Received ; | Revised ; | Accepted

DOI: xxx/xxxx

SAND2020-7171C

ARTICLE TYPE

Evaluating causal-based feature selection for fuel property prediction models

Bernard Nguyen*[1] | Leanne S. Whitmore[1,2] | Anthe George[1] | Corey M. Hudson[1]

[1]Sandia National Laboratories, Livermore, CA, 94551
[2]University of Washington, Seattle, WA, 98115

Abstract

In-silico screening of novel biofuel molecules based on chemical and fuel properties is a critical first step in the biofuel evaluation process due to the significant volumes of samples required for experimental testing, the destructive nature of engine tests, and the costs associated with bench-scale synthesis of novel fuels. Machine learning structure-property models are limited by training sets of few existing measurements and high dimensionality, leading to poor predictive performance. Feature selection has been shown to improve machine learning models, but correlation-based feature selection fails to provide scientific insight into the underlying mechanisms that determine structure-property relationships. In this study, we investigate the impact that causal-based feature selection might have on improving model predictions while also helping to identify key molecular substructures for optimizing chemical and fuel properties. We found that causal-based feature selection performed better than alternative filtration methods for predicting one of three fuel properties, and that a structural causal model provides valuable scientific insights into the relationships between molecular substructures, chemical properties, and fuel properties.

KEYWORDS:
causal inference, Bayesian networks, QSPR, feature selection

1 | INTRODUCTION

Evaluating candidates for commercial biofuels can often be an expensive process that requires significant bench-scale chemical synthesis and destructive testing of samples. Ultimately, promising biofuel candidates may be ruled out due to poor fuel property performance, environmental toxicity, or the inability to cost-effectively scale up production. For these reasons, *in-silico* screening of molecules is an essential first step to save both time and money. Chemical kinetic models and quantum chemical simulations are based on known physical principles and provide accurate predictions for some fuel properties, but are computationally expensive, often taking thousands of computation hours to generate predictions for a single molecule. Machine Learning (ML) models can quickly generate predictions, but are limited by existing measurements, with numerous structural properties for each compound, resulting in short, high-dimensional datasets. Training ML models on these datasets often results in inaccurate predictions, however with tempered expectations, predictions generated by models with reduced errors can still be leveraged for *in-silico* screening.

While there are many fuel properties that are important to evaluate novel biofuel candidates, properties related to autoignition tendencies are often the most challenging to predict due to autoignition dependencies on complex combustion kinetics. Research Octane Number (RON) and Motor Octane Number (MON) are two autoignition properties developed in the early 20^{th} century

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

that are still used today to evaluate the antiknock performance of a fuel mixture. Automobile drivers are likely familiar with the Anti-Knock Index (AKI), often described as the octane rating, which is displayed on the pump at fueling stations and is determined by

$$\mathrm{AKI} = \frac{\mathrm{RON} + \mathrm{MON}}{2}. \tag{1}$$

However, many recent studies [1] have determined that modern spark ignition (SI) engines are better served by an antiknock rating called Octane Index (OI), given by

$$\mathrm{OI} = \mathrm{RON} - K * (\mathrm{RON} - \mathrm{MON}), \tag{2}$$

where K refers to an engine-specific constant that depends on the pressure and temperature of unburned fuel in the cylinder. The octane index utilizes K to map RON and MON values to engines that operate at conditions not accurately described by the AKI. These studies show that for many modern engine architectures with negative K values, high Octane Sensitivity (OS), given by

$$\mathrm{OS} = \mathrm{RON} - \mathrm{MON}, \tag{3}$$

is an important feature to avoid engine knock. The OI still requires the determination of both the RON and MON of the fuel according to ASTM methods, which generally requires more than a liter for each fuel candidate. In this context, it is clear that RON and OS are important fuel properties to be able to predict *in-silico*.

BiocompoundML [20] is a suite of binary classification models developed by Sandia National Laboratories for predicting fuel properties such as RON, MON, and OS. These models are capable of accurately predicting whether a molecule will have high/low RON, MON, or OS, but fail to provide concrete values necessary for confidently screening molecules. BiocompoundML is trained on a dataset that encodes molecular structure into an mxf binary matrix, **D**, such that D_{ij} represents the presence or absence of feature j in molecule i. Due to the infinite number of possible permutations that atoms can be arranged, this dataset is extremely sparse and contains thousands of molecular signatures and fingerprints. As shown in **FIGURE 1** these molecular descriptors range in specificity from any single carbon bond to very specific substructures.

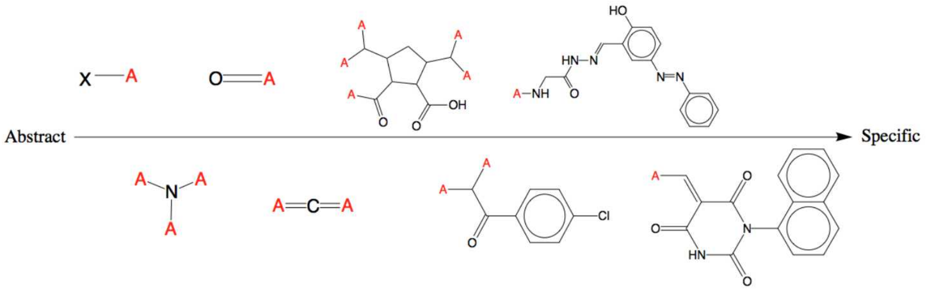

FIGURE 1 Examples of molecular descriptors. **A** refers to any aliphatic (non-aromatic) carbon and **X** refers to any connection.

