A Novel Methodology for Molecular Design via Data Driven Techniques

Wan Qi Woo,[1] Lik Yin Ng,[2] Umaganeswaran Sivaneswaran[1] and
Nishanth G. Chemmangattuvalappil[1*]

[1]Department of Chemical and Environmental Engineering, Centre of Excellence for
Green Technologies, The University of Nottingham, Malaysia Campus,
Broga Road, 43500 Semenyih, Selangor, Malaysia
[2]Universiti Tunku Abdul Rahman, Jalan Sungai Long, Bandar Sungai Long,
Cheras 43000, Kajang, Selangor, Malaysia

[*]Corresponding author: Nishanth.c@nottingham.edu.my

Published online: 15 February 2017

To cite this article: Woo, W. Q. et al. (2017). A novel methodology for molecular design via data driven techniques. *J. Phys. Sci.*, 28(Supp. 1), 1–24, https://doi.org/10.21315/jps2017.28.s1.1

To link to this article: https://doi.org/10.21315/jps2017.28.s1.1

ABSTRACT: *The design of chemical products that satisfies customer requirements commences with the identification of desirable properties for a specific application. Molecular design techniques have been traditionally used to find the products that meet the identified properties. Conventionally, product design based on properties is done based on the assumption that property prediction models are available for the target properties. However, in many design problems encountered in industry, such prediction models may not be readily available. In this paper, we have developed a systematic framework for the design of chemical products by targeting the attributes defined by customer even if there are no property prediction models available for target properties. In addition, a methodology has been developed for the understanding of the global interactions between properties and their impact on environmental and technical performance. Different kinds of chemometric techniques have been used to define the needs of the customer in terms of physical properties. In the next step, computer aided molecular design techniques have been integrated with the data driven methodologies to design the optimum products. To illustrate the applicability of the developed framework, a case study has been discussed to design environmentally benign chemical products to satisfy property requirements of a biofuel additive. This has been achieved in conjunction with the developed property models that represent consumer defined attributes of biofuel additives.*

Keywords: Molecular design, chemometric techniques, group contribution models, data driven techniques, product design

1. INTRODUCTION

1.1 Chemical Product Design

Chemical products consist of a very wide range of scopes and they can be generally categorised into three classes, namely basic chemical products, industrial products and consumer products. The first class is the basic chemical products, which usually consists of well-defined molecules and mixtures of molecules. Secondly, industrial products are those typically categorised by thermophysical and transport properties. Finally, consumer products are manufactured by basic chemical and industrial chemical products. Unlike other products, configured consumer chemical products are usually sold to the consumers. Though different in functionality and quantity produced, the procedure in designing these different classes of chemical products is similar.[1]

Due to the changes in chemical industry over the recent years, the design of chemical products has become more important and essential. Chemical product design is the process of choosing the right product to be made for a specific application.[1] It can be defined as the identification of molecule/mixture that possesses properties which fulfils a set of desired customer requirements. Moggridge and Cussler proposed a general framework of chemical product design process.[1] Figure 1 illustrates the four stages of chemical product design.

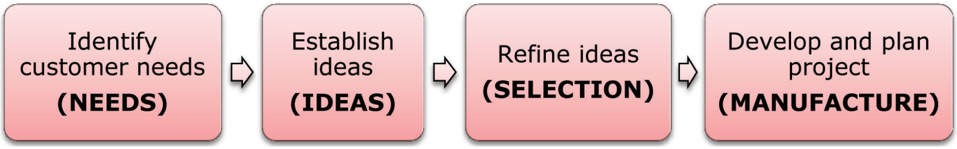


Figure 1: Chemical product design stages.

Based on the framework, the chemical product design consists of four stages. They are the needs, ideas, selections and manufacture. Chemical product design begins by identifying the desirable properties of the target chemical. This is required in order to satisfy a specific industry or the consumer's needs. It can be seen that the whole chemical product design process is driven and governed by the requirements of the customers. In many occasions, the customers' interests can be quantified in terms of product properties. Therefore, it is important to have a tool that systematically identifies the chemicals or blends with desirable properties.[2]

1.2 Property Based Design

In most cases, functionality in a chemical product is defined in terms of physical properties rather than the chemical structure of the product. Since customer requirements are the driving force for chemical product design, it is important to convert their qualitative requirements into quantitative product specification in order to design a chemical product.[3] This can be done by computer aided molecular design (CAMD) techniques. CAMD techniques are important for chemical product design for their ability to design and estimate molecules with a given set of product target specifications.[4] Therefore, these techniques are often used in the early stage of chemical product design for screening purpose. While utilising CAMD techniques in chemical product design problems, property models are used to estimate the properties of the product. These property models are computational tools developed to predict the molecular properties from structural descriptors, which are used to quantify molecular structure of a molecule. Some of the frequently used structural descriptors are chemical bonds and molecular geometry. However, most of the formulations for chemical product design problem involve highly non-ideal mixtures and thus, the property models are usually have not been developed yet.[4] The common models that are available for prediction of controlled chemistry are the thermodynamic models. Therefore, those non-thermodynamic properties are necessary to be further investigated or must be correlated to the thermodynamic properties.

1.3 Challenges and Motivation

Most of the current product design algorithms are based on the assumption that the property prediction models are available for the target properties. However, if the required property models are not available, the current available techniques are difficult to be used for the chemical product design problems. Therefore, it is necessary to develop a framework to solve product design problems by addressing the absence of property models. In this work, property models are generated by utilising experimental data for target properties where property prediction models are unavailable. The model development is based on the hypothesis that these properties can be represented as a function of other physical properties. Hence, the property model can be generated by identifying the correlation between the properties with other physical properties.

2. METHODOLOGY

2.1 Proposed Framework

As mentioned earlier, the objective of this study is to develop a framework that integrates data-driven techniques with molecular design methodologies for optimal chemical product design. Through this, it will be possible to identify the optimal molecules with specific properties even if there are no models available to predict those properties. The developed integrated framework is divided into three interconnected phases, namely the problem formulation, statistical model development and molecular design.

2.1.1 Problem formulation

In problem formulation, the objective of the chemical product design problem is defined. The sources of this objective can either be a brand new chemical product with high market demand or a requirement for updating an existing chemical product with better functionalities. For both cases, the satisfaction of the need is formulated as a problem. In the next step, it is required to understand and define the influential design target parameters in terms of measurable product properties (e.g., boiling point, melting point, flash point, surface tension, viscosity, etc.) that affect the main objective such that it satisfies the customer needs.

