On estimating physical and chemical properties of hydrocarbon fuels using mid-infrared FTIR spectra and regularized linear models

Yu Wang[a,*], Yiming Ding[a], Wei Wei[a], Yi Cao, David F Davidson[a], Ronald K Hanson[a]

[a]*Stanford University, Stanford, California 94305*

Abstract

The concept of a compact, economical FTIR-based analyzer for estimating the properties of hydrocarbon fuels with small amounts of fuel is proposed. The high correlations between mid-IR FTIR absorption spectra of fuel vapor in the range 3300 to 3550 nm and 15 physical and chemical properties, such as density, initial boiling point, surface tension, kinematic viscosity, number of carbon and hydrogen per average molecule, and derived cetane number, for 64 hydrocarbon fuels are demonstrated. Lasso-regularized linear models based on linear combination of absorption cross sections at selected wavelengths are built for each of these physical and chemical properties, yielding accurate estimations.

Keywords: Hydrocarbon fuel, alternative fuel, physical and chemical property, combustion, mid-IR spectroscopy, generalized linear model, machine learning

*Corresponding author
Email address: yuwangme@stanford.edu (Yu Wang)

Nomenclature

IR	Infrared
FTIR	Fourier-transform infrared
GC	Gas chromatography
GC×GC	Two-dimensional GC
CS	Cross section
DCN	Derived cetane number by ASTM D6890
IDT	Ignition delay time
LBO	Lean blow-out
FP	Flash point by ASTM D93
Total C	Total number of carbon atoms per average molecule
Total H	Total number of hydrogen atoms per average molecule
MW	Molecular weight
IBP	Initial boiling point by ASTM D86
ρ	Density at 15°C by ASTM D4052
ST	Surface tension at 22°C by ASTM D1331
NHC	Net heat of combustion by ASTM D4809
KV	Kinematic viscosity at -20°C by ASTM D445
Total cyclo	Total cycloparaffin weight percentage
CV	Cross validation
CVE	Cross validation error
NJFCP	National jet fuel combustion program

1. Introduction

In the past two decades, Alternative Jet Fuel (AJF) development has been a popular combustion research topic due to an increasing interest in reducing combustion emissions, mitigating climate change, and improving energy supply security [1]. In particular, the National Jet Fuel Combustion Program (NJFCP) was established by the Federal Aviation Administration (FAA) in 2014, supporting a collaborative effort involving over 30 institutions to understand the impact of jet fuel physical properties and chemical composition on combustion behavior. A particular goal of this program is to eventually streamline the process of AJF certification, which is currently a major hurdle for market penetration [1, 2, 3]. According to market research [4], the global alternative fuel and hybrid vehicle market is expected to reach $614 billion (about 3% of the US GDP in 2017) by 2022 and it is currently growing at compound average growth rate of 12.9%.

Within the kinetics working group of NJFCP, there has been remarkable progress [2, 3] in applications of advanced laser diagnostics techniques in shock tube experiments [5, 6, 7, 8, 9, 10, 11] to characterize fuel pyrolysis and combustion behavior of a range of conventional, alternative, and synthetic jet fuels [12], enabling useful correlations between DCN and other jet fuel properties and contributing both to detailed chemical kinetic modeling and the hybrid chemistry (HyChem) approach [13, 14, 15, 16]. However, it is still difficult to physically model and calculate jet fuel properties. One of the main difficulties comes from jet fuels' complicated compositions that usually consist of hundreds of components. To overcome this challenge, direct estimation of these properties of hydrocarbon fuels from relatively accessible and available infrared spectral data has been proposed and studied by various researchers.

Zanier-Szydlowski et al. proposed in [17] methods using multivariate linear regression and liquid-phase near-IR spectra in the range 1562 to 2222.2 nm to predict the refractive index at 20°C, the density at 15°C, the weight percentage of hydrogen, the percentage of aromatic carbon and the weight percentage

of mono-, di- and total aromatics for hydrotreated gas oils. Balabin and Lomakina discussed in [18] the need for rapid, robust, and cheap quality control of industrial production in real time and online, which motivates the combination of informative liquid-phase near-IR spectra in the range of 909 to 2500 nm with advanced machine learning tools to predict properties of interest. They pointed out that such a need stands out especially in the multi-trillion dollar but environmentally-unfriendly petroleum industry and also in the fast-growing biofuel industry. They also discussed the potential nonlinearity in the spectrum-property relations due to strong intermolecular and intramolecular interactions, and shift of vibrational bands. It is expected that even relatively weak van der Waals force can affect the accuracy of linear models. Torres et al. proposed to apply support vector machine (SVM) and partial least square (PLS) models on liquid-phase mid-FTIR spectra in the range 2500 nm to 16.7 μm to estimate density, refractive index, and cold filter plugging point of biodiesel samples and their blends [19]. They reported the advantage of SVM over PLS for predicting non-linear properties. Alves et al. applied SVM to liquid-phase near-IR spectrum in the range 2096.9 to 2535.5 nm to predict flash point and cetane number and compared the results against those of PLS [20]. Da Silva et al. used liquid-phase spectra in both near- and mid-IR (833.3 nm to 15.4 μm) and machine learning models to classify if a gasoline contains dispersant and detergent additives [21]. More related work can be found in [22, 23, 24, 25, 26].

Similarly, estimation methods for jet and diesel fuel properties using nuclear magnetic resonance (NMR) spectra as inputs were also proposed in studies such as [27, 28]. In a similar way, the authors of [29, 30] proposed estimation methods based on SVM and PLS using gas chromatography and mass spectrum data. In addition, various estimation methods based on quantitative structure property relationship (QSPR) were proposed in [31, 32, 33, 34]. A more comprehensive review can be found in [35] by Dryer. To use QSPR to estimate properties, a wide number of one-dimensional (1D), two-dimensional (2D), and 3D molecular descriptors are first calculated. Then these quantitative descriptors, or features, are fed into both linear and nonlinear machine learning models such as multi-

variate linear regression, PLS, artificial neutral network, SVM, etc. Procedures such as recursive feature elimination are employed to down-select the features. In term of general procedure of building machine learning models, the step of calculating quantitative descriptors is essentially feature engineering, which refers to using domain-specific knowledge to create effective features that correlate with properties of interest. Feature engineering is a powerful tool that has proven critical to many machine learning problems, but it also introduces several potential issues. Firstly, it highly depends on the domain knowledge of the modeler. Secondly, it is often difficult to evaluate the relevance and importance of features due to the overlap in information provided by them. Thirdly, feature selection is often at the discretion of modeler and lacks a unified approach.

In this study, a data-driven approach inspired by data science and statistics is taken to directly develop correlations between fuel properties without compositional and kinetic modeling of jet fuels. The main goal is to develop practical and meaningful estimation methods for properties of hydrocarbon fuels using mid-IR spectra of fuel vapor. More specifically, in section 3, correlations between FTIR spectra from 3350 to 3450 nm and 15 physical and chemical properties are studied. Regularized linear models are proposed for each of these properties in section 4. This method has four practical advantages compared to the methods reviewed above. Firstly, feature engineering is not needed. A measured mid-IR FTIR spectrum is used as input without the need of preprocessing. The absorption at each wavelength is effectively one feature. Secondly, feature selection is performed systematically through model regularization. Manual selection is not needed. Thirdly, as demonstrated in section 4, instead of using more complicated non-linear models, linear models using mid-IR spectra data can achieve high estimation accuracy. Fourthly, this method provides estimation of multiple physical and chemical properties with one FTIR spectrum for various types of vaporized hydrocarbon fuels including pure hydrocarbons and their blends, distillate and synthetic jet fuels and their blends. These advantages will be discussed in greater detail in the following text.

Both spectral and property data presented in this paper come from var-

ious sources. Out of the 64 vapor-phase FTIR spectra of hydrocarbon fuels examined in the study, 22 (distillate and synthetic jet fuels) were measured at Stanford University using a Nicolet 6700 FTIR spectrometer; 18 (pure hydrocarbons, including n-, iso-, cyclo- paraffins and toluene) are taken from the Pacific Northwest National Laboratory (PNNL) gas-phase database for quantitative infrared spectroscopy [36]; and the spectra of 24 blends of jet fuels or single hydrocarbons are calculated from the aforementioned 22 plus 18 FTIR spectra. IDT and C_2H_4 yield were measured in Stanford's Flexible Applications Shock Tube (FAST) facility [37]. Derived cetane number (DCN), net heat of combustion (NHC), two-dimensional GC (GC×GC) and all physical properties of all jet fuels were measured by the Air Force Research Laboratory (AFRL) and provided through the NJFCP. LBO data were taken from literature [38, 39]. All properties of pure hydrocarbons and their blends were taken from various literature and online sources, including [40, 41, 42, 43].

2. Fourier-transform infrared spectroscopy

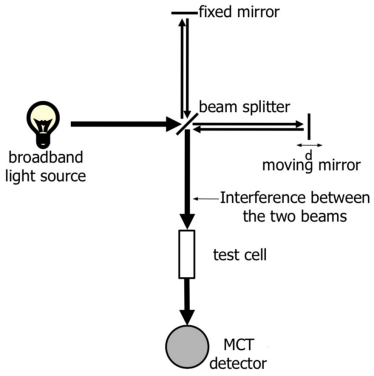


Figure 1: Example optical setup of FTIR spectrometer [44, p. 41-47]

Fourier-transform infrared spectroscopy is a widely used, mature technology (a detailed introduction can be found in [44]). Here we provide a brief review. A FTIR spectrometer typically employs a broadband IR light source, a Michelson

interferometer, and an IR detector to measure absorbance of a test medium (in our case fuel vapor) that can be used to calculate absorption coefficients and cross sections. A schematic for a typical optical setup of an FTIR is shown in Figure 1. As the moving mirror travels (measured by displacement d), different wavelengths from the light source are modulated due to inference. During this process, the spectrometer records the light signal in voltage vs the mirror displacement d, producing an interferogram. By performing a Fourier transformation on the interferogram, the absorption spectrum of the test gas can be inferred. For the PNNL database, the spectral resolution is about 0.1 cm^{-1} and the 1σ statistical uncertainty in absorbance value is $< 2\%$ [36]; jet fuels spectra were measured at Stanford with spectral resolution about 0.06 cm^{-1} and uncertainty around 2%. The details on experimental procedure for measurements of vapor-phase spectra at Stanford are provided in [45].

