Fuel 408 (2026) 137588

Contents lists available at ScienceDirect

Fuel

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

Full length article

A structured framework for predicting sustainable aviation fuel properties using liquid-phase FTIR and machine learning

Ana E. Comesana[a], Sharon Chen[a], Kyle E. Niemeyer[b], Vi H. Rapp[a,*]

[a] Energy Technologies Area, Lawrence Berkeley National Laboratory, Berkeley, CA, United States
[b] School of Mechanical, Industrial and Manufacturing Engineering, Oregon State University, Corvallis OR, United States

HIGHLIGHTS

- We created a method for using liquid-phase FTIR with machine learning models.
- Non-negative Matrix Factorization preserves spectral attributes for interpretation.
- We developed 5 property prediction models for SAFs and SAF blends.
- Models presented outperform or are aligned with previously developed models.
- The method supports discovering new relationships between chemicals and properties.

ARTICLE INFO

Keywords:
Fourier transform infrared spectroscopy
Machine learning
Sustainable aviation fuels
Non-negative matrix factorization

ABSTRACT

Sustainable aviation fuels have the potential to improve efficiency, reduce emissions, and enhance energy security. To help identify viable sustainable aviation fuels and accelerate research, machine learning models have been developed to predict relevant physicochemical properties. However, many models have limited applicability, leverage data from complex analytical techniques with confined spectral ranges, or use feature decomposition methods that offer limited interpretability. Using liquid-phase Fourier Transform Infrared (FTIR) spectra, this study presents a structured method for creating accurate and interpretable property prediction models for neat molecules, aviation fuels, and blends. Liquid FTIR spectra can be collected quickly and consistently, offering high reliability, sensitivity, and component specificity using less than 2 ml of sample. The method first decomposes FTIR spectra into fundamental building blocks using non-negative matrix factorization (NMF) to enable scientific analysis of FTIR spectra attributes and fuel properties. The NMF features are then used to create five ensemble models for predicting final boiling point, flash point, freezing point, density at 15 °C, and kinematic viscosity at −20 °C. All models were trained using experimental property data from neat molecules, aviation fuels, and blends. The models accurately predict key properties across a broad range of neat molecules and representative fuels and blends, while enabling interpretation of relationships between compositional elements, such as functional groups or chemical classes, and their resulting properties. This demonstrates strong potential to support sustainable aviation fuel research and development. The models and data are available on an interactive web tool.

1. Introduction

Sustainable aviation fuels (SAFs) offer the opportunity to improve efficiency, reduce emissions, and enhance energy security by diversifying the energy supply. Development and use of SAFs are challenging because fuels must meet strict safety, performance, and chemical composition standards. SAFs must also be fully compatible with petroleum-derived jet fuels, since all certified synthetic blending components are required to be blended with petroleum jet fuel prior to use [1–3].

Given the stringent requirements and high costs associated with SAF development, researchers have explored a range of modeling strategies to predict SAF performance before production, aiming to reduce experimental effort, lower costs, and minimize risks throughout the process. For example, Heyne et al. present a multi-tiered modeling and

* Corresponding author.
Email address: vhrapp@lbl.gov (V.H. Rapp).

https://doi.org/10.1016/j.fuel.2025.137588
Received 4 June 2024; Received in revised form 29 October 2025; Accepted 15 November 2025
Available online 8 December 2025

screening approach to predict and evaluate the physical and chemical properties of SAF candidates, using both bulk property correlations and molecular-level composition analysis [3]. They found that combining complementary modeling methods enhances the reliability of property predictions, enabling more efficient identification of viable SAFs that meet stringent aviation fuel specifications.

Various modeling approaches have been used to predict physicochemical properties of fuels (both petroleum-derived and certified/viable SAFs), blendstocks (chemicals combined with other chemicals to create fuel), neat molecules, and blends of these. Simple models that use linear-by-volume or linear-by-molar-mass calculations have been employed to predict physicochemical properties; however, these approaches can be inaccurate for blends exhibiting non-ideal mixing behavior, thereby limiting their applicability [4–6]. Models that correlate chemical descriptors or functional-group counts with properties are more sophisticated but may not fully capture the complex interactions within a fuel blend, leading to prediction inaccuracies [7–23].

Due to these limitations, researchers developed machine learning models that use different analytical techniques to characterize fuels and fuel blends, such as nuclear magnetic resonance (NMR), gas chromatography (GC), and mass spectrometry (MS) [24–27]. These analytical techniques can characterize neat molecules and blends by providing insights into the functional groups, bonds, and component classes present. Many techniques are also able to differentiate between isomers and provide fundamental chemical information about a fuel's components that can be used as inputs for machine learning models [24]. For example, one study used liquid chromatography/1H nuclear magnetic resonance (LC/1H NMR) with multiple regression analysis to develop a group property approach, similar to functional group theory, for predicting properties of distillate fuels, including flash point, density, and final boiling point [25]. Another study predicted similar properties using partial least squares with near-infrared (NIR) spectroscopy, Raman spectroscopy (RM), and gas chromatography with mass selective detection (GC-MS) [26]. Two-dimensional GC (or GC × GC) and Monte Carlo sampling have also been used to predict properties of SAFs [27]. Others evaluated four spectroscopic techniques—NMR spectroscopy, IR spectroscopy, Raman spectroscopy, and gas chromatography—for correlating fuel composition with physicochemical properties including density, viscosity, flash point, and freezing point. Their results indicate that GC × GC yields the most accurate correlations among all analytical instruments [28].

Fourier-transform infrared (FTIR) spectroscopy is another popular technique for developing predictive models because it offers rapid data acquisition, component specificity, and sensitivity [4–6,29–36]. Many researchers have developed property prediction models using a limited range of vapor-phase FTIR spectra. For example, Wang et al. used a limited region of vapor-phase FTIR spectra with LASSO-regularized linear models to predict physical and chemical properties including molecular weight, density, and initial boiling point of fuels [5]. Expanding on this work, they later added functional group theory information for a selected region of the full FTIR spectra to their LASSO-regularized models to predict the same properties for hydrocarbon fuels [30]. Boddapati et al. used a limited region of vapor-phase FTIR spectra and elastic-net regularized linear models to predict similar properties of pure hydrocarbons [4]. In addition, liquid-phase FTIR spectra have been used with partial least squares models to determine cetane number, net heat of combustion, viscosity, carbon-to-hydrogen ratio, and density of middle distillate fuels [31].

Although the literature presents well-performing FTIR-based models for predicting various fuel properties, the methods have several limitations [4–6,29–36]. For example, some studies developed predictive models using narrow FTIR spectral ranges, which may omit important information necessary for accurately predicting a variety of fuel properties [4–6,30]. Additionally, measuring vapor-phase FTIR spectra is time consuming and complicated since any variations in temperature or pressure in the sample lines or gas cell will result in significant errors [37]. Further, many of the models do not provide an easily interpretable analysis of the spectral features (e.g., how much a feature contributes to the predictions) and use non-standard error metrics, making it difficult to assess the model's performance on new data [4–6].