Based on the targeted properties defined above, property models are required to generate and solve the defined objective. The most common property prediction models are based on group contributions (GC) techniques.[5–7] These techniques consider a molecule as a collection of various molecular groups. The properties of the molecule can then be estimated as a summation of the contributions of the molecular groups and their frequency in the molecule. The property estimation model developed by using GC method can be represented by the following equation:[7]

$$f(X) = \sum_i N_i C_i + w \sum_j M_j D_j + z \sum_k O_k E_k \tag{1}$$

where *f(X)* is a function of the property *X*, *w* and *z* are binary coefficients depending on the levels of estimation, N_i, M_j, O_k are the number of occurrence of first, second and third-order group contribution correspondingly and C_i, D_j, E_k are contribution of first, second and third-order group subsequently. Second and third order molecular groups were developed as improvement to the first order molecular groups. By utilising these higher order molecular groups in GC methods, different functional groups and isomers can be distinguished and properties of molecules which involve polyfunctional groups can be predicted.[6,7] GC methods have been widely applied in estimating a number of thermodynamic properties of organic compounds.[6,7] However, not every physical property/attribute has specific property prediction models available. For these properties/attributes, it is necessary to develop new models based on experimental data. However, development of new group contribution models is a time consuming and tedious task. Therefore, in this work, the focus is on developing a prediction model based on the underlying properties of the attributes. The method used to develop target property models will be discussed in the following phase.

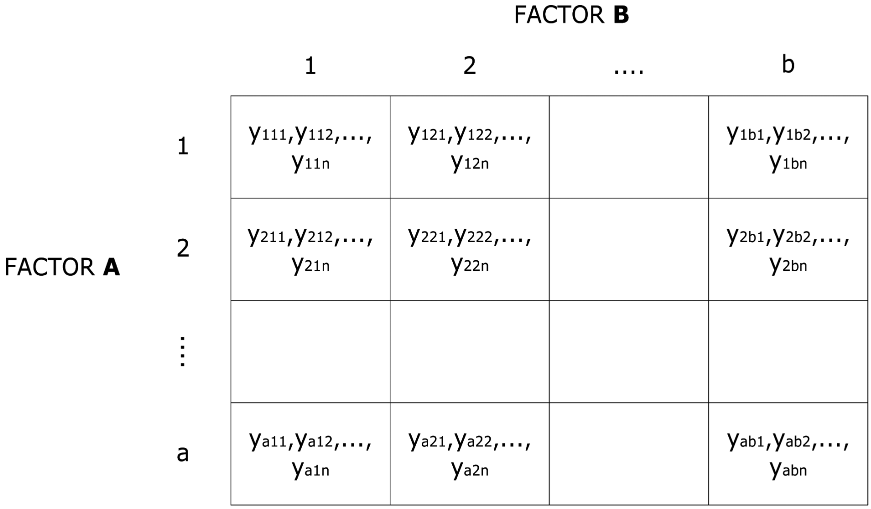
2.1.2 Statistical model development

In this stage, property models for properties which do not have available property models are developed. These physical properties can either be obtained from experimental data or estimated by utilising property prediction methods such as GC methods or TIs. These experimental data or predicted values are then used as the source for the development of the property model.

There are several strategies available develop a statistical model such as fitting regression models, factorial design, and analysis of variance (ANOVA), etc. Among all the strategies, a combination of factorial design technique and regression analysis has been chosen in this research methodology to evaluate the relationships between physical properties and attributes. The interactions between different properties are first addressed by utilising factorial designs. These interactions could have significant influence on the target attributes. Once the effect of different factors has been identified, the exact functional relationship between the properties and the attributes can then be developed by using regression analysis.

2.1.2.1 Factorial design

An experiment that involves the study of two or more factors is known as a factorial design, with the simplest type of factorial designs involving only two factors.[8] Factorial design technique appears to be very efficient for its possibility to systematically combine the levels being investigated for each complete trial or replication of the experiments.[9,10] In order to develop a property model, it is important to identify factors or independent variables that affect the response of interest or dependent variable. By utilising factorial design, effects and influence of the independent factors on the output of the dependent variable can be studied and investigated. Generally, a two-factor factorial experiment design can be represented as Figure 2.



		FACTOR **B**			
		1	2	....	b
FACTOR **A**	1	$y_{111}, y_{112}, ..., y_{11n}$	$y_{121}, y_{122}, ..., y_{12n}$		$y_{1b1}, y_{1b2}, ..., y_{1bn}$
	2	$y_{211}, y_{212}, ..., y_{21n}$	$y_{221}, y_{222}, ..., y_{22n}$		$y_{2b1}, y_{2b2}, ..., y_{2bn}$
	⋮				
	a	$y_{a11}, y_{a12}, ..., y_{a1n}$	$y_{a21}, y_{a22}, ..., y_{a2n}$		$y_{ab1}, y_{ab2}, ..., y_{abn}$

Figure 2: General arrangement for a two-factor factorial design.

Let y_{ijk} be the observed response with ith level of factor A (i = 1, 2, ..., a) and jth level of factor B (j = 1, 2, ..., b) for the kth replicate (k = 1, 2, ..., n). When the design factors are quantitative such as viscosity, pressure, density, etc., then a regression model representation of the two factorial experiments can be developed and written as:

$$y = \beta_0 + \beta_1(x_1) + \beta_2(x_2) + \beta_{12}(x_1 x_2) + \varepsilon \qquad (2)$$

where y is the response of interest, β_0, β_1, β_2 and β_{12} are the parameters values to be determined, x_1 represents factor A, x_2 represents factor B, x_1x_2 represents the interaction between A and B, and ϵ is random error term for the regression model.

Equation 2 above would be an illustration for the concept of interaction between two factors.[8] β_{12} represents the interaction coefficient between x_1 (factor A) and x_2 (factor B). This interaction coefficient might be ignored if it is relatively small as compared to the main effect coefficients β_1 and β_2; hence, this shows that there is no significant interaction occurring between both factors.

These concepts can be illustrated graphically by using a response surface plot and a contour plot for the specific model. The 3D response surface plot shown in Figure 3 illustrates a plane of y values created by various combinations of x_1 and x_2. On the other hand, Figure 4 shows the contour lines of constant response y in the x_1 versus x_2 plane. Notice that both example plots show a plane response surface and a contour plot containing parallel straight lines respectively. This indicates that no interaction was contributed between x_1 and x_2 along with a small coefficient of β_{12}. However, based on Figures 5 and 6, there is a significant interaction effect that twisted the plane in the response surface plot. This twisting effect of the response surface will cause the contour lines of constant response in the x_1 versus x_2 plane to be curved. Therefore, these two kinds of plot are very useful to represent an experiment model.

