fractions on Figure 11. We observe a high impact of the saturate and aromatic groups on the sensitivity which follow the expected behaviour. The impact of the oxygenate group is less pronounced although it tends to slightly increase the sensitivity.

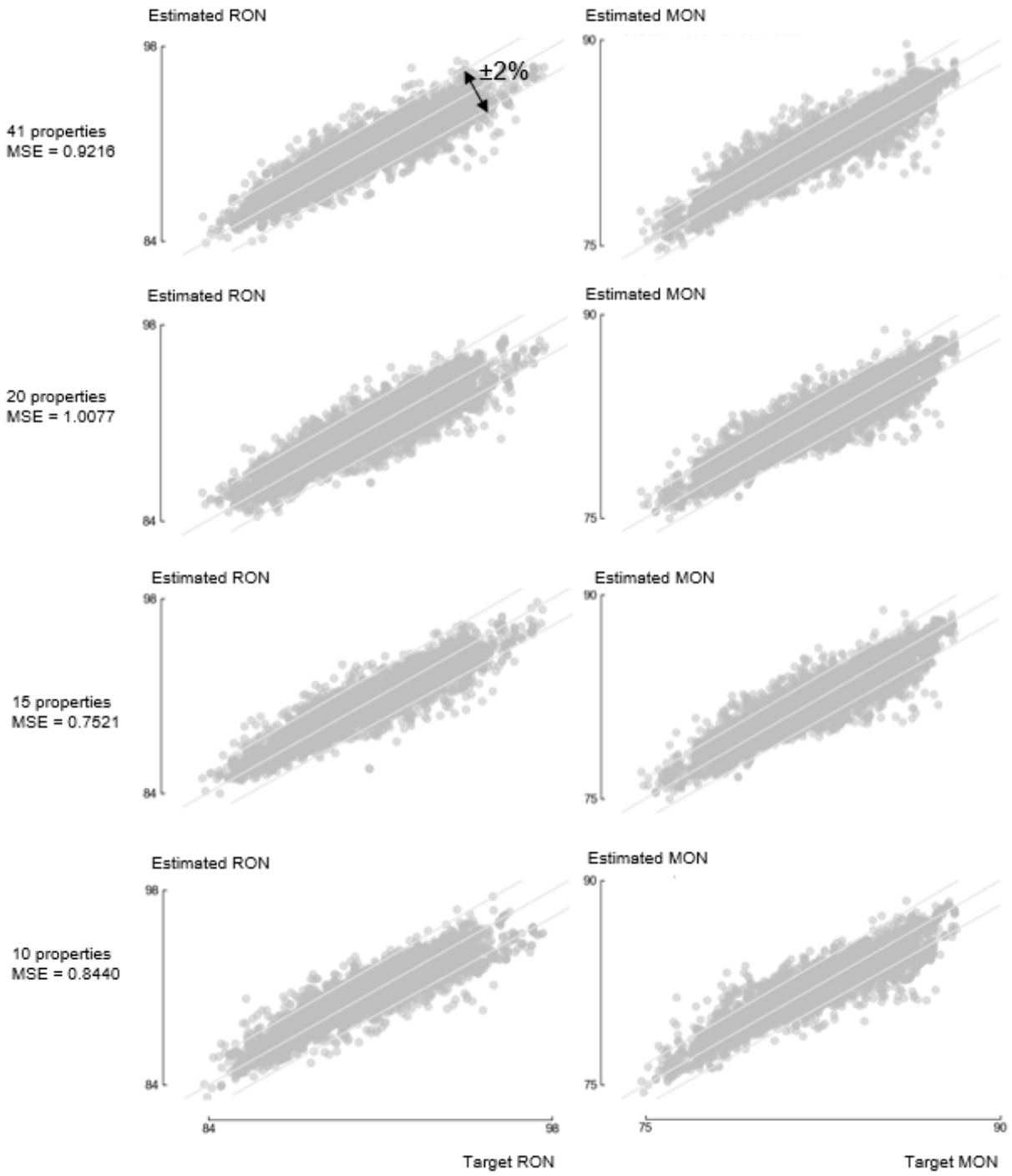

Figure 8: Estimated VS target octane numbers with the Procrustes variable selection method and the Auto scaling method.

