Energy Technologies Area
Lawrence Berkeley National Laboratory

Can machine learning predict fuel properties accurately?

Morgan A. Mayer[1], Tyler Huntington[2], Ana Comesana[2], Vi H. Rapp[2], and Kyle E. Niemeyer[1],

[1]School of Mechanical, Industrial and Manufacturing Engineering, Oregon State University
[2]Advanced Development & Optimization, Lawrence Berkeley National Laboratory

October 2019

This work was supported by the Department of Energy's Bioenergy Technologies Office under Lawrence Berkeley National Laboratory Contract No. DE-AC02-05CH11231.

Disclaimer:

Sub Topic: Other

2019 WSSCI Fall Technical Meeting
Organized by the Western States Section of the Combustion Institute
October 14–15, 2019
Albuquerque, New Mexico

Can machine learning predict fuel properties accurately?

Morgan A. Mayer[1], *Tyler Huntington*[2], *Ana Comesana*[2], *Vi H. Rapp*[2], *and Kyle E. Niemeyer*[1,*]

[1]*School of Mechanical, Industrial and Manufacturing Engineering, Oregon State University, Corvallis OR, United States*
[2]*Advanced Development & Optimization, Lawrence Berkeley National Laboratory, Berkeley CA, United States*
[*]*Corresponding author:* kyle.niemeyer@oregonstate.edu

Abstract: High-potential molecules derived from biomass sources may suitably replace or supplement traditional nonrenewable hydrocarbon fuels to reduce pollution and fuel processing cost. Experimental property testing of these bioproducts is usually conducted years after initial bench-scale experiments, due to high experimental costs and/or high volume requirements. However, neglecting to conduct property testing early in the pathway development cycle can lead to investments spent on scaling-up production of bioproducts and biofuels that do not perform as expected. Instead, machine-learning techniques can be used to develop quantitative structure–property relationships for molecules using a relatively large training set of molecular descriptor data. For this study, we compiled measured properties, IR spectra, and molecular descriptors of bio-based molecules from databases and published studies for training models of bioproduct properties. We trained regression models with molecular descriptors and will compare results of different estimators. This study describes the first steps towards a performance prediction tool for bio-based alternative fuels.
Keywords: Machine learning, biofuels, jet fuels, fuel properties

1. Introduction

Recent research efforts have been aimed at securing a more reliable energy supply with reduced emissions for the transportation industry. Alternative fuels, such as those derived from biomass, may supplement or replace traditional transportation fuels in an effort to reduce carbon emissions and increase energy yield. Alternative fuels research is primarily advanced through complex chemical and biological production techniques. Biomass conversion to fuel offers a variety of chemical pathways and molecules that may be able to serve as an alternative fuel. These synthetic fuels may be produced through Fischer-Tropsch synthesis, hydrotreatment of vegetable oils (HVO), or fermentation.

The U.S. Department of Energy's (DOE's) Co-Optimization of Fuels & Engines (Co-Optima) initiative [1] is taking a fuel property-based approach to identify a set of blendstocks that meet performance and fuel quality standards. Fuel testing can be costly even at the bench scale and often requires large quantities of fuel. To reduce the associated investment risk, machine learning is being investigated as a means of predicting fuel properties that reflect the ignition behavior. A property prediction tool could help vet fuels through an initial screening before they enter the certification process. It would ultimately enable faster, less expensive bioprocess optimization and

scale-up. Relevant to diesel fuels, the cetane number is one of the most important properties as it is a measure of the ignition quality. Cetane number (CN) can be obtained directly via cooperative fuel research (CFR) engine or indirectly via an ignition quality test (IQT) as derived cetane number (DCN). IQT is advantageous because it only requires a tenth of fuel used in a CRF engine test and is faster. However, the accuracy between the two methods is not well developed and varies between fuels. An accurate prediction of cetane number could accelerate the assessment of potential diesel alternatives.

Well-known machine learning algorithms have already been used to develop quantitative structure property relationship (QSPR) models. Estimators are tested and selected based on the target property and class of molecule. These techniques have been applied with molecular descriptors or optical measurements as the inputs. Molecular descriptors are quantified molecular-based theoretical parameters mathematically derived solely from molecular structures [2]. A molecular descriptor is a scalar value that may be considered one-dimensional, two-dimensional, or three-dimensional depending on the information used to compute it. The descriptor can be a simple atom count to a more complex quantum mechanical parameter that describes electron distribution. There are several open-source and proprietary software that can compute molecular descriptors. Mordred [3], used in this study, is an open-source Python application that can compute more than 1800 descriptors based on a molecule's SMILES, simplified molecular-input line-entry system.