The purpose of this study is to develop a structured approach for creating accurate and interpretable property prediction models for viable SAFs and SAF blendstocks using liquid-phase FTIR spectra of neat molecules, petroleum-based jet fuels, synthetic and alternative aviation fuels, and their blends. Liquid-phase FTIR spectra are used because they can be collected quickly and consistently, requiring only a few drops of fuel. The structured approach uses non-negative matrix factorization (NMF) to deconstruct FTIR spectra into fundamental building blocks that enable scientific analysis of spectral characteristics and fuel properties. Because of its non-negativity constraints, NMF is well-suited for spectral data and can create meaningful, additive features that are easy to interpret. Using the NMF features, we developed five ensemble models for predicting final boiling point, flash point, freezing point, density at 15 °C, and kinematic viscosity at −20 °C. In the following sections, we outline the methods for creating features from FTIR spectra using NMF and developing models with these features to predict fuel properties. Next, model performance metrics are presented, along with an analysis and interpretation of the features used in the models. Conclusions and future recommendations are also provided.

2. Methods

2.1. Experimental property data

Experimental property data for neat molecules, fuels, and blends were collected from multiple published sources [7,38–48]. Data include alkanes, isoalkanes, cycloalkanes, aromatics, alcohols, conventional fossil-based fuels (e.g., Jet-A, F-24), American Society for Testing and Materials (ASTM) approved SAFs (e.g., HEFA, ATJ, FT-SPK, and farnesane), and other viable SAF blendstocks, including 1,4-dimethylcyclooctane (DMCO), pinane, and limonene. Property data for F-24 were provided by the supplier, while a local certification lab measured properties for Jet A, HEFA, and their blends. Properties for other, neat-molecule blends that include n-heptane and isooctane were calculated using the linear-by-volume method. Because ethanol blending with n-heptane and/or isooctane follows linear property trends with low volumes of ethanol, only blends with less than 10 % ethanol were calculated using the linear-by-volume method [49].

Table 1 shows the number of neat molecules, fuels, and blends with experimental data for each property. Data for the 87 neat molecules, fuels, and blends used in this study can be found in Tables S3 and S4, and on the Feedstock to Function Website: feedstock-to-function.lbl.gov. Although some molecules and blends in the dataset—such as those containing oxygen—are not suitable for aviation use, we included them in model training to better characterize bond interactions that may occur in blends. For example, FTIR spectra of ethanol and 1-pentene show peaks in the 3000 cm^{-1} to 2850 cm^{-1} range (see Fig. 1), corresponding to C−H stretching vibrations; excluding ethanol would reduce the model's ability to recognize these interactions and degrade its performance. Moreover, developing the model with a more diverse fuel dataset increases its robustness and may aid in identifying the impact of contaminants during processing or transportation.

2.2. Liquid-phase FTIR spectra measurements

We collected FTIR spectra of 87 neat molecules, fuels, and blends from published literature (62) and measurements (25) using the method described by Daly et al. [32,33] with minor modifications. We modified the method to ensure spectra collected were consistent, repeatable, and matched spectra from previous literature [32,33]. Specifically, spectra of molecules, fuels, and blends were collected using a Thermo Scientific Nicolet iS50 FTIR spectrometer located at Berkeley Lab's Molecular

Table 1
Number of samples with experimental data for each of the properties. Note that density is measured at 15 °C and kinematic viscosity is measured at −20 °C. (See Tables S4 and S5 for details).

	Final boiling point	Flash point	Freezing point	Density	Kinematic Viscosity
Neat molecules	42	29	40	38	17
Fuels	4	4	4	5	4
Fuel blends	4	4	4	4	4
Other blends	22	22	22	22	6
	72	59	70	69	31

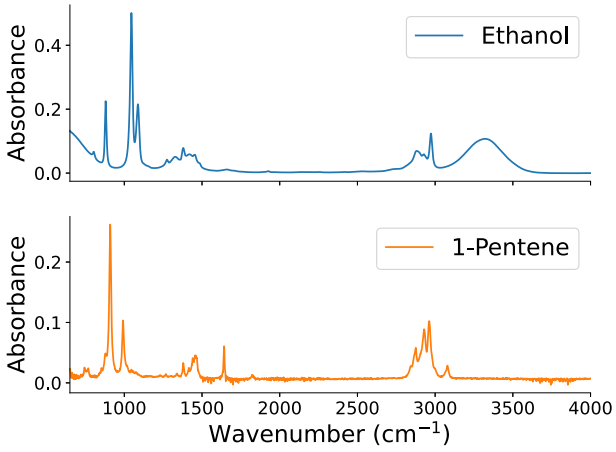


Fig. 1. Spectra of ethanol and 1-pentene.

Foundry with a Smart iTR Attenuated Total Reflectance sampling accessory containing diamond crystal with ZnSe lens. Viton square-profile o-rings with 316 stainless steel tags prevented the liquid fuels from spreading and evaporating during spectral measurements. Spectra were collected between 4000 cm^{-1} and 650 cm^{-1} at a 2 cm^{-1} resolution. Further details about the fuels and spectral measurements are provided in Section S-1 of the Supplemental Information (SI).

2.3. Spectra data Preprocessing and dimensionality reduction

FTIR spectra require preprocessing to minimize noise and variability between measurements before they can be used to train machine learning models [26,31,50]. We used Python libraries (Pandas and NumPy) to implement our spectral processing pipeline, which involved organizing the spectra into consistent ranges (binning), smoothing random fluctuations, adjusting baselines, clipping negative values, and scaling (normalizing) the spectra to enable integration into a unified database [51,52]. Because bin ranges vary slightly between instruments, we created wider bins to easily combine measurements at similar wavelengths. For example, creating a bin ranging from 650.4 cm^{-1} to 649.9 cm^{-1} combines the published spectra sampled at 650.162 cm^{-1} and measured data sampled at 650.145 cm^{-1} so they could be input into the machine learning model. We then applied a moving average with a window width of 15 bins to smooth experimental noise in the spectra while still retaining the shape of major peaks.

To remove background noise caused by mechanical vibrations or other external factors, we uniformly shifted the baselines of spectra by the mean of the values between 4000 cm^{-1} and 3600 cm^{-1} [26,50]. We selected this range because it does not contain absorption bands for fuels or compounds of interest in this study, providing a good proxy for background noise. Remaining negative values were clipped to zero and the data between 2500 cm^{-1} and 2000 cm^{-1} were removed because diamond crystals cause absorption bands in this region. We normalized the spectra by their sum to ensure that peak magnitudes for all spectra remain comparable. Eq. (1) shows the calculation used to normalize a spectrum, with X_{wn} representing the spectrum of interest at wavenumber *wn* and X_i representing the spectrum at wavenumber *i*:

$$\text{Normalized } X_{wn} = \frac{X_{wn}}{\sum_{i=\min}^{max} X_i} \tag{1}$$

The final shape and magnitude of the normalized spectra used for model training reflect the proportions of bonds, functional groups, and component classes [5].

