Title

A systematic method for selecting molecular descriptors as features when training models for predicting physiochemical properties

Authors

Comesana, Ana E
Huntington, Tyler T
Scown, Corinne D
et al.

Publication Date

2022-08-01

DOI

10.1016/j.fuel.2022.123836

Peer reviewed

Fuel 321 (2022) 123836

Contents lists available at ScienceDirect

Fuel

journal homepage: www.elsevier.com/locate/fuel

Full length article

A systematic method for selecting molecular descriptors as features when training models for predicting physiochemical properties

Ana E. Comesana [a], Tyler T. Huntington [b,c], Corinne D. Scown [a,b,c,d], Kyle E. Niemeyer [e], Vi H. Rapp [a,*]

[a] Energy Technologies Area, Lawrence Berkeley National Laboratory, Berkeley, CA, United States
[b] Biosciences Area, Lawrence Berkeley National Laboratory, Berkeley, CA, United States
[c] Joint BioEnergy Institute, Emeryville, CA, United States
[d] Energy & Biosciences Institute, University of California, Berkeley, CA, United States
[e] School of Mechanical, Industrial, and Manufacturing Engineering, Oregon State University, Corvallis, OR, United States

ARTICLE INFO

Keywords:
Chemical descriptors
Machine learning
Biofuel
TPOT

ABSTRACT

Machine learning has proven to be a powerful tool for accelerating biofuel development. Although numerous models are available to predict a range of properties using chemical descriptors, there is a trade-off between interpretability and performance. Neural networks provide predictive models with high accuracy at the expense of some interpretability, while simpler models such as linear regression often lack in accuracy. In addition to model architecture, feature selection is also critical for developing interpretable and accurate predictive models. We present a method for systematically selecting molecular descriptor features and developing interpretable machine learning models without sacrificing accuracy. Our method simplifies the process of selecting features by reducing feature multicollinearity and enables discoveries of new relationships between global properties and molecular descriptors. To demonstrate our approach, we developed models for predicting melting point, boiling point, flash point, yield sooting index, and net heat of combustion with the help of the Tree-based Pipeline Optimization Tool (TPOT). For training, we used publicly available experimental data for up to 8351 molecules. Our models accurately predict various molecular properties for organic molecules (mean absolute percent error (MAPE) ranges from 3.3% to 10.5%) and provide a set of features that are well-correlated to the property. This method enables researchers to explore sets of features that significantly contribute to the prediction of the property, offering new scientific insights. To help accelerate early stage biofuel research and development, we also integrated the data and models into a open-source, interactive web tool.

1. Introduction

Machine learning can leverage a breadth of experimental data in the public domain to accelerate fundamental and applied research for biofuel development. For example, machine learning models can predict relevant biofuel production pathways, as well as physical and chemical properties of potential biofuel molecules [1–3]. Preliminary fuel screening tools based on machine learning have already proven useful in programs worldwide and help identify promising molecules early in the fuel development cycle [1–6]. Many of these property prediction models also use chemical descriptors as features to provide a mathematical link between physiochemical properties and molecular structure [1–3,7–24]. Additionally, chemical descriptors can be easily generated using available software, which provides a feature-rich resource and simplifies model development [25–28]. Coupling molecular descriptors with automated machine learning tools, such as TPOT and Auto-Keras, can further streamline model development by automatically finding the highest performing model architecture among thousands of different algorithms and architectures [29–31].

While numerous machine learning models are available to predict a variety of fuel properties using chemical descriptors, there is a trade-off between interpretability and performance. Interpretability can enhance trust in machine learning models, especially when important features conform to existing knowledge about the target property, justifying the predictions of the model. Additionally, investigating important

* Corresponding author.
E-mail address: vhrapp@lbl.gov (V.H. Rapp).

https://doi.org/10.1016/j.fuel.2022.123836
Received 13 December 2021; Received in revised form 6 February 2022; Accepted 9 March 2022
Available online 4 April 2022

Nomenclature

FD	Fragmental Descriptor.
FGC	Functional Group Count Descriptors.
MAE	mean absolute error.
MAPE	mean absolute percent error.
MD	Molecular Descriptors.
MedAE	median absolute error.
MLR	Multiple Linear Regression.
NI	Not included or reported by authors.
PCA	Principal Component Analysis.
RFE	Recursive Feature Elimination.
RMSE	root mean squared error.
SI	Supplemental Information.
SMILES	simplified molecular input line entry specification.
TPOT	Tree-based Pipeline Optimization Tool.
VEM	valence electron mobile.

features in quantitative structure–property relationship models can provide new information about the property predicted, and lead to better understanding of the link between physical or chemical properties and molecular structure [32]. Many predictive models developed for identifying correlations between descriptors and properties of interest (i.e., interpretability) use multi-linear equations or least-squares regression [3,7–9,15,17–21]. These models are used because the linear coefficients directly correlate the descriptors' contributions to the prediction. However, the models may lack the predictive performance of more complex models such as neural networks, especially when applied to systems that exhibit nonlinear relationships [3,9,18,19]. For example, Kessler et al. [3] compared the performance of Multiple Linear Regression (MLR), Artificial neural network, and Graph neural network models for predicting yield sooting index. The MLR model provided insights into the structural components of a molecule that contribute to increasing yield sooting index, but had a mean absolute error (MAE) about seven times larger than the two neural network models.

In exchange for some interpretability, many researchers have used neural networks to develop fuel property models with high accuracy [2,3,9–11,14,16,18,19]. Although neural networks often perform better than other model architectures, they do not impose restrictions on input variables (i.e., features) or data relationships. This lack of restrictions can lead to combined features and hidden relationships that are not easily explained, making the models difficult to interpret from a mechanistic standpoint. For example, a neural network model may combine several molecular descriptors to create a new feature that minimizes error but does not directly correlate with a molecule's physical characteristics. Also, many neural networks do not show how correlations between features and the property are made during model development. As a result, a whole field of research has been dedicated to understanding their hidden relationships and model structure [33–36].

Feature selection is also critical for developing interpretable and accurate predictive models. It can reduce the risk of overfitting and identify important features with meaningful property relationships in the data [37]. Feature selection can also reduce multicollinearity in feature sets (i.e., features that are linearly correlated), which is often present in molecular descriptors and may promote unstable model coefficients [38]. Commonly used feature selection methods include statistical methods (e.g., using Pearson or Spearman correlation coefficients) or manually selected features based on experimental observations about the property of interest. Although these methods are effective at reducing features, they may overlook important features that offer new scientific insights about the property. For example, using statistical methods that ignore the property of interest may result in a set of non-correlated features irrelevant for that property. Manually selecting features could exclude important features that better characterize the data (e.g., distinguishing between isomers) [1,11,13,14,23,39]. Given the numerous methods for feature selection with varying drawbacks and constraints, it may be daunting to select one for developing a predictive model.

The purpose of this study is to develop a systematic method for creating accurate and interpretable fuel property prediction models that use molecular descriptors. Unlike previously published literature, this approach can be applied to a broad range of properties from physical to complex, and can be used to support scientific discovery. Specifically, the method enables researchers to identify, rank, and validate important property structure relationships that may accelerate fuel development. The method focuses on reducing the number of features by minimizing correlations between chemical descriptors to develop high-performing models. It also ranks the features based on their importance, enabling researchers to identify dominant chemical-structure features that impact property values.

To demonstrate our approach, we created predictive models for five common jet fuel properties that are considered when certifying new jet fuels [40]: melting point, boiling point, flash point, yield sooting index, and net heat of combustion. We selected these properties because they range from physical to complex, with some having a clear relationship with molecular structure (e.g., heat of combustion, yield sooting index) [3,22] and others having ambiguous structure–property relationships [7,10]. We used automated machine learning (TPOT) to develop the models, which leverage chemical descriptors from Mordred [28] and experimental property data of organic molecules from publicly available sources. We validated model accuracy using test data withheld from training and published literature. To demonstrate interpretability of our approach, we provide an in-depth discussion of the important features and how they correlate with properties. The data and models have been integrated into a user-friendly, interactive web tool (feedstock-to-function.lbl.gov) and are publicly available to help accelerate early stage biofuel development research [41,42]. Following this section, we provide a summary of previously published models. We then describe our methods for selecting an optimal number of features and developing interpretable machine learning models. Next, we present our results, draw conclusions, and provide recommendations for future extensions of this work.

2. Previous models

Several researchers have used molecular descriptors, molecular fragments, or functional groups as inputs (or features) for regression, or classification models to predict molecular properties relevant to biofuels.

Table 1 provides a summary of their work. Classical machine learning is based on mathematical algorithms such as linear regression, decision trees, or support vector machines. Deep learning algorithms are based on neural networks, and include perceptrons, artificial neural networks, or adversarial networks. Bayesian methods include Gaussian process regression, or Bayesian linear regression. Some studies only reported training or overall errors that averaged training, testing, and validation (when available) errors [1,2,12,15,39]. As noted by other researchers, reporting the test error provides a better measure of the model's predictive capability because external validation is necessary for determining the true predictive ability of the model [17].