A machine learning model is only as stellar as the quality and selection of inputs, or features, that train the model. A major consideration in a machine learning project is choosing which features to include in training, known as feature engineering. Too many irrelevant features can cause overfitting to noise, and too few features leads to a model that is underfit. In addition, two features may not have strong independent correlations with the target, but have a stronger combined correlation to the target. Certain features may be closely correlated with each other, so it's preferable to also combine these features to reduce overfitting. More samples may be required with more features to obtain a certain performance, so features are often combined when possible. For QSPR studies, it is difficult to intuit all of the correlations between 1800 descriptors, so researchers often use statistical studies in addition to physical intuition for feature engineering. Guyon and Elisseeff [4] suggest using both domain knowledge and physical intuition is a sound approach. Creton et al. [5] recommend selecting descriptors based on mathematical criteria first then ensuring that set has the descriptors corresponding to chemical intuition.

Previous QSPR studies implement various feature and model selection techniques based on molecule classes and target properties. Kessler et al. [6] used a backpropagating neural network to predict cetane number for a diverse molecule set including furanic compounds. They trained artificial neural network model using 290 molecules and 15 descriptors as features. Features were narrowed down from 1667 to 15 using an iterative regression analysis technique. The overall RMSE for the model was 5.95 CN units. St. John et al. [7] developed a model for sooting index using a multi-layer perceptron function in support with Scikit-learn regression functions. A recursive feature elimination strategy with support vector machine regression yielded 390 descriptors that produced the lowest median absolute error in cross-validation. Elton et al. [8] tested the predictive performance of five regression techniques and six featurization methods on nine properties of energetic materials. With a relatively small but diverse dataset of 109 molecules, they determined that kernel ridge regression produced the best model for properties of CNOHF energetic materials. A hand-picked set of 21 molecular descriptors was included as a tested featurization method, however the sum over bonds vector produced the lowest mean absolute error. Wang et al.

[9] evaluated property estimation methods for hydrocarbon fuels using mid-IR spectra as features. They optimized the Lasso regularization hyperparameters and determined the optimal number of discretized wavelength features for the best model performance. Creton et al. [5] developed four CN models for four different classes of molecules using multi-linear regression of molecular descriptors. They generated a correlation matrix to select features and only kept features that highly correlated with the target, CN, and poorly correlated with another descriptor. The models were trained on a relatively small number of molecules ranging from 21–38 molecules and five to seven features were used. Saldana et al. [2] used consensus modelling to combine linear and nonlinear QSPR models and found the averaged consensus model had the best property predictions for hydrocarbons.

This study aims to evaluate machine learning regression methods in predicting cetane number of bio-based molecules.

2. Methods

Regression methods via Scikit-learn were used to evaluate cetane number prediction. Scikit-learn [10] is a Python-based automated machine learning (AutoML) suite. The following regression functions from Scikit-learn were used: `ExtraTreesRegressor(n_estimators=10)`, `RandomForestRegressor(n_estimators=10)`, `DecisionTreeRegressor()`, `KernelRidge(alpha=1.0)`, `Lasso(alpha=0.1)`, and `ElasticNet()`. Default hyperparameters were used unless otherwise specified. K-fold cross evaluation from Scikit-learn divided the full dataset into five subsets and calculated RMSE of the five splits of data with four subsets used as training and one was set aside for validation. Molecular descriptors were calculated by Mordred, an open-source Python software for molecular descriptor calculations. Mordred [3] was used to calculate the full dataset of 941 molecular descriptors for 481 molecules. Experimental cetane number data was compiled from the ECNet database [11].

Performance of regression models were found using the full dataset of 941 features and also with smaller subsets of 20 and 5 of the most important features. The `ExtraTreesRegressor` class was used on 1000 random subsets of 80% of the data to determine a set of most important parameters (via the `feature_importances_` attribute). For cetane number, roughly 200–220 of the most important features encompassed 95% of the total feature importance when evaluating feature importance with the full dataset. A pitfall of using `ExtraTreesRegressor.feature_importances_` is that a value is blindly, independently varied in a range and performance is assessed, which may create a set of parameters that is not a real scenario for a molecule. The 20 most important features using this method are listed in Table 2.

