energy&fuels

Article

pubs.acs.org/EF

BioCompoundML: A General Biofuel Property Screening Tool for Biological Molecules Using Random Forest Classifiers

Leanne S. Whitmore,[†,‡] Ryan W. Davis,[†] Robert L. McCormick,[§] John M. Gladden,[†,‡] Blake A. Simmons,[†,¶] Anthe George,[†,‡] and Corey M. Hudson*[,†,‡]

[†]Sandia National Laboratories, Livermore, California 94551, United States

[‡]Joint BioEnergy Institute, Emeryville, California 94608, United States

[§]National Renewable Energy Laboratory, Golden, Colorado 80401, United States

[¶]Lawrence Berkeley National Laboratory, Berkeley, California 94720, United States

Supporting Information

ABSTRACT: Screening a large number of biologically derived molecules for potential fuel compounds without recourse to experimental testing is important in identifying understudied yet valuable molecules. Experimental testing, although a valuable standard for measuring fuel properties, has several major limitations, including the requirement of testably high quantities, considerable expense, and a large amount of time. This paper discusses the development of a general-purpose fuel property tool, using machine learning, whose outcome is to screen molecules for desirable fuel properties. BioCompoundML adopts a general methodology, requiring as input only a list of training compounds (with identifiers and measured values) and a list of testing compounds (with identifiers). For the training data, BioCompoundML collects open data from the National Center for Biotechnology Information, incorporates user-provided features, imputes missing values, performs feature reduction, builds a classifier, and clusters compounds. BioCompoundML then collects data for the testing compounds, predicts class membership, and determines whether compounds are found in the range of variability of the training data set. This tool is demonstrated using three different fuel properties: research octane number (RON), threshold soot index (TSI), and melting point (MP). We provide measures of its success with these properties using randomized train/test measurements: average accuracy is 88% in RON, 85% in TSI, and 94% in MP; average precision is 88% in RON, 88% in TSI, and 95% in MP; and average recall is 88% in RON, 82% in TSI, and 97% in MP. The receiver operator characteristics (area under the curve) were estimated at 0.88 in RON, 0.86 in TSI, and 0.87 in MP. We also measured the success of BioCompoundML by sending 16 compounds for direct RON determination. Finally, we provide a screen of 1977 hydrocarbons/oxygenates within the 8696 compounds in MetaCyc, identifying compounds with high predictive strength for high or low RON.

BACKGROUND

The assertion that fuel performance is controlled by a finite number of meaningful fuel properties is essential in the study, down-selection, and *a priori* assessment of various fuels and fuel-related compounds. The challenge in evaluating from thousands to hundreds of millions of potential known substances and compounds, without direct experimental characterization poses a daunting task. A number of techniques exist for dealing with this problem. A common approach is the use of chemical intuition to filter compounds by known, established, and experienced properties.[1] Although powerful, the chief limitation of chemical intuition is the limitation of all manual human efforts, namely, the efficient and tedious evaluation of large amounts of data.[2] To this end, a considerable effort has been expended over the last half-century in generating computational methods for predicting quantitative structure−property relationship (QSPR) and quantitative structure−activity relationship (QSAR) of various fuel properties.[3] The techniques for studying these vary considerably: from regression-based methods (e.g., multiple linear regression, neural network general regression, and partial least squares) to classification methods (e.g., ensemble trees, support vector machines, logistic regression, and linear discriminant analysis).[4] They also cover a large number of compound properties related to fuel performance, including cetane number,[5] octane number,[6] threshold soot index (TSI),[7] flashpoint,[5] melting point (MP),[8] and others. In addition to these machine-learning methods, other more recent approaches have developed model-based screens for fuel candidates using computer-aided molecular design.[9]

In this paper, we present a general-purpose open-source machine-learning classifier for screening compounds for desirable fuel properties. In compound screening, we do not predict specific values of fuel properties but rather the probability that a property value is above or below a defined threshold. For example, spark-ignition (SI) engines require fuel with a minimum research octane number (RON). Our classifier can predict the probability that a compound with unknown RON has a value above the required minimum. In contrast to more resource-intensive computational tools (e.g., quantum mechanical methods and density functional theory), our classifier trains in minutes and processes individual compounds

Received: August 5, 2016
Revised: September 13, 2016
Published: September 15, 2016

DOI: 10.1021/acs.energyfuels.6b01952
Energy Fuels 2016, 30, 8410−8418

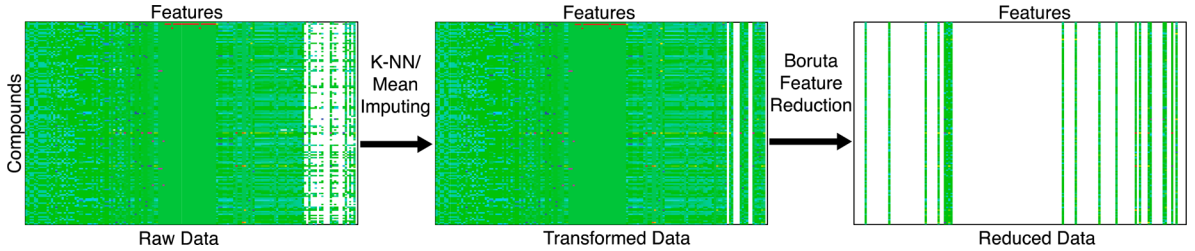

Figure 1. Feature imputation and selection. The first panel of this figure shows the raw data for RON-selected compounds using the fingerprint of NCBI and experimental data. In this panel, white cells correspond to missing data. These data are then imputed using *k*-NN and mean value in the second panel. Boruta feature selection/reduction is this used to remove uninformative features. The third panel shows the removal (in white) of uninformative features from the data set.

in seconds. While BiocompoundML may be used for any chemical property, we use this tool to present our predictive classifiers for RON, TSI, and MP. These three properties have special importance in that they define three key aspects of fuel performance, namely, the anti-knock performance (RON), the atmospheric/health impacts (TSI), and the ability for a pure compound to handle and blend into a fuel (MP).

