doi:10.1115/ICEF2017-3614 Pre-Print

Proceedings of the ASME 2017 Internal Combustion Fall Technical Conference
ICEF2017
October 15-18, 2017, Seattle, Washington, USA

APPLICATION OF A RECTIFIED LINEAR UNIT (RELU) BASED ARTIFICIAL NEURAL NETWORK TO CETANE NUMBER PREDICTIONS

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

Gregory Dorian
University of Massachusetts Lowell
Lowell, Massachusetts, United States

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

ABSTRACT

Due to the high cost and time required to synthesize alternative fuel candidates for comprehensive testing, an Artificial Neural Network (ANN) can be used to predict fuel properties, allowing researchers to preemptively screen desirable fuel candidates. However, the accuracy of an ANN is limited by its error, measured by the root mean square error (RMSE), standard deviation, and r-squared values derived from a given input database. The present work improves upon an existing model for predicting the Cetane Number (CN) by changing the neuron activation function of the ANN from sigmoid to rectified linear unit (ReLU). This change to the ANN's architecture provides an increase in accuracy by reducing the RMSE by 21.4% (1.35 CN units), the average standard deviation across models by 28%, and increasing the r-squared value by 0.0492 across a wide range of molecular structures. Additionally, by using the ReLU activation function, input data is not required to be normalized, which reduces the likelihood of an inaccurate prediction on future fuel candidates which may have input parameters outside the range of normalization. Increasing the accuracy of the predictive ANN in this way will allow researchers to obtain more accurate fuel property predictions for promising fuel candidates.

INTRODUCTION

Investigating the next generation of fuels has gained considerable interest, due in part to concerns over decreasing reserves of traditional fuel sources as well as their impact on the global climate. Biofuels, such as those derived from renewable plant sources, have complex oxygenated functional groups that are inherently more complex than conventional hydrocarbon fuels. Because of the high cost and time required for chemical synthesis, the next generation of biofuels are challenging to produce at the scale required for comprehensive experimental testing. However, through the use of a predictive model trained to determine key fuel properties such as cetane number, researchers can preemptively determine promising fuel compounds to expedite the research and development of fuels, resulting in both lower production costs and accelerated development timelines.

Cetane Number

The cetane number (CN) of a fuel is a measurement of its ignition quality when used in a diesel engine. The measurement is derived from the ignition delay from the time of injection. The two most common methods for verifying CN are using a Cooperative Fuel Research (CFR) engine or an Ignition Quality Tester (IQT). Experimental methods for determining CN with a CFR is outlined through American Society for Testing and Materials (ASTM) Standard D613 [1]. This method uses two reference compounds: n-hexadecane and isocetane (2,2,4,4,6,8,8-heptamethyl-nonane, or HMN), with CN values of 100 and 15 respectively. Using an equivalent blend of n-hexadecane and isocetane, the resultant volume fractions of each compound after a fuel in question is ignited is linearly related to the CN of the fuel.

Alternatively, the method using an IQT specified in the ASTM Standard D6890 measures the time delay between fuel injection and combustion [2]. Multiple tests are performed at constant conditions in the IQT, with the final ignition delay being an average of the test results. The Derived Cetane Number (DCN) is then obtained from the ignition delay using an empirical equation; however, this equation has limited accuracy when used across multiple test fuels [3]. While the method using CFR is preferred as it reflects typical engine behavior, the method using IQT requires a smaller volume of test fuel (typically 100 mL).

Even though the IQT provides a method for determining CN using a small volume of fuel, the large numbers of potential

fuels that researchers identify still take a considerable amount of time and investment to test. This bolsters the need for a predictive model for key fuel properties, allowing researchers to preemptive screen fuel candidates.

Predicting the Cetane Number

Historically, quantitative structure-property relationship (QSPR) values, which are measurements of physical and chemical qualities of molecules, in conjunction with feed-forward neural networks have proven to be successful in predicting CN when applied to branched paraffins [4] and pure hydrocarbons [5]. However, while the predictive models applied to these compounds were adequate for predicting CN within the range of compounds considered, they remained inaccurate when applied to other compounds and compound groups.