TPOT [12, 13], the Tree-based Pipeline Optimization Tool, is an AutoML software openly available on Github. TPOT randomly searches various data pipelines to determine the best data processing pipeline for predicting a target variable. TPOT was used to calculate RMSE for an input dataset of 481 molecules and 941 descriptors. TPOT was run for 24 hours to reveal a high-performing machine learning pipline.

IR spectra of 40 fuels was performed at Oregon State University Cascades campus. FTIR measurements were done in the range of 650–4000 nm. With the IR spectra used as input features, each of the 13898 absorbances for a fuel was considered an attribute.

Descriptor	Definition
RotRatio	rotatable bonds ratio
ATSC2m	centered moreau-broto autocorrelation of lag 2 weighted by mass
ATSC2Z	centered moreau-broto autocorrelation of lag 2 weighted by atomic number
ATSC2v	centered moreau-broto autocorrelation of lag 2 weighted by vdw volume
NssCH2	number of ssCH2
SssCH2	sum of ssCH2
nRot	rotatable bonds count
ATSC2i	centered moreau-broto autocorrelation of lag 2 weighted by ionization potential
ATSC2p	centered moreau-broto autocorrelation of lag 2 weighted by polarizability
GATS2c	geary coefficient of lag 2 weighted by gasteiger charge
C2SP3	SP3 carbon bound to 2 other carbons
SIC1	1-ordered structural information content
GATS2m	geary coefficient of lag 2 weighted by mass
GATS2Z	geary coefficient of lag 2 weighted by atomic number
ETA_dBeta	ETA delta beta
SpMAD_Dzp	spectral mean absolute diviation from Barysz matrix weighted by polarizability
CIC1	1-ordered complementary information content
CIC3	3-ordered complementary information content
GATS2se	geary coefficient of lag 2 weighted by sanderson EN
GATS2are	geary coefficient of lag 2 weighted by allred-rocow EN

Table 1: The 20 most important descriptors as features using the `ExtraTreesRegressor.feature_importances_` module and randomly splitting the data 1000 times. Descriptor definitions from Mordred website [3].

RMSE was taken as the model performance metric. RMSE was computed as

$$RMSE = \sqrt{\frac{1}{N}\Sigma_{i=1}^{N}\left(y_i^{\text{true}} - y_i^{\text{pred}}\right)^2}\,, \tag{1}$$

where N is the number of samples, y_i^{true} is the measured target value, and y_i^{pred} is the predicted target value. Percent error of limonene cetane number was calculated with

$$E = 100\% \times \frac{CN^{\text{meas}} - CN^{\text{pred}}}{CN^{\text{meas}}}\,, \tag{2}$$

where E is the percent error, CN^{meas} is the measured cetane number, and CN^{pred} is the predicted cetane number.

3. Results and Discussion

K-fold cross validation with $k = 5$ folds was done with six regression methods using the full set of 480 molecules, limonene excluded, and 941 descriptors used as features. In addition, smaller subsets of the 20 and 5 most important features trained regression models. In nearly all of the 1000 random subsets of the data, the RotRatio, or rotatable bonds ratio, importance described between 20 - 30% of the total importance. Figure 1 shows the average RMSE of the five validation sets. Hyperparameters were set to default values. The `RandomForestRegressor()` and `ExtraTrees()` regressors have the lowest average validation RMSE when all 941 descriptors are used as features.

Regression models for cetane number based on the same three sets of features described above were used to predict limonene cetane number. A random subset of 80% of the data was used for model training. Using the full set of descriptors, the `RandomForestRegressor()` and `ExtraTrees()` regressors method predicts limonene cetane number the best. In all cases, the `RandomForestRegressor()` using 20 features seems to predict limonene cetane number the best. The `Lasso()` regressor also well predicts limonene cetane number using just five features. The `DecisionTreeRegressor()` most poorly predicts limonene cetane number with five features and 20 features.

