PREDICTING THE CETANE NUMBER OF FURANIC BIOFUEL CANDIDATES USING AN IMPROVED ARTIFICIAL NEURAL NETWORK BASED ON MOLECULAR STRUCTURE

Travis Kessler
University of Massachusetts Lowell
Lowell, Massachusetts, United States

Eric R. Sacia
University of California at Berkeley
Berkeley, California, United States

Alexis T. Bell
University of California at Berkeley
Berkeley, California, United States

J. Hunter Mack
University of Massachusetts Lowell
Lowell, Massachusetts, United States

ABSTRACT

The next generation of alternative fuels is being investigated through advanced chemical and biological production techniques for the purpose of finding suitable replacements to diesel and gasoline while lowering production costs and increasing process yields. Chemical conversion of biomass to fuels provides a plethora of pathways with a variety of fuel molecules, both novel and traditional, which may be targeted. In the search for new fuels, an initial, intuition-driven evaluation of fuel compounds with desired properties is required. Due to the high cost and significant production time needed to synthesize these materials at a scale sufficient for exhaustive testing, a predictive model would allow chemists to preemptively screen fuel properties of potentially desirable fuel candidates. Recent work has shown that predictive models, in this case artificial neural networks (ANN's) analyzing quantitative structure property relationships (QSPR's), can predict the cetane number (CN) of a proposed fuel molecule with relatively small error. A fuel's CN is a measure of its ignition quality, typically defined using prescribed ASTM standards and a cetane testing engine. Alternatively, the analogous derived cetane number (DCN), obtained using an Ignition Quality Tester (IQT), is a direct measurement alternative to the CN that uses an empirical inverse relationship to the ignition delay found in the constant volume combustion chamber apparatus. DCN data points acquired using an IQT were utilized for model validation and expansion of the experimental database used in this study. The present work improves on an existing model by optimizing the model architecture along with the key learning variables of the ANN and by making the model more generalizable to a wider variety of fuel candidate types, specifically the class of furans and furan derivatives, by including specific molecules for the model to incorporate. The new molecules considered include tetrahydrofuran, 2-methylfuran, 2-methyltetrahydrofuran, 5,5'-(furan-2-ylmethylene)bis(2-methylfuran), 5,5'-((tetrahydrofuran-2-yl)methylene)bis(2-methyltetrahydrofuran), tris(5-methylfuran-2-yl)methane, and tris(5-methyltetrahydrofuran-2-yl)methane. Model architecture adjustments improved the overall root-mean-squared error (RMSE) of the base database predictions by 5.54%. Additionally, through the targeted database expansion, it is shown that the predicted cetane number of the furan-based molecules improves on average by 49.21% (3.74 CN units) and significantly for a few of the individual molecules. This indicates that a selected subset of representative molecules can be used to extend the model's predictive accuracy to new molecular classes. The approach, bolstered by the improvements presented in this paper, enables chemists to focus on promising molecules by eliminating less favorable candidates in relation to their ignition quality.

INTRODUCTION

Research into next-generation alternative fuels has gained significant interest due to concern over global warming, decreasing reserves of conventional fossil fuels, and drawbacks associated with first-generation biofuels like corn ethanol. Biofuels are typically derived from renewable sources such as sugars, starch and vegetable oil; however, the oxygenated functional groups in biofuel molecules add an additional layer of complexity over traditional hydrocarbons. Though these fuels offer many benefits, especially when derived from cellulosic biomass, the next-generation of biofuels have proven challenging to produce at scale cost-effectively. Providing predictive insight into key properties, such as the cetane number, can accelerate the development of new alternative fuels. By shortening the feedback loop inherent to research, scientists can quickly identify the most promising compounds and focus on increasing yield and decreasing costs.

Cetane Number

One of the most important parameters for evaluating a fuel for use in a diesel engine is the cetane number (CN), a measure of the fuel's ignition quality. It is a correlation based on ignition delay from the start of injection and includes both physical (vaporization) and chemical components. There are two widely used methods in determining CN, either using a Cooperative Fuel Research (CFR) engine or an Ignition Quality Tester (IQT). Experimental determination of CN using the single-cylinder CFR is specified through American Society for Testing and Materials (ASTM) Standard D613 [1]. An alternative approach uses the IQT test procedure, which is specified in the ASTM D6890 standard [2]. The method determines the ignition delay in a constant volume combustion chamber by measuring the time between the start of fuel injection and the onset of combustion. Both methods provide accurate CN measurements, although the CN obtained on the CFR is preferred since it reflects combustion behavior in an actual engine. Furthermore, the correlation between DCN and CN is based on an empirical relation and has limited accuracy when used across a range of fuels [3]. The potential impact of this limitation in relation to novel fuels has not been fully characterized. However, the IQT offers a distinct advantage in terms of increased speed and lower volumetric requirements, typically about 100 mL.