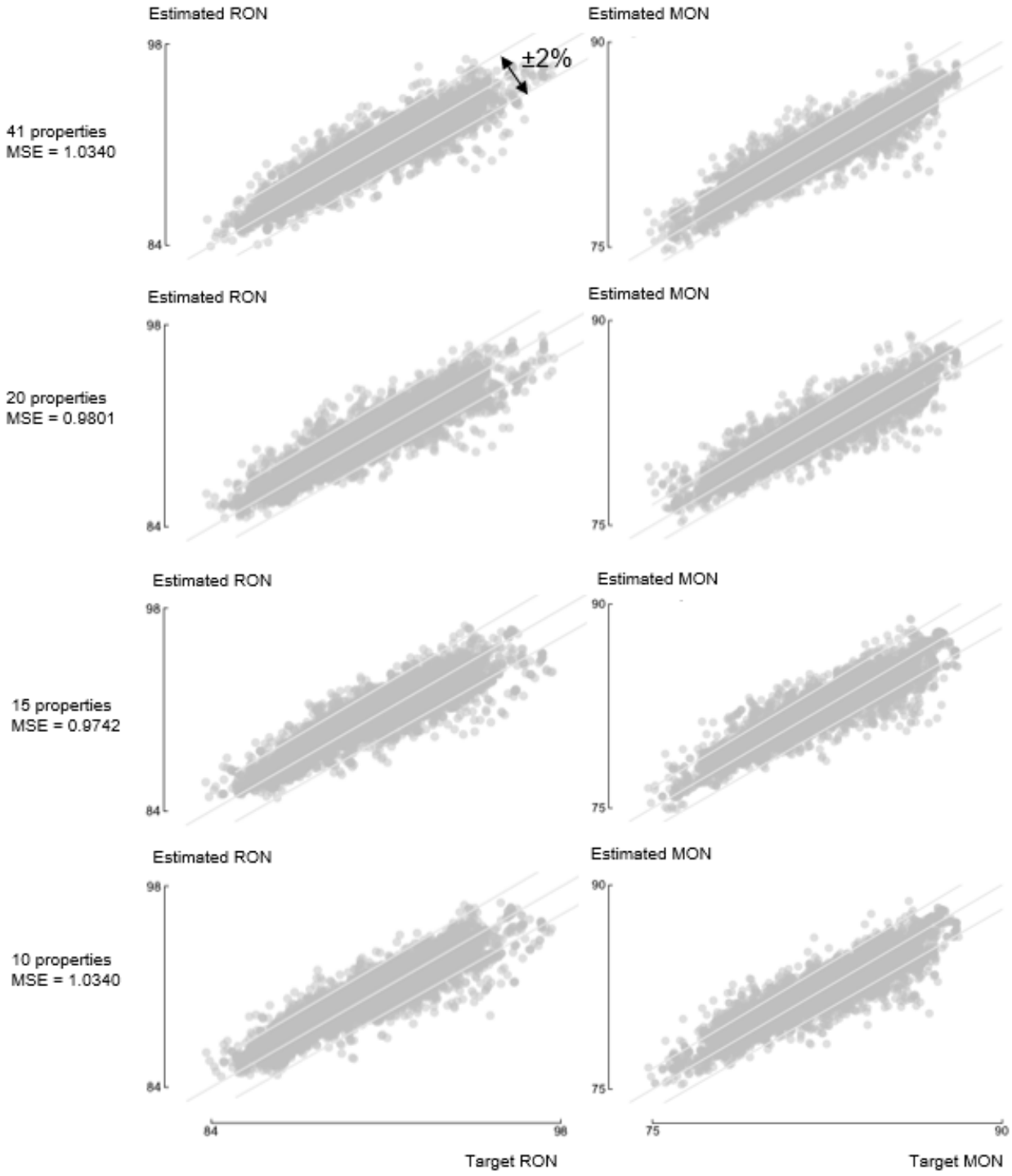

Figure 9: Estimated VS target octane numbers with the Procrustes variable selection method and the Range scaling method.