Feature selection has been shown to improve the accuracy, speed, and robustness of machine learning models [3], and feature selection methods generally fall into one of three categories – filters, wrappers, and embedded methods. Filters select the most appropriate features prior to training of the models, while wrappers involve iterative feature subset selection and model training until the model error approaches a minimum. An example of an embedded method is the multilayer perceptron, where feature selection is built into the ML model architecture itself. Because filter methods are applied prior to training, they are generally less computationally expensive than either wrappers or embedded feature selection methods, and are independent of the predictive model architecture. Additionally, due to the size limitations of our datasets, wrapper and embedded methods would likely result in overfitting, whereas filters could improve the robustness of our models to new data. Even the simplest correlation-based filter methods have been shown to perform comparably to wrapper methods [4]. For these reasons, this study will focus on different filter-based feature selection methods.

Filter-based feature selection generally involves methods of ranking feature importance, which are generally dependent on how correlated a given feature is to the metric being predicted. This provides insight into how certain features may be related but does not provide any insight into the underlying mechanisms that determine *how* the feature is related (i.e. correlation vs causation). Knowing which features *cause* the fuel property to go up or down in addition to which features are associated with the fuel property could potentially improve both the accuracy and the interpretability of ML models. The scientific insights generated from causal discovery in existing data could also drive future fuel selection and design decisions.

In his book, *Causality* [11], Judea Pearl differentiates between statistical association and causation by introducing the concept of causal assumptions. Causal assumptions take advantage of *a priori* knowledge in order to build a structural causal model (SCM) of the relationships between each variable (i.e. which variable is the cause and which is the effect). If *a priori* knowledge about the direction of causal relationships is unknown, it is possible to identify the causal direction by comparing the dependence between the predictive residuals and each variable [16]. The relationship between two variables $x \rightarrow y$ is *causal* if given otherwise identical samples, a change in x *always* results in a quantifiable change in y. The amount that changing x results in a change in y is the causal effect, and the direction of the causal relationship is determined by the causal assumption that $x \nleftarrow y$. One method to visualize a SCM is to generate a Bayesian causal network [10, 19] of treatments (features) and outcomes (fuel properties). For an example of a causal Bayesian network representing the current pandemic, see **FIGURE** 2 , where the nodes represent variables and the edges represent causal relationships. Public health officials are currently very concerned about what causes COVID-19 deaths and how they can be prevented. According to public data from the European CDC [15], COVID-19 deaths are 95% positively correlated with the total number of COVID-19 tests. Nevertheless, public health officials are not advocating for a decrease in testing in order to prevent further death. That is because officials understand that the COVID-19 deaths are actually fairly unrelated to the number of tests, and the relationship is definitely not *directly* causal. In fact, it can be argued that increasing the amount of testing can influence public perception of the virus and therefore change social behaviors that actually *reduce* the spread of COVID-19, a fundamental cause for the prevalence of COVID-19 illness, and as a result, COVID-19 deaths. Instead of advocating for a reduction in testing, health officials recommend measures to reduce the spread of the virus such as social distancing and the use of face masks to reduce aerosol transmission between individuals. Our hypothesis is that by understanding the causal relationships between molecular properties and fuel properties, we could train our models based on features that are directly causal in order to improve both accuracy and robustness.

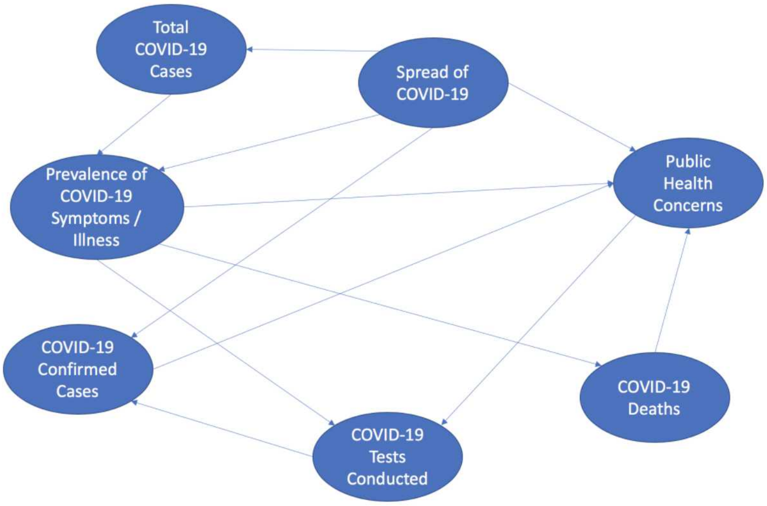

FIGURE 2 Bayesian network example for COVID-19.