In general, smaller or molecular family-specific (e.g., alcohols, alkanes, hydrocarbons, unsaturated hydrocarbons, and heterogeneous molecules) data sets result in lower errors for the boiling point, flash point, and yield sooting index models [11,12,16,17]. Melting point and heat of combustion models, however, have lower errors with the largest data sets [10,22]. For most of the properties, model type does not have an obvious impact on model accuracy. For example, most

Table 1
Published quantitative structure property relationship models for predicting melting point, boiling point, flash point, yield sooting index, and heat of combustion.

Property	Model type	Dataset size	Feature type	Number of features	Feature selection method	Test					Train + Test					Reference
						MAE	RMSE	MAPE	R^2	MedAE	MAE	RMSE	MAPE	R^2	MedAE	
Melting point (K)	Classical	277	MD	16	Multiple	–	44.6	–	–	–	–	–	–	–	–	[7]
	Classical	323	MD	8	Multiple	–	42	–	0.789	–	–	–	–	–	–	[8]
	Bayesian	1003	FG	42	Manual	–	–	–	–	–	9.5	15.2	–	0.965	–	[1]
	Classical	1097	MD & FG	35	Forward selection	25.9	39.1	11.3%	0.718	–	18.9	30.2	7.6%	0.875		[9]
	Deep	4173	MD	26	PCA	38.2	49.3	–	0.658	–	–	–	–	–	–	[10]
Boiling point (K)	Deep	134[a]	MD	7	Manual	–	–	1.19%	–	–	–	–	–	–	–	[11]
	Deep	150[b]	MD	10	Manual	–	–	–	–	–	–	–	–	0.998	–	[12]
	Classical	155	MD	8	Forward selection	–	7.3	–	–	–	–	–	–	–	–	[13]
	Deep	223	MD	16	Multiple	–	1.4	0.26%	0.999	–	–	2	0.54%	0.999	–	[14]
	Classical	298	MD	4	Multiple	–	–	–	–	–	–	–	2.3%	0.973	–	[15]
	Deep	1116	MD	6	Manual	11.6	–	4.33%	–	–	–	–	–	–	–	[16]
	Bayesian	1238	FGC	42	Manual	–	–	–	–	–	10.6	20	–	0.948	–	[1]
	Deep	7367	Functional Group Count Descriptors (FGC)	24	Random Forest	–	–	–	–	–	5.352	–	0.741	0.991	–	[43]
	Classical	80[b]	MD	3	Manual	–	–	–	0.999	–	–	–	–	–	–	[17]
	Classical	65[a]	MD	3	Manual	–	–	–	0.986	–	1.62	–	–	–	–	[17]
	Classical	70[c]	MD	3	Manual	–	–	–	0.937	–	–	–	–	–	–	[17]
	Deep	17768	MD	44	Forward selection	–	21	–	0.947	–	–	22	–	0.943	–	[44]
Flash point (K)	Classical	268	MD	9	n/a	9.9	–	–	0.9314	–	–	–	–	–	–	[18]
	Deep	513	MD	15	n/a	–	–	–	–	–	–	14	–	0.934	–	[18]
	Multiple	625	MD & FGC	28	Same as [21]	8.4	13.2	2.5%	0.944	–	7.1	10.9	2.2%	0.959	–	[19]
	Classical	1030	MD	4	Genetic algorithm-MLR	–	–	–	0.971	–	10.2	–	–	0.967	–	[20]
	Bayesian	1065	FG	42	Manual	–	–	–	–	–	6.2	10.1	–	0.980	–	[1]
	Deep	2934	FGC	24	Random Forest	4.461	5.715	1.325	0.988	–	3.952	5.084	1.158	0.991	–	[43]
Yield sooting index	Deep	297	MD	390	Multiple	–	–	–	–	–	–	–	–	–	3.08	[2]
	Bayesian	457	FCG	42	Manual	–	–	–	–	–	2.7	7.6	–	0.999	–	[1]
	Deep	567	MD	1800	Multiple	–	–	–	–	4.34	–	–	–	–	–	[3]
	Classical	441	FD	66	Manual	–	–	–	–	–	–	–	–	–	2.35–28.6	[39]
Heat of combustion (kJ/mol)	Classical	1650	MD	4	Multiple	104.1	163.2	–	0.996	–	–	–	–	–	–	[21]
	Classical	1714	MD	4	Genetic algorithm-MLR	104.13	156.9	15.18%	0.996	–	117.8	196.74	11.80%	0.995	–	[22]
	Multiple	2767	MD & FGC	35	Forward selection	35.7	62.7	0.80%	0.999	–	32.2	52.4	0.70%	0.999	–	[9]
	Deep	4590	FGC	142	Manual	–	17.23	0.20%	0.999	–	–	12.57	0.16%	0.999	–	[23]

[a]Hydrocarbons.
[b]Alkanes.
[c]Alcohols
Average of test, train, and validation (when available) errors.
Molecular Descriptors (MD); FGC; Fragmental Descriptor (FD); Principal Component Analysis (PCA); MLR;
Not included or reported by authors (NI).

of the classical melting point models report an root mean squared error (RMSE) of about 40 K for their test data [7–9], with the neural network model having the largest RMSE (49 K) with the largest data set [10]. Boiling point shows a similar trend with the smaller data-set, classical models having lower errors than the larger data-set, neural network models. Due to lack of reported test errors, there is not enough information to determine how model type might impact flash point and yield sooting index model accuracy.

When comparing feature types, heat of combustion model performance was significantly better when functional groups (i.e., group contribution methods) were used as features instead of molecular descriptors [9,23]. Saldana et al. [9] developed 11 models using both descriptors and functional groups. To achieve the best performing model, they used a consensus modeling approach, where results from the best performing (lowest error) descriptor model and functional group model were averaged to produce the final prediction. For the remaining properties, there was not enough data to identify any additional trends with feature types.

Despite the significant advancements for developing models for predicting properties, a different feature selection method was used for almost each study. Approaches ranged from manually selecting features to using varied, and sometimes multiple, statistical methods. Approximately half of the prior studies we surveyed discussed the relationship of the model features to the property and most of them used classical learning models [7,8,13–15,17,20–22]. This is likely because classical learning models can provide direct information on how features impact model predictions. For example, two classical models used to better understand how specific chemical features affect melting point of potential drugs provided an extensive discussion on the relationship between individual descriptors and their effect on melting point [7,8]. When using a deep learning model, obtaining detailed information to understand the scientific relevance of the models may be more difficult [3].

3. Methods

3.1. Experimental property data

To train the property prediction model, we aggregated and coalesced experimental property data for organic molecules from publicly available, published sources [2,4,19,39,45–51]. Property data included flash point, boiling point, melting point, heat of combustion, and yield sooting index. While heat of combustion has been linked to molecular structure, properties such as flash point and yield sooting index also depend on combustion chemistry. The wide range of these properties illustrates the versatility of this approach, and the potential to be applied to other properties. For flash point, we only included open-cup data (i.e., excluded Pensky–Martens, Abel, Tag, and Setaflash methods of testing flash point) to ensure results were comparable with each other. Because our models focus on biofuel and bioproduct development, we only included organic molecules containing carbon, hydrogen, oxygen, and/or nitrogen and restricted molecules to those with 30 or fewer carbon atoms. If multiple sources reported experimental data for the same molecule, we compared measurements and either averaged or removed the data. Specifically, measurements within 15% or five units of each other were averaged, while measurements differing by more than 15% and five units were considered unreliable, and the molecule was removed. The final database comprises a variety of chemical classes, such as alkanes, cycloalkanes, alkenes, cycloalkenes, alkynes, alcohols, cycloalcohols, aldehydes, ketones, cyclic ketones, esters, ethers, and aromatics (see Table S2 in the supplementary material for more information). Compared to published literature, our dataset has considerably more (up to 30 times) alcohols, cycloalcohols, aldehydes, ketones, cyclic ketones, esters, ethers, carboxylic acids, and aromatics, a comparable number of alkanes, cycloalkanes, alkenes, cycloalkenes, and alkynes to other datasets [1].

Using Pandas [52], Scikit-learn [53], and RDKit [54], we retrieved simplified molecular input line entry specification (SMILES) for each molecule to index and merge data from different sources. To index the database by SMILES, two challenges arise. First, isomers need unique SMILES to be correctly identified. For this reason, we obtained isomeric SMILES for the molecules with isomers. Second, a single molecule can have multiple SMILES. To avoid duplicated molecules when merging data, we standardized the SMILES using the `MolToSmiles()` and `MolFromSmiles()` functions in `RDKIT.Chem`. These functions map different SMILES belonging to the same molecule to a single molecule object in Rdkit, then return a single standardized SMILES. If a published database did not index molecules by SMILES, such as Co-Optima [4], we used the names of molecules to find SMILES via PubChem [46]. The complete database used for developing our models can be found at feedstock-to-function.lbl.gov.

