Prediction of Research/Motor Octane Number and Octane Sensitivity Using Artificial Neural Networks

Travis J. Kessler[1,*], **Corey Hudson**[2], **Leanne Whitmore**[3], and **J. Hunter Mack**[4]

[1]Department of Electrical & Computer Engineering, University of Massachusetts Lowell, Lowell, MA
[2]Sandia National Laboratories, Livermore, CA
[3]Department of Immunology, University of Washington, Seattle, WA
[4]Department of Mechanical Engineering, University of Massachusetts Lowell, Lowell, MA
[*]Corresponding author: travis_kessler@student.uml.edu

Abstract

Octane sensitivity (OS), defined as the research octane number (RON) minus the motor octane number (MON) of a given fuel, has gained interest among researchers due to its apparent effect on knocking conditions in internal combustion engines. Compounds with a high OS enable higher efficiencies, especially with respect to advanced compression ignition engines. RON/MON must be experimentally tested to determine OS; however, the experimental methods utilized require a substantial amount of time, a significant monetary investment, and specialized equipment. To this end, predictive computational models trained with existing experimental data and molecular properties would allow for the preemptive screening of compounds prior to performing these experiments. The present work proposes two methods for predicting the OS of a given compound: using artificial neural networks (ANNs) trained with quantitative structure-property relationship (QSPR) descriptors to predict RON and MON individually to compute OS from RON/MON predictions (derived octane sensitivity, dOS), and using an ANN trained with QSPR descriptors to directly predict OS. ANNs trained to predict RON and MON achieved test set root-mean-square errors (RMSEs) of 10.499 and 7.551 respectively. dOS calculations were found to have a test set RMSE of 6.432 while predicting OS directly resulted in a test set RMSE of 7.019, showing it is more beneficial to obtain OS from RON/MON predictions than predicting it directly. Furthermore, relationships between individual QSPR descriptors and RON/MON/OS are discussed, highlighting correlations between specific molecular features and these properties.

Keywords: Artificial neural network, research octane number, motor octane number, octane sensitivity, quantitative structure-property relationship

1 Introduction

1.1 Octane Number, Octane Sensitivity

Octane number is representative of the ignition quality of a gasoline-like fuel in a spark ignition engine; specifically, it is a measurement of the fuel's ability to resist knock (occurrence of autoignition before spark ignition due to high pressure) [1]. Two distinct octane number indices are used to quantify knock, research octane number (RON) and motor octane number (MON). RON is typically utilized to represent real-world engine conditions while MON is typically utilized to represent high-performance engine conditions, and their experimental procedures reflect this with respect to Cooperative Fuel Research engine inlet temperature and revolutions per minute [2] [3]. Each procedure requires two reference fuels to measure the RON/MON of a given compound, the most common being n-heptane and isooctane (2,2,4-trimethylpentane) which form a 0-100 scale. Tetraethyllead may be introduced to isooctane to increase the range of the scale to 0-120. A higher value of RON/MON indicates a higher resistance to knock. These experimental procedures require large quantities of each reference fuel and the compound to be measured (approximately 500 mL), as well as a considerable amount of time to carry out the procedures [4].

The octane sensitivity (OS) of a compound is defined as the difference between its RON and MON ($RON - MON$), and is used to measure the difference in performance of the compound at varying engine conditions. Studies have shown that the OS of a compound affects its ability to resist knock at varying pressures and temperatures [5] [6], specifically that a lower sensitivity at low load/pressure and a higher sensitivity at high load/pressure both yield a higher resistance to knock [7]. Knock is generally seen as a barrier for researchers who work towards increasing the efficiency of spark-ignition

engines, and a comprehensive understanding of autoignition and knock has yet to be achieved. It is therefore paramount that further studies of knock, including the nuanced role of OS, are performed to produce a better understanding of knock and ultimately raise the efficiency of spark-ignition engines.