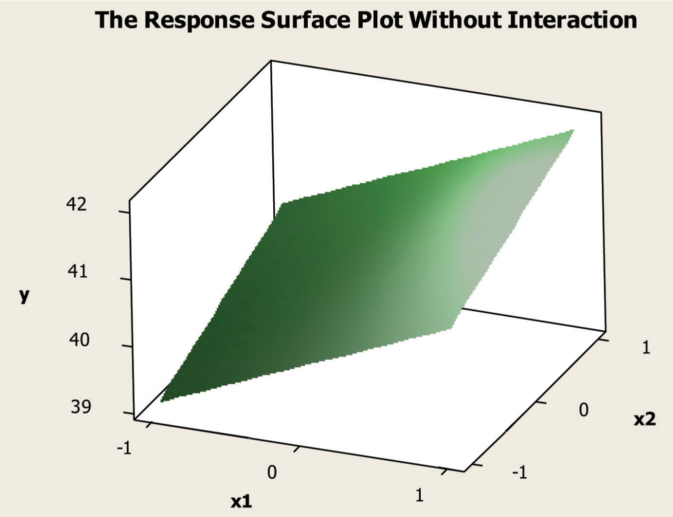


Figure 3: The response surface plot without interaction.

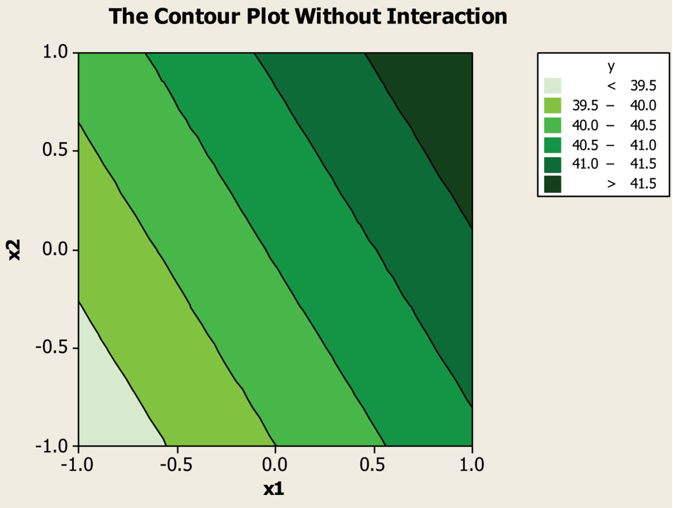


Figure 4: The contour plot without interaction.

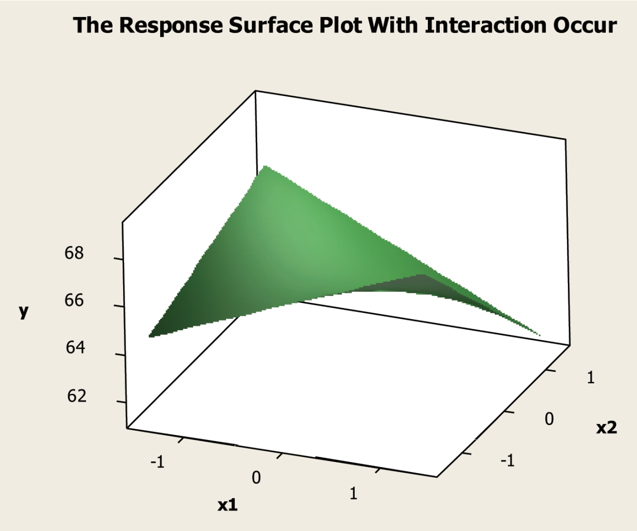


Figure 5: The response surface plot with interaction occurring.

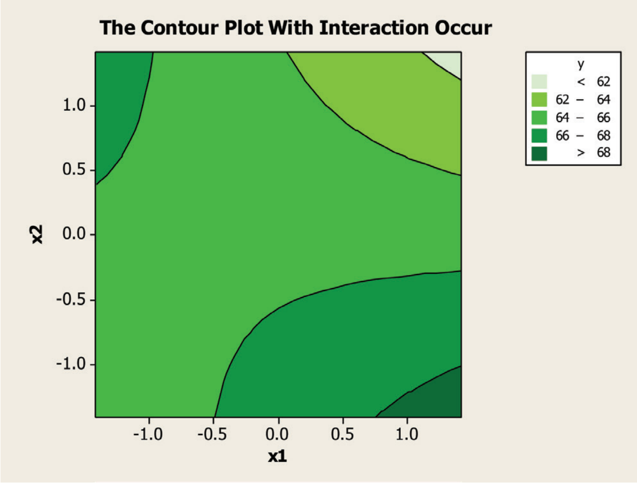


Figure 6: The contour plot with interaction occurring.

Once the underlying properties and their interactions have been identified, the next step is to use this information in developing the attribute-property relationship model. Based on the identified properties and the interactions, regression analysis can be used to generate the required model. Here, the attribute/property with no direct model is now represented in terms of other properties for which there are models. Therefore, the attribute can now be estimated from the molecular structure.

2.1.2.2 Model adequacy checking

The purpose in carrying out model adequacy checking is to ensure the robustness of the developed property models before utilising the model into molecular design problem. Decision is best indicated by the regression estimate for coefficient of determination (R^2)· As such, the high value of R^2 will lead to a high precision.[11] Besides that, residual analysis is a very helpful technique for model adequacy checking. Residual which is defined as the difference between the actual value and the predicted value can be calculated by the developed model. In Figure 7, the residuals are plotted against their expected values and residuals should be normal distributed. The data points plotted in this plot should form an approximate straight line to indicate that the normal distribution is a good model for this data set, whereas data points who far away from the line is consider as an outlier.[12]

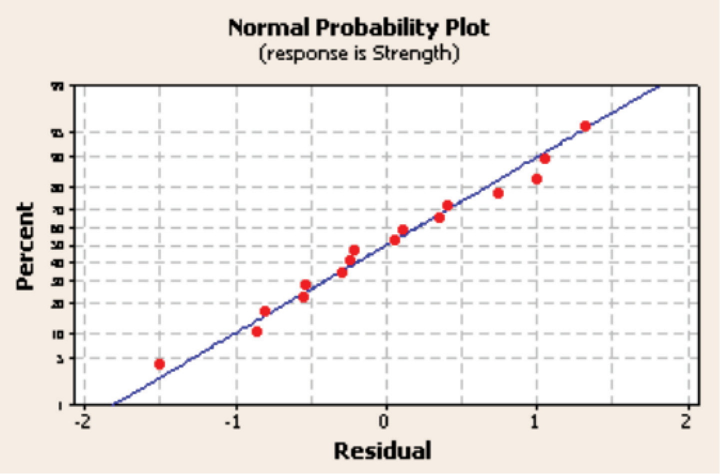


Figure 7: Normal probability plot of residuals.