2.1. Advantages of using vapor-phase mid-IR spectrum

Our vapor-phase mid-IR FTIR measurements utilize fuel vapor at 50 or 80°C (rather than liquid as in [17, 18, 20, 21, 22, 23, 24, 25, 26]) to characterize the hydrocarbon fuels' spectral features. These spectra are not sensitive to temperature from 50 to 80 °C. This method has several advantages. Firstly, the spectrum-property relations for fuel vapor are less susceptible to non-linearity than liquid as described in [18], and the calculation of mid-IR spectra for fuel mixtures is straightforward provided that the mole fraction and spectrum of each individual component is available as discussed in [46]. Beer-Lambert's Law for ideal gas and ideal gas mixtures shown in Equation 1 and Equation 2 can be utilized to calculate the absorption cross section σ_λ at each wavelength λ.

$$\alpha(\lambda) = -\ln\frac{I_t}{I_0} = \sigma_\lambda \frac{PL}{RT} \tag{1}$$

$$\alpha(\lambda) = -\ln\frac{I_t}{I_0} = \sigma_\lambda \frac{PL}{RT} = \sum_i \sigma_{\lambda,i} \frac{Px_i}{RT} L \tag{2}$$

where the summation is over all components i in a mixture; $\alpha(\lambda)$ is absorbance; I_t and I_0 are the laser intensities before and after passage through the absorbing gas; $\sigma_{\lambda,i}$ is the absorption cross section for component i at wavelength λ; P is pressure; x_i is the mole fraction of component i; R is the universal gas constant; T is temperature. Equation 2 enables a simple calculation of a fuel mixture's spectrum provided that the mole fractions x_i and the spectrum of all components $\sigma_{\lambda,i}$ are known. This is of practical importance for the development of alternative jet fuels since they are often mixtures of other hydrocarbon fuels. Secondly, the strong absorption features in the mid-IR region enables high signal-to-noise ratio spectra with small amounts of hydrocarbon fuel. This is again of practical importance for alternative fuel development as their supply is typically limited and may be available only in cubic-centimeter volumes. Thirdly, FTIR measurements are relatively simple and economical compared with NMR and GC×GC methods. Lastly, mid-IR spectra in the range 3300 to 3550 nm provide rich quantitative information on the molecular structure of hydrocarbon fuels. As shown in section 4, a full mid-IR spectrum in this range, subjected to statistical analysis, allows simultaneous and high-fidelity estimations of multiple physical and chemical properties.

3. Demonstration of correlation between FTIR spectra and physical and chemical properties

Table 1: Four absorption features of hydrocarbons

λ [μm]	Dominant Motion [47, 48, 49]
3.32	stretch of benzene H
3.37	asymmetric stretch of $-CH_3$
3.41	asymmetric stretch of $-CH_2-$
3.49	symmetric stretch of $-CH_3$, $-CH_2-$

As mentioned in previous sections, mid-IR spectra in the wavelength range 3350 to 3450 nm are utilized in this study. As shown in Figure 2 and Table 1

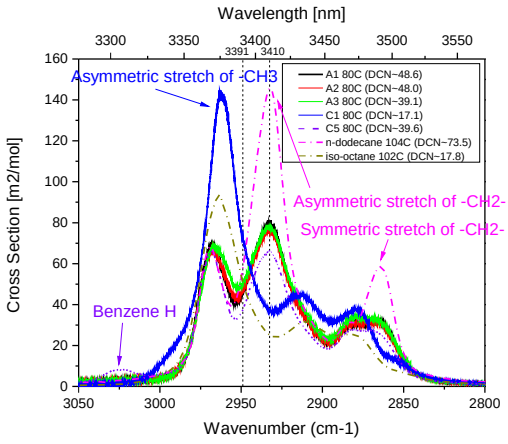


Figure 2: Jet fuel FTIR spectra at 80°C

(or Figure 1 and Table 1 of [46]), there are four main absorption features in this range due to the vibration of C−H bonds from different functional groups: benzene ring, $-CH_3$, $-CH_2-$, etc. Since the physical and chemical properties depend strongly on the functional groups in hydrocarbons, the spectrum-functional group relations provide the physical foundation for using mid-IR spectra to estimate fuel properties. In the rest of this section, the strong correlations between mid-IR spectra and fuels' properties are demonstrated.

3.1. Normalization to the FTIR spectrum

Both the absolute and normalized (by the integrated area) FTIR spectra in the wavelength range 3350 to 3450 nm are utilized. The shape of the normalized spectrum reflects the proportions of chemical component classes and functional groups; the absolute absorption cross section in the unnormalized spectrum reflects average molecule size. It is of note that the shapes of unnormalized and normalized spectra are the same for each fuel. In this paper, we attempt the correlations with both unnormalized and normalized spectra and select the one with the best correlation. In general, properties (such as C_2H_4 yield as defined in the caption of Figure 15, total cycloparaffin weight percentage, and density)

that strongly depend on molecular structure correlate best with the normalized spectrum; properties (such as total number of carbon/hydrogen atoms per average molecule, molecular weight) that strongly depend on molecule size correlate best with the unnormalized spectrum. Table 2 summarizes all physical and chemical properties studied in this study and whether the spectrum used is normalized or not.

3.2. Correlations between physical and chemical properties and absorption cross section at a single wavelength

Algorithm 1: Calculate sample Pearson correlation coefficient $\rho(\sigma_\lambda, P)$ of training dataset $\mathcal{F}$ for each property P, and for each wavelength λ. An example of $\rho(\sigma_\lambda, P)$, where P is total hydrogen per average molecule, is shown in the bottom figure of Figure 3b.

Result: $\rho(\sigma_\lambda, P)$

for *each property* P **do**

  for *each* λ *in 3350 to 3450 nm* **do**

    1. generate a vector of cross sections σ_λ at wavelength λ by interpolating the measured FTIR spectrum
    2. calculate sample Pearson correlation coefficient between σ_λ and P, as defined in Equation 3

  end

end

Here we use the sample Pearson correlation coefficient $\rho(\sigma_\lambda, P)$ defined in Equation 3 as a measure of sensitivity and linearity of the quantitative relation between a physical/chemical property P and absorption cross section σ_λ at a wavelength λ.

$$\rho(\sigma_\lambda, P) = \frac{\sum_{f\in\mathcal{F}} (\sigma_{\lambda,f} - \bar{\sigma}_\lambda)\left(P_f - \bar{P}\right)}{\sqrt{\sum_{f\in\mathcal{F}} (\sigma_{\lambda,f} - \bar{\sigma}_\lambda)^2 \sum_{f\in\mathcal{F}} \left(P_f - \bar{P}\right)^2}} \tag{3}$$

where f denotes a fuel in dataset $\mathcal{F}$; $\sigma_{\lambda,f}$ is the absorption cross section of fuel f at wavelength λ; $\bar{\sigma}_\lambda$ is the average of absorption cross sections at wavelength

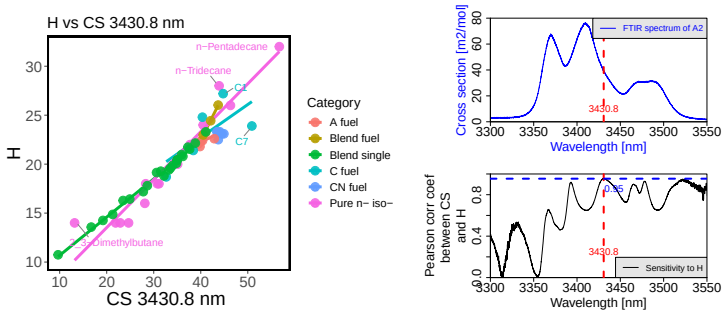
λ over all fuels in $\mathcal{F}$; P_f is the property P of fuel f; $\bar{P}$ is the average of property P over all fuels in $\mathcal{F}$. For a dataset $\mathcal{F}$ of fuels listed in Table 18, $\rho(\sigma_\lambda, P)$ measures the linearity between σ_λ and P and the quality of linear regression between them; ρ is always within ± 1 and $\rho = \pm 1$ indicates a perfect linear relation; $\rho = 0$ implies zero correlation between σ_λ and P for dataset $\mathcal{F}$. In this section, dataset $\mathcal{F}$ includes only fuels that are not pure aromatics or pure cycloparaffins from Table 18.

The most sensitive wavelength λ^* is selected such that the absorption cross section σ_λ has the highest sample Pearson correlation coefficient with the target property P, i.e.

$$\lambda^* = \arg\max \rho(\sigma_\lambda, P). \tag{4}$$

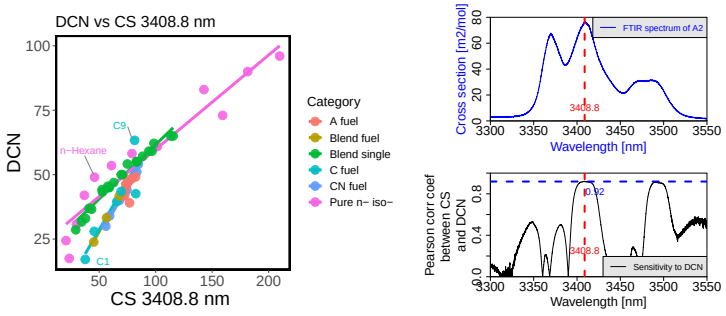
The procedure of selecting the most sensitive wavelength λ^* is outlined in Algorithm 1. For each property P of interest, the algorithm iterates though all 1600 wavelengths in the range of 3350 to 3450 nm (width of wavelength slice equals $100/1600 = 0.0625$ nm) and examines the sample Pearson correlation coefficient between the absorption cross section σ_λ and property P. Then it picks the most sensitive wavelength λ^* defined by Equation 4. It is of note that we only reported the most sensitive wavelength in region 3350 - 3450 nm due to a consideration of signal to noise ratio, i.e. the absorption is much stronger in the range of 3350 - 3450 nm. In addition, the absorption cross sections in 3450 - 3550 nm strongly correlate with those in 3350 - 3450 nm, since they are both due to molecular motions of $-CH_2-$ and $-CH_3$. Hence, when seeking for the most sensitive single wavelength, there is not much to gain by including 3450 - 3550 nm.

The correlations between mid-IR FTIR spectra and 15 physical and chemical properties listed in Table 2 are analyzed. Detailed descriptions of each column of Table 2 are included in its caption. All property values are listed in Table 19. Here we present two examples: total number of hydrogen per average molecule (Figure 3) and DCN (Figure 4). Figure 3 shows that the total number



(a) (b)

Figure 3: Total number of hydrogen per average molecule estimation using absorption cross section at $\lambda^* = 3430.8$ nm. (a) Total number of hydrogen per average molecule vs absorption cross section at 3430.8 nm. (b) Top: example absorption spectrum of a nominal jet fuel A2 (see Table 18); bottom: Pearson correlation coefficient $\rho(\sigma_\lambda, P)$ for $\lambda \in [3300, 3550]$ nm, where P stands for total number of hydrogen per average molecule, for all 64 fuels in Table 18.



