

# Cetane Number Estimation of Pure Compound using Group Contribution Method

Shah Aznie Ariffin Kashinath, Haslenda Hashim, Azizul Azri Mustaffa, Nor Alafiza Yunus\*

Process Systems Engineering Centre (PROSPECT), Research Institute for Sustainable Environment, School of Chemical and Energy Engineering, Faculty of Engineering, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia  
[alafiza@utm.my](mailto:alafiza@utm.my)

Water-in-diesel emulsion, also known as WIDE fuel, is one of the promising alternative fuels for diesel engines due to its positive impact on the performance and combustion characteristics of the engine, while at the same time reducing the emissions of NO<sub>x</sub> and particulate matter, without the need to modify the diesel engine. Cetane number is a key property that is considered in the formulation of diesel fuel. Cetane number is used to determine the ignition quality of fuel. One of the limitations in designing the diesel fuel formulation is the limited experimental data. The objective of this paper is to develop the cetane number property estimations for a pure compound. A property model was developed using the group contribution approach. In this approach, the molecular structure of the chemical was represented and divided into three levels, namely first-order (220 groups), second-order (130 groups), and third-order (74 groups). The cetane number with the group contribution occurrences of 271 chemicals were regressed using linear regression in MATLAB software to generate the contribution values of the three group levels. The regression step yielded contribution values of 43 groups for the first-order, 35 groups for the second-order, and 7 for the third-order. The coefficient of determination, R<sup>2</sup>, for the cetane number property models was 0.9447, indicating that the proposed model had a good correlation and is reliable to use.

## 1. Introduction

The diesel engine is a type of internal combustion engine, which offers better fuel-to-power conversion efficiency. Many studies have explored the potential of diesel blends or water-in-diesel emulsion fuel as a promising fuel due to its positive impact on the performance, combustion characteristics, and emission factors of the engine. Cetane number (CN) is one of the most significant properties for determining the ignition quality of an engine. According to Eloisa et al. (2011), CN influences the engine start ability, emissions, and peak cylinder pressure and combustion noise. When the diesel fuel mixture or emulsion is formulated, all the chemicals involved during formulation will determine the physical properties of the fuel. Dhinesh and Annamalai (2018) found that by adding cerium oxide nanoparticles in emulsion fuel, the quality of the cetane number of the emulsion fuel was improved. Leng et al. (2018) reported that adding the surfactant to the emulsion improved the cold-flow properties and ignition delay (cetane number) of the diesel fuel. Physical and thermodynamic property prediction models for pure compounds are one of the vital prerequisites for performing tasks such as simulation and optimization and computer-aided molecular/mixture (product) design, especially when the experimental value for the property of the pure compound is not available or limited (Hukkerikar et al., 2012a). In the domain of property prediction models, a few researchers have implemented group-contribution (GC) methods to predict pure compound properties such as open cup flash point (Constantinou and Gani, 1994), melting point, boiling point (Marrero and Gani, 2001), viscosity and surface tension (Conte et al., 2008), lethal concentration, LC50 (Hukkerikar et al., 2012a), and heat of combustion (Yunus and Zahari, 2017). All these proposed models are generally suitable to obtain the needed property values since these methods provide the advantage of quick estimates without requiring substantial computational work. In GC methods, the property of a component is a function of structurally dependent parameters, which are determined as a function of the frequency of the groups that represent the molecules

and their contributions (Hukkerikar et al., 2012b). The GC method has proven able to provide a good prediction and only requires chemical structure as input (Yunus and Zahari, 2017). Due to its predictive capability, the GC method was considered for the CN estimation in this study. The objective of this paper is to propose a CN property model using the group contribution (GC) method. In the group-contribution approach, the molecular structure of the chemical is divided into a set of functional groups, where each group contributes to the value of the property (Conte et al., 2008).

## 2. Methodology

There are five steps involved in the CN property estimation, as illustrated in Figure 1 below:

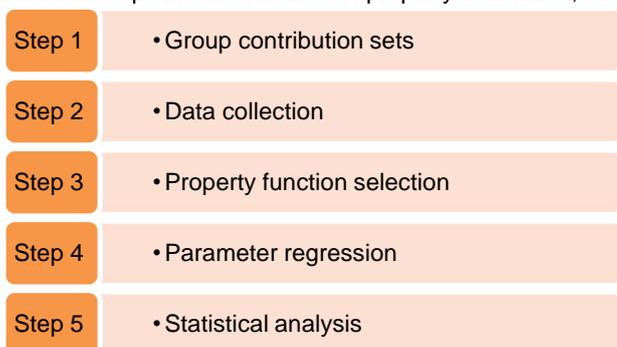


Figure 1: Methodology for cetane number (CN) property estimation

### 2.1 Group contribution sets

In this study, the functional group of the chemical was defined according to Marrero and Gani, (2001). This definition was employed to predict the variety of properties such as critical temperature, critical pressure, the Standard Gibbs energy, and the Standard Enthalpy of Vaporization covering more than 2,000 compounds. For the GC method, generally, the property estimation of chemicals is defined as three levels, namely first-order, second-order, and third-order groups. The basic (first) level uses the contributions of first-order groups that describe a wide variety of organic compounds. The higher (second) level provides additional structural information, which is not covered by the first-order groups, and thus provides corrections to the estimation at the first-level. The final level (third-order) provides the adjustment to the prediction made from the first and second level, where the contributions from the structure of complex molecules are calculated (Conte et al., 2008). More detailed information and distribution of each level are given in Figure 2.