Our software is a discrete classification tool, which allows it to be used to quickly screen compounds. Included in this package are two additional important features: (1) a clustering tool and (2) a feature reduction function. The clustering tool was designed with screening in mind. Because we anticipate one common use case is the screening of a large number of compounds, our automated clustering algorithm helps ensure that tested compounds bear structural resemblance to training compounds, prior to classification. Our feature reduction function implements Boruta feature selection[10] to automatically reduce the number of features in the model. The Boruta selection feature is important because our feature selection protocol includes an automated feature collection package for obtaining data directly from the open chemical repository PubChem,[11] an implementation of PaDEL-Descriptor (http://padel.nus.edu.sg/software/padeldescriptor/) for generating QSAR descriptors (called using the chemofeatures parameter in BioCompoundML), and an interface for providing user-supplied features. The interface is especially useful when including experimental, proprietary, or closed-licensed data sources. Users can, in fact, provide separate substance data files (SDFs) in training and testing directories to collect QSAR descriptors for prediction using PaDEL-Descriptor.

This software is evaluated using three strategies. The first strategy is leave-out resampling, using 50% build versus 50% test data sets, for RON, TSI, and MP. The second strategy directly measured RON for 16 compounds not included in our training data set. The third strategy involved testing the 8696 compounds stored in the MetaCyc database.[12] Nearly all of these compounds are known to be biologically produced and were thus evaluated to screen for potentially high-performance biofuel blendstocks.

This software differs in a variety of ways from other general-purpose chemical model development tools. Unlike camb, an open-source chemical model development tool, written in R, BioCompoundML does not focus specifically on protein activity or quantification focusing but rather on chemical characterization and classification.[13] BioCompoundML is a fully packaged distribution, which allows it to be run on user data, directly from a personal computer, server, or cloud instance. This separates BioCompoundML from other tools, such as OCHEM, which provide an online development environment for chemical property prediction and classification.[14] BioCompoundML is a fully functional executable program, created in Python. With the exception of selecting parameters and providing training and testing data, users should not have to directly interact with the provided libraries, unlike scikit-chem (https://github.com/richlewis42/scikit-chem) and RDKit (https://github.com/rdkit/rdkit), which provide a set of Python libraries for designing chemical analysis workflows, predicting common features, and converting between data formats.

METHODS AND IMPLEMENTATION

Feature Collection. BioCompoundML is designed to be used in conjunction with PubChem Entrez application programming interface (API) of the National Center for Biotechnology Information (NCBI). We include with this package a feature extraction package that directly collects experimental and fingerprint data from NCBI using either the compound identifier (CID) of NCBI or standard Chemical Abstracts Service (CAS) Registry Numbers. Two main sets of features are extracted: experimental/estimated features and fingerprints. The experimental/estimated features that are used by the feature extraction package include experimentally measured properties (e.g., MP, boiling point, and vapor pressure), inferred structural features (e.g., rotatable bond count and heavy atom count), chemical properties (e.g., molecular weight and formal charge), and inferred chemical properties (e.g., XLogP3, which estimates the octanol−water partitioning, a property directly related to hydrophobicity). PubChem fingerprints are 881 properties relating to the compound structure. These include, for example: >4H, ≥1 any ring size 6, and simplified molecular-input line-entry system (SMILES) patterns [e.g., C(−C)(−C)(═C)]. A full list of fingerprints is available at ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.txt. BioCompoundML is designed to connect directly to PubChem and extract these features for any compound in its training or testing set. These retrievals can, however, also be turned off directly. Additionally, this package also has the capacity to retrieve SDFs from NCBI. These may be useful in downstream QSPR/QSAR feature extraction.

BioCompoundML additionally includes a copy of PaDEL-Discriptor, a molecular descriptor calculator, that takes as input a SDF and provides thousands of individual QSPR and QSAR descriptors for each compound.[15] By default, BioCompoundML calculates 1444 of these descriptors (1D/2D descriptors). This software is provided with its open source Apache 2.0 license.

Commonly, analysts with interest in fuel properties will have access to proprietary data, experimental results, curated features, or licensed calculators. BioCompoundML includes the option of adding these features to the training and testing data sets. These features may be provided to the program as separate columns in the input data files.

Feature Reduction/Selection and Value Imputing. The most obvious means of feature reduction is to remove all non-variable features. This is performed as a first step, following feature collection. The next step of imputing missing values is achieved using a two-step approach. The first step is to perform *k*-nearest neighbor (*k*-NN)

DOI: 10.1021/acs.energyfuels.6b01952
Energy Fuels 2016, 30, 8410−8418

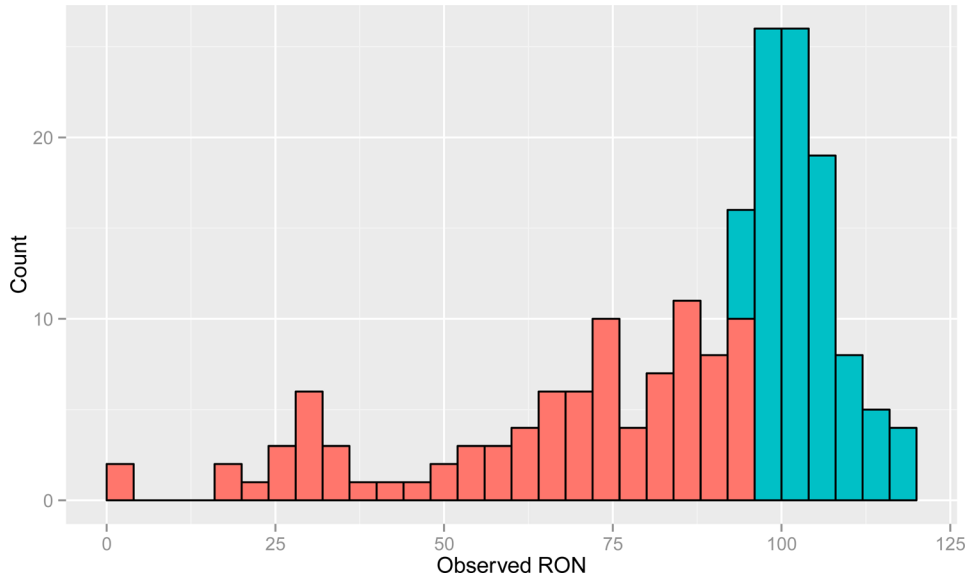

Figure 2. Distribution of RON values in training data. RONs in the literature range from 0 to 120. Our classifier is particularly interested in compounds that are higher than 94.4. The change in color on this figure corresponds to this boundary.

imputation.[16] This process takes a distance matrix and imputes missing values using the *k*-NN. The distance matrix in this tool is calculated using the Jaccard distance/Tanimoto score,[17] using the 881 NCBI fingerprint variables. This allows for the distance matrix to be collected separately from value imputation. This matrix is used to identify the nearest neighbors. The default for BioCompoundML is $k = 5$. The distance matrix is then used to assign a weight to each value for the nearest neighbors and return a weighted average, such that nearer neighbors (in this case, compounds) are more heavily weighted. This approach is generalizable and has shown consistent success as an approach to missing data.[18] In cases where features were too sparse to fully resolve using *k*-NN, we used the mean value for the feature as a minimum information imputer (see Figure 1).