Table 5: Comparison of the target versus predicted octane numbers with the Procustes and Auto methods

	RON target/predicted	MON target/predicted
42 variables		
Best prediction	90.6/90.6	80.3/80.3
Worst prediction	95.8/90	86.9/80.6
20 variables		
Best prediction	89.6/89.6	82.4/82.4
Worst prediction	95.2/90.7	86.3/81.1
15 variables		
Best prediction	86.7/86.7	80.2/80.2
Worst prediction	95.2/90.4	86.9/81.4
10 variables		
Best prediction	88.0/88.0	86.7/86.7
Worst prediction	95.2/89.9	86.9/81.1

Table 6: Comparison of the target versus predicted octane numbers with the Procustes and Range methods

	RON target/predicted	MON target/predicted
42 variables		
Best prediction	88.6/88.6	77.3/77.3
Worst prediction	95.2/90.2	86.9/81.1
20 variables		
Best prediction	90.4/90.4	76.3/76.3
Worst prediction	95.2/90.6	86.3/80.9
15 variables		
Best prediction	93.6/93.6	79.1/79.1
Worst prediction	95.2/90.1	86.9/81.1
10 variables		
Best prediction	86.5/86.5	87.1/87.1
Worst prediction	94.7/89.6	86.5/80.5

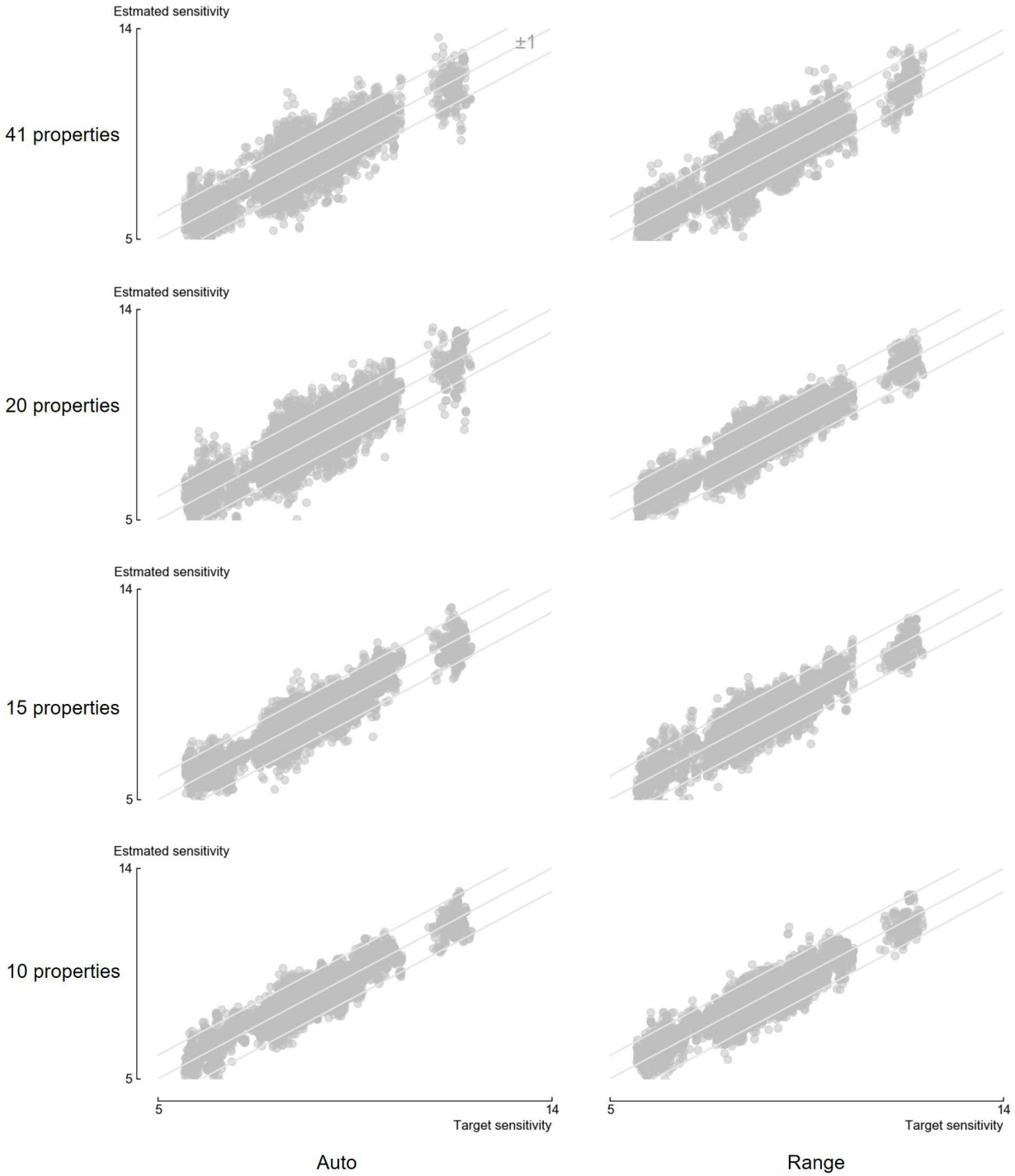

Figure 10: Estimated VS target sensitivity (RON - MON) with the Procrustes variable selection method and the Auto and the Range scaling methods.