3.2. Feature selection

For our model features, we generated molecular descriptors using Mordred [28] because it is an open-source library that offers a wide variety of descriptors. Mordred includes more than 1800 features grouped by 50 categories called modules. Descriptors are generated from SMILES and can be two- or three-dimensional. After generating all descriptors for each molecule, we removed descriptors with non-numerical values (e.g., containing strings or `NA` values) and descriptors with matching values across most (>95%) molecules. We also removed the `Autocorrelation` descriptors because they may not directly correspond to structural or physical properties of the molecule [55].

When reducing the number of features, we used Recursive Feature Elimination (RFE) with an ensemble model estimator (Random Forest Regressor) because it was used in other studies, is easy to implement, and can outperform other approaches [2,56,57]. RFE is a supervised feature selection method. Fig. 1 shows a flowchart of our model development process. To effectively use RFE, we first removed correlations between descriptors by developing correlation matrices for each descriptor module using the Spearman coefficient [58]. Although the Pearson coefficient is more commonly used [14,19], the Spearman coefficient captures not only linear but all monotonic relationships. The Spearman coefficient is also appropriate for discrete ordinal and continuous variables, both of which are present in molecular descriptors.

For each descriptor in the module correlation matrix, we counted the number of descriptors with a correlation coefficient of at least 0.7. We then ranked the descriptors from highest to lowest by number of correlated module descriptors. Next, we selected the top descriptor as a feature descriptor and removed that feature descriptor and its correlated descriptors from the correlation matrix. Last, we regenerated the correlation matrix and repeated these steps until all descriptors in the module were either selected as a feature descriptor or deleted. If descriptors had the same number of correlated descriptors and therefore the same ranking, they were selected by their order in Mordred. We repeated this process for individual modules and then a final time after combining all feature descriptors. This process generated the feature descriptor set for RFE.

To minimize bias in the model performance, we randomly divided each property data set into training (80%) and testing (20%) prior to using RFE. Fig. S2 and Table S1 in the Supplemental Information (SI) show the distribution of training and testing data and the number of molecule types (e.g., hydrocarbons, oxygenated hydrocarbons) used for each model.

Next, we implemented RFE using `scikit-learn` [53], which iteratively groups and removes the 10 least important descriptors until no descriptors remain. The importances of the descriptors are derived by fitting a random forest estimator at each step and looking at the accumulation of the impurity decrease within each tree [59]. Upon changing the random seed 10 times for heat of combustion, 7 out

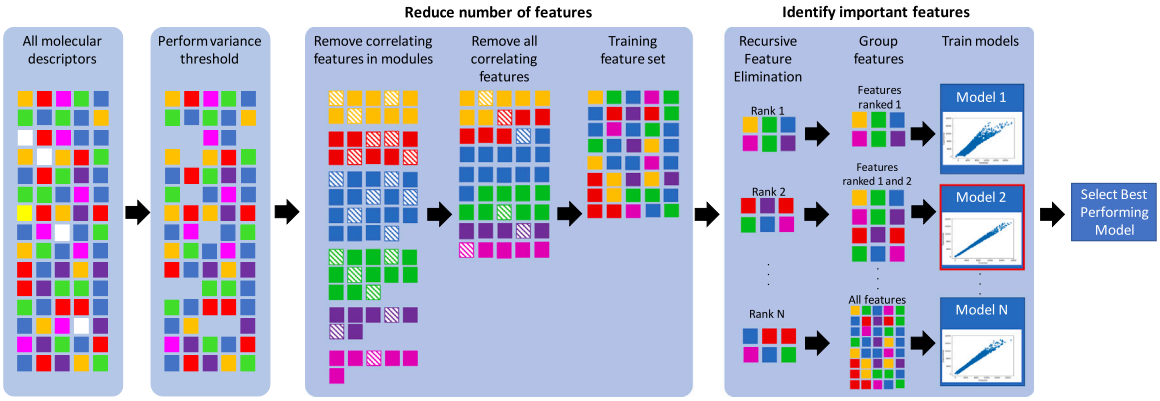

Fig. 1. Overview of feature selection and model optimization pipeline.

of the top 10 features were ranked 1st all 10 times, while the other three were ranked either 1st or 2nd. This shows that the method produces consistent results when randomizing the estimator. It then ranks the descriptor groups based on their removal order. For example, the last remaining descriptors will be ranked 1, while the second-to-last remaining descriptors will be ranked 2, and so forth. This method has been shown to be robust and useful for identifying strong predictors in low dimensional data [60].

We then created multiple feature sets, with the first feature set including only rank 1 descriptors, by sequentially adding higher ranked descriptor groups until the last feature set included descriptors of all ranks. We then trained TPOT on all the generated feature sets.

To evaluate the robustness of this approach, we varied the random seed ten times, retrained the models, and observed the changes in feature importances. We found that random seed did not have more than 20% variability in any of the feature importances, with the maximum change being between 0.48 and 0.6, in the yield sooting index model (see Fig. S1 in the SI for more information).

3.3. Model development using automated machine learning

TPOT is a tree-based automated machine learning tool that finds the best model for a given data set by exploring thousands of model pipelines [29,30]. For our models, we set TPOT to optimize only ensemble models because they can provide prediction intervals with quantile regression forests or quantile extra trees. This allowed us to estimate conditional quantiles and evaluate the reliability of the prediction for each molecule [61]. When training the models, TPOT performed five-fold cross-validation with each fold comprising 20% of the training set. TPOT performs the cross-validation internally, using the average accuracy of all five folds to select the highest performing model architecture. Then, it trains the highest performing model architecture with the entire training set and returns the final model [29,30]. For all feature sets described in Section 3.2, we used TPOT to develop models and then compared model performances. Like previous studies [1,3,9], we selected the final predictive model for the property, and its corresponding feature set, with the lowest cross-validation error. Fig. 1 provides a graphical representation of the feature and model selection process.

3.4. Model performance metrics

To evaluate the prediction performance and accuracy of each model, we calculated the RMSE, MAE, MAPE, and median absolute error (MedAE). These metrics do not indicate if the model predictions over- or under-estimate experimental data. Instead, they measure model performance and accuracy. They can be used on the train, test, or validation sets, but the performance on the test or validation sets should be used to estimate model's performance on unseen data.

The RMSE is frequently used to estimate model error:

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}\,, \tag{1}$$

where n is the number of samples (or molecules), y_i is the dependent variables, with $i = 1, \ldots, n$, and $\hat{y}_i$ is the predicted value of y_i. Due to the quadratic nature of RMSE, this metric is especially sensitive to large errors and outliers.

MAE is also a popular performance metric because it is intuitive and easy to interpret:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|\,. \tag{2}$$

MAE is less sensitive to large errors and outliers than RMSE, which may be desirable if the data span a large range.

MAPE is similar to MAE and provides a dimensionless (i.e., unit-less) measure of model accuracy:

$$\mathrm{MAPE} = 100 \times \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|\,. \tag{3}$$

However, MAPE captures errors that may seem large in terms of magnitude, but are small compared to the overall range of the data. This metric places more penalty on errors at the lower end of the target range. For example, MAPE will demonstrate that an error of 20 units when the expected value is 5 is less accurate than an error of 20 units when the expected value is 500.

Because MedAE is the median of all the absolute errors, it is robust to outliers and provides a measure of model performance for the majority of the data (i.e., excluding outliers):

$$\mathrm{MedAE} = \mathrm{median}\left(|\mathbf{y} - \hat{\mathbf{y}}|\right)\,, \tag{4}$$

where median($\mathbf{x}$) represents the median value of set $\mathbf{x}$.

4. Results and discussion

To develop accurate and interpretable machine learning models, we determined the optimal descriptor sets and created predictive models for five physiochemical properties using the method described in Section 3. We developed models for melting point, boiling point, flash point, yield sooting index, and heat of combustion. A summary of the models and their data performance (training and testing) is shown in Table 2.

Interestingly, the optimal model architecture identified by TPOT for all properties was ExtraTrees. This algorithm creates a large number of regression trees each trained on a random subset of features, then averages the prediction of each tree to come up with a final prediction. It has been shown to provide near optimal accuracy and good computational complexity [62]. In addition, variable importances derived from

Table 2
Property prediction molecules, features, and model test performance using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Median Absolute Error (MedianAE), and Mean Absolute Percentage Error (MAPE) for test data.

	Melting point	Boiling point	Flash point	Yield sooting index	Heat of combustion
MAE	29.2 K	15.7 K	13.4 K	17.32	91.9 kJ/mol
RMSE	39.9 K	28 K	20.4 K	40.54	238.3 kJ/mol
MedianAE	22.6 K	8.7 K	8.3 K	4.08	23.1 kJ/mol
MAPE	8.4%	3.3%	3.8%	10.5%	4.4%
Number of features	73	45	31	15	10
Number of molecules	8351	2431	1130	481	2489
Data range	68–822 K	225–903 K	85–546 K	−3.1–1339 1	116.4–17,893 kJ/mol
Model architecture	ExtraTrees	ExtraTrees	ExtraTrees	ExtraTrees	ExtraTrees

ensemble models such as ExtraTrees can properly assess the relevance of a variable [59].