After processing the spectra, we used `scikit-learn`'s `NMF` class to reduce the complexity of the spectra and identify important features for our model [53]. NMF works by breaking down the spectra into components that represent underlying chemical structures in the FTIR spectra. The contributions of these components to each sample (coefficients) were used as features to predict properties.

NMF decomposes the spectra into the product of two lower-rank matrices, $\mathbf{W}$ and $\mathbf{H}$, such that $\mathbf{X} \approx \mathbf{W} \times \mathbf{H}$, where $\mathbf{X}$ represents the original spectra dataset [54,55]. This decomposition is achieved through an iterative process that modifies the values within the $\mathbf{W}$ and $\mathbf{H}$ matrices to create the best approximation of the entire dataset [56]. NMF is a non-convex optimization problem with no closed-form solution, meaning that the optimal solution cannot be found using a simple formula and requires numerical methods. The process starts by randomly initializing two matrices, $\mathbf{W}$ and $\mathbf{H}$, that serve to factorize the spectra dataset. $\mathbf{H}$ is the coefficient matrix, representing the original spectra samples as additive linear combinations of features in $\mathbf{W}$, which is the basis matrix containing the non-negative features of the spectra. In other words, each row of $\mathbf{H}$ represents the weights or coefficients of the features in $\mathbf{W}$ that are used to reconstruct the corresponding spectra samples. The matrices $\mathbf{W}$ and $\mathbf{H}$ are then iteratively updated using an optimization algorithm, such as the multiplicative update rules, to minimize the reconstruction error between $\mathbf{W} \times \mathbf{H}$ and the original spectra dataset. For a given update n and original spectra dataset matrix $\mathbf{X}$, the updates for $\mathbf{W}$ and $\mathbf{H}$ are computed using

$$\mathbf{W}^n = \mathbf{W}^{n-1} \circ \frac{\mathbf{X}(\mathbf{H}^{n-1})^\top}{\mathbf{W}^{n-1}\mathbf{H}^{n-1}(\mathbf{H}^{n-1})^\top} \quad \text{and} \tag{2}$$

$$\mathbf{H}^n = \mathbf{H}^{n-1} \circ \frac{{\mathbf{W}^{n-1}}^\top \mathbf{X}}{({\mathbf{W}^{n-1}})^\top \mathbf{W}^{n-1}\mathbf{H}^{n-1}}. \tag{3}$$

This update is repeated until the reconstruction error (calculated as the Frobenius norm between $\mathbf{W} \times \mathbf{H}$ and the original spectra dataset $\mathbf{X}$) converges. At the end of the optimization, $\mathbf{W}$ represents the basis components and contains non-negative features of the spectra dataset, while $\mathbf{H}$ contains the coefficients for reconstructing the original spectra dataset, and also represents the original spectra samples as additive linear combinations of features in $\mathbf{W}$. This process decomposes the spectra into features that preserve spectral characteristics, making them easy to interpret [54].

The resulting features can be used to reconstruct the spectra and were used for the models. The non-negativity constraint imposed on the matrices $\mathbf{W}$ and $\mathbf{H}$ ensures that the features extracted are meaningful and interpretable, as they represent the presence or absence of

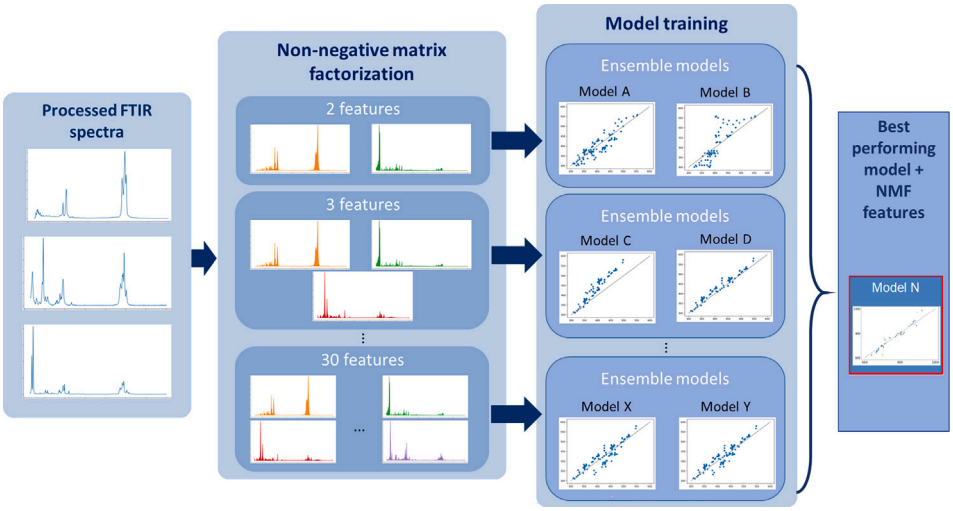


Fig. 2. Schematic of the model development process.

specific spectral components. Additionally, the dimensionality reduction achieved through NMF helps reduce the noise and complexity of the spectra, making it more suitable for modeling and analysis. Because NMF does not rely on property labels to decompose the spectra, we can include more samples in the NMF training set, even if no property data are available for them (i.e., unlabeled data). This is a significant advantage, as it allows us to expand the spectra dataset without needing property information for every sample. By including spectra from other relevant molecules and blends without known property data, NMF can identify common spectral features that might be missed if we only used samples with known properties. This improves the robustness of the model and helps prevent overfitting. The spectra without direct property data used to enhance our NMF model include blends with diisobutylene, n-heptane, 1-hexene, isooctane, 2-methylbutane, methylcyclohexane, toluene, and xylene. We split the property data into training (68 %), validation (12 %), and testing (20 %) sets, and fit the NMF models to the training data using the `NMF` class from `scikit-learn.decomposition` [53]. The validation set was used for hyperparameter tuning and model selection, while the test set was used to estimate the error of the whole pipeline on new data. Each model is trained on 4–7 fuels and fuel blends, with one blend held out for testing. This split evaluates whether the model can generalize to a new fuel whose spectrum is well represented by the training mixtures, while keeping training and test samples completely separate. Under NMF, spectra are modeled as additive combinations of basis features, so performance on the held-out blend reflects interpolation within the spectral-compositional space covered by the training data.

We trained the NMF models with up to 30 components for each property, using ensemble models as predictors, and then selected the number of components that yielded the best performance on the validation set. We also found that randomizing NMF initialization does not significantly alter the components or the number of components selected for the models. When varying the random state used for initialization, the final number of components and components themselves remain stable. This suggests that the NMF algorithm reaches stable solutions across different initializations, indicating that it likely converges to a global or well-defined local minimum.