4. Conclusions

Traditional regression techniques were evaluated on a dataset of 480 molecules using k-fold validation. Using the `ExtraTrees()` regressor, the rotatable bonds ratio regularly described the highest fractional importance in random subsets of 80% of the data. Limonene's cetane number was predicted to within 1% using the Elastic Net regression method, however, performance on varied training splits should be evaluated further. With hundreds of features there is a large change of overfitting the data to noise, so feature selection in a key element to model creation.

Next steps will be to evaluate model performances based on changing hyperparameters and also discovering better featurization techniques. The `ExtraTreesRegressor.feature_importances_` module changes features individually in a range to assess model performance which may simulate a molecule with unrealistic parameters and judge the importance based off of impossible descriptor combinations for molecules. This is especially amplified in features that are coupled that could be condensed into one attribute. In addition, performance will be assessed for different individual molecules and different classes of molecules.

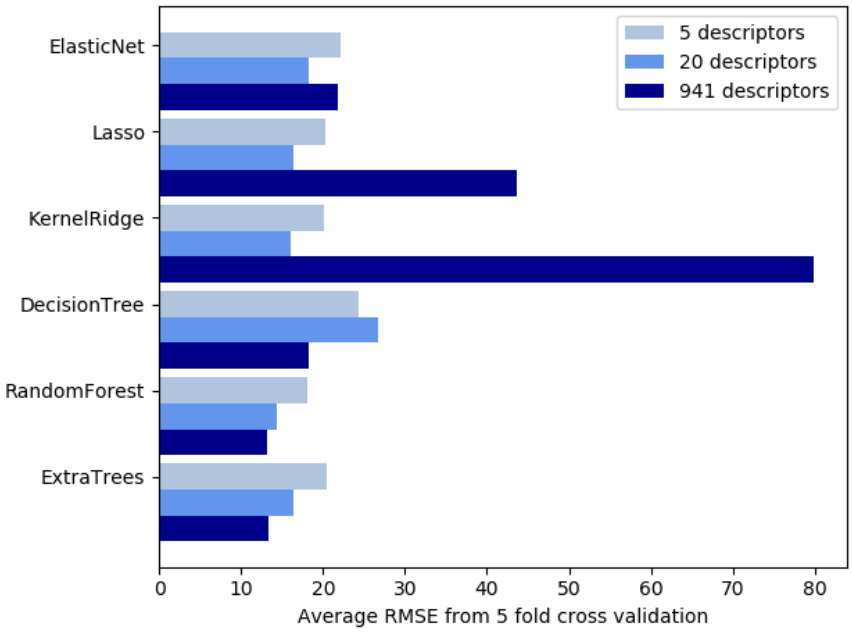

Figure 1: Average RMSE of validation sets for 5-fold cross validation for regression methods. Full dataset of 480 molecules (limonene excluded) with different subsets of features tested: all 941 descriptors, the 20 most important, and the 5 most important. 20% of the data was used for validation in each split.

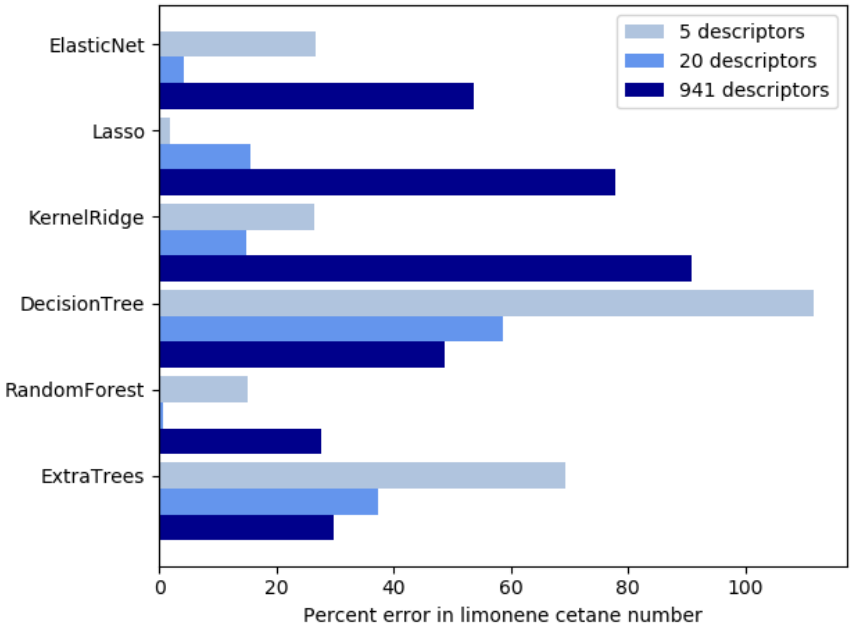

Figure 2: Absolute percent error in limonene cetane number prediction for six regression methods. A random subset of 80 % of the full dataset of 480 molecules (limonene excluded) trained the models with different subsets of features tested: all 941 descriptors, the 20 most important, and the 5 most important. Hyperparameters were set to default values for all regressors.