Consensus modeling, where results from both linear and nonlinear models (including artificial neural networks), improved upon the predictive power of previous methods, allowing 7 chemical families to be considered with a RMSE of 6.3 CN units [6]. However, multiple linear regression models are shown to be outperformed by ANN's when predicting cetane number using fatty acid methyl esters [7].

This paper adopts feed-forward ANN's trained through backpropagation and QSPR input data, as they have recently shown to be more flexible for predicting CN across multiple molecular classes, in part due to the ANN's nonlinear architecture allowing complex analysis of multidimensional input and output vectors, and the range of physical and chemical attributes represented in the QSPR input data [8]. The goal of this paper is to improve the accuracy and robustness of predictive models on a diverse dataset of molecular compounds. This is done through introducing an alternative activation function to the neural network's architecture, resulting in a reduction of the RMSE of each model, an increase in the r-squared coefficient of determination, and a decrease in the variance of predictions across multiple models.

ANN Activation Functions

ANN's use activation functions at each neuron to transform the input signal to the neuron into an output signal. Before modern support vector machines and random forest classification methods were introduced in the 1990's, ANN's utilizing the sigmoid activation function were common for machine learning tasks [9]. Due to the mathematical nature of the sigmoid function seen in Equation 1, exceedingly negative and positive input signals result in an output signal of 0 and 1 respectively. Therefore, input signals for sigmoid-based neurons typically need to be normalized between 0 and 1 using min/max normalization in Equation 3. However, when an input signal falls outside the parameters of normalization, namely the minimum and maximum of the dataset parameters used for training, the resultant output signal is not representative of the input signal. This proves to be a considerable problem when trying to predict the CN of new molecules whose QSPR descriptors fall outside their ranges of normalization.

$$f(x) = 1 / (1 + e^{-x}) \quad (1)$$

$$g(x) = \max(0, x) \quad (2)$$

$$z_i = (x_i - \min(x))/(\max(x) - \min(x)) \quad (3)$$

Alternatively, the Rectified Linear Unit (ReLU) activation function seen in Equation 2 does not require normalization of input data. Additionally, the output of a ReLU neuron is equal to 0 when the input is less than zero and is equal to the input otherwise, creating a natural dropout among the neurons of an ANN during training. Having dropout neurons helps avoid overfitting of the training data, as well as decreases the computation time needed to train [10]. This paper adopts the ReLU activation function for the ANN's used to predict CN, to avoid normalization and decrease overfitting of the training data.

METHODS

Data for cetane numbers used in the core data set was obtained from sets present in the NREL Compendium of Experimental Cetane Number Data [11] and other sources [6, 12, 13]. The core data set contains 423 molecules, with their cetane numbers derived from blend measurements, CFR, IQT or other ignition delay methods. Most of the cetane numbers were derived from IQT and CFR, as these methods provide higher accuracy relative to blend measurements and other methods. We have chosen not to exclude compounds that are common outliers across ANN's, as our goal is to provide a generalizable model for a wide variety of compound groups including new unknown data. Instead, common outliers were targeted for further research and their cetane numbers were updated using more recent sources [13].

SMILES (Simple Molecular-Input Line-Entry System) data was obtained using compound structures and MarvinSketch (ChemAxon Ltd.) [14]. 1667 QSPR molecular descriptors were obtained by feeding SMILES data into the E-Dragon software with CORINA three-dimensional conversion [15]. The QSPR descriptors and their corresponding cetane numbers compose the core data set.

Through an iterative regression analysis technique, the number of QSPR input descriptors was reduced from 1667 to 15. This

was done to shorten the computation time needed to construct ANN's while retaining low error. For each QSPR input descriptor, ANN's were constructed using each descriptor exclusively. The best performing descriptor was retained, and used in conjunction with every other individual descriptor. The best pair was retained and cycled again. This process was repeated until 15 parameters were obtained. This was done multiple times to confirm common parameter outcomes.

Figure 1 displays the resultant 15 descriptors and the ANN's RMSE after each reduction step. Between 15-25 iterations, the RMSE does not significantly improve. As the number of iterations increases further, the RMSE starts to increase. This relationship can be observed in Figure 2. This is due in part to many QSPR descriptors being equal to zero for most molecules, which is detrimental to an ANN and does not allow it to capture nonlinear relationships.