(a) (b)

Figure 4: DCN estimation using absorption cross section at 3408.8 nm. (a) DCN vs absorption cross section at $\lambda^* = 3408.8$ nm. (b) Top: example absorption spectrum of a nominal jet fuel A2 (see Table 18); bottom: Pearson correlation coefficient $\rho(\sigma_\lambda, P)$ for $\lambda \in [3300, 3550]$ nm, where P stands for DCN, for 61 fuels in Table 18 for which DCN is available.

of hydrogen per average molecule can be well estimated using the unnormalized absorption cross section $\sigma_{3430.8nm}$ for various types of hydrocarbon fuels with
250 Pearson correlation coefficient $\rho = 0.95$. Figure 4 shows the correlation between derived cetane number (DCN) and the absorption cross section $\sigma_{3408.8nm}$ with $\rho = 0.92$. As a summary of the most sensitive wavelength to each property, Fig-

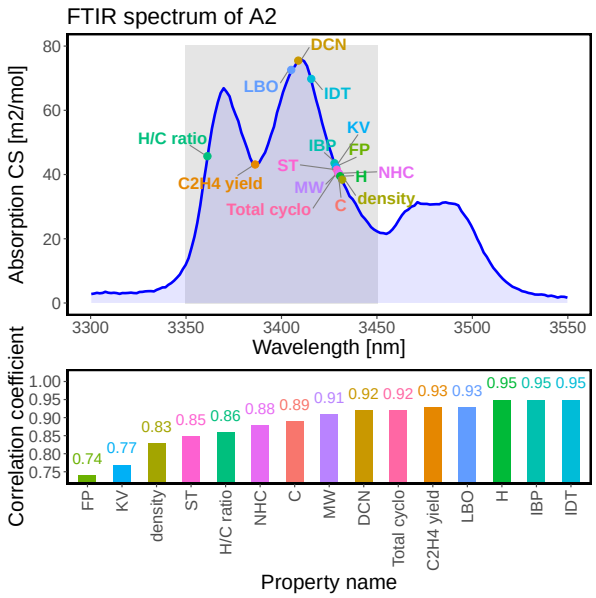


Figure 5: Top: most sensitive wavelength from 3350 to 3450 nm (shaded area) for 15 physical and chemical properties over the set of 64 fuels. Plotted here is the FTIR spectrum for A2 fuel (see Table 18). All correlations have sample Pearson correlation coefficient $\rho \in [0.74, 0.95]$; bottom: the sample Pearson correlation coefficient at λ^* for each of the 15 properties.

ure 5 visualizes their spectral location on the unnormalized FTIR spectrum of a nominal distillate jet fuel A2 (POSF10325, with detailed description in Table 18 and [12]). In Figure 5, the sample Pearson correlation coefficients range from 0.74 to 0.95 at the most sensitive wavelength for each property. Not surprisingly, important combustion properties, such as LBO, DCN, IDT, strongly correlate with the absorption peak corresponding to the $-CH_2-$ functional group (around 3410 nm) [46]; physical properties that depend strongly on molecule size, such as total carbon/hydrogen per average molecule, molecular weight, initial boiling point, correlate well with wavelengths in between absorption features of the $-CH_2-$ and $-CH_3$ functional groups. The clustering around 3425 nm could be due to a clustering around density and boiling point. As pointed out in [50], many properties such as surface tension and molecular weight can be estimated with density and average boiling point. These strong correlations demonstrate

the potential of using mid-IR FTIR spectra of fuel vapor to estimate physical and chemical properties of hydrocarbon fuels.

Table 2: Column "F." shows the figure number for each regularized linear model; column "P." is the name of physical/chemical properties; column "N." indicates if the normalized spectrum is used or not ("F" stands for false and "T" for true); m is the total number of data points in the training dataset with corresponding property data; column "CVE" shows the 10-fold cross validation error; column "%" is defined as CVE divided by the average of positive property values then multiplied by 100; N_λ is the number of wavelengths used in the regularized linear model.

F.	P.	N.	m	CVE	%	N_λ
7	Total C	F	64	0.315	3.2	10
8	Total H	F	64	0.428	2.1	10
9	MW	F	64	4.22	3.1	10
10	H/C ratio	F	64	0.0389	1.9	19
11	IBP	F	33	11.3	7.5	11
12	ρ	T	27	0.0172	2.3	16
13	ST	F	16	0.669	2.8	6
14	NHC	T	21	0.105	0.24	7
15	C_2H_4 yield	T	23	0.121	8.8	7
16	FP	F	19	6.64	14	6
17	LBO	F	11	6.49E-4	0.79	6
18	DCN	F	61	3.66	7.9	10
19	IDT	F	20	108	8.7	3
20	KV	F	15	0.697	14	7
21	Total cyclo	T	65	4.77	14	10

4. Regularized linear model for improved prediction accuracy

The predictive power of mid-IR FTIR spectra towards physical and chemical properties of hydrocarbon fuels is demonstrated in section 3. To obtain an accurate and practical estimation method for these properties, we choose to

use multiple wavelengths selected (by algorithm) from the full spectrum in 3300 to 3550 nm instead of using single wavelengths as in section 3. In the following sections, we present cross-validated linear models with Lasso regularization trained for each of the properties. In this section, all 64 fuels listed in Table 18 are included in the training dataset. All property values are listed in Table 19. The procedure of model development is outlined in Algorithm 2. For each property of interest, the algorithm generates an optimal model (and an optimal β_μ as defined in Equation 5) for each μ (as defined in Equation 5) in a sequence of μ's that eventually results in a different number of selected wavelengths. Then the algorithm compares these optimal models by their 10-fold cross validation error [51] (denoted e_μ) and picks the one with the lowest error.

Algorithm 2: Calculate coefficients β^* for each property with the best 10-fold cross validation error

Result: $\beta^*(P)$ for each property P

Generate a sequence S_μ of μ's, such that $\log_{10}\mu \in [-10, 10]$

for *each property* P **do**

  for *each* μ *in* S_μ **do**

    1) generate vector Y and matrix X as defined for Equation 5

    2) solve minimization problem as defined in Equation 5 and obtain β_μ

    3) perform 10-fold cross validation for μ, obtain cross validation error e_μ

  end

  1. plot e_μ against μ and obtain Figure 6

  2. find $\mu^* = \arg\min e_\mu$ (left dashed line in Figure 6)

  3. save $\beta^*(P) = \beta_{\mu^*}$

end

4.1. Lasso regularization and cross validation

In this section, we denote the discretized FTIR spectrum as matrix $X \in \mathbb{R}^{m \times n}$ and the properties as vector $Y \in \mathbb{R}^m$, where m is the number of fuels in the training dataset with corresponding property data and n is the number of wavelengths plus one (intercept). The FTIR spectrum is discretized by keeping 24 evenly separated wavelengths (and hence $n = 24+1 = 25$). The discretization helps to reduce noise in the spectrum while retaining key spectral features. It is of note that n could be larger than m for some properties in Table 2 (note that the number of fuels is also denoted as m in Table 2).

In an ordinary least square (OLS) regression setup, the following optimization problem is solved to obtain the optimal coefficients β:

$$\beta = \arg\min_{\beta \in \mathbb{R}^n} \; ||Y - X\beta||_2,$$

where $||Y - X\beta||_2$ denotes the L2-norm of vector $Y - X\beta$. However, OLS regression is not suitable for problems with $n > m$. In addition, down-selection of wavelengths is preferred as information about molecular structure is not evenly distributed across all wavelengths in 3300 to 3550 nm. Hence we choose to solve the following optimization with Lasso regularization ([52, p. 68-69]) as defined in Equation 5:

$$\beta_\mu = \arg\min_{\beta \in \mathbb{R}^n} \; ||Y - X\beta||_1 + \mu||\beta||_1, \tag{5}$$

where $\mu > 0$ is a hyper-parameter chosen by 10-fold cross validation ([52, p. 241-247]); $||\beta||_1$ is the L1-norm of β defined as $\sum_{i=1}^{n} |\beta_i|$; similarly $||Y - X\beta||_1$ is the L1-norm of vector $Y - X\beta$. The term $\mu||\beta||_1$ in the objective function in Equation 5 penalizes the magnitude of β and serves to limit the degree of freedom of the linear model and reduce overfitting. It also has the benefit of promoting sparsity in β_μ and hence selecting the most informative wavelengths. As mentioned above, 10-fold cross validation is performed by first partitioning fuels into ten partitions, denoted as $d_1, d_2, \cdots, d_{10}$, and then for each partition of data d_i, a model is trained using the other nine partitions of data

$d_1, \cdots, d_{i-1}, d_{i+1}, \cdots, d_{10}$ and the trained model is evaluated on d_i to obtain the cross validation error. The best hyper-parameter μ is chosen to be the one corresponding to the smallest cross validation error. The choice of optimal μ reflects the tradeoff between using more wavelengths for improved estimation accuracy and less wavelengths to control overfitting for better generality. This tradeoff is demonstrated in Figure 6. The cross validation error is high when too many or too few wavelengths are utilized, which corresponds to overfitting to noise in data and underfitting to signal in data. Cross validation error is chosen as the metric to compare linear models with different numbers of wavelength because it estimates future estimation error on unseen data.

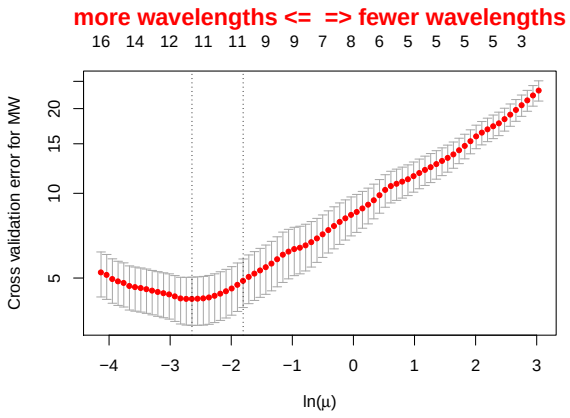


Figure 6: Cross validation error vs μ for estimating molecular weight. Top axis shows the number of wavelengths used corresponding to each μ on the bottom axis. Larger μ corresponds to fewer wavelengths in the linear model. The selected μ corresponds to the left dashed line, where the cross validation error is minimized. The region to the left of the left dashed line is where the model is too large and it is overfitting to the data noise; the region to the right of the right dashed line is where the model is too small and it does not capture all the signal.