After finalising the developed target properties model, the next step is to develop a molecular product design problem by using the target properties model develop together with other models that available in literatures.

2.1.3 Molecular design

The last phase of the methodology is to utilise the developed property model into a molecular product design problem to find the molecules with optimum target properties. The first step in solving the chemical product design problem is the identification of appropriate property models. For the case in which a suitable property model is not available for the property of interest, the methodology discussed in the previous section is utilised to identify the respective property model. Once all the chemical product properties can be identified by utilising either the existing or developed property models, the next step is to identify the property target ranges for the chemical product properties. These target ranges define the upper and lower limit which the product properties will fall within. Once the property target ranges have been determined, possible molecular groups are selected as potential building blocks. In this step, possible types of atoms and bonds of the final product are selected. Next, structural constraints are identified to eliminate the combination of infeasible solution and ensure the formation of complete molecular structures. Finally, to generate the optimal molecular structure subjected to the property models and structural constraints, the chemical product design problem solved as an optimisation programming problem by solving the objective function.

In designing the molecules to serve the specific purpose, some constraints are needed to be imposed on the molecule. These constraints include structural and some specific properties depending on the process requirements.

The structural constraints that are to be imposed are:

1. The Free Bond Number of the final molecule is zero to ensure that number of bonds attached to each group is equal to its valance and hence eliminating the possibility of existence any free hanging bonds in the molecule. The mathematical expression for the Free Bond Number (FBN), which is the number of free bonds in each acyclic molecular string is shown in Equation 3:[13]

$$FBN = \sum_{g=1}^{N_g} n_g FBN_g - 2\left(\sum_{g=1}^{N_g} b_g - 1\right) \tag{3}$$

   where *FBNg* is free bond number associated with group *g*.

2. The following expressions can be developed to ensure the existence of a meaningful molecule:

$$n_g \geq 0 \tag{4}$$

   In order to make sure that the designed molecules will meet the property targets identified in section 2.1.2, additional constraints must be imposed. In addition, it is important to make sure that the final structures will meet the constraints corresponding to environmental regulations as well.

$$\psi_{m}^{lower}\left(P_{m}\right) \leq \psi m\left(P_{m}\right) \leq \psi_{m}^{upper}\left(P_{m}\right) \tag{5}$$

where ψm is the function of the target property, and *m* that can be represented using a group contribution model.

With this input of process constraints, similar constraints for physical, environmental properties and preselected groups along with the structural constraints of:

$$N_f \leq N_{max}, \quad n_g \geq 0 \tag{6}$$

The number of first order groups that could be possibly present in the to-be designed molecules is maximised. The reason behind maximising these groups is to ensure that no potential molecule is left behind.

2.1.3.1 Enumeration of higher order groups

In order to increase the accuracy of property-based molecular design techniques, the effects of higher order molecular groups are to be considered while designing molecules. Higher order groups give a better description of molecules and differentiation of structural isomers is possible using these groups. The following methodology has been developed to estimate the contributions of higher order groups based on work of Chemmangattuvalappil et al.:[14]

Rule 1: Higher order groups can only be formed from complete molecular fragments. For instance, to form the higher order group CH (CH3) CH (CH3), there must be two CH and two (CH3) groups. It is not possible to consider a CH (CH3) group as a half higher order group.

Rule 2: If any of the higher order groups completely overlap some other higher order group, only the larger group must be chosen in order to prevent redundant description of the same molecular fragment.

So, if (l: *n*) is the set of first order groups that are the building blocks of one higher order group, *h*, (n_1: n_n) is the number of those first order groups present in the molecule, *η* is the number of occurrences of one particular first order group in a selected higher order group, n_h is the number of possible higher order groups from those first order groups, then:

$$n_h = Min\left(\frac{(n_l)}{\eta_l} : \frac{(n_n)}{\eta_n}\right) \tag{7}$$

n_h must be rounded down to the nearest integer number according to Rule 1. Moreover, some higher order groups may share a part of the higher order groups. For instance, 2 OH and 1 CH group can form 2CHOH groups. Hence, the possibility of sharing of various combinations of available first order groups participating in the given higher order group is considered.

So, If (i_l: i_n) number groups of (l: n) groups are shared, for all combinations of (i_l: i_n) such that (i_l: i_n)∈(n_1: n_n) and (n_1: n_n) ≥ (η_1,η_n) , the number of possible higher order groups is given by:

$$n_{gh} = Min\left(\frac{(n+i_l)}{\eta_l} : \frac{(n_{gn}+i_n)}{\eta_n}\right) \tag{8}$$

The groups with valance one alone is restricted from being shared. Highest value of n_{gh} is considered to be the maximum number of higher order groups n_g possible from given first order groups contributing to the presence of that higher order group in the molecule.

If $\psi_j^M\left(P_{j,mh}\right)$ is the property contribution from the higher order groups, it is calculated as:

$$\psi_j^M\left(P_{j,mh}\right) = \sum_{h=1}^{N_k} n_h P_{j,h} \tag{9}$$

The property for molecule i can now be estimated using Equation 10:[15]

$$\psi_j^M\left(P_{j,mh}\right) = \psi_j^M\left(P_{j,mf}\right) + \psi_j^M\left(P_{j,mh}\right) \tag{10}$$

Since all the combinations of all first order groups whose maximum is taken as the limit are considered, no potential molecule is left behind. Introduction of higher order groups in initial molecular design model would lead to a nonlinear model and their introduction initially doesn't increase the accuracy of the model fairly. So, finding the maximum of these higher order groups from different combinations of first order groups which obey the structural constraints would serve the purpose of finding potential molecules for a given process performance. Hence, all possible combinations of the numbers [0,n_g] and [0,n_h] are generated subject to following constraint.

$$FBN = 0 \tag{11}$$

The above method of generation enables the identification of structural isomers to some extent as the possibility of nonexistence of each higher order groups is considered. The possible molecules are screened out by checking if the combination of all the groups satisfies all structural, property and process constraints. Rule 2 indicates that a higher order group is not completely overlapped by any other higher order group. Hence, after screening out the molecules based on above constraints, an extra condition of whether any second order group is completely overlapped by another is checked before the final screening.

Since the molecule is designed by targeting the optimal properties, it can be ensured that the customer requirements have been addressed. Figure 8 illustrates the whole framework in a flow chart to summarise the developed approach. Therefore, this framework will be applied into a case study to illustrate the application of the novel approach.