Inspection of the resultant descriptors may yield insight into how the composition of a molecule influences the molecule's cetane number. Table 1 provides definitions for each resultant descriptor.

Descriptors such as nOHp (number of primary alcohols), nROH (number of hydroxyl groups), and nROR (number of ethers) are more physical in nature, and are associated with the foundation of chemical kinetics that occur during combustion.

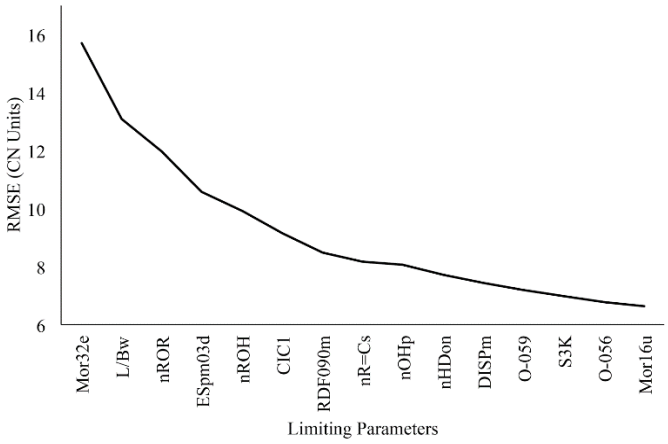

Figure 1. Iterative Addition of Descriptors to ANN

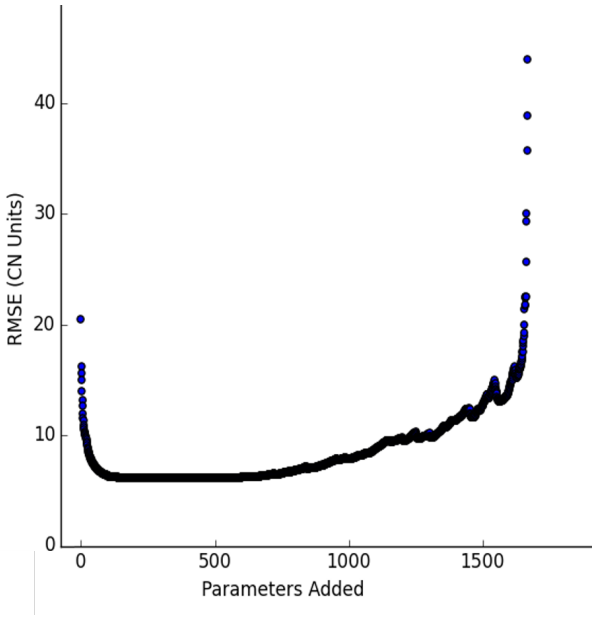

Figure 2. Iterative Addition of All Descriptors to ANN

Table 1. Glossary of Descriptor Definitions

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

ANN's were implemented in Python 3.5 utilizing Levenberg-Marquardt backpropagation, a common training method for ANN's [16]. The function for determining the mean squared error (MSE) was used as the optimization function for the ANN. During the training process, the ANN actively works to

reduce the MSE to a minimum relative to the core data set. The architecture of the ANN, seen in Figure 3, has 15 inputs (for 15 QSPR descriptors), two hidden layers containing 32 neurons each, and one output (CN prediction). Two hidden layers are used in the architecture to capture the nonlinear relationship between the inputs and the output.

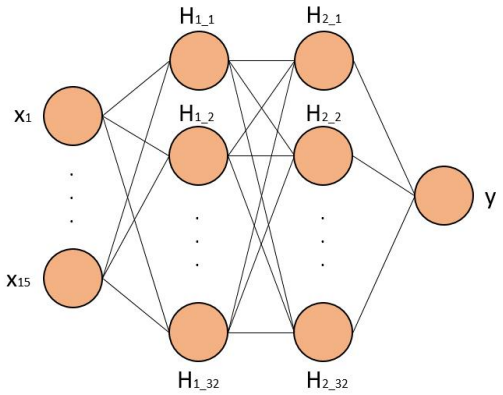

Figure 3. Model architecture with inputs (x_1 - x_{15}), two hidden layers with 32 neurons each, and an output y.