In addition to feature reduction, we were also interested in reducing data complexity through feature selection. We used the Boruta algorithm for selecting features for random forest classification.[10] The original Boruta algorithm was written in R. We use a separate package, written by Daniel Homola for Python https://github.com/danielhomola/boruta_py/. This algorithm works by generating a set of shadow random features, by duplicating and then shuffling the variables. These are used to create *z*-score distributions for each feature, through random iteration, followed by classification using random forest. Each original feature is compared to the maximum *z* score for shadow features. Features that do not score significantly better than the shadow features are excluded from the model. These can dramatically reduce the complexity of the model, by eliminating uninformative features (see Figure 1).

Random Forest Classification. Random forest classifiers work by creating ensembles of decision trees using randomized "bags of features".[19] These features are randomly sampled and classified into decision trees. These trees are then used in the process of "voting" for a machine-learned model. The incorporation of randomness into the training stage[19] means that, as the error rate decreases in the training set, through the addition of trees, the error rate in the testing set remains constant.[20] This is due to strong classifiers dominating the decision process, at the expense of weak (overfitting-prone) classifiers.[21]

An additional feature of the random forest classifier is its ability to separate tree-solving steps on parallel processes.[22] This makes the steps prior to voting, embarrassingly parallel. This time-consuming step can be handled by separate central processing units (CPUs), allowing the classifier to scale efficiently.

Use of Clustering To Limit Tests. As part of BiocompoundML, we include a clustering tool, which screens and selects compounds that are structurally similar to the training set. First, our clustering tool evaluates similarities and differences between compounds in the training data set. PubChem fingerprints, prior to feature reduction by Boruta, are transformed using the random trees embedding algorithm.[23] This unsupervised transformation algorithm uses a forest of random trees to encode data (compounds) by the indices of leaves that each compound falls into, resulting in a high dimensional sparse binary code. Dimensionality reduction by truncated singular value decomposition (SVD) is then implemented on the sparse binary codes.[24] Using this transformed model of the training data set, we identify the distinct number of clusters in the training data by calculating the silhouette coefficient and *k*-means clustering for 2−8 clusters. The cluster number with the highest average silhouette score is selected as the number of clusters that exist within the training data.

This clustering tool can then be used to evaluate the training data in the context of the testing data and extracts test data similar in structure to the compounds in the training. Initially, all testing data, prior to feature reduction by Boruta, are transformed using random trees embedding and dimensions are reduced via truncated SVD, creating a model of the testing data. Training data are then fitted to the testing data model. Each cluster (determined in the first step) of the fitted training data is used to next create a OneClassSVM model for each of the clusters. OneClassSVM is an unsupervised outlier detection method, which allows for the prediction of whether test data are similar to the training clusters or outlie them. Each of the fitted test compounds are run through each of the OneClassSVM training cluster models. If a compound is determined to be similar to at least one of the clusters, further classification and prediction of the compound occurs; otherwise, the compound is disregarded.

RESULTS

RON Classifier Development. A principle fuel property of interest for contemporary SI engines is the RON. RON is a measure of the resistance of a chemical to autoignition, which can occur in the fuel−air mixture ahead of the flame front created by the spark if temperature and pressure become too high for the autoignition resistance of the fuel. RON is one of the hallmark properties of SI engine fuels, and future, high-efficiency engines may require significantly higher RON than available in the market today.[25]

For this analysis, we collected RON measurements from 148 pure training compounds from our review of the literature[26−29]

DOI: 10.1021/acs.energyfuels.6b01952
Energy Fuels 2016, 30, 8410−8418

and 36 internally tested compounds. All but three of the internally tested compounds were tested according to the ASTM D2699 standard. The other compound RON values were estimated using derived cetane numbers, which is not exact but allows for discrete classification.[30] These compounds had a median RON of 94.4 (see Figure 2). We chose this RON as the threshold for the classifier, because it (1) evenly split the data set and (2) generally measures high versus low RON, as it is experimentally evaluated. To build our BioCompoundML model, we selected 881 NCBI fingerprint features and 26 NCBI experimental features. Only 157 of these features were variable in our training set. After Boruta feature selection (following *k*-NN imputation), only 9 features proved to be useful for model prediction (Table 1).

Table 1. RON Predicting Features by Weight and Type

features	weight	type
XLogP3	0.2689	physical
autoignition	0.1845	physical
C−C−C−C−C−C	0.1360	structural
LogP	0.1144	physical
XLogP-AA	0.0928	physical
complexity	0.0918	physical
boiling point	0.0488	physical
vapor pressure	0.0415	physical
C−C−C−C−C	0.0213	structural

Using 100 random 2-fold splits of the data, we observed high accuracy, precision, and recall. The average accuracy in these 50% leave-out experiments was 87.79%, with a standard deviation of 5.8%. Precision was 88.16%, with a standard deviation of 11.4%. Recall was 88.17%, with standard deviation of 11.3%. The receiver operator characteristic (ROC) area under the curve (AUC) was 0.88, with 0.056 standard deviations. Figure 3 shows the ROC curves for 10 random 50% leave outs. It is expected that the performance of the RON classifier will dramatically improve as the diversity of training data increases, expanding the functional group space. The current functional groups for the training set include alkenes, alkanes, alkynes, benzenes, cycloalkenes, cycloalkanes, alcohols, esters, ketones, and naphthalenes.

