Available online at www.sciencedirect.com

ScienceDirect

Proceedings of the Combustion Institute 000 (2020) 1–9

Proceedings of the Combustion Institute

www.elsevier.com/locate/proci

A comparison of computational models for predicting yield sooting index

Travis Kessler [a,*], Peter C. St. John [b], Junqing Zhu [c], Charles S. McEnally [c], Lisa D. Pfefferle [c], J. Hunter Mack [d]

[a] *Department of Electrical and Computer Engineering, University of Massachusetts Lowell, Lowell, MA, United States*
[b] *National Renewable Energy Laboratory, Golden, CO, United States*
[c] *Department of Chemical and Environmental Engineering, Yale University, New Haven, CT, United States*
[d] *Department of Mechanical Engineering, University of Massachusetts Lowell, Lowell, MA, United States*

Received 7 November 2019; accepted 2 July 2020
Available online xxx

Abstract

Sooting propensity, a measurement of how much particulate matter is produced when a fuel is burned, is a property of significant interest among researchers who are striving to discover the next generation of cleaner, more efficient fuels and fuel additives. Many compounds are not viable as fuels and/or fuel additives, and as a result, designing cleaner-burning biofuels using only experimental techniques is inefficient. Predictive models have been instrumental in reducing this inherent difficulty, providing researchers with a tool to pre-emptively screen compounds before production and testing. The present work compares the accuracies and interpretabilities of existing models used to predict a particular measure of sooting propensity, Yield Sooting Index (YSI). These models include artificial neural networks, graph neural networks, and multivariate equations. A novel equation for predicting YSI based on atom path count and bond order is proposed, which can highlight key structural components that contribute to YSI. It was found that artificial neural networks slightly outperform graph neural networks and greatly outperform multivariate equations in blind (test set) prediction accuracy; however, graph neural networks and multivariate equations provide significantly more interpretability as to how compound structure relates to YSI. Predictions of YSI are compared to experimental measurements for previously un-tested compounds with cetane numbers comparable to diesel fuel (50-60) (butyl decanoate, ethyl decanoate, 1,4-bis(ethenoxymethyl)cyclohexane, and 5-heptyloxolan-2-one), and it was found that these compounds produce significantly less soot compared to diesel fuel.

Keywords: Fuel property prediction; Yield sooting index; Artificial neural network; Graph neural network; Multivariate regression

* Corresponding author.
E-mail address: travis_kessler@student.uml.edu (T. Kessler).

https://doi.org/10.1016/j.proci.2020.07.009
1540-7489

1. Introduction

1.1. Yield sooting index

The numerical indices used to represent sooting propensity and the experimental procedures utilized to measure sooting propensity are diverse. The Threshold Sooting Index (TSI) is derived from smoke point measurements by using two reference compounds to create a 0-100 scale [1], and correlates well with engine and turbine soot formation [2]. While the procedure to measure smoke point is relatively simple to carry out and can be performed in a bench-top setting, the experiment can require up to 20 mL of the fuel in question [3]. Consequently, the Yield Sooting Index (YSI) was developed, which uses the maximum soot volume fraction measured in a flame doped with a given fuel as a numerical indication of sooting propensity [4]. Since the flame is typically doped with only 1000 ppm of the test fuel, YSI measurements require only a small sample volume (~0.1 mL) [5]. YSI also enables high sample throughput, so numerous data sets totaling in over 500 measured values are available [6,7]. These values span four orders of magnitude, from very low sooting compounds like methanol (YSI = 3) to very high sooting compounds like 1,2-diphenylbenzene (YSI = 1340). The large number and range of data points in these data sets is valuable for developing and testing predictive models.

1.2. Prediction of chemical properties

Various types of predictive models have historically been successful in predicting numerous chemical properties of a variety of compounds. The cetane number (CN) can be accurately predicted using consensus modeling, comprised of linear and non-linear models, as well as artificial neural networks based on cheminformatic descriptors (ANNs) [8,9]. ANNs are exceptional at forming correlations between multidimensional input and target data, ultimately allowing them to generalize predictions for data similar to the training data that is not observed during training [10]. Moreover, ANNs trained with quantitative structure-property relationship (QSPR) descriptors, which are numerical representations of a variety of physical and chemical traits for a given compound, have proven to be successful in predicting the CN of a variety of molecular classes including bio-mass derived furanic compounds [11]. QSPR-based ANNs have also been successful in predicting YSI, achieving 95% confidence (R^2 correlation between predicted and experimental values) in test set prediction accuracy (predictions for compounds not observed during training) [12].