2.4. Model development using machine learning

To train the models, we employed NMF to generate groups of 2 to 30 components, and used the resulting coefficient vectors as feature representations for each sample. For kinematic viscosity, groups of up to 15 features were created to account for the smaller property dataset size and to avoid overfitting. We then fit `ExtraTreesRegressor` and `RandomForestRegressor` models from the `sklearn.ensemble` module to the training set with a grid search over hyperparameters including number of estimators, maximum tree depth, and minimum samples per split [53]. Ensemble models aggregate the predictions of multiple decision trees to improve accuracy and reduce overfitting. Each tree in the ensemble is trained on a bootstrapped subset of the data, and the final prediction is the average of the individual tree outputs. This averaging process reduces the variance of the model compared to individual decision trees. We selected these regressors because of their high interpretability, and their ability to provide quantile-based predictions that allow us to inspect the distribution of individual tree outputs during development, helping evaluate model stability and potentially guide hyperparameter optimization [57]. All model training and hyperparameter tuning were performed using `scikit-learn` utilities [53]. The validation set was used to optimize the number of NMF components, ensemble model architecture, and hyperparameters. For each property, we selected the number of features and model architecture that yielded the best performance on the validation set. We then used the test set to estimate the model's error on new data. Fig. 2 shows a schematic of the model development process.

3. Results and discussion

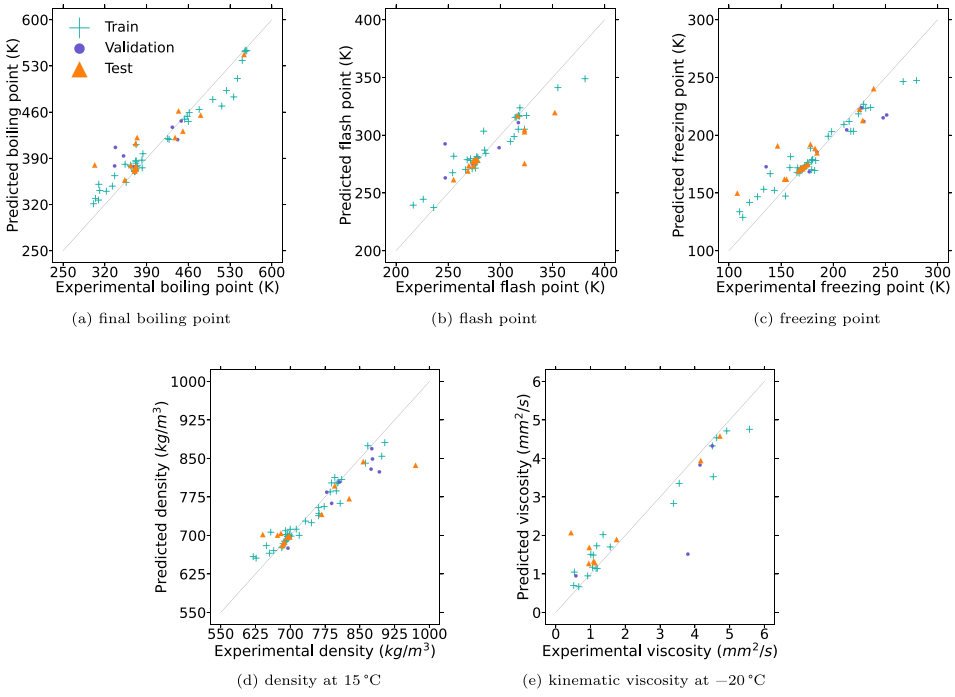
3.1. Model performance

To validate this approach, we used the method described in Section 2.4 to train models for final boiling point, flash point, freezing point, kinematic viscosity, and density of neat molecules, fuels, and blends. Table 2 summarizes each model's characteristics and test performance, while Fig. 3 shows parity plots for the models. Additional performance parameters such as training errors and overall model errors are provided in Section S-2 of the SI. Even though data used in this study are more diverse and include a larger range of property values, molecule types, and blends, the models perform consistently with other FTIR spectra-based property prediction models [4,25,26,30,31]. Specifically, the R^2 for boiling point is 0.89 (published literature: 0.73–0.91), flash point 0.8 (published literature: 0.71–0.8), density 0.88 (published literature: 0.96), and viscosity 0.86 (published literature: 0.58–0.95). Because of variations in data and methods for reporting and calculating error metrics, we are unable to directly compare our results with all of the studies, and are unable to compare metrics for freezing point. In addition, the models in this work exhibit higher accuracy on the blend samples in the test set than on the non-blend

Table 2
Characteristics and performance for the five property prediction models. Density is measured at 15 °C and viscosity is measured at −20 °C.

	Final boiling point (K)	Flash point (K)	Freezing point (K)	Density (kg/m^3)	Kinematic viscosity (mm^2/s)
Test MAE	17.17	9.6	8.9	22.2	0.45
Test RMSE	26.39	16.7	15.9	41	0.66
Test MedAE	6.25	3.51	2.65	2.91	0.23
Test MAPE	4.6 %	3 %	6.1 %	2.7 %	66.8*%
R^2	0.89	0.80	0.84	0.88	0.86
Number of features	16	8	15	9	10
Total sample count	72	59	70	69	31
Property range	300–559	216–382	108–278	620–971	0.4–5.6

*Most viscosity data lie between 0 and 2 mm^2/s.



(a) final boiling point (b) flash point (c) freezing point

(d) density at 15 °C (e) kinematic viscosity at −20 °C

Fig. 3. Parity plots showing experimental and predicted values of all models, split by training, validation, and testing sets.

samples, indicating that their applicability domain aligns well with the method's intended objective. Given the limited number of certified fuels and blends with measured property data, these results highlight the model's consistent behavior across the available real-fuel samples, and its potential for broader validation as additional data become available.

Further, the results show that the models are capable of accurately predicting properties when trained on clustered property data. For example, the test set Mean Absolute Error (MAE) for the viscosity model is 0.45 mm^2/s despite exhibiting two clusters between 0 and 2 and above 3.6 (Fig. 3e). The viscosity model could be improved with additional experimental property data, particularly between the two clusters. Mean Absolute Percentage Error (MAPE) does not seem to be an appropriate performance metric for kinematic viscosity, since more than one quarter of the viscosity data falls between 0 and 1, which inflates this metric due to division by smaller values (MAPE = mean(|experimental − predicted|/experimental)). For density, nearly 63 % of the test data has an error below the test MAE. Although the density model performs comparably to literature models, the MAE of 22 kg/m^3 is due to 1,2,3,4-tetrahydronaphthalene, which has the largest density in the dataset (970 kg/m^3 and nearly 100 kg/m^3 larger than the next largest compound). Without this compound, the model MAE decreases nearly 36 % (~ 14 kg/m^3), indicating that additional data is needed to improve accuracy of this model for higher density molecules that may occur in SAF blends. High-density molecules like 1,2,3,4-tetrahydronaphthalene may occur in jet fuel blends only in limited quantities, since ASTM D1655 requires the finished fuel to have a density between 775 and 840 kg/m^3 at 15 °C [2]. The NMF model approach enables greater flexibility to capture appropriate features provided by additional data and interpret how those features contribute to variations in fuel properties.