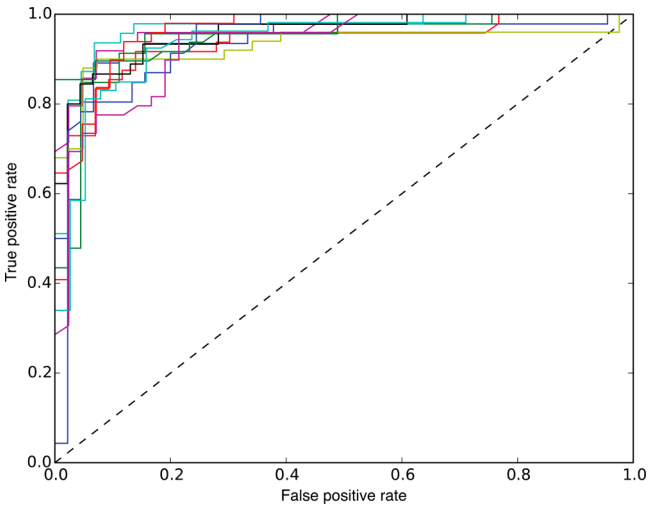

Figure 3. ROC curves for 10 random 50% leave-out experiments. Average AUC is 0.82. This figure shows 10 random ROC curves with 50% for training and 50% left out for testing.

The heaviest weighted features in the classifier are shown in Table 1. The most heavily weighted feature, XLogP3, is a measure of water−octanol partitioning. A high XLogP3 is predictive of a high RON value. The other LogP-related features (i.e., LogP and XLogP-AA) are alternatively calculated estimates of LogP. The temperature of autoignition, boiling point (at 760 mmHg), and vapor pressure (mmHg at 25 °C) are also important features in predicting whether a compound has a RON value greater than or less than 94.4. The structural features C−C−C−C−C−C and C−C−C−C−C, which are SMILES patterns that refer to the presence of six- and five-carbon structures, are also important. It is necessary to note that, although some features correlated directly with RON, ensembles of decision trees work hierarchically, meaning that feature interactions are not necessarily additive or direct. These weightings can, however, give insight into the physical attributes that govern the property of interest (in this case, RON) and, hence, the attributes that could merit further experimental investigation.

Using our RON model, we tested our results using 16 additional compounds not used to build the model. RON for these compounds was measured using the ASTM D2699 standard and has not previously been measured. They were selected on the basis of our interest in them as potential targets for synthetic bioengineering.

From the SI engine fuel perspective, there are essentially two classes: RON significantly >94.4 and RON not significantly >94.4. These compounds considerably expanded the range of variability, from the training data, and serve as a test of the expandability of our model. A linear regression model of classification probability of RON > 94.4 by observed RON has a $R^2 = 0.653$ ($p < 0.001$) (see Figure 4). On the basis of this linear regression, the classification decision boundary (the point where the model intersects 94.4, given $y = 0.0074x + 0.1547$) is 0.849. Using this decision point, all tested molecules have been correctly assigned. However, including the 95% confidence interval at the decision boundary (94.4), the range of marginal values covers from $p > 0.800$ to $p < 0.897$. One compound (linalool with an observed RON = 96.7) falls within this marginal region. This suggests that the observed accuracy [(true positives + true negatives)/total] is between 93.75 and 100% (see Table 2).

RON Evaluation for Large Numbers of Compounds Using BioCompoundML. We collected a large corpus of biologically produced compounds, using MetaCyc.[12] This database includes 8696 potentially biologically produced compounds. A much smaller proportion is of interest to us, and therefore, we included a feature in BioCompoundML that automatically clusters compounds using a Tanimoto fingerprint similarity criterion, which has been shown to be a powerful functional clustering technique.[17] Clustering training data to the total MetaCyc database allowed us to exclude a large set of compounds that would otherwise stretch the limits of any machine-learned model (see Figure 5).

We used the classifier to investigate the 1977 hydrocarbons and oxygenates in MetaCyc, to identify which compounds had the highest and lowest probability for being high RON compounds. Supplemental Table 1 of the Supporting Information shows 107 compounds that score as having high RON (on the basis of the threshold defined above of $p > 0.897$). It is important to note that the choice of compounds for

DOI: 10.1021/acs.energyfuels.6b01952
Energy Fuels 2016, 30, 8410−8418

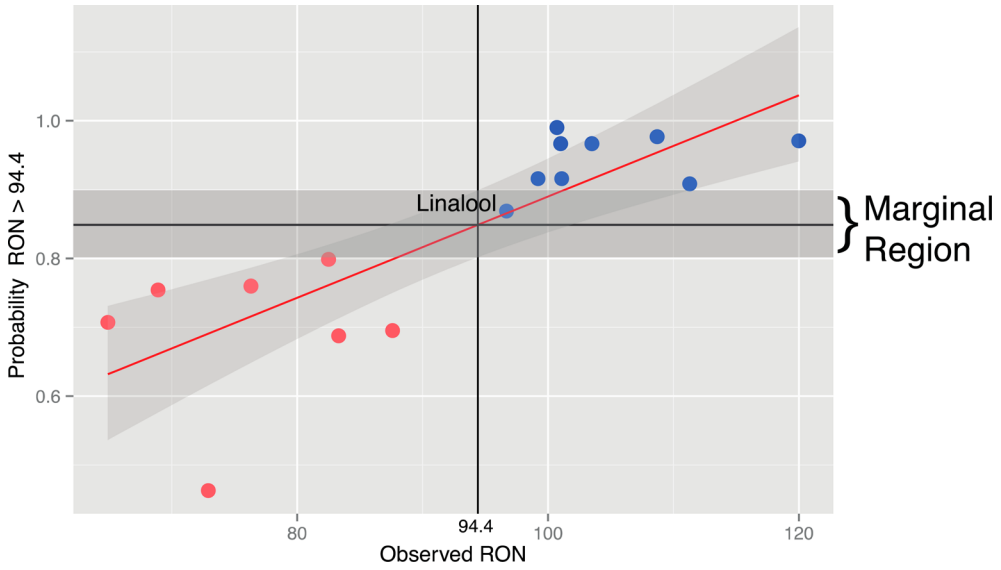

Figure 4. Regression of observed RON by probability in a high RON class (RON > 94.4) for experimentally measured compounds. The 95% confidence intervals are included. On the basis of this regression, the expected threshold for classification in high/low RON is $p = 0.8487$.