1.2 Fuel Property Prediction

A diverse set of methods exists for predicting numerous combustion-related properties of hydrocarbons and oxygenated compounds. Consensus modeling, comprised of non-linear and linear models, has been utilized to predict the cetane number (CN) of alkanes and cycloalkanes [8]. Additional studies have shown that artificial neural networks (ANNs) trained using experimental data and cheminformatic descriptors can accurately predict the CN of isoparaffins and diesel fuels [9]. Furthermore, ANNs trained with experimental data and quantitative structure-property relationship (QSPR) descriptors, numerical variables describing various physical, chemical, and electromechanical aspects of a given compound, have been shown to accurately predict the CN of a variety of molecular classes including furanic compounds derived from renewable resources [10]. ANNs have also proven to be successful in predicting the RON/MON of gasolines and gasoline blends based on chromatographic analysis and volumetric content respectively [11] [12], however the application of ANNs for predicting OS is currently very limited.

ANNs are capable of generalizing predictions for data not observed during training based on relationships it observes between multidimensional (multivariate) input/target training data [13]. The present work utilizes ANNs trained with experimental RON/MON/OS and QSPR descriptors to create predictive models for each property. QSPR descriptors are employed due to the extensive range of attributes they describe for a given compound, allowing the ANN to form complex correlations between multiple compounds and experimental RON/MON/OS. Additionally, using predicted values of RON/MON to derive OS, or derived octane sensitivity (dOS) ($RON_{pred} - MON_{pred}$), is investigated, specifically whether a higher accuracy can be achieved compared to using an ANN to predict OS directly. Furthermore, individual relationships between QSPR descriptors and RON/MON/OS are illustrated, highlighting key components of compounds that relate to these properties.

2 Experimental Procedure

2.1 Experimental Data

Experimental data for RON and MON was collected from a multitude of sources [1, 14–27], totaling in 344 unique compounds each with an experimental value for RON and MON. It was determined that multiple sources reported predicted RON/MON data [1, 15, 27], and this data was therefore not utilized during the ANN training procedure. Furthermore, compounds that presented themselves as outliers were not considered during the procedure, specifically *n-dodecane* with RON/MON values of -40, and *undecane* with RON/MON values of -35. Upon removing compounds with predicted RON/MON values and outliers the data set consisted of 308 unique compounds. Their OS were calculated given their experimental RON/MON values ($RON_{exp} - MON_{exp}$).

The 308 compounds in the data set were split into three subsets, denoted as the training set, validation set, and test set, 80%, 10%, and 10% of the total data set respectively. Compounds for each subset were chosen such that each subset contained a proportionally equal number of compounds based on the range of experimental OS values. Each property (RON/MON/OS) utilized these subsets for ANN training, and each subset remained constant to ensure an adequate comparison of ANN accuracy, specifically the ability of the ANNs to generalize predictions for data not observed during training (test set predictions). Simple molecular-input line-entry system (SMILES) strings were produced/aggregated for all 308 compounds. *alvaDesc* was used to generate 5305 QSPR descriptors for each compound using the SMILES strings, forming unique sets of quantitative values for each compound (https://www.alvascience.com/alvadesc/). QSPR descriptors and known experimental data for each compound represent the input and target data used by the ANNs during training.

2.2 Artificial Neural Network Training

Random forest regression from the *Scikit-learn* Python package was utilized to rank each QSPR descriptor by its correlation to RON, MON, and OS, measured by a random forest regression-derived value, importance [28]. A higher value of importance implies a larger correlation between a given QSPR descriptor and RON/MON/OS [29]. The sum of all 5305 QSPR descriptor importance values is equal to 1, for each property. Regression, and the subsequent ranking of QSPR descriptors, was performed with respect to the training and validation subsets. The 250 most important QSPR descriptors for RON, MON, and OS were chosen as input variables for each property's ANN to balance accuracy and training time. Figure 1 shows the importance values of the 250 most important QSPR descriptors for RON, MON, and OS. Including fewer QSPR descriptors decreases ANN accuracy, while including more QSPR descriptors increases training time with relatively insignificant improvements in ANN accuracy. Appendix Tables A1, A2, and A3 list the 10 most important descriptors for RON, MON, and OS respectively.