In general, our predictive models demonstrate comparable performance to previously published models. Our melting point model uses two to eight times more data than previous studies and has test errors comparable to Saldana et al. [9] (about 4 K higher MAE and RMSE) and lower than Karthikeyan et al. [10] (about 10 K lower MAE and RMSE). Our boiling point and flash point models have similar errors to previous models of larger size (greater than 1000 molecules) and diverse data-sets (i.e., models that were trained using multiple classes of organic compounds) [16,18,19]. The yield sooting index model also has errors that match previous studies, but uses only a fraction of the features (15 for our model versus 390 or 1800 for others), which may help prevent overfitting [2,3]. Although the heat of combustion model performed better than linear models that used chemical descriptors as features (MAE of 91.9 kJ/mol versus 104.1 kJ/mol) [21,22], it did not perform as well as models that used functional group counts as inputs [9,23]. For heat of combustion, it seems that using a group contribution approach compared to molecular descriptors as model inputs generally produces better model performance [63,64]. This is likely due to the fact that fuel enthalpy, which is related to heat of combustion, has been shown to correlate to functional groups. As shown in Table 1, this may not be consistent for other physiochemical properties.

In some cases it was difficult to compare performance of our model to the published literature because they only reported overall errors that averaged both training and testing errors (i.e., errors for the testing data were not provided) [1,12,15,17,18,20]. Reporting model performance using the average of training and testing errors may be misleading because the model is likely to have lower errors with training data than testing data. Therefore, the overall error may overestimate the predictive capabilities of the model, especially with new data. As discussed by other researchers, external validation is an indispensable validation method for determining the true predictive ability of the model, and reporting the test error will provide a better measure of the model's predictive capability [17,65,66].

When evaluating model performance for different molecular families (e.g., hydrocarbons, oxygenated hydrocarbons, organic nitrogen molecules), Fig. 2 shows that most models predict hydrocarbon properties better than nitrogen- and oxygen-containing compounds. Hydrocarbons may be easier to predict because their properties tend to correlate highly with bond structure and intermolecular connectivity [67]. For nitrogen- and oxygen-containing compounds, additional intermolecular forces (such as dipole–dipole moments) and degrees of freedom may influence properties in ways that are difficult to capture with available molecular descriptors. Yield sooting index is the only model with higher errors for hydrocarbons than other molecule families; however, the yield sooting index range for hydrocarbons is more than 10 times larger than oxygenated hydrocarbons and organic nitrogen molecules (see Tables S1 and S3 in the SI). The remaining models have the largest error among organic nitrogen molecules. For information about RMSE for these groups, see Table S4.

The following sections discuss in detail the chemical descriptors used for each model and interpret their relationships to the properties. Additionally, the SI contains a full list of descriptors, their definition, and their feature importance values for each model.

4.1. Melting point model

To interpret the melting point model features (i.e., descriptors) and understand which characteristics are captured by the model, Fig. 3 shows the feature importance of the descriptors grouped by module. Melting point depends on molecule properties and strength of the crystalline lattice. These are functions of molecular packing in the crystals (e.g., shape, size, symmetry) and intermolecular forces such as charge transfer and dipole–dipole interactions in the solid phase [10,68–70]. Melting point also depends on many entropic parameters such as oligomerization and other self-organization processes that may not be captured by chemical descriptors [7,8,10,68,69].

The highest ranked descriptor in the melting point model is `piPC3`, defined as the 3-ordered pi-path count (log scale) [28]. `piPC3` is the only descriptor used from the `PathCount` module and captures aspects of the molecular structure. Specifically, it measures the amount of branching in the bonded structure of the molecule, where aromatic bonds are weighted more heavily than single bonds [71]. As shown in Fig. S3, melting point tends to increase as the number of 3-ordered pi-path count (log scale) increases. Additionally, given a fixed value for `piPC3`, oxygenated hydrocarbons and organic nitrogen compounds have higher melting points than their hydrocarbon counterparts.

The second highest ranked descriptor is `MZ`, the only descriptor used from the `Constitutional` module. `MZ` is defined as the constitutional mean and does not contain geometric information. It is determined by calculating the mean atomic number of the molecule and then normalizing the mean by the atomic number of carbon (i.e., 6) [28]. In general, melting point increases with `MZ` (see Fig. S4). An organic molecule with a nitrogen or oxygen atom has a greater `MZ` value than a hydrocarbon with only carbon and hydrogen atoms because nitrogen and oxygen have greater atomic numbers than carbon. The dipole–dipole interactions between either the lone pairs in nitrogen or oxygen and hydrogen are stronger than the dipole–dipole interactions between carbon and hydrogen. This effect is owed to nitrogen and oxygen having greater electronegativity than carbon, which results in stronger intermolecular forces that require more energy to break (i.e., a greater melting point). This could explain why `MZ` is an important predictor of melting point in our model. The module with the largest aggregated feature importance is `MOEType`, which comprises 34 descriptors, including those that characterize intermolecular interactions. The `MOEType` module collects two-dimensional descriptors from the Molecular Operating Environment Software, based on precalculated surface area values (or VSA) derived from a list of functional groups that approximate van der Waal surface area [72,73]. `MOEType` descriptors capture many fundamental characteristics needed to predict melting point, including hydrophobic and hydrophilic effects, polarizability, electrostatic interactions, and steric effects [7,10,72,74,75].

Most of the remaining descriptors and modules in our melting point model capture additional features that characterize molecular shape, atoms, bond types, functional groups, and polarizability. For example, topological descriptors count numbers of atoms, bonds, branching, and electrons, and can be used to characterize polarizability, dipole moment, and some steric effects [76]. Polarizability and induced dipole

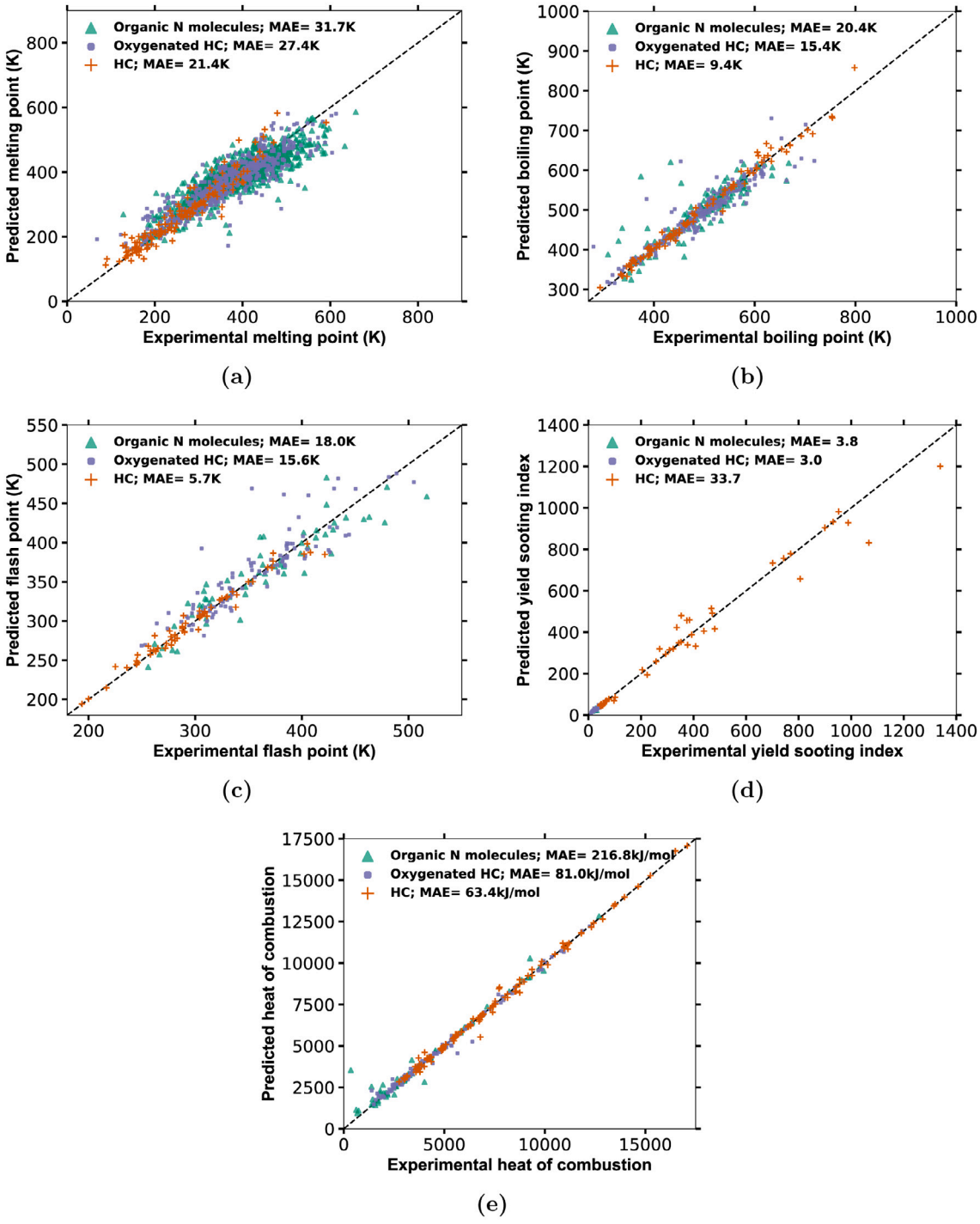

Fig. 2. Parity plots showing test set experimental and predicted values of **(a)** Melting Point, **(b)** Boiling Point, **(c)** Flash Point, **(d)** Yield Sooting Index, and **(e)** Heat of Combustion models for hydrocarbons, oxygenated hydrocarbons, and all other organic molecules.

interactions influence intramolecular interactions that are important determinants of melting point, especially for larger molecules with electrons that are easier to polarize [8,10].