Table 2. Experimentally Tested Compounds with Model Predictions

CAS Registry Number	compound	measured RON	prediction	probability in high RON class	accurate
106-21-8	3,7-dimethyl-1-octanol	64.9	not high RON	0.707	yes
13466-78-9	3-carene	68.9	not high RON	0.754	yes
13877-91-3	ocimene	72.9	not high RON	0.463	yes
78-69-3	3,7-dimethyl-3-octanol	76.3	not high RON	0.76	yes
123-35-3	myrcene	82.5	not high RON	0.799	yes
80-56-8	α-pinene	83.3	not high RON	0.63	yes
5989-27-5	(*R*)-(+)-limonene	87.6	not high RON	0.695	yes
78-70-6	linalool	96.7	unclear	0.869	marginal
470-82-6	eucalyptol	99.2	high RON	0.916	yes
142-62-1	butyl acetate	100.7	high RON	0.99	yes
123-92-2	isoamyl acetate	101	high RON	0.967	yes
93-58-3	methyl benzoate	101.1	high RON	0.998	yes
115-18-4	2-methyl-3-buten-2-ol	103.5	high RON	0.967	yes
110-19-0	isobutyl acetate	108.7	high RON	0.977	yes
67-64-1	acetone	111.3	high RON	0.908	yes
209-117-3	isopropyl acetate	>120	high RON	0.971	yes

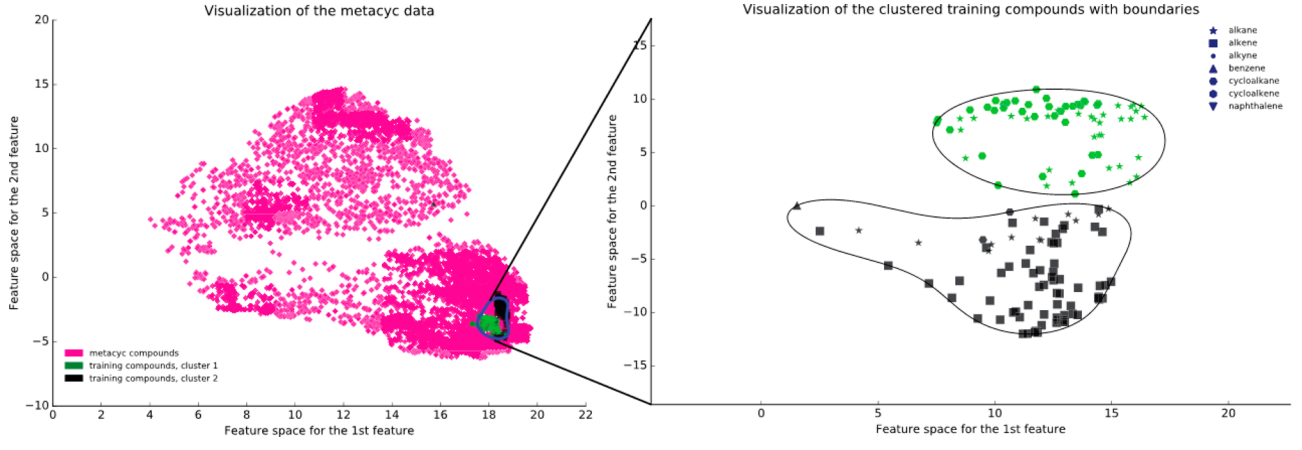

Figure 5. Non-metric multiple dimensional scaling of the first two principal components. RON training data maximally clustered into two classes (using the *k*-means algorithm and maximizing silhouette score). Black- and green-colored regions illustrate boundaries around these two clusters. Only 4.5% ($n = 398$) of 8696 compounds within MetaCyc fall into this category.

DOI: 10.1021/acs.energyfuels.6b01952
Energy Fuels 2016, 30, 8410−8418

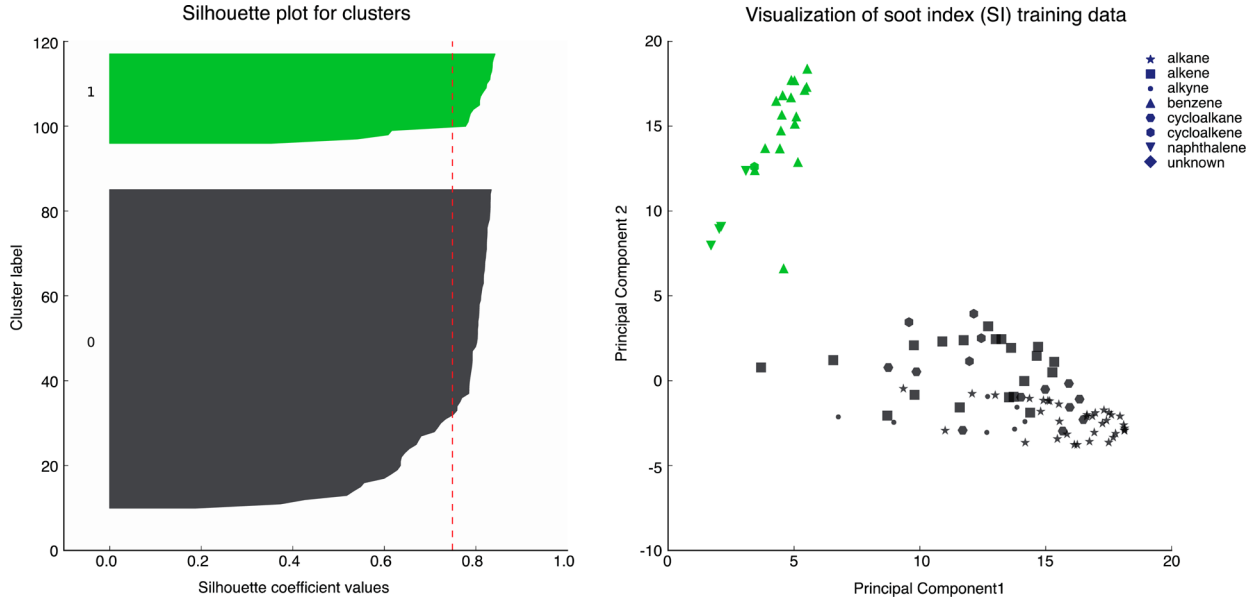

Figure 6. Clustering of TSI training compounds. The left panel shows the cumulative silhouette score for each class, with the red line representing the average silhouette score. The right panel shows the multiple dimensional scaling plot for the first two principle components. Compound types are labeled by point shape.

inclusion into a fuel is complicated, and there are a variety of reasons why a compound would be chosen for inclusion or exclusion from a fuel blend. However, the general purpose of this tool is to rank and classify a large number of compounds based on individual properties. It is the combination of these properties that ultimately inform us about which compounds will be useful in fuels.