2 | METHODS

2.1 | Digital Encoding of Molecular Structure

The dataset, **D**, used in this study consists of information from three major sources such that

$$\mathbf{D} = [\mathbf{P}, \mathbf{M}, \mathbf{C}]. \tag{4}$$

where **P** and **M** refer to molecular descriptors that were generated using the Pharmaceutical Data Exploratory Laboratory (PaDEL-Descriptor) [21] and MolSig Stereo Signature Molecular Descriptor [2] packages. PaDEL checks for the presence or absence of 1875 types of descriptors and 12 types of fingerprints in the molecule, whereas MolSig generates a graph-based representation of the molecule at varying heights. The PaDEL descriptors were encoded into the binary matrix **P**, where the value P_{ij} refers to the presence (1) or absence (0) of the descriptor j in molecule i. The MolSig descriptors were similarly encoded in **M**, but instead of binary representations, M_{ij} refers to the number of occurrences of descriptor j in molecule i. The final component, **C**, refers to chemical properties for each molecule pulled from PubChem [6], such as melting point or rotatable bond count.

2.2 | Methods of Ranking Feature Importance

2.2.1 | Correlation Coefficient

The most obvious way to rank features is to prioritize features that correlate with the fuel property being predicted. The three most common correlation coefficients are the Pearson, Spearman, and Kendall correlation coefficients. The Pearson correlation coefficient $\rho_{X,Y}$ between two variables X and Y is defined as

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}, \tag{5}$$

where $cov(X,Y)$ is the covariance between X and Y, and σ_X, σ_Y, μ_X, and μ_Y are the standard deviations and means of X and Y, respectively. The Spearman correlation coefficient, r_s, differs from the Pearson correlation coefficient in that it compares the *rankings*, rg, between features rather than the linear relationship between the values themselves, and is given by the equation

$$r_s = \rho_{rg_X, rg_Y}. \tag{6}$$

The Spearman correlation is robust to outliers and independent of the linearity between the variables, making it more suitable than the Pearson correlation for ordinal data. The Kendall correlation coefficient, like the Spearman correlation, is also a rank-order correlation method, but rather than taking the Pearson correlation of rankings of X and Y, the Kendall correlation τ is given by

$$\tau = \frac{c-d}{\binom{n}{2}} = \frac{c-d}{\frac{n(n-1)}{2}}, \tag{7}$$

where n is the number of observations and c and d are the number of concordant and discordant pairs between X and Y. Given the ranking pairs (x_i, y_i), (x_j, y_j), ...(x_n, y_n) where $i < j$, a pair is considered to be concordant if both $x_i > x_j$ and $y_i > y_j$ or both $x_i < x_j$ and $y_i < y_j$. The pair is considered to be discordant if either $x_i > x_j$ and $y_i < y_j$ or $x_i < x_j$ and $y_i > y_j$. The Kendall correlation coefficient can also be written as

$$\tau = \frac{2}{n(n-1)} \sum_{i<j} \text{sgn}(x_i - x_j)\text{sgn}(y_i - y_j), \tag{8}$$

where sgn refers to the sign function

$$\text{sgn}(x) = \begin{cases} -1, & if\, x < 0 \\ 0, & if\, x = 0 \\ 1, & if\, x > 0 \end{cases} \tag{9}$$

The Kendall correlation coefficient tends to be more conservative than its Spearman counterpart, making it even more robust to outliers and having better p-values for smaller datasets. Due to the ordinal nature of the PaDEL molecular descriptor data and the limited size of our dataset, we opted to use the Kendall rank-order correlation to measure correlation between descriptors and fuel properties.

2.2.2 | Mutual Information

Mutual information, I, represents the statistical dependence between two variables and is calculated by

$$I_{X,Y} = \sum_y \sum_x p_{X,Y}(x,y) log\left(\frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)}\right), \tag{10}$$

where $p_{X,Y}$ is the joint probability mass function of X and Y, and p_X and p_Y are the marginal probability mass functions of X and Y. Intuitively, the best features for predicting a given fuel property are likely the features that share the most "mutual information" with that property. Because of this, mutual information is one of the most common methods for ranking feature importance in filter-based feature selection [5, 13, 18].

2.2.3 | Gini Impurity

Gini impurity is the most common metric for measuring prediction accuracy in random decision tree models [7], and describes the probability that a sample would be incorrectly labeled if it was randomly labeled according to the distribution of the labels in the training set. In a decision tree, the Gini impurity G_l for a given leaf node l is

$$G_l = 1 - f_p^2 - f_n^2, \tag{11}$$

where f_p and f_n refer to the fraction of positive and negative samples, and a leaf containing only positive or negative samples would have a Gini impurity of 0. The Gini impurity for a given feature in the decision tree is the weighted average of the two leaf nodes which correspond to a feature. The Gini impurity is so common that the popular ML python package, scikit-learn, includes a *feature_importances_* attribute with all random forest models, which is the normalized inverse Gini impurity, such that

$$\sum_f G_f = 1 \tag{12}$$

for all features f. While Gini impurity may be the obvious choice for measuring feature importance in ensemble decision tree models, it is important to note that Gini has also been shown to have a selection bias when comparing variables of different scale or type (discrete vs continuous) [17].