Even with the advantages provided by the IQT, the sheer number of potential fuel molecules makes testing prohibitive in terms of both cost and time. This reinforces the need for a rapid and robust screening method for predicting CN, and potentially other properties, in order to aid in alternative fuel development.

Predicting the Cetane Number

Predicting cetane numbers and other fuel properties from molecular structure has an extensive history. Prior models based on quantitative structure property relationships (QSPR) have been developed to predict the CN of different compounds, which included an early, but limited, application of backpropagating neural networks for predicting the CN of isoparaffins and diesel fuels [4]. Though the study was limited to branched paraffins, the model showed a superior predictive power compared to conventional equations [5]. A subsequent study used QSPR software to generate 100 molecular descriptors for a set 275 compounds including 147 hydrocarbons and 128 oxygenates [6]; a genetic algorithm, or a search heuristic mimicking natural selection in regards to optimization problems, was used to identify which descriptors might influence CN. Although the model did not accurately predict CN (RMSE = 9.1 CN units), the work served as a basis for future models focused on predicting CN using QSPR inputs.

Other types of models have been used to predict CN. One approach utilized an inverse function method to predict the CN of pure hydrocarbons [7]. Though the model is accurate for the range of compounds considered, it is unable to predict the CN of compounds outside the test range. A recent model considered chemical families likely found in diesel fuels using the genetic function approximation (GFA), an iterative approach to generate relationships between molecular descriptors and CN [8]. Though the approach could not satisfactorily predict CN when including all 147 molecules in the data set, it utilized an approach of dividing the set into four different groups based on their chemical families to improve the model's predictive power. The method provides a sufficient local predictive tool for compounds within the same chemical family, but is unable to extend predictions to a larger and disparate data set.

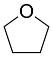
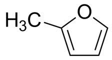
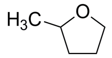
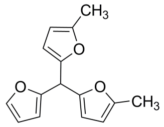
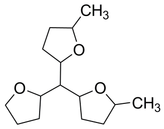
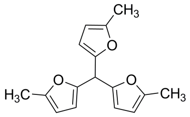
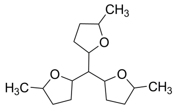
Another recent model extended the applicability to include alcohols and esters using "consensus" modeling, which averaged results from the outputs of various linear and nonlinear models (including neural networks) [9]. The approach considered 279 compounds from 7 chemical families and predicted CN with a RMSE of 6.3.

In light of the advances and drawbacks inherent to previous models, this paper adopts a backpropagation neural network approach since it appears to be more robust across multiple molecular classes/families due to their nonlinear architecture, which allows for a representation of very complex relationships between input and output vectors [10]. The goal of this paper is two-fold: (1) improve upon the state-of-the art models for predicting CN for a diverse data set, and (2) extend the model to consider a new molecular class (furanic compounds). The model's accuracy in regards to the furanic compounds can be compared for two cases, without new experimental data and with new experimental data. As a model's predictive power is only as good as the input data, it is expected that the inclusion of some new furanic compounds will increase the accuracy of the model without affecting the overall RMSE. Ultimately, the objective is to minimize the RMSE as much as possible. In practice, it was found that an RMSE of about 6 is acceptable.

Furanic Biofuels

Many strategies exist for converting the sugar units produced by biomass via photosynthesis into fuels. One particularly attractive method is to generate furan derivatives through acid-catalyzed dehydration reactions. Using this method, sugars containing five carbon atoms, such as the xylose sub-units that compose the hemicellulose portion of lignocellulosic biomass, can be converted into furfural, and sugars containing six carbon

