8th U. S. National Combustion Meeting
Organized by the Western States Section of the Combustion Institute
and hosted by the University of Utah
May 19-22, 2013

Cetane Number Prediction from Molecular Structure Using Artificial Neural Networks

T. Sennott[1] C. Gotianun[1] R. Serres[2] J.H. Mack[1] R. Dibble[1]

[1] Combustion Analysis Laboratory, University of California Berkeley, Berkeley, CA 94720 USA
[2] Arts et Métiers ParisTech, 151 Boulevard de l'Hôpital, 75013 Paris, France

The production of next-generation biofuels is being explored through a variety of chemical and biological approaches, all aiming at lowering costs and increasing yields while producing viable alternatives to gasoline or diesel fuel. Chemical synthesis can lead to a huge variety of different fuels and the guidelines for which molecules would yield desirable properties as a fuel are largely based on intuition. One such property of interest is the cetane number, a measure of the ignition quality of diesel fuel. The present work improves on existing models and extends them to more oxygenates (primarily ethers) as an interim step in extending the model for further compounds of interest such as furans and tetrahydrofurans. Various members of these classes are being considered as fuels of interest by chemists and engineers in biofuel production. The present model uses artificial neural networks (ANN's) as a tool for quantitative structure property relationship (QSPR) analysis. Predicting the cetane number of a fuel is especially important because a large volume of pure sample (100mL or more) is typically required for lab testing, the production of which can be difficult and time-consuming at the lab scale. To this end, a predictive model will allow chemists to eliminate unlikely targets and focus their attention on promising candidates.

1. Introduction

With increased concern about global warming, coupled with the increasing scarcity of conventional fossil fuels, more effort has been placed in researching alternative fuel sources, specifically biofuels. Biofuels are derived from renewable sources such as sugars, starch and vegetable oil, and as such typically contain oxygen and are more structurally complex and varied than traditional hydrocarbons. These fuels can offer many benefits, especially when derived from cellulosic biomass, but efforts to develop next-generation biofuel have proven challenging.

For diesel applications, one of the most important indicators for a fuel is its cetane number (CN). This number indicates the combustibility of a fuel, the fuel's ignition delay from injection to compression ignition, including both physical delay (vaporization) and chemical delay. Two of the most widely used methods consist of testing using a Cooperative Fuel Research (CFR) engine and an Ignition Quality Tester (IQT). Obtaining CN through a single-cylinder CFR is specified in the American Society for Testing and Materials (ASTM) D613. This measurement is compared to a CN scale defined by two standard compounds: cetane (n-hexadecane) and isocetane (2,2,4,4,6,8,8-heptamethyl-nonane, also referred to as HMN), which have CN values of 100 and 15 respectively.

Thus the volumetric percentage of a blend of cetane and HMN, which exhibit the same ignition delay as the fuel, defines its CN. Likewise, the IQT test procedure is specified in the ASTM-D6890, which derives the CN by determining ignition delay in a constant volume combustion chamber. The IQT method measures the ignition delay time, from fuel injection in the combustion chamber to the initiation of combustion. From a set of repeated tests, at constant conditions, an equation is used to convert the set of ignition delay times to derive a cetane number referred to as the Derived Cetane Number (DCN). Though both methods provide accurate CN measurements, the IQT offers the advantage of more rapid testing with substantially less fuel required (typically ~100mL). Still, when considering large numbers of potential fuels made at benchtop scale, even these tests can represent substantial investment of time and resources. Therefore a screening method for predicting this and other properties would be a desirable tool for researchers in the field.

The idea of predicting cetane numbers and other fuel properties from molecular structure is not new. Previously, models based on quantitative structure property relationships (QSPR) have been developed to predict the CN of various compounds. Early research in understanding the relationship between the molecular structure and CN is presented in literature by the National Renewable Energy Laboratory (NREL) [1]. Using a Quantitative Structure Activity Relationship (QSAR) software, 100 molecular descriptors, such as geometry, connectivity, functional groups, were generated from a CN data set of 275 compounds: 147 hydrocarbons and 128 oxygenates. From this, descriptors thought to influence the CN, were determined using a genetic algorithm and the resulting models were generated for both the entire dataset and specific subclasses (Olefins, Cyclic compounds, Oxygenates and Esters). Though CNs were not predicted with great accuracy (entire dataset standard error of 9.1 CN units and an $R^2 = 0.91$), this paper laid a good foundation for future work on the use of QSPR predictive models in determining CN of various compounds.