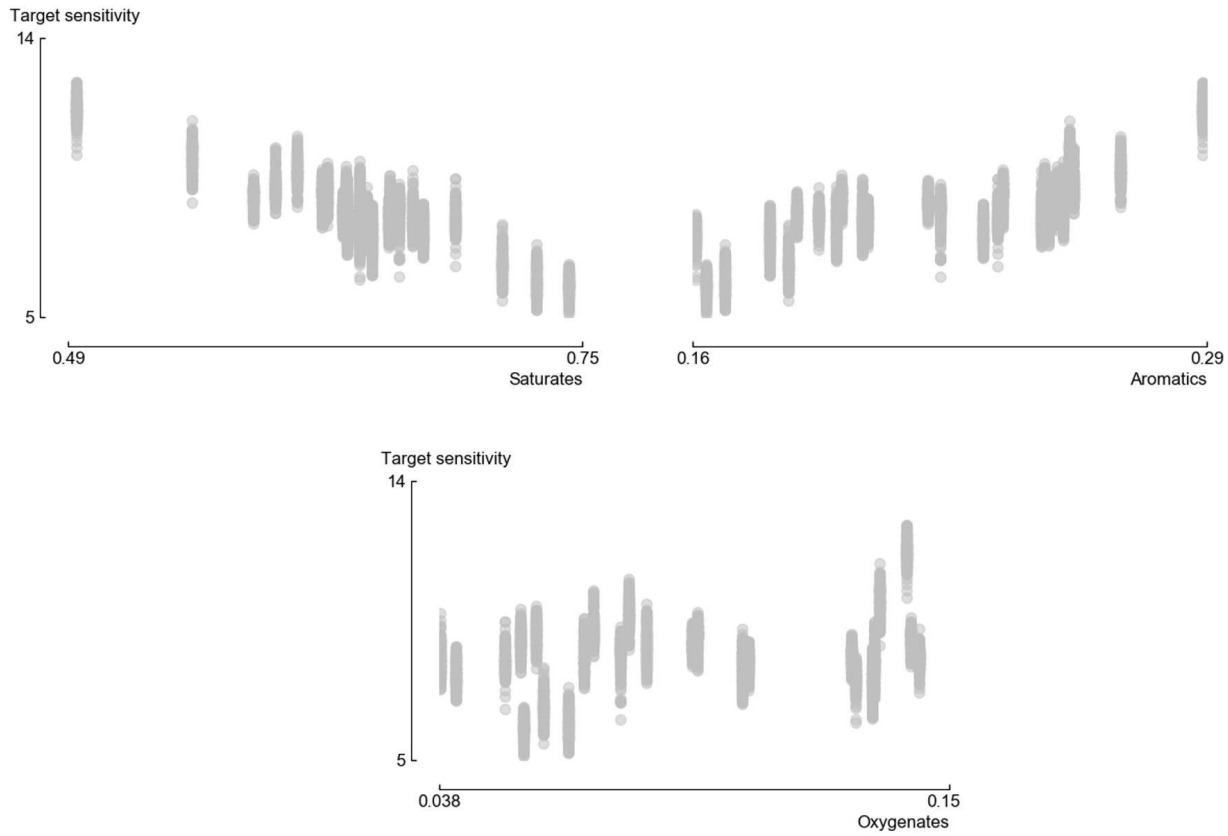

Figure 11: Evolution of the estimated sensitivity with the different hydrocarbon class fractions.

Conclusion

Relying on alternative sources of energy and new energy vectors is a way to achieve the sustainable development scenario of the world energy outlook. Among the different sources of energy, heavy alcohols have a particular role to play as they are environment-friendly, they are an alternative source of energy than fossil fuels and they can reduce the fuel consumption and the greenhouse gas emissions. Heavy alcohols can be produced via fermentation or gasification.

In the present paper, a prediction method based on ANN was developed to predict the octane numbers of gasoline blendstocks mixed with a alcohol among 1-propanol, 2-propanol, 1-butanol, 2-butanol and 2-methyl-1-propanol. Predicting the octane numbers is of major importance as it is a way to prevent phenomena such as knock that damages the engine. Moreover, being able to have an easy method to monitor the octane number is crucial due to the fluctuation of the fuel properties depending on the raw matter to produce the fuel.