The core database is split into testing, validation and learning (T/V/L) subsets, chosen randomly with proportions 0.1, 0.25 and 0.65 respectively. The learning subset is used for backpropagation learning by the ANN, and the validation subset is used to test the performance of the ANN after each training epoch (iteration). The training halts once there is no further improvement in predictions on the validation set. This cutoff point is determined by when the mean-delta-root-mean-square-error (mdRMSE) falls below a predetermined threshold value. The mdRMSE shows the average value of the change in RMSE of the validation subset between training epochs, and approaches zero as training progresses. The testing set is then used to determine the overall generalizability of the trained ANN. The overall performance of the ANN is determined by measuring the RMSE when it is tested on the entire core data set. A lower RMSE indicates a better performing ANN.

The number of ANN's that need to be constructed to achieve an optimal one is increased due to random T/V/L splitting. Ultimately, this provides greater accuracy for the final model, as the ANN can choose the best performing T/V/L split that lead to the best performing model. However, because the T/V/L subsets are assigned randomly to the core data set, it is possible that the final ANN may not be completely representative of the core data set regarding compound groups. Targeted T/V/L splitting may lead to better representation and fewer ANN's needing to be built, but would result in a questionable ANN without a systematic selection technique that retains accuracy.

The final predictive model (build set) is comprised of five individually trained ANN's from five computational nodes. Each node trains 75 ANN's and chooses the best performing ANN for the final model. A final prediction is equivalent to the average of the five ANN predictions. Due to each ANN being trained using different random T/V/L splits, each ANN's predictions are unique, tending to be either slightly higher or slightly lower than the actual CN of the molecule in question. The final averaged prediction tends to be closer to the actual CN, improving the accuracy of the model. Figure 4 shows a simulation diagram of the final predictive model's construction.

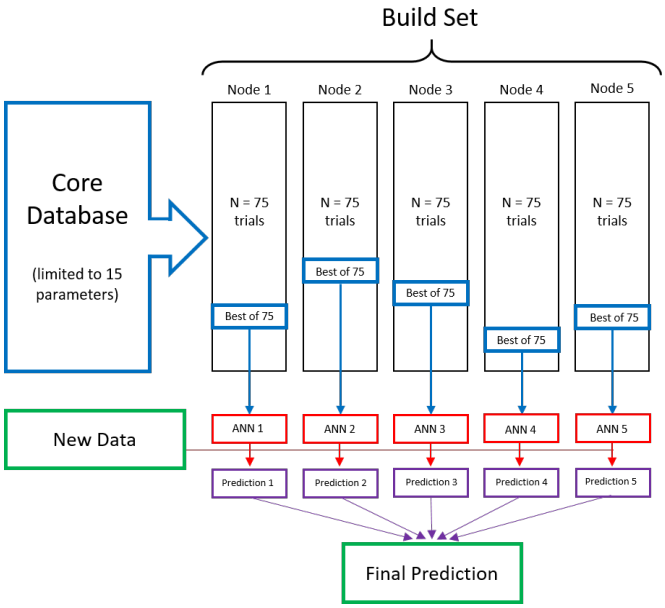

Figure 4. Build set construction simulation diagram

To compare activation functions, ten ReLU-based build sets and ten sigmoid-based build sets were constructed. Key statistical information was gathered from the ReLU build sets, namely the RMSE and the R^2 correlation coefficient of the best performing ReLU build set, and the standard deviation across the ten build sets for each activation function.

Additionally, cetane numbers from a set of 17 molecules withheld from the build process were predicted using both ReLU and sigmoid build sets. This was done to test the predictive accuracy of each build set type on new molecules not seen by the model.

RESULTS & DISCUSSION

Figure 5 shows a parity plot between experimental CN values and predicted CN values of the entire core data set obtained from the best performing build set from the ten constructed ReLU-based build sets, and Figure 6 for the sigmoid-based build sets. The solid line indicates a 1:1 parity, with the dashed lines indicating the bounds imposed by the RMSE. The RMSE

of this build set was found to be 4.95 CN units, showing an improvement over sigmoid-based predictive models by 21.4% (RMSE of 6.3 CN units).

Additionally, the R^2 correlation coefficient of the ReLU build set was found to be 0.963, showing an improvement over sigmoid-based predictive models by 0.0492 (R^2 of 0.9138). A higher R^2 value indicates better goodness of fit of the build set. Alternatively, this means the build set is better able to approximate the CN of input molecules.