2.2.4 | LIME

The open-source Local Interpretable Model-agnostic Explainer (LIME) [14] package is an explainable AI tool that shows which features are most important for decisions made by ML models. LIME achieves this by perturbing an input variable around its local neighborhood and measuring how the model predictions differ. Input variables are weighted based on how much changing that variable causes the model to change its predictions. It is important to note that while LIME is a valuable tool for explaining how decisions are made by black-box ML models, LIME feature weights are especially dependent on limitations in the completeness of the training set.

2.2.5 | Causality

Our hypothesis is that knowing the causal effect between features and fuel properties could potentially improve feature selection. Ideally, the causal effect would be measured through counterfactual examples in the data, where counterfactuals [9, 12, 19] are data points in which only the variable in question is changed. By controlling only one variable at a time, we could determine the causal effect that changing that variable has on various fuel properties. While this may be possible when designing an experiment (e.g. randomized, double-blind drug trials), real-world datasets rarely contain sufficient counterfactuals for every feature, and that is especially true in the case of datasets containing molecular signatures, where almost every feature shares a dependence with another. Instead, we could estimate causal effects using conditional Bayesian statistics. To do this, we first constructed the causal Bayesian network containing features, fuel properties, and the potential causal relationships between them. We identified potential causal relationships between features based off correlation, then determined the direction of those features based on *a priori* knowledge. For example, molecular substructure may have a causal impact on melting point, but not the other way around. Similarly, adding another benzene ring may increase total carbon count, but adding carbon to a molecule does not necessarily require a benzene ring to be present. In general, we determined that structural features could *cause* chemical properties, and that both structural features and chemical properties could *cause* fuel properties.

Once the causal relationships were mapped out in our Bayesian network, we could estimate the causal effect of features on fuel properties. Because we lacked sufficient data to estimate causal effect using counterfactuals, we instead relied on an alternative method - propensity score stratification. Samples were separated into strata based on the probability (propensity) that the feature of interest will have the value that it does based on the common causes between that feature and the fuel property. Existing causal inference packages such as Microsoft's DoWhy [8] include methods to estimate causal effects via propensity score stratification, but assume binary treatments and outcomes. We developed a method similar to the one used in DoWhy for propensity score stratification, but used linear regression to enable estimation of causal effects due to continuous variables. Any continuous variables in the dataset were discretized into bins, and the propensity score for each data point was calculated using multi-class logistic regression. Samples were then stratified based on k-means clustering of the probability that each data point will have a given value for a particular feature based on the common causes between the feature and fuel property. In order to ensure relative confidence that the causal effect was actually present in the data, strata with less than 5 data points were removed, as well as strata with only one unique value for the feature in question. Removal of strata with only one unique value was also necessary in order to perform linear regression, which was done on the remaining strata using only that feature as the input. The causal effect and weight for strata were recorded as the coefficient of regression (slope) and the $max(0, \mathrm{R}^2)$ between the predicted and actual values. Using the R^2 value as the weight when estimating the weighted average of the overall causal effect ensured that causal effects from low-confidence strata had minimal impact on the overall estimate. In other words, the overall causal effect C_f that a feature has on the fuel property was estimated to be

$$C_f = \frac{\sum_S n_s \cdot c_s \cdot \mathrm{R}_s^2}{\sum_S n_s \cdot \mathrm{R}_s^2}, \tag{13}$$

where S is the set of strata, n_s is the number of points in each strata, c_s is the causal effect of that strata, and R_s^2 is the $max(0, \mathrm{R}^2)$ for that strata. Estimating the causal effect for each strata helps to eliminate confounding dependence between the feature, the fuel property, and their shared causes, emphasizing the dependence that the fuel property has on that particular feature.

2.3 | Quantitative Evaluation of Various Ranking Methods

In order to compare across different methods of ranking feature importance, we normalized the feature importance as the relative weight for each ranking. The feature weight, w_f, for any given ranking method, *r*, is given by

$$w_{f_r} = \frac{|i_{f_r}|}{\sum_f |i_{f_r}|},$$

where i_{f_r} is the feature importance as defined by ranking method, *r*, and $\sum_F w_f = 1$. Once the feature weights were determined, we tested the top features listed by each ranking using a fairly standard random forest regressor (n_estimators=100,max_depth=5). In addition to the ranking methods mentioned above, we also included the results from a random feature selection method, as well as the mean of weights between all ranking methods. We trained the model on a subset of features starting with only the *best* feature, then adding features one-by-one with each iteration, measuring the root-mean-squared-error (RMSE) along the way. The expectation was that the most important features being included would result in the lowest RMSE at first, and that the RMSE would quickly decrease as more and more features were included in the training.

Rather than a traditional splitting of training and test sets, we used 10-fold cross validation for this process. Because each sample in the dataset is a unique molecule, removal of any given molecule could have removed key information regarding one of the features, resulting in significantly higher prediction error. The use of 10-fold cross validation over distinct training and test sets ensured that any negative impact from removing a key sample was averaged across the other 9 folds that included that molecule.