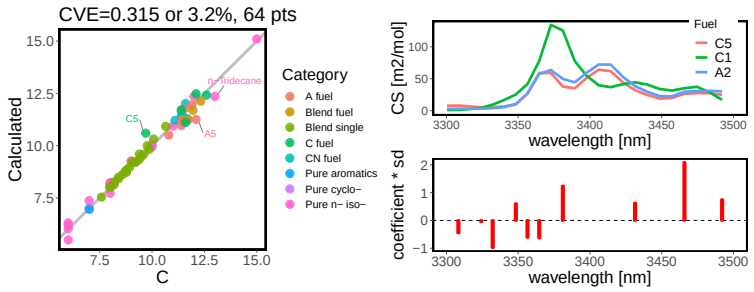
It is worth mentioning the equivalence between Equation 5 and Equation 6 (details provided in [52, p. 68]), where $t(\mu) > 0$ is a decreasing function in $\mu > 0$. The regularization term $\mu||\beta||_1$ effectively limits the possible values of β. Since the objective function of Equation 5 is convex in β, effective optimization

algorithms are available.

$$\beta^* = \arg\min_{\beta \in \mathbb{R}^n} ||Y - X\beta||_1 \quad \text{subject to } ||\beta||_1 \leq t(\mu) \tag{6}$$

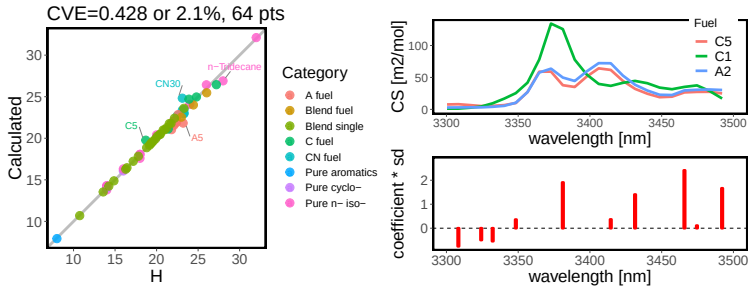
The results are presented below in Figure 7-21 and Table 3-17. Each property corresponds to a figure and table pair. For instance, the regularized linear model for estimating the total number of carbon atoms in an average molecule is shown in Figure 7 and Table 3. In Figure 7, Figure 7a demonstrates the performance of the model on the training dataset. The cross validation error (denoted CVE, both in absolute value and in percentage) and the number of fuels with this property value (total carbon per average molecule) in the training dataset are shown in the title of the figure. A larger CVE indicates potentially larger future estimation error. CVE should be viewed as a lower bound of future prediction error, i.e. the estimation error of total carbon atoms per average molecule is estimated to be at least 3.2%. Figure 7b shows example spectra of three jet fuels (C5, C1, A2, with detailed description available in Table 3 of [46]) and the selected wavelengths and contribution of each wavelength to the variation of total number of carbon. The contribution is calculated as the coefficient of cross section at wavelength λ multiplied by the sample standard deviation of cross sections of all fuels at this wavelength. Table 3 summarizes the selected wavelengths and the coefficients β of the regularized linear model for estimating the total number of carbon per average molecule.

The performance and parameter statistics, including cross validation error (in absolute value and in percentage) and number of wavelengths, of the 15 models for the 15 properties are summarized in Table 2. As shown in Table 2, each model utilizes at most 15 wavelengths. It is worth emphasizing that the regularized linear models presented in this study apply to fuel types in the training dataset, i.e. pure hydrocarbons and their mixtures, distillate and synthetic jet fuels. Caution is advised in extending the use of these models to other fuel types such as oxygenated fuels.



(a) (b)

Figure 7: Total carbon per average molecule. (a) Calculated C using unnormalized spectrum. (b) Example spectra and selected λs and variation calculated at each λ.



(a) (b)

Figure 8: Total hydrogen per average molecule. (a) Calculated H using unnormalized spectrum. (b) Example spectra and selected λs and variation calculated at each λ.

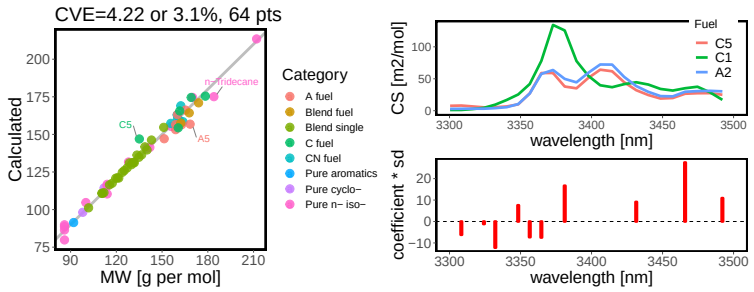
4.2. Linear additivity

The optimal model takes the following mathematical form

$$\text{property} = \beta_0^* + \sum_{i=1}^{N^*} \beta_i^* \sigma_{\lambda_i^*}, \tag{7}$$

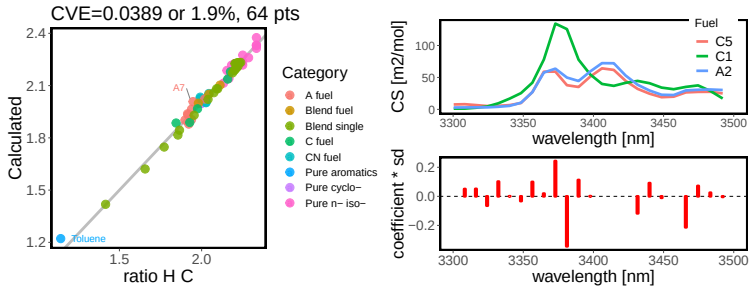
where optimal parameters $N^*, \beta_0^*, \beta_i^*, \lambda_i^*$ are all fitted by the training algorithm.

One observation following the linearity of physical and chemical property in σ_{λ_i} (Equation 7) is that to calculate a property for a fuel mixture one can simply take the average of the property of each component weighted by its mole fraction. This implies that linear interpolation is a reasonable approximation for



(a) (b)

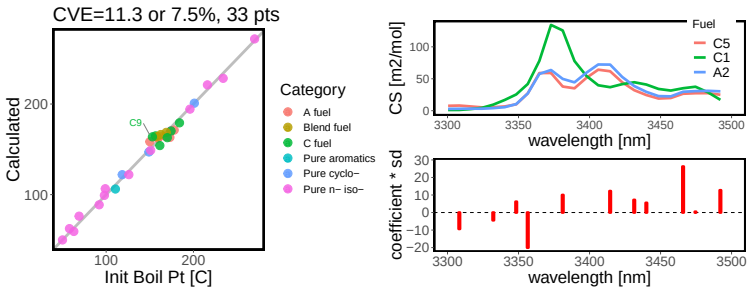
Figure 9: Molecular weight [g/mol]. (a) Calculated MW using unnormalized spectrum. (b) Example spectra and selected λs and variation calculated at each λ.



(a) (b)

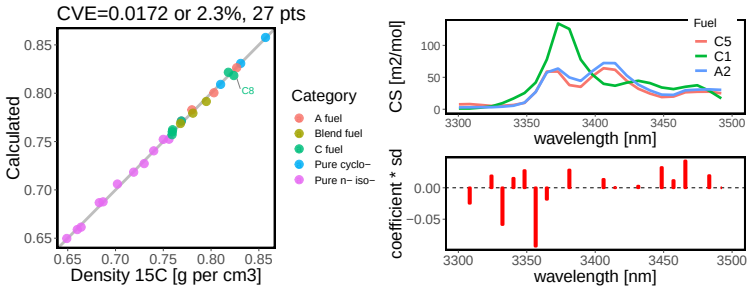
Figure 10: Hydrogen to carbon ratio. (a) Calculated H/C ratio using unnormalized spectrum. (b) Example spectra and selected λs and variation calculated at each λ.

this property and this training dataset regardless of whether it is truly linear in mole fractions. The percentage cross validation error (column "%") in Table 2 is a measure of the approximation quality. For instance, denote two fuels with average molecular formula $C_{m_1}H_{n_1}$, $C_{m_2}H_{n_2}$ with hydrogen to carbon ratio (H/C ratio) $r_1 = \frac{n_1}{m_1}, r_2 = \frac{n_2}{m_2}$ and consider their mixture with mole fractions $x_1, x_2 = 1 - x_1$. Then the H/C ratio of the mixture, as derived in the equations below, is clearly not linear in mole fractions x_1, x_2, but interpolation $x_1 r_1 + x_2 r_2$ can be used as a reasonable approximation considering that the percentage cross validation error is 1.9% (Table 2). The quality of approximation can also be seen from Figure 10a.



(a) (b)

Figure 11: Initial boiling point [°C] by ASTM D86. Data are taken from [43]. (a) Calculated IBP using unnormalized spectrum. (b) Example spectra and selected λs and variation calculated at each λ.



(a) (b)

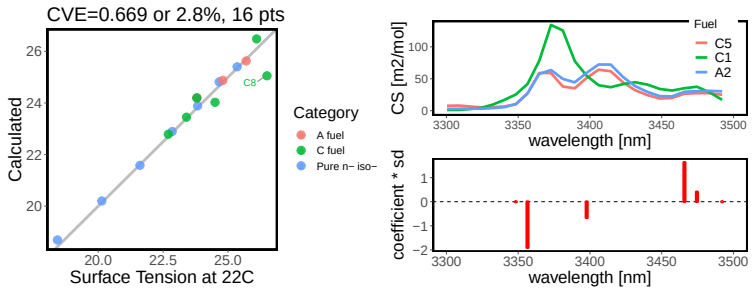
Figure 12: Density [g/cm^3] at 15° C by ASTM D4052, or at 20°C. Data are taken from [43]. (a) Calculated density using normalized spectrum. (b) Example spectra and selected λs and variation calculated at each λ.

$$r_m = \frac{x_1 m_1 r_1 + x_2 m_2 r_2}{x_1 m_1 + x_2 m_2} \tag{8}$$

$$= \frac{x_1}{x_1 + x_2 \frac{m_2}{m_1}} r_1 + \frac{x_2}{x_1 \frac{m_1}{m_2} + x_2} r_2 \tag{9}$$

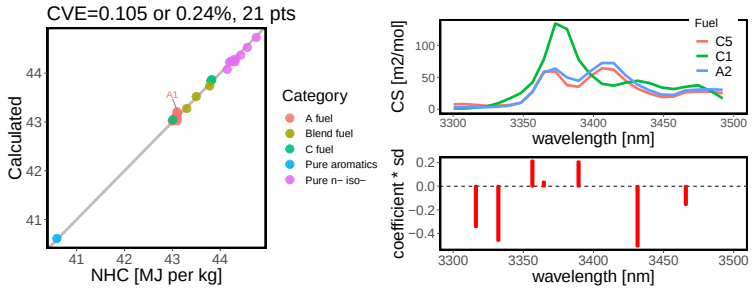
$$= \frac{x_1}{1 + x_2 (\frac{m_2}{m_1} - 1)} r_1 + \frac{x_2}{1 + x_1 (\frac{m_1}{m_2} - 1)} r_2 \tag{10}$$

Importantly, the regularized linear models proposed above can still estimate physical and chemical properties of hydrocarbon fuels based on its measured FTIR spectrum even if the property data for each component is not available.



(a) (b)

Figure 13: Surface tension [dynes/cm] by ASTM D1331. (a) Calculated surface tension using unnormalized spectrum. (b) Example spectra and selected λs and variation calculated at each λ.