To ensure the model did not overfit to the training data, we performed independent two-sample *t*-tests to compare the error distributions across the training, validation, and test sets for each property. The *t*-tests assessed whether there were statistically significant differences between the errors on the training and validation sets, as well as between the validation and test sets. A significant difference in error distributions—particularly between the test set and either the training or validation sets—would suggest potential overfitting, and limit generalization to new data. The *t*-test results (see Table 3) show that there are no statistically significant differences in errors across these sets for any property, as all *p*-values are greater than 0.05. The models perform consistently across the training, validation, and test sets, indicating good generalization to new data and a low risk of overfitting.

Table 3
P-values for *t*-tests comparing test errors to train and validation errors for each property.

	Train-test	Validation-test
Final boiling point	0.45	0.61
Flash point	0.53	0.77
Freezing point	0.80	0.15
Density	0.11	0.83
Viscosity	0.44	0.45

3.2. Model features

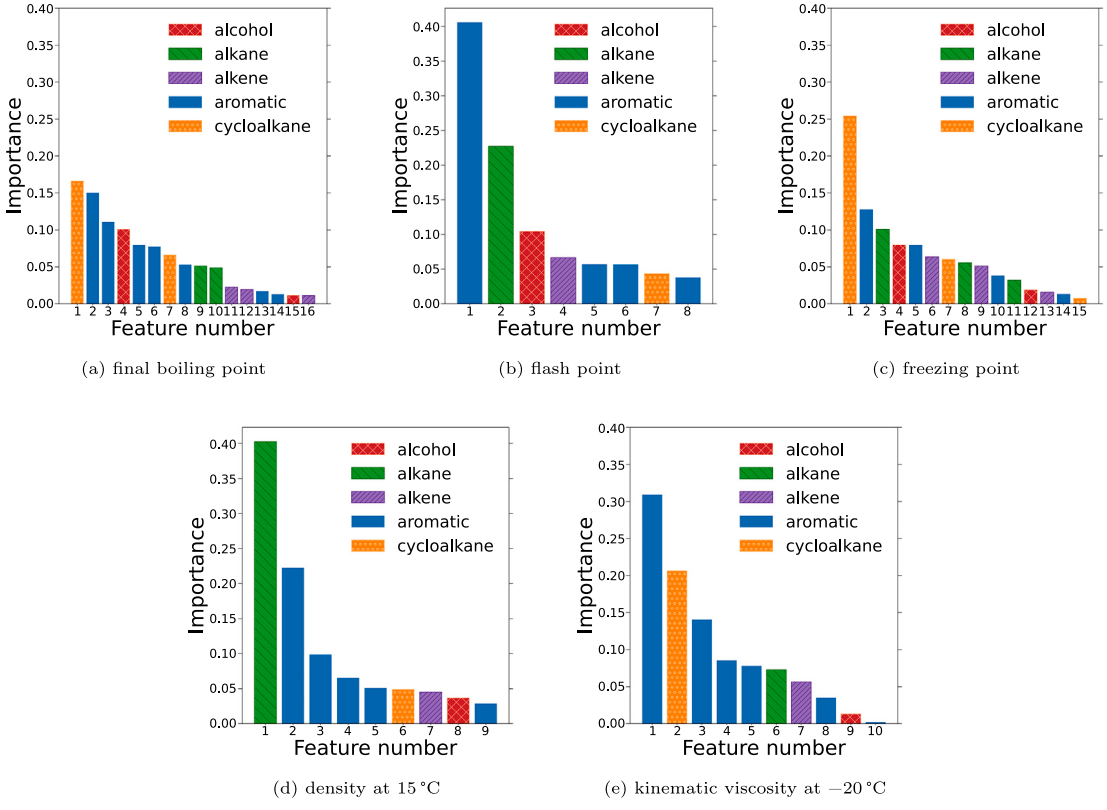
In addition to creating accurate predictive models, this method also provides interpretable features that highlight possible attributes most affecting each property. For example, Fig. 4 shows the types of features used in each property model and their importances. Components are grouped by compound class using identified spectral bands and regions associated with alcohols, alkanes, alkenes, aromatics, and cycloalkanes. The NMF-derived features capture recurring spectral patterns associated with key structural motifs such as aromatics, branched chains, and cyclic hydrocarbons. Because these spectral-structural relationships correspond to compositional aspects that govern fuel behavior across both molecular and blended systems, the learned feature-property correlations remain physically meaningful and transferable beyond the specific composition space of the training data. Feature importance is calculated by the model and based on the reduction in variance when a feature is used to split data at each node in the trees [53]. For each tree, the importance of a feature is determined as the sum of variance reduction across all nodes where the feature is used, weighted by the proportion of samples reaching those nodes. The final feature importances for the ensemble are the average of these values across all trees in the forest, normalized to sum to 1.

Fig. 4 shows that most models have between one and three dominant features, with the exception of final boiling point, which contains a more diverse set of equally important features. The top two features in the flash point model (Fig. 4b) contain spectral peaks similar to aromatics and alkanes with an aggregated feature importance of about 0.64. For the freezing point model (Fig. 4c), the highest-ranked feature contains cycloalkane attributes and has an importance of 0.25, about twice the importance of the second ranked feature. The density model's highest ranking features contain alkane and aromatic characteristics, with the alkane feature having an importance of 0.4 (see Fig. 4d). The top three features in the viscosity model (Fig. 4e) have an aggregated feature importance of over 0.65 and contain spectral peaks similar to aromatics and cycloalkanes.

All models rank features with aromatic characteristics either first or second. Interestingly, Fig. 5 shows that the top aromatic feature in the flash point and viscosity models is almost indistinguishable, exhibiting a prominent peak near 770 cm^{-1}, characteristic of terminal ethyl groups [58]. Several other out-of-plane C–H bending bands associated with aromatics can also be seen in the region 900–675 cm^{-1} [59]. Smaller peaks representing C–C in-ring stretching are located over 1500–1400 cm^{-1} and 1600–1585 cm^{-1} [59].

The top aromatic features match between the freezing point and density models, which are also the second-highest ranking features in both models (see Fig. 6). Both features contain prominent peaks between 800–650 cm^{-1} representing out-of-plane C–H bands [59]. Other peaks near 1600 cm^{-1} and from 1500–1400 cm^{-1} represent carbon-carbon stretching vibrations in the aromatic rings [59]. These features also include a prominent peak around 1244 cm^{-1}, likely representing C–C stretching from terminal tert-butyl groups [58].

Features containing cycloalkane attributes are ranked either first or second in the final boiling point, freezing point, and viscosity models. Similar to the aromatic features, the top cycloalkane features in the final boiling point and viscosity models exhibit nearly identical peaks (see



(a) final boiling point (b) flash point (c) freezing point

(d) density at 15 °C (e) kinematic viscosity at −20 °C

Fig. 4. Model feature importances for all models.