3. CASE STUDY: DESIGN OF BIOFUEL ADDITIVES

3.1 Introduction

The application and effectiveness of the developed framework is shown by solving a biofuel additives design problem. Biofuels are found to contain lower carbon footprints as compared with fossil fuels on the basis of equivalent amount of energy produced.[16,17] This is a huge advantage of biofuels over fossil fuels as environmental issues are considered. However, there are a few major drawbacks for biofuels to be used in commercial applications. Although biofuels are well known for producing less harmful pollutant such as carbon monoxide (CO), particulate matter (PM) and hydrocarbon (HC), the emission of mono-nitrogen oxides (NOx) of biofuels are higher compared to conventional fuel.[18] One of the approaches in reducing this drawback is by adding fuel additives.[19]

3.2 Parameters Affect NO_x Emissions

The NO_x emissions are harmful due to its contribution to a wide range of environmental effects such as the formation of acid rain and also the formation of greenhouse gases. Many researchers have suggested various possible factors that affect the NO_x emission in biodiesel. According to Nettles-Anderson and Olsen,[20] viscosity of the biodiesel has a significant effect on the NO_x emission. At low temperatures, an increase of biodiesel viscosity will increase the emission of NO_x as the increase of viscosity will increase the NO_x emission at low temperatures. Besides that, the density of biofuel will influence NO_x emission. An increase in biofuel density will result in an increase of the NO_x formation. Besides, the NO_x concentration is also affected by the surface tension of the fuels.

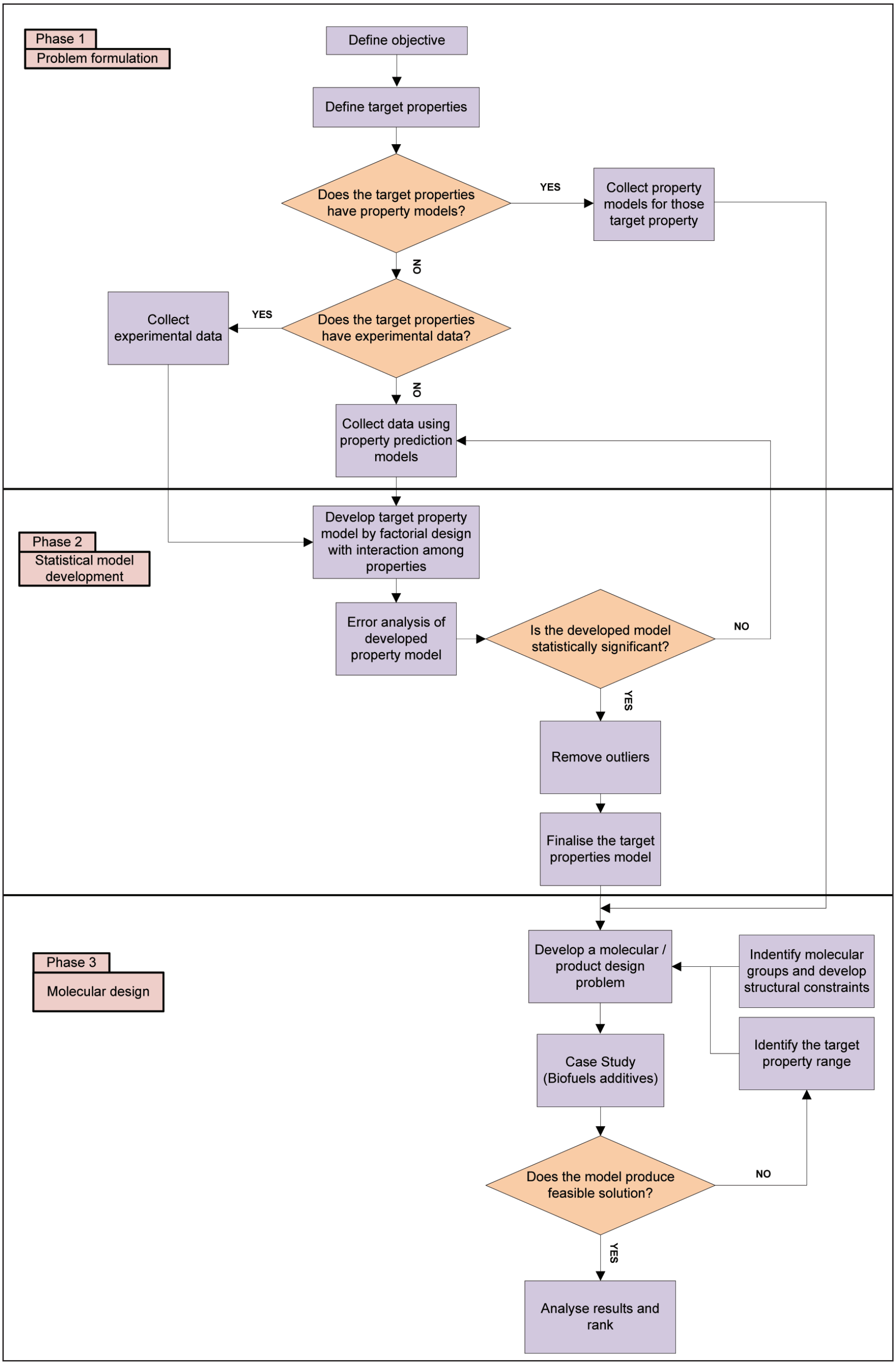

Phase 1
Problem formulation
Define objective
Define target properties
Does the target properties have property models?
YES
Collect property models for those target property
NO
Collect experimental data
YES
Does the target properties have experimental data?
NO
Collect data using property prediction models
Phase 2
Statistical model development
Develop target property model by factorial design with interaction among properties
Error analysis of developed property model
Is the developed model statistically significant?
NO
YES
Remove outliers
Finalise the target properties model
Phase 3
Molecular design
Develop a molecular / product design problem
Indentify molecular groups and develop structural constraints
Case Study (Biofuels additives)
Identify the target property range
Does the model produce feasible solution?
NO
YES
Analyse results and rank

Figure 8: The methodology flow chart for chemical product design.

Diesel spray properties include the Sauter mean diameter (SMD) and droplet size distribution.[21] The increase of SMD is affected by the increase of viscosity and surface tension of the fuel. During the premixed combustion phase, an increase of droplet size reduces the fraction of burning fuel that leads to the increment of the diffusion flame combustion duration, hence, the NO_x concentration increases. Last but not least, the increase in cetane number (CN) of biodiesel will reduce the NO_x formation. This is because ignition delay has been shortened and this reduces the combustion temperature along with the residence time.

3.3 Biodiesel Additives Design

Based on the previous explanation, there are several fuel properties that could be controlled in order to produce a superior biodiesel additive. In this section, it is desired to design a biofuel additive that can be used to increase the biodiesel cetane number for NO_x reduction. Besides that, other physical properties such as melting point, boiling point and flash point have been included in the design of biodiesel additives as well. These properties are to ensure the additive designed is in a liquid state where it will be suitable to mix with the biofuel itself. Flash point estimation is to determine the stability of the biofuel additive. Property and structural constraints were set to determine solutions that lie within the feasible region. The following subsection will discuss the process of designing the additives molecules by applying the framework developed in section 2.