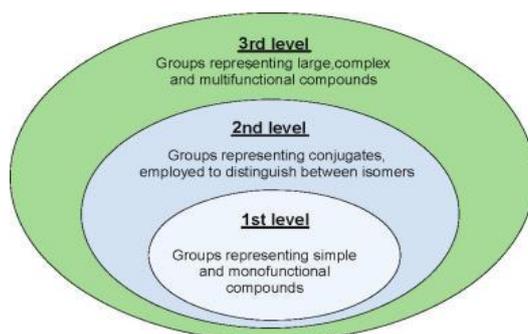


Figure 2: The description of the multilevel approach for the group contribution method (Conte et al., 2008).

Eq(1) below, as given by Marrero and Gani (2001), was used to represent the general form of the function  $f(X)$  of the target property  $X$  for the property estimation model:

$$f(X) = \sum_i N_i C_i + \sum_j M_j D_j + \sum_k O_k E_k \quad (1)$$

Where  $C_i$ ,  $D_j$  and  $E_k$  are the contribution values for the first-order, second-order, and third-order groups with  $N_i$ ,  $M_j$ , and  $O_k$  being the occurrences of each group.

## 2.2 Data collection

A CN dataset was collected as the initial database from Yanowitz et al., (2017) Compendium of Experimental Cetane Numbers, which contains 333 chemicals from the common families (alkanes, alkenes, ethers, esters, aldehydes, ketones, alcohols, and furan). The chemical structures of the compounds were identified and defined for the first-, second- and third-order levels according to Marrero and Gani (2001). Also, the occurrences of the chemicals were collected using ICAS software. The occurrences of the group contribution for all chemicals were used for the model regression in Step 4.

## 2.3 Property function selection

The property function was selected based on the CN trend observed in the data collected in Step 2. This function must show the best possible fit with the experimental data and should also provide good extrapolation capability. The collected experimental data on CN was plotted against the occurrences of the CH<sub>2</sub> group for various families of compounds to identify the best model for CN property. The resulting trend from the data collected shows that the property function is a linear function of the CN property function. Hence, the CN model is represented by Eq(2):

$$CN = CN_0 + \sum_i N_i C_i + \sum_j M_j D_j + \sum_k O_k E_k \quad (2)$$

## 2.4 Parameter regression

The Levenberg-Marquardt method was selected for the regression step to minimize the sum of squares of the differences between the experimental and estimated values of the CN property, as per the method outlined in Conte et al., (2008) for the parameter regression of surface tension and viscosity. The regression step was done using MATLAB. For Eq(2), CN is the cetane number of the chemical and CN<sub>0</sub> is the universal constant for the model. The contribution values of the contribution groups  $C_i$ ,  $D_j$ , and  $E_k$  was determined in three steps. The first step involved determining the universal constant,  $CN_0$ , and the contribution value of the first-order groups. The results of the universal constant and the group contribution value of the first-order group were used to determine the contribution value of the second-order group,  $D_j$ . The final step was the regression of the contribution value for the third-order group,  $E_k$ , using the results of the first and second steps. The final results of the regression step, which consists of the universal constant and the contribution values of the three levels, were analyzed to identify the outliers in the experimental data. The identified outliers were removed and the GC model parameters regressed again. Following that, 35 chemicals were identified as the outliers, as these did not follow the average trend. Meanwhile, 298 chemicals that fulfilled the trend were used for parameter regression. The training and testing data were divided randomly, with 271 chemicals as the training set and 26 chemicals as the testing set. The testing step is known as the validation step to verify the capability of the model.

## 2.5 Statistical analysis

The statistical analyses employed in this study are the Standard Deviation (SD), the Relative Deviation (RD), the Average Absolute Error (AAE), and the Average Relative Error (ARE), as defined by Eq(3) to Eq(6). All these equations are commonly used to verify the property model developed using the GC-based method (Marrero and Gani, 2001).

$$SD = \sqrt{\frac{\sum (\theta_i^{est} - \theta_i^{exp})^2}{N}} \quad (3)$$

$$RD = \frac{|\theta_i^{est} - \theta_i^{exp}|}{\theta_i^{exp}} 100 \quad (4)$$

$$AAE = \sum \frac{|\theta_i^{est} - \theta_i^{exp}|}{N} \quad (5)$$

$$ARE = \frac{\sum RD_i}{N} \quad (6)$$

Where  $N$  is the number of experimental data and  $\theta_i^{est}$  and  $\theta_i^{exp}$  are the predicted cetane number and experimental cetane number. The results of the statistical analysis are shown in Section 3 below, where a good prediction model is demonstrated with an  $R^2$  value close to unity.