A key difference between our melting point model and other published models is that it can distinguish between isomers because it includes three-dimensional descriptors. Only two three-dimensional descriptors are used in the melting point model: Plane of Best Fit (PBF) and `GeometricalShapeIndex`. Interestingly, these descriptors have a low feature-importance ranking. Previous research has suggested that melting point models without three-dimensional descriptors perform better than those with both two- and three-dimensional descriptors [10]. Here, we found that the addition of three-dimensional descriptors improved accuracy.

Given the diversity of the data used to train the model, common descriptors between many molecules may be more heavily weighted than other descriptors, resulting in larger prediction errors. For example, more than half of the molecules used for training contain a nitrogen atom and the third highest ranked descriptor is the number of nitrogen atoms (`nN`) in the `AtomCount` module. As another example, the

Fig. 3. Melting point descriptor importances grouped by module.

model predicts that 1,2-epoxyhexane ($C_6H_{12}O$) has a low melting point (171 K) like 1,3-epoxybutane (C_4H_8O), but its actual melting point is more than double (367 K). This indicates that some of the highest-ranked modules, such as `Estates`, `RingCount`, and `AtomCount`, that characterize the types of atoms and fragments or functional groups present in the molecules are missing some features, resulting in higher model predictions. Specifically, the descriptors may not accurately characterize molecules with stronger intermolecular interactions in the solid state due to oligomerization or other self-organization processes. Other researchers have observed similar phenomenon using descriptors to predict melting point, and indicate that available descriptors may not be sufficient to accurately predict melting point of molecules with strong, solid-state intermolecular interactions [7,10,68,69].

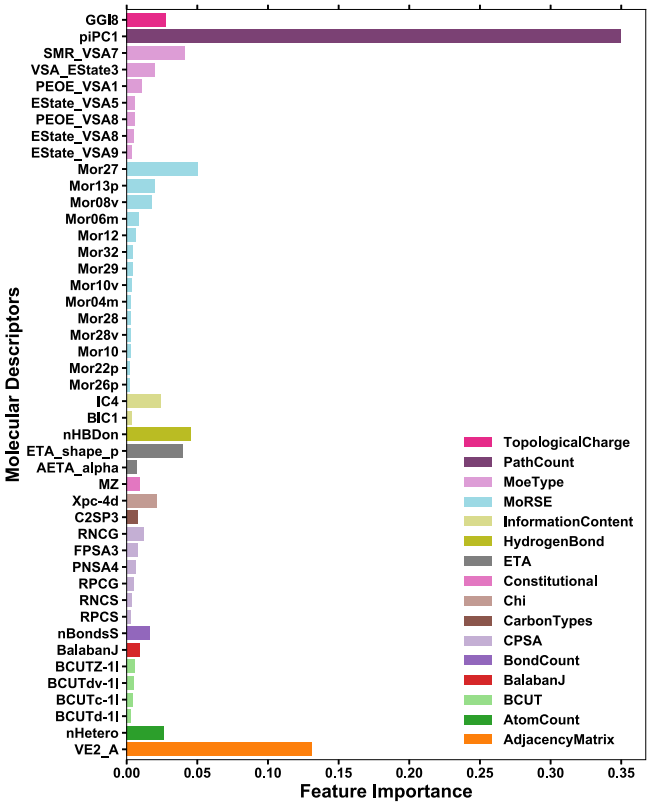

Fig. 4. Boiling point descriptor importance values grouped by module.

4.2. Boiling point model

To understand which characteristics are captured by the boiling point model, we compared the importance of individual descriptors (see Fig. 4). Boiling point depends on physical characteristics such as branching in the molecular structure, different measures of intermolecular connectivity, and the dipole–dipole interactions within the molecule [13,14,17]. Therefore, it is not surprising that `piPC1` is the highest ranked descriptor for the boiling point model, since it measures the amount of branching in the molecular structure. `piPC1` is defined as the 1-order pi-path count (log scale) and is the only descriptor used from the `PathCount` module [28]. As Fig. S5 shows, boiling point generally increases with increasing path count values, which agrees with previous research by Dai et al. [17]. The second-highest ranked descriptor is `VE2_A`, from the `AdjacencyMatrix` module, which describes intermolecular connectivity and vertex centrality [14,17,28]. Similar to previous studies, Fig. S6 shows that boiling point is inversely proportional to `VE2_A` values [14,17]. The `nHBDon` descriptor, which represents the number of hydrogen bond donors in the molecule, is ranked fourth-most important in our boiling model.

The `MoRSE` module was ranked second highest in the boiling point model, after `PathCount`. This module of descriptors was developed to encode the three-dimensional structure of a molecule without Cartesian or internal coordinates. The calculation involves the Euclidean distance between atoms, the total number of atoms, and different atomic properties [77]. When unweighted, the descriptors in this module are just functions of number of atoms in the molecules. These descriptors can also be weighted by atomic properties such as mass, van der Waals volume, or Sanderson electronegativity, which will highlight or diminish the role of certain atoms in the molecule. In the boiling point prediction model, the descriptors used are unweighted, and weighted by mass, van der Waal volume, and polarizability. `MoRSE` three-dimensional descriptors have been used for predicting boiling point, specifically weighted by van der Waal volume [14].

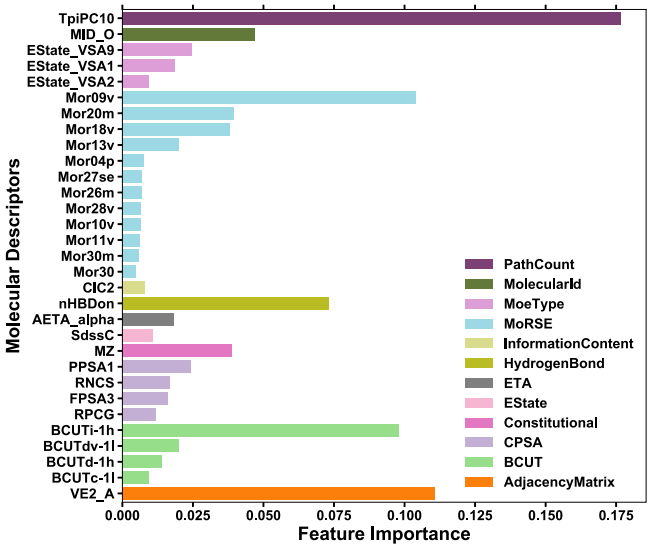

Fig. 5. Flash point descriptor importances grouped by module.

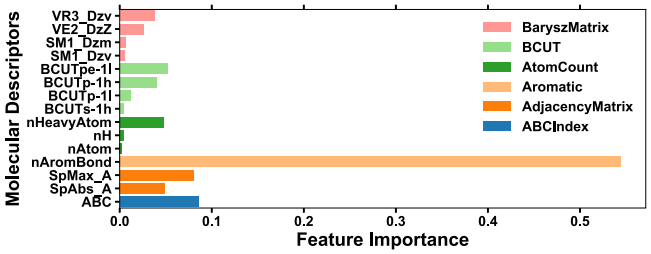

Fig. 6. Yield sooting index descriptor importances grouped by module.

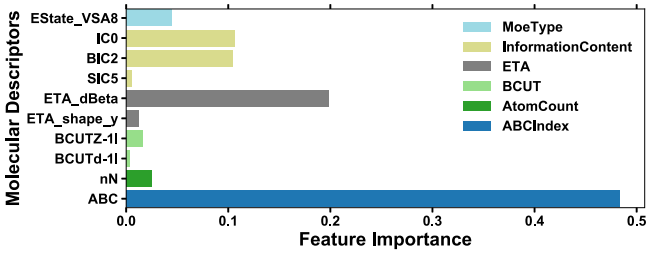

Fig. 7. Heat of combustion descriptor importances grouped by module.