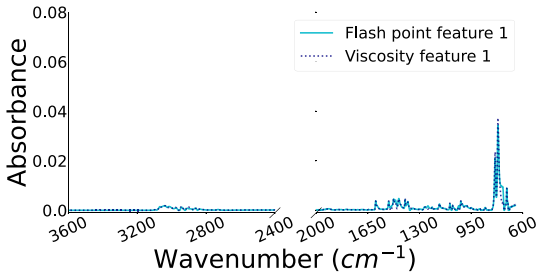


Fig. 5. Highest ranked aromatic features for flash point and viscosity models.

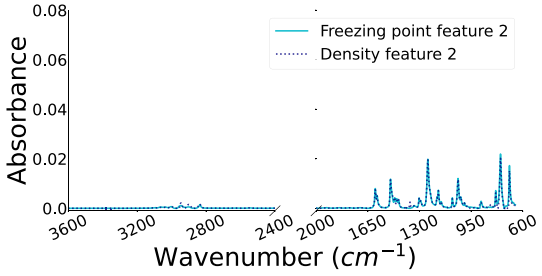


Fig. 6. Highest ranked aromatic features for freezing point and density models.

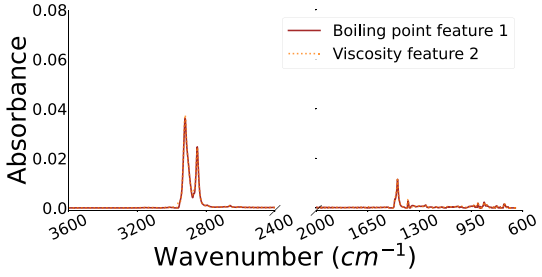


Fig. 7. Highest ranked cycloalkane features for final boiling point and viscosity models.

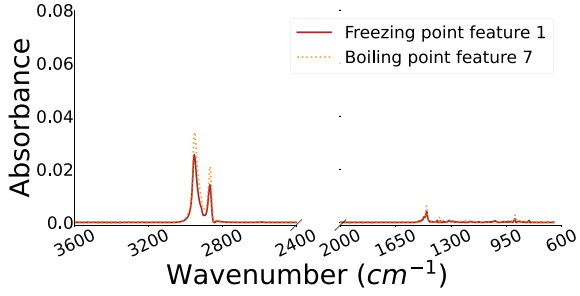


Fig. 8. Highest ranked cycloalkane feature in the freezing point model and second-highest ranked cycloalkane feature in the final boiling point model.

Fig. 7). The features exhibit two peaks near 2850 cm^{-1} and 2920 cm^{-1} representing asymmetric and symmetric stretching vibrations of CH_2 groups, as well as a smaller peak near 1450 cm^{-1} representing C–H bending or scissoring [59,60]. The peaks captured by these features closely match spectra from cyclohexane-based compounds. In addition, the top cycloalkane feature in the freezing point model closely resembles the second cycloalkane feature in the final boiling point model (see Fig. 8). Both of these features contain a peak between 1470 cm^{-1} and 1450 cm^{-1} attributed to C–H bending or scissoring, and C–H stretching peaks near 3000 cm^{-1} [59].

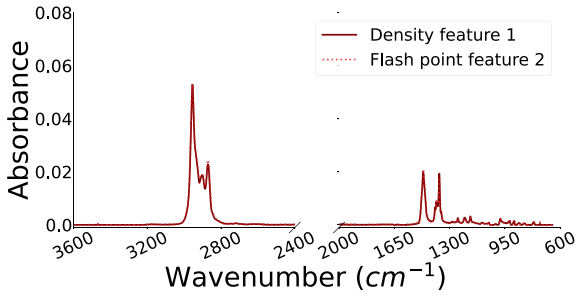


Fig. 9. Highest ranked alkane features for the density and flash point models.

Alkane-like features also play an important role in the density and flash point models, ranking first and second, respectively. The high importance of alkanes for predicting density could be due to alkanes generally having lower densities. The model indicates that alkane-like features in a spectrum would result in lower predicted density. Fig. 9 shows the top features with alkane characteristics for both models are almost indistinguishable. The features exhibit C–H stretching between 3000 cm^{-1} and 2850 cm^{-1}, and C–H bending or scissoring between 1470 – 1450 cm^{-1} [59]. In addition, the features contain C–H rocking bands attributed to methyl groups between 1370 – 1360 cm^{-1} [59]. The features lack C–H rocking bands seen in long-chain n-alkanes from 725 – 720 cm^{-1} and most likely represent shorter n-alkanes or branched alkanes such as isooctane [59]. Further details for other model features used in each property can be found in Section S-4 of the SI.

4. Conclusions

This study provides a structured method for creating accurate and interpretable property prediction models using liquid-phase FTIR spectra of neat molecules, petroleum-based jet fuels, synthetic and alternative aviation fuels, and their blends. Liquid-phase FTIR spectra can be collected rapidly and consistently, characterizing compounds with only a few drops of a sample. The method uses NMF to deconstruct FTIR spectra into fundamental building blocks that are used as features to train machine learning models. Deconstruction occurs intuitively, preserving inherent characteristics of spectra by leveraging the non-negativity constraint. Furthermore, this method facilitates interpretation for discovering new relationships between compositional elements of a fuel, such as functional groups or chemical classes, and its properties.

To demonstrate the method, we developed five ensemble models to predict final boiling point, flash point, freezing point, density at 15 °C, and kinematic viscosity at −20 °C. The number of neat molecules, petroleum-based jet fuels, synthetic and alternative aviation fuels, and blends in the property databases ranged from 31 (kinematic viscosity) to 72 (final boiling point). The models' performances align with published literature, having test mean absolute errors of 17.17 K, 9.6 K, 8.9 K, 22.2 kg/m^3, and 0.45 mm^2/s for final boiling point, flash point, freezing point, density, and kinematic viscosity, respectively. Because some of the experimental property data were discretized due to the large ranges in values, more data—especially additional certified fuels and representative blends—would improve accuracy of the models and further validate their performance and reliability.

This method provides interpretable features that enable the investigation of chemical classes, functional groups, and bonds contributing most to property prediction. Specifically, all models rank features with aromatic characteristics either first or second, with flash point and viscosity models sharing top aromatic features that are almost indistinguishable. The top aromatic features for the flash point and viscosity models exhibit a prominent peak characteristic of terminal ethyl groups, several other out-of-plane C–H bending bands associated with aromatics, and smaller peaks representing C–C in-ring stretching. Top aromatic

features for freezing point and density models also show nearly identical characteristics, including prominent peaks representing out-of-plane C−H bands, carbon–carbon stretching vibrations in the aromatic rings, and C−C stretching from terminal tert-butyl groups. Additional important features for boiling point, freezing point, and viscosity models include cycloalkane characteristics, while density and flash point models highly rank features with alkane characteristics.