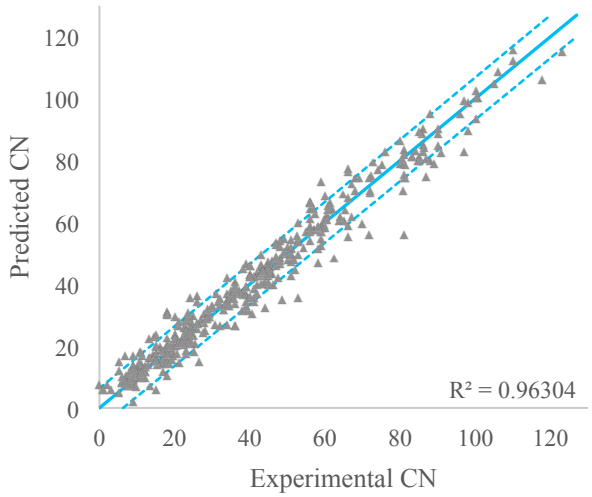

Figure 5. Parity plot of Predicted CN vs. Experimental CN (ReLU)

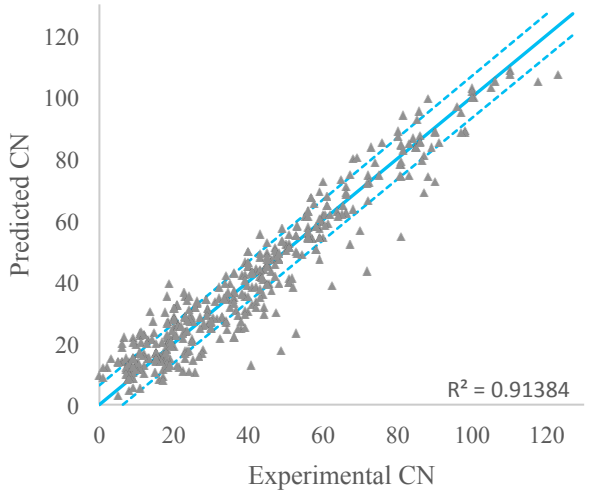

Figure 6. Parity plot of Predicted CN vs. experimental CN (Sigmoid)

The standard deviation of predictions across the ten ReLU-based build sets was found to be 2.06 CN units. Additionally, the standard deviation of predictions across the ten sigmoid-based build sets was found to be 2.87 CN units. The improvement of 28% indicates less variation in predictions across multiple build sets, illustrating consistent performance of the build set architecture.

Table 2 shows cetane number prediction results for 17 molecules withheld from the build process, their experimental values, and the RMSE's of each build set type. The predictions indicate that overall, the ReLU based model performs better than the sigmoid based model when predicting cetane numbers of molecules not seen by the model during training. This is reflected by the RMSE values for the ReLU predictions and the sigmoid predictions, 7.94 and 9.23 CN unites respectively.

Table 2. Test Set Cetane Number Predictions

Molecule Name	Experimental Value	ReLU Prediction	Sigmoid Prediction
methyl linolelaidate	43.0	47.98	57.09
methyl petroselinate	58.6	62.40	74.88
methyl ricinoleate	37.4	42.14	41.74
methyl-5(Z)8(z)11(Z)14(Z)-eicosatetraenoate	29.6	27.85	40.72
methyl-9-decenoate	38.3	52.17	46.44
octadecyl acetate	90.0	84.40	95.05
octyl laurate	84.0	69.04	69.11
octyl myristate	71.0	81.94	77.77
octyl valerate	49.0	48.54	50.59
palmitoleyl alcohol	46.0	50.99	53.44
pentyl pentanoate	28.2	32.36	32.74
Perhydro-phenanthrene	38.8	24.96	22.38
propyl laurate	71.0	60.38	65.03
propyl myristate	71.0	68.90	73.22
propyl oleate	72.0	63.28	71.24
tetradecyl acetate	81.0	78.65	89.96
triethyleneglycol dimethyl ether	120.0	117.39	116.36
	RMSE:	**7.94**	**9.23**

Figure 7 shows the delta values (predicted CN with ReLU minus experimental CN) for each compound group present in the core data set, sorted by mean prediction error. It is worth noting that underrepresented compound groups in the core data set such as aldehydes tend to deviate from parity more than other compound groups. Alternatively, well-represented compound groups such as esters are found to perform the best as a group. The outcome of including compounds of an individual group has shown to increase the model's accuracy on that group [17]. This reinforces the need for additional experimental data for underrepresented compound groups to improve upon the overall accuracy of the predictive model.