Table 1. Investigated Furanic Compounds

Entry	Compound	Formula	T_{fp} [^oC]	ν_{40^oC} [$mm^2\ s^{-1}$]	Lubricity [mm]	ρ_{40^oC} [$g\ mL^{-1}$]	ΔH_{comb} [$MJ\ L^{-1}$]	Structure
1	tetrahydrofuran	C_4H_8O	-108.4	-	-	0.883	-	
2	2-methylfuran	C_5H_6O	-89	-	-	0.91	-	
3	2-methyltetrahydrofuran	$C_5H_{10}O$	-136	-	-	0.854	-	
4	5,5'-(furan-2-ylmethylene)bis(2-methylfuran)	$C_{15}H_{14}O_3$	11	11.8	160	1.102	35.9	
5	5,5'-((tetrahydrofuran-2-yl)methylene)bis(2-methyltetrahydrofuran)	$C_{15}H_{26}O_3$	< -40	7.45	180	1.007	35.6	
6	tris(5-methylfuran-2-yl)methane	$C_{16}H_{16}O_3$	31	18.1	160	1.086	35.9	
7	tris(5-methyltetrahydrofuran-2-yl)methane	$C_{16}H_{28}O_3$	< -40	7.33	220	0.983	1.007	

atoms, such as fructose or the glucose sub-units that are present in starches and cellulose, can be converted into 5-hydroxymethyl furfural (HMF) [11-13].

A popular reason for targeting furfural and HMF as fuel intermediates is that they provide useful molecular functionalities to continue to upgrade these molecules in high yields to produce increasingly valuable fuels. Such coupling reactions are especially critical in growing diesel markets to meet the volatility specifications of existing fuels. The scope of available reaction pathways from these furan derivatives is enormous. For the purposes of this work, the set of biomass derivatives shown in Table 1 containing furan and tetrahydrofuran rings will be referred to as furanic compounds. Decarbonylation of furfural produces furan, which may be readily hydrogenated to produce tetrahydrofuran (Table 1, Entry 1). Selective hydrogenation of furfural produces 2-methylfuran (Table 1, Entry 2) [14], a valuable chemical intermediate, which may further be hydrogenated to 2-methyltetrahydrofuran (Table 1, Entry 3). 2-methyltetrahydrofuran, a gasoline additive, may also be formed from levulinic acid, another product of sugar dehydration reactions, and has been approved as a gasoline blend component by the United States Department of Energy [11].

The listed furanic compounds, along with many other available molecules, produce a suite of intermediates that may readily be combined via acid-catalyzed electrophilic aromatic substitution [14-15], base-catalyzed aldol condensation [13], or acid-catalyzed etherification [16], among others. The coupling of two 2-methylfuran molecules by furfural or 5-methyl furfural via electrophilic aromatic substitution produces the molecules shown in Entries 4 and 6, respectively in Table 1 [14-15]. These reactions can occur at mild conditions with >90% selectivity with no reaction solvent, making them quite attractive [15]. The fuel value of these particular molecules has been previously reported and is intuitively expected to be low since the aromatic furan rings can stabilize radicals during the combustion process, slowing the combustion reactions and lowering the cetane number [17]. The aromaticity also causes π-stacking, raising the melting point of these pure components. A selective hydrogenation of the molecules shown in Entries 4 and 6 of Table 1 has been shown to produce Entries 5 and 7, respectively [17]. After the reaction, the tetrahydrofuran rings on the products are no longer aromatic, leading to substantially improved fuel characteristics, including high cetane numbers, very good lubricity, and good cold flow properties [17].

While prior work has shown schemes for hydrodeoxygenation of the molecules in Entries 4 and 6 of Table 1 completely to traditional hydrocarbon alkanes [14], doing so imparts an additional cost by requiring 55-60% more hydrogen in the overall process than producing the tetrahydrofuranyl analogs in Entries 5 and 7. Therefore, improvement of predictive cetane methods to include the scope of oxygenated fuels, especially those with furan and tetrahydrofuran rings, will assist in directing the path of fuel research toward novel targets, instead of solely to more traditional hydrocarbon products.

METHODS

The cetane number data used for the core data set was obtained from sets found in the NREL Compendium of Experimental Cetane Number Data [18] and other sources [6, 9]. It contains 284 molecules in total. The NREL Compendium, used as the primary source of data, lists experimental cetane number values attained from multiple methods, including CFR, IQT, octane-to-cetane correlations, and blend measurements. Values from the latter two approaches are less accurate (+/- 5 CN units). Furthermore, multiple compounds tested have numerous reported values within sizable ranges. Therefore, the quality of the reported data used for the core data set limits the accuracy of the predictive model. CN data for the set of furanic compounds was determined using an Ignition Quality Tester (IQT) using ASTM Standard D6890. The evaluations were conducted at Intertek (San Francisco, California, USA) and the National Renewable Energy Laboratory (Golden, Colorado, USA).