Overall, the method can consistently predict properties of neat molecules, fuels, and blends, while providing interpretable features that enable scientific discovery and help accelerate SAF research. To support research and development, the models and data have been integrated into an existing web tool, located at feedstock-to-function.lbl.gov.

We recommend that future fuel-property models report consistent test metrics in addition to overall errors which combine testing and training errors. This will improve transparency in model performance, especially with new data, and facilitate comparisons across studies, as prior studies often lack standardized metric definitions or clear reporting of evaluation procedures. To improve and expand the method, future work should explore additional analytical techniques with a substantial amount of spectral and experimental property data.

CRediT authorship contribution statement

Ana E. Comesana: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Data curation, Conceptualization. **Sharon Chen:** Writing – review & editing, Methodology, Data curation. **Kyle E. Niemeyer:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization. **Vi H. Rapp:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Funding acquisition, Conceptualization.

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.

Acknowledgements

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.

Work at the Molecular Foundry was supported by the Office of Science, Office of Basic Energy Sciences, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

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-AC05-00OR22725. 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, worldwide 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 also like to acknowledge Ben Harvey, Paul Wrzesinski, and Milissa Flake for supplying aviation fuels and compounds critical to the success of this project; Josh Heyne for providing a wealth of data, insights, and feedback; Tonghun Lee for support with obtaining fuel and fuel blend property data; Jaspreet Singh and Mark McMahon for their support with interpreting FTIR spectra data, and for helping us obtain, operate, and calibrate equipment for FTIR spectra measurements; Liana Klivansky for her support with operating and interpreting data from the FTIR spectrometer; Cody Bastien for ensuring the fuel property certification measurements were repeatable and robust; and Michael Golden for his support with gaining access to published literature and data.

Appendix A. Supplementary data

Supplementary data for this article can be found online at doi:10.1016/j.fuel.2025.137588.

Data availability

Data and models are available on our website, feedstock-to-function.lbl.gov, as stated in the manuscript. Data will also be available upon request.

References

[1] American Society of Testing and Materials. ASTM d4054-19: standard practice for evaluation of new aviation turbine fuels and fuel additives. West Conshohocken, PA: ASTM International; 2019. https://doi.org/10.1520/D4054-19
[2] American Society of Testing and Materials. ASTM d7566-20: standard specification for aviation turbine fuel containing synthesized hydrocarbons. West Conshohocken, PA: ASTM International; 2020. https://doi.org/10.1520/D7566-20
[3] Heyne J, Rauch B, Le Clercq P, Colket M. Sustainable aviation fuel prescreening tools and procedures. Fuel 2021;290:120004. https://doi.org/10.1016/j.fuel.2020.120004
[4] Boddapati V, Ferris AM, Hanson RK. Predicting the physical and chemical properties of sustainable aviation fuels using elastic-net-regularized linear models based on extended-wavelength FTIR spectra. Fuel 2024;356:129557. https://doi.org/10.1016/j.fuel.2023.129557
[5] Wang Y, Ding Y, Wei W, Cao Y, Davidson DF, Hanson RK. On estimating physical and chemical properties of hydrocarbon fuels using mid-infrared FTIR spectra and regularized linear models. Fuel 2019;255:115715. https://doi.org/10.1016/j.fuel.2019.115715
[6] Boddapati V, Ferris AM, Hanson RK. On the use of extended-wavelength FTIR spectra for the prediction of combustion properties of jet fuels and their constituent species. Proc Combust Inst 2023;39(1):1347–55. https://doi.org/10.1016/j.proci.2022.08.041
[7] Comesana AE, Huntington TT, Scown CD, Niemeyer KE, Rapp VH. A systematic method for selecting molecular descriptors as features when training models for predicting physiochemical properties. Fuel 2022;321:123836. https://doi.org/10.1016/j.fuel.2022.123836
[8] Espinosa G, Arenas A, Giralt F. Prediction of boiling points of organic compounds from molecular descriptors by using backpropagation neural network. In Carbó-Dorca R, Gironés X, Mezey PG, editors. Fundamentals of molecular similarity. Boston, MA: Springer US; 2001, Ch. 1, pp. 1–10. https://doi.org/10.1007/978-1-4757-3273-3, Mathematical and Computational Chemistry.
[9] Dai YM, Zhu ZP, Cao Z, Zhang YF, Zeng JL, Li X. Prediction of boiling points of organic compounds by QSPR tools. J Mol Graph Model 2013;44:113–9. https://doi.org/10.1016/j.jmgm.2013.04.007
[10] 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. https://doi.org/10.1021/ci00020a017
[11] Cherqaoui D, Villemin D. Use of a neural network to determine the boiling point of alkanes. J Chem Soc Faraday Trans 1994;90:97. https://doi.org/10.1039/ft9949000097
[12] 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. https://doi.org/10.1021/ci0500132
[13] Gharagheizi F, Alamdari R. Prediction of flash point temperature of pure components using a quantitative structure-property relationship model. QSAR Comb Sci 2008;27:679–83. https://doi.org/10.1002/qsar.200730110
[14] 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. https://doi.org/10.1021/ci050307n
[15] 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. https://doi.org/10.1021/ci020280x
[16] 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 November. 2021;304. https://doi.org/10.1016/j.fuel.2021.121437
[17] 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. https://doi.org/10.1080/1062936X.2013.766634
[18] 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. https://doi.org/10.1016/j.fluid.2007.09.022
[19] 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 Model 2019;87:109–20. https://doi.org/10.1016/j.jmgm.2018.11.013
[20] 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. https://doi.org/10.1021/jp953224q

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

[22] Saldana DA, Starck L, Mougin P, Rousseau B, Pidol L, Jeuland N, Creton B. Flash point and cetane number predictions for fuel compounds using quantitative structure property relationship (QSPR) methods. Energy Fuels 2011:9.

[23] St. John PC, Kairys P, Das DD, McEnally CS, Pfefferle LD, Robichaud DJ, Nimlos MR, Zigler BT, McCormick RL, Foust TD, Bomble YJ, Kim S. A quantitative model for the prediction of sooting tendency from molecular structure. Energy Fuels 2017;31(9):9983–90. https://doi.org/10.1021/acs.energyfuels.7b00616

[24] Aslani S, Armstrong DW. High information spectroscopic detection techniques for gas chromatography. J Chromatogr A 2022;1676:463255. https://doi.org/10.1016/j.chroma.2022.463255

[25] Caswell KA, Glass TE, Swann M, Dorn HC. Rapid prediction of various physical properties for middle distillate fuels utilizing directly coupled liquid chromatography/proton nuclear Magnetic resonance. Anal Chem 1989;61(3):206–11. https://doi.org/10.1021/ac00178a004

[26] Johnson K, Morris R, Rose-Pehrsson S. Evaluating the predictive powers of spectroscopy and chromatography for fuel quality assessment. Energy Fuels February 2006;20. https://doi.org/10.1021/ef050347t

[27] Yang Z, Kosir S, Stachler R, Shafer L, Anderson C, Heyne JS. A GC×GC tier α combustor operability prescreening method for sustainable aviation fuel candidates. Fuel 2021;292:120345. https://doi.org/10.1016/j.fuel.2021.120345

[28] Vozka P, Kilaz G. A review of aviation turbine fuel chemical composition-property relations. Fuel 2020;268:117391. https://doi.org/10.1016/j.fuel.2020.117391

[29] Bradley M. Advantages of FT-IR in polymer manufacturing. June 2021. https://www.thermofisher.com/blog/materials/advantages-of-ft-ir-in-polymer-manufacturing/.