ANNs were constructed using the *ECNet* Python package, an open-source Python package compiled specifically for constructing predictive models for fuel properties [30]. The ANN architecture for RON, MON, and OS consisted of 250

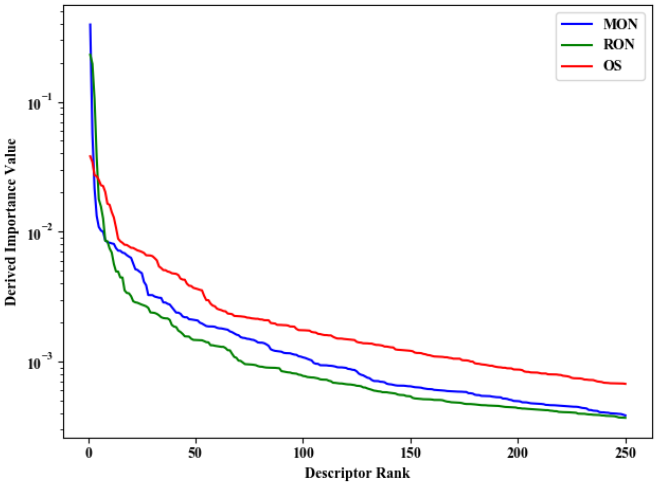

Figure 1: QSPR descriptor importance of RON, MON, and OS, ranked from most-to-least important

input neurons (one for each QSPR descriptor), two hidden layers with 128 and 64 neurons, and an output layer with 1 neuron (corresponding to RON, MON, or OS). The rectified linear unit activation function was used at each layer with the exception of the output layer which utilized a linear scaling function. The backpropagation algorithm in conjunction with the Adam optimization function were used for regression with hyper-parameter values of 0.9 for *beta 1*, 0.999 for *beta 2*, $1.0e-8$ for *epsilon*, 0.01 for *learning rate*, and 0.0 for learning rate decay [31]. ANNs for each property regressed with respect to the training subset. After each training iteration (epoch), the mean squared error of the validation subset's predictions were evaluated; training was terminated once performance ceased to improve for the validation subset to prevent overfitting. Performance of the ANN was determined given the root mean-squared error (RMSE) of predictions for the test set, providing a metric for how well the ANNs are able to generalize predictions for data not observed during training. 25 ANNs were constructed for each property to ensure consistency in results.

In addition to constructing ANNs for OS, the ANNs constructed to predict RON and MON were used to derive OS from RON/MON predictions, denoted as derived octane sensitivity dOS ($dOS = RON_{pred} - MON_{pred}$). 25 calculations of dOS were performed for each subset as a result of the 25 ANNs trained for RON and MON. The RMSEs of each subset were compared to the subset RMSEs resulting from ANNs trained directly with experimental OS data.

3 Results and Discussion

Figures 2(a-d) show parity plots for training, validation, and test set predictions, averaged across 25 trained ANNs, for RON, MON, dOS, and OS. The test set RMSEs for RON, MON, dOS, and S were found to be 10.499, 7.551, 6.432, and 7.019 respectively. Center dashed lines represent 1:1 parity, and outside dashed lines represent $\pm$ the test set's RMSE. It is seen that the test set RMSE for dOS is lower than the test set RMSE for OS, indicating that deriving OS from RON and MON predictions is more beneficial than predicting OS directly; however, it is apparent that both methods for predicting OS are relatively similar in accuracy, highlighting the viability of both methods.

Figures 3(a) and 3(b) show the relationships between RON/MON and GATS2m, a QSPR descriptor of high importance to both RON and MON. Geary autocorrelation indices, such as GATS2m, are quantifiable measurements of resemblance in neighboring point values; when autocorrelation indices are used with respect to compound structure, they indicate repeating patterns within a given compound [32]. Furthermore, a higher value of the autocorrelation index implies a more significant resemblance in neighboring atoms [33]. The visual relationships of GATS2m, an autocorrelation index weighted by mass, to both RON and MON indicate that a compound with a more uniform distribution of mass (similar mass distributions at neighboring atoms) leads to the compound having a lower value of RON/MON.