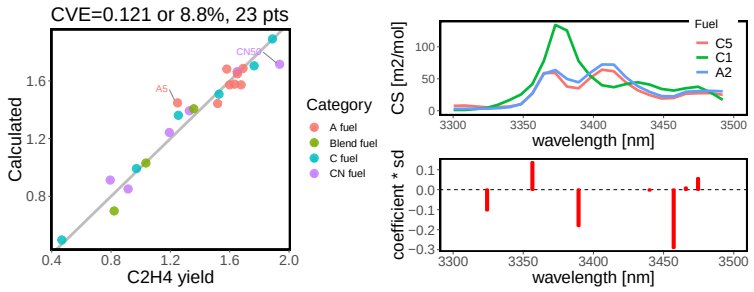
(a) (b)

Figure 14: Net heat of combustion [MJ/kg] by ASTM D4809. (a) Calculated NHC using normalized spectrum. (b) Example spectra and selected λs and variation calculated at each λ.

This is one of the advantages of using vapor phase spectra as described in more detail in subsection 2.1.

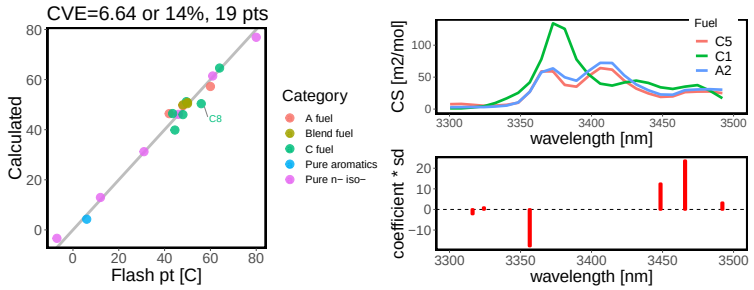
4.3. R language and RStudio

Training and cross validation of the regularized linear models are performed with the R language [53] using RStudio, specifically the `glmnet` package [54, 55] and `cv.glmnet` function, which were developed by researchers in the statistics department at Stanford University.



(a) (b)

Figure 15: C_2H_4 yield at 1300 K, 4 atm and 2 ms. It is defined as the mole fraction of C_2H_4 produced at 2 ms in a jet fuel pyrolysis experiment at 1300 K and 4 atm divided by the initial jet fuel mole fraction. Data are taken from [46]. (a) Calculated C_2H_4 yield using normalized spectrum. (b) Example spectra and selected λs and variation calculated at each λ.

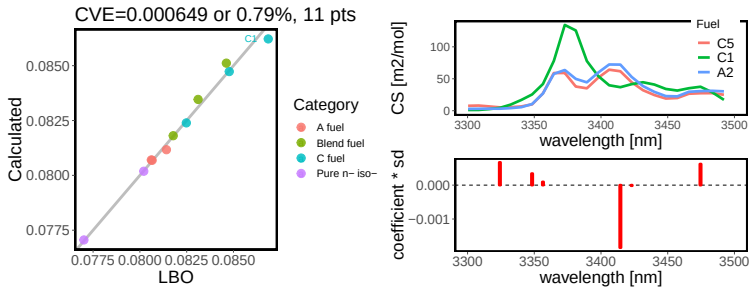


(a) (b)

Figure 16: Flash point [°C] by ASTM D93. Data are taken from [43]. (a) Calculated FP using unnormalized spectrum. (b) Example spectra and selected λs and variation calculated at each λ.

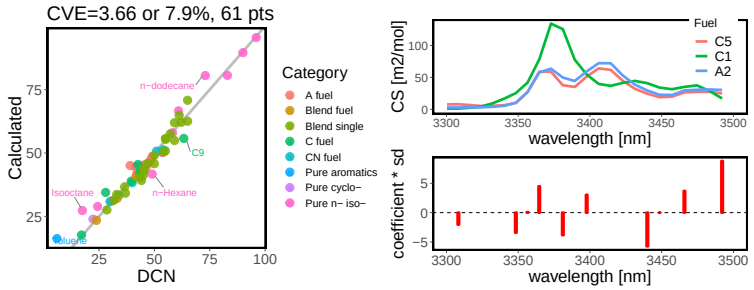
5. Conclusion

FTIR spectroscopy is used to provide the complete spectrum for unreacted hydrocarbon fuel vapor in the range 3300 to 3550 nm. Absorption cross sections in this wavelength region contain quantitative information about molecular structure. Different properties are most sensitive to different wavelengths, which in turn confirms the benefit of using the full spectrum. Spectral data can be combined with more sophisticated statistical models, such as the regularized



(a) (b)

Figure 17: LBO. (a) Calculated LBO using unnormalized spectrum. (b) Example spectra and selected λs and variation calculated at each λ.



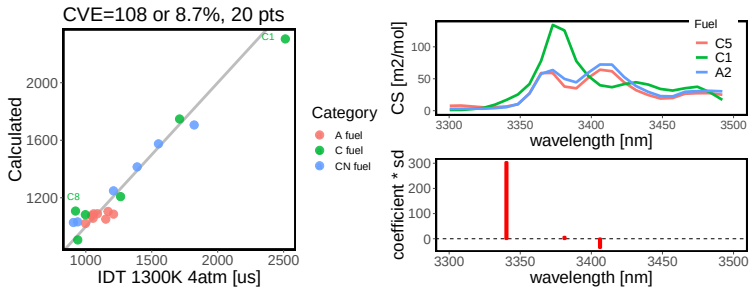
(a) (b)

Figure 18: DCN by ASTM D6890. Data are taken from [40, 41, 42]. (a) Calculated DCN using unnormalized spectrum. (b) Example spectra and selected λs and variation calculated at each λ.

linear model as demonstrated, to provide accurate estimations.

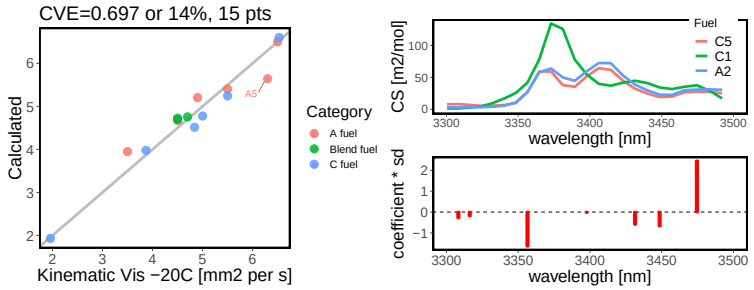
6. Acknowledgement

This work was funded in part by the U.S. Federal Aviation Administration (FAA) Office of Environment and Energy as a part of ASCENT Project 25 under FAA Award Number: 13-C-AJFE-SU-016. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the FAA or other ASCENT sponsors.



(a) (b)

Figure 19: IDT at 1300 K, 4 atm, with equivalence ratio 1. Data are taken from [46]. (a) Calculated IDT using unnormalized spectrum. (b) Example spectra and selected λs and variation calculated at each λ.



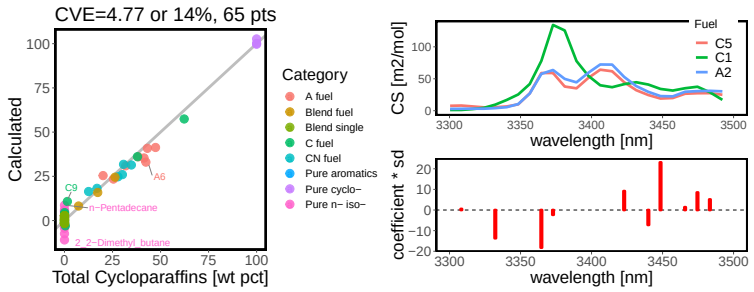
(a) (b)

Figure 20: Kinematic viscosity [mm/s] at -20°C by ASTM D445. (a) Calculated kinematic viscosity using unnormalized spectrum. (b) Example spectra and selected λs and variation calculated at each λ.

This work was also supported by the U.S. Air Force Office of Scientific Research through AFOSR Grant No. FA9550-16-1-0195, with Dr. Chiping Li as contract monitor. DISTRIBUTION A. Approved for public release: distribution unlimited.

Finally, much of the spectral data utilized were obtained in work supported by the U.S. Army Research Laboratory and the U.S. Army Research Office under contract/grant number W911-NF-17-1-0420.

The authors thank Dr. Christopher L. Strand for reviewing the paper.



(a) (b)

Figure 21: Total cycloparaffin weight percentage. (a) Calculated total cycloparaffin weight percentage using normalized spectrum. (b) Example spectra and selected λs and variation calculated at each λ.

Declaration of interest

None.

References

[1] M. Colket, J. Heyne, M. Rumizen, M. Gupta, T. Edwards, W. M. Roquemore, G. Andac, R. Boehm, J. Lovett, R. Williams, J. Condevaux, D. Turner, N. Rizk, J. Tishkoff, C. Li, J. Moder, D. Friend, V. Sankaran, Overview of the National Jet Fuels Combustion Program, AIAA Journal 55 (2017) 1087–1104.

[2] J. S. Heyne, M. B. Colket, M. Gupta, A. Jardines, J. P. Moder, J. T. Edwards, M. Roquemore, C. Li, M. Rumizen, Year 2 of the National Jet Fuels Combustion Program: Towards a Streamlined Alternative Jet Fuels Certification Process, 55th AIAA Aerospace Sciences Meeting (2017) 1–14.

[3] J. S. Heyne, E. Peiffer, M. B. Colket, A. Jardines, C. Shaw, J. P. Moder, W. M. Roquemore, J. T. Edwards, C. Li, M. Rumizen, M. Gupta, Year 3 of the National Jet Fuels Combustion Program: Practical and Scientific Impacts of Alternative Jet Fuel Research, in: 2018 AIAA Aerospace Sci-

ences Meeting, volume 812, 2018. URL: https://arc.aiaa.org/doi/10.2514/6.2018-1667. doi:10.2514/6.2018-1667.

[4] Allied Market Research, Global Opportunity Analysis and Industry Forecast, 2014 - 2022, 2016. URL: https://www.alliedmarketresearch.com/alternative-fuel-and-hybrid-vehicle-market.

[5] T. Parise, D. F. Davidson, R. K. Hanson, Shock tube/laser absorption measurements of the pyrolysis of a bimodal test fuel, Proceedings of the Combustion Institute 36 (2017) 281–288.

[6] J. Shao, R. Choudhary, Y. Peng, D. F. Davidson, R. K. Hanson, A shock tube study of n-heptane, iso-octane, n-dodecane and iso-octane/n-dodecane blends oxidation at elevated pressures and intermediate temperatures, Fuel 243 (2019) 541–553.

[7] A. M. Ferris, D. F. Davidson, R. K. Hanson, A combined laser absorption and gas chromatography sampling diagnostic for speciation in a shock tube, Combustion and Flame 195 (2018) 40–49.

[8] S. Wang, R. K. Hanson, Ultra-sensitive spectroscopy of OH radical in high-temperature transient reactions, Optics Letters 43 (2018) 3518.

[9] W. Wei, W. Y. Peng, Y. Wang, R. Choudhary, S. Wang, J. Shao, R. K. Hanson, Demonstration of non-absorbing interference rejection using wavelength modulation spectroscopy in high-pressure shock tubes, Applied Physics B: Lasers and Optics 125 (2019) 9.