A variety of other models, both linear and non-linear, have been used to predict CN. Some of these models apply a functional group contribution approach, while others use molecular descriptors, as the NREL group did. Smolenskii et al [2] utilized an inverse function approach for accurately predicting CN of pure hydrocarbons (R2 value of ~0.99) Though this curve-fitting model is accurate for the range of compounds included, it does not have the flexibility to include other chemical families. More recently, Creton et al [3] developed models for each chemical family (Paraffinic, Napthenic, Aromatic and Olefinic compounds) using genetic function algorithms. Further work by Saldana et al [4] combined linear and nonlinear models (including neural networks) through "consensus" modeling, and extended the model to include alcohols and esters. This approach resulted in a robust CN prediction model for 279 compounds from 7 chemical families. The model calculated an R2 value of about 0.934 and total root mean square error (RMSE) of 6.3, and represents the current state of the art for prediction over a broad range of compounds.

The goal of this paper is to develop a more accurate and robust predictive tool by fine-tuning the predictive model and extending the data set to other chemical families. This is done by first identifying the most influencing molecular descriptors that contribute to a compound's CN through the use of sensitivity analyses studies. This paper will then present a model that analyzes a larger data set, while at the same time improving its predictive accuracy.

2. Methods

Cetane Number and Derived Cetane Number data used for this analysis came from data sets provided in the NREL Compendium of Experimental Cetane Number Data [5] and two publications (Saldana et al [4] and NREL [1]). The accuracy of some of the quoted values from these sources is questionable, due largely to the uncertainty of CN tests (commonly taken to be +/- 5 CN units) and the difficulty ensuring the purity of some compounds. The largest amount of this information comes from the NREL "Compendium" [5], which takes data from various sources in the literature. Compounds tested may be of uncertain purity, and are obtained via different methods, some less accurate than others (such as blend tests or octane to cetane correlations). Some compounds have multiple reported values with large ranges, while others are grouped closely. Previous authors such as Saldana et al have selectively excluded a number of experimental CN values for these reasons, a procedure we will also employ.

Overall, the quality of the data is the greatest limiting factor in training predictive models. This limitation can be mitigated by appropriate "trimming" of the data set, as will be discussed later.

Additional updated DCN data was furnished by Brad Zigler and Matt Ratcliff from recent IQT testing at NREL. Further data is being sought, from both experimental sources as well as further literature review.

Compound structures were converted to SMILES using MarvinSketch (ChemAxon Ltd.), if they were not readily available from the Saldana database. SMILES structures were converted to 2-D structures (MDL .sdf files) using the NCI online calculator [6]. With the generated .sdf file, 1667 QSAR molecular descriptors were generated using e-Dragon [7], a free online Java port of Talete's Dragon software, which also handled calculation of the 3-D structures using CORINA.

Selection of initial candidate set of Molecular Descriptors was accomplished using literature [3], [4], [8], as well as attempts to capture the range of features that seemed likely to impact reactivity. The Handbook of Molecular Descriptors [9] was a key aid in understanding the various molecular descriptors. Functional group counts were also included to allow the network to account for the effects of various functional groups (such as ether linkages) on reactivity. The complete initial set of 31 descriptors selected was as follows (listed by Dragon abbreviation): MW, Sv, Se, Ss, nBM, nCIC, Xt, VDA, MSD, MEcc, SPH, ASP, nRCOOR, nRCO, nOHp, nROR, nOHs, C-001, C-002, C-003, C-004, H-051, Ui, W, J, TIE, S0K, S1K, S2K, S3K, FDI, and JGT.

Input parameter reduction was performed using an iterative search described in the Results section, reducing the number of parameters depending on the data set being evaluated.

Regression analysis was performed using Artificial Neural Networks implemented in MATLAB. The topology employed was a feed-forward network trained by Levenberg-Marquadt backpropogation, one of the most common types of ANN's. As studies continue, other topologies may be investigated in order to deal with the increasing complexity of a growing data set. Presently, investigation has focused on the validity of the data set itself, since it is likely that it is the largest factor limiting the learning algorithm.

The figure of merit selected for the regression was the mean squared error (MSE) or equivalently, root mean square error (RMSE), since these have the effect of weighting larger errors more heavily while penalizing small errors less; this behavior is desired in this case since the model is designed as an estimation tool, where small errors are tolerated (and expected) while large ones are much more undesirable.

Prior to each regression, the data was randomly assigned to one of three conditions: learning (50%), validation (35%), and testing (15%). In this scheme, learning data is used to train the network, until performance on the validation set stops improving (validation stop). Test data is held in reserve as a measure of how "generalizable" the model is. In cases where validity of the model was being tested on new data sources (i.e. DCN data from NREL), 10% was assigned to testing, and then the compounds of interest were forced into the test set as well.

Allowing random sorting of compounds into these different groups required more trails, but should yield greater peak accuracy since choosing the best network should select the training set which captures the greatest amount of information for the network to learn. In the future a set may be selected and used persistently for this purpose.

3. Results and Discussion

3.1 Initial Studies

First studies were conducted with the database of 279 hydrocarbons, alcohols and esters from Saldana et al [4] and the full list of 31 parameters listed in the Methods section. During these studies an optimum network structure of 2 hidden layers with 8 nodes per layer was found to offer the highest accuracy with minimum regression time. Additionally, it was found that averaging the predictive results of the top five (5) simulations for a given condition gave better predictions than the best model alone (typical reductions in RMSE of ~30%), and that 150 regression trials for a given condition was sufficient to produce these best networks.

Initial studies on the Saldana database yielded accuracies commensurate with the results of that group (RMSE of 5.4 units vs. 6.3 units). A possible reason for the slight improvement over the Saldana result is their model pre-prescribed the distribution of molecules into subsets randomly, without regard for benefit each molecule might confer to the learning algorithm. It also should be noted that Saldana also had a different distribution of training/validation/learning molecules (70/20/10%) and used a "consensus" model consisting not only of a feed-forward neural network, but also a generalized regression neural network (GRNN) and a support vector machine (SVM) network. The difference between their single feed forward artificial neural network (FF-ANN) and their consensus model (RMSE 11.3 vs. 6.3) is roughly similar with the difference we observed between our single-network results and our averaged result, so it appears much of the benefit of "consensus" model may be obtained with only a single regression architecture repeated several times, provided it is a stochastic one like a neural network.

A parity plot for the initial model using the Saldana database may be seen in Figure 1. Lines surrounding parity line are plus or minus 5 CN units, commonly assumed to be experimental uncertainty of CN measurements, though in many cases uncertainty in experimental values is actually much greater.

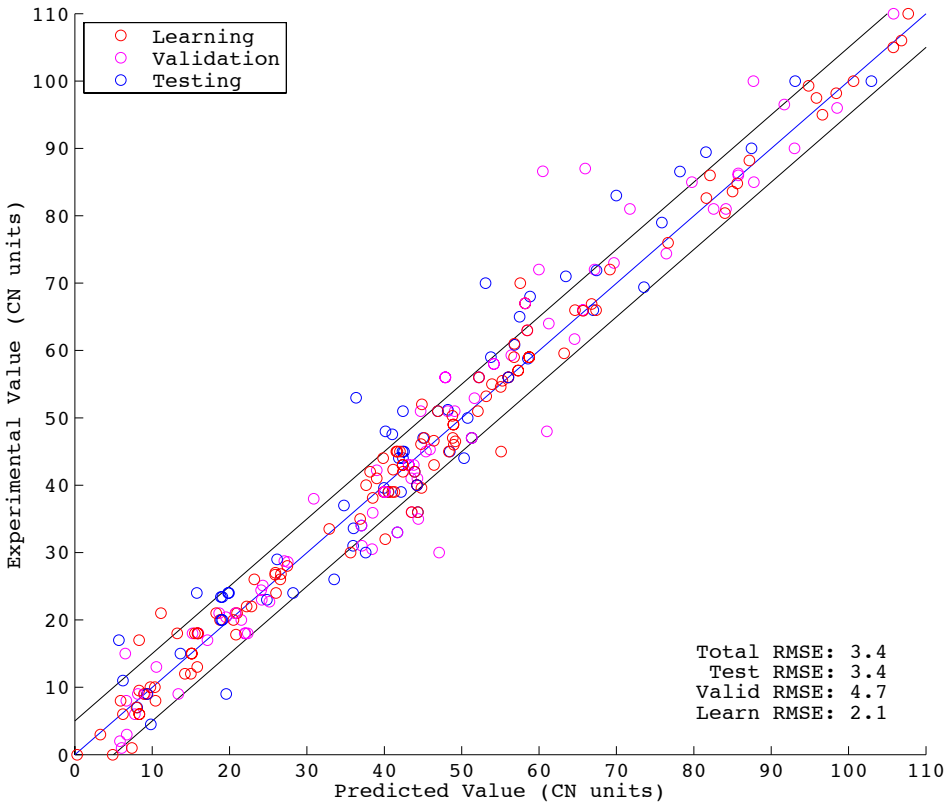

Figure 1: Parity Plot for CN prediction for 279 Hydrocarbons, Esters, and Alcohols

3.2 Input Parameter Reduction

To determine which subset of the initially selected molecular descriptors most of the variance in the data, an iterative strategy was devised. For each possible parameter, networks were repeatedly regressed using only that single parameter (5 runs per parameter, 150 networks per run). The parameter that produced the lowest average MSE was retained, and the next trial used this parameter plus any of the remaining parameters. Repeating this process through the list of all parameters yields a useful insight into the large amount of covariance between the possible inputs.

Figure 2 shows the parameter selection process applied to the initial Saldana dataset. It may be seen that additional parameters rapidly reach a point of diminishing returns, and that after eight (8) parameters there is no additional gain. This plot only shows the first fifteen (15) descriptors, but the same behavior continues out to the full parameter list (31 descriptors), with repeatability actually

decreasing with more than 20 descriptors. For comparison, the 2004 NREL model used between 14 and 23 descriptors depending on the subset of molecules they were analyzing.

It should be noted that repeating this parameter selection process yielded the same order of key descriptors for the early list (up to 7 descriptors), giving good confidence that these are actually the most predictive descriptors in the list. Once performance stopped improving, descriptor order was no longer repeatable.

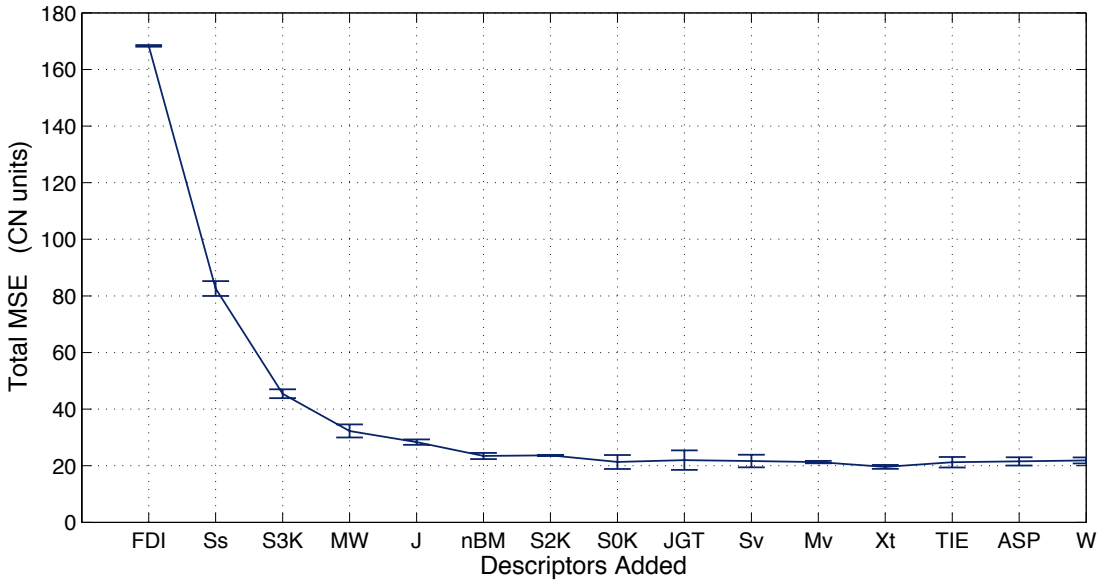

Figure 2: Iterative Addition of Descriptors to Model (Original Dataset)

3.3 Elimination of Questionable Data

As noted in the introduction, there is a good deal of uncertainty associated with the CN database sources used. As such, Saldana et al [4] discredited 45 "outliers" from their set that were poorly predicted and had experimental values that were questionable, presumably based on a combination of using expected value, various sources, methods and range of reported values.

We further analyzed the experimental dataset to determine if there were other questionable values that were limiting the accuracy of the model. By repeatedly regressing the networks and tracking which molecules that were consistently mispredicted, we identified lists of molecules to further investigate. Information on their structure, source(s), and also on the value predicted by Saldana et al was aggregated. These compounds were then interrogated on these bases, and if they failed multiple tests they were discarded. In cases where reported values spanned large ranges, or where the only methods were ones with questionable accuracy (ON to CN correlation, or blend extrapolation), this was counted against them. If similar compounds existed in the dataset that exhibited clear trends in contrast to the value in question, this was also noted. If the value seemed strange but it seemed like there could be chemistry that we were not accounting for, or if the compound had nothing similar to compare it to, the value was left in.

Applying this methodology several times, the dataset was eventually reduced by 34 compounds, to a total of 245 compounds. Regression on this dataset produced an improved RMSE of 3.3. The parity plot is shown in Figure 3, and is notably cleaner.

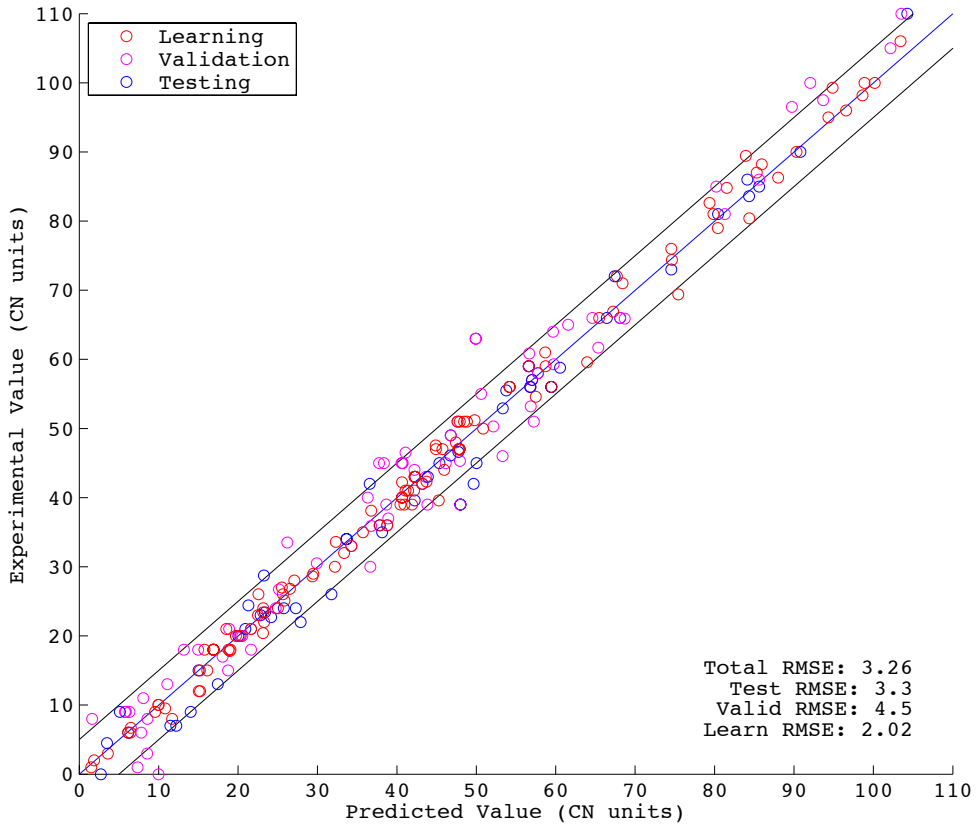

Figure 3: Parity Plot for Reduced Set of 245 HC's, Esters, and Alcohols

Whether it was truly appropriate to remove these compounds is difficult to say for certain. The model can accurately achieve this same RMSE even with set limits extended to 35% test, 25% validate and 40% learn, which would seem to indicate that it has good predictive power even with large amounts of data held unseen.

But of course the best test will be applying the reduced set model to new compounds and seeing if predictive power is unaffected or improves. This would indicate no important information was lost with these compounds, and ideally, the network was able to learn more about the set due to reduced noise from bad values. We will see a limited example of this shortly, and work will continue in this vein in the future.

It is also interesting to note that the response of adding descriptors to the model is somewhat different for the reduced data set. Figure 4 shows the parameter selection process repeated for the reduced set. Note there is less variability of the results in the "fully-fit" region, and also a small but steady improvement in performance with added parameters. This could indicate that the model for

the full dataset is hindered by bad data points, and thus oscillates while trying to fit the contradictory data, unable to improve with additional physical information.

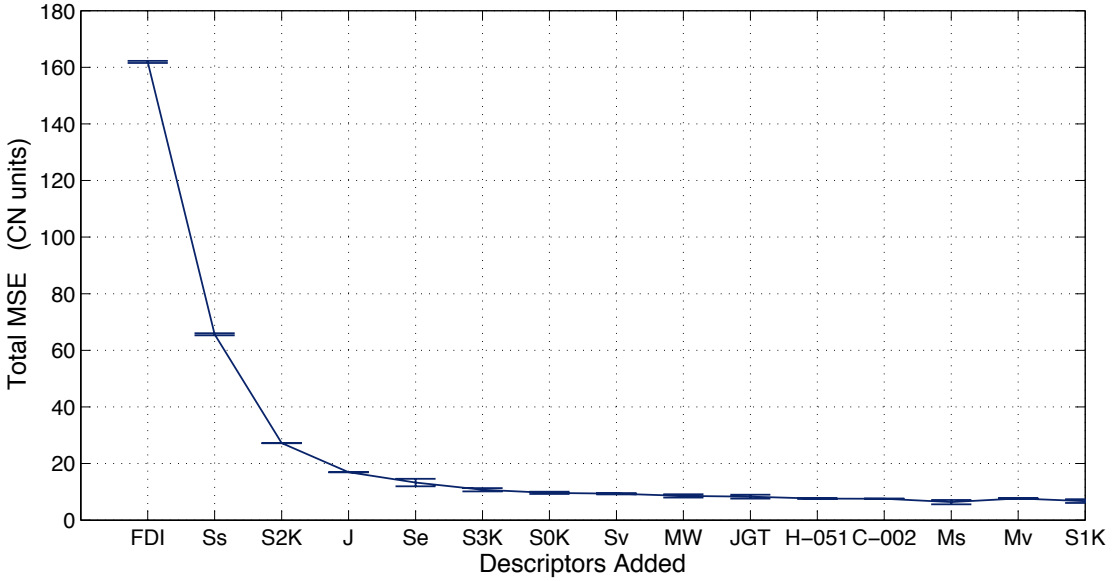

Figure 4: Iterative Addition of Descriptors to Model (Reduced Dataset)

3.4 Addition of New Compounds to Data Set

The Saldana database did not include ethers, aldehydes or ketones, although CN data on compounds from these groups were available in the compendium (21 ethers and 3 aldehydes/ketones) [4]. These values were added to the "reduced" set of HC's, alcohols and ethers, to form a new expanded set of 268 compounds. Given the future interest in furans and tetrahydrofurans, the ether compounds are especially important to include, even though there is greater uncertainty in the data given their tendencies to form peroxides in air [10]**.**

Additional data was received from NREL in the form of DCN numbers from high-precision IQT studies they have been conducting. Of these 52 compounds, 14 were totally new and added to the database, while 38 were values for compounds in the database. In most cases, these CNs matched with 4-5 CN units, but in some cases (~25%) they were greater than this, sometimes as high as 15 CN units. Some of the discrepancies may be attributed to these being DCNs instead of true CNs, but this should be investigated further, especially those with larger DCN, to determine if they indicate more compounds which should be omitted from our set.

It was desired to extend the model to these compounds purely as a predictive exercise, so except where noted, these additional compounds were reserved as "test" values and not used in training or validation.

Beginning with the parameter study shown in Figure 5, we can see that a greater number of parameters (11 vs. 8) are required, as would be expected to account for a wider variety of molecules.