Another experiment we conducted to evaluate ranking methods was to remove the top 10 features identified by each ranking method and to repeat the RMSE experiment for the top 50 features that remained. If a ranking method *correctly* identified a feature as being important for accurate prediction, the RMSE curve should be higher than its counterpart with no features removed. On the other hand, if the ranking method *incorrectly* identified a feature as being important, removing that feature should have minimal effect on the RMSE curve.

3 | RESULTS

Due to some ranking methods attributing more features as having zero weight than others, different ranking methods could only be compared *up to* the minimum number of feature inclusions. **FIGURE** 3 shows the RMSE results for the top 80 features for RON, MON, and OS. While all ranking methods showed significant improvement over random feature inclusion, the causality-based feature selection method performed relatively poorly for all fuel properties.

Another test to see if each ranking method successfully identified the most important features was to remove the top 10 features of each ranking method and compare the models' error with the control where no features were removed. **FIGURE** 4 shows the RMSE results of this experiment and compares the RMSE curves for when the top 10 features were removed according to

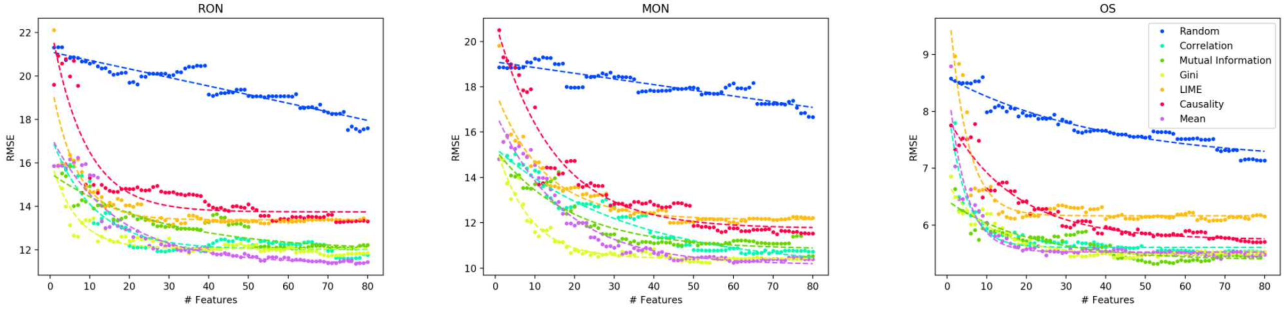

FIGURE 3 Comparison of ranking methods for RON, MON, and OS.

each ranking method. The order of feature importance for this experiment was the mean of the feature weights for each ranking method. While causality-based feature selection still did not show improvement in the RON and MON models, it did perform well for predicting OS. It is also worth noting that while Gini impurity performed the best in the previous experiment, methods based on correlation and mutual information appeared to perform the best in this experiment.

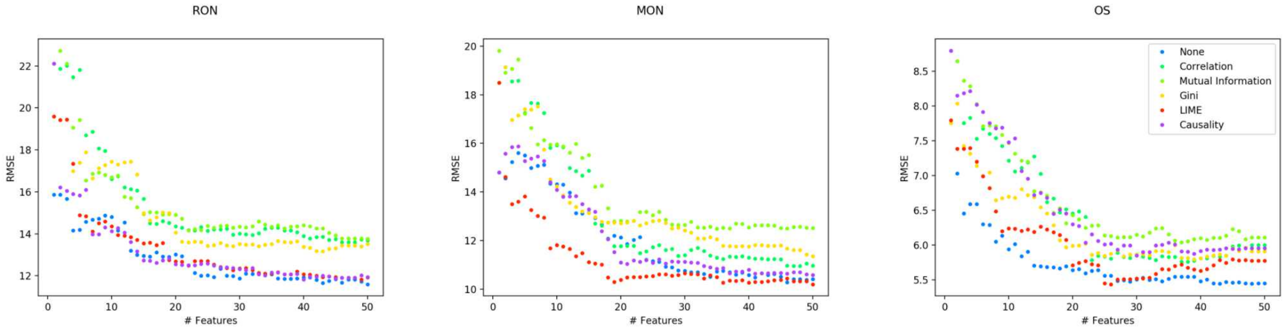

FIGURE 4 RMSE with top 10 features removed from each ranking method.

As expected, the best ranking method for each fuel property was different, reflecting the common sentiment that there is no "best" feature selection method for all datasets. However, feature selection in general remains an important preprocessing step for machine learning models based on sparse and limited datasets. **FIGURE** 5 and 6 show the difference between models trained on all features and models trained on only the top 80% cumulative feature weights for each fuel property, which ended up being 605, 603, and 606 features for RON, MON, and OS respectively. **TABLE** 1 shows the cross validation results (R^2 and RMSE values) for the models based on the amount of feature inclusion, but it is important to note that **FIGURE** 5 and 6 show predictions where the training and test set are the same. Samples in both **FIGURE** 5 and 6 are colored from green to red such that the red samples represent errors outside of the acceptable range of ±10 for RON and MON and ±5 for OS.