TSI Classification. Soot formation is important in many types of combustion for both environmental and engineering reasons. Soot formation can be measured as the smoke point (ASTM D1322 standard, the height in millimeters of a smokeless flame when the fuel is burned in a specific lamp), with lower values indicating a higher sooting tendency. Aviation turbine fuels are limited to a minimum smoke point because high levels of soot cause a higher level of the fuel energy content to be released as thermal radiation, reducing engine durability and efficiency. The TSI was developed to provide a comparison of the soot-forming tendency of different fuels that takes into account differences in air−fuel combustion stoichiometry and different experimental setups. TSI is defined as a value between 0 and 100 for evaluating the onset of soot formation in both premixed and diffusion flames[31] and is defined mathematically by the equation

$$\mathrm{TSI} = a\left(\frac{\text{molecular weight}}{\text{smoke point}}\right) + b$$

where a and b are experimental constants. TSI corresponds to the inverse of the smoke point, scaled by the molecular weight and experimental constants to between 0 and 100. In some cases, SI engines that employ direct injection can produce high levels of soot particles, and TSI is one metric that has been proposed for predicting fuel effects.[32] Our study used the TSI value of a previous study to determine success of our classification technique in predicting this value using 98 compounds.[33] Training data for this feature were found to optimally cluster into two classes (using the k-means algorithm and optimizing silhouette score; see Figure 6). The TSI classification boundary was chosen as the median, which for this data set was 5.9.

Using NCBI fingerprint and NCBI experimental and computed features, Boruta reduced the total number of features from 907 to 16 important features. The majority of selected features correspond to SMILES patterns (10 of 16). These may ultimately correspond to bond availability. Complexity is unsurprisingly the most strongly weighted feature (weight = 0.1639). It has been well-known for decades that increasing complexity in general leads to greater resistance to oxidation and corresponds to increased soot formation.[34,35] Features relating to mass and molecular weight are also heavily weighted, which is unsurprising, given that they are important features in the definition of TSI. XLogP3 is also heavily weighted (weight = 0.0639).

Using 100 random 2-fold splits of the data, we observed high accuracy, precision, and recall. The average accuracy in these 50% leave-out experiments was 87.21%, with a standard deviation of 9.2%. Precision was 90.71%, with a standard deviation of 12.4%. Recall was 82.30%, with standard deviation of 19.3%. ROC AUC was 0.87, with 0.09 standard deviations.

MP Classification. MP is an example of a fuel property in which there is a large body of experimental data from which to build a machine-learning classifier. MP is important in that fuels are required to be in the liquid state for handling and blending. Additionally, MP of a proposed bioblendstock is related to the minimum wintertime operability temperature of the final fuel blend. We set the classification criterion at 20 °C. This temperature corresponds to identification of compounds that are liquid at room temperature and is the maximum allowable boundary for a fuel blendstock. To build our classifier, we used the Jean-Claude Bradley Open Melting Point Dataset.[36] After the removal of redundancies and low-quality measurements, this data set provided 14 869 individual MP measurements.

To classify MP, we adopted two distinct approaches: we classified using exclusively NCBI fingerprints and separately using NCBI fingerprints and a small number of "computed properties". These computed properties include 20 basic

DOI: 10.1021/acs.energyfuels.6b01952
Energy Fuels 2016, 30, 8410−8418

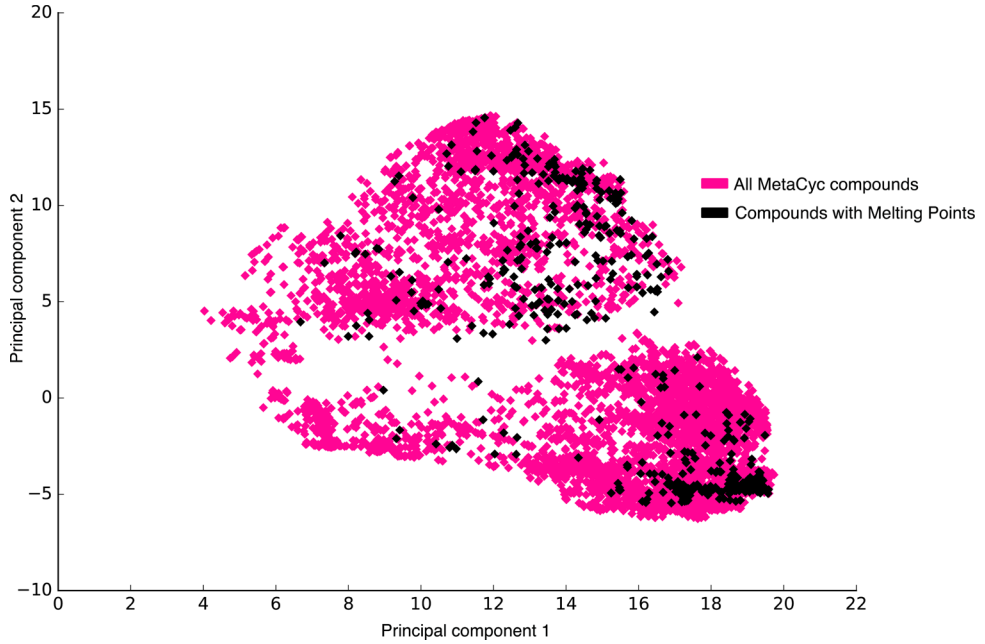

Figure 7. MetaCyc compounds with MP compounds mapped. Data are visualized using non-metric multiple dimensional scaling of the first two principal components. Compounds with available MPs are plotted in black.

chemical properties: density, hydrogen-bond donor count, rotatable bond count, XLogP3, formal charge, undefined atom stereocenter count, molecular weight, hydrogen-bond acceptor count, XLogP3-AA, LogP, defined atom stereocenter count, complexity, covalently bonded unit count, isotope atom count, undefined bond stereocenter count, heavy atom count, exact mass, monoisotopic mass, topological polar surface area, and defined bond stereocenter count. Notably, this custom feature list excludes boiling point, MP, autoignition, vapor density, flash point, and vapor pressure, which were likely to dominate feature weights.

Running Boruta feature selection on the fingerprint-only data set reduced the number of features from 881 to 288. This data set resulted in a large number of features of very small classifying power. The five most heavily weighted features include the SMARTS patterns: C(~H)(~H)(~H) (weight = 0.0259), ≥2 O (weight = 0.021), ≥1 O (weight = 0.0205), ≥1 any ring size 6 (weight = 0.0201), and ≥2 N (weight = 0.0178). The case of many features of limited discriminating power can lead to a problem of overfitting. In this case, having a wide breadth of data can be a huge benefit. Figure 7 maps MP data relative to MetaCyc. In contrast to RON, this map covers a huge span of likely data. The average accuracy using 100 random 50% leave-out experiments was 92.3% (±0.5%); average precision was 94.7% (±0.6%); average recall was 96.2% (±0.5%); and average ROC AUC was 0.8438 (±0.01).