Further work to improve on compound group representation and their resultant biases would include analysis on the optimal number of molecules of each group needed to obtain accuracy within the group. Additionally, analysis on compound group representation within specified CN ranges could yield insightful information into prediction bias within the CN ranges. A segmented graph indicating compound frequency within CN ranges can be seen in Figure 8.

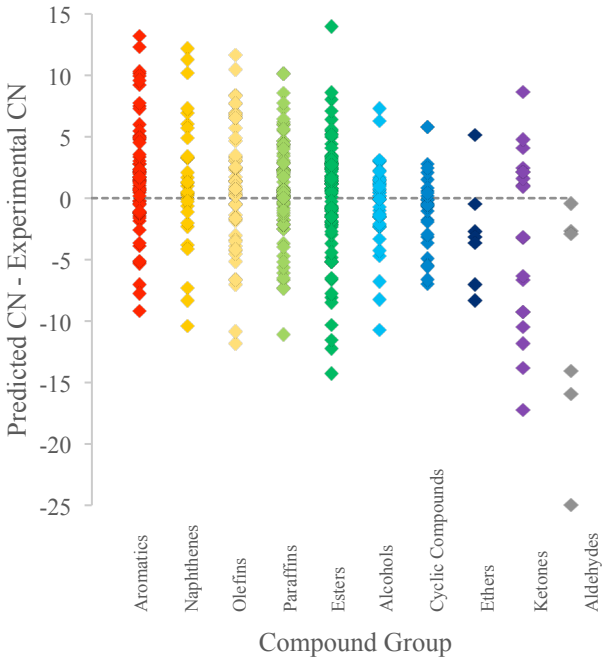

Figure 7. Prediction performance for compound groups

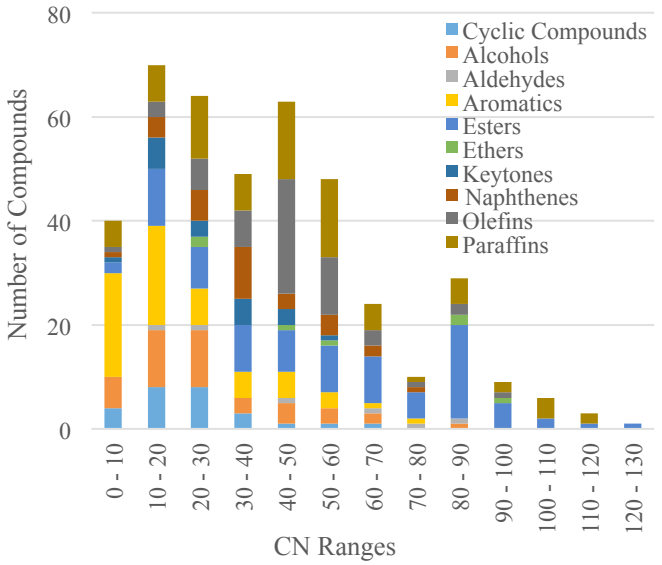

Figure 8. Compound group representation within CN ranges

Figure 9 depicts key statistical information about prediction error within specified CN ranges (mean, median, interquartile ranges (IQT) and outliers). The IQT ranges for each CN range greater 60 CN are considerably larger than the CN ranges less than 60 CN, which indicates more variation in predictions within these ranges. Additionally, the mean value is noticeably less than the median value for most ranges greater than 60 CN. This indicates that the predictive model predicts lower for compounds within these ranges. These phenomena are likely due to underrepresentation of compounds within the affected CN ranges, and overrepresentation for other ranges. The total compound frequencies in Figure 8 seem to bolster this theory.