4 | DISCUSSION

Although causal-based feature ranking performed much better than random feature selection, it generally performed worse than other common methods for identifying the most important features for fuel property predictions. The exception was that causal-based feature selection seemed to perform the best for identifying the top features for predicting OS according to the results of the second experiment. Gini importance appeared to perform well in the first experiment, but other methods quickly

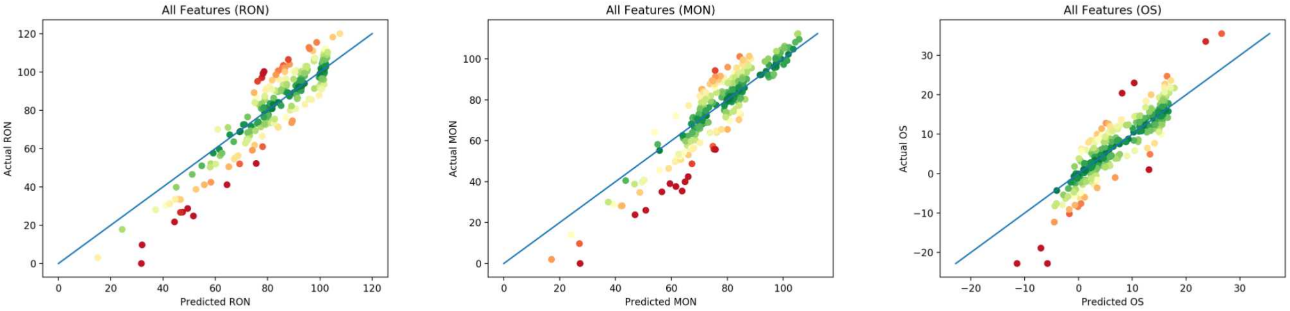

FIGURE 5 Actual vs predicted values for RON, MON, and OS with all features included.

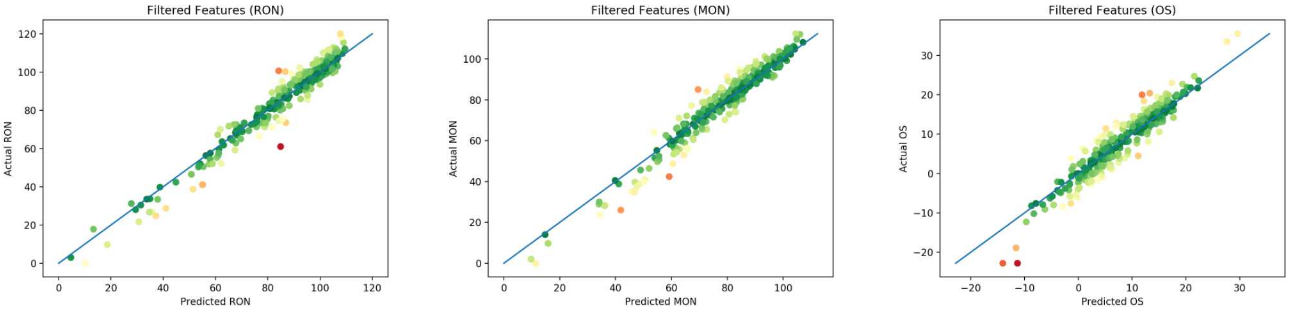

FIGURE 6 Actual vs predicted values for RON, MON, and OS with filtered features.

TABLE 1 R^2 and RMSE for models based on feature inclusion

	RON		**MON**		**OS**	
Features included	$\mathbf{R^2}$	**RMSE**	$\mathbf{R^2}$	**RMSE**	$\mathbf{R^2}$	**RMSE**
All	0.555	13.849	0.544	12.451	0.525	5.723
Top 80% Weight	0.669	11.491	0.666	10.535	0.550	5.556

caught up as the number of features increased. Gini importance also did not perform as consistently in the second experiment, indicating that features identified as important by Gini may have contained redundant information, and removing those top 10 features did not have a large impact on predictive performance.

In order to compare between features and fuel properties of different scales, the dataset was normalized prior to estimating causal effects. **FIGURE** 7 shows examples of some of the molecular features with the greatest impact on fuel properties, as well as examples of those substructures in the dataset. It is clear that features identified to have a causal effect on fuel properties most consistently identified changes in OS, as expected from the results of the second experiment. While causal-based feature selection may not have improved feature selection for all fuel properties, it performed on par with other feature selection methods as the number of included features increased. Causal-based feature selection and the associated construction of SCMs also have the additional benefit of increased understanding behind the underlying mechanisms that influence fuel properties, as well as a simpler understanding of how different features affect each other and the fuel properties. **FIGURE** 8 shows the relationships between features according to correlation (with a threshold of $|\pm 0.5|$) on the left vs causal relationships between each feature and the fuel properties on the right. The correlation map shows that 612 features share at least one "significant" correlation with