Sub Topic: Other

5. Acknowledgements

This project was funded by Lawrence Berkeley National Laboratory funded by the Department of Energy's Bioenergy Technologies Office, under prime contract number DE-AC02-05CH11231.

References

[1] J. Farrell, J. Holladay, and R. Wagner, Fuel Blendstocks with the Potential to Optimize Future Gasoline Engine Performance: Identification of Five Chemical Families for Detailed Evaluation, tech. rep. Report No. DOE/GO-102018-4970, U.S. Department of Energy, Washington, DC, 2018.

[2] D. A. Saldana, B. Creton, P. Mougin, N. Jeuland, B. Rousseau, and L. Starck, Rational Formulation of Alternative Fuels using QSPR Methods: Application to Jet Fuels, Oil & Gas Science and Technology – Rev. IFP Energies nouvelles, 68 (2013) 651–662. DOI: 10.2516/ogst/2012034.

[3] H. Moriwaki, Y.-S. Tian, N. Kawashita, and T. Takagi, Mordred: a molecular descriptor calculator, Journal of Cheminformatics 10 (2018) 4. DOI: 10.1186/s13321-018-0258-y.

[4] I. Guyon and A. Elisseeff, An Introduction to Variable and Feature Selection, Journal of Machine Learning Research 3 (2003) 1157–1182.

[5] B. Creton, C. Dartiguelongue, T. de Bruin, and H. Toulhoat, Prediction of the Cetane Number of Diesel Compounds Using the Quantitative Structure Property Relationship, 24 (2010) 5396–5403. DOI: :10.1021/ef1008456.

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

[7] P. C. St. John, P. Kairys, D. D. Das, C. S. McEnally, L. D. Pfefferle, D. J. Robichaud, M. R. Nimlos, B. T. Zigler, R. L. McCormick, T. D. Foust, Y. J. Bomble, and S. Kim, A Quantitative Model for the Prediction of Sooting Tendency from Molecular Structure, Energy and Fuels 31 (2017), DOI: 10.1021/acs.energyfuels.7b00616.

[8] D. C. Elton, Z. Boukouvalas, M. S. Butrico, and P. W. Fuge Mark D.and Chung, Applying machine learning techniques to predict the properties of energetic materials, Scientific Reports (2018), DOI: 10.1038/s41598-018-27344-x.

[9] Y. Wang, Y. Ding, W. Wei, Y. Cao, D. F. Davidson, and R. K. Hanson, On estimating physical and chemical properties of hydrocarbon fuels using mid-infrared FTIR spectra and regularized linear models, Fuel 255 (2019) 115715.

[10] 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, and E. Duchesnay, Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[11] T. Kessler and J. H. Mack, ECNet: Large scale machine learning projects for fuel property prediction, Journal of Open Source Software 2 () 401. DOI: 10.21105/joss.00401.

[12] R. S. Olson, R. J. Urbanowicz, P. C. Andrews, N. A. Lavender, L. C. Kidd, and J. H. Moore, "Applications of Evolutionary Computation: 19th European Conference, EvoApplications 2016, Porto, Portugal, March 30 – April 1, 2016, Proceedings, Part I", in: ed. by G. Squillero and P. Burelli, Springer International Publishing, 2016, chap. Automating Biomedical Data Science Through Tree-Based Pipeline Optimization, pp. 123–137.

[13] R. S. Olson, N. Bartley, R. J. Urbanowicz, and J. H. Moore, Evaluation of a Tree-based Pipeline Optimization Tool for Automating Data Science, Proceedings of the Genetic and Evolutionary Computation Conference 2016 ACM, Denver, Colorado, USA, (2016), pp. 485–492.