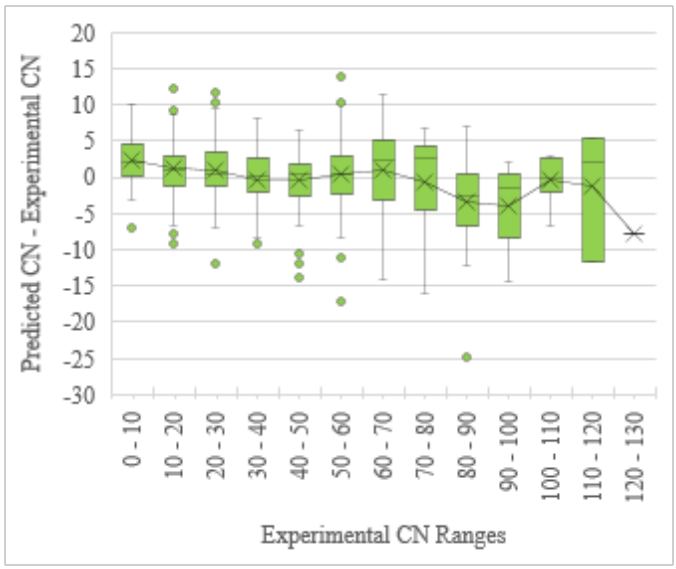

Figure 9. CN deltas within specified CN ranges

CONCLUSIONS

From the statistical data obtained from both ReLU-based and sigmoid-based predictive models, it can be concluded that:

- Using the ReLU activation function when predicting the CN of biofuel candidates yields an overall improvement across the core data set of 21.4% compared to when the sigmoid activation function is used, and provides a build set RMSE of 4.95 CN units.
- The use of the ReLU activation function decreases the standard deviation of predictions across multiple build sets by 28%, indicating consistent performance of the build set architecture.
- Using the ReLU activation function increases the R^2 correlation coefficient of the core data set, indicating better performance for future predictions.
- ANN's utilizing the ReLU activation function generally perform better than those using the sigmoid activation function for predicting the cetane number of new candidate molecules.

These results indicate that the current predictive model proposed is both accurate and robust when predicting the CN of potential biofuels across a variety of compounds represented in the core data set, as well as future biofuel candidates.

REFERENCES

[1] ASTM D613-16a, Standard Test Method for Cetane Number of Diesel Fuel Oil, ASTM International, West Conshohocken, PA, 2016, www.astm.org

[2] ASTM D6890-16e1, 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, 2016, www.astm.org

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

[6] D.A. Saldana, L. Starck, P. Mougin, B. Rousseau, L. Pidol, N. Jeuland, 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.

[7] R. Piloto-Rodríguez, Y. Sánchez-Borroto, M. Lapuerta, L. Goyos-Pérez, S. Verhelst, "Prediction of the cetane number of biodiesel using artificial neural networks and multiple linear regression," Energy Conversion and Management 65, pp 255-261, 2013

[8] T. Sennott, C. Gotianun, R. Serres, M. Ziabasharhagh, J.H. Mack, 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.

[9] J. Ma, R. Sheridan, A. Liaw, G. Dahl, V. Svetnik, "Deep Neural Nets as a Method for Quantitative Structure-Activity Relationships," J. Chem. Inf. Model., 55 (2), pp 263-274, 2015

[10] G.E. Dahl, T.N. Sainath, G.E. Hinton, "Improving deep neural networks for LVCSR using rectified linear units and dropout," Acoustics Speech and Signal Processing (ICASSP) 2013 IEEE International Conference on, pp. 8609-8613, 2013.

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

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

[13] M. Dahmen, W. Marquardt, "A Novel Group Contribution Method for the Prediction of the Derived Cetane Number of Oxygenated Hydrocarbons," Energy Fuels, 29 (9), pp 5781-5801, 2015

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

[15] 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, V. V Prokopenko, "Virtual computational chemistry laboratory--design and description," Journal of computer-aided molecular design, Vol. 19, No. 6, pp. 453-63, 2005.

[16] 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.

[17] T. Kessler, E. Sacia, A. T. Bell, J. H. Mack, "Predicting the Cetane Number of Furanic Biofuel Candidates Using an Improved Artificial Neural Network Based on Molecular Structure," ASME 2016 Internal Combustion Engine Fall Technical Conference, Greenville, SC, 2016