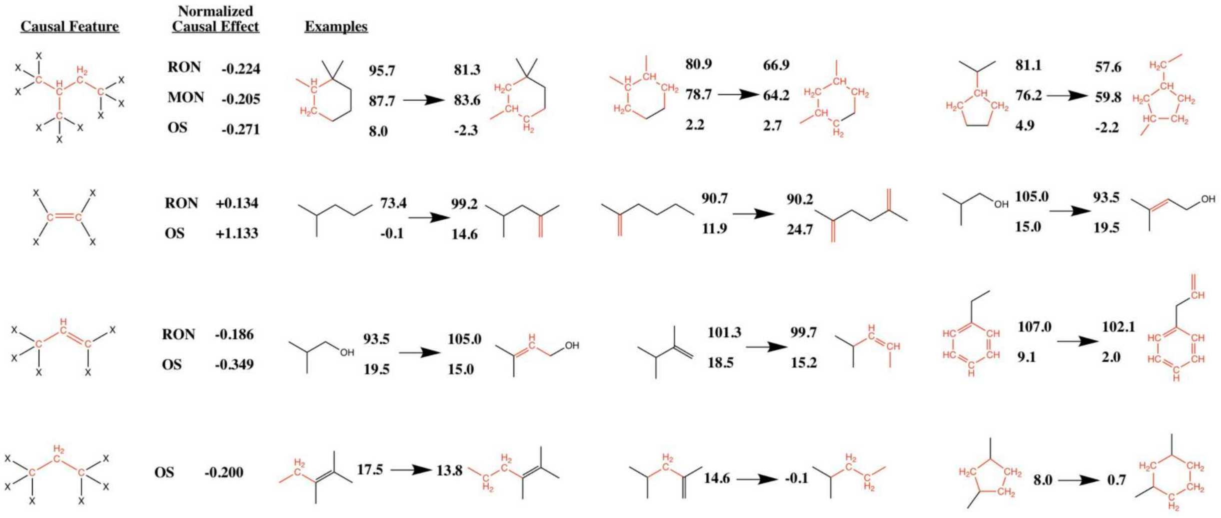

FIGURE 7 Examples of causal features and their effect on fuel properties.

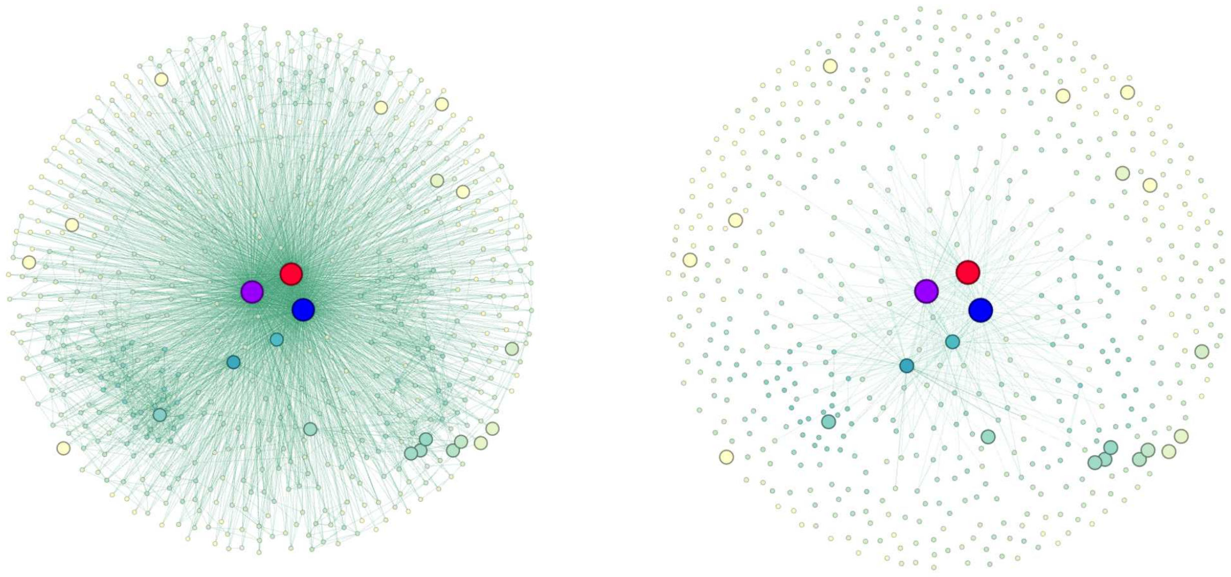

FIGURE 8 Correlative vs causal feature relationship map.

another feature, but the causal map indicates that only 214 of those nodes share a *causal* relationship with another feature. The causal map also reduces the number of potential relationships by roughly 90% from 3465 edges to 349. The simpler structural causal model allows us to construct a Bayesian causal network to visualize the role that each feature plays in determining fuel properties and develop a better understanding of how to design fuels with improved fuel properties in the future.

Because the two experiments to evaluate feature ranking methods did not yield consistent results, we have decided to use the mean feature weight for all ranking methods for feature selection of RON, MON, and OS models. Because this approach is an aggregate of several different ranking methods, it is robust to different datasets regardless of sparsity or size. **FIGURE** 5 and **FIGURE** 6 clearly show the importance of feature selection for sparse and limited datasets, and that the mean feature weight between several ranking methods is sufficient for accurate model predictions. Note that the predicted fuel properties in **FIGURE** 5 shows the tendency for such datasets to predict values closer to the mean when redundant and irrelevant features are included. Intelligent feature selection enables us to generate robust models that neither overfit nor underfit the available data, such as the ones used in **FIGURE** 6. While feature selection may not be as important for large and robust datasets with limited dimensionality, it is certainly important in many applications with limited data which are common across scientific research. Causal-based feature selection may not be the best ranking method for every application, but it is a useful tool which helps to pull back the curtain on the underlying mechanisms that a black-box model cannot otherwise identify.