The novelty of the proposed model is to be based on chemical and physical properties which are cheap to evaluate, so that it can be applied even if the composition is unknown, which is useful for complex fuel whose composition cannot be entirely defined. In contrast, a predicting models from the literature rely on the whole fuel composition[12].[17] The methods from the literature either require the whole fuel composition and the detailed octane number of each molecule or PIONAOx hydrocarbon classes while measuring the composition can be expensive[15].[16] Also, the methods from the literature has never been developed for large oxygenate molecules[14 15].[16] Some other models were developed only for pure molecules[18] or for blends of gasoline streams.[13]

Additionally, knowing the whole composition or all the octane numbers is not feasible for complex fuels. Moreover, studying 41 different properties from the literature is also new and it helped to select some of them applying PCA to study the principal properties. The principal properties that were found out during this study constitute a new result. The number of input properties was investigated, from 10 to 41 with a selection method

driven by PCA. A large number of properties were investigated, so, with the results that were collected, the properties useful for ON prediction were identified. It was shown that the number of input properties is not correlated with the size of the hidden layer. Hidden layers of the same size allowed us to reach similar results whatever is the number of input properties. Additionally, as long as the size of the hidden layer of the ANN is big enough (>180 neurons), 10 properties predict the octane number accurately. The 10 properties were calculated only based on the distillation curve, on the atomic content and on the specific gravity. So, another novelty of the method comes from the small number of properties that must be measured and the simplicity of the related measurement methods.

A new feature from the applied methodology is the use of PCA to study the principal variables in a fuel blend. This kind of study could be realized with the chemical data of fuels obtained with advanced analytical methods such as NMR or GC×GC.

The current methodology is only applicable for a given type of fuel: a gasoline blendstock mixed with an oxygenated molecule. As prospects, it would be interesting to compare the required properties for ON prediction with another type of fuel. Moreover, an experimental campaign with a large number of fuels would be useful to provide data to validate the current methodology.

Acknowledgement

This work was carried out within the framework of an EU project and benefited from a grant from la Region wallonne.

GD has received funding from the Fonds National de la Recherche Scientifique (FRS-FNRS) through a FRIA fellowship.

This research has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No 714605.

Supporting Information Available

The following file is available free of charge.

- SupportingInformation.xlsx: input, target and estimated octane numbers, and composition of the simulated fuels

References

(1) Ferrarotti, M.; Bertolino, A.; Amaduzzi, R.; Parente, A. On the Influence of Kinetic Uncertainties on the Accuracy of Numerical Modeling of an Industrial Flameless Furnace Fired With NH3/H2 Blends: A Numerical and Experimental Study. *Frontiers in Energy Research* **2020**, *8*, 333.

(2) Rixhon, X.; Limpens, G.; Coppitters, D.; Jeanmart, H.; Contino, F. The Role of Electrofuels under Uncertainties for the Belgian Energy Transition. *Energies* **2021**, *14*.

(3) Cassiers, S.; Boveroux, F.; Martin, C.; Maes, R.; Martens, K.; Bergmans, B.; Idczak, F.; Jeanmart, H.; Contino, F. Emission Measurement of Buses Fueled with Biodiesel Blends during On-Road Testing. *Energies* **2020**, *13*.

(4) Tipler, S.; Parente, A.; Coussement, A.; Contino, F.; Symoens, S. H.; Djokic, M. R.; Geem, K. M. V. Prediction of the PIONA and oxygenate composition of unconventional fuels with the Pseudo-Component Property Estimation (PCPE) method. Application to an Automotive Shredder Residues-derived gasoline. SAE Technical Paper Series. 2018.

(5) Tipler, S.; Mergulhão, C. S.; Vanhove, G.; Van Haute, Q.; Contino, F.; Coussement, A. Ignition Study of an Oxygenated and High-Alkene Light Petroleum Fraction Produced from Automotive Shredder Residues. *Energy & Fuels* **2019**, *33*, 5664–5672.

(6) Christensen, E.; Yanowitz, J.; Ratcliff, M.; McCormick, R. L. Renewable Oxygenate Blending Effects on Gasoline Properties. *Energy & Fuels* **2011**, 4723–4733.

(7) Ma, Y.; Huang, S.; Huang, R.; Zhang, Y.; Xu, S. Ignition and combustion characteristics of n-pentanol–diesel blends in a constant volume chamber. *Applied Energy* **2017**, *185*, 519–530.