As opposed to other models, we have chosen not to eliminate any data from the core data set based on concerns with the provided experimental data. It has already been shown that carefully eliminating data seen as questionable can improve a model, but the overarching goal is a fully generalizable model based on all available experimental data. Therefore, all experimental data in the core set is retained and targeted reductions are not implemented at this point.

Compound structures were first converted to SMILES (Simple Molecular-Input Line-Entry System) using MarvinSketch (ChemAxon Ltd.) [19]. SMILES structures were then converted to 2-D structures using the NCI online calculator [20], which allowed for the generation of 1667 QSAR molecular descriptors using e-Dragon [21]. A database for the core data set and new furanic compound data set were constructed using these descriptors and their known cetane numbers.

The number of input parameters was reduced from 1667 to 15 using an iterative regression analysis technique. This was done to reduce build-time of the neural networks while retaining low error. For each parameter, networks regressed using only that single parameter. The parameter that produced the lowest average RMSE was retained, and the next trial was run using this parameter plus each of the remaining parameters, until a list of 15 parameters was obtained. This was considered one run. It is worth noting that each run has independent learning/validation/testing data splitting. A total of 15 additional runs were completed, and the most frequently chosen parameters at all 15 intervals across 15 runs were taken to be the final parameters. This was done to ensure an accurate representation of parameters for multiple data splits. Regression using all of the parameters yields useful insight into the large amount of covariance between the possible inputs.

Figure 1 shows the results of the parameter reduction. Between 15-25 parameters, RMSE does not improve significantly. As the list of included parameters is increased further, RMSE begins to rise. This is due to the fact that many of the values for some parameters are equal for the majority of molecules, which is detrimental to the neural network and unhelpful in capturing the nonlinear behavior. Historically, 14-23 descriptors have been chosen for similar approaches in the literature. The process is repeatable across multiple attempts, which suggests that the chosen set is likely to be the most influential descriptors in regards to CN prediction.

A closer look at the included descriptors can give some insight into how a molecule's geometry might influence a property like CN. Table 2 lists the definitions of each of the descriptors retained after the reduction process. Additional information on each of the descriptors can be found in the literature [22-23].

While these descriptors are relatively nuanced, others such as nROR, (number of ethers), nROH (number of hydroxyl groups), and nOHp (number of primary alochols) are more physical and align with the foundation of chemical kinetics in combustion.

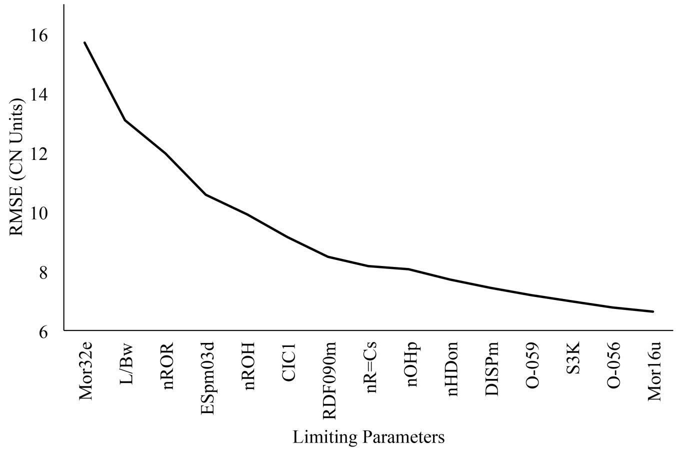


Figure 1. Iterative Addition of Descriptors to Model

Using artificial neural networks (ANN's) implemented in Python, a regression analysis of the core data set was performed using Levenberg-Marquardt backpropagation involving stochastic gradient descent, a common learning technique for ANN's [24]. The optimization function used for the regression was the mean squared error function, where the network converges to the point of least error relative to the core data set as a whole. The model architecture, shown in Figure 2, includes input data (the 15 retained molecular descriptors), two hidden layers of 32 neurons each, and a single output (CN). Two hidden layers, rather than one, are used in order to capture the highly nonlinear relationship between QSPR descriptors and CN.