[10] S. Wang, D. F. Davidson, R. K. Hanson, Shock tube measurements of OH concentration time-histories in benzene, toluene, ethylbenzene and xylene oxidation, Proceedings of the Combustion Institute 37 (2019) 163–170.

[11] W. Y. Peng, Y. Wang, J. Cassady, C. L. Strand, R. K. Hanson, Single-Ended Sensor for Thermometry and Speciation in Shock Tubes Using Native Surfaces, IEEE Sensors XX (2019) 1–8.

[12] J. T. Edwards, Reference Jet Fuels for Combustion Testing, 55th AIAA Aerospace Sciences Meeting (2017) 1–58.

[13] H. Wang, R. Xu, K. Wang, C. T. Bowman, R. K. Hanson, D. F. Davidson, K. Brezinsky, F. N. Egolfopoulos, A physics-based approach to modeling real-fuel combustion chemistry - I. Evidence from experiments, and thermodynamic, chemical kinetic and statistical considerations, Combustion and Flame 193 (2018) 502–519.

[14] R. Xu, K. Wang, S. Banerjee, J. Shao, T. Parise, Y. Zhu, S. Wang, A. Movaghar, D. J. Lee, R. Zhao, X. Han, Y. Gao, T. Lu, K. Brezinsky, F. N. Egolfopoulos, D. F. Davidson, R. K. Hanson, C. T. Bowman, H. Wang, A physics-based approach to modeling real-fuel combustion chemistry – II. Reaction kinetic models of jet and rocket fuels, Combustion and Flame 193 (2018) 520–537.

[15] Y. Tao, R. Xu, K. Wang, J. Shao, S. E. Johnson, A. Movaghar, X. Han, J.-W. Park, T. Lu, K. Brezinsky, F. N. Egolfopoulos, D. F. Davidson, R. K. Hanson, C. T. Bowman, H. Wang, A Physics-based approach to modeling real-fuel combustion chemistry – III. Reaction kinetic model of JP10, Combustion and Flame 198 (2018) 466–476.

[16] K. Wang, R. Xu, T. Parise, J. Shao, A. Movaghar, D. J. Lee, J.-W. Park, Y. Gao, T. Lu, F. N. Egolfopoulos, D. F. Davidson, R. K. Hanson, C. T. Bowman, H. Wang, A physics-based approach to modeling real-fuel combustion chemistry – IV. HyChem modeling of combustion kinetics of a bio-derived jet fuel and its blends with a conventional Jet A, Combustion and Flame 198 (2018) 477–489.

[17] N. Zanier-Szydlowski, A. Quignard, F. Baco, H. Biguerd, L. Carpot, F. Wahl, Control of Refining Processes on Mid-Distillates by Near Infrared Spectroscopy, Oil and Gas Science and Technology 54 (1999) 463–472.

[18] R. M. Balabin, E. I. Lomakina, Support vector machine regression

(SVR/LS-SVM) - An alternative to neural networks (ANN) for analytical chemistry? Comparison of nonlinear methods on near infrared (NIR) spectroscopy data, Analyst 136 (2011) 1703–1712.

[19] A. R. Torres, G. M. Xavier, M. L. Paredes, C. L. Cunha, A. S. Luna, R. C. Oliveira, Predicting the properties of biodiesel and its blends using mid-FT-IR spectroscopy and first-order multivariate calibration, Fuel 204 (2017) 185–194.

[20] J. C. L. Alves, C. B. Henriques, R. J. Poppi, Determination of diesel quality parameters using support vector regression and near infrared spectroscopy for an in-line blending optimizer system, Fuel 97 (2012) 710–717.

[21] M. P. F. Da Silva, L. R. E. Brito, F. A. Honorato, A. P. S. Paim, C. Pasquini, M. F. Pimentel, Classification of gasoline as with or without dispersant and detergent additives using infrared spectroscopy and multivariate classification, Fuel 116 (2014) 151–157.

[22] P. A. Pantoja, J. López-Gejo, C. A. O. do Nascimento, G. A. C. L. Roux, Application of Near-Infrared Spectroscopy to the Characterization of Petroleum, Technical Report, 2017. URL: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781119286325.ch8. doi:10.1002/9781119286325.ch8.

[23] C. Pasquini, A. F. Bueno, Characterization of petroleum using near-infrared spectroscopy: Quantitative modeling for the true boiling point curve and specific gravity, Fuel 86 (2007) 1927–1934.

[24] U. Hoffmann, N. Zanier-Szydlowski, Portability of near infrared spectroscopic calibrations for petrochemical parameters, Journal of Near Infrared Spectroscopy 7 (1999) 33–45.

[25] Z.-N. Xing, J.-X. Wang, Y. Ye, G. Shen, Rapid Quantification of Kinematical Viscosity in Aviation Kerosene by Near-Infrared Spectroscopy (2006).

[26] P. Felizardo, P. Baptista, M. S. Uva, J. C. Menezes, M. J. Neiva Correia, Monitoring biodiesel fuel quality by near infrared spectroscopy, Journal of Near Infrared Spectroscopy 15 (2007) 97–105.

[27] D. J. Cookson, B. E. Smith, Calculation of Jet and Diesel Fuel Properties Using 13C NMR Spectroscopy, Energy and Fuels 4 (1990) 152–156.

[28] T. H. DeFries, R. V. Kastrup, D. Indritz, Prediction of cetane number by group additivity and carbon-13 Nuclear Magnetic Resonance, Industrial & Engineering Chemistry Research 26 (1987) 188–193.

[29] W. F. d. C. Rocha, D. A. Sheen, Determination of physicochemical properties of petroleum derivatives and biodiesel using GC/MS and chemometric methods with uncertainty estimation, Fuel 243 (2019) 413–422.

[30] P. Vozka, B. A. Modereger, A. C. Park, W. T. J. Zhang, R. W. Trice, H. I. Kenttämaa, G. Kilaz, Jet fuel density via GC x GC-FID, Fuel 235 (2019) 1052–1060.

[31] B. Creton, C. Dartiguelongue, T. De Bruin, H. H. Toulhoat, Prediction of the Cetane Number of Diesel Compounds Using the Quantitative Structure Property Relationship, Energy & Fuels 24 (2010) 5396–5403.

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

[33] D. A. Saldana, L. Starck, P. Mougin, B. Rousseau, B. Creton, On the rational formulation of alternative fuels: Melting point and net heat of combustion predictions for fuel compounds using machine learning methods, SAR and QSAR in Environmental Research 24 (2013) 525–543.

[34] D. A. Saldana, L. Starck, P. Mougin, B. Rousseau, N. Ferrando, B. Creton, Prediction of density and viscosity of biofuel compounds using machine learning methods, Energy and Fuels 26 (2012) 2416–2426.

[35] F. L. Dryer, Chemical kinetic and combustion characteristics of transportation fuels, Proceedings of the Combustion Institute 35 (2015) 117–144.

[36] S. W. Sharpe, T. J. Johnson, R. L. Sams, P. M. Chu, G. C. Rhoderick, P. A. Johnson, Gas-phase databases for quantitative infrared spectroscopy, Applied Spectroscopy 58 (2004) 1452–1461.

[37] Y. Wang, Y. Cao, D. F. Davidson, R. K. Hanson, Ignition delay time measurements for distillate and synthetic jet fuels, in: AIAA Scitech 2019 Forum, American Institute of Aeronautics and Astronautics, Reston, Virginia, 2019. URL: http://arc.aiaa.orghttps://arc.aiaa.org/doi/10.2514/6.2019-2248. doi:10.2514/6.2019-2248.

[38] R. Q. Casselberry, E. Corporan, M. J. DeWitt, Correlation of combustor lean blowout performance to supercritical pyrolysis products, Fuel 252 (2019) 504–511.

[39] E. Corporan, J. T. Edwards, S. Stouffer, M. DeWitt, Z. West, C. Klingshirn, C. Bruening, Impacts of Fuel Properties on Combustor Performance, Operability and Emissions Characteristics, in: 55th AIAA Aerospace Sciences Meeting, American Institute of Aeronautics and Astronautics, Reston, Virginia, 2017. URL: http://arc.aiaa.org/doi/10.2514/6.2017-0380. doi:10.2514/6.2017-0380.

[40] M. J. Murphy, J. D. Taylor, R. L. McCormick, Compendium of Experimental Cetane Number Data, Technical Report, 2004. URL: http://www.osti.gov/servlets/purl/1086353/. doi:10.2172/1086353.

[41] S. Dooley, S. H. Won, M. Chaos, J. Heyne, Y. Ju, F. L. Dryer, K. Kumar, C. J. Sung, H. Wang, M. A. Oehlschlaeger, R. J. Santoro, T. A. Litzinger, A jet fuel surrogate formulated by real fuel properties, Combustion and Flame 157 (2010) 2333–2339.

[42] S. H. Won, S. Dooley, P. S. Veloo, H. Wang, M. A. Oehlschlaeger, F. L. Dryer, Y. Ju, The combustion properties of 2,6,10-trimethyl dodecane and

a chemical functional group analysis, Combustion and Flame 161 (2014) 826–834.

[43] Engineering ToolBox, Hydrocarbons - physical data, 2017. URL: https://www.engineeringtoolbox.com/hydrocarbon-boiling-melting-flash-autoignition-point-density-gravity-molweight-d{_}1966.html.

[44] B. Smith, Fundamentals of Fourier Transform Infrared Spectroscopy, Second Edition, CRC Press, 2011. URL: https://www.taylorfrancis.com/books/9781420069303. doi:10.1201/b10777. arXiv:arXiv:1011.1669v3.

[45] A. E. Klingbeil, J. B. Jeffries, R. K. Hanson, Temperature-dependent mid-IR absorption spectra of gaseous hydrocarbons, Journal of Quantitative Spectroscopy and Radiative Transfer 107 (2007) 407–420.

[46] Y. Wang, Y. Cao, W. Wei, D. F. Davidson, R. K. Hanson, A new method of estimating derived cetane number for hydrocarbon fuels, Fuel 241 (2019) 319–326.

[47] J. J. Workman, Interpretive spectroscopy for near infrared, Applied Spectroscopy Reviews 31 (1996) 251–320.

[48] G. Socrates, Infrared and Raman characteristic group frequencies, John Wiley & Sons, 2004. doi:10.1002/jrs.1238. arXiv:arXiv:1011.1669v3.

[49] Q. Wang, Y. Hou, W. Wu, M. Niu, S. Ren, Z. Liu, The relationship between the humic degree of oil shale kerogens and their structural characteristics, Fuel 209 (2017) 35–42.

[50] M. R. Riazi, Characterization Petroleum and Properties of Fractions, 2005. URL: http://www.copyright.com/.https://www.cambridge.org/core/product/identifier/CBO9781107415324A009/type/book{_}part.

[51] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2 2 (1995) 1137–1143.

