

**Proceedings of the ASME 2019 Internal Combustion Fall Technical Conference
ICEF2019
October 20-23, 2019, Chicago, IL, USA**

Screening Compounds for Fast Pyrolysis and Catalytic Biofuel Upgrading Using Artificial Neural Networks

Travis Kessler¹, Thomas Schwartz², Hsi-Wu Wong³, and J. Hunter Mack⁴

¹Department of Electrical and Computer Engineering, University of Massachusetts Lowell

²Department of Chemical and Biomedical Engineering, University of Maine

³Department of Chemical Engineering, University of Massachusetts Lowell

⁴Department of Mechanical Engineering, University of Massachusetts Lowell

ABSTRACT

There is significant interest among researchers in finding economically sustainable alternatives to fossil-derived drop-in fuels and fuel additives. Fast pyrolysis, a method for converting biomass into liquid hydrocarbons with the potential for use as fuels or fuel additives, is a promising technology that can be two to three times less expensive at scale when compared to alternative approaches such as gasification and fermentation. However, many bio-oils directly derived from fast pyrolysis have a high oxygen content and high acidity, indicating poor performance in diesel engines when used as fuels or fuel additives. Thus, a combination of selective fast pyrolysis and chemical catalysis could produce tuned bioblendstocks that perform optimally in diesel engines. The variance in performance for derived compounds introduces a feedback loop in researching acceptable fuels and fuel additives, as various combustion properties for these compounds must be determined after pyrolysis and catalytic upgrading occurs. The present work aims to reduce this feedback loop by utilizing artificial neural networks trained with quantitative structure-property relationship values to preemptively screen pure component compounds that will be produced from fast pyrolysis and catalytic upgrading. The quantitative structure-property relationship values selected as inputs for models are discussed, the cetane number and sooting propensity of compounds derived from the catalytic upgrading of phenol are predicted, and the viability of these compounds as fuels and fuel additives is

analyzed. The model constructed to predict cetane number has a test set prediction root-mean-squared error of 9.874 cetane units, and the model constructed to predict yield sooting index has a test set prediction root-mean-squared error of 13.478 yield sooting index units (on the unified scale).

1 INTRODUCTION

The emphasis on finding renewable and cleaner sources of energy has become prevalent throughout the greater scientific community due to concerns regarding global climate change and decreasing reserves of traditional petroleum-based fuels. Consequently, researchers have focused their efforts on discovering novel fuels derived from renewable sources (i.e. biofuels). One method for converting biomass into bio-oil is fast pyrolysis; however, not all compounds derived from fast pyrolysis perform optimally in existing engines and/or produce high amounts of negative byproducts, as do the compounds resulting from catalytically upgrading the products of fast pyrolysis. This highlights the need to screen compounds that will be produced from fast pyrolysis and catalytic upgrading before the procedures occur, ultimately providing researchers with more promising candidate compounds. To reduce the inherent feedback loop associated with alternative fuel research, machine learning (specifically artificial neural networks, ANNs) can be employed to predict the cetane number and yield sooting index of the products of fast pyrolysis and catalytic upgrading.

1.1 Background Information

Fast pyrolysis converts biomass into bio-oils by rapidly heating (>1000 °C/sec) the biomass to temperatures ranging from 400-600 °C for between 1-5 seconds and may yield up to 70 wt% of liquid phase products (bio-oils) [1]. It has been shown in recent techno-economic analyses that the cost to produce bio-oils via fast pyrolysis is comparable to the cost of producing bio-oils via alternative methods such as gasification, indicating fast pyrolysis is viable at scale [2]. The products of fast pyrolysis can range from alcohols and ethers to furanic and phenolic compounds depending on the feedstock utilized in the procedure [3]. Some products that form as a result of using lignocellulosic biomass as the primary feedstock in fast pyrolysis have been identified as potential replacements for fossil-derived fuel additives and drop-in fuels [4], however many of these compounds have a high oxygen content (indicating poor performance as a drop-in fuel additive). For these compounds with sub-optimal characteristics, selective catalytic upgrading may result in compounds that perform better as fuel additives while decreasing soot formation.

Catalytic upgrading is significant because it increases the hydrogen-to-carbon ratio, reduces oxygen content and increases molecular weights for oxygenated compounds derived from fast pyrolysis through a process called hydrodeoxygenation [5]. Many hydrocarbon products were observed as a result of hydrotreating oxygen-rich compounds using Pd/C catalysts at 310-375 °C and then subjecting them to hydrocracking at 400°C and 2000 PSI [6]. Hydrogenolysis of phenolic compounds using Ru/TiO₂ catalysts results in aromatic ring hydrogenation occurring on the Ru/Ti nanoparticle surface, in turn forming compounds with fewer oxygen atoms and that are aromatic in nature (both important qualities for indicating favorable performance as drop-in fuel additives) [7] [8]. The present work focusses on evaluating compounds resulting from catalytically upgrading products of fast pyrolysis using Ru/TiO₂ catalysts and phenol as an additive during the reaction.

Two properties will be evaluated for products of fast pyrolysis and catalytic upgrading – cetane number and yield sooting index. Cetane number (CN) measures a fuel's ignition quality in a diesel engine and is derived from the ignition delay after the fuel is injected. Diesel fuel typically has a CN of 40-50. Two common techniques for measuring CN are with an Ignition Quality Tester (IQT) and a Cooperative Fuel Research (CFR) engine. An IQT ascertains the ignition delay of a given fuel in a constant volume combustion chamber by measuring the time between injection and combustion [9]. Alternatively, a CFR engine utilizes two reference compounds, n-hexadecane and isocetane (with CN's of 15 and 100, respectively) – igniting a given fuel in an equivalent blend of these compounds results in a volume fraction of these compounds linear related to the fuel's CN [10]. A CFR is preferred over the IQT, as the

CFR reflects typical engine behavior; however, the IQT requires less test fuel (around 100 mL). While methods like the IQT utilize a small volume of fuel for determining CN, the number of tests required for determining CN's of a suite of potential alternative fuels results in a considerable time and monetary investment.

Various sooting indices are used to measure soot formation and how much particulate matter is emitted by a fuel during combustion. The Threshold Sooting Index (TSI) was developed to standardize the measurement of soot through smoke point and ranks fuels on a 0-100 scale using reference molecules [11]. Measuring smoke point involves measuring the maximum flame height attainable by a fuel combusting in a test lamp without smoking and has been shown to be a dependable indicator of sooting propensity of aviation fuels in turbines [12] as well as emissions from SI engines [13]. To account for reduced stoichiometric air required by oxygenated fuels, the oxygen extended sooting index (OESI) was defined as an extension to TSI [14]. While these smaller, bench-scale methods of measuring sooting propensity through smoke point are convenient, they suffer from a few disadvantages; operator bias in estimating an appropriate flame shape may occur, as well as the requirement of up to 20 mL of the fuel in order to measure [12].

Due in part to these disadvantages the yield sooting index (YSI) was developed, whose measurement is not based on smoke point, rather the maximum soot volume fraction measured in a flame ignited by a fuel doped with the molecule of interest [14]. YSI is measured on a 0-100 scale, using reference fuels n-hexane and benzene with YSI values of 0 and 100 respectively. Diesel fuel typically has a YSI of 235-250. This method of measurement requires a significantly smaller volume of the sample, and correlates adequately with TSI [15] [16] [17] [18]. A "high sooting" YSI scale was also developed to emphasize mass-fraction-based fuel doping, and recent research has provided a unified YSI scale, standardizing measurements from a variety of compounds and compound groups [19]. Regardless of the advances made in measuring soot formation, a considerable time investment exists when testing a large number of compounds.

1.2 Predicting Cetane Number & Yield Sooting Index

Using computational techniques to predict CN has a pervasive history and includes a variety of methods. Such methods include consensus modeling, where linear and non-linear models are employed in parallel to obtain an averaged predicted CN value – these models can predict CN for a variety of molecular classes with a blind prediction root-mean-squared error of 6.5 [20], however multiple linear models are surpassed by the accuracy of ANNs when fatty acid methyl esters are used to predict CN [21]. Many methods rely on quantitative structure-property relation-

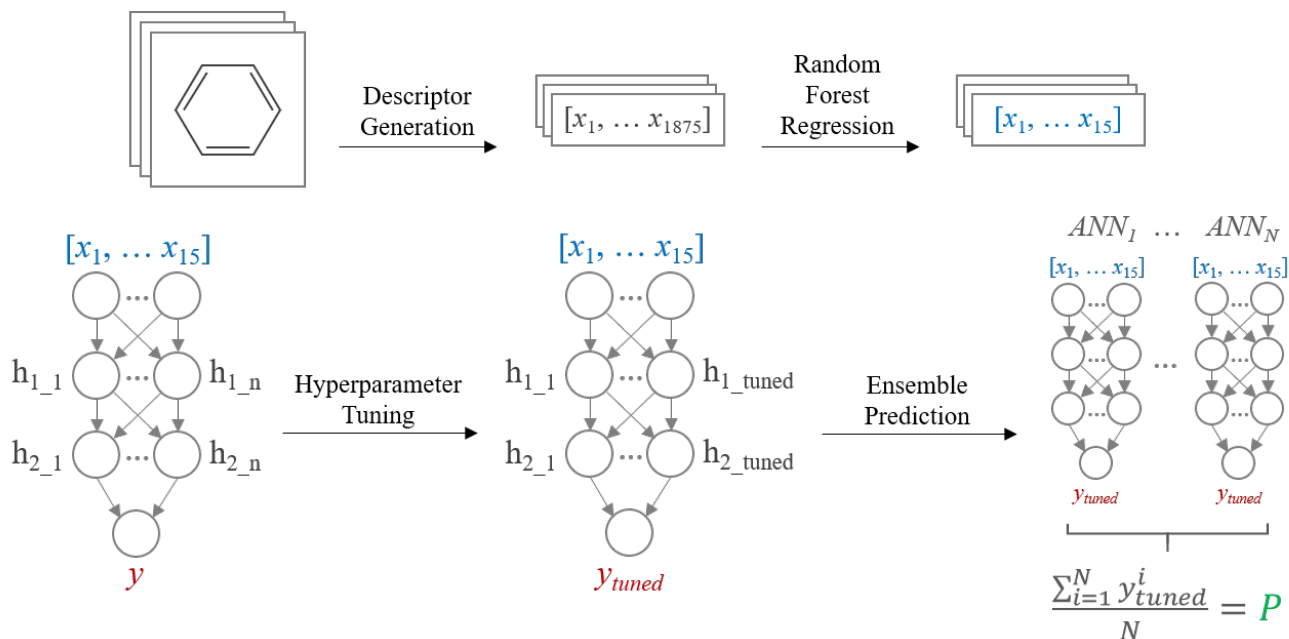


Figure 1: Workflow diagram illustrating the experimental procedure utilized by the present work

ship (QSPR) descriptors, which are numerical measurements of an assortment of physical and chemical properties relating to molecules, and have proven to be successful when applied to predicting the CN of pure hydrocarbons [22] and branched paraffins [23]. Additionally, ANNs have been applied to predicting YSI using QSPR descriptors as training data and obtained 95% confidence in test set prediction accuracy (r-squared value of predictions for compounds not used in ANN training) [24].