Table 2. Glossary of Descriptor Terminology

Descriptor	Definition
Mor32e	Signal 32 / Weighted by Sanderson electronegativity
ESpm05u	Spectral moment of order 2 from edge adjacency mat.
CIC1	Complementary Information Content Index (neighborhood symmetry of 1-order)
RDF035u	Radial Distribution Function - 035 / unweighted
nROR	Number of ethers
nROH	Number of hydroxyl groups
L/Bw	Length-to-breadth ratio by WHIM
RDF090m	Radial Distribution Function - 090 / weighted by mass
nHDon	Number of donor atoms for H-bonds (N and O)
RDF020p	Radial Distribution Function - 020 / weighted by polarizability
nOHp	Number of primary alcohols
EEig08x	Eigenvalue n. 8 from augmented edge adjacency mat. weighted by bond order
O-059	Al-O-Al / Atom-centered fragments
G3s	3rd component symmetry direction WHIM index / weighted by l-state
GATS8m	Geary autocorrelation of lag 8 weighted by mass

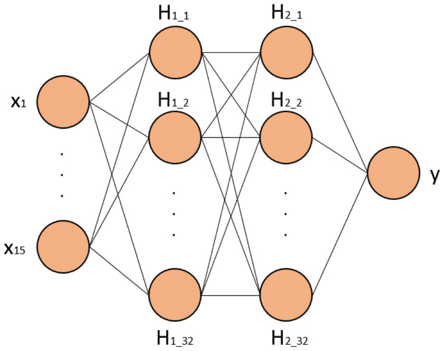


Figure 2. Model architecture including inputs (x_1-x_{15}), two hidden layers of 32 neurons, and an output y.

Each ANN randomly assigned each molecule of the core data set to one of three conditions: learning, validation, and testing, with proportions of 65%, 25%, and 10% respectively. The testing proportion of the data was used to evaluate the final generalizability of the network after training. The ANN was trained on the learning proportion of the data set until there was no significant improvement in the performance of the validation proportion. This cutoff point was determined by the mean-delta-root-mean-squared-error (mdRMSE) falling below a predetermined threshold value. The mdRMSE represents the mean value of the change in RMSE of the validation data between learning epochs (iterations), and approaches zero as the number of learning epochs increases. Final performance of the ANN is determined by the overall RMSE of the ANN when tested on the entire core data set. A lower overall RMSE indicates a more optimized ANN.

Using random learn/validate/test splitting increased the number of ANN's needing to be constructed to achieve an accurate final network. This ultimately provided greater accuracy due to the ANN being able to choose what it learns, allowing itself to determine the learning set that provides the least error. Due to the learning data being randomly chosen, it is possible that an optimal ANN may not be completely representative of the entire database in regard to compound types. Hand-picking learn/validate/test sets may reduce the number of ANN's that need to be built, however the accuracy of the ANN would be questionable without an enhanced selection technique.

The architecture of the final predictive model (build set) consists of averaged results from the best five ANN's from five nodes. Each node of the build set was subject to 75 trials, where each trial was an independent ANN. From each node, the best performing trial was selected based on the previously listed criteria. Averaging predictions across five ANN's decreases the overall RMSE of the core data set. Because each ANN was trained with different learning data splits, each ANN tends to predict CN values slightly different than the others; either slightly higher or slightly lower than the desired CN for some molecules. When the predictions of five ANN's are averaged, the average result tends to be closer to the desired CN, lowering the overall RMSE. A simulation diagram illustrating the construction of build sets is shown in Figure 3.