[52] T. Hastie, R. Tibshirani, J. Friedman, Springer Series in The Elements of Statistical Learning, 2009. doi:10.1007/978-0-387-98135-2. arXiv:arXiv:1011.1669v3.

[53] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2019. URL: https://www.R-project.org.

[54] J. Friedman, T. Hastie, R. Tibshirani, Regularization Paths for Generalized Linear Models via Coordinate Descent, Journal of Statistical Software 33 (2015).

[55] N. Simon, J. Friedman, T. Hastie, R. Tibshirani, Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent, Journal of Statistical Software 39 (2015).

Table 3: Wavelengths [nm] and coefficients for average number of carbon atoms. Intercept $\beta_0^* = -1.64$.

	w1	w2	w3	w4	w5	w6	w7	w8	w9	w10	w11
λ^*	3300.4	3308.3	3324.2	3332.2	3348.4	3356.5	3364.7	3381.1	3431.5	3465.9	3492.1
β^*	9.6E-01	-3.2E-01	-5.3E-02	-5.9E-01	1.2E-01	-6.3E-02	-3.5E-02	5.5E-02	5.9E-02	3.2E-01	4.8E-02

Table 4: Wavelengths [nm] and coefficients for average number of hydrogen atoms. Intercept $\beta_0^* = -2.91$.

	w1	w2	w3	w4	w5	w6	w7	w8	w9	w10	w11
λ^*	3300.4	3308.3	3324.2	3332.2	3348.4	3381.1	3414.5	3431.5	3465.9	3474.6	3492.1
β^*	1.3E+00	-5.3E-01	-5.5E-01	-3.3E-01	7.3E-02	8.4E-02	1.1E-02	1.3E-01	3.7E-01	1.7E-02	1.1E-01

Table 5: Wavelengths [nm] and coefficients for MW [g/mol]. Intercept $\beta_0^* = -22.7$.

	w1	w2	w3	w4	w5	w6	w7	w8	w9	w10	w11
λ^*	3300.4	3308.3	3324.2	3332.2	3348.4	3356.5	3364.7	3381.1	3431.5	3465.9	3492.1
β^*	1.3E+01	-4.4E+00	-1.3E+00	-7.4E+00	1.5E+00	-7.6E-01	-4.2E-01	7.4E-01	8.5E-01	4.2E+00	6.9E-01

Table 6: Wavelengths [nm] and coefficients for ratio H/C. Intercept $\beta_0^* = 2.14$.

	w1	w2	w3	w4	w5	w6	w7	w8	w9	w10	w11	w12	w13	w14	w15	w16	w17	w18	w19	w20
λ^*	3300.4	3308.3	3316.2	3324.2	3332.2	3340.3	3348.4	3356.5	3364.7	3372.9	3381.1	3389.4	3397.7	3431.5	3440.0	3448.6	3465.9	3474.6	3483.3	3492.1
β^*	-1.1E-01	3.6E-02	5.4E-02	-7.1E-02	6.1E-02	2.5E-04	-6.6E-03	1.0E-02	1.1E-03	9.7E-03	-1.5E-02	9.4E-03	8.2E-05	-1.1E-02	9.7E-03	-1.6E-03	-3.2E-02	1.1E-02	2.5E-03	-2.6E-04

Table 7: Wavelengths [nm] and coefficients for initial boiling point [°C]. Intercept $\beta_0^* = -20.0$.

	w1	w2	w3	w4	w5	w6	w7	w8	w9	w10	w11	w12
λ^*	3300.4	3308.3	3332.2	3348.4	3356.5	3381.1	3414.5	3431.5	3440.0	3465.9	3474.6	3492.1
β^*	1.3E+01	-5.7E+00	-2.0E+00	1.0E+00	-1.7E+00	3.9E-01	2.8E-01	5.9E-01	5.0E-01	3.6E+00	6.8E-02	6.4E-01

Table 8: Wavelengths [nm] and coefficients for density at 15°C [g/cm^3]. Intercept $\beta_0^* = 0.314$.

	w1	w2	w3	w4	w5	w6	w7	w8	w9	w10	w11	w12	w13	w14	w15	w16	w17
λ^*	3300.4	3308.3	3324.2	3332.2	3340.3	3348.4	3356.5	3364.7	3381.1	3406.1	3414.5	3431.5	3448.6	3457.2	3465.9	3483.3	3492.1
β^*	1.4E+02	-1.0E+02	1.2E+02	-2.1E+02	3.1E+01	2.4E+01	-3.9E+01	-5.9E+00	8.7E+00	3.5E+00	6.0E-02	2.2E+00	4.5E+01	1.6E+01	5.9E+01	2.6E+01	-2.7E-03

Table 9: Wavelengths [nm] and coefficients for Surface Tension at 22°C [dynes/cm]. Intercept $\beta_0^* = 19.7$.

	w1	w2	w3	w4	w5	w6	w7
w	3300.4	3348.4	3356.5	3397.7	3465.9	3474.6	3492.1
c	1.3E-01	-1.6E-03	-3.0E-01	-4.6E-02	4.1E-01	1.2E-01	-2.9E-04

Table 10: Wavelengths [nm] and coefficients for NHC [MJ/kg]. Intercept $\beta_0^* = 48.5$.

	w1	w2	w3	w4	w5	w6	w7
λ^*	3316.2	3332.2	3356.5	3364.7	3389.4	3431.5	3465.9
β^*	-2.9E+02	-6.2E+02	1.9E+02	2.0E+01	1.8E+02	-8.9E+02	-4.6E+02

Table 11: Wavelengths [nm] and coefficients for C2H4 yield. Intercept $\beta_0^* = 5.05$.

	w1	w2	w3	w4	w5	w6	w7	w8
λ^*	3300.4	3324.2	3356.5	3389.4	3440.0	3457.2	3465.9	3474.6
β^*	9.3E+01	-9.8E+02	2.6E+02	-2.0E+02	-1.5E+01	-1.7E+03	7.1E+01	5.8E+02

Table 12: Wavelengths [nm] and coefficients for flash point [°C]. Intercept $\beta_0^* = -51.9$.

	w1	w2	w3	w4	w5	w6	w7
λ^*	3300.4	3316.2	3324.2	3356.5	3448.6	3465.9	3492.1
β^*	4.1E+00	-2.0E+00	1.1E+00	-2.0E+00	1.7E+00	3.4E+00	2.1E-01

Table 13: Wavelengths [nm] and coefficients for LBO. Intercept $\beta_0^* = 0.0762$.

	w1	w2	w3	w4	w5	w6
λ^*	3324.2	3348.4	3356.5	3414.5	3423.0	3474.6
β^*	7.9E-04	5.8E-05	1.3E-05	-6.1E-05	-1.8E-06	1.8E-04

Table 14: Wavelengths [nm] and coefficients for DCN. Intercept $\beta_0^* = 26.7$.

	w1	w2	w3	w4	w5	w6	w7	w8	w9	w10	w11
λ^*	3300.4	3308.3	3348.4	3356.5	3364.7	3381.1	3397.7	3440.0	3448.6	3465.9	3492.1
β^*	-6.9E-02	-1.4E+00	-7.6E-01	9.2E-03	2.8E-01	-1.7E-01	1.7E-01	-6.9E-01	-2.0E-03	5.8E-01	5.6E-01

Table 15: Wavelengths [nm] and coefficients for IDT at 1300K, 4atm [μs]. Intercept $\beta_0^* = 748$.

	w1	w2	w3
λ^*	3340.3	3381.1	3406.1
β^*	9.8E+01	3.0E-01	-3.1E+00

Table 16: Wavelengths [nm] and coefficients for kinematic viscosity at -20°C [mm^2/s]. Intercept $\beta_0^* = -5.21$.

	w1	w2	w3	w4	w5	w6	w7	w8
λ^*	3300.4	3308.3	3316.2	3356.5	3397.7	3431.5	3448.6	3474.6
β^*	3.0E-03	-1.7E-01	-1.6E-01	-2.5E-01	-3.7E-03	-1.5E-01	-1.6E-01	8.9E-01

Table 17: Wavelengths [nm] and coefficients for Total Cycloparaffins [wt %]. Intercept $\beta_0^* = -138$.

	w1	w2	w3	w4	w5	w6	w7	w8	w9	w10	w11
λ^*	3300.4	3308.3	3332.2	3364.7	3372.9	3423.0	3440.0	3448.6	3465.9	3474.6	3483.3
β^*	1.1E+03	4.6E+02	-2.9E+04	-7.6E+03	-7.2E+02	5.0E+03	-8.1E+03	3.9E+04	2.4E+03	1.7E+04	8.1E+03

Table 18: List of fuels and their GC×GC compositions. The labeling of fuels is consistent with [46].