3.3.1 Attributes model development

CN is a widely used fuel quality parameter. It is a dimensionless descriptor that measures the ignition quality of a fuel in a diesel engine.[22] However, there is no linear property model available to predict the cetane number from the chemical structure. As a hypothesis, NO_x emissions have strong relationship between cetane number and it can be represented as an empirical function of two physical properties which is viscosity and molar volume. Hence, data-driven approach has been used to correlate the viscosity and molar volume to generate a property model equation for cetane number.

As mentioned earlier, factorial design technique was performed to determine the effect and interactions between two or more factors on the respond of interest, whereby viscosity and molar volume are the factors that influenced the cetane number in this case.

$$CN = f\left(\mu, V_m\right) \tag{12}$$

The property model of cetane number developed by using factorial design and linear regression analysis can be represented as shown in Equation 13:

$$CN = -20.2075 + 0.224983\mu + 0.341760V_m - 0.00179108(\mu \times V_m) \qquad (13)$$

where μ is the viscosity and V_m is the molar volume.

The coefficient of determination (R^2) obtained based on 115 data points is 90.91%. In addition, from Figure 11, it can be deduced that the data follow normal distribution. A response surface (Figure 9) and a contour plot (Figure 10) have been plotted based on the predicted property model of cetane number to represent the interaction between viscosity and molar volume. The interaction effect in this case is not very strong.

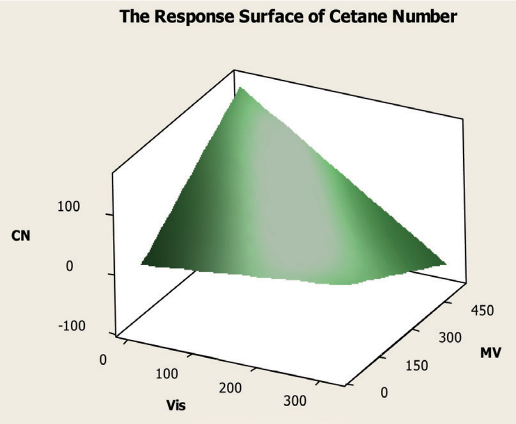


Figure 9: The response surface of cetane number.

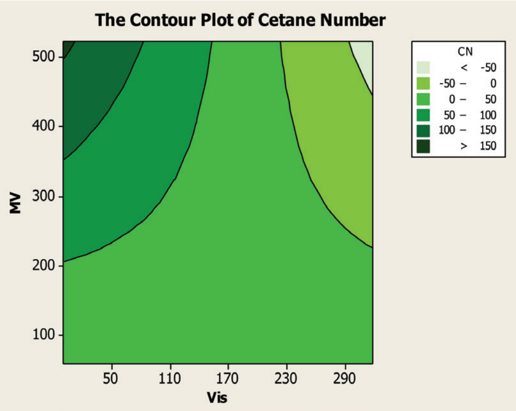


Figure 10: The contour plot of cetane number.

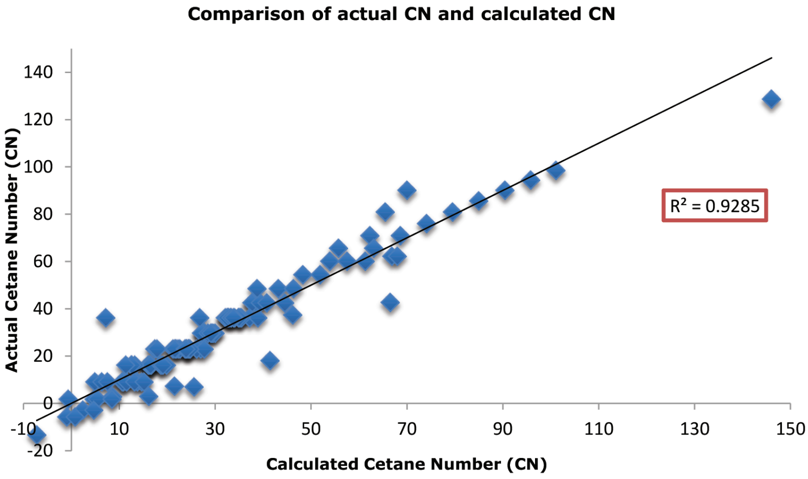


Figure 11: Comparison of actual CN and calculated CN.

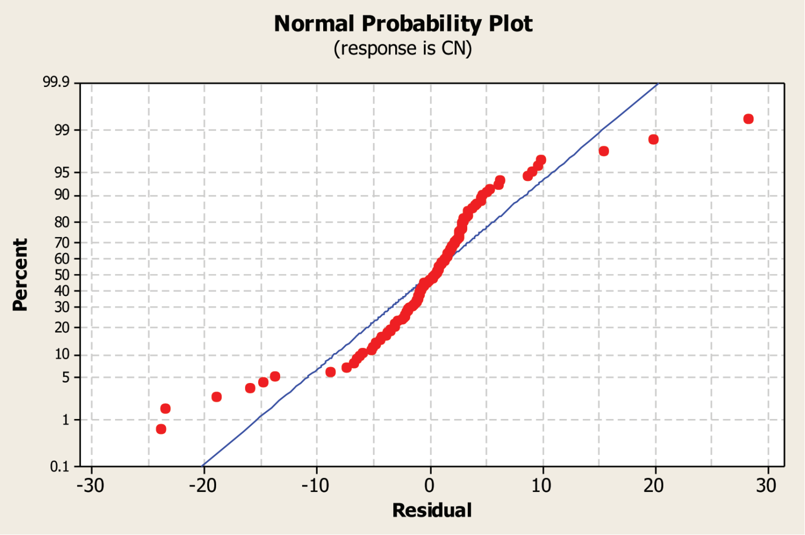


Figure 12: Normal probability plot of residuals obtained from statistical analysis software.

By referring to Figure 11, almost all observation data fall on the best fit line. On the other hand, Figure 12 shows the normal probability plot of cetane number. The plot of the data points is approximately a straight line. This target property model is utilised in the following steps on molecular design to develop additives with high cetane number.

3.3.2 Additives property estimation

As property models developed by using GC methods are available for boiling point, melting point, viscosity, molar volume and flash point, these properties are estimated by using the existing property prediction models. Below are the list of property models available in literatures and these models will cooperated in the molecular design problem section:

Viscosity (μ): Dynamic viscosity estimated at 300K, only first order group contributions are considered.[23]

$$\ln(\mu) = \sum\left(n_1\mu_1 + n_2\mu_2 + n_3\mu_3 + \ldots + n_n\mu_n\right) \tag{14}$$

where n_1 to n_n is the number of first groups, and μ_1 to μ_n is the first order group contributions values.