Using fingerprint plus computed properties let Boruta reduce the number of features from 901 to 194. The most heavily weighted features primarily include structural fingerprints but weighs the selected features much more heavily than fingerprint only. The five most heavily weighted features include complexity (weight = 0.0903), computed property topological polar surface area (weight = 0.0670), molecular weight (weight = 0.0582), exact mass (weight = 0.0578), and monoisotopic mass (weight = 0.0556). For this classifier, the average accuracy using 100 random 50% leave-out experiments was 93.7% (±0.5%), the average precision was 95.4% (±0.6%), the average recall was 97.1% (±0.4%), and the average ROC AUC was 0.8662 (±0.01).

DISCUSSION

In this paper, we present a general-purpose fuel property characterization tool. This tool uses a variety of machine-learning techniques to perform automatic random forest classification of compounds. BiocompoundML can be adapted to rapidly screen compounds for any desired compound property. We used this technique on a variety of compounds to classify three key fuel properties (i.e., RON, TSI, and MP). We found that this technique provided accurate and precise classification of compounds into high and low classes. We also found that it accurately classified high RON compounds and, when it failed to accurately classify low RON compounds, most errors were marginal. We further ran this tool on a large set of biologically producible hydrocarbons and identified numerous strong predictions of compounds with high and low RONs. This tool proved accurate and precise in predicting TSI class. For data sets with a large number training set (MP), BioCompoundML proved accurate and precise, even with exclusively structural data.

This tool is fully open-source and is usable by a broad range of fuel researchers and industry. The use of this tool is not restricted to the three properties chosen, and it is intended that, in the future, it will find applicability to a wide number of chemical properties.

Name of tool, BioCompoundML; tool home page, http://www.github.com/sandialabs/BioCompoundML; operating system, Linux/Unix; programming language, Python; compatible versions, Python 2.6, 2.7, 3.3, and 3.4; license, BSD-2 clause; and documentation, http://sandialabs.github.io/BioCompoundML/.

DOI: 10.1021/acs.energyfuels.6b01952
Energy Fuels 2016, 30, 8410−8418

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.energyfuels.6b01952.

MetaCyc hydrocarbons with score above $p > 0.897$ threshold (Supplemental Table 1) (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: cmhudso@sandia.gov.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This research was conducted as part of the Co-Optimization of Fuels & Engines (Co-Optima) Project sponsored by the Bioenergy Technologies and Vehicle Technologies Offices, Office of Energy Efficiency and Renewable Energy (EERE), U.S. Department of Energy (DOE). Co-Optima is a collaborative project of multiple national laboratories initiated to simultaneously accelerate the introduction of affordable, scalable, and sustainable biofuels and high-efficiency, low-emission vehicle engines. Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the National Nuclear Security Administration, DOE, under Contract DE-AC04-94AL85000. This work was part of the DOE Joint BioEnergy Institute (http://www.jbei.org) supported by the Office of Biological and Environmental Research, Office of Science, DOE, through Contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the DOE. This work was also supported by the Vehicle Technologies Office, DOE, under Contract DE347AC36-99GO10337 with the National Renewable Energy Laboratory. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript or allow others to do so for United States Government purposes.

■ NOMENCLATURE

NCBI = National Center for Biotechnology Information
RON = research octane number
TSI = threshold soot index
MP = melting point
QSPR = quantitative structure−property relationship
QSAR = quantitative structure−activity relationship
CID = compound identifier
CAS = Chemical Abstracts Service
SDF = substance data file
k-NN = *k*-nearest neighbor
SVD = singular value decomposition
ROC = receiver operator characteristic
AUC = area under the curve
SMILES = simplified molecular-input line-entry system

■ REFERENCES

(1) Keserű, G. M.; Soós, T.; Kappe, C. O. Anthropogenic reaction parameters−the missing link between chemical intuition and the available chemical space. *Chem. Soc. Rev.* **2014**, *43* (15), 5387−5399.

(2) Gani, R. Chemical product design: Challenges and opportunities. *Comput. Chem. Eng.* **2004**, *28* (12), 2441−2457.

(3) Do, P. T. M.; Crossley, S.; Santikunaporn, M.; Resasco, D. E. Catalytic strategies for improving specific fuel properties. *Catalysis* **2007**, *20*, 33−64.

(4) Yee, L. C.; Wei, Y. C. Current modeling methods used in QSAR/QSPR. *Statistical Modelling of Molecular Descriptors in QSAR/QSPR* **2012**, *2*, 1−31.

(5) Saldana, D. A.; Starck, L.; Mougin, P.; Rousseau, B.; Pidol, L.; Jeuland, N.; Creton, B. Flash point and cetane number predictions for fuel compounds using quantitative structure property relationship (QSPR) methods. *Energy Fuels* **2011**, *25* (9), 3900−3908.

(6) Smolenskii, E. A.; Ryzhov, A. N.; Bavykin, V. M.; Myshenkova, T. N.; Lapidus, A. L. Octane numbers (ONs) of hydrocarbons: A QSPR study using optimal topological indices for the topological equivalents of the ONs. *Russ. Chem. Bull.* **2007**, *56* (9), 1681−1693.

(7) Yang, Y.; Boehman, A. L.; Santoro, R. J. A study of jet fuel sooting tendency using the threshold sooting index (TSI) model. *Combust. Flame* **2007**, *149* (1), 191−205.

(8) Saldana, D. A.; Starck, L.; Mougin, P.; Rousseau, B.; Creton, B. On the rational formulation of alternative fuels: Melting point and net heat of combustion predictions for fuel compounds using machine learning methods. *SAR and QSAR in Environmental Research* **2013**, *24* (4), 259−277.

(9) Dahmen, M.; Marquardt, W. Model-Based Design of Tailor-Made Biofuels. *Energy Fuels* **2016**, *30* (2), 1109−1134.

(10) Kursa, M. B.; Rudnicki, W. R. Feature selection with the Boruta package. *Journal of Statistical Software* **2010**, *36* (11), 1−13.

(11) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem substance and compound databases. *Nucleic Acids Res.* **2016**, *44* (D1), D1202−D1213.