[30] Wang Y, Wei W, Zhang Y, Hanson RK. A new strategy of characterizing hydrocarbon fuels using FTIR spectra and generalized linear model with grouped-lasso regularization. Fuel 2021;287:119419. https://doi.org/10.1016/j.fuel.2020.119419

[31] Fodor GE, Mason RA, Hutzler SA. Estimation of middle distillate fuel properties by FT-IR. Appl Spectrosc 1999;53(10):1292–8. https://doi.org/10.1366/0003702991945542

[32] Daly SR, Niemeyer KE, Cannella WJ, Hagen CL. Predicting fuel research octane number using fourier-transform infrared absorption spectra of neat hydrocarbons. Fuel 2016;183:359–65. https://doi.org/10.1016/j.fuel.2016.06.097

[33] Daly SR, Tran K, Niemeyer KE, Cannella WJ, Hagen CL. Predicting fuel low-temperature combustion performance using fourier-transform infrared absorption spectra of neat hydrocarbons. Fuel 2019;242:343–4. https://doi.org/10.1016/j.fuel.2019.01.054

[34] Yang S-B, Moreira J, Li Z. Predicting crude oil properties using fourier-transform infrared spectroscopy (FTIR) and data-driven methods. Digit Chem Eng 2022;3:100031. https://doi.org/10.1016/j.dche.2022.100031

[35] Moro MK, Neto ÁC, Lacerda V, Romão W, Chinelatto LS, Castro EVR, Filgueiras PR. FTIR, 1h and 13c NMR data fusion to predict crude oils properties. Fuel 2020;263:116721. https://doi.org/10.1016/j.fuel.2019.116721

[36] Rivera-Barrera D, Rueda-Chacón H, Molina V D. Prediction of the total acid number (TAN) of Colombian crude oils VIA ATR–FTIR spectroscopy and chemometric methods. Talanta 2020;206:120186. https://doi.org/10.1016/j.talanta.2019.120186

[37] Stec AA, Fardell P, Blomqvist P, Bustamante-Valencia L, Saragoza L, Guillaume E. Quantification of fire gases by FTIR: experimental characterisation of calibration systems. Fire Saf J 2011;46(5):225–33. https://doi.org/10.1016/j.firesaf.2011.02.004

[38] Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, Zaslavsky L, Zhang J, Bolton EE. Pubchem 2019 update: improved access to chemical data. Nucleic Acids Res 2018;47(D1):D1102–D1109. https://doi.org/10.1093/nar/gky1033

[39] Chemspider. https://www.chemspider.com/.

[40] Springer Materials. https://materials.springer.com/.

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

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

[43] Edwards J. Reference jet fuels for combustion testing, In: 55th AIAA Aerospace Sciences Meeting. American Institute of Aeronautics and Astronautics; 2017. https://doi.org/10.2514/6.2017-0146

[44] Rosenkoetter K, Kennedy C, Chirik P, Harvey B. [4 + 4]-cycloaddition of isoprene for the production of high-performance bio-based jet fuel. Green Chem September 2019;21. https://doi.org/10.1039/C9GC02404B

[45] Olivier Rolland, Olivier Rolland. Evaluation of synthesized iso-paraffins produced from hydroprocessed fermented sugars (SIP fuels), Final report, TOTAL New Energies, Amyris, Inc., and United States Air Force Research Laboratory (February. 2014.

[46] Yang J, Xin Z, He QS, Corscadden K, Niu H. An overview on performance characteristics of bio-jet fuels. Fuel 2019;237:916–36. https://doi.org/10.1016/j.fuel.2018.10.079

[47] National Renewable Energy Laboratory. Co-optimization of fuels & engines: fuel properties database. https://www.nrel.gov/transportation/fuels-properties-database.

[48] Yaws CL. Chapter 1 - physical properties – organic compounds. In Yaws CL, editor. The Yaws Handbook of physical properties for hydrocarbons and chemicals (Second Edition) Boston: Gulf Professional Publishing; 2015. pp. 1–683. https://doi.org/10.1016/B978-0-12-800834-8.00001-3

[49] Singh E, Tingas E-A, Goussis D, Im HG, Sarathy SM. Chemical ignition characteristics of ethanol blending with primary reference fuels. Energy Fuels 2019;33(10):10185–96. https://doi.org/10.1021/acs.energyfuels.9b01423

[50] Zhang F, Tang X, Tong A, Wang B, Wang J. An automatic baseline correction method based on the penalized least squares method. Sensors 2020;20(7). https://doi.org/10.3390/s20072015

[51] McKinney W. Data structures for statistical computing in Python, In: van der Walt S, Millman J, editors. Proceedings of the 9th Python in Science Conference; 2010. p. 51–6.

[52] Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, Kern R, Picus M, Hoyer S, van Kerkwijk MH, Brett M, Haldane A, del Río JF, Wiebe M, Peterson P, Gérard-Marchant rard-Marchant P, Sheppard K, Reddy T, Weckesser W, Abbasi H, Gohlke C, Oliphant TE. Array programming with numpy. Nature 2020;585(7825):357–62. https://doi.org/10.1038/s41586-020-2649-2

[53] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12:2825–30.

[54] Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature 1999;401(6755):788–91. https://doi.org/10.1038/44565

[55] Lee D, Seung H. Algorithms for non-negative matrix factorization. Adv Neural Inform Process Syst February 2001;13.

[56] Paatero P, Tapper U. Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. Environmetrics 1994;5:111–26. https://doi.org/10.1002/env.3170050203

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

[58] Scheuermann SS, Forster S, Eibl S. In-depth interpretation of mid-infrared spectra of various synthetic fuels for the chemometric prediction of aviation fuel blend properties. Energy Fuels 2017;31(3):2934–43. https://doi.org/10.1021/acs.energyfuels.6b03178

[59] Morsch, Farmer, Cunningham, Sharrett, Kennepohl. Organic Chemistry. 2024. https://chem.libretexts.org/Bookshelves/Organic_Chemistry/Organic_Chemistry_(Morsch_et_al.).

[60] Puhan D, Casford MTL, Davies PB. Evaluation of structural and compositional changes of a model monoaromatic hydrocarbon in a benchtop hydrocracker using GC, FTIR, and NMR spectroscopy. ACS Omega 2023;8(39):35988–6000. https://doi.org/10.1021/acsomega.3c03833