In accordance with these findings, this paper utilizes feed-forward ANNs trained with QSPR descriptors to predict CN and unified YSI. The present work leverages the unified YSI scale, as the models constructed aim to predict YSI for a variety of compounds and compound groups. It has recently been shown that ANNs can extend their predictive capabilities to a variety of molecular classes, including predicting the CN of biomass-derived furanic compounds [25]. ANNs provide a non-linear model architecture, allowing a multi-dimensional input vector containing a suite of individual QSPR values to be correlated to an experimental property value. QSPR descriptors are utilized due to the wide range of physical and chemical property representations available, subsequently distinguishing one molecule from another [26].

Traditionally, QSPR descriptors used as ANN inputs are chosen via an iterative regression analysis technique, where QSPR descriptors are added to the ANN based on the ANN's performance when the value is used in conjunction with previously selected values [25]. While this method highlights the ANN's ability to determine complex relationships between QSPR descriptor values as they are included, it fails to provide an explanation as to why/how individual QSPR descriptors correlate to a given property. Random

forest regression has been shown to be a viable method for reducing the number of quantitative structure-activity relationship values used as inputs in a binary classification decision tree while retaining predictive accuracy [27], and provide a quantifiable measurement (importance) of individual value correlation to a given target value [28]. Therefore, the present work leverages importance resulting from random forest regression to select QSPR descriptors as ANN inputs.

2 MATERIALS AND METHODS

The experimental procedure followed in the present work is illustrated in Figure 1. Each of these tasks is outlined in more detail in the following subsections. In summary:

1. QSPR descriptors were generated for training/testing data
2. Random forest regression was utilized to select highly influential QSPR descriptors
3. Hyperparameters for ANNs using highly influential descriptors as input variables were tuned to optimize performance
4. An ensemble of selected candidate ANNs was employed to obtain a prediction

2.1 Experimental Data

Experimental CN data was obtained from the NREL Compendium of Experimental Cetane Number data [29] and

Table 1: Products/wt% of fast pyrolysis of lignocellulosic biomass, catalytically upgrading with Ru/TiO₂ and Phenol

Fast Pyrolysis Product	wt% of FP	Catalytically Upgraded Product
phenol	0-5	phenoxybenzene
methanol	0-1	anisole
1-hydroxypropan-2-one	0-1	1-phenoxy-2-propanone
furan-2-ylmethanol	0-1	2-phenoxyethylfuran
5-(hydroxymethyl)furan-2-carbaldehyde	0-5	5-phenoxyethylfurfural
2-hydroxy-cyclopent-2-en-1-one	0-1	2-phenoxy-2-cyclopentene-1-one
2-methoxy phenol	0-5	1-methoxy-2-phenoxybenzene
2-methyl phenol	0-5	1-methyl-2-phenoxybenzene
4-methyl phenol	0-5	1-methyl-4-phenoxybenzene
2-methoxy-4-methyl phenol	0-5	2-methoxy-4-methyl-1-phenoxybenzene
3,5-dimethyl phenol	0-5	1,3-dimethyl-5-phenoxybenzene
3-ethyl phenol	0-5	1-ethyl-3-phenoxybenzene
4-ethyl-2-methoxy phenol	0-5	1-ethyl-4-phenoxybenzene

other sources [20] [30] [31] totaling in 445 unique compounds. Methods for obtaining experimental CN values include derivations from blend measurements, the use of an IQT/CFR and other, unknown ignition delay methods. Most experimental values were obtained using an IQT/CFR, as these methods are more accurate than blending/other methods. Experimental YSI data was obtained from a variety of sources [15] [16] [17] [18], was measured using laser-induced incandescence and other gas-phase steady-state measurement techniques, and totaled in 421 compounds.

Simple molecular-input line-entry system (SMILES) strings were obtained for all compounds comprising the CN and YSI databases using MarvinSketch [32] and validated using compound entries on PubChem [33]. SMILES strings were then converted to MDL Molfiles using Open Babel to generate three-dimensional geometry for the compound [34], and the Molfiles were fed into PaDEL-Descriptor to generate 1444 1D/2D and 431 3D QSPR descriptors for each compound [35]. QSPR descriptor values and experimental property values comprise input and target data respectively during ANN training.

Table 1 displays a list of compounds that are expected be produced by performing fast pyrolysis on lignocellulosic biomass, and includes phenolic compounds, furanic compounds and benzenes, with expected wt%s [36][37]. Also included in Table 1 are the expected products of performing hydrogenolysis using Ru/TiO₂ catalysts on each compound resulting from fast pyrolysis, with phenol included in the catalysis process. For each compound listed in Table 1, QSPR descriptors were generated using the previously mentioned techniques. The purpose of this exercise is to focus on a particular byproduct of fast pyrolysis and determine whether it can be effectively upgraded into a more suitable fuel using catalysis.

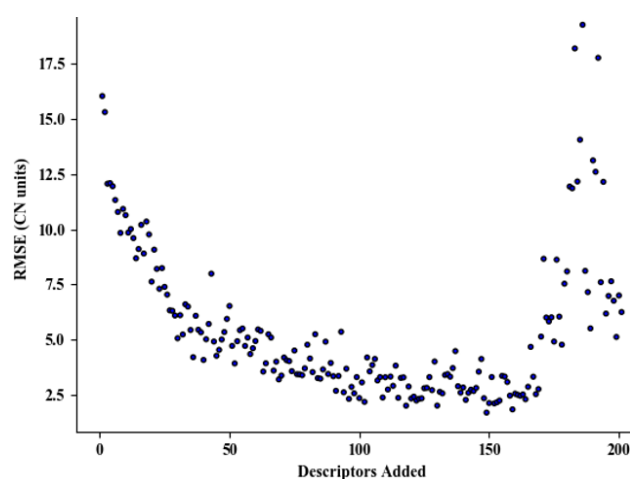


Figure 2: RMSE of CN predictions vs. number of important descriptors added to the ANN's input

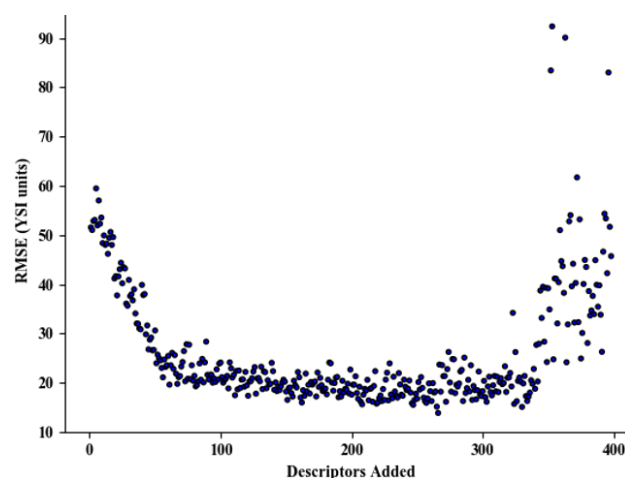


Figure 3: RMSE of YSI predictions vs. number of important descriptors added to the ANN's input

2.2 QSPR Descriptor Selection

Random forest regression using Scikit-learn was employed to determine QSPR descriptor "importance", a numerical measurement indicating correlation between descriptor values and experimental property values for CN and YSI – the higher the importance of a descriptor, the more it contributes to a correlation to a given property [28]. The sum of all descriptor importances is equal to one. It was found that 15-25 descriptors balances computation time and ANN predictive accuracy. Appendix Tables A1 and A2 show the selected QSPR descriptors for CN and YSI respectively. The descriptors selected to represent CN and YSI were used as input variables in ANNs constructed for predicting CN and YSI, both during the hyperparameter tuning process and candidate training process.

Figure 2 illustrates ANN performance (RMSE of CN predictions for training data after 2500 epochs) versus the number of important descriptors added as inputs to the ANN, and Figure 3 illustrates the same behavior for YSI. Performance for CN degrades significantly past 150 descriptor additions, and performance for YSI degrades significantly past 300 descriptor additions. Degrading performance can be attributed in part to the number of constant-value descriptors, as they are pernicious to the ANN; i.e., the ANN is unable to determine any relationship between them and a given experimental value. For example, "Nn", or the number of nitrogen atoms present in the compound, plays no significant role during training as all training data is comprised of hydrocarbons. Additionally, descriptor importances where performance degradation occurs are very small in comparison to descriptors with high importance. The descriptor ranked 150th in importance for CN has an importance of 0.00044, and the descriptor ranked 300th in importance for YSI has an importance of 0.000029. Descriptors with very low importances likely contribute to performance degradation, as the relationships between their values and an experimental property value are more abstract than values of more important descriptors.

2.3 ANN Hyperparameter Tuning

ANNs were trained using the Adam optimization function, which possesses five hyperparameters that affect the quality of the ANN’s training [38]:

- β_1 : exponential decay rate 1 for moment estimates
- β_2 : exponential decay rate 2 for moment estimates
- ϵ : number to prevent division by zero in the algorithm implementation
- α : learning rate (i.e. step size)
- learning rate decay: degradation of learning rate after each learning epoch

Table 2: Tuned hyperparameters for CN model

Hyperparameter	Tuned Value
β_1	0.0482
β_2	0.8011
ϵ	0.2752
α	1.0
Decay	0.2710
Neurons in 1st hidden layer	35
Neurons in 2nd hidden layer	31

Table 3: Tuned hyperparameters for YSI model

Hyperparameter	Tuned Value
β_1	0.3241
β_2	0.6434
ϵ	0.9485
α	0.2622
Decay	0.0145
Neurons in 1st hidden layer	39
Neurons in 2nd hidden layer	10

In addition to these five training variables, the optimal number of neurons in the ANN’s hidden layer(s) must be determined. Given the number of hyperparameters that must be manually tuned under normal circumstances, an artificial bee colony (ABC) was utilized to algorithmically determine the optimal sets of hyperparameters for both CN and YSI models. ABCs mimic the foraging behavior of honeybees to search a multidimensional search space of tunable variables and have been shown to out-perform genetic algorithms and other particle swarm optimization algorithms in tuning various hyperparameters in ANNs [39].

The ABC was supplied with a fitness function to determine the ability of ANNs to predict either CN or YSI for test set molecules, where a lower test set RMSE resulting from a given set of hyperparameters was deemed better performing. The fitness function supplied to the ABC constructed an ANN using the QSPR descriptors selected with random forest regression as input variables and was trained for 2000 epochs. 70% of experimental data was used in the learning set, and 20% of experimental data was used in the validation set. Learning/validation data was shuffled for each bee to mimic the ensuing candidate ANN training procedure. Test set data, the remaining 10%, remained constant throughout the hyperparameter tuning process. The same test set data was used during candidate ANN training. The ABC was run for 20 search cycles with 50 employer bees for both CN and YSI datasets. Tables 2 and 3 show the tuned hyperparameters for CN and YSI respectively given the use of the Adam optimization function and two hidden layers. These hyperparameters were used during candidate ANN training and ensemble model construction.

2.4 Candidate ANN Training/Selection

ANN training was performed with ECNet, an open source machine learning toolkit created to predict fuel properties of potential next-generation fuels [40]. A model is considered a collection of ANNs whose predictions are averaged to obtain a final prediction, producing a more accurate prediction in a similar fashion to classifier ensembles [41]. Each ANN was chosen from a pool of ANN candidates, where the pool's goal was to optimally predict either CN or YSI. The ANN that was chosen from each pool achieved the lowest RMSE in predicting unseen data across all the pool's candidates. 5 pools, each with 75 candidate ANNs were trained for both CN and YSI.

Each candidate ANN was supplied with a random learning set and a random validation set, 70% and 20% of the total data respectively, while the remaining 10% of data remained constant for all candidates to measure their performance in predicting unseen data. The ANNs were trained using the backpropagation algorithm and the learning set, while the validation set measured the progress of the ANNs' learning. Once performance stopped improving on the validation set, learning was terminated – this was done to prevent any overfitting on the learning set. Figure 4 illustrates predictive accuracy (mean absolute error of CN predictions) versus the number of ANN training epochs for the learning, validation and test sets. In this example, training would terminate after 3000 epochs. Establishing an early stopping point for training has proven to be successful in allowing ANNs to generalize predictions for unseen data [42]. Shuffling learning/validation sets allowed each ANN to learn from a different representation of data, and each ANN's predictions are unique (often predicting slightly higher or slightly lower than the known experimental value). Consequently, the final prediction of the model tends to be more accurate. Model performance was measured by determining RMSE and the r-squared correlation coefficient when predicting unseen data (the constant test set).

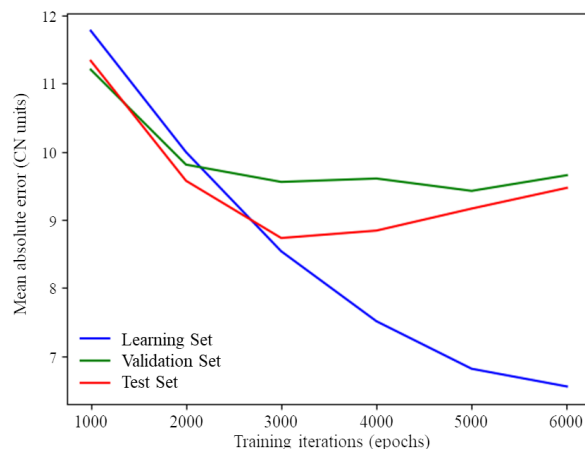


Figure 4: Mean absolute error of CN predictions vs. training epochs

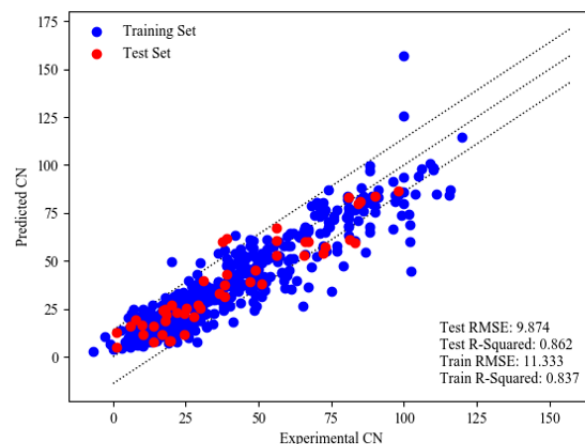


Figure 5: CN parity plot illustrating predicted values vs. experimental values

3 RESULTS AND DISCUSSION

Figures 5 and 6 show parity plots between predicted values and experimental values for CN and YSI respectively, illustrating the performance of the training (learning and validation) set and the blind test set. The center dashed lines illustrate a 1:1 parity between predicted values and experimental values, and the outer dashed lines show bounds imposed by the test set's RMSE. Test set RMSE and r-squared are significantly better for both CN and YSI, as candidate ANN's were selected based on their ability to predict unseen data.

It is worth noting that experimental data, notably experimental CN values, may have inherent error associated with the experimental methods used to obtain the data. For example, some CN values used in the present work that were

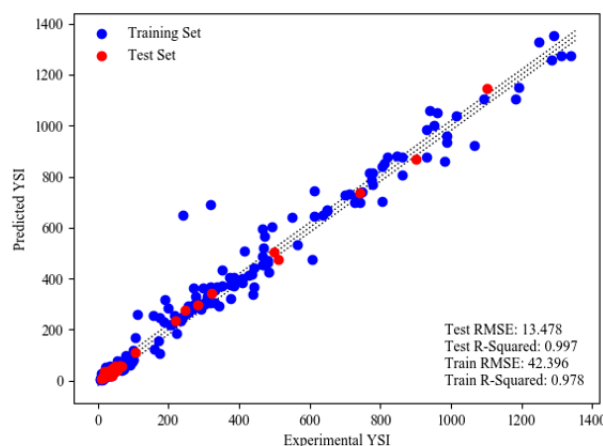


Figure 6: YSI parity plot illustrating predicted values vs. experimental values

Table 4: Predicted CN/YSI for products of fast pyrolysis

Compound Name	CN (± 9.874)	YSI (± 13.478)
phenol	1.84	161.71
methanol	9.10	1.55
1-hydroxypropan-2-one	8.76	14.00
2-furanmethanol	7.16	81.89
5-hydroxy methyl furfural	7.02	215.34
2-hydroxy-cyclopent-2-en-1-one	7.82	71.24
2-methoxy phenol	8.75	330.34
2-methyl phenol	2.44	237.00
4-methyl phenol	2.79	244.16
2-methoxy-4-methyl phenol	8.38	403.78
3,5-dimethyl phenol	1.42	357.99
3-ethyl phenol	5.60	315.51
4-ethyl-2-methoxy phenol	11.61	479.04

measured with an IQT are on average 4.8 CN units different than CN values for the same compounds measured with a CFR [29]. Other ignition delay methods are more drastic, having an average difference of 15.7 compared to IQT measurements. ANNs are reliant on the accuracy of training data, and it is currently unknown how errors in experimental CN/YSI values affect ANN training (other than reducing predictive accuracy overall).

Tables 4 and 5 display CN and YSI predictions for the products of fast pyrolysis and catalytic upgrading respectively. Potential error in these predictions is defined by the properties' respective test set RMSEs. As seen by the disparity between pre- and post-catalytic upgrading CN values, upgraded compounds such as phenoxybenzene (a result of phenol/phenol upgrading), 1-methoxy-2-phenoxybenzene (a result of 2-methoxy phenol/phenol upgrading) and 2-methoxy-4-methyl-1-phenoxybenzene (a result of 2-methoxy-4-methyl phenol/phenol upgrading) have significantly higher CNs. Furan/phenol-based upgrades also have higher CNs. Ethyl/phenol-based upgrades did not improve CN significantly. All upgraded compounds have significantly higher YSIs than their pre-upgraded products except for phenoxybenzene.

Figures 7 and 8 display known property values versus values of the QSPR descriptor with the highest importance for CN and YSI respectively (CN vs. RotBFrac, YSI vs. piPC5). An exponential trend is apparent in both relationships, represented by Equations 1 and 2.

$$CN = 0.435e^{5.822(RotBFrac)} + 19.131 \quad (1)$$

$$YSI = 6.732e^{0.912(piPC5)} + 13.980 \quad (2)$$

If Equations 1 and 2 are used to predict the CN and YSI for known test set values, CN predictions have an r-squared correlation coefficient of 0.706 and YSI predictions

Table 5: Predicted CN/YSI for products of cat. upgrading

Compound Name	CN (± 9.874)	YSI (± 13.478)
phenoxybenzene	12.01	168.85
anisole	9.71	649.02
1-phenoxy-2-propanone	11.11	342.27
2-phenoxyethylfuran	12.11	451.96
5-phenoxyethylfurfural	12.94	525.03
2-phenoxy-2-cyclopentene-1-one	11.80	541.33
1-methoxy-2-phenoxybenzene	16.06	689.18
1-methyl-2-phenoxybenzene	6.77	684.55
1-methyl-4-phenoxybenzene	5.83	696.12
2-methoxy-4-methyl-1-phenoxybenzene	17.06	746.00
1,3-dimethyl-5-phenoxybenzene	5.75	747.66
1-ethyl-3-phenoxybenzene	8.90	714.75
1-ethyl-4-phenoxybenzene	8.51	718.21

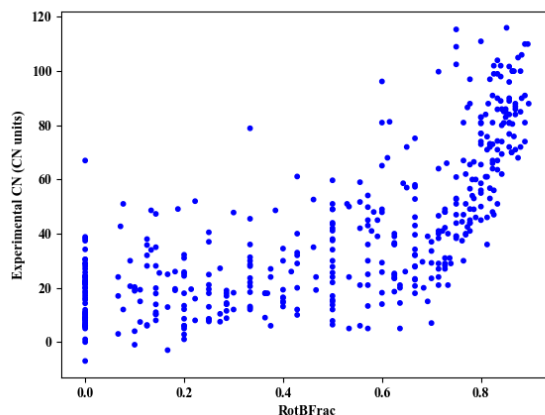


Figure 7: Relationship between CN and RotBFrac

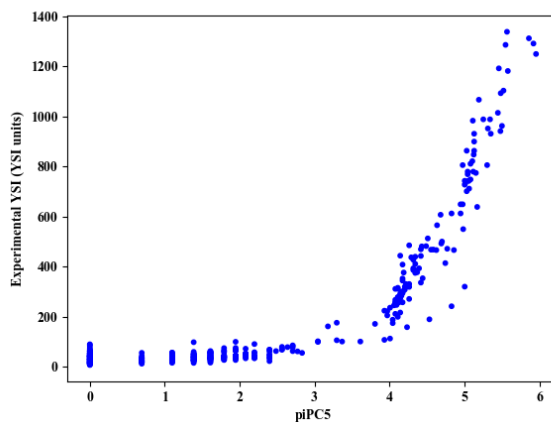


Figure 8: Relationship between YSI and piPC5

have an r-squared correlation coefficient of 0.979. While it appears the equation derived from piPC5 is relatively robust in its ability to provide accurate YSI predictions, the considerably lower r-squared value resulting from predicting CN using RotBFrac compared to using an ANN highlights the ANN's ability to infer information about compounds from a multidimensional input vector of QSPR descriptors.

Based on the relationship between experimental CN and RotBFrac (fraction of non-terminal rotatable bonds), compounds with an abundance of rotatable bonds tend to have higher CN. This correlates well with investigations into the bulk modulus of fuels, where a higher bulk modulus resulting from fewer rotatable bonds in saturated compounds indicates a fuel is relatively incompressible [43]. This incompressibility leads to a faster injection time, increasing the ignition delay, and results in a lower cetane number.

Various autocorrelation indices, such as GATS and ATSC, were selected as important descriptors to represent CN. These autocorrelation indices measure the resemblance of neighboring point values, and when applied to the analysis of a given compound indicate repeating patterns in the compound [44]. The higher the value of the autocorrelation index, the more significant a resemblance in two neighboring points (atoms) is [45]. The various weightings of the selected autocorrelation indices, being electronegativities, mass and charges, likely indicate that patterns of electron and mass distributions play a role in determining a compound's CN.

The relationship between experimental YSI and piPC5 (a measurement of atom path lengths in a given compound) indicates that longer chains of atoms and larger compounds correlate to higher soot formation. This observation aligns with existing studies indicating fuel blends containing long carbon chains tend to emit more particles as a result of combustion [46]. 8 of 15 descriptors selected for YSI measure path length of varying orders, strengthening this argument.

Figures showing relationships between properties and selected QSPR descriptors were included in Appendix Figures F1-F30.

4 CONCLUSIONS

The models proposed in the present work can generalize CN/YSI predictions for molecules not observed during training with RMSEs of 9.874 and 13.478 respectively. CN/YSI of expected products of fast pyrolysis and catalytic upgrading were predicted, and pre- and post-upgraded products were compared. Additionally, the QSPR descriptor selection methodology utilized in the present work offers a considerable amount of insight into individual QSPR descriptor/experimental value relationships.

From the predicted CN/YSI values displayed in Tables 4 and 5, it can be concluded that:

- Catalytically upgrading phenol, methoxy-phenolic, and furanic compounds using Ru/TiO₂ catalysts and

phenol during the catalysis process yields compounds with higher CN values, likely attributed to the lower oxygen content of the products.

- Catalytically upgrading ethyl-based compounds using Ru/TiO₂ catalysts and phenol during the catalysis process yields compounds with no significant improvement in CN values.
- Catalytically upgrading products of fast pyrolysis using Ru/TiO₂ catalysts and phenol during the catalysis process yields compounds with significantly higher sooting propensity, with the exception of diphenyl ether.

From analyzing experimental CN/YSI values and individual selected QSPR descriptor values, it can be concluded that:

- There is a positive correlation between the fraction of non-terminal rotatable bonds and CN.
- Autocorrelation indices, corresponding to patterns of atoms in compounds based on electronegativity, mass and charge, affect the CN of a given compound.
- There is a strong, positive correlation between atom path length and sooting propensity.

Based on these findings, further pursuit of catalytically upgrading phenolic/furanic compounds in an experimental setting is recommended. Additionally, different additives during the catalysis process besides phenol should be assessed, such as furanic additives. In particular, this study has shown that a single byproduct of fast pyrolysis (phenol) can be upgraded into hydrocarbons with more suitable CN values at the expense of YSI. Future research will investigate the entire matrix of bio-oil constituents and possible catalytic products.

Additional investigations into why the selected QSPR descriptors contribute to CN/YSI from a chemical standpoint, and what role upgraded compounds play in a mixture of traditional petroleum-based fuel should be performed.

5 ACKNOWLEDGEMENTS

This work was funded by the U.S. Department of Energy under award DE-EE0008479 as part of the Co-Optimization of Fuels & Engines (Co-Optima) project sponsored by the U. S. Department of Energy (DOE) Office of Energy Efficiency and Renewable Energy (EERE) Bioenergy Technologies Office (BETO).

References

- [1] D. M. Alonso, J. Q. Bond, J. A. Dumesic. *Catalytic conversion of biomass to biofuels*. Green Chemistry, vol. 12, pp. 1493-1513, 2010
- [2] R. P. Anex, A. Aden, F. K. Kazi, J. Fortman, R. M. Swanson, M. M. Wright, J. A. Satrio, R. C. Brown, D. E. Daugaard, A. Platon, G. Kothandaraman, D. D. Hsu, A. Dutta. *Techno-economic comparison of biomass-to-transportation fuels via pyrolysis, gassification, and biochemical pathways*. Fuel, vol. 89(S1), pp. S29-S35, 2010
- [3] P. R. Patwardhan, J. A. Satrio, R. C. Brown, B. H. Shanks. *Influence of inorganic salts on the primary pyrolysis products of cellulose*. Bioresource Technology, vol. 101(12), pp. 4646-4655, 2010
- [4] S. P. S. Chundawat, G. T. Beckham, M. E. Himmel, B. E. Dale. *Deconstruction of lignocellulosic biomass into fuels and chemicals*. Annual Review of Chemical and Biomolecular Engineering, vol. 2, pp. 121-145, 2011
- [5] J. D. Adjaye, N. N. Bakhshi. *Production of hydrocarbons by catalytic upgrading of a fast pyrolysis bio-oil*. Fuel Processing Technology, vol. 45(3), pp. 161-202, 1995
- [6] D. C. Elliott, T.R. Hart, G. G. Neuenschwander, L. J. Rotness, A. H. Zacker. *Catalytic hydroprocessing of biomass fast pyrolysis bio-oil to produce hydrocarbon products*. Env. Prog. & Sustainable Energy, vol. 28(3), pp. 441-449, 2009
- [7] C. Newman, X. Zhou, B. Goundie, I. T. Ghampson, R. A. Pollack, Z. Ross, M. C. Wheeler, R. W. Meulenbergh, R. N. Austin, B. G. Frederick. *Effects on support identity and metal dispersion in supported ruthenium hydrodeoxygenation catalysts*. Applied Catalysts A: General, vol. 477, pp. 64-74, 2014
- [8] R. C. Nelson, B. Baek, P. Ruiz, B. Goundie, A. Brooks, M. C. Wheeler, B. G. Frederick, L. C. Grabow, R. N. Austin. *Experiemntal and Theoretical Insights into the Hydrogen-Efficient Direct Hydrodeoxygenation Mechanism of Thenol over Ru/TiO₂*. ACS Catalysts, vol. 5, pp. 6509-6523, 2015
- [9] 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
- [10] ASTM 613-16a, *Standard Test Method for Cetane Number of Diesel Fuel Oil*. ASTM International, West Conshohocken, PA, 2016, www.astm.org
- [11] H. Calcote, D. Manos. *Effect of molecular structure on incipient soot formation*. Combust. Flame, vol. 49, pp. 289-304, 1983
- [12] D1322, *A. Test Method for Smoke Point of Kerosine and Aviation Turbine Fuel*. 2015
- [13] E. J. Barrientos, J. E. Anderson, M. M. Maricq, A. L. Boehman. *Particulate matter indices using fuel smoke point for vehicle emissions with gasoline, ethanol blends, and butanol blends*. Combust. Flame, vol. 167, pp. 308-319, 2016
- [14] E. J. Barrientos, M. Lapuerta, A. L. Boehman. *Group additivity in soot formation for the example of C-5 oxygenated hydrocarbon fuels*. Combust. Flame, vol. 160, 2013
- [15] C. S. McEnally, L. D. Pfefferle. *Improved sooting tendency measurements for aromatic hydrocarbons and their implications for naphthalene formation pathways*. Combust. Flame, vol. 148, pp. 210-222, 2007
- [16] C. S. McEnally, L. D. Pfefferle. *Sooting tendencies of nonvolatile aromatic hydrocarbons*. P. Combust. Inst., vol. 32, pp. 673-679, 2009
- [17] C. S. McEnally, L. D. Pfefferle. *Sooting Tendencies of Oxygenated Hydrocarbons in Laboratory-Scale Flames*. Environ. Sci. Technol., vol. 45, pp. 2498-2503, 2011
- [18] D. D. Das, C. S. McEnally, L. D. Pfefferle. *Sooting tendencies of unsaturated esters in nonpremixed flames*. Combust. Flame, vol. 162, pp. 1489-1497, 2015
- [19] D. D. Das, P. C. St. John, C. S. McEnally, S. Kim, L. D. Pfefferle. *Measurements and Prediction of Sooting Tendencies of Pure Hydrocarbons*. 2017
- [20] D. A. Saldana, L. Starch, P. Mougin, B. Rousseau, L. Pidol, N. Jeuland, B. Creton. *Flash Point and Cetane Number Prediuctions for Fuel Compounds Using Quantitative Structure Property Relationship (QSPR) Methods*. Energy and Fuels, Vol. 25, No. 9, pp. 3900-3908, 2011
- [21] R. Piloto-Rodriguez, Y. Sanchez-Borroto, M. Lapuerta, L. Goyos-Perez, S. Verhelst. *Prediction of cetane number of biodiesel using artificial neural networks and multiple linear regression*. Energy Conversion and Management 65, pp. 255-261, 2013
- [22] E. A. Smolenskii, V. M. Bavykin, A. N. Ryzhov, O. L. Slovokhotova, I. V. Chuvaeva, A. L. Lapidus. *Cetane number of hydrocarbons: calculations using optimal tological indices*. Russian Chemical Bulletin, Vol. 57, No. 3, pp. 461-467, 2008

- [23] H. Yang, C. Fairbridge, Z. Ring. *Neural Network Prediction of Cetane Number of iso-Paraffins and Diesel Fuel*. Petroleum Science and Technology, Vol. 19, No, 5-6, pp. 573-586, 2001
- [24] P. C. St. John, P. M. Kairys, D. D. Das, C. S. McNally, L. D. Pfefferle, D. J. Robichaud, M. R. Nimlos. *A quantitative model for the prediction of sooting tendency from molecular structure*. Energy and Fuels, 2017
- [25] T. Kessler, E. R. 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 Division Fall Technical Conference, 2016
- [26] 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
- [27] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, B. P. Feuston. *Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling*. J. Chem. Inf. Comput. Sci., vol. 43, pp. 1947-1958, 2003
- [28] L. Breiman. *Random forests*. Journal of Machine Learning, vol. 45, pp. 5-32, 2001
- [29] J. Yanowitz, M. A. Ratcliff, R. L. McCormick, J. D. Taylor, M. J. Murphy. *Compendium of Experimental Cetane Numbers*. NREL/TP-5400-61693, 2014
- [30] J. Taylor, R. McCormick, W. Clark. *Report on the relationship between molecular structure and compression ignition fuels*. NREL Technical Report, 2014
- [31] M. Dahmen, W. Marquardt. *A Novel Group Contribution Method for the Prediction of the Derived Cetane Number of Oxygenated Hydrocarbons*. Energy and Fuels, vol. 29(9), pp. 5781-5801, 2015
- [32] ChemAxon, 2015. *MarvinSketch, Version 15.10.19.0*. <http://www.chemaxon.com>
- [33] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, E. E. Bolton. *Pubchem 2019 update: improved access to chemical data*. Nucleic Acids Res. 2019 Jan 8; 47(D1):D1102-1109
- [34] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, G. R. Hutchinson. *Open Babel: An open chemical toolbox*. Journal of Cheminformatics, vol. 3, pp. 33, 2011
- [35] Yap CW. *PaDEL-Descriptor: An open source software to calculate molecular descriptors and fingerprints*. Journal of Computational Chemistry. 32 (7): pp. 1466-1474, 2011
- [36] P. R. Patwardhan, R. C. Brown, B. H. Shanks. *Understanding the Fast Pyrolysis of Lignin*. ChemSusChem, vol. 4(11), pp. 1629-1636, 2011
- [37] X. Zhou, M. W. Nolte, H. B. Mayes, B. H. Shanks, L. J. Broadbelt. *Experimental and Mechanistic Modeling of Fast Pyrolysis of Neat Glucose-Based Carbohydrates. 1. Experiments and Development of a Detailed Mechanistic Model*. Industrial & Engineering Chemistry Research, vol. 53(34), pp. 13274-13289, 2014
- [38] D. P. Kingma, J. Ba. *Adam: A Method for Stochastic Optimization*. International Conference for Learning Representations, 2015
- [39] Z. G. Cam, S. Cimen, T. Yildirim. *Learning parameter optimization of Multi-Layer Perceptron using Artificial Bee Colony, Genetic Algorithm and Particle Swarm Optimization*. IEEE 13th International Symposium on Applied Machine Intelligence and Informatics, 2015
- [40] T. Kessler, J. H. Mack. *ECNet: Large scale machine learning projects for fuel property prediction*. Journal of Open Source Software, 2(17), pp. 401, 2017
- [41] L. Kincheva, C. Whitaker. *Measures of diversity in classifier ensembles*. Machine Learning, vol. 51, pp. 181-207, 2003
- [42] L. Prechelt. *Automatic Early Stopping Using Cross Validation: Quantifying the Criteria*. Neural Networks, vol. 11(4), pp. 761-767, 1998
- [43] M. A. Hess, M. J. Haas, T. A. Foglia. *Attempts to reduce NO_x exhaust emissions by using reformulated biodiesel*. Fuel Processing Technology, vol. 88(7), pp. 693-699, 2007
- [44] S. A. Klein, C. W. Tyler. *Phase discrimination of compound gratings: generalized autocorrelation analysis*. Journal of the Optical Society of America A, vol. 3(6), pp. 868-879, 1986
- [45] D. Chessel. *The spatial autocorrelation matrix*. Vegetation dynamics in grasslands, healthlands and mediterranean ligneous formations, pp. 177-180, 1981
- [46] M. M. Rahman, S. Stevanovic, M. A. Islam, K. Heimann, M. N. Nabi, G. Thomas, B. Feng, R. J. Brown, Z. D. Ristovski. *Particle emissions from microalgae biodiesel combustion and their relative oxidative potential*. Environ. Sci.: Processes Impacts, vol. 17, pp. 1601-1610, 2015

Appendix

Table A1: QSPR descriptors selected for CN, their importances, descriptions

Descriptor Name	Importance	Description
RotBFrac	0.5747	Fraction of rotatable bonds, excluding terminal bonds
AVP-5	0.052	Average valence path, order 5
GATS2e	0.0159	Geary autocorrelation - lag 2 / weighted by Sanderson electronegativities
GATS2m	0.0137	Geary autocorrelation - lag 2 / weighted by mass
GATS2c	0.0121	Geary autocorrelation - lag 2 / weighted by charges
ATSC2c	0.0083	Centered Broto-Moreau autocorrelation - lag 2 / weighted by charges
AVP-6	0.0081	Average valence path, order 6
AlogP	0.0074	Ghose-Crippen LogKow
SssCH2	0.0066	Sum of atom-type E-State: -CH2-
ALogp2	0.0063	Square of ALogP
ATSC2m	0.0059	Centered Broto-Moreau autocorrelation - lag 2 / weighted by mass
ATSC1e	0.0051	Centered Broto-Moreau autocorrelation - lag 1 / weighted by Sanderson electronegativities
RDF140v	0.0044	Radial distribution function - 140 / weighted by relative van der Waals volumes
TDB3u	0.0041	3D topological distance based autocorrelation - lag 3 / unweighted
RDF140p	0.0041	Radial distribution function - 140 / weighted by relative polarizabilities

Table A2: QSPR descriptors selected for YSI, their importances, descriptions

Descriptor Name	Importance	Description
piPC5	0.2602	Conventional bond order ID number of order 5
piPC4	0.1248	Conventional bond order ID number of order 4
piPC3	0.1099	Conventional bond order ID number of order 3
TpiPC	0.0975	Total conventional bond order (up to order 10)
piPC6	0.0853	Conventional bond order ID number of order 6
SpMax1_Bhm	0.0575	Largest absolute eigenvalue of Burden modified matrix - n 1 / weighted by relative mass
SpMax1_Bhv	0.0291	Largest absolute eigenvalue of Burden modified matrix - n 1 / weighted by relative van der Waals volumes
R_TpiPCTPC	0.0201	Ratio of total conventional bond order (up to order 10) with total path count (up to order 10)
SpMax1_Bhp	0.0195	Largest absolute eigenvalue of Burden modified matrix - n 1 / weighted by relative polarizabilities
ETA_Eta_F	0.0181	Functionality index EtaF
piPC8	0.0133	Conventional bond order ID number of order 8
piPC7	0.0128	Conventional bond order ID number of order 7
SpMin1_Bhi	0.0126	Smallest absolute eigenvalue of Burden modified matrix - n 1 / weighted by relative first ionization potential
MLFER_E	0.0121	Excessive molar refraction
ETA_Beta	0.0100	A measure of electronic features of the molecule

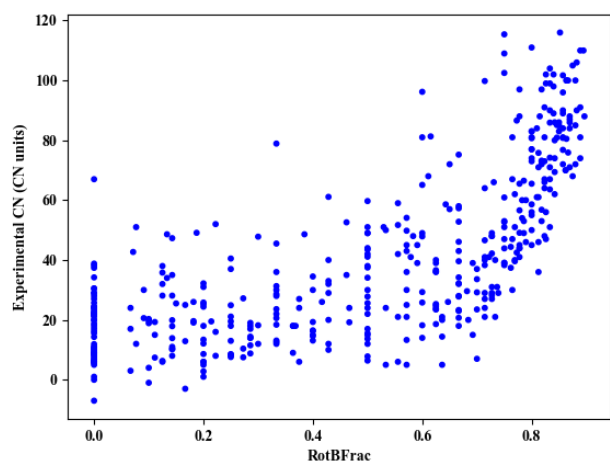


Figure A1: Relationship between CN and RotBFrac

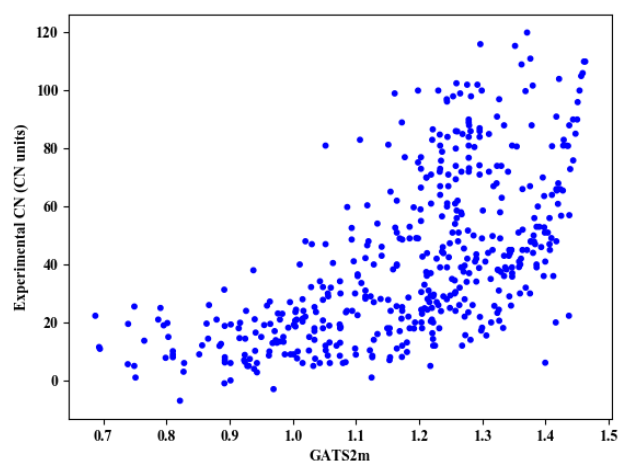


Figure A4: Relationship between CN and GATS2m

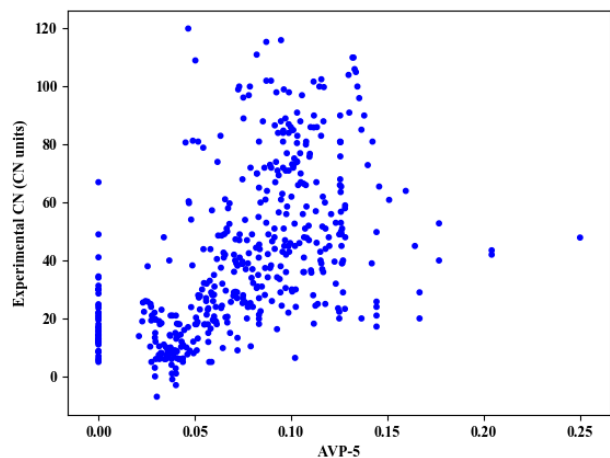


Figure A2: Relationship between CN and AVP-5

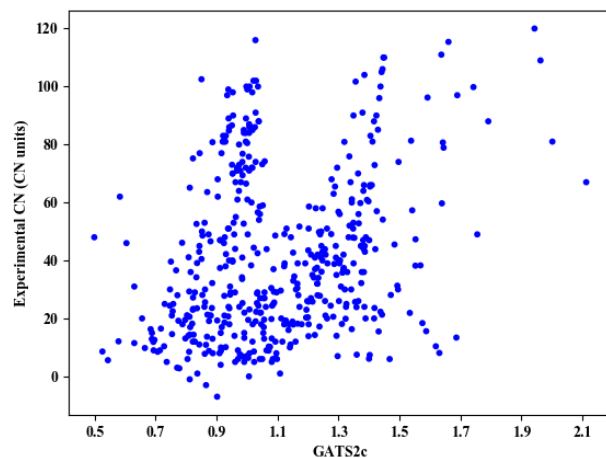


Figure A5: Relationship between CN and GATS2c

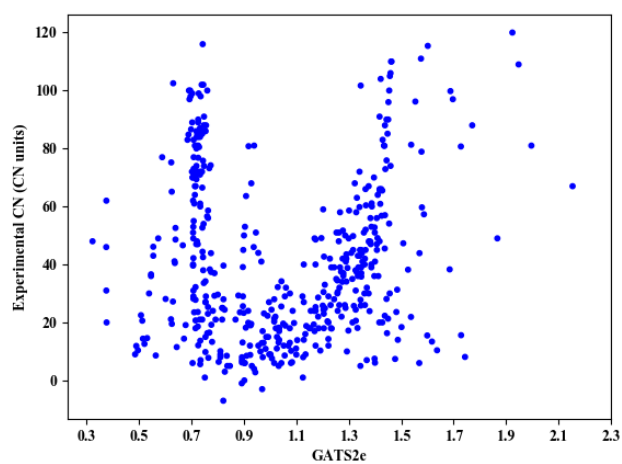


Figure A3: Relationship between CN and GATS2e

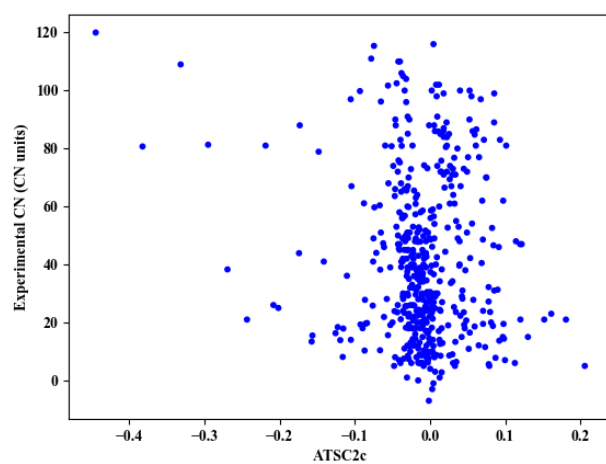


Figure A6: Relationship between CN and ATSC2c

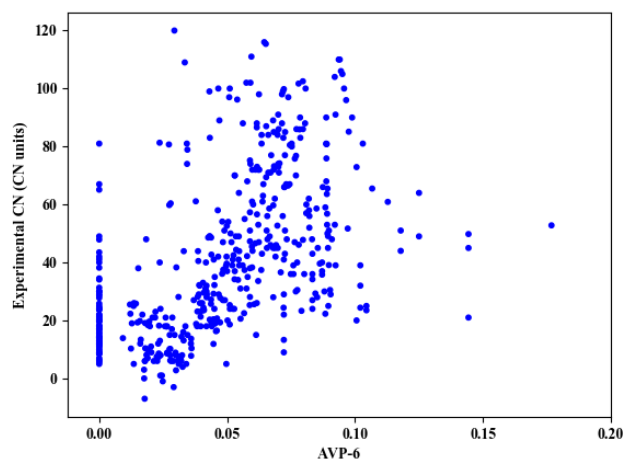


Figure A7: Relationship between CN and AVP-6

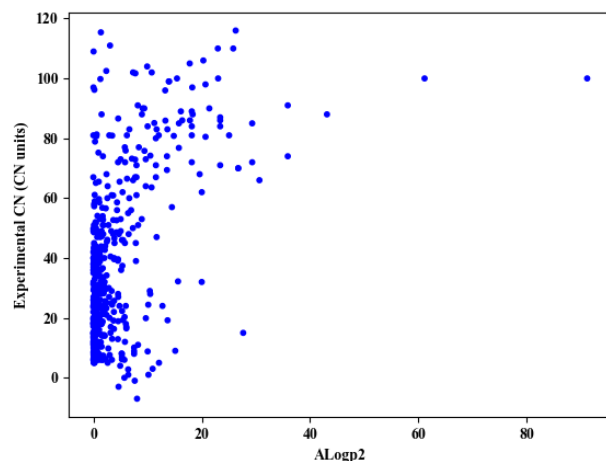


Figure A10: Relationship between CN and ALogp2

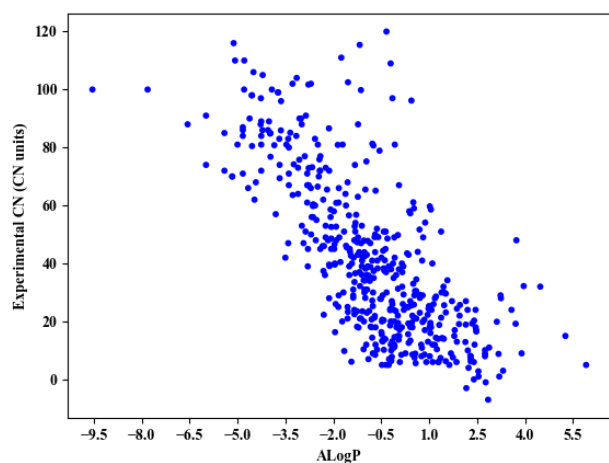


Figure A8: Relationship between CN and AlogP

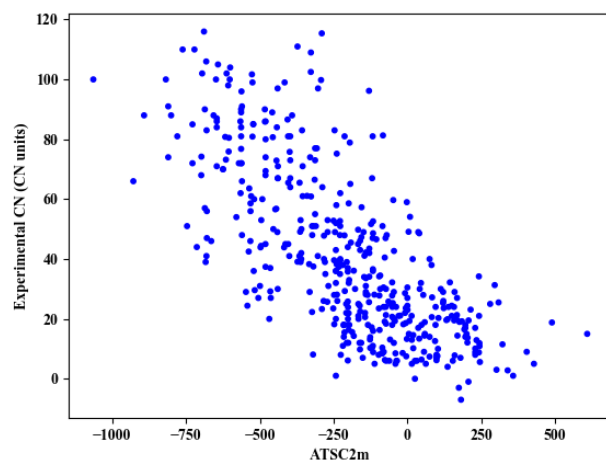


Figure A11: Relationship between CN and ATSC2m

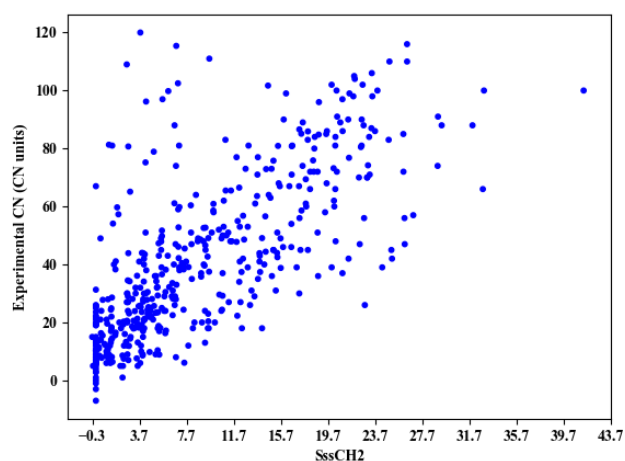


Figure A9: Relationship between CN and SssCH2

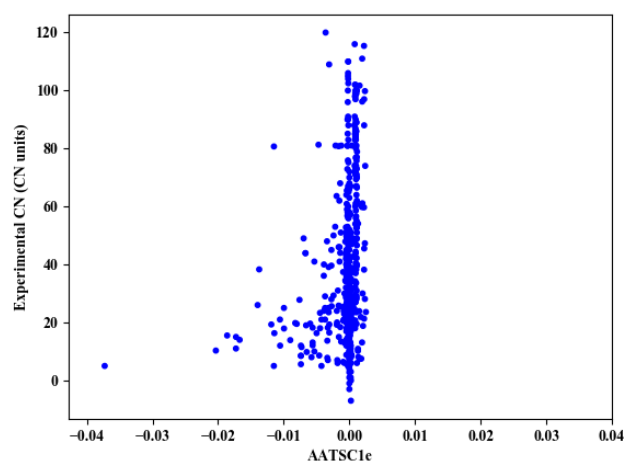


Figure A12: Relationship between CN and ATSC1e

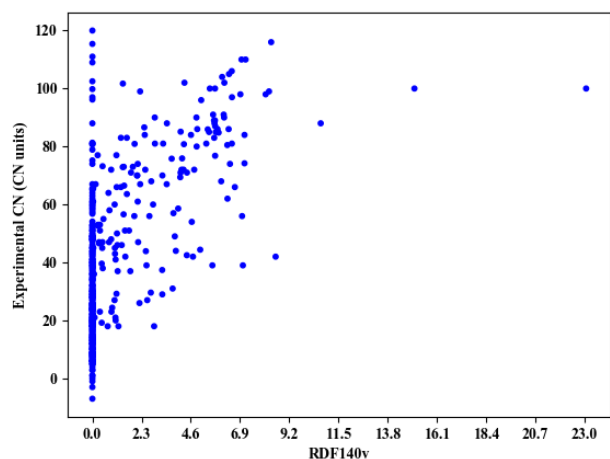


Figure A13: Relationship between CN and RDF140v

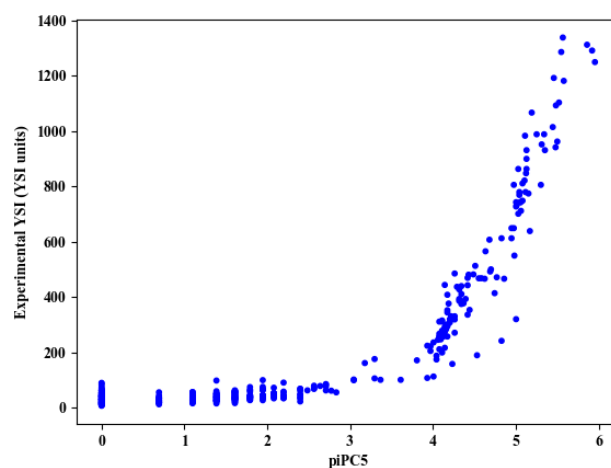


Figure A16: Relationship between YSI and piPC5

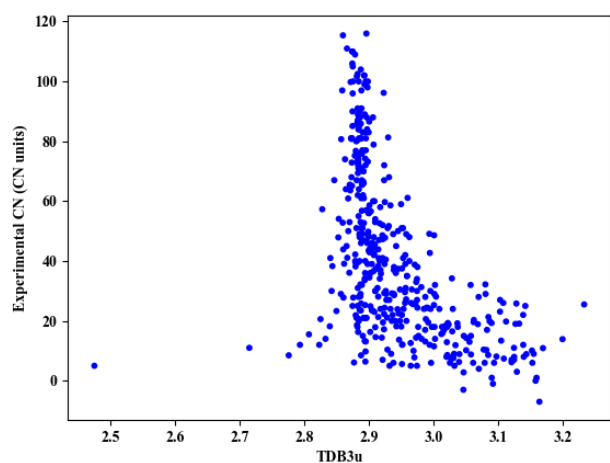


Figure A14: Relationship between CN and TDB3u

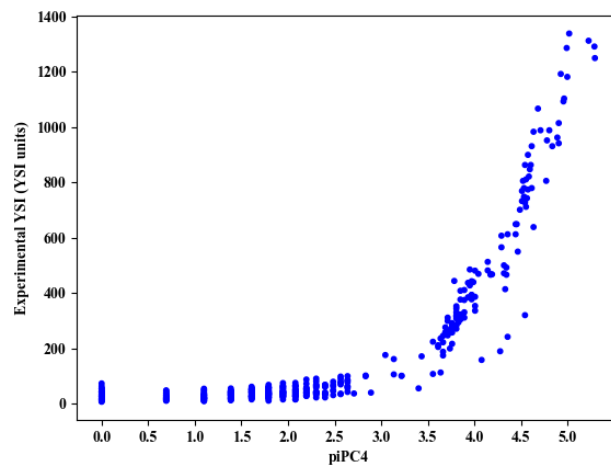


Figure A17: Relationship between YSI and piPC4

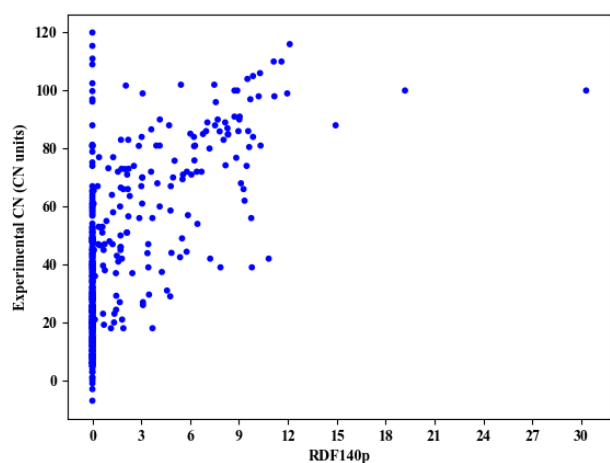


Figure A15: Relationship between CN and RDF140p

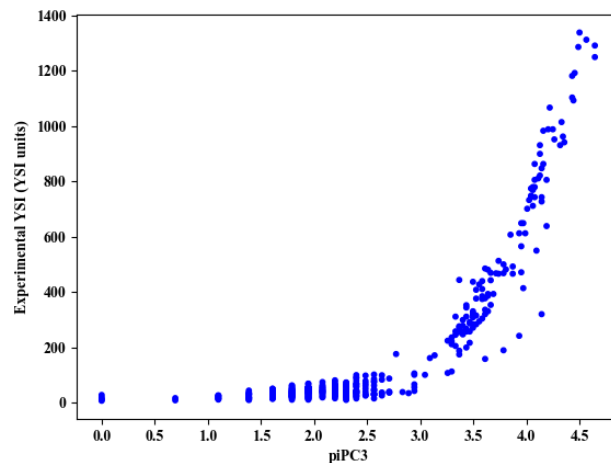


Figure A18: Relationship between YSI and piPC3

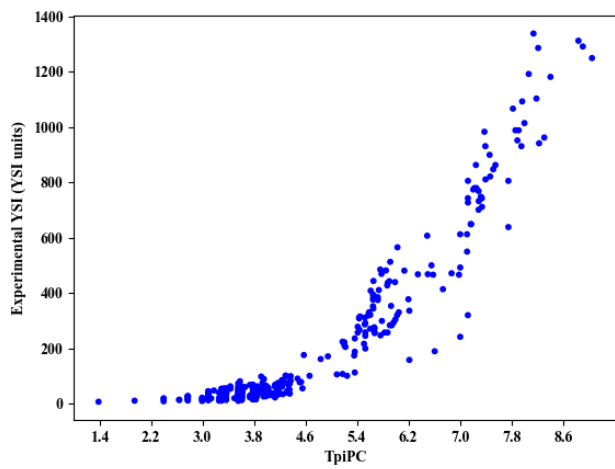


Figure A19: Relationship between YSI and TpiPC

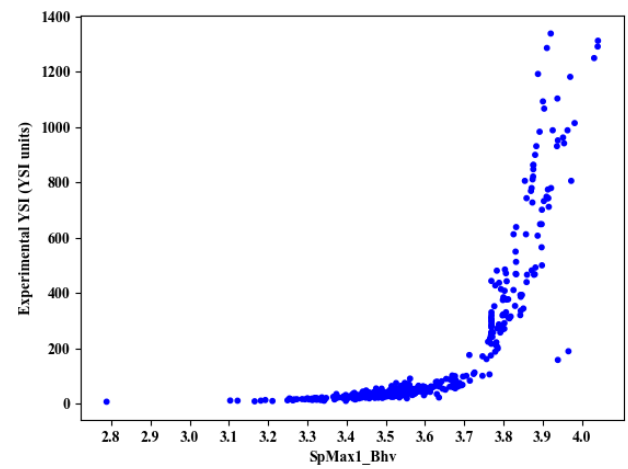


Figure A22: Relationship between YSI and SpMax1_Bhv

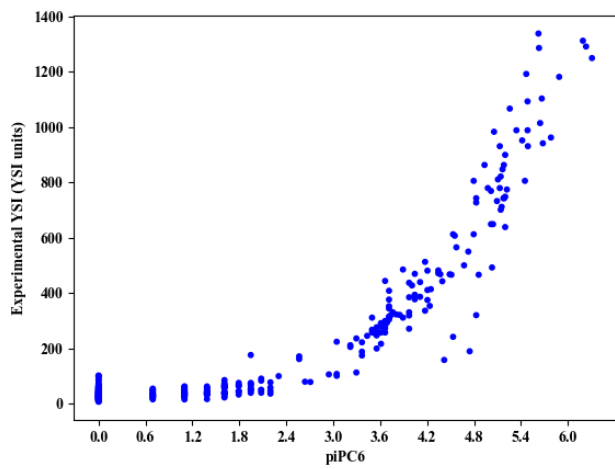


Figure A20: Relationship between YSI and piPC6

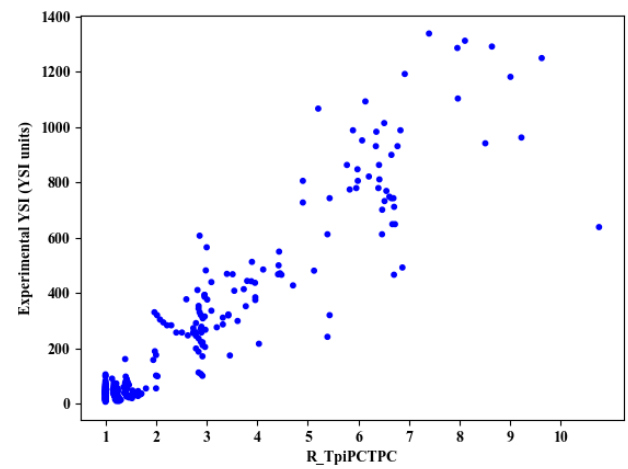


Figure A23: Relationship between YSI and R_TpiPCTPC

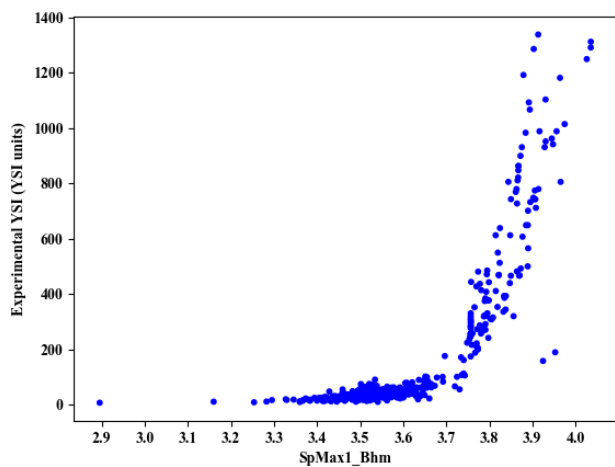


Figure A21: Relationship between YSI and SpMax1_Bhm

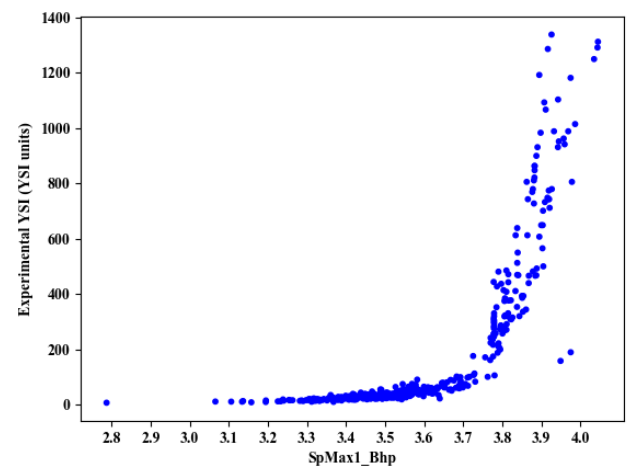


Figure A24: Relationship between YSI and SpMax1_Bhp

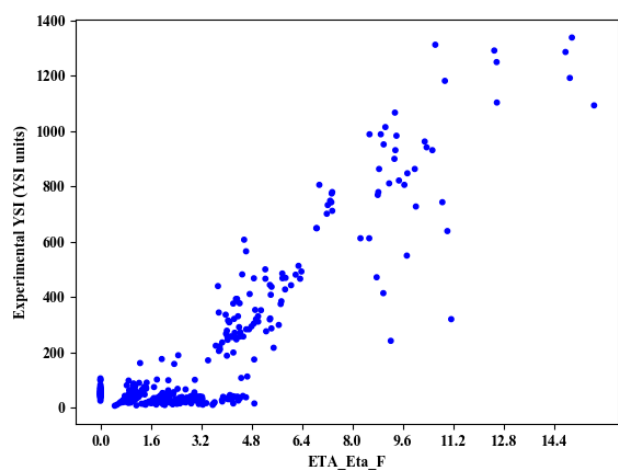


Figure A25: Relationship between YSI and ETA_Eta_F

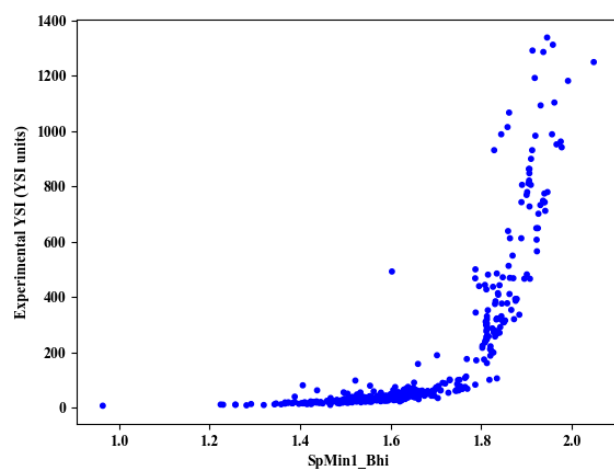


Figure A28: Relationship between YSI and SpMin1_Bhi

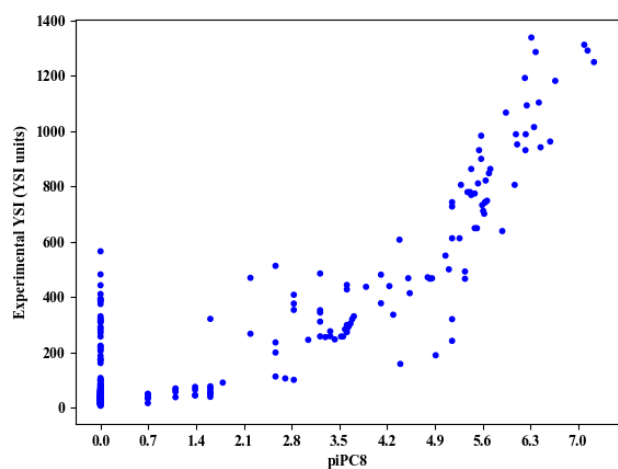


Figure A26: Relationship between YSI and piPC8

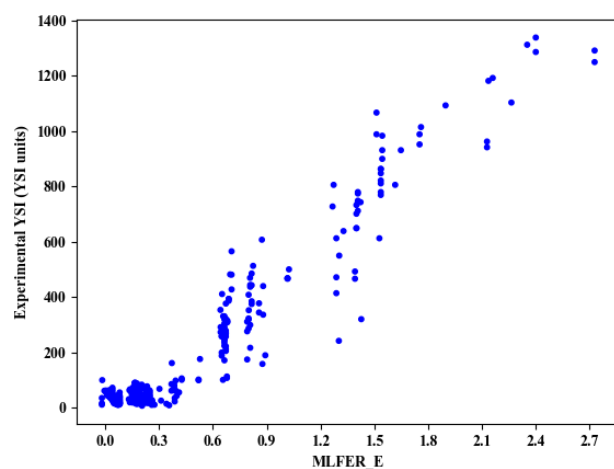


Figure A29: Relationship between YSI and MLFER_E

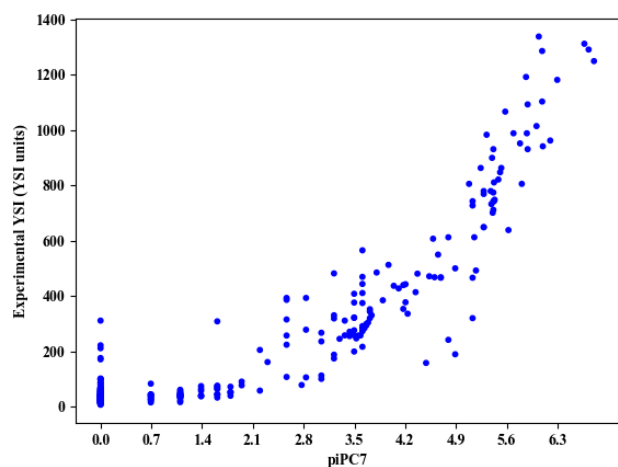


Figure A27: Relationship between YSI and piPC7

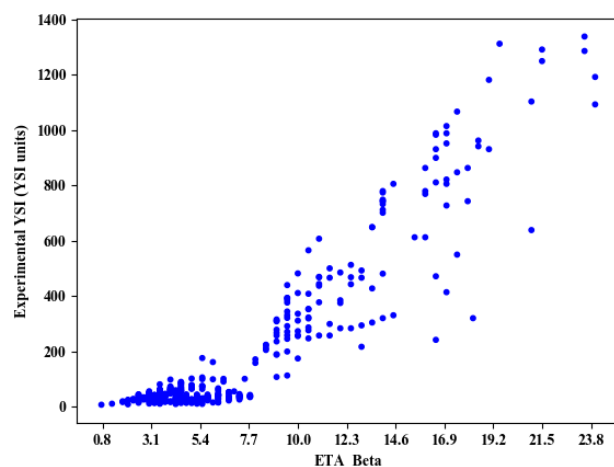


Figure A30: Relationship between YSI and ETA_Beta