5 | ACKNOWLEDGEMENTS

This research was conducted as part of the Co-Optimization of Fuels & Engines (Co-Optima) project sponsored by the U.S. Department of Energy (DOE) Office of Energy Efficiency and Renewable Energy (EERE), Bioenergy Technologies and Vehicle Technologies Office.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Security Administration under contract DE-NA0003525.

References

[1] Abdul-Manan, A. F. N., G. Kalghatgi, and H. Babiker, 2018: Exploring alternative octane specification methods for improved gasoline knock resistance in spark-ignition engines. *Frontiers in Mechanical Engineering*, **4**, doi:10.3389/fmech.2018.00020.

[2] Carbonell, P., L. Carlsson, and J.-L. Faulon, 2013: Stereo signature molecular descriptor. *Journal of Chemical Information and Modeling*, **53**, no. 4, 887–897, doi:10.1021/ci300584r.

[3] Chandrashekar, G. and F. Sahin, 2014: A survey on feature selection methods. *Computers Electrical Engineering*, **40**, no. 1, 16–28, doi:10.1016/j.compeleceng.2013.11.024.

[4] Hall, M. A. and L. A. Smith, 1999: Feature selection for machine learning: Comparing a correlation-based filter approach to the wrapper. *University of Waikato - Department of Computer Science*.
URL https://www.aaai.org/Library/FLAIRS/1999/flairs99-042.php

[5] Huang, J. and P. Rong, 2009: A hybrid genetic algorithm for feature selection based on mutual information. *Information Theory and Statistical Learning*, 125–152, doi:10.1007/978-0-387-84816-7_6.

[6] Kim, Sunghwan, Thiessen, P. A., Bolton, E. E., Chen, Jie, Gindulyte, Jane, and et al., 2015: *Pubchem substance and compound databases*.
URL https://academic.oup.com/nar/article/44/D1/D1202/2503131

[7] Menze, B. H., B. M. Kelm, R. Masuch, U. Himmelreich, P. Bachert, W. Petrich, and F. A. Hamprecht, 2009: A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics*, **10**, no. 1, 213, doi:10.1186/1471-2105-10-213.

[8] Microsoft, 2020: *microsoft/dowhy*.
URL https://github.com/microsoft/dowhy

[9] Mothilal, R. K., A. Sharma, and C. Tan, 2020: Explaining machine learning classifiers through diverse counterfactual explanations. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, doi:10.1145/3351095.3372850.

[10] Pearl, J., 1995: Causal diagrams for empirical research. *Biometrika*, **82**, no. 4, 702–710, doi:10.1093/biomet/82.4.702.

[11] — 2009: *Causality*. Cambridge University Press.

[12] — 2013: Structural counterfactuals: A brief introduction. doi:10.21236/ada580574.

[13] Peng, H., F. Long, and C. Ding, 2005: Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **27**, no. 8, 1226–1238, doi:10.1109/tpami.2005.159.

[14] Ribeiro, M., S. Singh, and C. Guestrin, 2016: Why should i trust you?: Explaining the predictions of any classifier. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, doi:10.18653/v1/n16-3020.

[15] Roser, M., H. Ritchie, E. Ortiz-Ospina, and J. Hasell, 2020: *Coronavirus pandemic (covid-19) - statistics and research*. URL https://ourworldindata.org/coronavirus

[16] Spirtes, P. and K. Zhang, 2016: Causal discovery and inference: concepts and recent methodological advances. *Applied Informatics*, **3**, no. 1, doi:10.1186/s40535-016-0018-x.

[17] Strobl, C., A.-L. Boulesteix, A. Zeileis, and T. Hothorn, 2007: Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, **8**, no. 1, doi:10.1186/1471-2105-8-25.

[18] Vergara, J. R. and P. A. Estévez, 2013: A review of feature selection methods based on mutual information. *Neural Computing and Applications*, **24**, no. 1, 175–186, doi:10.1007/s00521-013-1368-0.

[19] Weisberg, D. S. and A. Gopnik, 2013: Pretense, counterfactuals, and bayesian causal models: Why what is not real really matters. *Cognitive Science*, **37**, no. 7, 1368–1381, doi:10.1111/cogs.12069.

[20] Whitmore, L. S., R. W. Davis, R. L. Mccormick, J. M. Gladden, B. A. Simmons, A. George, and C. M. Hudson, 2016: Biocompoundml: A general biofuel property screening tool for biological molecules using random forest classifiers. *Energy Fuels*, **30**, no. 10, 8410–8418, doi:10.1021/acs.energyfuels.6b01952.

[21] Yap, C. W., 2010: *Padel-descriptor: An open source software to calculate molecular descriptors and fingerprints*. URL https://onlinelibrary.wiley.com/doi/full/10.1002/jcc.21707

□