### 3. Results and discussion

The results of the contribution values are shown in Tables 1, 2, and 3 for the first-order, second-order, and third-order groups. From these tables, it can be concluded that the regression of the experimental data generated (43, 35, and 7) group contribution values for the first, second, and third-order groups. It is important to highlight that all the groups listed in Tables 1 to 3 followed the functional groups presented in Marrero and Gani, (2001) for the first-, second- and third-order levels. The results in Tables 1 to 3 were used to predict the CN of the chemicals in the training dataset. The plots of the estimated values of CN were then compared to the experimental training and testing data, as shown in Figure 3. The model with the three-level groups predicted the CN accurately, with an  $R^2$  value equal to 0.9447 and 0.926 for the training set and the testing set. The  $CN_0$  value based on Eq(2) is 16.04.

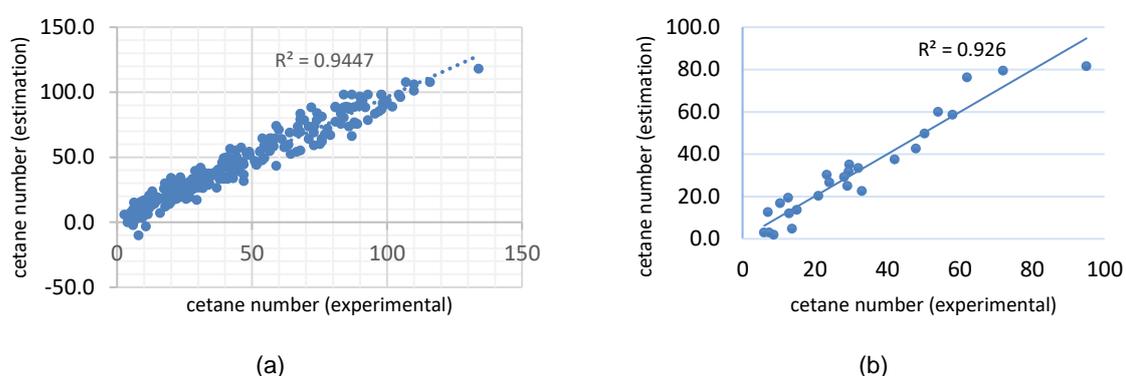


Figure 3: Estimated CN versus experimental CN for (a) training data and (b) testing data

Table 1: Contribution values of the first-order group

No.	Group	$C_i$	No.	Group	$C_i$
1	CH <sub>3</sub>	3.49	23	CHCO	-11.296
2	CH <sub>2</sub>	4.879	24	CHO	20.867
3	CH	-3.7	25	CH <sub>3</sub> COO	-5.795
4	C	-7.296	26	CH <sub>2</sub> COO	-12.851
5	CH <sub>2</sub> =CH	-1.174	27	aC-COO	-4.005
6	CH=CH	-8.088	28	COO	-6.28
7	CH <sub>2</sub> =C	-2.623	29	CH <sub>3</sub> O	15.253
8	CH=C	-3.646	30	CH-O	-2.696
9	aCH	-2.76	31	aC-O	-0.385
10	aC	0.09	32	OCH <sub>2</sub> CH <sub>2</sub> OH	-0.832
11	aC	-3.848	33	OCH <sub>2</sub> CHOH	-4.023
12	aC	-16.722	34	CH <sub>2</sub> cyc	1.557
13	aC-CH <sub>3</sub>	0.49	35	CHcyc	-1.354
14	aC-CH <sub>2</sub>	-3.601	36	Ccyc	-5.537
15	aC-CH	-11.132	37	CH=CHcyc	-2.253
16	aC-C	-23	38	CH=Ccyc	-4.766
17	aC-CH=CH <sub>2</sub>	-22.999	39	CH <sub>2</sub> =Ccyc	1.047
18	OH	-13.769	40	O	1
19	aC-OH	0.596	41	CO	-7.022
20	COOH	-21.342	42	-O-	26.861
21	CH <sub>3</sub> CO	-5.334	43	Ccyc=C	-3.811
22	CH <sub>2</sub> CO	-8.401			