4.3. Flash point model

Fig. 5 shows the importance of features used in the flash point model. Flash point correlates well with boiling point, and some studies even used one property to predict the other [78,79]. Therefore, it is not surprising that the top descriptors used in the flash point model, as well as their relative importances, mimic those in the boiling point model. Specifically, the top ranked descriptors are `TpiPC10` in the `PathCount` module and `VE2_A` in the `AdjacencyMatrix` module. Figs. S8 and S7 show that `VEA_2` may correlate better with hydrocarbons than `TpiPC10`. While the `PathCount` module has been successfully used in previous flash point models, the `AdjacencyMatrix` module has not [18]. However, similar descriptors that capture the number of bonds and bond strength of a molecule, such as edge adjacency indices, have been used in some models [20].

Like in the boiling point model, the `MoRSE` module has the highest aggregated importance in the flash point model and captures three-dimensional structure of the molecules. The `nHBDon` descriptor also captures similar effects, as increasing hydrogen bond donors increases flash point, and has been used in previous models [23].

4.4. Yield sooting index model

Fig. 6 shows the importance of features used in the yield sooting index model. Aromaticity in molecules correlates with higher sooting tendencies (i.e., higher yield sooting indices) [39,80]. As expected, the highest ranked descriptor, which has an importance greater than all other descriptors combined, is `nAromBond` in the `Aromatic` module. This descriptor counts the number of aromatic bonds [28]. As noted by previous researchers, molecules with aromatic atoms have significantly higher yield sooting index than their non-aromatic counterparts [39]. Our model shows a similar trend, with yield sooting index generally increasing with the number of aromatic bonds (see Fig. S9).

Although `nAromBond` is an important feature in the yield sooting index model, `nAromBond` does not capture structure of non-aromatic portions of the molecule or information about non-aromatic molecules. As such, the remaining descriptors in the model capture non-aromatic features. For example, the second-highest ranked descriptor was `ABC`, defined as the atom-bond connectivity index, and measures branching [28]. This is the most important descriptor in the heat of combustion model, and details about this descriptor are included in Section 4.5. Additionally, four descriptors come from the BCUT module, which is calculated from the Burden matrix (a vertex- and edge-weighted adjacency matrix) [81]. One of the descriptors that measures the highest eigenvalue of the burden matrix weighted by polarizability is also used in the yield sooting index model developed by Kessler et al. [3].

4.5. Heat of combustion model

Fig. 7 shows the importance of features used in the heat of combustion model. The highest ranked descriptor in the heat of combustion model is the `ABC` descriptor, with a feature importance of almost 0.5. The `ABC` descriptor is the Atom-bond Connectivity Index, or ABC Index, developed as an improved version of the Randić Connectivity Index, which measures branching in saturated hydrocarbons [67,82]. The ABC Index is a degree-based molecular structure descriptor used to model thermodynamic properties of organic chemical molecules and advance nanochemical applications [83]. The ABC Index correlates with the stability of linear and branched alkanes as well as the strain of energy of cycloalkanes. It also correlates with heat of formation of alkanes and cycloalkanes, and can predict their thermodynamic properties [67,84,85].

As shown in Fig. S10, `ABC` correlates well with heat of combustion, and clearly correlates with alkanes and cycloalkanes, agreeing with previous studies [67,84]. Almost 50% of our data comprise hydrocarbons (1164 out of 2490 molecules), with 25% being alkanes and cycloalkanes (498 alkanes and 121 cycloalkanes). This may partially explain the high importance of the `ABC` descriptor in the model.

The second-highest ranked descriptor is `ETA_dBeta` from the `ETA` module. β is the valence electron mobile (VEM) count, and sums contributions from sigma bonds, pi bonds, and a δ term that measures resonating lone pair electrons in an aromatic system [83]. `ETA_dBeta` is defined as the difference between the contribution from non-sigma bonds (i.e., pi-bonds and δ) and sigma bonds [28]. From Fig. S11, we can see that this descriptor better-isolates organic nitrogen molecules than `ABC`. This suggests that `ETA_dBETA` may encode structural features to better characterize organic nitrogen molecules.

5. Conclusions

This research establishes a comprehensive method for developing interpretable models that predict multiple molecular properties. The method can be applied to a broad range of properties from physical to

complex and can help researchers identify, rank, and validate important property structure relationships that may accelerate biofuel development. The method focuses on reducing the number of features by minimizing correlations between chemical descriptors to develop high-performing models. It also ranks the features based on their importance, enabling researchers to identify dominant chemical-structure features that impact property values.

To demonstrate our method, we developed molecular property prediction models for five common jet fuel properties: melting point, boiling point, flash point, yield sooting index, and net heat of combustion. The properties range from physical to complex, having known and unknown relationships with molecular structure or combustion chemistry. Data used to train the models contain organic molecules (specifically, carbon atoms attached to hydrogen, oxygen, and/or nitrogen atoms) with less than 30 carbon atoms. The numbers of molecules used to develop the models range from 481 molecules (yield sooting index) to 8351 molecules (melting point).

The MAPE for the models, based on the test data, range from 3.3% (boiling point) to 10.5% (yield sooting index), performing similarly to previously published models. A key advantage to these models is that they enable users to directly explore the relationships between properties and the importances of individual descriptors or modules used by the model. For example, we found that the Atom-bond Connectivity Index, a measure of molecule branching, well-predicts heat of combustion, especially for alkanes and cycloalkanes. We also observed that the number of aromatic bonds is a good predictor of yield sooting index, agreeing with previous research. The consistency of this method was tested by randomizing different parts of the algorithm and comparing the results, which are consistent and highlight the same features each time.

Overall, our method provides a consistent and robust approach for developing physiochemical property-prediction models. To accelerate early biofuel research and development, we integrated the data and models into a user-friendly, interactive webtool that is publicly archived and can be found at feedstock-to-function.lbl.gov [41,42]. From this research, we recommend that future models report test errors in addition to overall errors that combine testing and training errors. This will improve transparency for model performance, especially with new or unseen data. Additionally, this method could be used to develop models for other physiochemical properties (e.g., viscosity, density) or to reduce multicollinearity of other highly correlated feature sets similar to chemical descriptors (e.g., gene expression datasets) when developing models.

CRediT authorship contribution statement

Ana E. Comesana: Conceptualization, Methodology, Software, Validation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Tyler T. Huntington:** Software, Data curation, Writing – review & editing. **Corinne D. Scown:** Writing – review & editing, Project administration, Supervision, Funding acquisition. **Kyle E. Niemeyer:** Conceptualization, Writing – review & editing, Supervision, Funding acquisition. **Vi H. Rapp:** Conceptualization, Methodology, Writing – review & editing, Supervision, Funding acquisition.

Declaration of competing interest

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

Acknowledgments

This work was supported by the Bioenergy Technologies Office of the U.S. Department of Energy through Contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the US Department of Energy. This research also used resources of the Oak Ridge Leadership Computing Facility, which is a Department of Energy Office of Science User Facility supported under Contract DE-AC05-00OR22 725. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a nonexclusive, paidup, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

The authors would like to thank Josh Heyne for providing a wealth of data and answering all our questions, Charles Finney for computational support, Nicole Labbe for her feedback and recommendations, Elisa Vidales for her feedback, and Morgan Mayer for her feedback and support during model development.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.fuel.2022.123836.

References

[1] Li R, Herreros JM, Tsolakis A, Yang W. Machine learning-quantitative structure property relationship (ML-QSPR) method for fuel physicochemical properties prediction of multiple fuel types. Fuel 2021;304. http://dx.doi.org/10.1016/j.fuel.2021.121437.

[2] St. John PC, Kairys P, Das DD, McEnally CS, Pfefferle LD, Robichaud DJ, et al. A quantitative model for the prediction of sooting tendency from molecular structure. Energy Fuels 2017;31(9):9983–90. http://dx.doi.org/10.1021/acs.energyfuels.7b00616.

[3] Kessler T, St. John PC, Zhu J, McEnally CS, Pfefferle LD, Mack JH. A comparison of computational models for predicting yield sooting index. Proc Combust Inst 2021;38:1385–93. http://dx.doi.org/10.1016/j.proci.2020.07.009.

[4] National Renewable Energy Laboratory. Co-Optimization of Fuels & Engines: Fuel Properties Database. https://www.nrel.gov/transportation/fuels-properties-database/,

[5] vom Lehn F, Brosius B, Broda R, Cai L, Pitsch H. Using machine learning with target-specific feature sets for structure-property relationship modeling of octane numbers and octane sensitivity. Fuel 2020;281:118772. http://dx.doi.org/10.1016/j.fuel.2020.118772.

[6] vom Lehn F, Cai L, Tripathi R, Broda R, Pitsch H. A property database of fuel compounds with emphasis on spark-ignition engine applications. Appl Energy Combust Sci 2021;5:100018. http://dx.doi.org/10.1016/j.jaecs.2020.100018.

[7] Bergström CAS, Norinder U, Luthman K, Artursson P. Molecular descriptors influencing melting point and their role in classification of solid drugs. J Chem Inf Comput Sci 2003;43:1177–85. http://dx.doi.org/10.1021/ci020280x.