Molar Volume (***V***$_m$): Saturated liquid molar volume estimated at 298 K, in which only first order group contributions are considered.[24]

$$V_m - d = \sum\left(n_1V_{m1} + n_2V_{m2} + n_3V_{m3} + \ldots + n_nV_{mn}\right) \tag{15}$$

Where the d value is 0.01211 m^3 kmol^{-1}, n_1 to n_n is the number of first groups, and V_{m1} to V_{mn} is the first order group contributions values.

Normal melting point (*Tm*): Normal melting point temperature (T_m, K), only first order group contribution are considered.[7]

$$\mathrm{Exp}\left(T_m/T_{m0}\right) = \sum\left(n_1T_{m1} + n_2T_{m2} + n_3T_{m3} + \ldots + n_nT_{mn}\right) \tag{16}$$

where the value of T_{m0} is 147.450 K, n_1 to n_n is the number of first groups, and T_{m1} to T_{mn} is the first order group contributions values.

Normal boiling point (*Tb*): Normal boiling point temperature (T_b, K), only first order group contribution are considered.[7]

$$\mathrm{Exp}\left(T_b/T_{b0}\right) = \sum\left(n_1T_{b1} + n_2T_{b2} + n_3T_{b3} + \ldots + n_nT_{bn}\right) \tag{17}$$

where value T_{b0} is 222.543 K, n_1 to n_n is the number of first groups, and T_{b1} until Tbn is the first order group contributions values.

Flash point (*Tf*):[25]

$$T_f = 4.656 + 0.844T_b - 0.234 \times 10^{-3}\left(T_b^2\right) \tag{18}$$

where both flash point (T_f) and boiling point (T_b) are in unit Kelvin (K).

Molecular Weight (*M*):

$$M = \sum\left(n_1 M_1 + n_2 M_2 + n_3 M_3 + \ldots + n_n M_n\right) \tag{19}$$

where n_1 to n_n is the number of first groups, and M_1 to M_n is the molecular weight for respective group contribution.

3.3.3 Molecular design

Once the property models for all the properties are determined, the property target ranges for each of the property are identified as shown in Table 1.

Table 1: Constraints set to solve molecular design problem.

Property	Constraints
Viscosity (μ)	μ>0
Molar volume (V_m)	V_m>0
Normal melting point (T_m)	100K<T_b<260K
Normal boiling point (T_b)	450K<T_b<600K
Flash point (T_f)	*T_f>350K*

For the design of biofuel additives with maximum cetane number, 5 molecular groups have been considered as the building blocks, which are CH_3, CH_2, CH, C and OH. These molecular groups and their respective contributions for different properties are shown in Table 2, with various objectives stated previously which are Equations 13–19 together with their constraints to solve this additives molecular design.

Table 2: Group contributions values.

n_n	Group	Viscosity, η_n	Molar volume, V_{mn}	Melting point, T_{mn}	Boiling point, T_{bn}
*n*1	CH3	–1.0278	0.02614	0.6953	0.8491
*n*2	CH2	0.2125	0.01641	0.2515	0.7141
*n*3	CH	1.318	0.00711	–0.373	0.2925
*n*4	C	2.8147	–0.0038	0.0256	–0.0671
*n*5	OH	1.3057	0.00551	2.7888	2.567

Together with structural constraints, the objective's function is solved to obtain the optimal product. Multiple solutions are generated using integer cuts. The computational results shown below are the potential molecular structure of the biofuel additives with their estimated properties.

Table 3: Potential molecular structure of biofuel additives.

No	MS	CN	V	MV	MP	BP	FP	MW
1	C13H28O	60.57	9.86	242.40	258.67	530.35	386.45	200.13
2	C12H26O	55.59	7.97	225.99	252.11	515.18	377.36	186.12
3	C11H24O	50.45	6.45	209.58	245.24	498.91	367.49	172.11
4	C10H22O	45.18	5.21	193.17	238.04	481.35	356.70	158.10
5	C9H20O	39.41	5.52	175.90	241.23	477.71	354.44	144.09

MS = Molecular structure, *CN* = cetane number, *V* = Viscoscity (cP), *MV* = Molar volume (cm^3/mol), *MP* = Melting point (K), *BP* = Boiling point (K), *FP* = Flash point (K), *MW* = Molecular weight.

In Table 3, the estimated cetane number of potential molecular structure was arranged descending from the top. The first molecular structure has the highest cetane number and this molecule also exists in the reality namely 11-methyl-1-dodecanol. Figure 13 shows the molecular structure of 11-methyl-1-dodecanol:

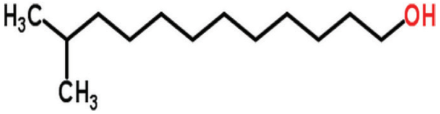


Figure 13: Molecular structure of 11-methyl-1-dodecanol.

The list of its physical properties obtained from RSC Chemical Database and the predicted physical properties were shown in Table 4 below.

Table 4: Comparison between physical properties from different sources.

Physical properties	Predicted values	Database values
Molecular formula	$C_{13}H_{28}O$	$C_{13}H_{28}O$
Average mass	200.13	200.36
Melting point (°C)	–14	–
Boiling point (°C)	257	260.8
Flash point (°C)	113	105.5
Cetane number	60.57	–
Viscosity (cP)	9.86	–
Molar volume (cm^3/mol)	242.4	241

Based on the comparison between the estimated data and the database obtained, it can be concluded that the estimation method gives the almost accurate figure with slight deviation in the results. Besides that, based on the overall set of solutions obtained, it can be assumed that a biodiesel additive is mainly made up from alcohol components.

4. CONCLUSION

In this work, a novel methodology has been developed for property-based product design when the property prediction models for some for the target properties are unavailable. This methodology effectively combined experimental data and property prediction models to develop new models. The hypothesis behind the developed method is that different consumer attributes can be represented as a function of physical properties for which there are models. The factorial design technique has been used to estimate the relationships between properties and to determine the underlying interactions between more than two factors with the target property. The proposed methodology has been applied on the design of a biofuel additive. In this design, a property model based on physical properties has been developed for cetane number and applied into a molecular design problem for the additive design. The developed model can be used in molecular design problem to estimate properties or attributes which cannot be predicted by property prediction methods. In addition, the benefit of the developed methodology can be applied in different scenarios and the property model can be easily developed by using factorial design or other statistical analysis method provided the experimental data of that specific property are available.