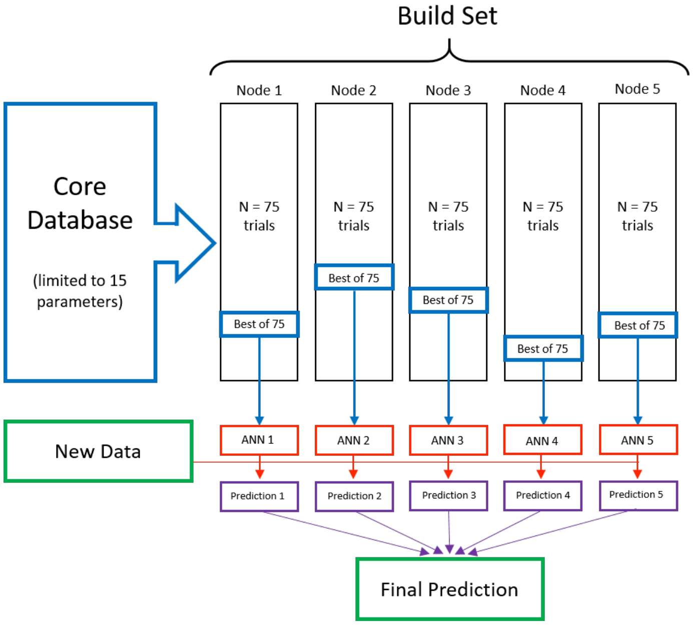


Figure 3. Simulation diagram of the build set construction procedure

Table 3. Summary of results for experimental and predicted CN

Compound	CN (Experimental)	Predicted CN (raw database)	Predicted CN (exp. database)	Error (raw)	Error (expanded)
tetrahydrofuran	26.8	31.09	29.72	4.29	2.92
2-methylfuran	8.30	4.00	4.41	4.30	3.89
2-methyltetrahydrofuran	20.5	22.40	20.53	1.90	0.03
5,5'-(furan-2-ylmethylene)bis(2-methylfuran)	25.5	6.72	17.98	18.78	7.52
5,5'-((tetrahydrofuran-2-yl)methylene)bis(2-methyltetrahydrofuran)	60.4	57.00	57.76	3.40	2.64
tris(5-methylfuran-2-yl)methane	22.3	9.99	17.57	12.31	4.73
tris(5-methyltetrahydrofuran-2-yl)methane	59.8	51.61	54.48	8.19	5.32

RESULTS & DISCUSSION

Cetane numbers for the seven furanic compounds included in this study were predicted using the core data set as inputs and the model architecture outlined above. Figure 4 depicts a parity plot of experimental CN versus predicted CN for all molecules using the core data set. The solid line indicates parity (perfect prediction) and the dashed lines specify the total RMSE of the model. Predictions for the core data are shown as crosses, while predictions for the furanic compounds are shown as solid circles. The overall RMSE of the model based on the core data set was 5.97 CN units, showing an improvement of 5.54% (0.35 CN units) over prior efforts. It is apparent that some of the furanic compounds fall well outside the RMSE bounds of the model. The average absolute error between predicted and experimental CN was 7.60 CN units for the model based on the core data set, well outside the RMSE of the model. The maximum absolute error was 18.78 CN units for 5,5'-(furan-2-ylmethylene)bis(2-methylfuran). This relatively high overall error is due to the absence of furanic compounds in the core data set, and hence the absence of furanic compounds during the learning processes. The absence of furanic compounds in the learning processes limits the model's accuracy in regards to predicting the cetane number of some furanic compounds.

Next, an expanded data set was created by adding experimental results for six of the seven furanic compounds to the core data. The remaining furanic compound was then predicted using a new model based on the expanded data set. This was done to attain a "blind" prediction of the compound left out, as the predictive model had no exposure to this compound in during the learning processes. As motivated in the introduction, the inclusion of additional similar molecules to the input data set should improve the generalizability of the model to other furanic compounds. The descriptor reduction step shows that the retained descriptors used in the model do not change between the core and expanded databases. This also makes sense intuitively; with a core data set of 284 molecules, adding only six would not change the descriptors used to predict all 290. A parity plot of experimental CN versus predicted CN for a model based on the expanded data set is shown in Figure 5. The total RMSE of the model improves slightly to 5.95 CN units. More importantly, the average absolute error between experimental and predicted cetane numbers for the furanic compounds improved to 3.86 CN units, with a maximum absolute error of 7.52 CN units. This represents an improvement of 49.21% when using the expanded data set over the core data set. This validates the hypothesis that a targeted expansion of the input data set can extend the applicability of the model to new molecular classes.

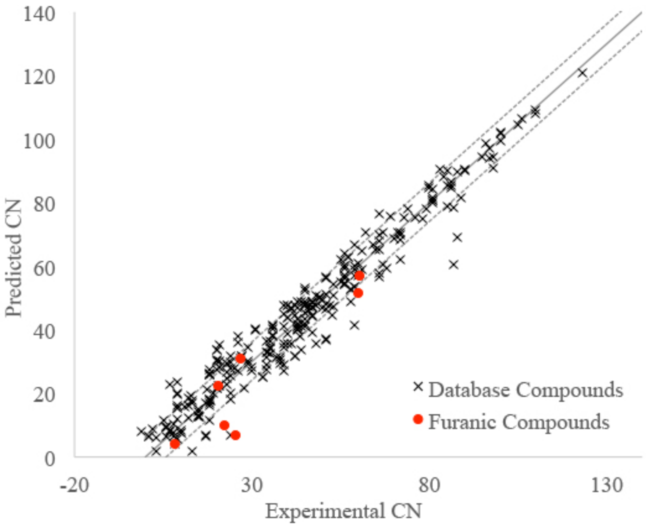


Figure 4. Parity plot of cetane numbers for all compounds in the core data set. Solid line indicates parity; dashed line indicates RMSE (5.97 CN units).