[8] Modarresi H, Dearden JC, Modarress H. QSPR correlation of melting point for drug compounds based on different sources of molecular descriptors. J Chem Inf Model 2006;46:930–6. http://dx.doi.org/10.1021/ci050307n.

[9] Saldana DA, 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 QSAR Environ Res 2013;24:259–77.

[10] Karthikeyan M, Glen RC, Bender A. General melting point prediction based on a diverse compound data set and artificial neural networks. J Chem Inf Model 2005;45:581–90. http://dx.doi.org/10.1021/ci0500132.

[11] Gakh AA, Gakh EG, Sumpter BG, Noid DW. Neural network-graph theory approach to the prediction of the physical properties of organic compounds. J Chem Inf Comput Sci 1994;34:832–9. http://dx.doi.org/10.1021/ci00020a017.

[12] Cherqaoui D, Villemin D. Use of a neural network to determine the boiling point of alkanes. J Chem Soc Faraday Trans 1994;90:97. http://dx.doi.org/10.1039/ft9949000097.

[13] Sola D, Ferri A, Banchero M, Manna L, Sicardi S. QSPR prediction of N-boiling point and critical properties of organic compounds and comparison with a group-contribution method. Fluid Phase Equilib 2008;263:33–42. http://dx.doi.org/10.1016/j.fluid.2007.09.022.

[14] Roubehie Fissa M, Lahiouel Y, Khaouane L, Hanini S. QSPR estimation models of normal boiling point and relative liquid density of pure hydrocarbons using MLR and MLP-ANN methods. J Mol Graph 2019;87:109–20. http://dx.doi.org/10.1016/j.jmgm.2018.11.013.

[15] Katritzky AR, Mu L, Lobanov VS, Karelson M. Correlation of boiling points with molecular structure. 1. A training set of 298 diverse organics and a test set of 9 simple inorganics. J Phys Chem 1996;100:10400–7. http://dx.doi.org/10.1021/jp953224q.

[16] Espinosa G, Arenas A, Giralt F. Prediction of boiling points of organic compounds from molecular descriptors by using backpropagation neural network. In: Carbo-Dorca R, Girones X, Mezey PG, editors. Fundamentals of molecular similarity. Mathematical and computational chemistry, Boston, MA: Springer US; 2001, p. 1–10. http://dx.doi.org/10.1007/978-1-4757-3273-3_1.

[17] Dai Y-m, Zhu Z-p, Cao Z, Zhang Y-f, Zeng J-l, Li X. Prediction of boiling points of organic compounds by QSPR tools. J Mol Graph 2013;44:113–9. http://dx.doi.org/10.1016/j.jmgm.2013.04.007.

[18] Zhokhova NI, Baskin II, Palyulin VA, Zefirov AN, Zefirov NS. Fragmental descriptors in QSPR: Flash point calculations. Russ Chem Bull 2003;52:1885–92. http://dx.doi.org/10.1023/B:RUCB.0000009629.38661.4c.

[19] Saldana DA, Starck L, Mougin P, Rousseau B, Pidol L, Jeuland N, et al. Flash point and cetane number predictions for fuel compounds using quantitative structure property relationship (QSPR) methods. Energy Fuels 2011;25(9):3900–8. http://dx.doi.org/10.1021/ef200795j.

[20] Gharagheizi F, Alamdari RF. Prediction of flash point temperature of pure components using a quantitative structure-property relationship model. QSAR Comb Sci 2008;27:679–83. http://dx.doi.org/10.1002/qsar.200730110.

[21] Pan Y, Jiang JC, Wang R, Jiang JJ. Predicting the net heat of combustion of organic compounds from molecular structures based on ant colony optimization. J Loss Prev Process Ind 2011;24:85–9. http://dx.doi.org/10.1016/j.jlp.2010.11.001.

[22] Gharagheizi F. A simple equation for prediction of net heat of combustion of pure chemicals. Chemometr Intell Lab Syst 2008;91:177–80. http://dx.doi.org/10.1016/j.chemolab.2007.11.003.

[23] Gharagheizi F, Mirkhani SA, Tofangchi Mahyari A-R. Prediction of standard enthalpy of combustion of pure compounds using a very accurate group-contribution-based method. Energy Fuels 2011;25:2651–4. http://dx.doi.org/10.1021/ef200081a.

[24] Roy K, Ambure P, Kar S. How precise are our quantitative structure–activity relationship derived predictions for new query chemicals? ACS Omega 2018;3:11392–406. http://dx.doi.org/10.1021/acsomega.8b01647.

[25] Tetko IV, Gasteiger J, Todeschini R, Mauri A, Livingstone D, Ertl P, et al. Virtual computational chemistry laboratory – Design and description. J Comput Aided Mol Des 2005;19(6):453–63. http://dx.doi.org/10.1007/s10822-005-8694-y.

[26] Yap CW. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. J Comput Chem 2011;32(7):1466–74. http://dx.doi.org/10.1002/jcc.21707.

[27] Masand VH, Rastija V. PyDescriptor: A New PyMOL plugin for calculating thousands of easily understandable molecular descriptors. Chemometr Intell Lab Syst 2017;169:12–8. http://dx.doi.org/10.1016/j.chemolab.2017.08.003.

[28] Moriwaki H, Tian Y-S, Kawashita N, Takagi T. Mordred: A molecular descriptor calculator. J Cheminformatics 2018;10:4. http://dx.doi.org/10.1186/s13321-018-0258-y.

[29] Le TT, Fu W, Moore JH. Scaling tree-based automated machine learning to biomedical big data with a feature set selector. Bioinformatics 2020;36(1):250–6.

[30] Olson RS, Urbanowicz RJ, Andrews PC, Lavender NA, Kidd LC, Moore JH. Automating biomedical data science through tree-based pipeline optimization. In: Squillero G, Burelli P, editors. Applications of evolutionary computation: EvoApplications 2016. Lecture notes in computer science, vol. 9597, Springer International Publishing; 2016, p. 123–37. http://dx.doi.org/10.1007/978-3-319-31204-0_9.

[31] Jin H, Song Q, Hu X. Auto-keras: An efficient neural architecture search system. 2019, arXiv:1806.10282.

[32] Chen C-H, Tanaka K, Kotera M, Funatsu K. Comparison and improvement of the predictability and interpretability with ensemble learning models in QSPR applications. J Cheminformatics 2020;12:19. http://dx.doi.org/10.1186/s13321-020-0417-9.

[33] Samek W, Montavon G, Lapuschkin S, Anders CJ, Müller K-R. Explaining deep neural networks and beyond: A review of methods and applications. Proc IEEE 2021;109(3):247–78. http://dx.doi.org/10.1109/JPROC.2021.3060483.

[34] He C, Ma M, Wang P. Extract interpretability-accuracy balanced rules from artificial neural networks: A review. Neurocomputing 2020;387:346–58. http://dx.doi.org/10.1016/j.neucom.2020.01.036.

[35] Montavon G, Samek W, Müller K-R. Methods for interpreting and understanding deep neural networks. Digit Signal Process 2018;73:1–15. http://dx.doi.org/10.1016/j.dsp.2017.10.011.

[36] Olah C, Mordvintsev A, Schubert L. Feature visualization. Distill 2017. http://dx.doi.org/10.23915/distill.00007.

[37] Ying X. An overview of overfitting and its solutions. J Phys Conf Ser 2019;1168:022022. http://dx.doi.org/10.1088/1742-6596/1168/2/022022.

[38] Rácz A, Bajusz D, Héberger K. Intercorrelation limits in molecular descriptor preselection for QSAR/QSPR. Mol Inform 2019;38. http://dx.doi.org/10.1002/minf.201800154.

[39] Das DD, St. John PC, McEnally CS, Kim S, Pfefferle LD. Measuring and predicting sooting tendencies of oxygenates, alkanes, alkenes, cycloalkanes, and aromatics on a unified scale. Combust Flame 2018;190:349–64. http://dx.doi.org/10.1016/j.combustflame.2017.12.005.

[40] Zhang C, Hui X, Lin Y, Sung C-J. Recent development in studies of alternative jet fuel combustion: Progress, challenges, and opportunities. Renew Sustain Energy Rev 2016;54:120–38. http://dx.doi.org/10.1016/j.rser.2015.09.056.

[41] Comesana A, Huntington T, Scown C, Niemeyer K, Rapp V. Berkeley Lab Feedstock to Function tool property database. http://dx.doi.org/10.5281/zenodo.5914847.

[42] Comesana A, Huntington T, Scown C, Niemeyer K, Rapp V. Berkeley Lab Feedstock to Function tool property models. http://dx.doi.org/10.5281/zenodo.6383369.

[43] Liu J, Gong S, Li H, Liu G. Molecular graph-based deep learning method for predicting multiple physical properties of alternative fuel components. Fuel 2021;122712. http://dx.doi.org/10.1016/j.fuel.2021.122712.