The current work can be further extended and improved by including mixture design techniques in the design chemical products. By doing this, it will be possible to accurately track the effect of additives on the products rather than relying on predefined target properties. Another important issue to be addressed in future is the interaction between the attributes themselves. The interaction between attributes using multivariate statistical analysis must be developed in future work in the area of chemical product design.

5. ACKNOWLEDGEMENTS

The financial support from the Ministry of Higher Education Malaysia through the Long-term Research Grant Scheme (LRGS) (Project Code: LRGS/2013/UKM-UNMC/PT/05) is gratefully acknowledged.

6. NOMENCLATURE

η = Number of occurrences of first order groups
μ = Viscosity
GC = Group contribution
M = Molecular weight
Tb = Boiling point
Tb0 = Adjustable parameter for boiling point
Tm = Melting point
Tm0 = Adjustable parameter for melting point
Tf = Flash point
K = Constant
P = Target property
Ni = Number of occurrence of first order group of type-i
Mj = Number of occurrence of second order group of type-j
Ok = Number of occurrence of third order group of type-k
Ci = Contribution of the first order group of type-i
Dj = Contribution of the second order group of type-j
Ek = Contribution of the third order group of type-k
a, b, c, d, e, f = Correlation constants

7. REFERENCES

1. Cussler, E. L. & Moggridge, G. D. (2001). *Chemical product design*. Cambridge: Cambridge University Press.
2. Hada, S. et al. (2012). Product and mixture design in latent variable space by chemometric techniques. *Comp. Aided Chem. Eng.*, 30, 147–151, https://doi.org/10.1016/B978-0-444-59519-5.50030-7.
3. Achenie, L. E. K., Gani, R. & Venkatasubramanian, V. (2003). *Computer aided molecular design: Theory and practice*. Amsterdam: Elsevier Science Inc.
4. Gani, R. (2004). Chemical product design: Challenges and opportunities. *Comput. Chem. Eng.*, 28, 2441–2457, https://doi.org/10.1016/j.compchemeng.2004.08.010.
5. Ambrose, D. (1978). Correlation and estimation of vapour-liquid critical properties: I. Critical temperatures of organic compounds. *Nat. Phys. Lab.*, 92, 1–35.
6. Constantinou, L. & Gani, R. (1994). New group contribution method for estimating properties of pure compounds. *AIChE J.*, 40(10), 1697–1710, http://doi.org/10.1002/aic.690401011.

7. Marrero, J. & Gani, R. (2001). Group-contribution based estimation of pure component properties. *Fluid Phase Equilib.*, 183–184, 183–208, http://doi.org/10.1016/S0378-3812(01)00431-9.
8. Montgomery, D. C. (2013). *Design and analysis of experiments*. New York: John Wiley & Sons.
9. Lewis, S. M. & Dean, A. M. (2001). Detection of interactions in experiments on large numbers of factors. *J. Royal Stat. Soc. Series B Stat. Method.*, 63(4), 633–672, http://doi.org/10.1111/1467-9868.00304
10. Wang, S., Huang, G. H. & Veawab, A. (2013). A sequential factorial analysis approach to characterise the effects of uncertainties for supporting air quality management. *Atmos. Environ.*, 67, 304–312, http://doi.org/10.1016/j.atmosenv.2012.10.066.
11. Tedeschi, L. O. (2006). Assessment of the adequacy of mathematical models. *Agric. Syst.*, 89(2–3), 225–247, http://doi.org/10.1016/j.agsy.2005.11.004.
12. Aggarwal, C. (2013). *Outlier analysis*. New York: Springer.
13. Eljack, F. T. et al. (2008). A property based approach for simultaneous process and molecular design. *Chin. J. Chem. Eng.*, 16(3), 424–434, http://doi.org/10.1016/S1004-9541(08)60100-7.
14. Chemmangattuvalappil. N. G. et al. (2010). Combined property clustering and GC+ techniques for process and product design. *Comput. Chem. Eng.*, 34(5), 582–591, http://doi.org/10.1016/j.compchemeng.2009.12.005.
15. Chemmangattuvalappil. N. G. et al. (2009). A novel algorithm for molecular synthesis using enhanced property operators. *Comput. Chem. Eng.*, 33(3), 636–643, https://doi.org/10.1016/j.compchemeng.2008.07.016.
16. Arvidsson, R. et al. (2011). Life cycle assessment of hydrotreated vegetable oil from rape, oil palm and Jatropha. *J. Clean Prod.*, 19, 129–137, http://doi.org/10.1016/j.jclepro.2010.02.008.
17. Shrestha, B. M. et al. (2014). Change in carbon footprint of canola production in the Canadian Prairies from 1986 to 2006. *Renew. Energy*, 63, 634–641, https://doi.org/10.1016/j.renene.2013.10.022.
18. Xue, J., Grift, T. E. & Hansen, A. C. (2011). Effect of biodiesel on engine performances and emissions. *Renew. Sustain. Energy Rev.*, 15, 1098–1116, https://doi.org/10.1016/j.rser.2010.11.016.
19. Ribeiro, N. M. et al. (2007). The role of additives for diesel and diesel blended (ethanol or biodiesel) fuels: A review. *Energy Fuels*, 21(4), 2433–2445, https://doi.org/10.1021/ef070060r.
20. Nettles-Anderson, S. & Olsen, D. (2009). Survey of straight vegetable oil composition impact on combustion properties. SAE technical paper, SAE International, 1–487, http://doi.org/10.4271/2009-01-0487.

21. Palash, S. M. et al. (2013). Impacts of biodiesel combustion on NOx emissions and their reduction approaches. *Renew. Sust. Energy Rev.*, 23, 473–490, http://doi.org/10.1016/j.rser.2013.03.003.
22. Lapuerta, M., Rodríguez-Fernández, J. & Mora, E. F. d. (2009). Correlation for the estimation of the cetane number of biodiesel fuels and implications on the iodine number. *Energy Policy*, 37(11), 4337–4344, http://doi.org/10.1016/j.enpol.2009.05.049.
23. Conte, E. et al. (2008). Combined group-contribution and atom connectivity index-based methods for estimation of surface tension and viscosity. *Ind. Eng. Chem. Res.*, 47(20), 7940–7954, http://doi.org/10.1021/ie071572w.
24. Constantinou, L., Gani, R. & O'Connell, J. P. (1995). Estimation of the acentric factor and the liquid molar volume at 298 K using a new group contribution method. *Fluid Phase Equilib.*, 103(1), 11–22, http://doi.org/10.1016/0378-3812(94)02593-P.
25. Patil, G. S. (1988). Estimation of flash point. *Fire Mater.*, 12(3), 127–131, http://doi.org/10.1002/fam.810120307.