Figure 3(c) shows the relationship between OS and nCsp2, a descriptor with high importance relative to OS. nCsp2 measures the number of sp^2 hybridized carbon atoms in a given compound. An sp^2 hybridized carbon atom will have (1) electrons with a higher potential energy than a non-hybridized carbon atom, and (2) trigonal structures, resulting in three bonds to the hybridized atom and bond angles of 120 degrees. Figure 3(d) illustrates the distribution of OS at varying values of nCsp2. For each distribution, top and bottom bars represent the minimum and maximum values of OS, the center bar illustrates the median value of OS, and the width of distribution shows the overall distribution of OS. It is observed that compounds with 2 and 4 sp^2 hybridized carbon atoms tend to have higher values of OS; specific interpretations of the significance of a compound containing 2 or 4 sp^2 hybridized carbon atoms as it relates to OS have yet to be formulated, and warrants further studies into how compound structure affects OS.

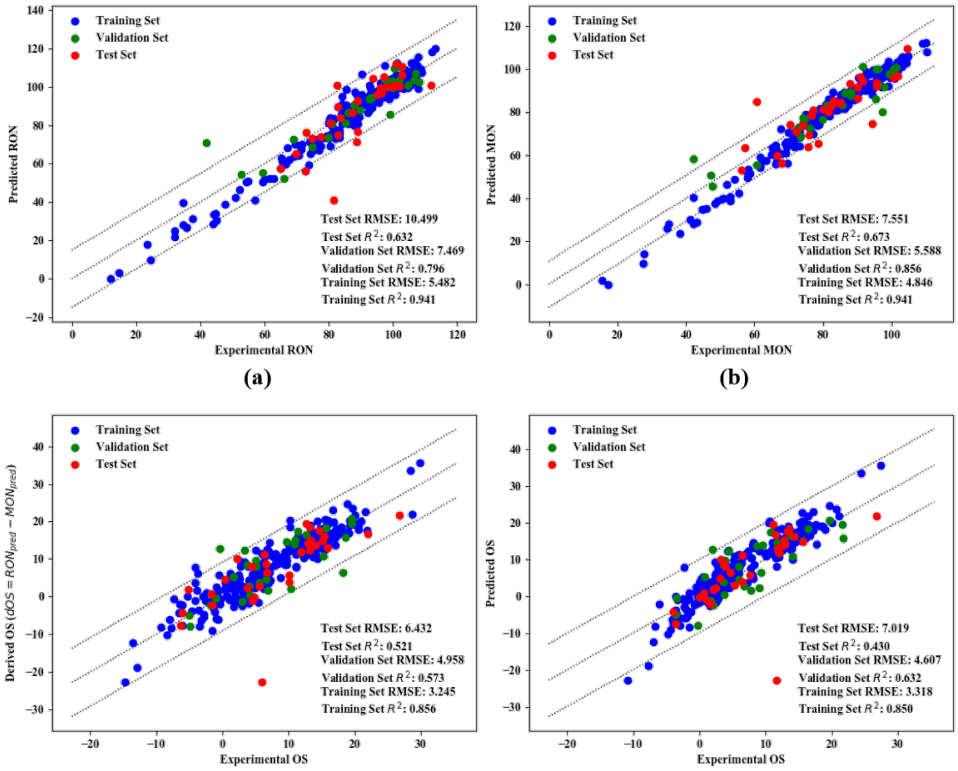

Figure 2: Parity plots showing predicted vs. experimental values for the training, validation, and test subsets for **(a)** RON, **(b)** MON, **(c)** dOS, and **(d)** OS