Message passing neural networks, specifically graph neural networks (GNNs), have shown to out-perform various predictive model architectures when predicting numerous properties represented in the QM9 computational data set (a reference data set for chemical property predictions) as well as properties specific to organic photovoltaic applications [13,14]. Recent work illustrates that the frequency of specific structural components in compounds correlates with YSI, supporting the theory that a structure-based machine learning model such as a GNN can successfully be applied to predicting YSI [6].

Numerous single variable and multivariate equations can estimate the sooting propensity of pure hydrocarbons with varying degrees of success. Structural groups, including six- and four-membered carbon aromatic ring increments (A6 and A4) and the number of branches within a given compound, correlate with the TSI of hydrocarbons [15]. Furthermore, the TSI of hydrocarbons can be estimated using equations derived from the number of carbon atoms, the hydrogen deficiency index, the Balaban connectivity matrix, and the molecular connectivity matrix [16]. While the accuracy of these equations is poor compared to more complex predictive models such as ANNs, the relationships they illustrate between compound structure and sooting propensity provide insight as to how sooting propensity is affected by compound structure.

ANNs are often considered a "black-box" solution to multivariate problems; while very strong correlations can be made between compound structure and sooting propensity, the way ANNs determine these correlations is opaque. It has been shown that middle-to-end learning algorithms based on encoded pattern recognition offer a significant amount of interpretation as to how predictive models learn from their training data [17]. Additionally, graphical training methods (graph convolutional neural networks) have shown to be successful in the design of chemical structures based on knowledge extracted from trained models [18]. A solution as to how a high degree of accuracy for ANNs can be retained while providing a similar level of interpretability as graphical learning models is a topic of considerable interest in the field of predictive modeling and should be explored further.

The present work leverages three methods of predicting the YSI for a variety of compounds: ANNs trained with QSPR descriptors, GNNs, and a newly-proposed multivariate equation. QSPR descriptors are used due to the wide range of chemical and physical representations available for a given compound, allowing an ANN to interpret multivariate QSPR-YSI relationships. Using a GNN reveals the dependence of YSI on compound structure, and provides insight as to which compound substructures contribute to YSI. A multivariate equation is derived using least squares regression from QSPR descriptors that have high correlations with YSI, and the significance of the selected QSPR

Please cite this article as: T. Kessler, P.C. St. John, J. Zhu et al., A comparison of computational models for predicting yield sooting index, Proceedings of the Combustion Institute, https://doi.org/10.1016/j.proci.2020.07.009

descriptors as they relate to compound structure and YSI is discussed. As a stringent test of these predictive models, their predictions of four previously un-tested compounds were compared to new experimental measurements.

2. Experimental procedure

2.1. Experimental data

Experimentally measured values of YSI were taken from [6,7] and supplemented with some additional data for furans [19]. The final database contains 567 distinct compounds. This data was split into three subsets, the training set (70%), validation set (20%), and testing set (10%). Each set was designed to house a proportionally equal number of compounds based on the range of experimental YSI values. Each set remains constant for all model/equation training/derivations to ensure a proper comparison between all predictive models/equations.

Compounds of interest that were not previously experimentally tested were chosen for this investigation as their experimental CNs reside in the 50-60 range (acceptable values as blending agents for diesel fuel) and were identified by the predictive models as having YSIs significantly less than the YSI of diesel fuel, ~235. CN values were obtained from the NREL Compendium of Experimental Cetane Number data and additional literature [20,21]. Their sooting propensities were measured by doping a methane/air nonpremixed flame with 1000 ppm of each compound and using line-of-sight spectral radiance to quantify the level of soot in the flames. Values for YSI were obtained using reference compounds toluene and heptane and known experimental YSI values of 170.9 and 36.0 respectively. The uncertainty of an experimental measurement for a given compound is 5% of the measured value; 3% is due to random uncertainty in the soot measurement, and 2% is due to uncertainty in the mass density of the test compounds (which is required to calculate the liquid phase flowrate of dopant corresponding to 1000 ppm in the gas phase fuel mixture). This experimental procedure is identical to the procedure performed by McEnally et al [22].

2.2. Artificial neural network training

Simple molecular-input line-entry system (SMILES) strings were gathered/generated for all 567 compounds present in the experimental YSI data set. 5305 QSPR descriptors were generated for each compound using alvaDesc (https://www.alvascience.com/alvadesc/). Random forest regression using Scikit-learn was utilized to rank each QSPR descriptor by its correlation with YSI using a derived value of importance; a higher importance value indicates a stronger correlation between a given QSPR descriptor and YSI [23]. The sum of all importance values for a set of variables is equal to 1. Regression and ranking were performed with respect to the training set. To balance accuracy and training time of the ANN, the optimal number of QSPR descriptors used as ANN input variables was determined. Including an excessive amount of QSPR descriptors dramatically increases computational training time for the ANNs while providing relatively insignificant increases in accuracy. The median absolute error (MAE) of an ANN's predictions for the training set was compared to the number of QSPR descriptors added as input variables (from most- to least-important). The relationship between the number of QSPR descriptors used as ANN input variables and prediction MAE was observed to be inversely exponential. The optimal number of QSPR descriptors used as ANN input variables was found to be about 1800. Table 1 lists the ten QSPR descriptors with the highest correlation to YSI out of 5305 total QSPR descriptors.

The ANN architecture consists of an input layer with 1800 neurons (one per input variable), two hidden layers, and an output layer with one neuron (corresponding to the target dimensionality, experimental YSI). Including two hidden layers ensures a non-linear architecture capable of determining relationships between all 1800 selected QSPR descriptors and experimental YSI values. To maximize the accuracy of the ANN, the optimal number of neurons per hidden layer must be determined. ANNs are trained with the Adam optimization function, which accepts user-defined parameters for learning rate and learning rate decay [24]; these parameters must also be optimized to maximize ANN performance. An artificial bee colony (ABC) was employed to optimize the parameters in this multidimensional search space. An ABC excels at optimizing multivariate problems, and has been shown to increase the predictive accuracy of ANNs by 20.4% by tuning ANN hyper-parameters [25]. Solutions (values for each parameter) were evaluated by the ABC by constructing an ANN and measuring the MAE of the validation set after 500 epochs. The solution with the lowest MAE was deemed to have the best parameter values for the ANN's architecture and learning parameters. By employing 25 bees for 25 search cycles, it was found that the optimal number of neurons for the first and second hidden layer are about 1900 and 1300 respectively, limiting the search space of the number of neurons per layer to [1, 3600]. The optimal learning rate was found to be about 1×10^{-3}, and the optimal learning rate decay was found to be about 1×10^{-6}, both limited to the search space of [0.0, 1.0].

ANN training was performed using ECNet, an open-source Python package tailored to predicting fuel properties [26]. ANNs were trained using a

Please cite this article as: T. Kessler, P.C. St. John, J. Zhu et al., A comparison of computational models for predicting yield sooting index, Proceedings of the Combustion Institute, https://doi.org/10.1016/j.proci.2020.07.009

Table 1
QSPR descriptors with highest correlation to YSI.

Descriptor	Importance	Definition
SM6_B(p)	0.173912	spectral moment of order 6 from Burden matrix weighted by polarizability
SpMax1_Bh(p)	0.088155	largest eigenvalue n. 1 of Burden matrix weighted by polarizability
piID	0.086094	conventional bond order ID number
piPC05	0.078171	molecular multiple path count of order 5
nCar	0.065015	number of aromatic C(sp2)
nCsp2	0.055157	number of sp^2 hybridized carbon atoms
piPC04	0.049709	molecular multiple path count of order 4
SM6_B(v)	0.042868	spectral moment of order 6 from Burden matrix weighted by van der Waals volume
SpMax1_Bh(v)	0.028548	largest eigenvalue n. 1 of Burden matrix weighted by van der Waals volume
piPC03	0.027874	molecular multiple path count of order 3

backpropagation algorithm to regress on the training set, and the validation set measured the performance of the ANN during training. Training was terminated once performance on the validation set ceased to improve. The ANN's ability to generalize predictions for data not observed during training is measured by the MAE of predictions for the test set. The ANN training procedure was carried out ten times to ensure consistency in predictive accuracy.

2.3. Graph neural network training

SMILES strings were gathered/generated for all 567 compounds present in the same experimental YSI data set. Key information about compound structure was determined from each SMILES string, including each compound's connectivity matrix, node (atom) classifications, and discrete edge (bond) classifications. Atom classifications are comprised of vectors containing information such as their atomic symbol, degree of bonding, and if they exist within a ring. Bond classifications are comprised of discrete vectors for bond order (single, double, triple, or aromatic). Connectivity, atom classifications, and bond classifications derived from SMILES strings are used as input vectors for GNN training.

The message passing algorithm was implemented using a method similar to the connectivity algorithm presented by Jørgensen et al., where feature vectors for atoms and bonds are passed to neighboring atoms and the atom and bond states are updating accordingly [27]. The compound/graph is defined as the sum of all atom states, and a predicted YSI is obtained by passing the compound vector through a series of densely-connected layers that are optimized using a backpropagation algorithm. These operations were implemented using the Keras/TensorFlow Python packages and are available on GitHub (http://github.com/nrel/nfp). GNNs were fit using the training set, and training was halted when validation set performance ceased improving. The GNN's ability to generalize predictions for data not observed during training is measured by the MAE of predictions for the test set. The GNN training procedure was carried out ten times to ensure consistency in predictive accuracy.

In addition to determining a value of YSI from the GNN, an additional readout step was performed with respect to edge/bond states instead of individual node/atom states similar to the procedure performed by Schütt et al. [28]. This provides metrics of contribution from each bond in a given compound as it relates to YSI, providing insight as to which structural components of a given compound contribute to a higher YSI value.

2.4. Path length equation derivation

Three of the ten QSPR descriptors present in Table 1 are defined as the molecular multiple path count *piPC* of varying orders *N*. While descriptors such as $SM6\_B(p)$ and $SpMax1\_Bh(p)$ rank higher with regards to correlation to YSI, their relation to specific components of a compound's structure is somewhat arbitrary. Alternatively, *piPC* descriptors are directly related to the structure/connectivity of a given compound, measuring the frequency of paths of length $N+1$ atoms that appear in a compound, weighted by bond order. Eq. (1) depicts the calculation for *piPC* of order *N*:

$$piPC(N) = ln\left(1 + \frac{\sum_{i=1}^{A}\sum_{j=1}^{P_{N,i}}\prod_{k=1}^{\alpha_{i,j}} B_{k,k+1}}{2}\right) \quad (1)$$

where A is the number of atoms present in the compound, $P_{N,i}$ is the number of paths of length $N+1$ starting from atom i, $\alpha_{i,j}$ is the number of atoms within the path j starting from atom i, and $B_{k,k+1}$ is a numerical representation of bond order between atom k and $k+1$. The numerical representations of bond order are 1 for a single bond, 2 for a double bond, 3 for a triple bond, and 1.5 for an aromatic bond. This equation was derived from the algorithm used to calculate *piPC* descriptors implemented in PaDEL-Descriptor [29]. Figure 1 illustrates the relationship between YSI and *piPC*05 for compounds in the training and validation sets, as well as an equation derived from the coefficient

Please cite this article as: T. Kessler, P.C. St. John, J. Zhu et al., A comparison of computational models for predicting yield sooting index, Proceedings of the Combustion Institute, https://doi.org/10.1016/j.proci.2020.07.009

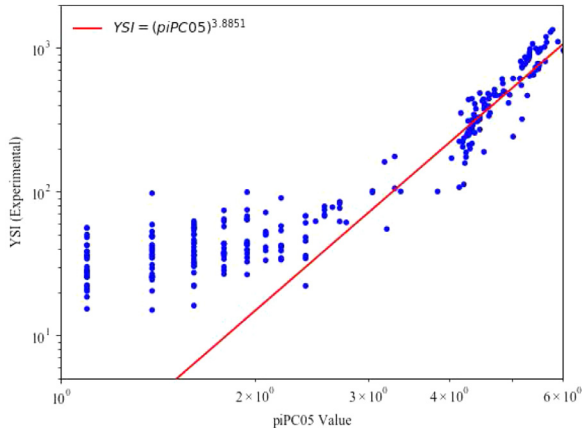

Fig. 1. Experimental YSI versus *piPC05* value for compounds in training and validation sets.

found for *piPC*05 presented the following section of this manuscript. An exponential relationship between YSI and *piPC*05 is observed, notably for compounds with a higher experimental YSI value, indicating regression can be used to derive coefficient(s) of an exponential equation relating YSI to *piPC*05. The same behavior occurs for other *piPC* descriptors at varying orders *N*.

A multivariate exponential equation was derived by performing the Levenberg-Marquardt method of least-squares regression (LSR) on *piPC*05, *piPC*04, and *piPC*03 [30]. Fitting an equation to these QSPR descriptors requires Eq. (2)'s coefficients *c* be optimized to minimize the error of calculated YSI values compared to experimental YSI values:

$$\mathrm{YSI}_{\mathrm{calc}} = \sum_{N=3}^{5} (piPC(N))^{c_N} \tag{2}$$

The optimal coefficients *c* are derived from performing LSR on the training and validation sets. LSR was performed using Scipy's *curve_fit* function [31]. The validation set is included to provide the LSR algorithm with additional samples not relevant to blind (test set) prediction accuracy measurements and comparisons. The ability of the derived equation to generalize predictions for data not observed during regression is determined by measuring the MAE of test set calculations compared to test set experimental values.

3. Results and discussion

Figures 2 and 3 show parity plots for predicted YSI values compared to experimental YSI values for an ANN and GNN resulting from the experimental procedure. While the ANN's test set MAE is lower than the GNN's test set MAE, the GNN's predictions for the training and validation sets are significantly more accurate than the ANN's. The optimal coefficients for Eq. (2), c_5, c_4, and c_3, were found to be 3.8851, 3.1567e-06, and 1.4326 respectively. It is worth noting that the coefficient c_4 is significantly lower than c_5 and c_3, indicating the contribution of *piPC*04 in this equation is minimal when *piPC*05 and *piPC*03 are utilized for LSR. Figure 4 shows a parity plot for equation-derived predictions compared to experimental YSI values. It is apparent, both from respective set MAE values and from visualizing prediction versus experimental parity, that the derived equation is not nearly as accurate as the ANN or GNN models. The standard deviations in YSI estimation for each coefficient, derived from the top-left to bottom-right values of the covariance matrix resulting from LSR, were found to be 0.0201, 9.4512, and 2.6802. These values inversely correlate with the apparent contribution of each of the coefficients to a predicted value of YSI. Further analysis of the relationship between path length and YSI should not consider *piPC04*, and instead opt to use higher or lower rank measurements of path length.

Table 2 illustrates the average MAE values for YSI predictions for the training, validation, and testing sets for ten ANNs, ten GNNs, and the coefficients resulting from optimizing Eq. (2). Overall, the MAE for test set predictions resulting from ANNs is 0.48 YSI units lower than the MAE of test set predictions resulting from GNNs. Training and validation MAE values are significantly lower for GNNs compared to ANNs. MAE values resulting from predicting YSI with the derived equation, while consistent, are significantly higher than both the average ANN and GNN MAE values. Table 3 shows the measured YSI values of the four previously un-tested compounds, as well as the average YSI prediction for each compound across the ten constructed ANNs, ten constructed GNNs, and the coefficients used to optimize Eq. (2). Experimental error is given as ±5%,

Please cite this article as: T. Kessler, P.C. St. John, J. Zhu et al., A comparison of computational models for predicting yield sooting index, Proceedings of the Combustion Institute, https://doi.org/10.1016/j.proci.2020.07.009

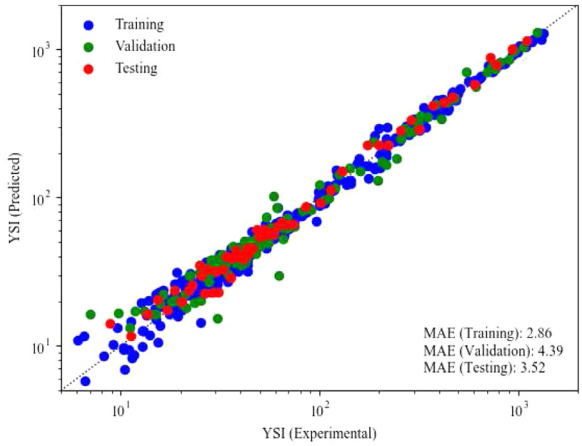

Fig. 2. Comparison of predicted YSI values resulting from ANN predictions and experimental YSI values.

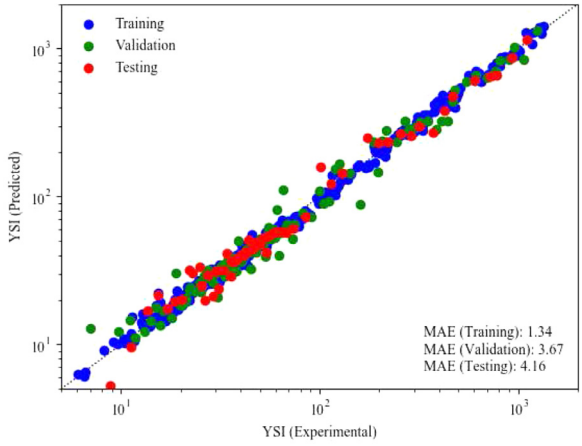

Fig. 3. Comparison of predicted YSI values resulting from GNN predictions and experimental YSI values.

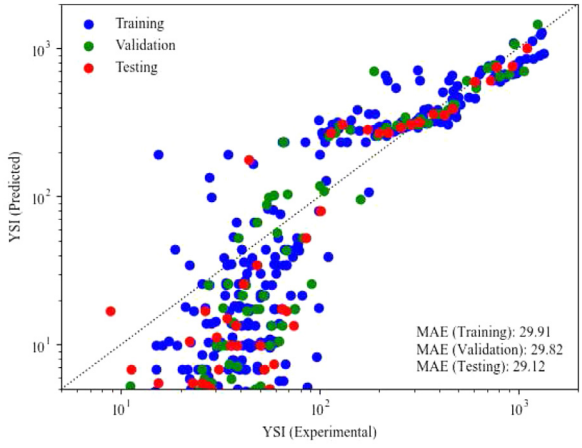

Fig. 4. Comparison of predicted YSI values resulting from the optimized Eq. (2) predictions and experimental YSI values

Please cite this article as: T. Kessler, P.C. St. John, J. Zhu et al., A comparison of computational models for predicting yield sooting index, Proceedings of the Combustion Institute, https://doi.org/10.1016/j.proci.2020.07.009

Table 2
MAEs and standard deviations for training, validation and test sets across ten ANNs, ten GNNs, and Eq. (2) with optimized coefficients.

Method	Train MAE	Train Stdev.	Valid. MAE	Valid. Stdev.	Test MAE	Test Stdev.
ANN	3.38	0.57	4.50	0.50	4.34	0.62
GNN	1.65	0.47	3.69	0.32	4.82	0.76
Equation	29.91	N/A	29.82	N/A	29.12	N/A

Table 3
Experimental YSI measurements and predictions for previously un-tested compounds.

Compound Name (IUPAC)	CN	Measured YSI (±5%)	Ave. ANN Pred. (±4.34)	Ave. GNN Pred. (±4.81)	Eqn. Pred. (±29.12)
butyl decanoate	55	82.6	75.0	73.3	53.4
ethyl decanoate	51	58.0	57.1	56.8	35.1
1,4-bis(ethenoxymethyl) cyclohexane	61.1	130.5	135.5	103.0	117.8
5-heptyloxolan-2-one	52.6	58.2	52.7	60.2	76.7

and error in predictions for the ANN, GNN, and derived equation are defined as the model/equation test set MAE values. If up to 5% error is assumed for all experimental measurements, the ANN model is capable of predicting the YSI for all listed compounds within its average test set MAE. Alternatively, the GNN is unable to predict butyl decanoate and 1,4-bis(ethenoxymethyl)cyclohexane within its average test set MAE given up to 5% experimental error. Further, the prediction for 1,4-bis(ethenoxymethyl)cyclohexane resulting from the derived equation is more accurate than the prediction resulting from the GNN. Predictions from the derived equation are relatively inaccurate compared to the ANN and GNN, but are within the equation's test set MAE given 5% experimental error.

The CN and YSI of a fuel mixture is the concentration-weighted linear combination of the CNs and YSIs of the individual components [2,32]. The compounds present in Table 3 have CN values similar to traditional diesel fuel, however their YSI values are significantly less than that of diesel fuel ($\sim$235). While using a 100% concentration of these compounds is impractical both economically and logistically, employing these compounds as blending agents up to $\sim$30% has the potential to decrease soot formation during combustion without causing any degradation in cetane number. Further experiments to analyze the behavior of these compounds as additives should be performed, and the viability of producing these compounds at scale should be investigated from both a technoeconomic and life-cycle perspective.

The exponential relationship between YSI and *piPC* descriptors illustrated in Fig. 1 offers significant insight as to which types of compounds yield higher values of YSI. The derivation of *piPC* values represented by Eq. (1) shows that (1) a large compound with significantly more paths of length $N+1$ present, and (2) a compound with a higher number of high-order bonds both result in a higher value for *piPC* descriptors, and consequentially higher YSI values for the compounds. A similar observation has been noted with respect to soot formation in blends, where blends containing long carbon chains tend to produce more soot during combustion [33].

Figure 5(a–d) show the weighted bond contributions obtained from the GNN procedure for cyclohexene, 1,4-cyclohexadiene, propylbenzene, and 1,1-diphenylethylene. Their respective experimental YSI values are 45.6, 101.4, 235.7, and 743.1. Figure 5(d) shows that the bonds present at the junction between two rings contribute to a higher YSI. When *piPC* values of varying order are calculated for this compound, the frequency of atom paths along this junction is significantly higher compared to atoms further from the junction. The value of *piPC*05 for 1,1-diphenylethylene is 5.1922. A similar behavior is observed in Fig. 5(c), where the junction between the compound's ring and chain has higher bond contributions and is expected to house more paths during *piPC* calculation. The value of *piPC*05 for propylbenzene is 4.2185. GNN bond contributions also correlate with bond order, as Fig. 5(b) has higher bond contributions than in Fig. 5(a) due to the addition of a second double bond. This observation also correlates with *piPC* calculations, where higher order bonds increase the weighting of atoms paths during calculation. The values of *piPC*05 for 1,4-cyclohexadiene and cyclohexene are 3.2958 and 2.4849 respectively. The visualization of bond contribution in Fig. 5 and their comparison to *piPC*05 values highlights the correlation between GNN bond contribution, *piPC*, and experimental YSI.

Please cite this article as: T. Kessler, P.C. St. John, J. Zhu et al., A comparison of computational models for predicting yield sooting index, Proceedings of the Combustion Institute, https://doi.org/10.1016/j.proci.2020.07.009

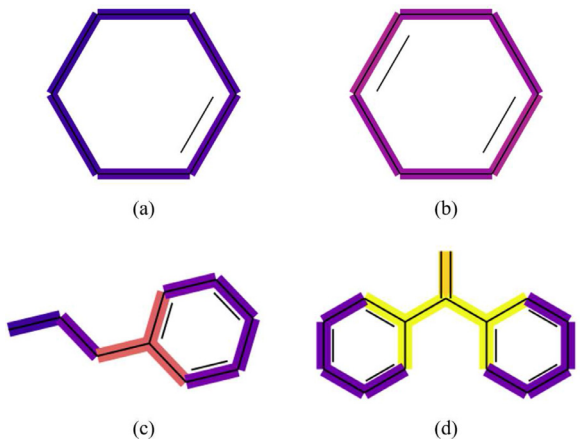

Fig. 5. Contributions of bonds to YSI for **(a)** cyclohexene, **(b)** 1,4-cyclohexadiene, **(c)** propylbenzene, and **(d)** 1,1-diphenylethylene, derived from GNN model. Bond contribution increases as the color hues shift from blue, through red, into yellow.

4. Conclusions and recommendations

The present work illustrates the disparity between accuracy and interpretability for three YSI predictive model designs, ANNs, GNNs, and a novel multivariate equation. ANNs are shown to be highly accurate at generalizing YSI predictions for data not observed during training, however no significant interpretations of QSPR descriptors as they relate to YSI can be drawn from the trained ANN. While this highlights the ANN's ability to form relationships between multivariate input/target data, its lack of interpretability leaves much to be desired. GNNs and the derived equation offer insight as to which structural components of a compound contribute to YSI, although neither is as accurate as an ANN in generalizing predictions for data not observed during training.

The compounds investigated in this study, butyl decanoate, ethyl decanoate, 1,4-bis(ethenoxymethyl)cyclohexane, and 5-heptyloxolan-2-one, each offer significantly lower sooting propensities compared to traditional diesel fuel while retaining optimal values of CN.

It has been shown that the number of paths of certain lengths and the number of higher order bonds, represented by piPC descriptors, show a strong correlation to YSI. Additionally, there is a correlation between piPC and how a GNN interprets bond contributions as they relate to YSI.

While there is a trade-off between accuracy and interpretability between the modeling approaches presented here, their application to fuel property prediction is beneficial from a screening standpoint and can also provide additional insight into structural influences.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This material is based upon work supported by the U.S. Department of Energy (DOE) Office of Energy Efficiency and Renewable Energy (EERE) Bioenergy Technologies Office (BETO) under award DE-EE0008479 as part of the Co-Optimization of Fuels & Engines (Co-Optima) project, and by the U.S. Department of Energy (DOE) Office of Energy Efficiency and Renewable Energy (EERE) Bioenergy Technologies Office (BETO) and Vehicle Technologies Office (VTO) Program under award DE-EE0007983.

References

[1] H. Calcote, D. Manos, *Combust. Flame* 49 (1983) 289–304.
[2] E.J. Barrientos, J.E. Anderson, M.M. Maricq, A.L. Boehman, *Combust. Flame* 167 (2016) 308–319.
[3] ASTM International, D1322-18: ASTM international, 2018,
[4] C.S. McEnally, L.D. Pfefferle, *Combust. Flame* 148 (2007) 210–222.
[5] C.S. McEnally, L.D. Pferfferle, *Proc. Combust. Inst.* 32 (2009) 673–679.
[6] D.D. Das, P.C.S. John, C.S. McEnally, S. Kim, L.D. Pfefferle, *Combust. Flame* 190 (2018) 349–364.

Please cite this article as: T. Kessler, P.C. St. John, J. Zhu et al., A comparison of computational models for predicting yield sooting index, Proceedings of the Combustion Institute, https://doi.org/10.1016/j.proci.2020.07.009

[7] C.S. McEnally, D.D. Das, L.D. Pfefferle, 2017, https://doi.org/10.7910/DVN/7HGFT8 Harvard Dataverse, V1.
[8] E.A. Smolenskii, V.M. Bavykin, A.N. Ryzhov, O.L. Slovokhotova, I.V. Chuvaeva, A.L. Lapidus, *Russ. Chem. Bull.* 57 (3) (2008) 461–467.
[9] H. Yang, C. Fairbridge, Z. Ring, *Petrol. Sci. Technol.* 19 (5–6) (2001) 573–586.
[10] S. Baluja, D. Pomerleau, *Adv. Neural Inf. Process. Syst.* (1994) 753–760.
[11] T. Kessler, E.R. Sacia, A.T. Bell, J.H. Mack, *Fuel* 206 (2017) 171–179.
[12] P.C.S. John, P.M. Kairys, D.D. Das, C.S. McEnally, L.D. Pfefferle, D.J. Robichaud, M.R. Nimlos, *Energy Fuels* (2017).
[13] P.C.S. John, C. Phillips, T.W. Kemper, A.N. Wilson, Y. Guan, M.F. Crowley, M.R. Nimlos, R.E. Larsen, *J. Chem. Phys.* 150 (23) (2019) 234111.
[14] J. Gilmer, S.S. Schoenholz, P.F. Riley, O. Vinyals, G.E. Dahl, in: Proceedings of the 34th International Conference on Machine Learning, 70, 2017, pp. 1263–1272.
[15] S. Yan, E.G. Eddings, A.B. Palotas, R.J. Pugmire, A.F. Sarofim, *Energy Fuels* 19 (6) (2005) 2408–2415.
[16] M.P. Hanson, D.H. Rouvray, *J. Phys. Chem.* 91 (11) (1987) 2981–2985.
[17] Q. Zhang, S. Zhu, *Front. Inf. Technol. Electron. Eng.* 19 (1) (2018) 27–39.
[18] T. Xie, J.C. Grossman, *Phys. Rev. Lett.* 120 (14) (2018) 145301.
[19] J. Zhu, B. Hu, B.D. Etz, H. Kwon, C.S. McEnally, Y. Xuan, 2019, https://doi.org/10.6084/m9.figshare.7929899.v1.
[20] J. Yanowitz, M.A. Ratcliff, R.L. McCormick, J.D. Taylor, M.J. Murphy, 2014, NREL/TP-5400-61693.
[21] M. Dahmen, W. Marquardt, *Energy Fuels* 29 (9) (2015) 5781–5801.
[22] C.S. McEnally, Y. Xuan, P.C.S. John, D.D. Das, A. Jain, S. Kim, T.A. Kwan, L.K. Tan, J. Zhu, L.D. Pfefferle, *Proc. Combust. Inst.* 37 (1) (2019) 961–968.
[23] L. Breiman, *J. Mach. Learn.* 45 (2001) 5–32.
[24] D.P. Kingma, J. Ba, in: International Conference for Learning Representations, 2015.
[25] S. Sharma, H. Gelaf-Romer, T. Kessler, J.H. Mack, *J. Open Source Softw.* 4 (2019).
[26] T. Kessler, J.H. Mack, *J. Open Source Softw.* 2 (17) (2017) 401.
[27] P.B. Jørgensen, K.W. Jacobsen, M.N. Schmidt.
[28] K. Schütt, P.J. Kindermans, H.E.S. Felix, S. Chmiela, A. Tkatchenko, K.R. Müller, *Adv. Neural Inf. Process. Syst.* (2017) 991–1001.
[29] C. Yap, *J. Comput. Chem.* 32 (7) (2011) 1466–1474.
[30] H.P. Gavin, 2013, Dept. Civil Environ. Eng. Duke Univ., Durham, NC, USA.
[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, *J. Mach. Learn. Res.* 12 (2011) 2825–2830.
[32] A.C. Hansen, Q. Zhang, P.W.L. Lyne, *Bioresour. Technol.* 96 (3) (2005) 277–285.
[33] M.M. Rahman, S. Stevanovic, M.A. Islam, K. Heimann, M.N. Nabi, G. Thomas, B. Feng, R.J. Brown, Z.D. Ristovski, *Environ. Sci.* 17 (2015) 1601–1610.

Please cite this article as: T. Kessler, P.C. St. John, J. Zhu et al., A comparison of computational models for predicting yield sooting index, Proceedings of the Combustion Institute, https://doi.org/10.1016/j.proci.2020.07.009