A summary of the individual results for the furanic compounds included this study is shown in Table 3. It is worth noting that the error is defined as the magnitude of the difference between the predicted cetane number and the experimental cetane number. As defined by the expanded set's RMSE (5.95), the desired tolerance is also equal to 5.95 CN units. It can be seen that all compounds improved when the model was based on the expanded data set. However, some molecules experienced a greater improvement in the predicted CN than others.

Predictions for 5,5'-(furan-2-ylmethylene)bis(2-methylfuran) and tris(5-methylfuran-2-yl)methane were the most pronounced.

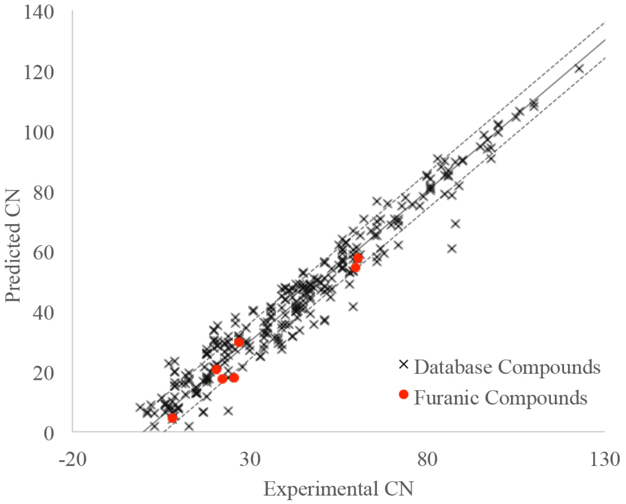


Figure 5. Parity plot of cetane numbers for all compounds using the expanded data set. Solid line indicates parity; dashed line indicates RMSE (5.95 CN units).

CONCLUSIONS

Several furanic compounds were evaluated as potential alternative fuels for use in diesel engines. Major conclusions include:

- Two of the biofuel candidates posses CN's in a suitable range for use in traditional diesel engines. These compounds are produced via hydrogenation of the furan moieties in these 15 and 16-carbon containing compounds to their tetrahydrofuranyl analogs, providing cetane numbers of 60.4 and 59.8, respectively.
- Improvements in model architecture improved the overall accuracy of CN predictions by 5.54% (0.35 CN units) over prior efforts, with a total RMSE of 5.97 for the core data set.
- The use of an expanded data set, based on a targeted expansion of the input data to include similar molecules, improved the predictive accuracy by 49.21%. This represents an improvement in absolute error between predicted and experimental CN from 7.60 CN units for the core data set and 3.86 CN units for the expanded data set.

The results indicate that the current model is accurate and robust in predicting the CN of furanic molecules. Improvements in the overall RMSE of the model can be obtained through a few of the aforementioned approaches including elimination of questionable input data. Furthermore, the model can be confidently applied to other furanic compounds under consideration for use as alternative fuels.

ACKNOWLEDGMENTS

Alexis T. Bell acknowledges funding from the Energy Biosciences Institute (EBI), which is supported by BP. The authors would also like to acknowledge the National Renewable Energy Laboratory for their assistance in testing some of the furanic compounds and Juan Pablo Trelles at the University of Massachusetts Lowell for allowing the use of his computational cluster.

REFERENCES

[1] ASTM D613 Standard Test Method for Cetane Number of Diesel Fuel Oil, ASTM International, West Conshohocken, PA, 2015.

[2] ASTM D6890 Standard Test Method for Determination of Ignition Delay and Derived Cetane Number (DCN) of Diesel Fuel Oils by Combustion in a Constant Volume Chamber, ASTM International, West Conshohocken, PA, 2015.

[3] A.D.B. Yates, C.L. Viljoen, and A. Swarts, "Understanding the Relation Between Cetane Number and Combustion Bomb Ignition Delay Measurements." SAE Technical Paper 2004-01-2017, 2004.

[4] H. Yang, C. Fairbridge, and Z. Ring, "Neural Network Prediction of Cetane Number for iso-Paraffins and Diesel Fuel," Petroleum Science and Technology, Vol. 19, No. 5–6, pp. 573-586, 2001.

[5] T.H. DeFries, R.V. Kastrup, and D. Indritz, "Prediction of cetane number by group additivity and carbon-13 nuclear magnetic resonance," Ind. Eng. Chem. Res., Vol. 26, pp. 188-193, 1987

[6] J. Taylor, R. McCormick, and W. Clark, "Report on the relationship between molecular structure and compression ignition fuels," NREL Technical Report, 2004.