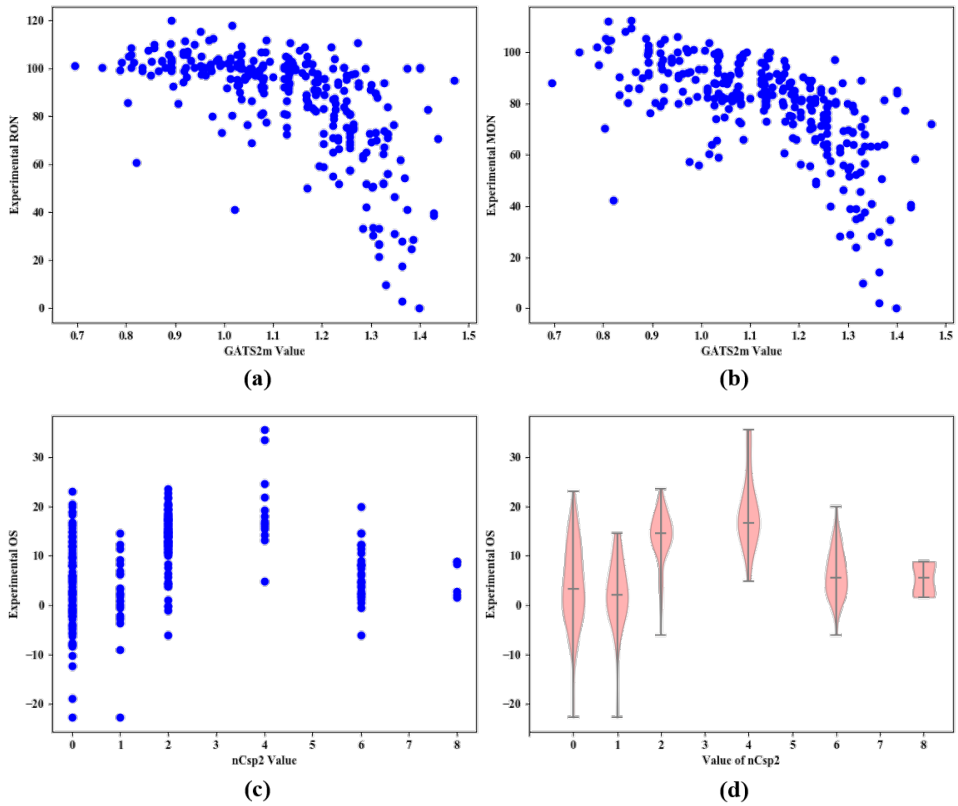

Figure 3: **(a)** relationship between RON and GATS2m **(b)** relationship between MON and GATS2m **(c)** relationship between OS and nCsp2 **(d)** distribution of OS at varying values of nCsp2

4 Conclusions and Recommendations

Based on the results outlined in the present work, it can be concluded that:

- Accurate predictive models can be constructed for RON and MON, with test set RMSEs of 10.499 and 7.551 respectively.
- Predictive models can be constructed for OS; moreover, while using RON/MON predictions to derive OS predictions shows more accuracy than predicting OS directly, both methods prove viable.
- There is an inverse relationship between mass uniformity within a compound and its RON/MON.
- Compounds with 2 or 4 sp^2 hybridized carbon atoms tend to have a higher OS.

It is recommended that (1) the proposed models be utilized for screening compounds theorized to result from renewable resources to accelerate the discovery of cleaner, more efficient, and economically viable alternatives to gasoline, and (2) further analysis be performed with respect to QSPR-property relationships, specifically to identify key structural components that relate to RON/MON/OS.

References

[1] A. Demirbas, M. A. Balubaid, A. M. Basahel, W. Ahmad, M. H. Sheikh. *Octane Rating of Gasoline and Octane Booster Additives.* Petroleum Science and Technology, 33(11), pp. 1190-1197 (2015).

[2] *ASTM D2699-18a Standard Test Method for Research Octane Number of Spark-Ignition Engine Fuel.* ASTM International (2018).

[3] *ASTM D2700-18a Standard Test Method for Motor Octane Number of Spark-Ignition Engine Fuel.* ASTM International (2018).

[4] G. Mendes, H. G. Aleme, P. J. S. Barbeira. *Determination of octane numbers in gasoline by distillation curves and partial least squares regression.* Fuel, 97, pp. 131-136 (2012).

[5] A. Yates, A. Swarts, C. Viljoen. *Correlating auto-ignition delays and knock-limited spark-advance data for different types of fuel.* SAW Technical Paper 2005-01-2083 (2005).

[6] M. Mehl, T. Faravelli, F. Giavazzi, E. Ranzi, P. Scorletti, A. Tardani. *Detailed chemistry promotes understanding of octane numbers and gasoline sensitivity.* Energy Fuels, 20(6), pp. 2391-2398 (2006).

[7] J. P. Szybist, D. A. Splitter. *Pressure and temperature effects on fuels with varying octane sensitivity at high load in SI engines.* Combustion and Flame, 177, pp. 49-66 (2017).

[8] 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 topological indices.* Russian Chemical Bulletin, 57(3), pp. 461-467 (2008).

[9] H. Yang, C. Fairbridge, A. Ring. *Neural Network Prediction of Cetane Numbers for Isoparaffins and Diesel Fuel* Petroleum Science and Technology, 19(5-6), pp. 573-586 (2001).

[10] T. Kessler, E. R. Sacia, A. T. Bell, J. H. Mack. *Artificial neural network based predictions of cetane number for furanic biofuel additives.* Fuel, (206), pp. 171-179 (2017).

[11] J. A. van Leeuwen, R. J. Jonker, R. Gill. *Octane number prediction based on gas chromatographic analysis with non-linear regression techniques.* Chemometrics and Intelligent Laboratory Systems, 25, pp. 325-340 (1994).

[12] N. Pasadakis, V. Gaganis, G. Foteinopoulos. *Octane number prediction for gasoline blends.* Fuel Processing Technology, 87(6), pp. 505-509 (2006).

[13] S. Baluja, D. Pomerleau. *Non-intrusive gaze tracking using artificial neural networks.* Advances in Neural Information Processing Systems, pp. 753-760 (1994).

[14] *Knocking Characteristics of Pure Hydrocarbons, STP225-EB.* ASTM International, West Conshohocken, PA (1958).

[15] T. E. Daubert, R. P. Danner. *API technical data book-petroleum refining.* American Petroleum Institute (API), Washington DC (1997).

[16] P. Ghosh, K. J. Hickey, S. B. Jaffe. *Development of a detailed gasoline composition-based octane model.* Industrial & Engineering Chemistry Research, 45(1), pp. 337-345 (2006).

[17] J. H. Mack, V. H. Rapp, M. Broeckelmann, T. S. Lee, R. W. Dibble. *Investigation of biofuels from microorganism metabolism for use as anti-knock additives.* Fuel, 117, pp. 939-943 (2014).

[18] E. Christensen, J. Yanowitz, M. Ratcliff, R. L. McCormick. *Renewable oxygenate blending effects on gasoline properties.* Energy & Fuels, 25(10), pp. 4723-4733 (2011).

[19] R. W. Jenkins, M. Munro, S. Nash, C. J. Chuck. *Potential renewable oxygenated biofuels for the aviation and road transport sectors.* Fuel, 103, pp. 593-599 (2013).

[20] S. R. Daly, K. E. Niemeyer, W. J. Cannella, C. L. Hagen. *Predicting fuel research octane number using Fourier-transform infrared absorption spectra of neat hydrocarbons.* Fuel, 183, pp. 359-365 (2016).

[21] J. Scherzer. *Octane-enhancing, zeolitic FCC catalysts: Scientific and technical aspects.* Catal. Rev.: Sci. Eng., 31(3), pp. 215-354 (1989).

[22] *sandialabs/BioCompoundML.* Retrieved from https://github.com/sandialabs/BioCompoundML/blob/ master/bcml/data/RON.txt (2019).

[23] R. L. McCormick, G. Fioroni, L. Fouts, E. Christensen, J. Yanowitz, E. Polikarpov, K. Albrecht, D. J. Gaspar, J. Gladden, A. George. *Selection Criteria and Screening of Potential Biomass-Derived Streams as Fuel Blendstocks for Advanced Spark-Ignition Engines.* SAE International Journal of Fuels and Lubricants, 10(2), pp. 442-460 (2017).

[24] R. L. McCormick, M. A. Ratcliff, E. Christensen, L. Fouts, J. Luecke, H. M. Chupka, J. Yanowitz, M. Tian, M. Boot. *Properties of oxygenates found in upgraded biomass pyrolysis oil as components of spark and compression ignition engine fuels.* Energy and Fuels, 29(4), pp. 2453-2461 (2015).

[25] . M. J. Pilling. *Low-temperature combustion and autoignition.* 35 (1997).

[26] A. De Klerk. *Fischer-Tropsch Refining.* John Wiley & Sons (2012).

[27] *Cloudflame.* Retrieved from https://cloudflame.kaust.edu.sa/fuel/octane_calc (2019).

[28] 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. *Scikit-learn: Machine Learning in Python* Journal of Machine Learning Research, 12, pp. 2825-2830 (2011).

[29] L. Breiman. *Random Forests.* Machine Learning, 45(1), pp. 5-32 (2001).

[30] 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).

[31] D. P. Kingma, J. Ba. *Adam: A method for stochastic optimization.* International Conference for Learning Representation (2015).

[32] S. A. Klein, C. W. Tyler. *Phase discrimination of compound gratings: generalized autocorrelation analysis.* Journal of the Optical Society of America A, 3(6), pp. 868-879 (1986).

[33] D. Chessel. *The spatial autocorrelation matrix.* Vegetation dynamics in grasslands, healthlands and mediterranean ligneous formations, pp. 177-180 (1981).

Appendix

Table A1: 10 most influential QSPR descriptors for MON, their importances, and descriptions

Descriptor Name	Importance	Description
GATS2m	0.3927	Geary autocorrelation of lag 2 weighted by mass
SssCH2	0.0571	Sum of ssCH2 E-states
SpMaxA_EA(bo)	0.0215	normalized leading eigenvalue from edge adjacency mat. weighted by bond order
X0Av	0.0132	average valence connectivity index of order 0
SIC1	0.0109	Structural Information Content index (neighborhood symmetry of 1-order)
CIC1	0.0102	Complementary Information Content index (neighborhood symmetry of 1-order)
ChiA_B(s)	0.0100	average Randic-like index from Burden matrix weighted by I-State
GATS6e	0.0086	Geary autocorrelation of lag 6 weighted by Sanderson electronegativity
GATS2e	0.0083	Geary autocorrelation of lag 2 weighted by Sanderson electronegativity
ATSC3s	0.0082	Centred Broto-Moreau autocorrelation of lag 3 weighted by I-state

Table A2: 10 most influential QSPR descriptors for RON, their importances, and descriptions

Descriptor Name	Importance	Description
ChiA_B(s)	0.2300	average Randic-like index from Burden matrix weighted by I-State
SssCH2	0.1957	Sum of ssCH2 E-states
GATS2m	0.1042	Geary autocorrelation of lag 2 weighted by mass
SpMaxA_EA(bo)	0.0373	normalized leading eigenvalue from edge adjacency mat. weighted by bond order
Eta_L_A	0.0178	eta average local composite index
SIC1	0.0156	Structural Information Content index (neighborhood symmetry of 1-order)
SpMin1_Bh(s)	0.0126	smallest eigenvalue n. 1 of Burden matrix weighted by I-state
NssCH2	0.0085	Number of atoms of type ssCH2
BIC1	0.0084	Bond Information Content index (neighborhood symmetry of 1-order)
GATS6s	0.0074	Geary autocorrelation of lag 6 weighted by I-state

Table A3: 10 most influential QSPR descriptors for OS, their importances, and descriptions

Descriptor Name	Importance	Description
AVS_B(s)	0.0380	average vertex sum from Burden matrix weighted by I-State
nCsp2	0.0344	number of sp2 hybridized Carbon atoms
SIC1	0.0272	Structural Information Content index (neighborhood symmetry of 1-order)
Chi_Dz(p)	0.0268	Randic-like index from Barysz matrix weighted by polarizability
CIC1	0.0250	Complementary Information Content index (neighborhood symmetry of 1-order)
SpMax1_Bh(s)	0.0228	largest eigenvalue n. 1 of Burden matrix weighted by I-state
SpMax_B(s)	0.0225	leading eigenvalue from Burden matrix weighted by I-State
Eta_D_epsiB	0.0202	eta measure of unsaturation
Chi1_EA(ed)	0.0164	connectivity-like index of order 1 from edge adjacency mat. weighted by edge degree
LLS_01	0.0161	modified lead-like score from Congreve et al. (6 rules)