Table 2: Contribution values of the second-order group

No.Group	$D_j$	No.	Group	$D_j$
1 (CH <sub>3</sub> ) <sub>2</sub> CH	2.010	19	aC-CH <sub>n</sub> -O- (n in 1..2)	-2.804
2 (CH <sub>3</sub> ) <sub>3</sub> C	-3.767	20	aC-CH(CH <sub>3</sub> ) <sub>2</sub>	4.235
3 CH(CH <sub>3</sub> )CH(CH <sub>3</sub> )	-7.785	21	(CH <sub>n</sub> =C)cyc-CH <sub>3</sub> (n in 0..2)	2.141
4 CH(CH <sub>3</sub> )C(CH <sub>3</sub> ) <sub>2</sub>	-7.836	22	(CH <sub>n</sub> =C)cyc-CH <sub>2</sub> (n in 0..2)	-4.525
5 CH <sub>n</sub> =CH <sub>m</sub> -CH <sub>p</sub> =CH <sub>k</sub> (k,m,n,p in 0..2)	4.047	23	CHcyc-CH <sub>3</sub>	0.025
6 CH <sub>3</sub> -CH <sub>m</sub> =CH <sub>n</sub> (m,n in 0..2)	-0.199	24	CHcyc-CH <sub>2</sub>	0.384
7 CH <sub>2</sub> -CH <sub>m</sub> =CH <sub>n</sub> (m,n in 0..2)	-0.595	25	CHcyc-CH	-2.602
8 CH <sub>p</sub> -CH <sub>m</sub> =CH <sub>n</sub> (m,n in 0..2; p in 0..1)	0.252	26	CHcyc-C	1.716
9 CHCHO or CCHO	-21.095	27	CHcyc-C=CH <sub>n</sub> (n in 1..2)	-1.943
10 CH <sub>3</sub> COCH <sub>2</sub>	-1.461	28	CHcyc-OH	2.654
11 CHCOOH or CCOOH	12.637	29	Ccyc-CH <sub>3</sub>	0.043
12 CH <sub>3</sub> COOCH or CH <sub>3</sub> COOC	8.058	30	AROMRINGs1s2	0.226
13 CHOH	-1.507	31	AROMRINGs1s3	0.618
14 COH	5.850	32	AROMRINGs1s4	-5.665
15 COO-CH <sub>n</sub> -CH <sub>m</sub> -OOC (n, m in 1..2)	-3.885	33	AROMRINGs1s2s3	0.867
16 OOC-CH <sub>m</sub> -CH <sub>m</sub> -COO (n, m in 1..2)	10.422	34	AROMRINGs1s2s4	0.717
17 CH <sub>m</sub> -O-CH <sub>n</sub> =CH <sub>p</sub> (m,n,p in 0..3)	-0.036	35	AROMRINGs1s3s5	-2.293
18 CH <sub>n</sub> =CH <sub>m</sub> -COO-CH <sub>p</sub> (m,n,p in 0..3)	-2.349			

Table 3: Contribution values of the third-order group

No .	Group	$E_k$
1	aC-CH <sub>n</sub> cyc (fused rings) (n in 0..1)	4.569
2	CH multiring	0.183
3	C multiring	-2.779
4	aC-CH <sub>n</sub> -O-CH <sub>m</sub> -aC (different rings) (n,m in 0..2)	5.608
5	AROM.FUSED[2]	-9.330
6	AROM.FUSED[2]s1	8.782
7	AROM.FUSED[2]s2	9.603

Table 4 shows the statistical results of the regression procedure for Step 1 to Step 3. The statistical parameter values ( $R^2$ , AAE, ARE, and SD) are given for all three levels of the compounds in the dataset containing 271 chemicals. The result better predicted the property if the regression step considered all three group contribution values to predict CN, as proven by the improvement in the  $R^2$  values and the reduced standard deviation.

Table 4: Comparison of the statistical analysis of training data for the contribution of all group levels

Statistical analysis	1 <sup>st</sup> order	1 <sup>st</sup> and 2 <sup>nd</sup> order	1 <sup>st</sup> , 2 <sup>nd</sup> and 3 <sup>rd</sup> order
$R^2$	0.9348	0.9445	0.9447
Average absolute error, AAE	5.930	5.4206	5.39
Average relative error, ARE	23.10	20.83	20.48
Standard deviation, SD	7.482	6.904	6.8933

To prove its capability, the model was tested using a testing dataset. The contribution values from Table 1 to Table 3 were used to predict the CN of the chemicals in the testing dataset. A comparison between the experimental CN and the estimated CN for 26 compounds is shown in Figure 3b. The  $R^2$  value of the predicted CN for the testing dataset was 0.926. Therefore, the predictive performance of the model is acceptable, at least for these compounds. Table 5 shows an example of the calculation of the CN value of 1-decanol using the developed property model. For reference, the experimental value of CN for 1-decanol is 50.3. By using this model, the prediction value of 1-decanol returned a value of 49.7. Therefore, the model developed using the group contribution method can accurately predict the cetane number (CN) of the compound.