It is also apparent that, as with the unreduced set, the standard deviations of the "fully developed" models are higher, indicating a greater amount of variation to the predictions, further indicating there are poorly predicted points, either due to bad data or to features not being captured by the model.

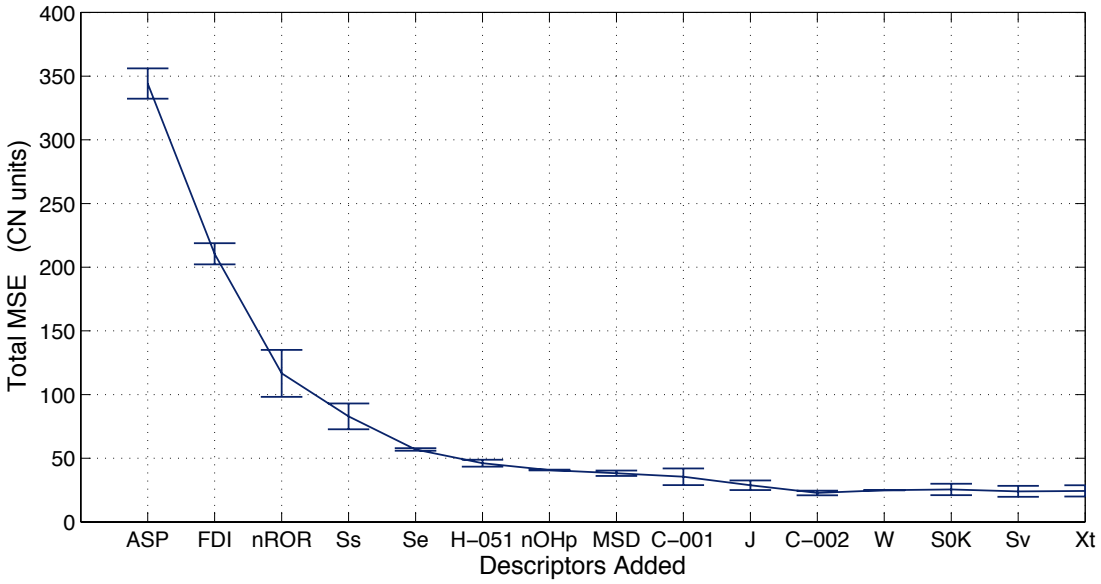

Figure 5: Iterative Addition of Descriptors to Model (Expanded Dataset)

Figure 6 shows the parity plot for the expanded dataset that presents a reduced accuracy with the new compounds (RMSE 6.3). Removing the constraint that the new DCN data from NREL be reserved as "test" values, we can improve the prediction slightly (RMSE 5.6), indicating there may not be enough data in the learning set without these compounds. Approximately 30 of the 38 compounds are predicted reasonably well (error less than 10 CN), with 8 being substantially mispredicted.

It is important to note that the poorly predicted compounds are not all new molecules; adding the new compounds also affects the performance of the network on the "original set". Considering the difficulty in oxygenate cetane number testing as well as the number of molecules discredited from the original dataset, it seems very likely that some of these new data points are unreliable. As noted, this can have the secondary effect of also reducing the predictive ability of the model on "good" data points; this is because the bad points may distort the network's knowledge of the effects of the input parameters.

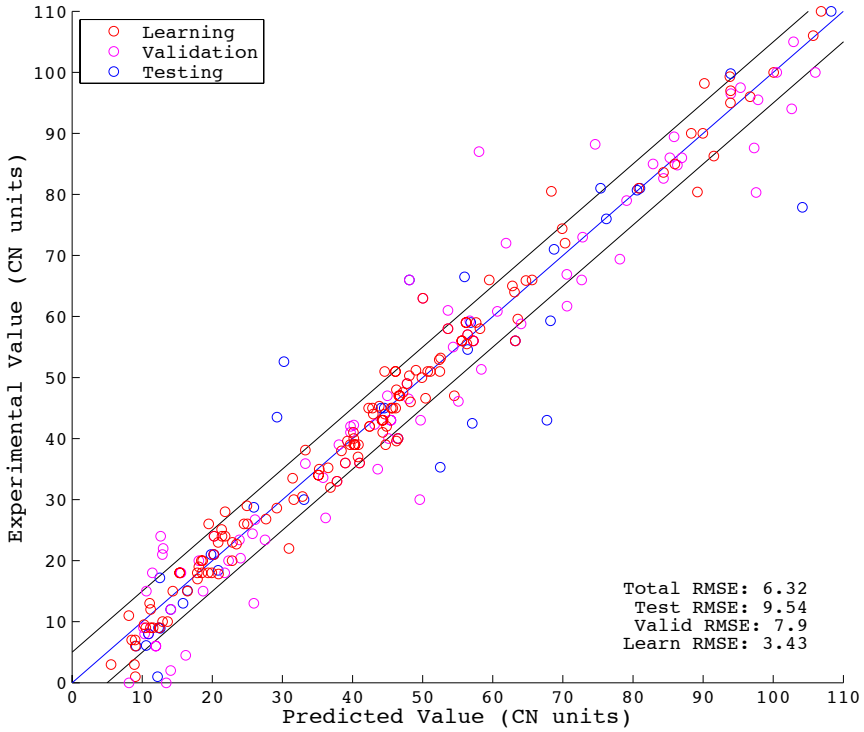

Figure 6: Parity Plot for 282 HC's, Esters, Alcohols and Ethers

The immediate next step for this research is a detailed study of the new compounds to determine what compounds may be removed. Care must be taken during this process given the small size of the new data set. If the resulting accuracy is still less than desired, an expanded parameter search may be necessary, or else different network topologies may need to be investigated.

It would also be desirable in the future to build tools to inspect the standard deviations of individual predictions between models, as this would give some better insight into how "stable" the predictions are, and possibly which molecules are presenting challenges to fit.

4. Conclusions

This work represents an interim step to developing a comprehensive model for predicting CN over a range of compounds including new oxygenates. The present model represents an improvement on current art, and is an effective tool for predicting HC's, alcohols and esters. It is also generally accurate for ethers although further improvement is desired and expected. Future work aims to improve and extend this model to new classes of compounds that are of interest from researchers, some of which are being tested presently or already have been tested, including furans and tetrahydrofurans. Eventually, this same methodology could be applied to other properties of interest (such as lubricity, viscosity, cloud point, etc.) to create a more comprehensive screening tool for desirable fuel candidates, provided sufficiently large data sets can be collected.

Acknowledgements

The authors would like to thank Romain Serres (ENSAM) and Robert Dibble (UCB) for their introduction to the topic and their prior work on predicting octane numbers of hydrocarbons and alcohols using neural networks. Brad Zigler and Matt Ratcliff (NREL) deserve great thanks for the additional DCN data they provided and their guidance on the project. We would also like to thank Alex Bell, Eric Sacia and the rest of the Bell group at Energy Biosciences Institute (EBI) for helping motivate the project's eventual goal of predicting new fuels, including furans and tetrahydrofurans. Finally, we would like to thank J.Y. Chen and Ben Wolk (UCB) for granting us computational time on one of their workstations.

References

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

[2] 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, Mar. 2008.

[3] 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, Oct. 2010.

[4] 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, Sep. 2011.

[5] M. J. Murphy, J. D. Taylor, and R. L. Mccormick, "Compendium of Experimental Cetane Number Data," 2004.

[6] 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.

[7] 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, Jun. 2005.

[8] R. Santana, P. Do, M. Santikunaporn, W. Alvarez, J. Taylor, E. Sughrue, and D. Resasco, "Evaluation of different reaction strategies for the improvement of cetane number in diesel fuels," *Fuel*, vol. 85, no. 5–6, pp. 643–656, Mar. 2006.

[9] R. Todeschini and V. Consonni, *Handbook of Molecular Descriptors*, 11th ed. Weinheim, Germany: , 2000.

[10] C. J. Mueller, W. J. Cannella, T. J. Bruno, B. Bunting, H. D. Dettman, J. a. Franz, M. L. Huber, M. Natarajan, W. J. Pitz, M. a. Ratcliff, and K. Wright, "Methodology for Formulating Diesel Surrogate Fuels with Accurate Compositional, Ignition-Quality, and Volatility Characteristics," *Energy & Fuels*, vol. 26, no. 6, pp. 3284–3303, Jun. 2012.