(8) ASTM International, ASTM D2699-15a, Standard Test Method for Research Octane Number of Spark-Ignition Engine Fuel. **2015**,

(9) ASTM International, ASTM D2700-16a, Standard Test Method for Motor Octane Number of Spark-Ignition Engine Fuel. **2016**,

(10) Pera, C.; Knop, V. Methodology to define gasoline surrogates dedicated to auto-ignition in engines. *Fuel* **2012**, *96*, 59–69.

(11) Tipler, S.; Fürst, M.; Van Haute, Q.; Contino, F.; Coussement, A. Prediction of the Octane Number: A Bayesian Pseudo-Component Method. *Energy & Fuels* **2020**, *34*, 12598–12605.

(12) Albahri, T. a. Structural Group Contribution Method for Predicting the Octane Number of Pure Hydrocarbon Liquids. *Industrial & Engineering Chemistry Research* **2003**, *42*, 657–662.

(13) Pasadakis, N.; Gaganis, V.; Foteinopoulos, C. Octane number prediction for gasoline blends. *Fuel Processing Technology* **2006**, *87*, 505 – 509.

(14) Doicin, B.; Onutu, I. Octane Number Estimation Using Neural Networks. *Revista de Chimie* **2014**, *65*, 599–602.

(15) Ibrahim, E. A.; Farooq, A. Octane Prediction from Infrared Spectroscopic Data. *Energy & Fuels* **2020**, *34*, 817–826.

(16) Abdul Jameel, A. G.; Van Oudenhoven, V.; Emwas, A.-H.; Sarathy, S. M. Predicting Octane Number Using Nuclear Magnetic Resonance Spectroscopy and Artificial Neural Networks. *Energy & Fuels* **2018**, *32*, 6309–6329.

(17) 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. *Industrial and Engineering Chemistry Research* **2017**, *56*, 12236–12245.

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

(19) Riazi, M. R. *Characterization and Properties of Petroleum Fractions*; ASTM International, 2005.

(20) American Petroleum Institute, Technical Data Book- Petroleum Refining. **1997**,

(21) Aspen Technology Inc., *Aspen Physical Property System: Physical Property Methods*; 2013; pp 1–234.

(22) Parente, A.; Sutherland, J. C. Principal component analysis of turbulent combustion data: Data pre-processing and manifold sensitivity. *Combustion and Flame* **2013**, *160*, 340 – 350.

(23) Parente, A.; Sutherland, J. C.; Tognotti, L.; Smith, P. J. Identification of low-dimensional manifolds in turbulent flames. *Proceedings of the Combustion Institute* **2009**, *32 I*, 1579–1586.

(24) Jolliffe, I. T. Discarding Variables in a Principal Component Analysis. II: Real Data. *Applied Statistics* **1973**, *22*, 21.

(25) Krzanowski, W. Selection of Variables to Preserve Multivariate Data Structure, Using Principal Components. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **1987**, *36*, 22–33.

(26) Coussement, A.; Gicquel, O.; Parente, A. MG-local-PCA method for reduced order combustion modeling. *Proceedings of the Combustion Institute* **2013**, *34*, 1117–1123.

(27) Isaac, B. J.; Coussement, A.; Gicquel, O.; Smith, P. J.; Parente, A. Reduced-order PCA models for chemical reacting flows. *Combustion and Flame* **2014**, *161*, 2785–2800.

(28) D'Alessio, G.; Attili, A.; Cuoci, A.; Pitsch, H.; Parente, A. Unsupervised Data Analysis of Direct Numerical Simulation of a Turbulent Flame via Local Principal Component Analysis and Procustes Analysis. International Workshop on Soft Computing Models in Industrial and Environmental Applications. 2020; pp 460–469.

(29) D' Alessio, G.; Attili, A.; Cuoci, A.; Pitsch, H.; Parente, A. *Data Analysis for Direct Numerical Simulations of Turbulent Combustion*; Springer, 2020; pp 233–251.

(30) D' Alessio, G.; Cuoci, A.; Parente, A. Feature extraction and artificial neural networks for the on-the-fly classification of high-dimensional thermochemical spaces in adaptive-chemistry simulations. *Data-Centric Engineering, 2, E2* **2021**,

(31) Leppard, W. R. The chemical origin of fuel octane sensitivity. *SAE Technical Paper 902137* **1990**, 16.