[7] E.A. Smolenskii, V.M. Bavykin, A.N. Ryzhov, O.L. Slovokhotova, I.V. Chuvaeva, and A.L. Lapidus, "Cetane numbers of hydrocarbons: calculations using optimal topological indices," Russian Chemical Bulletin, Vol. 57, No. 3, pp. 461-467, 2008.

[8] B. Creton, C. Dartiguelongue, T. de Bruin, and H. Toulhoat, "Prediction of the Cetane Number of Diesel Compounds Using the Quantitative Structure Property Relationship," Energy & Fuels, Vol. 24, No. 10, pp. 5396–5403, 2010.

[9] D.A. Saldana, L. Starck, P. Mougin, B. Rousseau, L. Pidol, N. Jeuland, and B. Creton, "Flash Point and

Cetane Number Predictions for Fuel Compounds Using Quantitative Structure Property Relationship (QSPR) Methods," Energy & Fuels, Vol. 25, No. 9, pp. 3900–3908, 2011.

[10] T. Sennott, C. Gotianun, R. Serres, M. Ziabasharhagh, J.H. Mack, and R.W. Dibble, "Artificial neural network for predicting cetane number of biofuel candidates based on molecular structure", ASME 2013 Internal Combustion Engine Division Fall Technical Conference, 2013.

[11] M.J. Climent, A. Corma, and S. Iborra, "Conversion of biomass platform molecules into fuel additives and liquid hydrocarbon fuels," Green Chem., Vol. 16, pp. 516-547, 2014.

[12] S. Dutta, S. De, B. Saha, and Md.I. Alam, "Advances in conversion of hemicellulosic biomass to furfural and upgrade to biofuels," Catal. Sci. Tech., Vol. 2, pp. 2025-2036, 2012.

[13] D.M. Alonso, J.Q. Bond, and J.A. Dumesic, "Catalytic conversion of biomass to biofuels," Green Chem., Vol. 12, pp. 1493-1513, 2010.

[14] A. Corma, O. de la Torre, and M. Renz, "Production of high quality diesel from cellulose and hemicellulose by the Sylvan process: catalysts and process variables," Energy Environ. Sci., Vol. 5, pp.6328-6344, 2012.

[15] M. Balakrishnan, E.R. Sacia, and A.T. Bell, "Syntheses of Biodiesel Precursors: Sulfonic Acid Catalysts for Condensation of Biomass-Derived Platform Molecules," ChemSusChem, Vol. 7, pp. 1078-1085, 2014.

[16] E.R. Sacia, M. Balakrishnan, and A.T. Bell, "Biomass conversion to diesel via the etherification of furanyl alcohols catalyzed by Amberlyst-15," J. Catal., Vol. 313, pp. 70-79, 2014.

[17] M. Balakrishnan, E.R. Sacia, and A.T. Bell, "Selective Hydrogenation of Furan-Containing Condensation Products as a Source of Biomass-Derived Diesel Additives," ChemSusChem, Vol. 7, pp. 2796-2800, 2014.

[18] J. Yanowitz, M.A. Ratcliff, R.L. McCormick, J.D. Taylor, and M.J. Murphy, "Compendium of Experimental Cetane Numbers," NREL/TP-5400-61693, 2014.

[19] MarvinSketch, Version 15.10.19.0, 2015. ChemAxon (http://www.chemaxon.com)

[20] C. G. C. T. and U. Services and W.-D. Ihlenfeldt, "Online SMILES Translator and Structure File Generator," 2011. [Online]. Available: http://cactus.nci.nih.gov/index.html.

[21] I. V Tetko, J. Gasteiger, R. Todeschini, A. Mauri, D. Livingstone, P. Ertl, V. a Palyulin, E. V Radchenko, N. S. Zefirov, A. S. Makarenko, V. Y. Tanchuk, and V. V Prokopenko, "Virtual computational chemistry laboratory--design and description," Journal of computer-aided molecular design, Vol. 19, No. 6, pp. 453-63, 2005.

[22] R. Todeschini and V. Consonni, "Handbook of molecular descriptors," Weinheim: Wiley-VCH, 2000.

[23] O. Devinyak, D. Havrylyuk, and R. Lesyk, "3D-MoRSE descriptors explained," Journal of Molecular Graphics and Modelling, Vol. 54, pp. 194-203, 2014.

[24] K. Levenberg, "A Method for The Solution of Certain Nonlinear Problems in Least Squares," The Quarterly of Applied Mathematics, Vol. 2, No. 2, pp. 164-168, 1944.