(12) Caspi, R.; Altman, T.; Billington, R.; Dreher, K.; Foerster, H.; Fulcher, C. A.; Holland, T. A.; Keseler, I. M.; Kothari, A.; Kubo, A.; Krummenacker, M.; Latendresse, M.; Mueller, L. A.; Ong, Q.; Paley, S.; Subhraveti, P.; Weaver, D. S.; Weerasinghe, D.; Zhang, P.; Karp, P. D. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.* **2014**, *42* (D1), D459−D471.

(13) Murrell, D. S.; Cortes-Ciriano, I.; van Westen, G.; Stott, I. P.; Bender, A.; Malliavin, T. E.; Glen, R. C. Chemically Aware Model Builder (camb): An R package for property and bioactivity modelling of small molecules. *J. Cheminf.* **2015**, *7* (1), 45.

(14) Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A. K.; Rupp, M.; Teetz, W.; Brandmaier, S.; Abdelaziz, A.; Prokopenko, V. V.; Tanchuk, V. Y.; Todeschini, R.; Varnek, A.; Marcou, G.; Ertl, P.; Potemkin, V.; Grishina, M.; Gasteiger, J.; Schwab, C.; Baskin, I. I.; Palyulin, V. A.; Radchenko, E. V.; Welsh, W. J.; Kholodovych, V.; Chekmarev, D.; Cherkasov, A.; Aires-de-Sousa, J.; Zhang, Q.-Y.; Bender, A.; Nigsch, F.; Patiny, L.; Williams, A.; Tkachenko, V.; Tetko, I. V. Online chemical modeling environment (OCHEM): Web platform for data storage, model development and publishing of chemical information. *J. Comput.-Aided Mol. Des.* **2011**, *25* (6), 533−554.

(15) Yap, C. W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* **2011**, *32* (7), 1466−1474.

(16) Troyanskaya, O.; Cantor, M.; Sherlock, G.; Brown, P.; Hastie, T.; Tibshirani, R.; Botstein, D.; Altman, R. B. Missing value estimation methods for DNA microarrays. *Bioinformatics* **2001**, *17* (6), 520−525.

(17) Baldi, P.; Nasr, R. When is chemical similarity significant? The statistical distribution of chemical similarity scores and its extreme values. *J. Chem. Inf. Model.* **2010**, *50* (7), 1205−1222.

(18) Batista, G. E.; Monard, M. C. An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence* **2003**, *17* (5−6), 519−533.

(19) Breiman, L. Random forests. *Machine learning* **2001**, *45* (1), 5−32.

(20) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random forest: A classification and regression tool for

DOI: 10.1021/acs.energyfuels.6b01952
Energy Fuels 2016, 30, 8410−8418

compound classification and QSAR modeling. *Journal of Chemical Information and Computer Sciences* **2003**, *43* (6), 1947−1958.
(21) Biau, G.; Devroye, L.; Lugosi, G. Consistency of random forests and other averaging classifiers. *J. Mach. Learn. Res.* **2008**, *9* (Sept), 2015−2033.
(22) Liaw, A.; Wiener, M. Classification and regression by randomForest. *R News* **2002**, *2* (3), 18−22.
(23) Moosmann, F.; Nowak, E.; Jurie, F. Randomized clustering forests for image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2008**, *30* (9), 1632−1646.
(24) Halko, N.; Martinsson, P.-G.; Tropp, J. A. *Finding Structure with Randomness: Stochastic Algorithms for Constructing Approximate Matrix Decompositions*; Applied & Computational Mathematics, California Institute of Technology: Pasadena, CA, Sept 2009; Technical Report 2009-05.
(25) Leone, T. G.; Anderson, J. E.; Davis, R. S.; Iqbal, A.; Reese, R. A.; Shelby, M. H.; Studzinski, W. M. The effect of compression ratio, fuel octane rating, and ethanol content on spark-ignition engine efficiency. *Environ. Sci. Technol.* **2015**, *49* (18), 10778−10789.
(26) Blurock, E. S. Automatic learning of chemical concepts: Research octane number and molecular substructures. *Comput. Chem.* **1995**, *19* (2), 91−99.
(27) Balaban, A. T.; Kier, L. B.; Joshi, N. Structure-property analysis of octane numbers for hydrocarbons (alkanes, cycloalkanes, alkenes). *MATCH* **1992**, *28*, 13−27.
(28) American Society for Testing Materials; American Petroleum Institute. *Knocking Characterstics of Pure Hydrocarbons: Developed under American Petroleum Institute Research Project 45*; American Society for Testing Materials: Philadelphia, PA, 1958; ASTM Special Technical Publication 225.
(29) Al-Fahemi, J. H.; Albis, N. A.; Gad, E. A. M. QSPR models for octane number prediction. *J. Theor. Chem.* **2014**, *2014*, 1.
(30) Yanowitz, J.; Ratcliff, M. A.; McCormick, R. L.; Taylor, J. D.; Murphy, M. J. *Compendium of Experimental Cetane Numbers*; National Renewable Energy Laboratory (NREL): Golden, CO, 2014; Technical Report NREL/TP-5400-61693, DOI: 10.2172/1150177.
(31) Calcote, H. F.; Manos, D. M. Effect of molecular structure on incipient soot formation. *Combust. Flame* **1983**, *49* (1), 289−304.
(32) Barrientos, E. J.; Anderson, J. E.; Maricq, M. M.; Boehman, A. L. Particulate matter indices using fuel smoke point for vehicle emissions with gasoline, ethanol blends, and butanol blends. *Combust. Flame* **2016**, *167*, 308−319.
(33) Olson, D. B.; Pickens, J. C.; Gill, R. J. The effects of molecular structure on soot formation II. Diffusion flames. *Combust. Flame* **1985**, *62* (1), 43−60.
(34) Schalla, R. L.; McDonald, G. E. Mechanism of smoke formation in diffusion flames. *Symp. (Int.) Combust., [Proc.]* **1955**, *5*, 316−324.
(35) Hunt, R. A. Relation of smoke point to molecular structure. *Ind. Eng. Chem.* **1953**, *45* (3), 602−606.
(36) Bradley, J.-C.; Williams, A.; Lang, A. *Jean-Claude Bradley Open Melting Point Dataset*, 2014; DOI: 10.6084/m9.figshare.1031637.

DOI: 10.1021/acs.energyfuels.6b01952
Energy Fuels 2016, 30, 8410−8418