[44] Gharagheizi F, Mirkhani SA, Ilani-Kashkouli P, Mohammadi AH, Ramjugernath D, Richon D. Determination of the normal boiling point of chemical compounds using a quantitative structure–property relationship strategy: Application to a very large dataset. 354:250–258. http://dx.doi.org/10.1016/j.fluid.2013.06.034.

[45] Bradley J-C, Williams A, Lang A. Jean-Claude Bradley open melting point dataset. 2014, http://dx.doi.org/10.6084/m9.figshare.1031637.v2.

[46] Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem 2019 update: Improved access to chemical data. Nucleic Acids Res 2018;47(D1):D1102–9. http://dx.doi.org/10.1093/nar/gky1033.

[47] Kessler T, Mack JH. Ecnet: Large scale machine learning projects for fuel property prediction. J Open Source Softw 2017;2:401. http://dx.doi.org/10.21105/joss.00401.

[48] Yaws CL. Chapter 1 - Physical properties – Organic compounds. In: Yaws CL, editor. The yaws handbook of physical properties for hydrocarbons and chemicals. 2nd ed.. Boston: Gulf Professional Publishing; 2015, p. 1–683. http://dx.doi.org/10.1016/B978-0-12-800834-8.00001-3.

[49] Das D, McEnally C, Kwan T, Zimmerman J, Cannella W, Mueller C, et al. Sooting tendencies of diesel fuels, jet fuels, and their surrogates in diffusion flames. Fuel 2017;197:445–58. http://dx.doi.org/10.1016/j.fuel.2017.01.099.

[50] McEnally CS, Das DD, Pfefferle LD. Yield Sooting Index Database Volume 2: Sooting Tendencies of a Wide Range of Fuel Compounds on a Unified Scale. 2017, http://dx.doi.org/10.7910/DVN/7HGFT8.

[51] Fisher Scientific. Material Safety Data Sheet- Fisher SCI, https://fscimage.fishersci.com/msds/96461.htm.

[52] McKinney W. Data Structures for Statistical Computing in Python. In: van der Walt S, Millman J, editor. Proceedings of the 9th Python in science conference. 2010. p. 51–6.

[53] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res 2011;12:2825–30.

[54] Landrum G. RDKit: Open-source cheminformatics. http://dx.doi.org/10.5281/zenodo.591637. http://www.rdkit.org.

[55] Hollas B. An analysis of the autocorrelation descriptor for molecules. J Math Chem 2003;33(2):91–101. http://dx.doi.org/10.1023/a:1023247831238.

[56] Xue Y, Li H, Ung CY, Yap CW, Chen YZ. Classification of a diverse set of tetrahymena pyriformis toxicity chemical compounds from molecular descriptors by statistical learning methods. Chem Res Toxicol 2006;19:1030–9. http://dx.doi.org/10.1021/tx0600550.

[57] Bahl A, Hellack B, Balas M, Dinischiotu A, Wiemann M, Brinkmann J, et al. Recursive feature elimination in random forest classification supports nanomaterial grouping. NanoImpact 2019;15:100179. http://dx.doi.org/10.1016/j.impact.2019.100179.

[58] Spearman rank correlation coefficient. In: The concise encyclopedia of statistics. New York, NY: Springer New York; 2008, p. 502–5. http://dx.doi.org/10.1007/978-0-387-32833-1_379.

[59] Louppe G, Wehenkel L, Sutera A, Geurts P. Understanding variable importances in forests of randomized trees. Adv Neural Inf Process Syst 2013;26.

[60] Darst BF, Malecki KC, Engelman CD. Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. BMC Genet 2018;19(S1). http://dx.doi.org/10.1186/s12863-018-0633-8.

[61] Meinshausen N. Quantile regression forests. J Mach Learn Res 2006;7:983–99.

[62] Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Mach Learn 2006;63(1):3–42. http://dx.doi.org/10.1007/s10994-006-6226-1.

[63] Setiawan R, Mohammadinia S. Toward estimating standard enthalpy of combustion of pure chemical compounds: Extreme learning machine approach. Energy Sources A 2021;1–9. http://dx.doi.org/10.1080/15567036.2021.1917730.

[64] Frutiger J, Marcarie C, Abildskov J, Sin G. A comprehensive methodology for development, parameter estimation, and uncertainty analysis of group contribution based property models—An application to the heat of combustion. J Chem Eng Data 2016;61(1):602–13. http://dx.doi.org/10.1021/acs.jced.5b00750.

[65] Ojha P, Mitra I, Das R, Roy K. Further exploring rm2 metrics for validation of QSPR models. Chemometr Intell Lab Syst 2011;107:194–205.

[66] Roy PP, Roy K. On some aspects of variable selection for partial least squares regression models. QSAR Comb Sci 2008;27:302–13. http://dx.doi.org/10.1002/qsar.200710043.

[67] Estrada E, Torres L, Rodriguez L, Gutman I. An atom-bond connectivity index: Modelling the enthalpy of formation of alkanes. Indian J Chem 1998;37A:849–55, URL http://nopr.niscair.res.in/handle/123456789/40308.
[68] Todeschini R, Consonni V. Hand book of molecular descriptors. Methods and principles in medicinal chemistry, 1st ed.. Wiley; 2000, http://dx.doi.org/10.1002/9783527613106.
[69] Mauri A, Consonni V, Todeschini R. Molecular descriptors. In: Leszczynski J, Kaczmarek-Kedziera A, Puzyn T, G. Papadopoulos M, Reis H, K. Shukla M, editors. Hand book of computational chemistry. Cham: Springer International Publishing; 2017, p. 2065–93. http://dx.doi.org/10.1007/978-3-319-27282-5_51.
[70] Katritzky AR, Jain R, Lomaka A, Petrukhin R, Maran U, Karelson M. Perspective on the relationship between melting points and chemical structure. Cryst Growth Des 2001;1:261–5. http://dx.doi.org/10.1021/cg010009s.
[71] Thurston BA, Ferguson AL. Machine learning and molecular design of self-assembling-conjugated oligopeptides. Mol Simul 2018;44(11):930–45. http://dx.doi.org/10.1080/08927022.2018.1469754.
[72] Labute P. A widely applicable set of descriptors. J Mol Graph 2000;18:464–77. http://dx.doi.org/10.1016/S1093-326300068-1.
[73] Guha R, Willighagen E. A survey of quantitative descriptions of molecular structure. Curr Top Med Chem 2012;12(18):1946–56. http://dx.doi.org/10.2174/156802612804910278.
[74] Johnson-Restrepo B, Pacheco-Londoño L, Olivero-Verbel J. Molecular parameters responsible for the melting point of 1, 2, 3-Diazaborine compounds. J Chem Inf Comput Sci 2003;43(5):1513–9. http://dx.doi.org/10.1021/ci020387k.
[75] Spowage BM, Bruce CL, Hirst JD. Interpretable correlation descriptors for quantitative structure-activity relationships. J Cheminformatics 2009;1(1):22. http://dx.doi.org/10.1186/1758-2946-1-22.
[76] Charton M. The nature of topological parameters. I. Are topological parameters 'fundamental properties'? J Comput Aided Mol Des 2003;17:197–209. http://dx.doi.org/10.1023/A:1025378125128.
[77] Devinyak O, Havrylyuk D, Lesyk R. 3D-MoRSE descriptors explained. J Mol Graph 2014. http://dx.doi.org/10.1016/j.jmgm.2014.10.006.
[78] Patil GS. Estimation of flash point. Fire Mater 1988;12:127–31. http://dx.doi.org/10.1002/fam.810120307.
[79] Liu X, Liu Z. Research progress on flash point prediction. J Chem Eng Data 2010;55(9):2943–50. http://dx.doi.org/10.1021/je1003143.
[80] Gülder OL. Influence of hydrocarbon fuel structural constitution and flame temperature on soot formation in laminar diffusion flames. Combust Flame 1989;78(2):179–94. http://dx.doi.org/10.1016/0010-2180(89)90124-7.
[81] Burden FR. Molecular identification number for substructure searches. J Chem Inf Model 1989;29:225–7. http://dx.doi.org/10.1021/ci00063a011.
[82] Information Resources Management Association, editor. Nanotechnology: Concepts Methodologies, Tools, and Applications. IGI Global; 2014, http://dx.doi.org/10.4018/978-1-4666-5125-8.
[83] Roy K, Das RN. On extended topochemical atom (ETA) indices for QSPR studies. In: Nanotechnology. IGI Global; 2014, p. 841–73. http://dx.doi.org/10.4018/978-1-4666-5125-8.ch037.
[84] Das K, Gutman I, Furtula B. On atom-bond connectivity index. Filomat 2012;26:733–8. http://dx.doi.org/10.1016/j.cplett.2011.06.049.
[85] Md Said MR, Mohammed M, Atan K, Khalaf AJM, Hasni R, Nawawi A. Atom bond connectivity index of molecular graphs of alkenes and cycloalkenes. J Comput Theor Nanosci 2017;14:5011–9. http://dx.doi.org/10.1166/jctn.2017.6912.