Category	Fuel	POSF	C	H	Total Aromatics [wt %]	Total Cycloparaffins [wt %]	Total iso Paraffins [wt %]	Total n Paraffins [wt %]
A fuel	A1	10264	10.8	21.8	13.4	20.1	39.7	26.8
A fuel	A2	10325	11.4	22.1	18.7	31.9	29.5	20.0
A fuel	A3	10289	11.9	22.6	20.6	47.4	18.1	13.9
A fuel	A4	12784	11.5	22.1	18.6	43.2	23.2	15.1
A fuel	A5	12831	12.1	23.2	18.2	41.4	25.2	15.2
A fuel	A6	12843	11.7	22.4	18.6	42.4	23.8	15.3
A fuel	A7	12905	11.5	22.4	21.2	25.5	29.6	23.8
A fuel	A8	12906	11.4	22.1	17.4	38.4	25.1	19.0
Blend fuel	20%A2-80%C1		12.3	26.0	4.3	7.3	83.6	4.6
Blend fuel	50%A2-50%C1		11.9	24.4	10.1	17.3	61.6	10.9
Blend fuel	80%A2-20%C1		11.6	23.0	15.4	26.3	41.7	16.5
Blend single	BF1		9.4	19.1	20.4	0.0	0.0	79.6
Blend single	BF10		8.4	16.4	47.9	0.0	26.0	26.1
Blend single	BF11		8.2	17.2	0.0	0.0	100.0	0.0
Blend single	BF12		8.0	14.9	29.8	0.0	35.4	34.9
Blend single	BF13		9.2	19.3	15.0	0.0	67.8	17.2
Blend single	BF14		8.6	17.8	38.4	0.0	43.9	17.7
Blend single	BF2		8.8	16.3	40.8	0.0	0.0	59.2
Blend single	BF3		7.6	10.7	80.4	0.0	0.0	19.6
Blend single	BF4		8.2	13.6	60.2	0.0	0.0	39.8
Blend single	BF5		9.6	21.2	100.0	0.0	0.0	0.0
Blend single	BF6		9.2	20.4	0.0	0.0	19.3	80.7
Blend single	BF7		8.8	19.6	0.0	0.0	39.7	60.3
Blend single	BF8		8.4	18.8	0.0	0.0	60.0	40.0
Blend single	BF9		8.0	14.3	0.0	0.0	79.9	20.1
Blend single	Won10		10.6	23.3	0.0	0.0	33.8	66.2
Blend single	Won11		8.7	19.4	0.0	0.0	64.6	35.4
Blend single	Won12		8.9	19.9	0.0	0.0	53.3	46.7
Blend single	Won13		9.2	20.4	0.0	0.0	39.0	61.0
Blend single	Won14		9.5	21.0	0.0	0.0	25.6	74.4
Blend single	Won15		9.9	21.7	0.0	0.0	6.6	93.4
Blend single	Won6		9.0	20.1	0.0	0.0	73.9	26.1
Blend single	Won7		9.4	20.8	0.0	0.0	65.4	34.7
Blend single	Won8		9.8	21.5	0.0	0.0	55.8	44.2
Blend single	Won9		10.1	22.2	0.0	0.0	48.1	51.9
		11498						
C fuel	C1	12368	12.6	27.2	0.0	0.1	99.6	0.0
		12384						
		12344						
C fuel	C4	12489	11.4	24.8	0.4	0.4	98.5	0.2
		12345						
		12713						
C fuel	C5	12789	9.7	18.7	30.7	0.1	51.6	17.7
		12816						
C fuel	C7	12925	12.1	23.9	4.9	62.3	29.5	3.3
C fuel	C8	12923	11.6	21.4	27.3	38.0	21.0	13.7
CN fuel	CN30	13197	11.6	23.1	13.1	12.6	65.0	9.4
CN fuel	CN35	13198	11.4	23.3	10.3	16.9	61.7	11.1
CN fuel	CN40	13199	11.7	23.3	12.8	27.8	47.8	11.6
CN fuel	CN45	13200	11.4	23.1	8.7	30.1	47.0	14.2
CN fuel	CN50	13201	11.1	22.5	8.3	34.8	39.4	17.5
CN fuel	CN55	13202	11.5	23.3	7.4	30.7	34.7	24.4
Pure aromatics	Toluene		7.0	8.0	100.0	0.0	0.0	0.0
Pure cyclo-	Cyclodecane		10.0	20.0	0.0	100.0	0.0	0.0
Pure cyclo-	Cycloheptane		7.0	14.0	0.0	100.0	0.0	0.0
Pure cyclo-	Cyclooctane		8.0	16.0	0.0	100.0	0.0	0.0
Pure n- iso-	2,2-Dimethyl_butane		6.0	14.0	0.0	0.0	100.0	0.0
Pure n- iso-	2,3-Dimethylbutane		6.0	14.0	0.0	0.0	100.0	0.0
Pure n- iso-	3-Methylhexane		8.0	18.0	0.0	0.0	100.0	0.0
Pure n- iso-	3-Methylpentane		6.0	14.0	0.0	0.0	100.0	0.0
Pure n- iso-	Isooctane		8.0	18.0	0.0	0.0	100.0	0.0
Pure n- iso-	n-Decane		10.0	22.0	0.0	0.0	0.0	100.0
Pure n- iso-	n-Dodecane		12.0	26.0	0.0	0.0	0.0	100.0
Pure n- iso-	n-Heptane		7.0	16.0	0.0	0.0	0.0	100.0
Pure n- iso-	n-Hexane		6.0	14.0	0.0	0.0	0.0	100.0
Pure n- iso-	n-Nonane		9.0	20.0	0.0	0.0	0.0	100.0
Pure n- iso-	n-Pentadecane		15.0	32.0	0.0	0.0	0.0	100.0
Pure n- iso-	n-Tridecane		13.0	28.0	0.0	0.0	0.0	100.0
Pure n- iso-	n-Undecane		11.0	24.0	0.0	0.0	0.0	100.0
Pure n- iso-	n-Octane		8.0	18.0	0.0	0.0	0.0	100.0

Table 19: Physical and chemical properties of fuels in the training dataset. The labeling of fuels is consistent with [46].

Category	Fuel	MW	ratio H/C	IBP [43]	Density [43]	ST	NHC	C2H4 yield [46]	FP	LBO [38, 39]	DCN [40, 41, 42]	IDT [46]	KV	Total cyclo
A fuel	A1	151.4	2.019	150	0.7799	23.8	43.1	1.58	42	0.08066	48.61	997.8	3.5	20.08
A fuel	A2	158.9	1.939	159.2	0.803	24.8	43.06	1.69	48	0.08061	48	1044	4.5	31.86
A fuel	A3	165.4	1.899	177.9	0.8268	25.7	43	1.599	60	0.08142	39.07	1059	6.5	47.39
A fuel	A4	160.1	1.922	168			43.1	1.518			41.52	1210	4.9	43.16
A fuel	A5	168.4	1.917	161			43.1	1.248			45.05	1151	6.3	41.4
A fuel	A6	162.8	1.915	173			43.1	1.633			41.91	1088	5.5	42.38
A fuel	A7	160.4	1.948					1.652			49.11	1169		25.48
A fuel	A8	158.9	1.939					1.677			46.34	1055		38.44
Blend fuel	20%A2-80%C1	173.9	2.112	168.8	0.768		43.78	0.821	50	0.08462	23.86		4.7	7.325
Blend fuel	50%A2-50%C1	167.8	2.045	162.5	0.781		43.5	1.035	50	0.08311	33.28		4.5	17.31
Blend fuel	80%A2-20%C1	162.3	1.98	158.3	0.795		43.3	1.358	48	0.08178	41.78		4.5	26.32
Blend single	BF1	131.8	2.039								62.14			0
Blend single	BF10	117.3	1.955								43.27			0
Blend single	BF11	115.5	2.098								32.99			0
Blend single	BF12	110.5	1.865								31.66			0
Blend single	BF13	129.9	2.091								57.14			0
Blend single	BF14	120.8	2.075								45.52			0
Blend single	BF2	121.6	1.856								54.14			0
Blend single	BF3	101.8	1.416								28.58			0
Blend single	BF4	111.9	1.656								44.18			0
Blend single	BF5	136.6	2.208								61.21			0
Blend single	BF6	130.9	2.217								55.08			0
Blend single	BF7	125.2	2.227								46.93			0
Blend single	BF8	119.6	2.238								36.59			0
Blend single	BF9	110.8	1.772								36.93			0
Blend single	Won10	151.1	2.188								65			0
Blend single	Won11	123.9	2.23								45			0
Blend single	Won12	127.1	2.224								50			0
Blend single	Won13	131.1	2.217								55			0
Blend single	Won14	134.8	2.211								59.1			0
Blend single	Won15	140.2	2.203								65			0
Blend single	Won6	128.6	2.221								45			0
Blend single	Won7	133.5	2.213								50			0
Blend single	Won8	138.8	2.205								55			0
Blend single	Won9	143.1	2.198								59.1			0
C fuel	C1	178.4	2.159	174.3	0.7597	23.4	43.82	0.468	49.5	0.08686	17.1	2513	5	0.05
C fuel	C4	161.6	2.175	161.5	0.7592	22.7	43.81	0.971	44.5	0.08477	28	1711	3.87	0.43
C fuel	C5	135.1	1.928	156.6	0.7689	23.8	43.01	1.764	43.5	0.08248	39.6	1264	1.96	0.07
C fuel	C7	169.1	1.975	184	0.8181	26.1		1.528	64		42.6	939	6.53	62.31
C fuel	C8	160.6	1.845	170	0.8238	26.5		1.254	56		43.5	922	4.84	37.97
CN fuel	CN30	162.3	1.991					0.915			30	1822		12.55
CN fuel	CN35	160.1	2.044					0.7946			34	1551		16.93
CN fuel	CN40	163.7	1.991					1.193			40	1390		27.83
CN fuel	CN45	159.9	2.026					1.328			44	1210		30.14
CN fuel	CN50	155.7	2.027					1.937			51	937.8		34.81
CN fuel	CN55	161.3	2.026					1.65			54	906		30.74
Pure aromatics	Toluene	92	1.143	110.6			40.59		6		6			0
Pure cyclo-	Cyclodecane	140	2	201	0.857									100
Pure cyclo-	Cycloheptane	98	2	118.4	0.81									100
Pure cyclo-	Cyclooctane	112	2	149	0.831						22.3			100
Pure n- iso-	2,2-Dimethylbutane	86	2.333	50	0.649						24.4			0
Pure n- iso-	2,3-Dimethylbutane	86	2.333	58										0
Pure n- iso-	3-Methylhexane	114	2.25	92	0.687						42			0
Pure n- iso-	3-Methylpentane	86	2.333	63	0.66						30.7			0
Pure n- iso-	Isooctane	114	2.25	99			44.31				17.5			0
Pure n- iso-	n-Decane	142	2.2	174	0.73	23.83	44.24		46					0
Pure n- iso-	n-dodecane	170	2.167	216	0.75	25.35	44.15		80	0.07701	73			0
Pure n- iso-	n-Heptane	100	2.286	98	0.683	20.14	44.57		-7	0.08021	53.5			0
Pure n- iso-	n-Hexane	86	2.333	69	0.664	18.43	44.75				49			0
Pure n- iso-	n-Nonane	128	2.222	151	0.719	22.85	44.31		31		60.9			0
Pure n- iso-	n-Pentadecane	212	2.133	270	0.769						96			0
Pure n- iso-	n-Tridecane	184	2.154	234	0.756						90			0
Pure n- iso-	n-Undecane	156	2.182	196	0.74	24.66	44.19		61		83			0
Pure n- iso-	Octane	114	2.25	126	0.702	21.61	44.43		12		58.2			0

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
j.fuel.2019.115715		

4. Title and Subtitle	5. Report Date
On estimating physical and chemical properties of hydrocarbon fuels using mid-infrared FTIR spectra and regularized linear models	July 2019
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
Yu Wang, Yiming Ding, Wei Wei, Yi Cao David F. Davidson Ronald K. Hanson	

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
Center of Excellence for: Alternative Jet Fuels and Environment Stanford University Stanford CA 94305	11. Contract or Grant No. 13-C-AFJE-SU-15

12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
Federal Aviation Administration Office of Environment and Energy	14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

The concept of a compact, economical FTIR-based analyzer for estimating the properties of hydrocarbon fuels with small amounts of fuel is proposed. The high correlations between mid-IR FTIR absorption spectra of fuel vapor in the range 3300–3550 nm and 15 physical and chemical properties, such as density, initial boiling point, surface tension, kinematic viscosity, number of carbon and hydrogen per average molecule, and derived cetane number, for 64 hydrocarbon fuels are demonstrated. Lasso-regularized linear models based on linear combination of absorption cross sections at selected wavelengths are built for each of these physical and chemical properties, yielding accurate estimations.

17. Key Words	18. Distribution Statement
Hydrocarbon fuel, Alternative fuel, Physical and chemical property, Combustion, Mid-IR spectroscopy, Generalized linear model, Machine learning, ASCENT	

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified		