Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

School of Physics

New Physics-Based Methods for Learning Sustainable Aviation Fuel Certification Fit-for-Purpose Properties

Robert Parker
20331754

A report submitted to the School of Physics for Master's Thesis

Under the supervision of Dr. Stephen Dooley

2026

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Declaration concerning plagiarism

I have read and I understand the plagiarism provisions in the General Regulations of the *University Calendar* for the current year, found at http://www.tcd.ie/calendar

I have completed the Online Tutorial in avoiding plagiarism 'Ready, Steady, Write', located at **http://tcd-ie.libguides.com/plagiarism/ready-steady-write**

STUDENT NUMBER: 20331754

SIGNED: Robert Parker

Abstract

The development of predictive models for key physical properties of aviation fuels using atom type mole fractions derived from ^{1}H ^{13}C HSQC NMR spectroscopy was investigated. The approach enables property estimation from small sample volumes, supporting the prescreening of sustainable aviation fuel (SAF) candidates.

Datasets were compiled for six properties: liquid density, surface tension, flash point, distillation curve, derived cetane number, and kinematic viscosity. Experimental and literature data provided 1,241 points for density, 1,260 for surface tension, 1,074 for distillation curve, 405 for derived cetane number, 184 for flash point, and 1,373 for viscosity. Models were developed using multiple linear regression, binomial regression, random forest regression, and artificial neural networks, with validation against unseen test data.

The density and surface tension models achieved the highest predictive accuracy, with mean errors of 0.47 % and 2.22 % respectively. The flash point and distillation curve models also performed well, showing mean errors below 1 % and R^2 values above 0.97. The derived cetane number model showed lower general correlation ($R^2 < 0.95$) but strong agreement for real-fuel validation. The viscosity model, developed with an artificial neural network, achieved a mean error of 3.61 %.

Overall, atom type based modelling demonstrates strong potential for efficient SAF property prediction, demonstrating that the properties are physical by nature. Future improvements will rely on expanding datasets to cover more complex mixtures and compositional ranges.

Acknowledgements

I would like to express my sincere gratitude to Professor Stephen Dooley for his guidance, insight, and continuous support throughout this research. I am grateful to Professor Mohammad Reza Ghaani for his valuable advice, encouragement, and contributions to the development of this work. I wish to thank the School of Physics at Trinity College Dublin for providing the resources, facilities, and academic environment that made this research possible. I would also like to acknowledge Conall McNamara, Alex Weiming Liu, Ailís O'Shea, Liam Mannion, Aron Bell, Tiarnán Watson-Murphy, Harry Dunne, Prajwal Rao and Mark Kelly from the low carbon technologies research group for their collaboration, technical discussions, and shared expertise. Finally, I extend my appreciation to all colleagues and collaborators whose assistance and contributions have been integral to the completion of this study.

List of Publications

1. Parker, R. P.; Kelly, M.; Watson-Murphy, T.; Ghaani, M. R.; Dooley, S. *Predictive Modeling of Liquid Density and Surface Tension for Sustainable Aviation Fuels Using Nuclear Magnetic Resonance Atom Types.* Energy & Fuels, 2025, 39 (7), 3690–3702. DOI: 10.1021/acs.energyfuels.4c05601

2. Bell, A.; Mannion, L. A.; Kelly, M.; Parker, R.; Ghaani, M. R.; Dooley, S. *Life Cycle CO_2e Intensity of Power-to-Liquid Sustainable Aviation Fuel Scenarios and Specific Use Cases.*
   Advances in Applied Energy, 2025, 20, 100248.
   DOI: 10.1016/j.adapen.2025.100248

3. Boehm, R. C.; Parker, R.; Yang, Z.; Dooley, S.; Heyne, J. S. *Blend Prediction Model for Vapor Pressure of Jet Fuel Range Hydrocarbons.* Sustainability, 2025, 17 (21), 9612.
   DOI: 10.3390/su17219612

List of Presentations

R. Parker, M. Kelly, T. Watson-Murphy, M. R. Ghaani, S. Dooley.
A Nuclear Magnetic Resonance Method for Tier α Prescreening of Sustainable Aviation Fuel Candidates.

- Oral Presentation, 1st International Conference on Sustainable Aviation Research**,** Dublin, Ireland**,** 8 July 2025**.**
- Oral Presentation, Coordinated Research Council Annual Meeting, Dayton, Ohio, USA, 29 April 2025**.**
- Oral Presentation, 18th International Conference on Stability, Handling and Use of Liquid Fuels, Louisville, Kentucky, USA, 12 September 2024.
- Poster Presentation**,** MaREI Conference, University College Cork, Ireland, 20 June 2024.

Table of Contents

List of Figures

most components in jet fuel are higher in carbon number than C7, which means they have higher boiling points and therefore less deviation from linearity at the highest relevant temperatures. The data presented in the figure is from NIST Thermodynamic Data for Pure Compounds.[114]

Figure 13. Experimental density temperature-dependence of Jet-A POSF 10325 and JP-5 POSF 10289. This plot shows that the temperature dependences of fuel density for these real jet fuels is essentially linear over the temperature range considered. The data present in the figure is from Edwards et al.

Figure 14. Example of a decision tree as part of a random forest, showing the first three layers of decisions for a model on kinematic viscosity.

Figure 15. Example of an artificial neural network model architecture.

Figure 16. Atom type mole fraction on a log-radar plot for select fossil aviation fuels and synthetic aviation fuels, showing considerable compositional differences. The axis represents the mass fraction of each atom type in the sample.

Figure 17. Atom type mass fractions of Gevo POSF 13718, S-8 POSF 5018, Amyris Farnesane POSF 11832, and World Energy HEFA H101 on a bar chart log plot with Jet A POSF 10325 for direct comparison with a conventional aviation turbine fuel.

Figure 18. Temperature dependent surface tension, predicted surface tension versus actual surface tension with an 80:20 split in training (unseen data, hollow symbols) and testing (seen data, full symbols) data respectively for normal paraffins.

Figure 19. Temperature dependent surface tension, predicted surface tension versus actual surface tension with an 80:20 split in training (unseen data, hollow symbols) and testing (seen data, full symbols) data respectively for iso-paraffins and normal paraffins.

Figure 20. Temperature dependent surface tension, predicted surface tension versus actual surface tension with an 80:20 split in training (unseen data, hollow symbols) and testing (seen data, full symbols) data respectively for cycloparaffins, iso-paraffins and normal paraffins.

Figure 21. Temperature dependent surface tension, predicted surface tension versus actual surface tension with an 80:20 split in training (unseen data, hollow symbols) and testing (seen data, full symbols) data respectively for aromatics, cycloparaffins, iso-paraffins and normal paraffins.

Figure 22. Temperature dependent surface tension, predicted surface tension versus actual surface tension for unseen testing data respectively for aromatics, cycloparaffins, iso-paraffins and normal paraffins using group atom type firewall methodologies.

Figure 23. Temperature dependent surface tension, predicted surface tension versus actual surface tension for training and unseen testing data, showing the model has high fidelity for training and testing case.

Figure 24. Temperature dependent liquid density, predicted surface tension versus actual liquid density for training and unseen testing data, showing the model has high fidelity for training and testing case.

Figure 25. Flash point, predicted versus actual flash point for training data only showing the model has high fidelity to training and testing case.

Figure 26. Parity plots for volume distillates required for prescreening: 10%, 50%, and 90%. The prescreening model uses volume as a parameter so overall there are 1,074 boiling points in the model.

Figure 27. R^2 reduction analysis of random forest number of decision trees. Number of estimators represents the decision trees, showing that the random forest converges at 36.

Figure 28. Derived Cetane Number (DCN), predicted DCN versus actual DCN for 405 experimental measurements.

Figure 29. Predicted kinematic viscosity against actual kinematic viscosity as computed from artificial neural network.

Figure 30. Surface tension versus temperature, literature experiment (full symbols) and unseen model predictions of this study (hollow symbols). For reference the ASTM fit-for-purpose properties are included as lines, which are derived from the CRC World Fuel Survey (2006). The bars represent the range of surface tensions measured for samples of Jet A, JP-5 and JP-8 in the survey.

Figure 31. Liquid density versus temperature, literature experiment (full symbols) and unseen model predictions of this study (hollow symbols). The ASTM fit-for-purpose properties are included as lines for reference. The ASTM D1655 specification is used as a qualification (limits @ 15°C) metric in the initial ASTM D4054 evaluation.

Figure 32. Flash point predictions of conventional and synthetic aviation turbine fuel. The error bars represent the ASTM D56 reproducibility of flash point via the closed cup method.

The predicted error range is 1.74 – 5.64%. The experimental data was taken from Edwards et al.[115]

Figure 33. Property – Property correlation matrix for flash point, freezing point, boiling point, density and surface tension using physical property data gathered from DIPPR database.

Figure 34. Temperature versus volume, literature experiment (full symbols) and unseen model predictions of this study (hollow symbols). The ASTM fit-for-purpose properties are included as lines for reference.

Figure 35 Derived Cetane Number validation data set of real fossil and synthetic aviation turbine fuels using atom types quantified from ^{1}H ^{13}C HSQC NMR spectroscopy. The experimental data was taken from Edwards et al.[115]

Figure 36. Kinematic viscosity validation data set of real fossil and synthetic aviation turbine fuels using atom types quantified from ^{1}H ^{13}C HSQC NMR spectroscopy. The experimental data was taken from Edwards et al.[115]

Figure 37. Average percentage error in the model prediction of temperature dependent density and temperature dependent surface tension of Jet-A POSF 10325 and JP-5 POSF 10289, relative to experiment, for different descriptions of the compositional space. The y-axis denotes the singular absence of each molecular class from the training data.

List of Tables

Glossary

SAF	Sustainable Aviation Fuel
ASTM	Previously, American Society of Testing and Materials. Now, ASTM International, with no formal definition of ASTM
DCN	Derived Cetane Number
OEM	Original Equipment Manufacturer
HSQC	Heteronuclear Single Quantum Coherence
NMR	Nuclear Magnetic Resonance
GC	Gas Chromatography
GCxGC	Two Dimensional Gas Chromatography
VUV	Vacuum Ultraviolet
FID	Flame Ionization Detection
FTIR	Fourier Transform Infrared Spectroscopy
AARD	Absolute Relative Deviation
ARE	Average Relative Error
QSPR	Quantitative Structure Property Relationship
MLR	Multilinear Regression
ANN	Artificial Neural Network
DEPT	Distortionless Enhancement by Polarisation Transfer
CRC	Coordinated Research Council
FACE	Fuels for Advanced Combustion Engines
GC-MS	Gas Chromatograph – Mass Spectrometer
RON	Research Octane Number
MON	Motor Octane Number
NIST	National Institute of Standards and Technology
AIChE	American Institute of Chemical Engineers
DIPPR	Design Institute for Physical Properties
MAPE	Mean Average Percentage Error
R^2	Coefficient of Determination
SPK	Synthetic Paraffinic Kerosene
HEFA	Hydroprocessed Esters and Fatty Acids
IPK	Isoparaffinic Synthetic Kerosene
IQT	Ignition Quality Tester
NJFCP	National Jet Fuels Combustion Program

1. Introduction

Sustainable aviation fuels (SAFs) or more formally, synthetic aviation turbine fuels, produced using carbon neutral feedstocks, are important in mitigating the environmental impact of aviation, reducing greenhouse emissions, and enabling long-term sustainability in the industry. Introducing new aviation fuels and fuel additives in accordance with ASTM D4054[1] poses challenges for fuel producers due to the number of fuel property measurements and the high volume of fuel required by the evaluation process, presenting a high economic barrier to entry and a deterrent for investment.

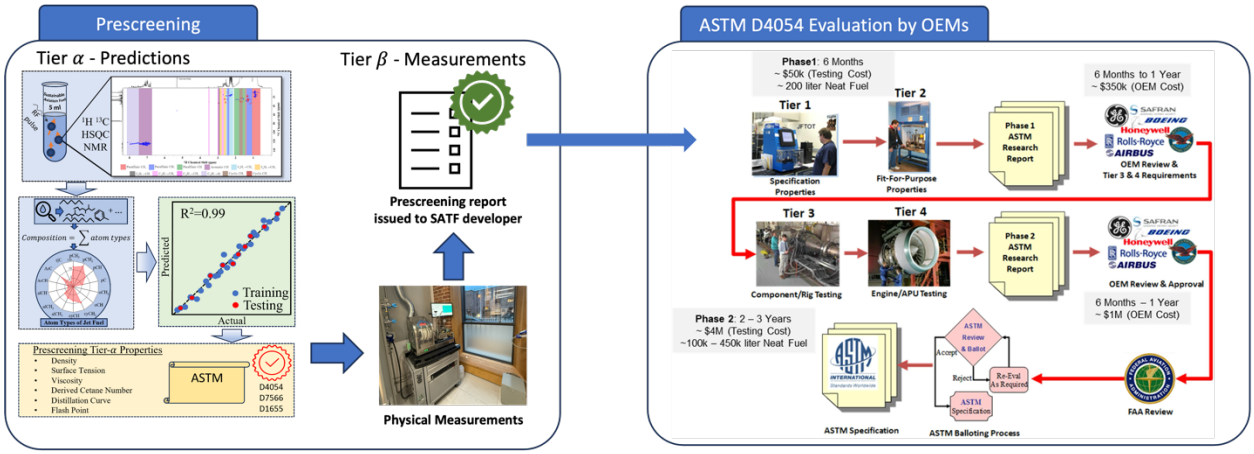


Figure 1. Process flowchart for a candidate sustainable aviation fuel, where prescreening is a tool used to develop confidence in developer/producer product conformity to successful D4054 OEM evaluation.[1, 126]

Prescreening is a novel technique where the properties of the SAF candidates can be predicted using statistical regression or other mathematical models based on some form of compositional information, this process is shown in Figure 1. Heyne et al.[2] have summarized six key properties to which prescreening should be first applied to aid in eventual fuel qualification. These include density at 15°C, surface tension at 22°C, viscosity at -20 and -40°C, Derived Cetane Number (DCN), Distillation Curve and Flash Point. Density, defined as the ratio of mass to volume, is a fundamental property that serves as an indicator for several other characteristics including specific energy, viscosity, heat of combustion and atomization.[3] Surface tension, the energy required to increase the surface area of a liquid, arises from the balance between cohesive forces that bind molecules within the liquid and adhesive forces at the liquid–interface boundary.[3] This property plays a critical role in determining droplet size and atomization dynamics in gas turbine combustors, thereby influencing both efficiency and

operability.[2] Viscosity, a measure of a liquid's resistance to flow, further affects the ease of fuel handling, loading and atomization.[3] The derived cetane number, obtained from ignition delay time, provides a widely used metric for assessing fuel quality and its propensity to knock.[4] Complementing this, the distillation curve characterizes the boiling point distribution of a fuel, offering insight into its chemical composition and volatility profile.[5, 6] Finally, flash point, the minimum temperature at which a fuel generates a flammable vapor phase above the liquid surface, constitutes a key safety parameter governing storage, handling and transportation.[7]

As SAFs are intended to serve as 'drop-in' replacements for conventional fuels, ASTM D4054[1] and ASTM D7566[8] outline specific fuel properties and acceptable range limits that a SAF must exhibit if the fuel is to be approved for use with aviation original equipment manufacturer (OEM) airframes and engines. The outlined properties are of primary importance in this context and are therefore central to the focus of the research undertaken.

In the context of SAF evaluation and approval according to ASTM D4054, prescreening methods that require minimal fuel quantities are advantageous, as they reduce costs while providing sufficient data for accurate property predictions. Heteronuclear Single Quantum Coherence (HSQC) Nuclear Magnetic Resonance (NMR) spectroscopy is a promising technique for such applications due to its non-destructive nature and capability to analyse complex mixtures. NMR spectroscopy probes molecular structures by utilizing the magnetic properties of specific atomic nuclei. For hydrocarbon fuels, HSQC-NMR enhances traditional NMR by correlating the chemical shifts of protons (^{1}H) with those of heteronuclei, such as carbon (^{13}C) or nitrogen (^{15}N), enabling detailed structural analysis. This method is highly effective for identifying and quantifying distinct atom types, where each atom type is defined by its unique electronic environment within the molecular structure of a liquid fuel.

For fuel analysis, ^{1}H ^{13}C HSQC-NMR offers several distinct advantages that are complementary to chromatographic and infrared spectroscopy methods. While certain gas chromatography detectors are capable of sub-ppm detection sensitivity, ^{1}H ^{13}C HSQC-NMR achieves similar sensitivity without requiring extensive sample preparation, analyte

derivatization, or calibration, which is often necessary in gas chromatography analyses. Although infrared spectroscopy can face challenges with highly complex mixtures due to overlapping spectral features, it remains a powerful tool for functional group characterization. Additionally, ^{1}H ^{13}C HSQC-NMR allows for in situ analysis and provides valuable structural information about molecular interactions within complex mixtures. Compared to gas chromatography (GC) and infrared spectroscopy, NMR spectroscopy generally involves higher initial costs due to instrument acquisition and maintenance.

In this study, ^{1}H ^{13}C HSQC-NMR is employed to quantify atom types in hydrocarbon aviation turbine fuel-like mixtures, offering a detailed breakdown of the mole fraction of each atom type. This information is of utility in understanding the composition and properties of the fuel. The ability of ^{1}H ^{13}C HSQC-NMR to handle the inherent complexity of hydrocarbon mixtures, while also providing both compositional and structural insights, makes it a valuable tool in fuel research and development.

2. Literature Review

2.1 Group Contribution Models

Liquid density and surface tension predictive models for small samples of aviation fuel using an array of analytical techniques are available in literature.[9-15] Yang et al.[15] used GCxGC to determine the hydrocarbon group-type composition of SAF, relying on group elution patterns based on volatility and polarity. Monte Carlo sampling was used to bound the range of properties of SAF. This method reported a coefficient of determination (R^2) of 0.93 for density at 15 ℃ and a R^2 of 0.64 for surface tension at 22 ℃ for 20 predicted SAF values. Heyne et al.[10] demonstrated that incorporating vacuum ultraviolet (VUV) identification into GCxGC analysis reduced the uncertainty in various fuel property predictions, with confidence intervals reduced up to 21 times compared to GCxGC with flame ionization detection (FID), though R^2 values were not reported. Wang et al.[14] has used Fourier Transform Infrared Spectroscopy (FTIR) to characterize the average fuel structure in terms of functional groups and use linear methods to estimate hydrocarbon fuel properties. This method reported a R^2 of 0.83 for density at 15 ℃ and a R^2 of 0.85 for surface tension at 22℃.

Group contribution methods are techniques that estimate the properties of pure components through the summation of contributions from specific molecular groups or structures within the compound.[16-19] Predictive models for density that utilize group contribution methods already exist in the literature.[20-25] Skander et al.[23] conducted a study which used molecular structures and a correction term based on the relative structure location to CH_3 groups and carbon atom placement in ring structures for the estimation of physical properties of hydrocarbons, including liquid density at 20 °C. The resulting model predicted liquid density of all hydrocarbon classes with an Average Absolute Relative Deviation (AARD) of 1.3%. Wakeham et al.[25] utilized carbon atoms/bonds and molecular energies within hydrocarbons to estimate the liquid density and critical properties with an Average Relative Error (ARE) of 1.17% for liquid density at 20 °C.

Similarly, group contribution methods for the estimation of surface tension are available in the literature.[21, 26-31] Gharagheizi et al.[27] conducted a study where the surface tension for ionic liquids were computed based on their chemical structures. These structures were defined by

the anions and cations within the ionic liquids. After evaluating 34 different chemical structures, they found that only 19 were effective for accurately estimating surface tension. The model's performance was characterized by an AARD of 3.20%. This study of Gharagheizi et al.[27] represents the first instance in literature where surface tension estimation of a liquid relied solely on group contributions and temperature. Building upon this approach, Shahsavari et al.[31] predicted the surface tension of ionic liquids using a group contribution method based on the amount of each constituent element present, rather than chemical structures. The resulting model displayed an AARD of 4.16%. These studies demonstrate that estimating surface tension through group contributions is feasible, however, its application is limited to known samples with discrete quantifiable chemical structures.

Flash point is a complex property based on the volatility of a sample, there are group contribution models and predictive methods in the literature.[15, 32-43] Pan et al.[36] developed group contribution and quantitative structure property relationship (QSPR) models for aliphatic and aromatic hydrocarbons, comparing multilinear regression (MLR) with artificial neural networks and finding that nonlinear methods improved accuracy. Atabati et al.[41] employed stepwise multilinear regression on molecular connectivity indices of alkanes, demonstrating reliable agreement with experimental data. Dai et al.[39] proposed group contribution-based models for alkanes using both MLR and artificial neural networks (ANN) approaches, achieving low average deviations. Saldana et al.[34] extended structure-based prediction to fuel mixtures, developing a machine-learning model paired with a novel equation to estimate flash points of fuel blends based on component structures and mole fractions.

DCN models based on molecular structure/composition exist in the literature[40, 44-48]. Li et al.[48] introduced an machine-learned regression group contribution framework that predicts cetane numbers for pure hydrocarbons and mixtures from fragment counts, establishing a structural baseline. Al Ibrahim and Farooq[47] used infrared spectral features assembled from pure-component spectra to predict DCN of hydrocarbon mixtures and real fuels, demonstrating mixture-aware, composition-driven accuracy. Dalmiya et al.[45] extended this idea by using liquid-phase IR absorption of jet fuels and functional-group surrogates to model DCN. Flora et al.[44] combined GCxGC-resolved composition with a QSPR/Machine learning model to

predict DCN for hydrocarbon species and blends representative of practical aviation fuels. Broader structure-based machine learning has also been shown effective: graph neural networks mapping molecular graphs to ignition-quality metrics including DCN.[40]

Viscosity prediction models based on molecular structure and composition exist in the literature[15, 21, 33, 40, 49-64]. Cai et al.[62] built a QSPR using group/fragment descriptors to predict temperature-dependent liquid viscosity of diverse hydrocarbons. Jovanović et al.[57] proposed a free-volume-based group-contribution model specific to n- and iso-alkanes. Lötgering-Lin and Gross[65] introduced entropy-scaling and group-contribution viscosity correlations that connect structure to transport via EOS-derived residual entropy. Panwar et al.[53] used Gaussian process regression with structural derived descriptors to model temperature-dependent viscosity of complex hydrocarbon base oils. For real fuels, Cáceres-Martínez et al. developed a least squares model that predicts jet-fuel kinematic viscosity at −20 °C from composition-informed features.[63]

Distillation curve prediction models have been widely developed in the literature[66-75]. Bell et al.[72] introduced general surrogate composition calculations that incorporated blending rules for the distillation curve, providing a framework to link molecular composition to bulk volatility behaviour. Dobbelaere et al.[75] use machine learning starting from paraffins, isoparaffins, olefins, naphthenes, aromatics, and oxygenates composition, and reconstruct molecule-level mixture representations to predict full distillation profiles. Hosseinifar et al.[69] proposed a predictive method for petroleum fluids that constructs ASTM D86 and true-boiling-point curves from limited data using a six-point input scheme. Pasadakis et al.[66] demonstrated that mid-infrared spectroscopy combined with artificial neural networks can accurately predict detailed diesel distillation profiles at 5% volume intervals. Bolanca et al.[76] further showed that FTIR and Raman spectroscopy paired with neural networks are effective for estimating key distillation parameters (T10, T50, T90), highlighting the utility of vibrational spectroscopic descriptors in volatility prediction. There is a noticeable gap in literature for group contribution methods for boiling point distribution profiles with the focus being on a molecular, as opposed to sub molecular, level.

While group contribution methods utilizing atom types, as proposed in this study, have been previously employed in the literature,[77, 78] their application to the prediction of the critical properties under study in complex fuel-like mixtures remains unexplored. In this regard, Prof. Dooley's research group have previously developed the ^{1}H ^{13}C HSQC NMR methodology for the purpose of atom type analysis of fuels.[77-80] It has been shown that this methodology can identify and precisely quantify discrete atom types by integrating peaks on the ^{1}H spectrum and assigning with assistance from the HSQC spectrum to identify the atom types.[80] A major gap in the literature, which this work seeks to address, is the absence of property prediction models trained on pure hydrocarbons using atom type mole fractions as predictors.

2.2 Nuclear Magnetic Resonance Spectroscopy of Complex Hydrocarbon Fuels

Nuclear Magnetic Resonance spectroscopy has become an invaluable tool for parsing the detailed molecular makeup of complex hydrocarbon fuels.[51, 77, 78, 80-110] Early foundational efforts used ^{1}H and ^{13}C NMR to decompose gasoline and diesel into functional group distributions. Bansal et al.[111] employed ^{1}H NMR alongside HSQC to estimate total aromatic content in diesel range products, distinguishing between mono-aromatics and polyaromatics. Subsequent work by Bansal et al.[106] applied HSQC and ^{13}C distortionless enhancement by polarisation transfer (DEPT) techniques to residual fluid catalytic cracking feedstocks, enabling semi-quantitative evaluation of aromatic content. Burger et al.[104] expanded this "composition-explicit" approach using the Advanced Distillation Curve method: they obtained distillate fractions from jet fuels, then used ^{1}H and ^{13}C NMR to quantify structural classes (e.g., aromatics, linear alkanes) in each fraction, thereby uncovering subtle shifts in branching across the distillation range. These studies collectively established NMR as a potent technique for discerning average structural features in fuel samples.

In 2019, Ure et al.[82] advanced the methodology with the most rigorous and quantitative NMR protocol for fuel analysis to date. Their work presented a careful workflow for collecting fully quantitative ^{1}H and ^{13}C NMR spectra of liquid hydrocarbon fuels, including the choice of solvent and relaxation delay times to ensure accuracy. The novel technique integrates the ^{1}H spectrum using HSQC spectroscopy to expedite region-specific integrals and resolve overlapping signals, thereby inferring carbon atom type quantification and assignment without

using a quantitative carbon spectrum. Validation against reference fuels of known composition demonstrated that their method could reproduce atom type distributions to within a 3% uncertainty (excluding overlapping peak effects). Application to CRC Fuels for Advanced Combustion Engines (FACE) Gasoline A, C, F, and I showed strong agreement with the Detailed Hydrocarbon Analysis reported by Daly et al.[112] A parity plot comparing the fuels examined in both studies is presented in Figure 2, alongside model fuels formulated at Trinity College Dublin for extremity testing and additional validation. Comparisons of this type are uncommon for novel methodologies such as the one employed here, because of the need for measurement of the same sample in both laboratories, and the equipment needed to perform the experiments. They are further complicated by the requirement for complete molecular characterisation of the fuel to enable direct comparison with NMR-derived atom types.

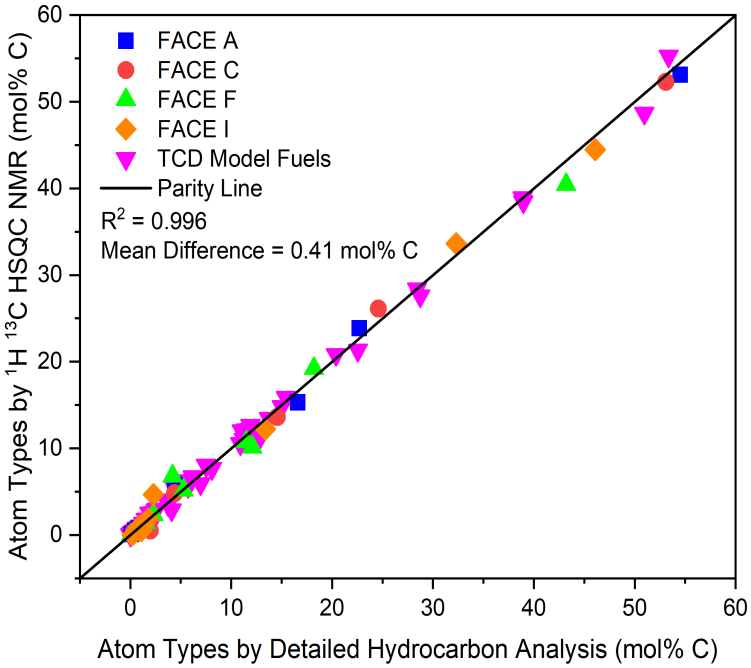


Figure 2. Atom types by ^{1}H ^{13}C HSQC-NMR vs. Atom Types by Detailed Hydrocarbon Analysis to examine the fidelity of the NMR methodology. With an R^2 of 0.996 and a Mean Difference of 0.41 mol% the methodology has high fidelity to GC-MS methods.

Figure 2 demonstrates that the ^{1}H ^{13}C HSQC NMR spectroscopy method for determining and quantifying atom types in complex mixtures agrees closely with GC-MS detailed hydrocarbon analysis, with R^2 of 0.996 and a mean difference of 0.41 mol%. A more detailed comparison for FACE A is provided in Table 1, which reports the differences in mol% carbon

between HSQC-NMR and GC-MS for each atom type, along with the corresponding absolute deviations. The smallest deviation observed is 1.39 mol% carbon for paraffinic methyl.

Other researchers have produced complementary NMR based analyses tailored to surrogate formulation and fuel property prediction. For example, Mondal et al.[88] developed a rapid ^{1}H NMR method to estimate paraffins, olefins, naphthenes, aromatic content in narrow-cut naphtha samples, similar in approach to Ure et al.'s atom-type quantification. Jameel et al.[108] employed high-resolution ^{1}H NMR to identify and quantify six hydrocarbon functional groups (various paraffinic CH_3, CH_2, CH; naphthenic CH–CH_2; aromatic C–CH) in gasoline, subsequently using these as inputs to predict key combustion behaviours such as RON, MON, DCN, and sooting indices with strong accuracy. Jameel et al.[77] earlier demonstrated that functional group distributions derived from ^{1}H NMR could effectively predict ignition quality (e.g., cetane number) using multilinear regression models.

Table 1. Detailed comparison of atom type carbon fractions in FACE A determined by ^{1}H ^{13}C HSQC NMR spectroscopy and GC-MS, including absolute differences for each atom type. The smallest difference observed is 1.39 mol% carbon for paraffinic methyl.

Atom Types		**FACE A**		
		HSQC-NMR (mol% C)	**GC-MS (mol% C)**	**Absolute Difference (mol% C)**
Paraffinic	1. CH_3	53.11	54.50	1.39
	2. CH_2	23.87	22.70	1.17
	3. CH	15.32	16.60	1.28
	4. C	5.77	4.40	1.37
Olefinic	5. CH_2	n/a	n/a	n/a
	6. CH	n/a	n/a	n/a
Cyclic	7. cyCH_2	0.82	0.80	0.02
	8. cyCH	0.41	0.40	0.01
Aromatic	9. CH	0.51	0.50	0.01
	10. C	0.10	0.10	0.00
Alpha Aromatic	11. CH_3	0.10	0.10	0.00
	12. CH_2	n/a	n/a	n/a
	13. CH	n/a	n/a	n/a

Polyaromatic	14. Junction	n/a	n/a	n/a

2.3 Validation Data for Model Verification

Validation data is essential for predictive models, as it serves as unseen testing data that enables verification of a model's generalisability and provides a means of assessing performance under limiting conditions. As previously discussed, the models proposed in this study, those using atom type to predict properties of jet aviation fuel, are inherently constrained by the availability of validation data. This limitation arises from the absence of literature addressing atom types using the NMR methodology applied here.

Several approaches can be taken to develop a database of validation data. One option is to apply blending rules to simulate complex mixtures. Another is to acquire samples with published experimental values that allow quantification of atom types. Alternatively, surrogate mixtures can be constructed in which atom types are known and subsequently characterised through experiment. A further possibility is to obtain samples for which atom types can be quantified and their properties determined experimentally. The four approaches outlined here will be examined in detail below.

Blending rules have been extensively investigated in the literature, with particular value demonstrated in prescreening through GCxGC methodologies.[2, 10, 15, 72, 113] The principle of a blending rule is to use proportionality or an experimentally derived parameter to estimate how two or more compounds will mix to produce a new property. For example, density can be estimated by assuming linear mixing on a volumetric basis, such that molecules with larger volumes exert a greater influence on the resulting density.[114] This approach is especially relevant to GCxGC analysis, where compositional information is obtained by identifying possible compounds in the fuel via mass spectrometry, while mass fractions are quantified using a splitter and flame ionisation detector. For relatively simple mixtures, this information is often sufficient to apply blending rules to calculate mixture properties. However, in more complex cases, particularly due to co-elution of compounds, notably the large number of isomeric paraffins within the carbon number ranges of interest, accurate molecular

identification becomes impractical. To address this, the prescreening technique developed at Washington State University employs a Monte Carlo strategy, in which sets of potential mixtures are generated using the compositional data from GCxGC. Since the mass fraction of each isomer class can be determined (for example, 1% iso-C10), the properties of these candidate mixtures can be calculated and averaged to provide a prescreened result. Blending rule's applicability to the present work, however, is less straightforward. By design, blending rules involve assumptions that introduce error, as they represent a substitution for direct experimental measurement. As a result, any model trained on properties derived from blending rules will inevitably accumulate error from multiple sources: experimental uncertainty, the inherent approximations of the blending rule, and the modelling approach itself. For this reason, in the present study, calculated properties are used only where absolutely necessary due to the scarcity of literature data, and only after a critical evaluation of the suitability of the blending rule in question.[15]

Experimental data for the properties under study is available in the literature[115, 116] and is more favourable than calculated data as the inaccuracies in the calculation methods would be inherited into the models produced. The main limitation in this case is the lack of compositional data, which serves as the predictors for the models developed in this study. Some fuels from the National Renewable Jet Fuel Program had already been sent to Trinity College Dublin, which enabled the use of real conventional fuels for verification. In addition, further reference fuels with published experimental data were obtained from Washington State University[5], as well as from the Alternative Jet Fuel Test Database.[116] Where these sources were accessible and appropriate, this approach provided the most effective verification data, as it required only compositional analysis rather than complete experimental determination. Each property considered in this study was verified using this method where possible.

Surrogate mixtures provide another route to obtaining validation data, though they are accompanied by distinct challenges.[117] Conventional jet fuels consist of thousands of components, making it necessary to simplify the chemical matrix in order to construct a representative surrogate. This reduction, while practical, inevitably compromises the chemical diversity of the mixture. Furthermore, the compounds required to produce surrogates are

costly, particularly isomeric paraffins, which are difficult to separate in pure form owing to their similar physical properties and the same co-elution issues discussed earlier. The expense and complexity of preparing accurate surrogates therefore limit their widespread use. Nevertheless, when carefully designed, surrogate mixtures can provide a controlled means of linking atom-type information with measured properties and can therefore contribute useful validation data in specific cases.

Finally, the most extensive and reliable method of gathering verification data is by acquiring a sample with enough volume to perform experiments. To this end, Monument Chemical in Texas, USA was contacted, they provided three samples of Hydrotreated Esters and Fatty Acids (HEFA), while also a sample of Jet A-1 was acquired from Weston Airport in Leixlip, Dublin.

3. Methodology

The basis of the methodology is the assumption that the surface tension, density, viscosity, DCN, distillation curve and flash point of fuels are principally separately dependent on the molecular structure of a fuel, which can be measured by NMR, and that this relationship can be learned by mathematical analysis of an appropriate database of molecular structures for each property.

3.1 Materials

As this study deals with a novel methodology for the prediction of jet aviation turbine fuel properties, a number of samples required and were necessary for testing and construction of these models. The compositional analysis methodology used in this work is currently unique to Trinity College Dublin in literature meaning that any fuel used for the evaluation of the models had to be attained and physically measured in the laboratory. In Table 2, a list of the materials used, and acquired for this study are listed along with their sources. During the project, I reached out to multiple sources to increase the laboratory stock of fuels in Trinity College Dublin.

Table 2. Material list of fuels used in the study.

Fuels	**Source**
FACE A Gasoline	TCD Laboratory
FACE C Gasoline	TCD Laboratory
FACE F Gasoline	TCD Laboratory
FACE I Gasoline	TCD Laboratory
Jet A-1 (Weston)	TCD Laboratory
Jet A POSF 10324	TCD Laboratory
JP-8 POSF 10264	TCD Laboratory
JP-5 POSF 10289	TCD Laboratory
Shell - SPK POSF 5729	TCD Laboratory
HRJ - Camelina POSF 7720	TCD Laboratory
Gevo - ATJ POSF 10151	TCD Laboratory
Sasol - IPK POSF 7629	TCD Laboratory
Gevo POSF 13718	Washington State University
S-8 POSF 5018	Washington State University
Amyris Farnesane POSF 11832	Washington State University
World Energy HEFA	Washington State University
M101	Monument Chemical
M102	Monument Chemical

M103	Monument Chemical

The NMR methodology was applied to the real aviation and gasoline fuels available in the laboratory: the FACE gasolines are part of the CRC Report AVFL-24, Jet A-1 (Weston), Jet-A POSF 10325, JP-8 POSF 10264 and JP-5 POSF 10289 are all fossil derived aviation fuels, POSF 5729 Shell Fischer-Tropsch Synthetic Paraffinic Kerosene derived from natural gas,[118] POSF 7720 Hydrotreated Esters and Fatty Acid (HEFA) derived from Camelina oil,[118] POSF 10151 Gevo Alcohol-to-Jet (ATJ) derived from iso-butanol, and POSF 7629 Sasol-IPK derived from coal.[118] POSF 13718 Gevo, POSF 5018 S-8, POSF 11932 Amyris Farnesane and World Energy HEFA are synthetic fuel samples received from Washington State University which were used in the study by Yang et al.[5] M101, M102, M103 are renewable diesel HEFA samples received from Monument Chemical, Texas. Figure 4 shows the assigned ^{1}H ^{13}C HSQC spectrum for a selection of these fuels. The atom types for each fuels are presented in Table 5.

3.2 Nuclear Magnetic Resonance Compositional Analysis

Nuclear magnetic resonance allows for the quantification of atom type mole fractions within a hydrocarbon mixture by exploiting the magnetic properties of atomic nuclei.[80] The atom types utilized in this study are presented in Figure 3. The method of Ure et al.[80] has been adapted to additionally enable the identification and quantification of aromatic-ring junction carbon atom types.[51] As introduced by Ure et al.[80] and further developed in this work, ^{1}H ^{13}C HSQC-NMR identifies the unique nuclear and electronic environments associated with each atomic type within hydrocarbon molecules that constitute typical liquid transportation fuels. This identification inherently includes the geometrical configurations of each atomic type and molecular group ensemble.

3.2.1 Atom Types by ^{1}H ^{13}C HSQC NMR Spectroscopy

To validate the full atom type compositional analysis procedure, a set of six real fuels with unknown atom type compositions were analysed. The ^{1}H ^{13}C HSQC NMR spectra for these real fuels are presented in Figure 4, the untruncated ^{1}H spectra as well as the data of each ^{1}H spectrum are provided in Figures 5 – 10. HSQC, quantitative ^{1}H, and qualitative ^{13}C NMR spectra were collected using a Bruker Avance III 400 NMR spectrometer with the parameters suggested by Ure et al.[80] HSQC spectroscopy produces a two-dimensional NMR spectrum

wherein the ^{1}H and ^{13}C spectrum are on the x and y axes respectively. The intersection of data from these two axes, termed a cross-peak, represents a specific C-H bond. Carbon atoms with an odd number of bonded hydrogen atoms (i.e., paraffinic CH, CH_3) have an opposite phase to carbon atoms with an even number of bonded hydrogen atoms (i.e., paraffinic CH_2). Hence, a phase editing procedure is applied to the data to allow for the quantification and display of both sets of data on the one spectrum. As the magnitude of the ^{1}H chemical shift is due to the local bonding environment of the nuclei, it is possible to assign regions of the x-axis (^{1}H spectrum) to each atom type. The ^{1}H ^{13}C HSQC spectrum was used to assign spectral regions for each atom type for each sample. These regions were then integrated in the quantitative ^{1}H spectrum to determine the atom type composition in mol% hydrogen, which was then used to calculate the inferred mol% carbon composition using the extended method of Watson-Murphy et al.[51] As mol % carbon is not a familiar term in the field of aviation, a conversion can be made to convert to mass %. The mol % carbon and mol % hydrogen is calculated from the various atom types in a sample, then the molecular weight associated with each atom type is found, and the product of the mol % gives the mass % for each element (hydrogen and carbon). The sum of these then represents the mass % for the atom type understudy.

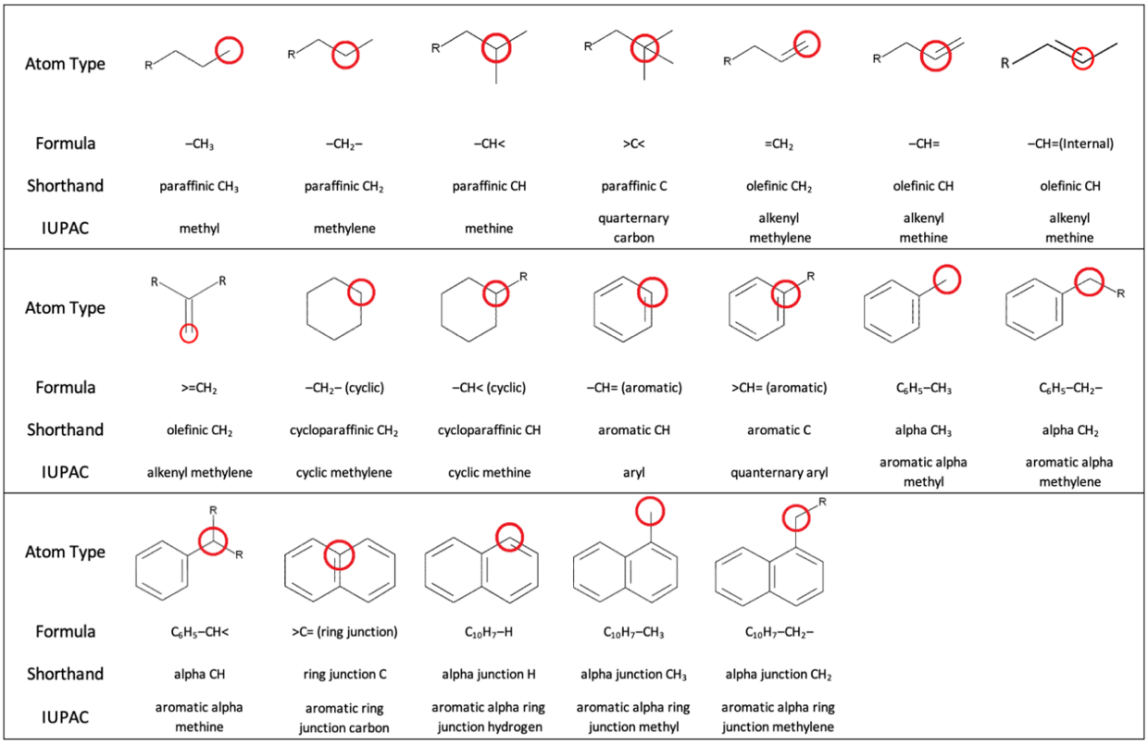
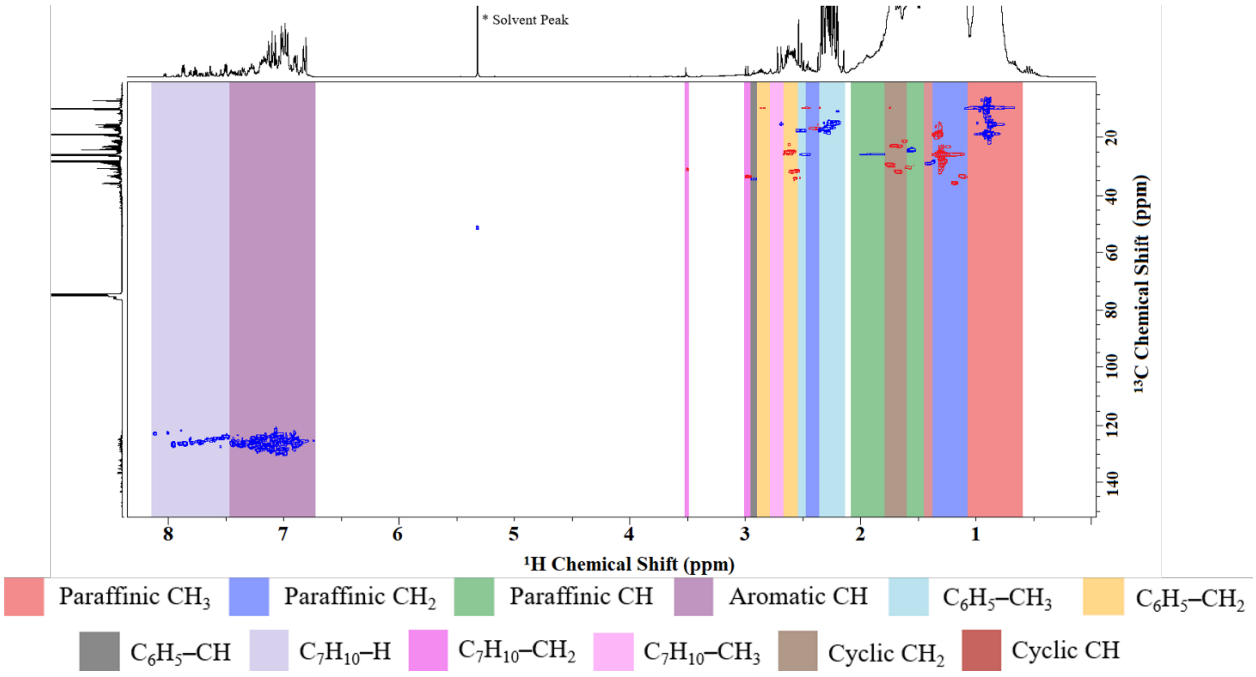
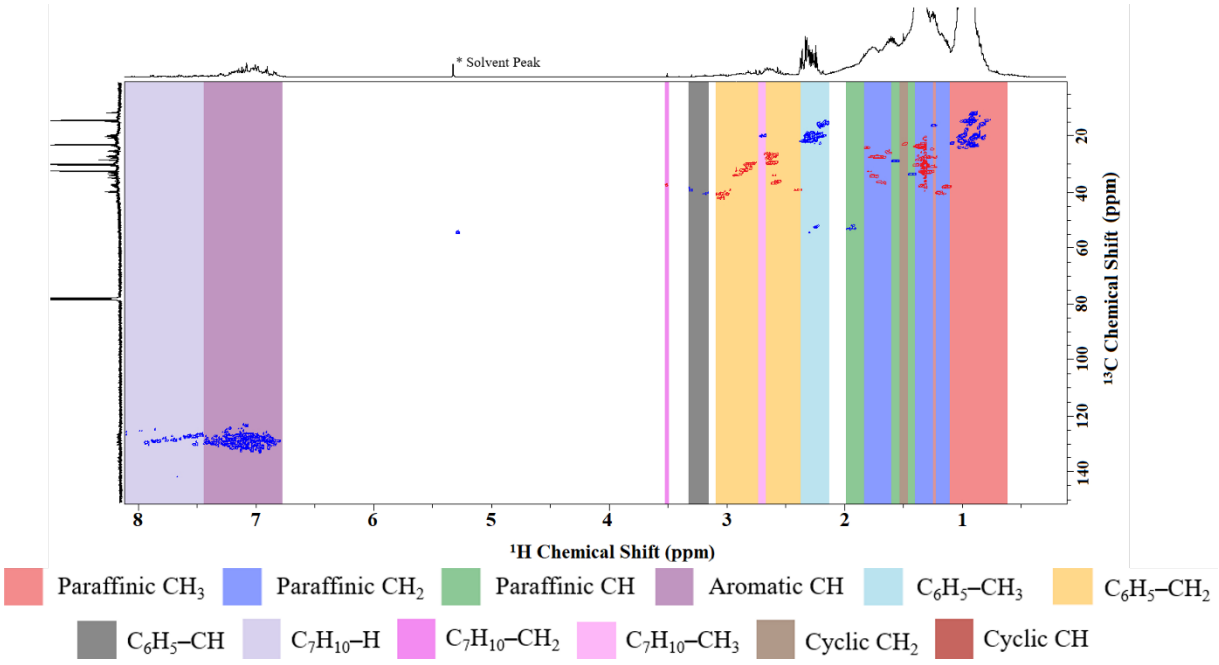


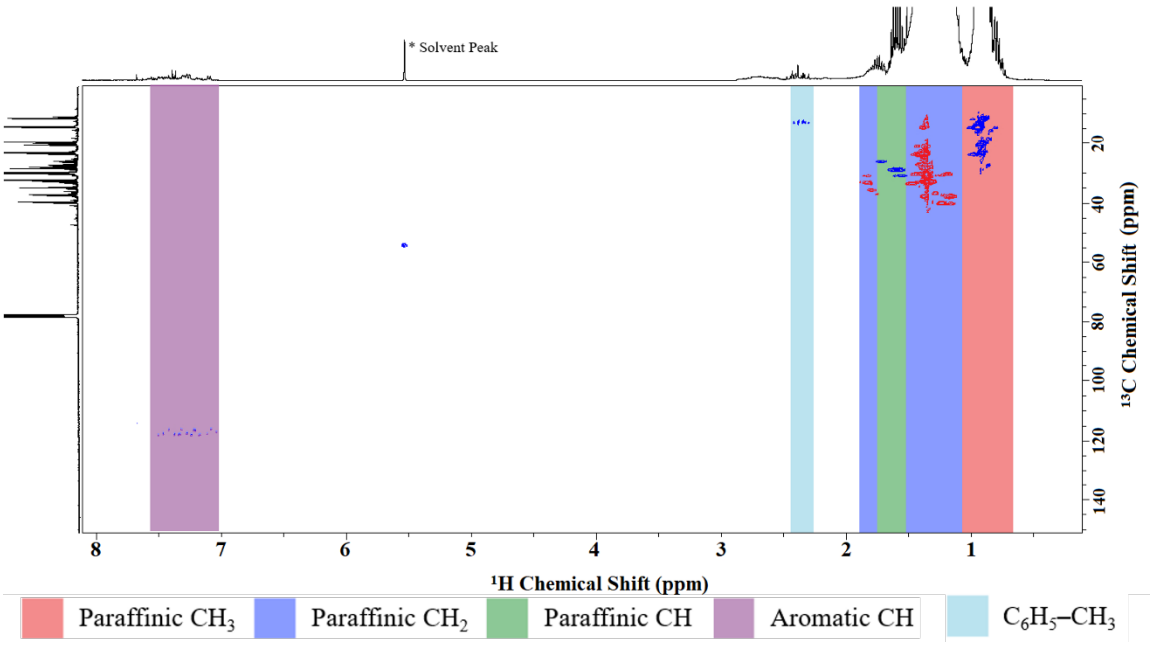
Figure 3. Atom types quantifiable through methodology proposed by Ure et al., and Watson-Murphy et al.[51, 80]



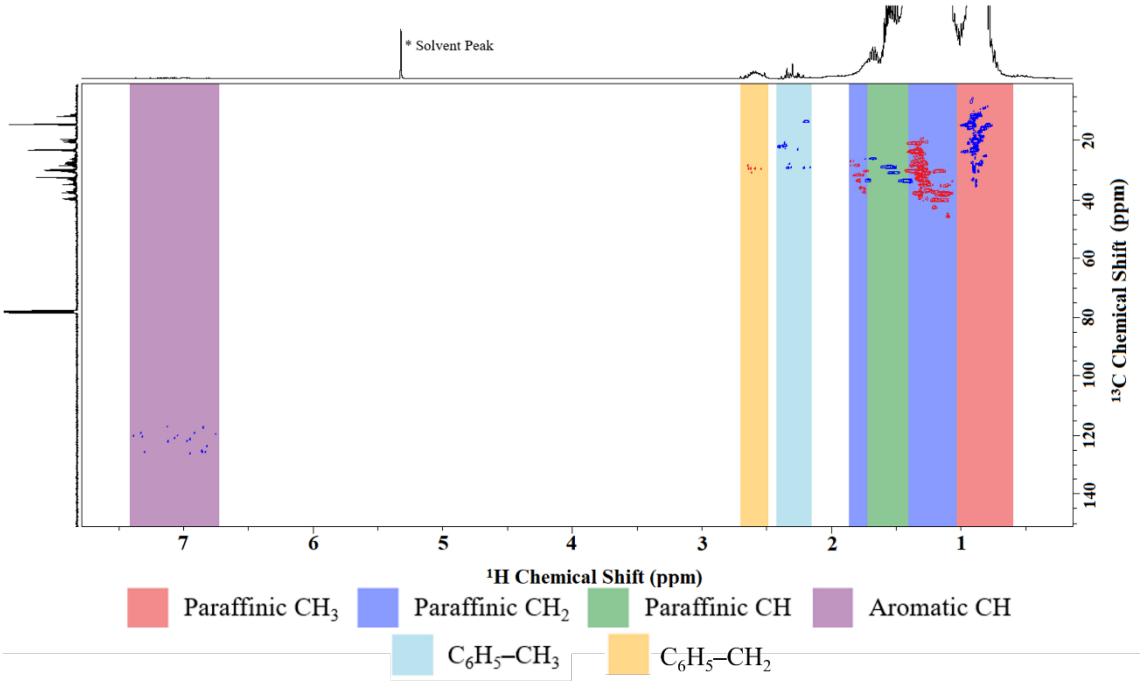
(A) Jet-A POSF 10325



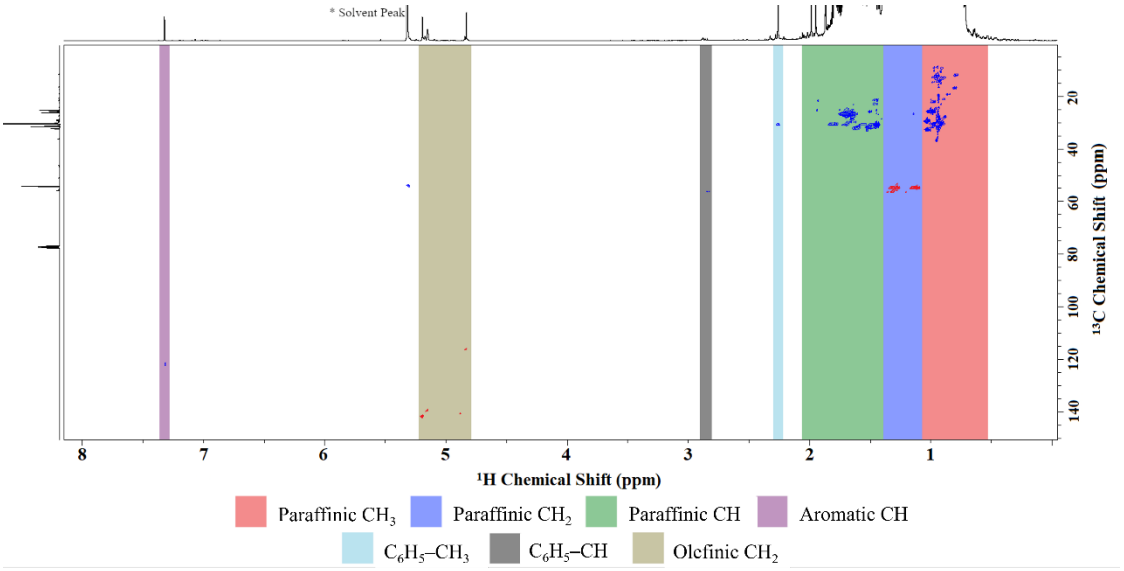
(B) JP-5 POSF 10289



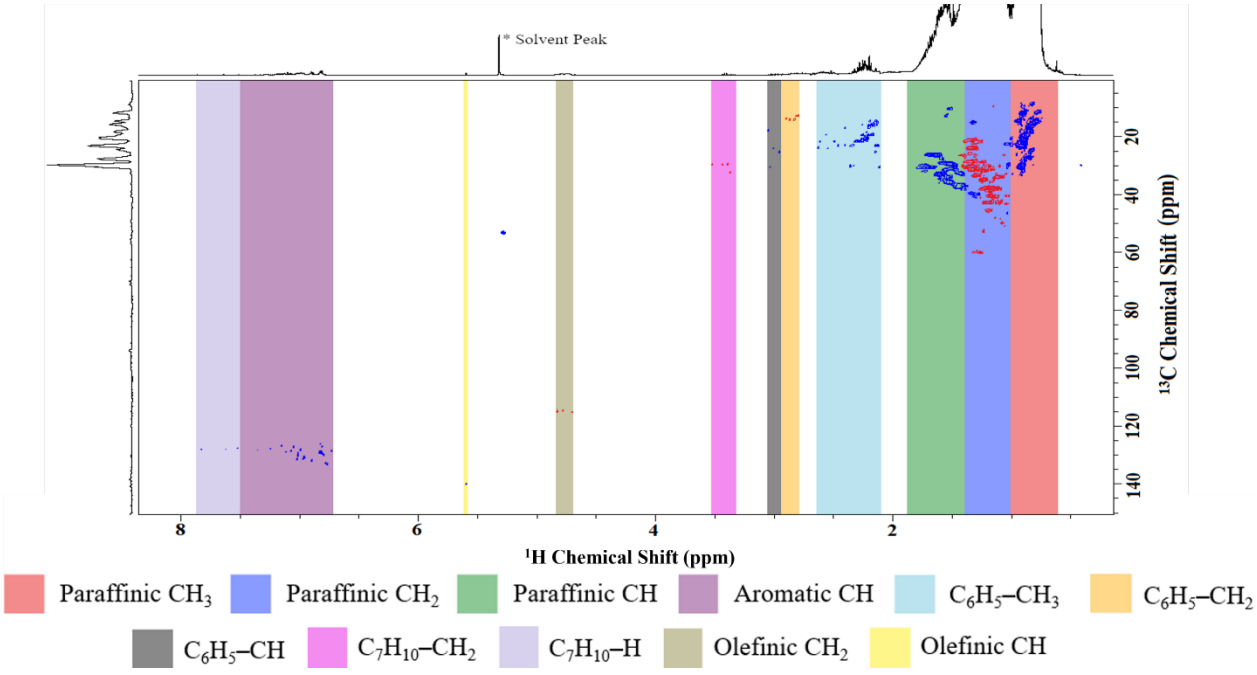
(C) Shell-SPK POSF 5729



(D) HRJ-Camelina POSF 7720



(E) Gevo-ATJ POSF 10151



(F) Sasol-IPK POSF 7629

Figure 4. ^{1}H ^{13}C HSQC-NMR Spectra for 2 fossil fuels (Jet A, JP5) and 4 synthetic aviation fuels (Shell-SPK, HRJ-Camelina, Gevo-ATJ, Sasol-IPK) which are used to assign atom types. These fuels are used to validate the models presented.

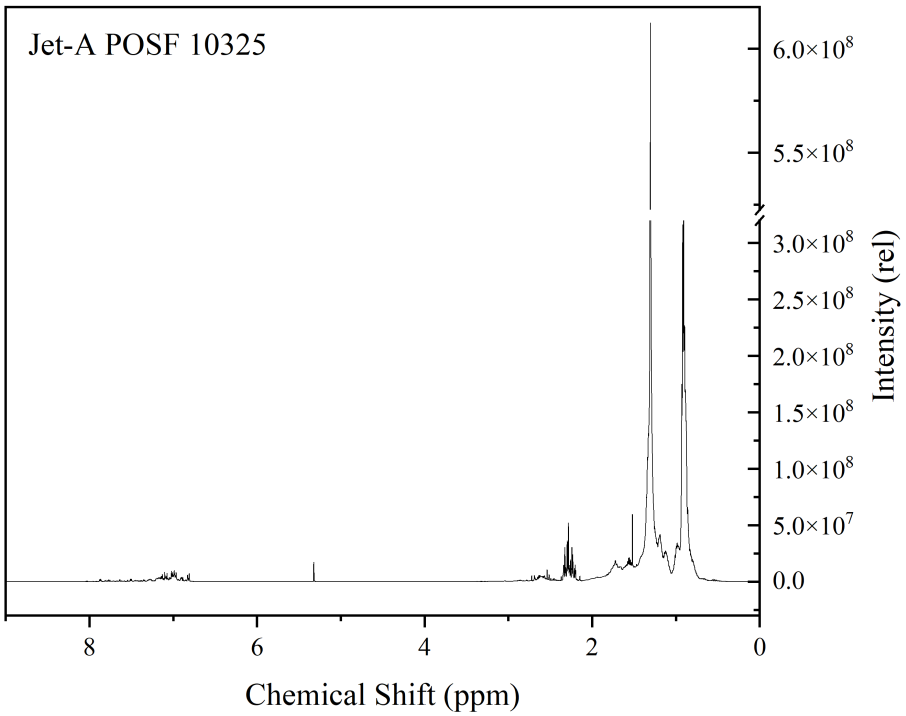


Figure 5. ^{1}H NMR spectrum of Jet-A POSF 10325 with an axis break to allow for the intensities of the aromatic region (+7 ppm) to be visible.

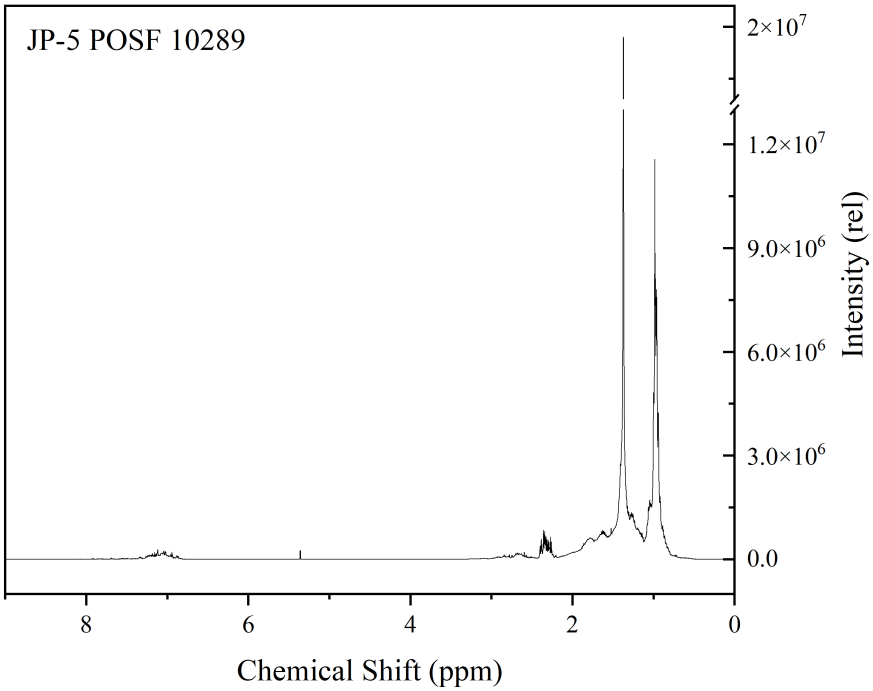


Figure 6. ^{1}H NMR spectrum of JP-5 POSF 10289 with an axis break to allow for the intensities of the aromatic region (+7 ppm) to be visible.

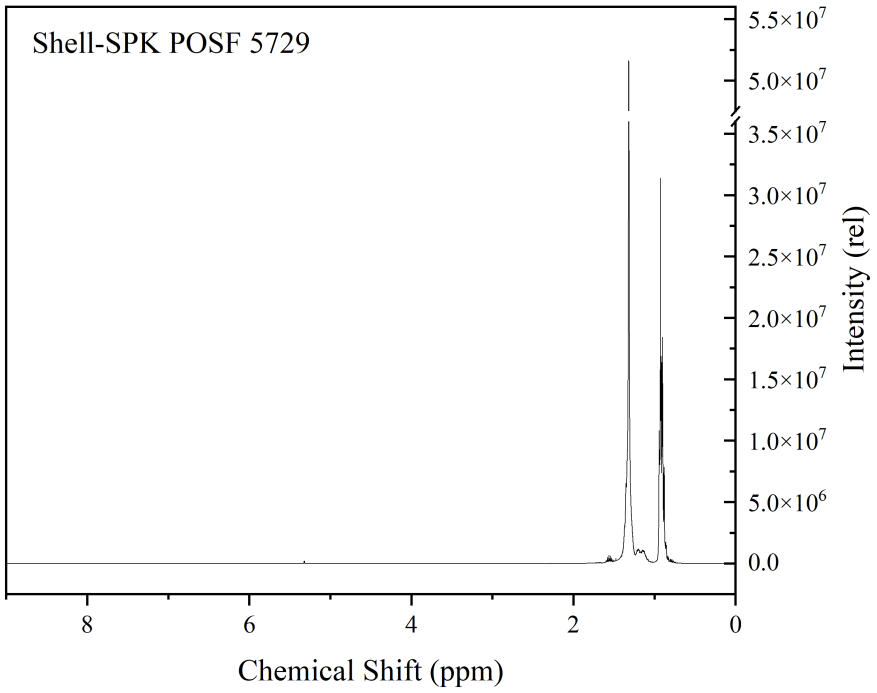


Figure 7. [1]H NMR spectrum of Shell-SPK POSF 5729 with an axis break to allow for the intensities of the lower paraffinic peaks (<2 ppm) to be visible.

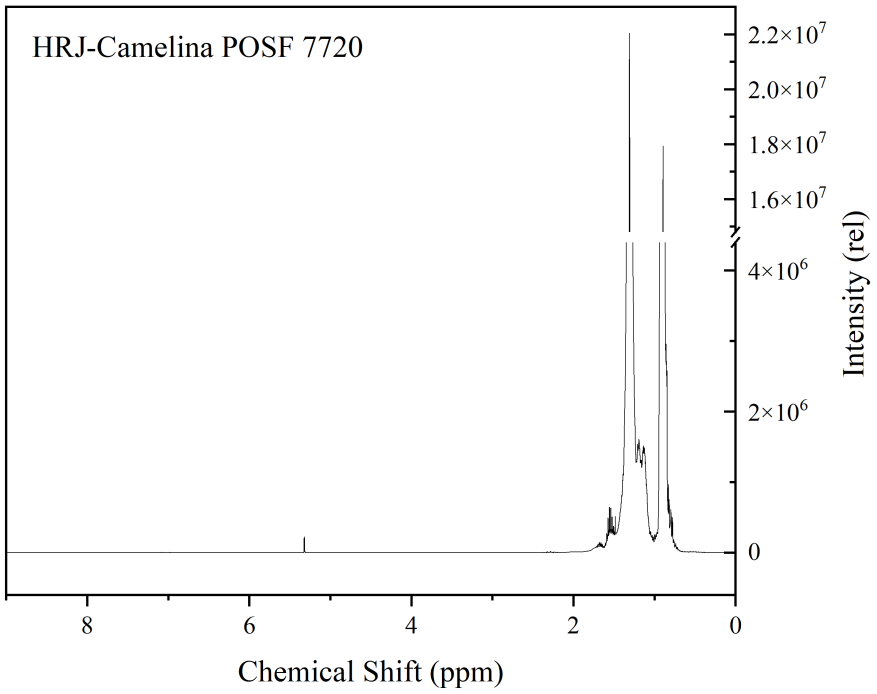


Figure 8. [1]H NMR spectrum of HRJ-Camelina POSF 7720 with an axis break to allow for the shapes of the full paraffinic peaks to be visible.

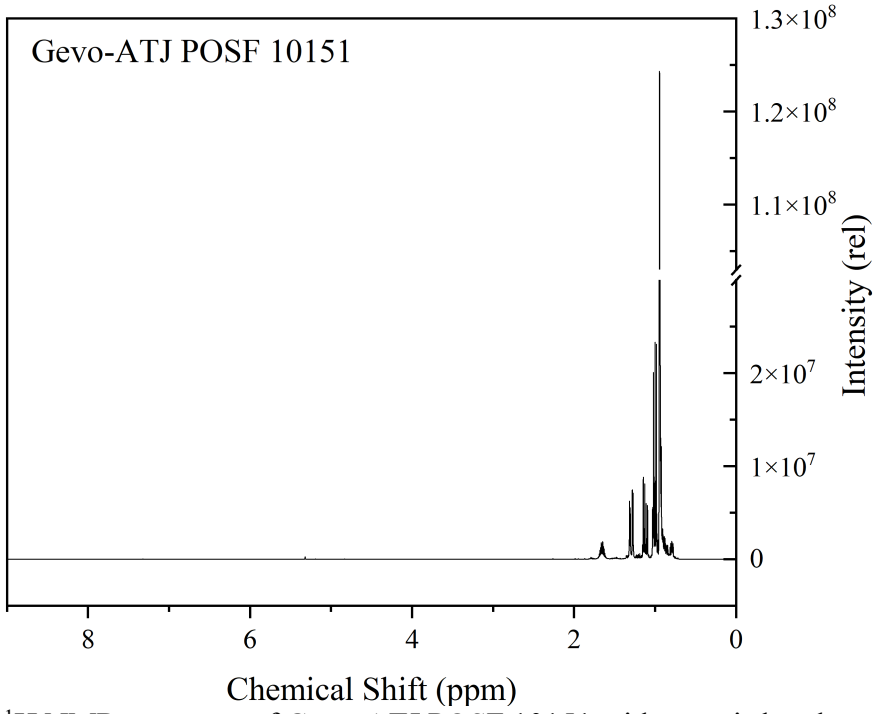


Figure 9. [1]H NMR spectrum of Gevo-ATJ POSF 10151 with an axis break to allow for the intensities of the <2 ppm to be visible.

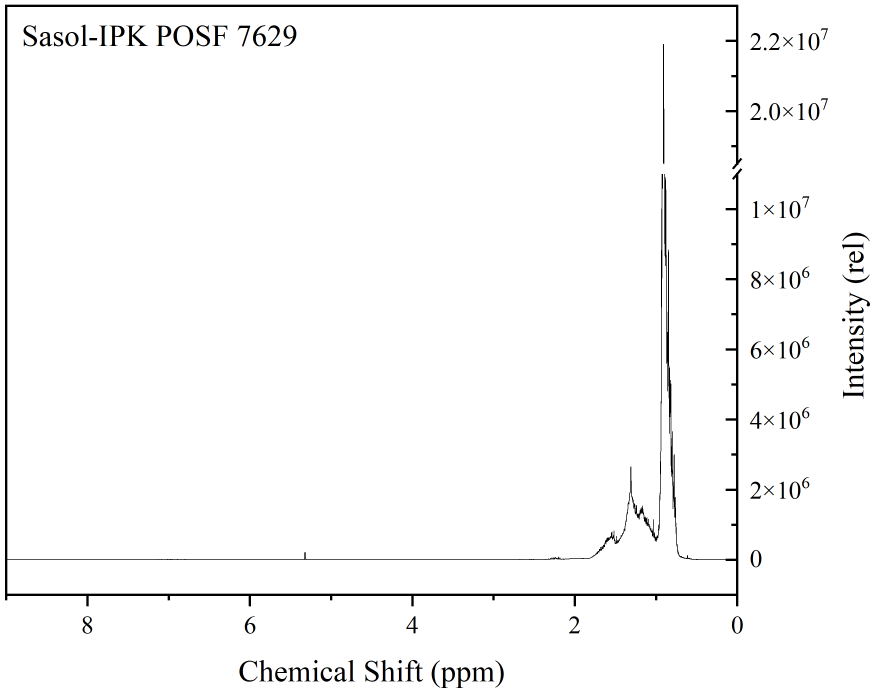


Figure 10. [1]H NMR spectrum of Sasol-IPK POSF 7629 with an axis break to allow for the intensities of the aromatic region (7+ppm) to be visible.

3.2.2 NMR Spectrometer Parameters

The NMR spectrometer requires parameter tuning for each adjustable parameter to ensure that a clear signal which is narrow and distinguishable from noise is attainable. The parameters listed in Table 3 are adapted from Ure et al.[80] The parameters are adjusted for the masses used for the solvent and solute. Originally, 0.05 g of solute was used, though this caused the signal to be too small for atom types in lower concentrations, such as aromatics or olefins, it was found that 0.2 g of solute allowed for maximum signal acquisition without saturation. The preparation of solutions and the choice of solvents are crucial factors in the NMR analysis. Solvents must be carefully selected to avoid spectral interference, and the sample-to-solvent ratio must be optimized to ensure clear and accurate atom type quantification. In this study, a 1:4 ratio of sample to deuterated solvent (0.2 g of sample to 0.8 g of deuterated dichloromethane) was used. This ratio was found to balance resolution and intensity, avoiding issues such as decreased peak intensity from excessive solvent or peak overlap from too much sample. The choice of deuterated dichloromethane was specifically made to minimize interference with the atom type distribution observed in jet fuels. An additional parameter was also included, the mass of an internal reference standard which was used to quantify the hydrogen content of the sample.[91] The hydrogen content is used to calculate the H:C ratio which then allows for the quantification of quaternary carbon and aromatic carbon atom types which cannot be inferred from a ^{1}H spectrum. The H:C ratio was verified measured using an Exeter Analytical CE 440 Elemental Analyzer, and is used to deduce the quantity of quaternary paraffinic carbon present, which do not appear on the spectra due to having no bonded hydrogens. Although the methodology employed accounts for all atom types described, it does not consider trace impurities that may be present in the fuel.

These impurities are described in ASTM D4054[1] to be in a ppm unit basis. However, only organic impurities (water, carbonyls, alcohols, esters, and phenols) which are directly bonded to carbon or hydrogen atoms would be detectable in the NMR spectra used in this analysis. These are not identifiable through the methodology employed and would be included in the integration for atom types if they are within the range.

Table 3. Nuclear magnetic resonance spectroscopy parameters for ^{1}H and ^{13}C acquisition.

Parameter	**^{1}H NMR**	**^{13}C NMR**
Solvent	CD_2Cl_2	$CDCl_3$
Mass solvent (g)	0.98	0.98
Mass solute (g)	0.02	0.02
Mass standard (g)	0.0015	-
Concentration $Cr(acac)_3$ (M)	0	0.05
Field strength (MHz)	400	101
Flip angle (°)	90	90
Acquisition time (s)	4	1.4
Delay time (s)	60	10
Time domain points ($\times 10^3$)	64	64
Sweep width (Hz)	8012	24038
Dwell time (μs)	62.4	20.8
Number of scans	16	3200
Line broadening (Hz)	0.3	1

Field strength is determined by the spectrometer being used and is tuned for the strength of the magnetic field for a proton nuclei. Flip angle, acquisition time, delay time, time domain points, sweep width and dwell time are carefully selected by the spectrometer specialist Dr. John O'Brien to ensure a clear and clean spectrum. The number of scans is an adjustable parameter which can decrease the signal to noise ratio in a square root proportion.

3.3 Predictive Modelling Development

There are three technical conditions that must be met when employing machine learning for this purpose: (1) the atom types which most significantly contribute to each property must be identifiable through the ^{1}H ^{13}C HSQC NMR spectrum of the real fuel requiring analysis, (2) there must be property data for pure components and/or hydrocarbon mixtures of sufficient quality and quantity to train the numerical model, (3) the resulting operating equation must be concise and only contain statistically significant parameters. This approach enhances our understanding of the importance of, and interactions between, atom types, the molecules they comprise, and the "physical" properties they define. Therefore, a simpler model (i.e. MLR) will be favoured over a more complex model (i.e. nonlinear regression, neural networks).

3.3.1 Data Sources and Preprocessing

To train the models, a database of experimental measurements was collected from the literature for both pure components and their mixtures for each property. The pure components atom types mole fractions were calculated based on their structure, and the mixtures were linear by mole calculations of the pure constituents. Where possible, the pure components were chosen to replicate the composition conventional jet aviation fuel as shown in Figure 11 from the National Alternative Jet Fuel Database.[119] This means that the carbon numbers range from 6 – 18 in chain length (i.e. n-hexane – n-octadecane). Additionally, the inclusion of cycloparaffins, iso-paraffins and aromatic components, as they each make up a large portion of jet fuel composition. For surface tension and density, 2,501 data points were compiled, consisting of normal paraffins, iso-paraffins, cycloparaffins, aromatics, and binary mixtures of these classes. Statistical descriptors including minimum, maximum, mean, and standard deviation for surface tension, density, temperature, and each atom type are provided in Table 4. Surface tension and density can be assumed, based on experimental measurements, to vary linearly with temperature over restricted ranges. Accordingly, only data at least 50 K below the boiling point of each component were included in the dataset. Discrete datasets were prepared for each property: the density dataset contained 1,241 points and the surface tension dataset contained 1,260 points. Each dataset was divided into a single random 80:20 split for training and testing. Figure 12 and 13 show this linearity within the temperature constraints for n-heptane surface tension and conventional aviation fuel density.

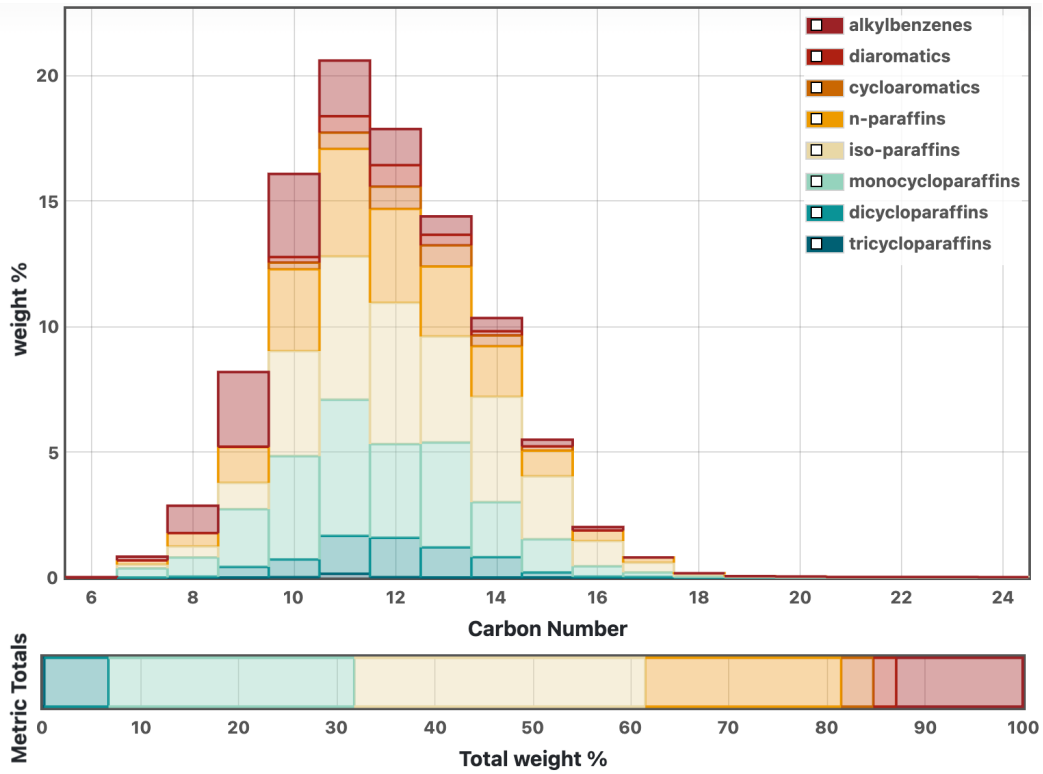


Figure 11. GCxGC-MS weight % against carbon number distribution for Jet A POSF 10325. This graph adapted from data acquired from the National Alternative Jet Fuel Database.[119]

Table 4. Training and testing statistical data for temperature dependent surface tension and density datasets of 1,260 and 1,241 values respectively.

Surface Tension				
Descriptor	**Min**	**Mean**	**Max**	**Standard Deviation**
Temperature	263.15	311.36	363.15	21.35
pCH3	0.00	0.21	0.71	0.20
pCH2	0.00	0.22	0.83	0.28
pCH	0.00	0.03	0.38	0.07
pC	0.00	0.02	0.14	0.04
cyCH2	0.00	0.17	1.00	0.31
cyCH	0.00	0.01	0.17	0.03
alCH3	0.00	0.04	0.33	0.09
alCH2	0.00	0.01	0.13	0.03
alCH	0.00	0.00	0.11	0.02
ArCH	0.00	0.25	1.00	0.34
ArC	0.00	0.05	0.33	0.09
rjC	0.00	0.00	0.20	0.02
Surface Tension	8.08	21.98	40.78	4.67

Density				
Descriptor	**Min**	**Mean**	**Max**	**Standard Deviation**
Temperature	263.55	310.37	363.15	20.53
pCH3	0.00	0.22	0.75	0.19
pCH2	0.00	0.33	0.88	0.32
pCH	0.00	0.03	0.25	0.06
pC	0.00	0.02	0.13	0.03
cyCH2	0.00	0.21	1.00	0.34
cyCH	0.00	0.01	0.14	0.03
alCH3	0.00	0.01	0.25	0.05
alCH2	0.00	0.00	0.13	0.01
alCH	0.00	0.00	0.11	0.01
ArCH	0.00	0.15	1.00	0.30
ArC	0.00	0.02	0.25	0.05
rjC	0.00	0.00	0.20	0.01
Density	527.26	735.50	993.41	66.83

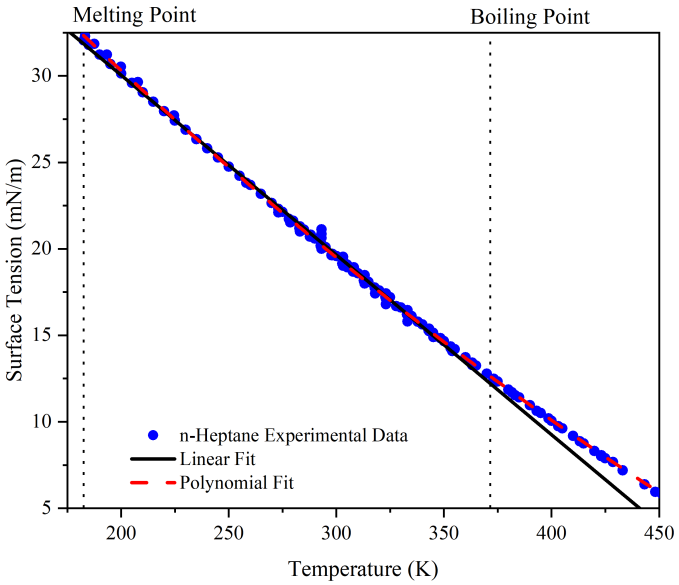


Figure 12. Surface tension temperature-dependence of n-heptane over a large range of temperature. This shows that until we near the phase boundary the surface tension temperature dependence can be assumed to be linear. n-Heptane was chosen as an extremal example here as most components in jet fuel are higher in carbon number than C7, which means they have higher boiling points and therefore less deviation from linearity at the highest relevant temperatures. The data presented in the figure is from NIST Thermodynamic Data for Pure Compounds.[114]

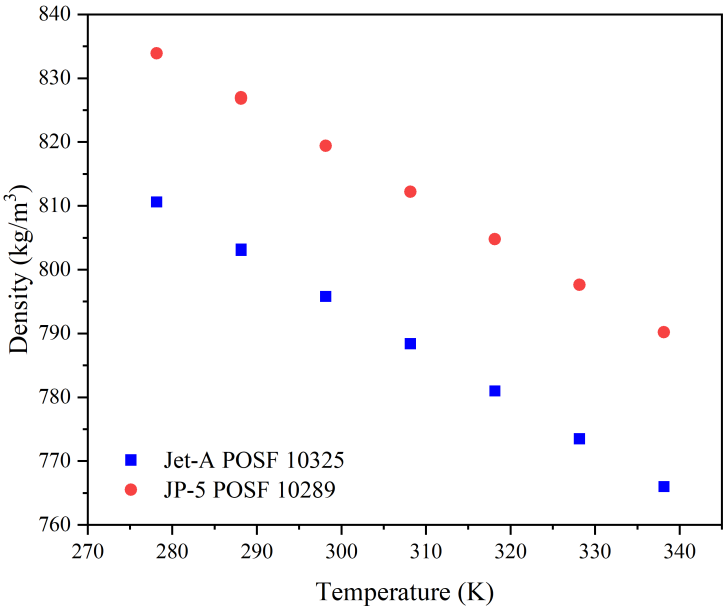


Figure 13. Experimental density temperature-dependence of Jet-A POSF 10325 and JP-5 POSF 10289. This plot shows that the temperature dependences of fuel density for these real jet fuels is essentially linear over the temperature range considered. The data present in the figure is from Edwards et al.[115]

The flash point dataset is composed of data from the DIPPR (Design Institute for Physical Properties) database from American Institute of Chemical Engineers (AIChE)[120] for pure component and data from literature for mixture measurements.[121] This dataset has 184 flash points, this is notably smaller than both density and surface tension. The reason for this is due to the non-temperature dependent nature of flash point. Meaning, that for a pure component or mixture there is only one flash point available, but for surface tension there is a different surface tension at each temperature change.

Distillation curve is a property of mixtures, as for a pure component there is one temperature at which it distils i.e. boiling point. This means that the dataset is composed only of mixtures. These mixtures are taken from Yang et al.[5] where a comprehensive database of fuel distillation curves are listed. The samples used in the study were requested so the atom types could be quantified allowing the data to be used in the model training. There are mixtures of the fuels used in the study, which was calculated using a linear by mole equation. The dataset is composed of 1,074 measurements at different temperatures of the fuel.

For the training of the DCN model, experimental values were gathered from studies employing an ignition quality tester. In total 405 derive cetane numbers are in the dataset, these data points are wholly mixtures as pure components were found to be destructive to the model, this is due to the use of different measurement techniques and the intermolecular interactions which are crucial to the derived cetane number.[4, 40, 44, 45, 47, 81, 85, 108, 117, 122, 123]

The viscosity dataset is composed of temperature dependent data from the DIPPR Database.[120] The experimental results for pure components were measuring dynamic viscosity, though the model is attempting to predict kinematic viscosity. The conversion is simple and requires the density at the measurement temperature, then the ratio of the dynamic viscosity and density gives the kinematic viscosity. The density for the pure components considered in this database were also gathered from DIPPR. It was found that just the pure component database was not sufficient for the construction of the model, to this end, measurements for kinematic viscosity were performed in Trinity College Dublin on jet fuel and pure component mixtures to populate the dataset. In total, the kinematic viscosity dataset has 1,373 data points.

Through training the model on the pure components an unsatisfactory model was produced in terms of model fidelity and performance. Thus, experimental measurements of real and surrogate jet fuel was added to the database, then using mixing rules to calculate surrogate mixtures and random mixtures of pure components and jet fuel. The following mixing rule was used, through an investigation by Boehm et al.[114] The equation was a linear by mole calculation

$$\nu_{fuel} = \langle \vec{X} | \vec{\nu} \rangle$$

where X is the mole fraction of each component, and ν is the kinematic viscosity of each component.

3.3.2 Machine Learning Techniques

MLR models were produced using operating equations of the form,

$$y = \alpha_0 + \sum_{n=1}^{N} \alpha_n x_n$$

where α_0 is the intercept, α_n are the coefficients learned for each independent variable, x_n are the atom type mole fractions or temperature, and the dependent variable, y. Individual models for temperature dependent liquid density, temperature dependent surface tension were then trained using the Linear Regression module from the SciKit Learn Python package.[124]

Independent variables were evaluated using the p-value of each coefficient, and backward selection was applied to remove atom types with p-values above 0.05. Models were retrained iteratively until the coefficient of determination decreased below 0.95. Ridge regression was employed to improve model stability, penalising excessive coefficient magnitudes and reducing the risk of overfitting. The resulting models, presented in Figures 4 and 5 and Table 1, achieved the highest determination coefficients while retaining only statistically significant atom types.

Where MLR did not achieve sufficient accuracy (R^2 above 0.95) and the dataset could not be expanded, more complex techniques were introduced. Polynomial regression was employed next, incorporating squared terms and cross-product terms between atom types to capture non-linear interactions. The general operating equation was

$$y = \alpha_0 + \sum_{i=1}^{atom\ types} a_i x_i + \sum_{i=1}^{atom\ types} b_i x_i^2 + \sum_{i=1}^{atom\ types} \sum_{j=1}^{atom\ types} c_{i,j} x_i x_j$$

where α_0 is the intercept, a_i are the coefficients learned for each independent variable, b_i are the coefficients learned for the square of each independent variable, c_i are the coefficients learned for the product of a variable with another variable, x_n are the atom type mole fractions or temperature, and the dependent variable is y. Polynomial regression extends multilinear regression by including non-linear terms that allow the model to capture interactions between variables and curvature in the data that linear models cannot represent. In binomial regression, each independent variable contributes both a linear and a squared term, and pairwise combinations of variables are included to describe potential interaction effects. These additions increase the model's flexibility and its ability to reproduce more complex relationships between atom types and physical properties. In this study, i has the range [1, 19] and j has range [1, 18] from the atom types available in Figure 3, this means that without statistical reduction the full binomial model would have 210 coefficients.

Although higher-order expansions are theoretically possible, the number of terms increases exponentially, which leads to overfitting and numerical instability, particularly when the dataset is limited. For this reason, only binomial models were investigated in this study. The significance of each term was assessed using p-value analysis, and non-significant terms were progressively removed through systematic reduction until the model retained only statistically meaningful predictors. Individual models for flash point, viscosity, and distillation curve were then developed using the Polynomial Regression module from the SciKit Learn Python package.[124]

The next step in model architecture complexity is the random forest regressor, an ensemble of decision trees that learn regression rules from the dataset. A decision tree can be understood as a structure composed of nodes with branching conditions; at each branch a split is made depending on whether the input satisfies a specified criterion. The algorithm evaluates all available features and possible split thresholds, and the split chosen is the one that minimises the mean square error

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2.$$

For each node, the dataset is partitioned into left and right branches, and the process repeats until the tree is complete. In a random forest, only a subset of features is assigned to each tree, which ensures diversity among the trees and prevents over-reliance on a single feature. The ensemble prediction is then obtained by averaging across the individual decision trees. Figure 14 shows a decision tree in a random forest, and how it can vary based on its initial architecture rules.

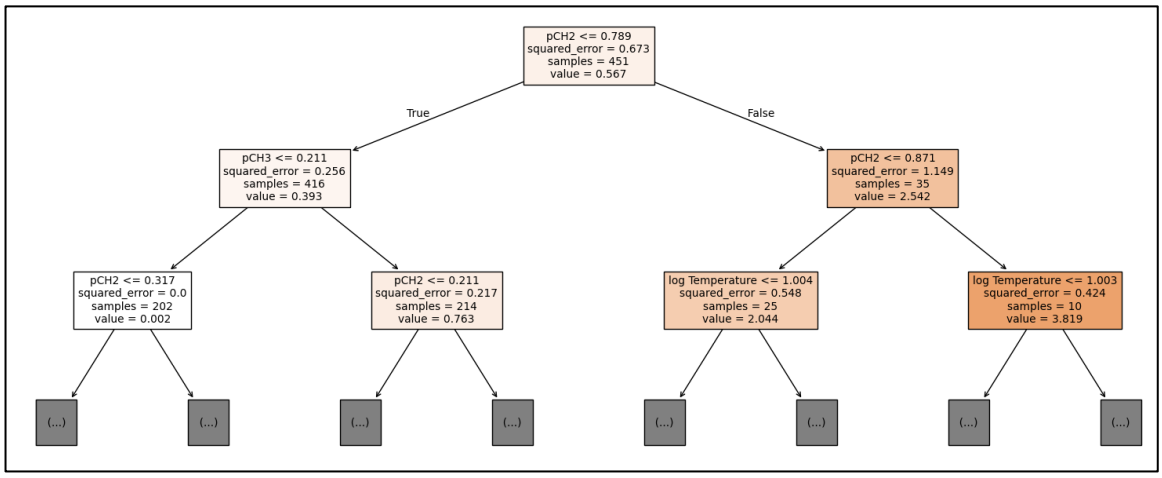


Figure 14. Example of a decision tree as part of a random forest, showing the first three layers of decisions for a model on kinematic viscosity.

In practice, a simplified one-layer decision tree can be thought of as a sorting mechanism, whereas deeper trees sort across multiple parameters to reach a final prediction. The predictive range of such models is inherently limited by the data on which they are trained, which constrains their generalisability. For this reason, reliable performance requires that the training dataset adequately spans the relevant composition and property space. In this study, random forest regressors were applied to viscosity and distillation curve data using the Random Forest Regressor module from the SciKit Learn Python package.[124]

Finally, the most complex model architecture investigated in this study are ANNs. Neural networks are considered black box models, as their internal decision processes cannot be easily or directly interpreted. Although the relative importance of individual features can be evaluated, the precise mechanism by which these features interact to influence the prediction is not accessible. For this reason, ANNs sacrifice interpretability in exchange for predictive

power and flexibility. They were therefore employed only in cases where simpler models such as multilinear regression, polynomial regression, or random forest regressors could not provide sufficient accuracy due to the limited information available in the training dataset. The performance of an ANN depends strongly on the network architecture. Key parameters include the number of hidden layers, the number of neurons per layer, the activation functions, and the choice of optimisation algorithm, an example simplified ANN is shown in Figure 15. The balance between network depth and dataset size is particularly important, as overly complex architectures can lead to overfitting, while overly simple architectures may fail to capture the necessary relationships. Considerable time is often required to refine the network architecture to achieve stability and accuracy. To streamline this process, the H2O AutoML Python package[125] was used, which automates hyperparameter tuning and architecture selection. A grid search procedure was implemented within a prescribed time frame, during which the algorithm explored multiple network configurations and selected the model with the highest fidelity to unseen test data. This ensured that model complexity was increased only as needed, while avoiding overfitting through systematic cross-validation.

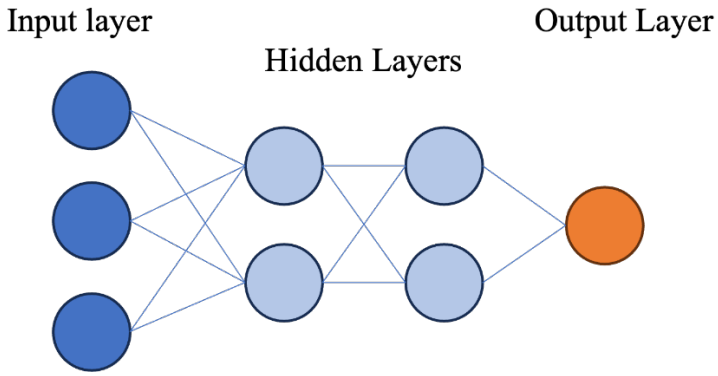


Figure 15. Example of an artificial neural network model architecture.

In this study, ANN models were trained for viscosity prediction. Their use was restricted to this property as it presented the greatest challenge for the simpler regression-based approaches, and the ANN framework offered the flexibility required to capture the complex, non-linear dependencies present in the datasets.

3.3.3 Validation and Performance Metrics

Model validation was carried out using multiple statistical criteria. The p-value of each coefficient was examined to ensure that only atom types with a statistically significant

influence were retained. Model accuracy was quantified using the mean absolute percentage error (MAPE) and the coefficient of determination (R^2). Models with R^2 values above 0.95 and low MAPE were considered sufficiently reliable for predictive purposes.

Unseen test data, representing 20% of each dataset, were used to confirm that the models generalised beyond the training set. Performance on the test data was compared directly with training performance to detect overfitting.

Finally, verification was extended to real fuels. Experimental property data from the National Renewable Jet Fuel Program,[116] Washington State University,[5, 15] and the Alternative Jet Fuel Database[119] were compared with predictions based on NMR-derived atom type distributions. This validation step provided an assessment of model performance in conditions representative of real jet fuels, beyond the simplified mixtures available in the training datasets.

3.4 Experimental Techniques for Populating Kinematic Viscosity Dataset

To facilitate the training and validation of the kinematic viscosity model, experiments were performed on a selection of the fuel samples available in the laboratory at Trinity College Dublin. The experiments were performed on the Anton Paar SVM 3001 Cold Properties Kinematic Viscometer at the SAFLAB in SMBC Aviation Capital HQ. To validate the measurements of the device 99.9% purity Toluene from Fischer Scientific was used and compared against the DIPPR database value.[120]

4. Results and Discussion

The temperature-dependent liquid density (ρ), temperature dependent surface tension (σ), and DCN regression model coefficients, are provided in Figures 23, 24, 28. Flash point was found to need a non-linear model, thus there is an extensive number of coefficients which are provided in appendix 1. The distillation curve model uses random forest regressor, the model architecture can be found in Table 6. Kinematic viscosity was developed using an ANN and the model architectures are shown in Table 7. A notable result is that the best-performing models for both liquid density and surface tension utilize the same set of atom types as independent variables, indicating that these atom types are fundamentally influential for both properties. The derived cetane number MLR model contains only four terms which are known to contribute strongly to ignition quality.[117] The p-values of all variables are below 0.01, underscoring the strong influence of each atom type within the predictive equations. This observation was rigorously interrogated and confirmed to be valid.

The predictive performance of each model was then assessed by evaluation against two discrete databases; (1) the training and testing dataset discussed previously comprised of pure components and pure component mixtures from the literature; and (2) the reference real fuels available at Trinity College Dublin as outlined above.

4.1 Atom Type Composition

To evaluate model performance against (2), atom type data is needed. The results of the ^{1}H ^{13}C HSQC NMR assignment are provided in Table 5 in mol % carbon and graphically in Figure 16 and 17 displayed in mass % to show the difference in each unitary representation, which we propose as a useful way to simplify complex fuel compositions, allowing for a visual inspection of "what a fuel looks like". The conventional fossil aviation fuels such as Jet-A POSF 10325 and JP-5 POSF 10289 have similar atom type distributions with a large quantity of paraffinic atom types, and a notable fraction of aromatic and cycloparaffinic types. The synthetic aviation fuels have a similarly large quantity of paraffinic atom types, but do not contain as high levels of aromatics, and strikingly no cycloparaffins for Shell-SPK POSF 5729.

The NJFCP fuels were included in the publication regarding the temperature dependent surface tension and density models.[126] Further NMR atom type quantification work on real conventional and synthetic aviation turbine fuel was done proceeding that work. Four real fuel samples were acquired from Washing State University, the samples were studied in the publication regarding distillation curve evaluation of ASTM D86 and D2887.[5] The four fuels were Gevo POSF 13718, S-8 POSF 5018, Amyris Farnesane POSF 11832, and World Energy HEFA H101. The atom type quantification is shown in Figure 17, a format which was developed for the presentation at the International Conference of Sustainable Avaition Research (ICSAR) to allow a synthetic aviation turbine fuel to be assessed in comparison with a conventional jet aviation turbine fuel such as Jet A POSF 10325 while being able to see the atom type mass % values. These SAFs show typical characteristics with high mass fractions of paraffinic atom types and virtually no other atom type. There are small concentrations of aromatic componets but none in quantities greater than 0.14 mass %.

Table 5. Atom type compositions determined using ^{1}H ^{13} HSQC and quantitative ^{1}H NMR spectroscopy for real synthetic and fossil aviation fuels. The rows represent the composition of the fuel as described through atom types.

		Atom Type Mol % Carbon																
Fuel Sample	**POSF**	**pCH_3**	**pCH_2**	**pCH**	**pC**	**oCH_2**	**oCH**	**$cyCH_2$**	**cyCH**	**$alCH_3$**	**$alCH_2$**	**alCH**	**ArCH**	**ArC**	**rjC**	**ajH**	**$ajCH_3$**	**$ajCH_2$**
FACE A Gasoline	-	53.11	23.87	15.32	5.77	0.00	0.00	0.82	0.41	0.10	0.00	0.00	0.51	0.10	0.00	0.00	0.00	0.00
FACE C Gasoline	-	52.27	26.13	13.61	4.80	0.00	0.00	0.00	0.00	1.03	0.31	0.00	0.52	1.34	0.00	0.00	0.00	0.00
FACE F Gasoline	-	40.44	19.22	10.13	6.79	1.41	1.41	10.74	0.00	2.16	0.19	0.00	5.18	2.34	0.00	0.00	0.00	0.00
FACE I Gasoline	-	44.46	33.61	12.24	4.65	1.01	1.01	1.82	0.71	0.00	0.00	0.00	0.51	0.00	0.00	0.00	0.00	0.00
Jet A-1 (Weston)	-	22.54	48.92	9.84	0.00	0.00	0.00	1.53	0.66	3.03	1.46	0.23	6.76	4.72	0.31	0.42	0.06	0.12
Jet A POSF 10324	10325	20.34	44.09	5.97	1.33	0.00	0.00	12.54	4.54	2.97	1.79	0.39	0.54	5.14	0.37	0.46	0.18	0.10
JP-8 POSF 10264	10264	24.62	54.61	7.43	0.76	0.00	0.00	1.39	0.00	2.32	0.63	0.18	4.65	3.12	0.28	0.45	0.06	0.05
JP-5 POSF 10289	10289	23.99	42.21	8.38	2.69	0.00	0.00	2.50	3.69	2.79	2.05	0.00	6.40	4.84	0.47	0.60	0.17	0.18
Shell - SPK POSF 5729	5729	27.11	65.30	5.83	1.64	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.06	0.03	0.00	0.00	0.00	0.00
HRJ - Camelina POSF 7720	7720	29.64	60.30	8.71	1.17	0.00	0.00	0.00	0.00	0.03	0.01	0.00	0.09	0.05	0.00	0.00	0.00	0.00
Gevo - ATJ POSF 10151	10151	58.46	17.02	10.09	14.29	0.05	0.00	0.00	0.00	0.01	0.00	0.02	0.02	0.03	0.00	0.00	0.00	0.00
Sasol - IPK POSF 7629	7629	48.68	27.17	17.57	5.69	0.02	0.01	0.00	0.00	0.21	0.07	0.03	0.22	0.32	0.01	0.02	0.00	0.00
Gevo POSF 13718	13718	56.86	17.90	8.63	16.61	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S-8 POSF 5018	5018	25.95	66.47	5.87	1.42	0.00	0.00	0.00	0.00	0.05	0.05	0.00	0.09	0.10	0.00	0.00	0.00	0.00
Amyris Farnesane POSF 11832	11832	29.22	42.00	12.41	16.09	0.00	0.00	0.00	0.00	0.13	0.00	0.00	0.02	0.13	0.00	0.01	0.00	0.00
World Energy HEFA	H101	29.12	62.43	7.99	0.23	0.00	0.00	0.00	0.00	0.07	0.00	0.00	0.10	0.07	0.00	0.00	0.00	0.00

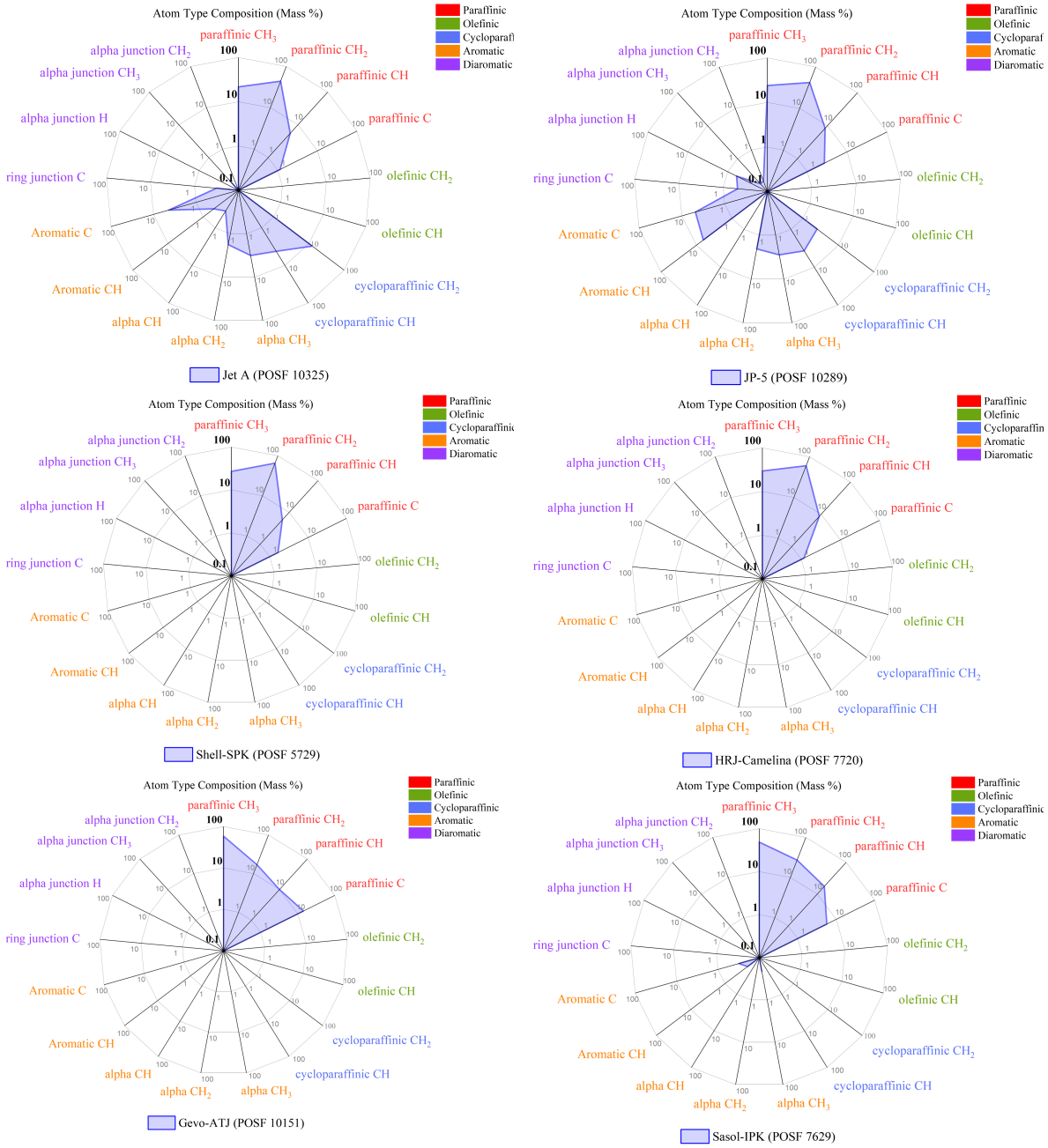


Figure 16. Atom type mole fraction on a log-radar plot for select fossil aviation fuels and synthetic aviation fuels, showing considerable compositional differences. The axis represents the mass fraction of each atom type in the sample.

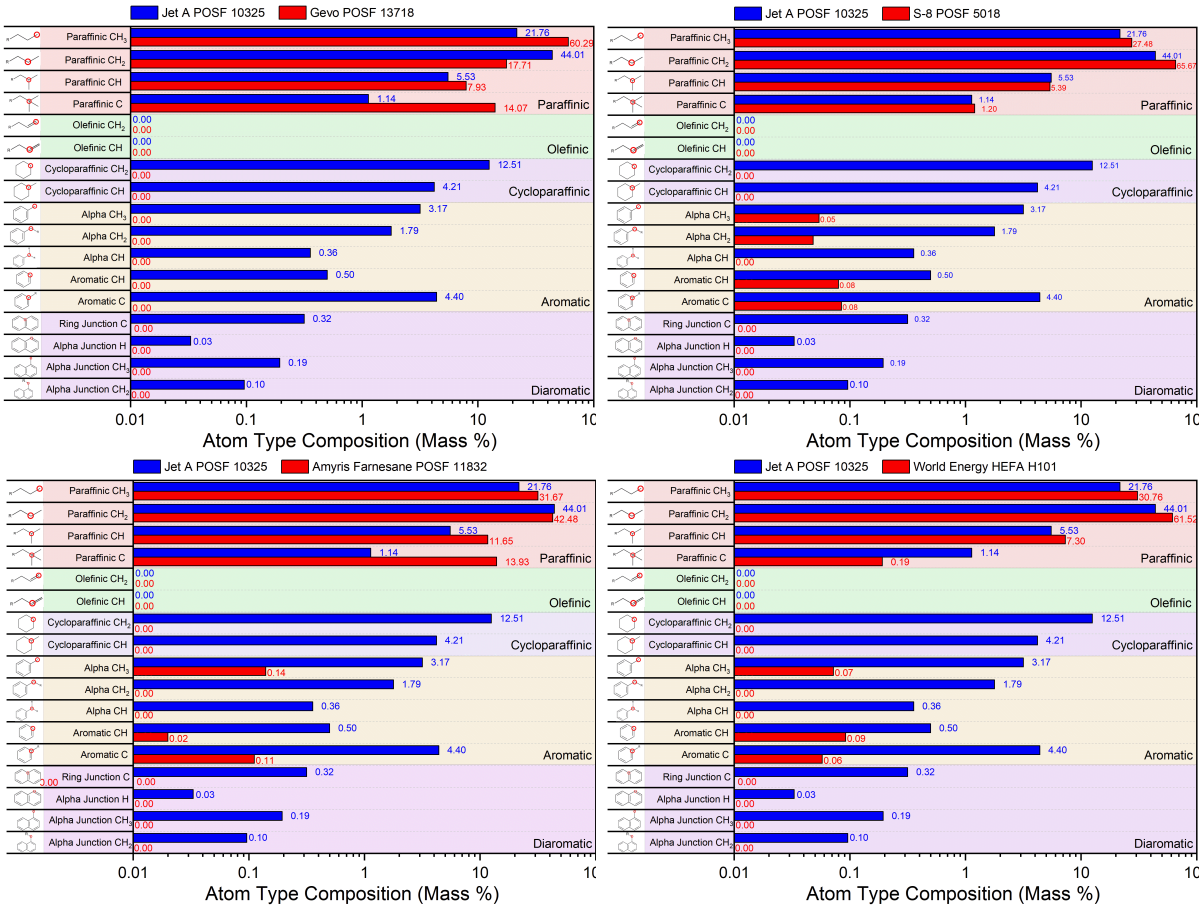


Figure 17. Atom type mass fractions of Gevo POSF 13718, S-8 POSF 5018, Amyris Farnesane POSF 11832, and World Energy HEFA H101 on a bar chart log plot with Jet A POSF 10325 for direct comparison with a conventional aviation turbine fuel.

4.2. Predictive Models

The following figures are the complete construction of the surface tension model, showing how each molecular class was added systematically and the operating equation evaluation. This full discussion was absent from the publication on the model due to length restrictions.[126] Following this, each model is presented in its final form, though the same careful evaluation has been done for each in the exact same manner.

Normal paraffins are the simplest pure component hydrocarbons so represent the ideal hydrocarbon class to begin the model construction. The longest equation of state with temperature and the atom types which can exist in a normal paraffin includes on three coefficients: temperature, paraffinic methyl and paraffinic methylene. This is shown on the left graph in Figure 18. Only unseen testing data (i.e. 20% of the entire dataset) is shown in the

figure to show the models can handle unseen cases and for ease of viewing. By using the unit of mole fraction or mass fraction the parameter inherently captures information about the other parameters, in other words, they are multicollinear. Though this works to the advantage of a simple model, especially with two atom types as, one atom type contains the information of both of them. The equation for calculating the mole fraction shows this $mol\,\%\ C\ of\ pCH_3 = \frac{number\ of\ pCH_3}{number\ of\ pCH_3 + pCH_2}$ clearly. Therefore, it was found that paraffinic methylene could be completely removed with no change in model performance. The model performance is acceptable with an R^2=0.993 for unseen testing data and a MAPE of 0.63%.

$$\sigma\left[\frac{mN}{m}\right] = \alpha_0 + \alpha_1 T + \alpha_2 pCH_3 + \alpha_3 pCH_2 \qquad \sigma\left[\frac{mN}{m}\right] = \alpha_0 + \alpha_1 T + \alpha_2 pCH_3$$

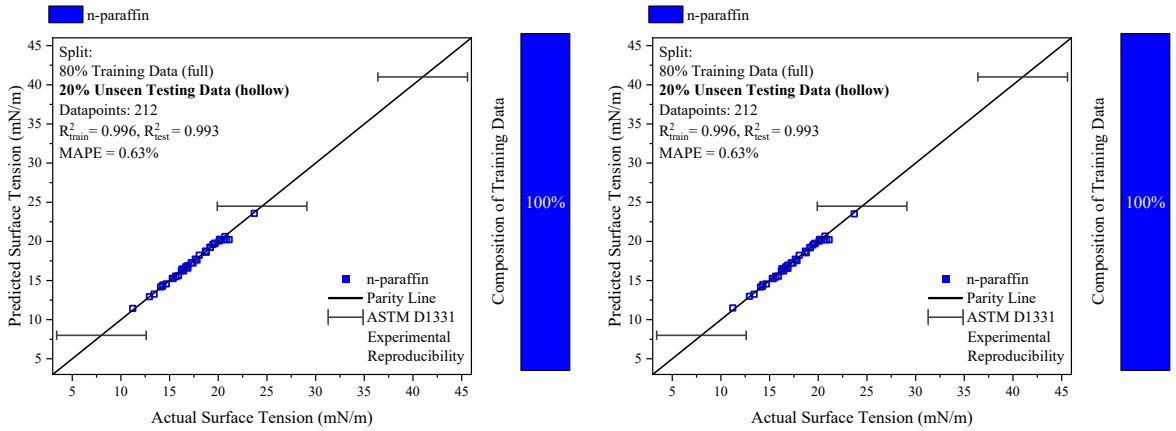


Figure 18. Temperature dependent surface tension, predicted surface tension versus actual surface tension with an 80:20 split in training (unseen data, hollow symbols) and testing (seen data, full symbols) data respectively for normal paraffins.

Branched paraffinic molecules (isomers) are the next step in hydrocarbon complexity as four atom types can be present: paraffinic methyl, paraffinic methylene, paraffinic methine, paraffinic quaternary carbon. In Figure 19, the left graph shows the unseen testing data for the full model considered for iso-paraffins. The R^2 has decreased slightly from the n-paraffins studied and there are less datapoints presents as iso-paraffins are less studied in literature. In the right-hand side graph, the combined dataset of n-paraffins and iso-paraffins has been used to train the model and term reduction was applied. It was found that to accurately predict normal and isomeric paraffins paraffinic methylene was not required. The R^2 with n-paraffins included increases to 0.981.

$$\sigma\left[\frac{mN}{m}\right] = \alpha_0 + \alpha_1 T + \alpha_2 pCH_3 + \alpha_3 pCH_2 + \alpha_4 pCH + \alpha_5 pC \qquad \sigma\left[\frac{mN}{m}\right] = \alpha_0 + \alpha_1 T + \alpha_2 pCH_3 + \alpha_3 pCH + \alpha_4 pC$$

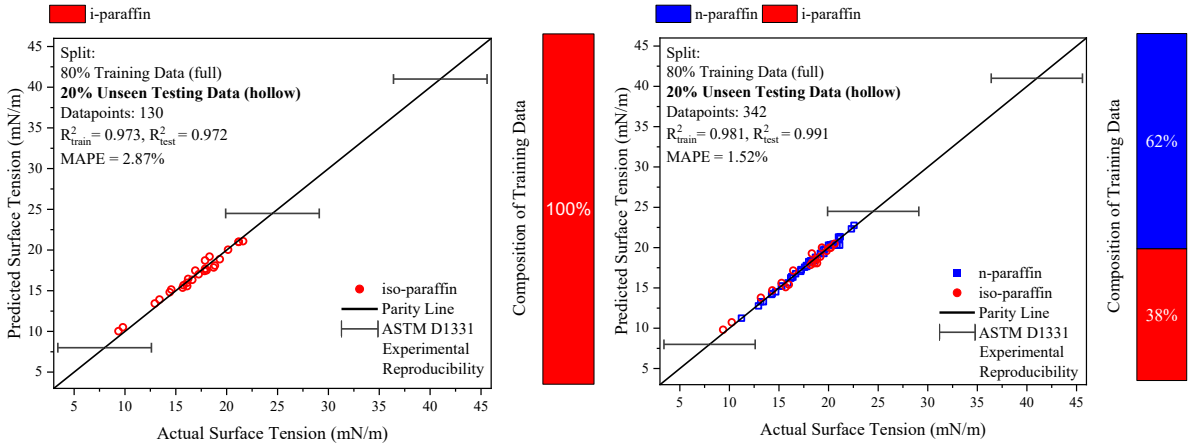


Figure 19. Temperature dependent surface tension, predicted surface tension versus actual surface tension with an 80:20 split in training (unseen data, hollow symbols) and testing (seen data, full symbols) data respectively for iso-paraffins and normal paraffins.

Cycloparaffins, and branched cycloparaffins were considered next in the model development. Cycloparaffins are rings of carbons, unlike aromatics they are not limited to 6 carbons in the ring. Of the cycloparaffinic structures considered the atom types that are present are: paraffinic methyl, paraffinic methylene, paraffinic methine, quaternary carbon, cyclic methylene and cyclic methine. Literature data for cycloparaffins is scarce and centres around cyclohexane and substituted cyclohexane. In Figure 20, the left graph shows only cycloparaffins modelled. The model maintains high accuracy (R^2>0.95) despite the increased complexity of the molecules in the training set. This acts as a validation that the methodology and the modelling approach is valid and generalisable. In fact, this claim has further support when the datasets for normal, iso, and cyclo paraffins are combined and when used to train a model it was found that the model maintains high fidelity and a low error as shown in the graph on the right.

It was found during the reduction of terms that both cyclic methylene and cyclic methine were required for the accurate prediction of cycloparaffins and cycloparaffinic isomers. This follows the information which is provided by each atom type logically. Cyclic methylene captures the carbons in the ring which are only connected to another cyclic atom type. This means it acts as a proxy for molecule size, i.e. the larger the mole fraction of cyclic methylene

the larger the ring. Though this information is saturated when dealing with a pure cycloparaffin, such as cyclopentane, cyclohexane, cycloheptane would all have only cyclic methylene meaning the information about the ring size is lost, this is inherently in the molecule shape as it is composed of only one atom type in a pure state. Though, when an isomer is studied, cyclic methylene shows information on the ring size, and cyclic methine shows information on the branching which is statistically significant, which explains why both atom types are needed in the operating equation on the right side of the figure.

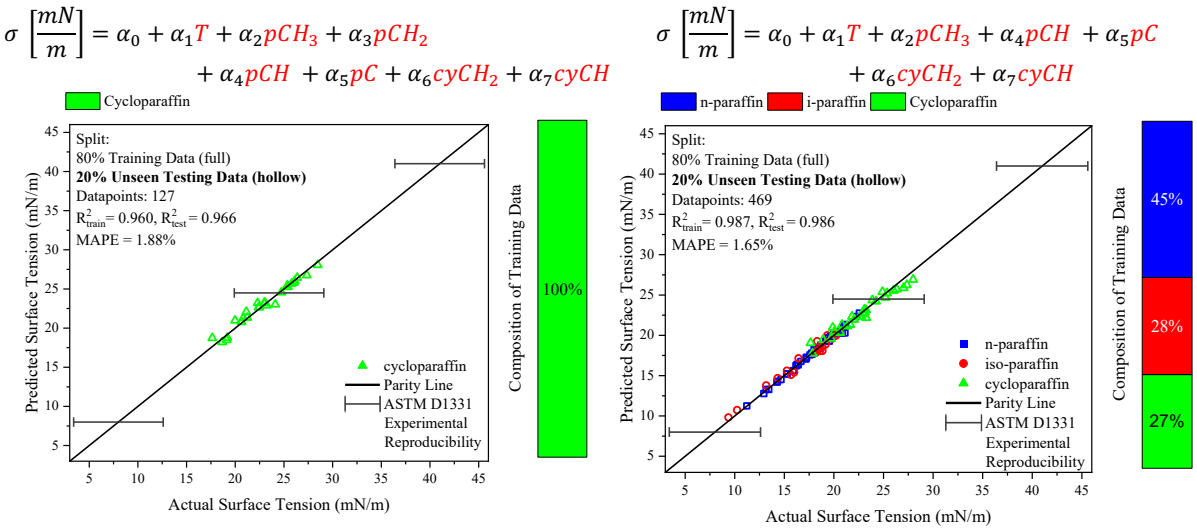


Figure 20. Temperature dependent surface tension, predicted surface tension versus actual surface tension with an 80:20 split in training (unseen data, hollow symbols) and testing (seen data, full symbols) data respectively for cycloparaffins, iso-paraffins and normal paraffins.

The last pure component hydrocarbon class to model are aromatics, which are composed of a benzene ring (6 carbons) and can have branched structures. Due to the deshielding nature the benzene ring has in NMR spectroscopy, it is possible to distinguish alpha atom types (nuclei one bond from an aromatic ring) from the paraffinic atom types. It is noted, that nuclei further on a branch from an aromatic ring can be identified in ^{1}H spectroscopy, due to overlapping peaks from the number of components in a fuel only the alpha nuclei were identifiable in this study, all other nuclei in the branch were treated as paraffinic atom types. The atom types which are considered for aromatic molecules are: paraffinic methyl, paraffinic methylene, paraffinic methine, quaternary carbon, alpha methyl, alpha methylene, alpha methine, aromatic methine, aromatic quaternary carbon, ring junction carbon. In Figure 21, on the left graph and equation, the model performs sufficiently well with an R^2 of 0.984 for unseen testing data, the

surface tension are noted to be to a higher value than the other classes, this is due to both molecular weight increases, and more uniform linear stacking in intermolecular distributions. Though, on the addition of the aromatic experimental data to the training dataset of the previously discussed hydrocarbon classes the model lost sufficient fidelity with an R^2=0.928 for unseen data. This decrease in prediction accuracy is seen in the right graph in Figure 21, where the tight range of predictions around the parity line broadened indicating a decrease in fidelity and model accuracy, as seen in the difference in the MAPE. This was an unseen problem as the addition of previous hydrocarbon classes did not warrant such an effect. It was found through term reduction that though six aromatic atom types were used in the original training, only two were statistically significant for the combine dataset model. Meaning, that regardless of if the four other atom types available were added the effect on fidelity was negligible. The two aromatic atom types are aromatic methine and the ring junction carbon, the former and indication of the level of substitution and the latter on ring size, i.e. whether naphthalene was present. The ring junction carbon is important as the unseen testing data which can be seen at 40 mN/m, by far the highest surface tension of the pure components is from these diaromatic components, 1-methylnapthtalene specifically.

$$\sigma\left[\frac{mN}{m}\right] = \alpha_0 + \alpha_1 T + \alpha_2 pCH_3 + \alpha_3 pCH_2 + \alpha_4 pCH + \alpha_5 pC + \alpha_6 alCH_3 + \alpha_7 alCH_2 + \alpha_8 alCH + \alpha_9 ArCH + \alpha_{10} ArC + \alpha_{11} rjC$$

$$\sigma\left[\frac{mN}{m}\right] = \alpha_0 + \alpha_1 T + \alpha_2 pCH_3 + \alpha_3 pCH + \alpha_4 pC + \alpha_5 cyCH_2 + \alpha_6 cyCH + \alpha_7 ArCH + \alpha_8 rjC$$

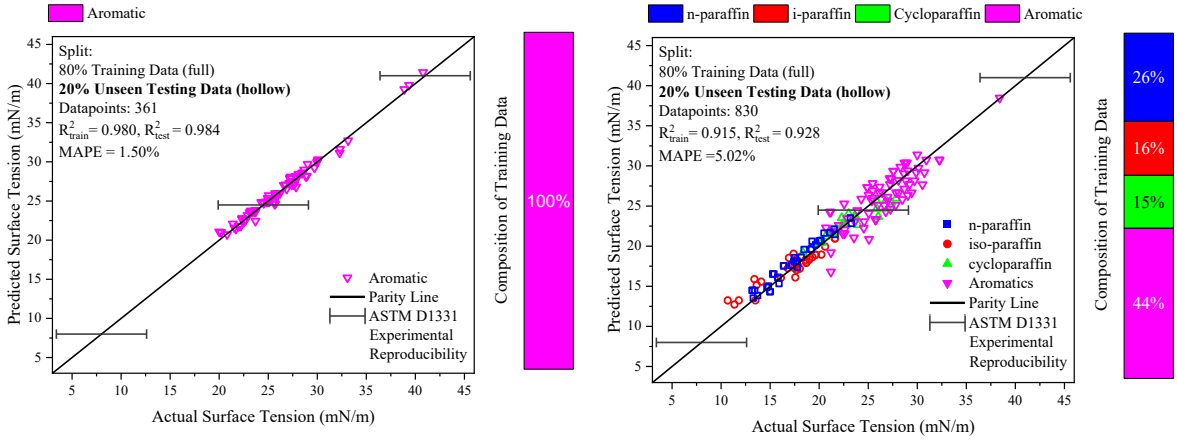


Figure 21. Temperature dependent surface tension, predicted surface tension versus actual surface tension with an 80:20 split in training (unseen data, hollow symbols) and testing (seen data, full symbols) data respectively for aromatics, cycloparaffins, iso-paraffins and normal paraffins.

The decrease in model fidelity when aromatics were trained with the normal paraffins, iso-paraffins and cycloparaffins revealed a characteristic of aromatic components. The substituents (any branch group from an aromatic) did not significantly contribute to the property when compared with the benzene ring. If it were possible to identify and quantify each atom type which is part of a substituted group then these contributions would be usable in the model. But this is not the case, and what happened instead was the relationship for paraffinic methyl, methylene, methine and quaternary carbon were being influenced by the substituted aromatics being trained. The aromatics compounds can be seen to have a higher surface tension than normal and iso-paraffins. The paraffinic coefficients are artificially raised by characteristics caused by the aromatic ring not the paraffinic substituents. This hypothesis required testing as it was a novel proposition which was not discussed in literature. There are two methods by which it could be tested; 1. Use substituent specific atom types in the operating equation, 2. Only train aromatic data on aromatic atom types. Method 1 is not possible with the current NMR methodology which does not allow for the quantification of atom types past the alpha atom and instead parses these atom types as paraffinic. This is because the resonance shift in the NMR spectrum overlaps with the paraffinic atom types as the deshielding effects of the aromatic ring do not extend along the bond past the alpha bond.

Method 2 is achievable as it involves only changing the way the aromatics are used in training the model. For each hydrocarbon class all of the statistically significant atom types are used in training, so for aromatics all seven atom types are used to train the model. Method 2 proposes that only aromatic atom types should train the model to aromatic components. Therefore, using this method, only aromatic methine and ring junction carbon mol % carbons would be used to train aromatic atom types. Then the relationship/coefficients for paraffinic atom types (which are still present in the aromatic compounds) are trained from normal and iso-paraffinic compounds. In summary, the paraffinic atom types are trained by normal and iso-paraffinic compounds, the cyclic atom types are trained by cyclic compounds, and the aromatic atom types are trained by aromatic compounds. Then the relations are applied to any compound or mixture regardless of what class they are. This means the most relevant class determines the contribution for the atom type it is associated with, this methodology was called group additivity firewall, as it stops the undue influence on coefficients from properties which

are not due to that atom type. The results of this are shown in Figure 22 where it can be seen that this hypothesis was confirmed. The model retains fidelity when aromatics are introduced to the training dataset with the other hydrocarbon classes when the atom types are screened and only aromatic types are allow to contribute.

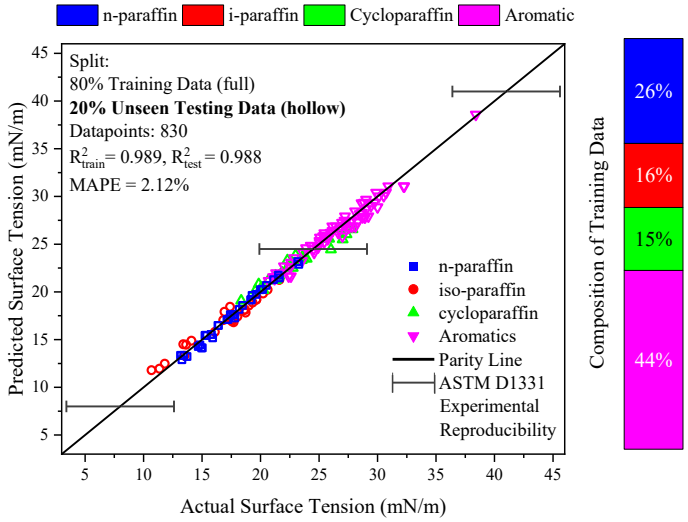


Figure 22. Temperature dependent surface tension, predicted surface tension versus actual surface tension for unseen testing data respectively for aromatics, cycloparaffins, iso-paraffins and normal paraffins using group atom type firewall methodologies.

$$\sigma\,[mN/m] = 61.9 - 0.1T - 36.7pCH_3 + 30.2pCH + 58.2pC - 6.3cyCH_2 + 14.4cyCH - 2.8ArCH + 49.9rjC$$

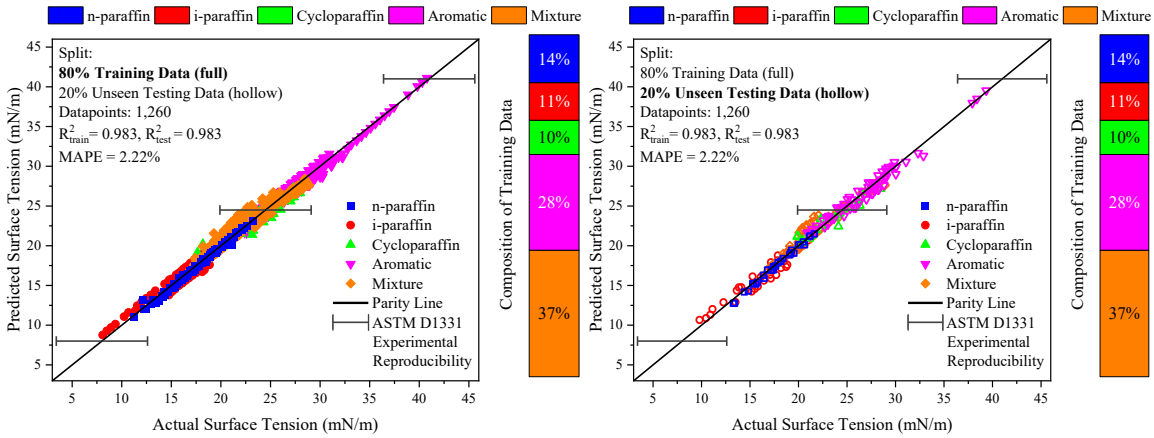


Figure 23. Temperature dependent surface tension, predicted surface tension versus actual surface tension for training and unseen testing data, showing the model has high fidelity for training and testing case.

The final addition to the training dataset was mixtures of pure components. This was used to populate the training dataset with mixtures where the atom types were calculated on a linear by mol basis in proportion with each compound. In Figure 23, the final model for surface tension is shown, training on the left and testing on the right. Both have the same R^2 of 0.983 this shows that the model is not overfitted to the training data as it maintains the same fidelity to unseen testing data. The ASTM D1331 shows the reproducibility limit for experimental measurements using a Du Nouy ring tensiometer. The predictive capability of the model is within these limits showing it is applicable to the prescreening application.

The full training methodology for the density model is not discussed, it follows the exact same path as the surface tension model. The progressive addition of hydrocarbon classes with the statistical reduction of the operating equation revealed the optimum model included the same atom types as the surface tension model. The density model is shown in Figure 24, the training on the left, and testing on the right. Again, as with the surface tension model, the R^2 was 0.996 for both testing and training showing the training data had not been overfit. The density dataset spanned a large range of 540 kg/m^3 to 995 kg/m^3 which is well outside the minimum and maximum range permissible for an aviation turbine fuel from the CRC Handbook of Aviation Fuel Properties. The predictive model has a mean average percentage error of 0.46% for the unseen testing data which shows its accuracy and applicability to prescreening.

$$\rho\left[\frac{kg}{m^3}\right] = 1101 - 0.9T - 556.9pCH_3 + 539.2pCH + 995.8pCH - 69.9cyCH_2 + 420.1cyCH + 28.3ArCH + 771.4rjC$$

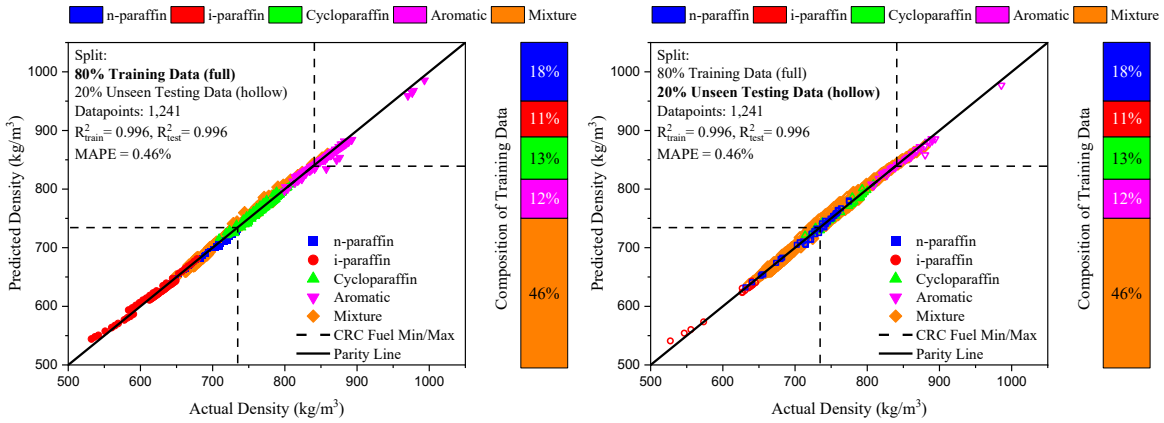


Figure 24. Temperature dependent liquid density, predicted surface tension versus actual liquid density for training and unseen testing data, showing the model has high fidelity for training and testing case.

In Figure 25, the flash point model is shown. The same linear regression procedure from the surface tension model development was followed. Though it was found that it was not appropriate for flash point which is an intrinsically more complicated property due to its close relationship to initial boiling point. A binomial non-linear regression model was found to be accurate for the prediction of flash point of pure components. The models have a training R^2 of 0.982 and a testing R^2 of 0.980, there is a small discrepancy which is not unexpected for a non-linear regression model as it is inherently more likely to overfit to the training data due to the increased number of predictors. This is subjugated by statistical analysis of the predictors, p-values were used to reduce the dimensionality of the coefficients so that only statistically significant values (p-value < 0.05) were included. This technique means that 21 parameters are used. It is noted that the training dataset (184 datapoints) is only 12% that of surface tension and density, 1,260 and 1,241 datapoints respectively. This is due to there only being a single flash point datapoint per compound, where surface tension and density are temperature dependent, meaning that one compound can have multiple data points. The flash point model is therefore more compositionally diverse, despite it having a lower number of datapoints over all. The testing data is composed of only mixtures as the limited number of pure component flash points were need to train the model, the mixtures are pure component in either binary or tertiary mixing configurations. The standard ASTM test method for flash point is D51 closed

cup tester, which has an experimental reproducibility within 3 °C for measurements near that of a jet fuel (~41 °C), therefore, the model's mean average percentage error of 0.81% shows that it is applicable to prescreening.

$$Flash\ Point\ [K] = \sum_{i=1}^{atom\ types} a_i x_i + \sum_{i=1}^{atom\ types} b_i x_i^2 + \sum_{i=1}^{atom\ types} \sum_{j=1}^{atom\ types} c_{i,j} x_i x_j$$

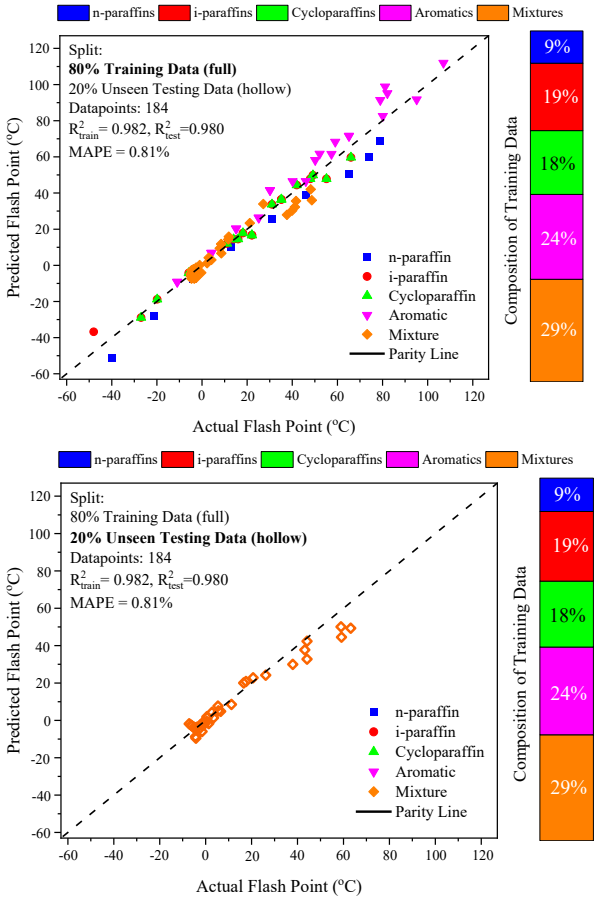


Figure 25. Flash point, predicted versus actual flash point for training data only showing the model has high fidelity to training and testing case.

Distillation curve is the most difficult property of a complex mixture to model and also one of the most important characteristics of fuel. It is very clear whether a fuel is compliant by evaluating the distillation curve as it is an indication of the range and quantity of components in the fuel, which is how the characteristics are derived (by physical nature of molecules). Linear and non-linear regression were not complex enough to model distillation curve. And, as the composition acquired by NMR spectroscopy is the entire solution, i.e. before reaching the initial boiling point, the only compositional data that is available is that of the entire

solution. Furthermore, because only the atom types are described from the compositional analysis, instead of the molecules, it is not currently possible to formulate a compositional matrix to allow for the modelling of a distillation curve by a piece-wise compound form is used by Yang et al.[5] and Bell at al,[72] these modelling approaches are appropriate for GCxGC data where the molecules are discernible. To achieve an accurate model using the NMR derived atom types a simple approach was used, where the data was treated like a dependent property, similar to how the surface tension and density models are temperature dependent, the distillation curve model is volume dependent. Where the volume is the % distilled. This is represented as at, for example, 180 °C 10 volume % of the sample has a evaporated.

Therefore, the predictors used for the model were the volume distilled and the atom types, the predicted value is the temperature at which that % volume distilled had been achieved. The training set is composed of liquid fuel mixtures primarily from the work of Yang et al.[5] where the atom types of the fuels are known through the Ure et al.[82] methodology. Samples used in the in the Yang study were requested so that the composition of atom types could be quantified and the data presented in their work could be used to model distillation curve in this work. A random forest regressor was developed, this was chosen as it was the first model to give high fidelity for distillation curve predictions. The model is shown in Figure 26, though random forest regressors are useful within the scope of their training data, and because the training data is all jet fuel, the model will have a bias towards outputting a positive result so compositional analysis must be used and evaluated before applying the model. But within the scope of jet fuel compositions, it is seen that the model performs well, a snapshot of 3 distillation volumes are shown and the $R^2 > 0.95$ showing good model performance. The overfitting of the training data can be seen slightly in 50% and 90% volume distilled where there are several testing points off of the parity line.

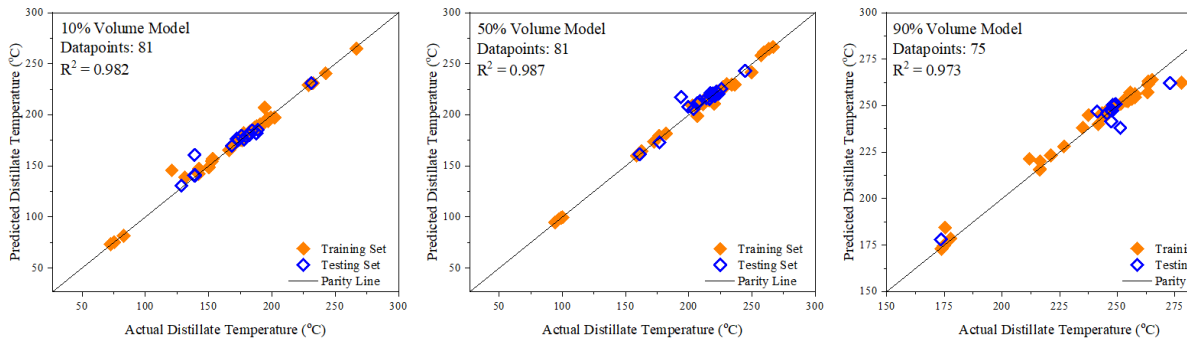


Figure 26. Parity plots for volume distillates required for prescreening: 10%, 50%, and 90%. The prescreening model uses volume as a parameter so overall there are 1,074 boiling points in the model.

The number of decision trees within a random forest regressor has a direct influence on the R^2 obtained from model predictions. As the number of trees increases, the model generally exhibits improved stability and predictive accuracy; however, beyond a certain point, further increases yield negligible gains.

Table 6. Model parameters and notes for random forest model architecture used for the training and testing of the distillation curve model, at 6 decision trees and 36 decision trees.

Setting	6 decision trees	36 decision trees	Definitions
Number of trees	6	36	Ensemble size
Criterion	squared error	squared error	For regression
Max depth	None	None	Trees grow until leaves are pure or cannot split further
Min samples split	2	2	Smallest node size considered for a split
Min samples leaf	1	1	Smallest leaf size
Max features per split	sqrt of p	sqrt of p	p is the number of predictors; reduces correlation among trees

As shown in Figure 27, the R^2 value converges at approximately n = 36 decision trees, indicating that additional trees do not significantly enhance model performance. A practical balance between computational efficiency and predictive fidelity can therefore be achieved with approximately n = 6 trees, beyond which improvements are marginal. This behaviour reflects the characteristic convergence of ensemble learning methods, where model variance decreases rapidly at low ensemble sizes before reaching a plateau.

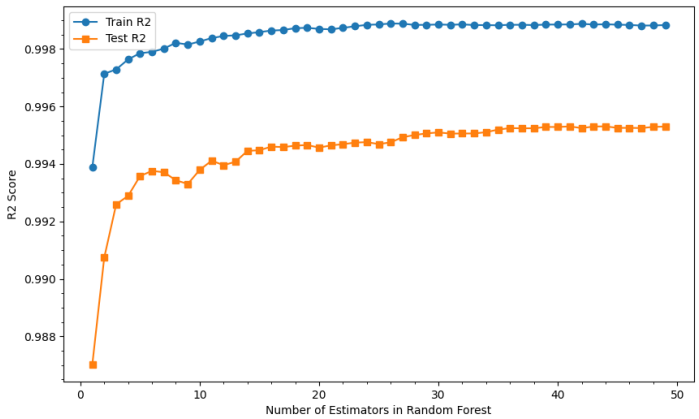


Figure 27. R^2 reduction analysis of random forest number of decision trees. Number of estimators represents the decision trees, showing that the random forest converges at 36.

Derived cetane number is a complex fuel characteristic to measure and predict as it requires the use of an ignition quality tester (IQT). It is not formally required by ASTM D4054 for new synthetic aviation fuel evaluation, though it is required for batch testing in ASTM D7566. Only mixtures measured on an IQT were used in the training and testing of the model. A linear regression model was used to construct the operating equation for DCN. Only three primary atom types are present, and one group of atom types. The group, aromatics, corresponds to the summation of all the atom types that are aromatic, i.e. aromatic methine, aromatic carbon and the alpha atom types. Aromatics are known to decrease DCN, with the lowest DCN values recorded from aromatic compounds. This is due to the poor combustion performance of polyaromatic hydrocarbons. DCN can be thought of as a proxy for ignition quality, as it is formally derived from it. The R^2 of the training and testing is 0.770 and 0.876, this does not

$$DCN = 1.43 + 9.48pCH_3 + 84.12pCH_2 + 7.38Aromatic + 13.74cyCH_2$$

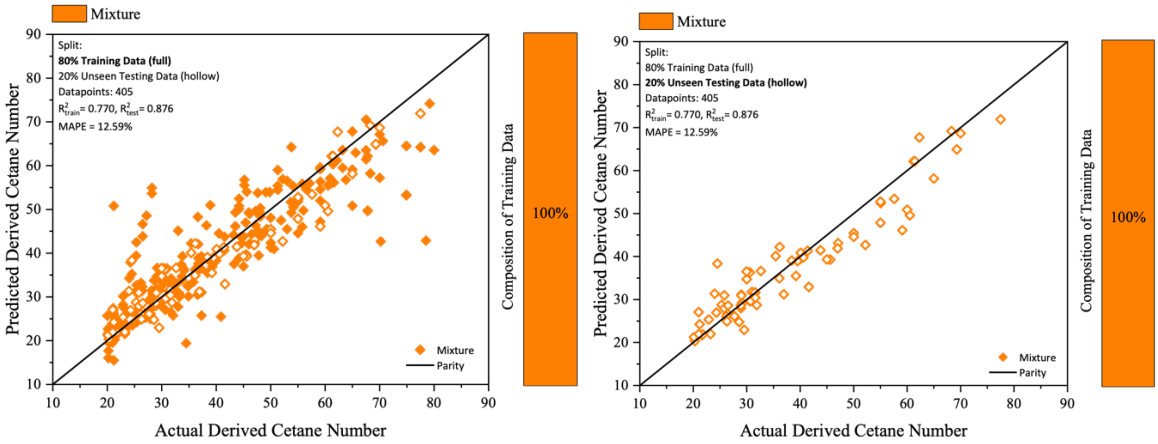


Figure 28. Derived Cetane Number (DCN), predicted DCN versus actual DCN for 405 experimental measurements.

reach the 0.95 as required for the other models. But it can be seen in Figure 28 that the model captures the general trend, and continues to capture it in testing unseen data. These are for gasolines, diesel and jet aviation fuels which show the diversity in the model's capabilities. The mean average percentage error of 12.59% is acceptable for this property as it is meant to serve only as an indicator.

The kinematic viscosity of jet fuel was found to not be predictable by a model trained exclusively on pure components. For this reason, a mixing rule was implemented to calculate kinematic viscosities of mixtures using the linear by mole mixing rule as described in 3.3.1. The model is an artificial neural network and the performance is shown in Figure 29 and model parameters in Table 7. ANN was chosen as linear and polynomial regression was not capable of capturing the complex nature of viscosity. Random forest was found to overfit the data, and was unable to capture the temperature dependent relationship. This lead to random forest not performing well outside of the mixture training set, which is due to the restricted set of temperatures used in training. A neural network was found to perform the best for training, testing and validation. The training and testing R^2 are 0.9989 and 0.9950 respectively. The model can generalise a temperature dependence, which allows it to outperform random forest regressor. Additionally, the training data spans a large range from 1 – 16 mm^2/s which captures the kinematic viscosities which can be possible in jet aviation fuel.

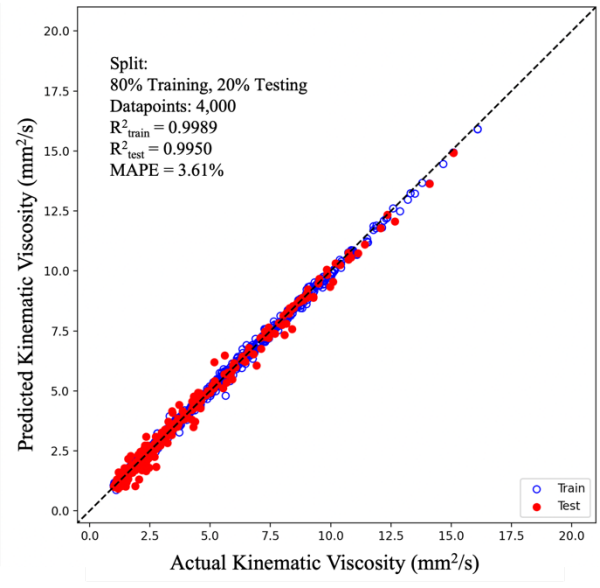


Figure 29. Predicted kinematic viscosity against actual kinematic viscosity as computed from artificial neural network.

Table 7. Model parameters for kinematic viscosity artificial neural network architecture as trained using H2O AutoML package in Python.

Parameter	Value	Notes
Network type	Fully connected feedforward	Single hidden layer
Input features	16	Standardise or normalise inputs
Hidden layers	1	100 neurons
Output units	1	Scalar regression target
Hidden activation	ReLU	Good default for regression with standardised inputs
Output activation	Linear	Suitable for regression
Loss function	Mean squared error	Optimises squared error
Optimiser	Adam	Stable default for tabular data
Learning rate	0.001	Tune on a validation set
Batch size	32	Tune between 32 and 256
Epochs	200	With early stopping

4.3 Experimental Results

Table 8 presents the measured densities and kinematic viscosities of a selection of jet fuels, surrogate mixtures, and reference fuels across the investigated temperature range. All samples exhibit the expected inverse temperature dependence of both density and viscosity, with values decreasing as temperature increases. Jet A-1 Weston showed a consistent trend from 0.8219 g cm^{-3} at −20 °C to 0.7768 g cm^{-3} at 40 °C, and from 3.76 $mm^2 s^{-1}$ to 1.21 $mm^2 s^{-1}$ in viscosity, demonstrating typical behaviour for middle-distillate fuels.

These results were used to build the training and testing dataset for the calculated mixtures for viscosity. The kinematic viscosity data from Edwards et al. was then used to validate the model.[115]

Table 8. Kinematic viscosity and density experimental values for real and surrogate jet fuel using Anton Paar SVM 3001 in Trinity College Dublin.

Sample	**Temperature °C**	**Density g/cm^3**	**Kinematic Viscosity mm^2/s**
Jet A-1 Weston	40	0.777	1.21
Jet A-1 Weston	20	0.792	1.61
Jet A-1 Weston	0	0.807	2.34
Jet A-1 Weston	-20	0.822	3.76
n-heptane (0.34) + 1,3,5-trimethylbenzene (0.33) + iso-octane (0.33)	40	0.714	0.53
n-heptane (0.34) + 1,3,5-trimethylbenzene (0.33) + iso-octane (0.33)	20	0.731	0.62
n-heptane (0.34) + 1,3,5-trimethylbenzene (0.33) + iso-octane (0.33)	0	0.748	0.77
n-heptane (0.34) + 1,3,5-trimethylbenzene (0.33) + iso-octane (0.33)	-20	0.764	1.00
FACE A Gasoline	40	0.675	0.53
FACE A Gasoline	20	0.693	0.62
FACE A Gasoline	0	0.710	0.77
FACE A Gasoline	-20	0.727	0.98
FACE F Gasoline	40	0.687	0.49
FACE F Gasoline	20	0.698	0.59
FACE F Gasoline	0	0.722	0.71
FACE F Gasoline	-20	0.739	0.90
FACE I Gasoline	40	0.670	0.49
FACE I Gasoline	20	0.688	0.57
FACE I Gasoline	0	0.705	0.69
FACE I Gasoline	-20	0.722	0.88
S8 POSF 5018	40	0.736	1.33
S8 POSF 5018	20	0.751	1.81
S8 POSF 5018	0	0.766	2.67
S8 POSF 5018	-20	0.781	4.47
Gevo POSF 13718	40	0.682	1.79
Gevo POSF 13718	20	0.694	2.47
Gevo POSF 13718	0	0.705	3.68
Gevo POSF 13718	-20	0.721	6.03
Amyris Farnesane POSF 11382	40	0.757	3.78
Amyris Farnesane POSF 11383	20	0.771	5.83
Amyris Farnesane POSF 11384	0	0.785	10.38
Amyris Farnesane POSF 11385	-20	0.799	23.08
FACE C Gasoline	40	0.670	0.49
FACE C Gasoline	20	0.688	0.58
FACE C Gasoline	0	0.705	0.70
FACE C Gasoline	-20	0.722	0.88

4.4 Validation of Prescreening Models

The validation dataset for the prescreening models was built from a database of reference jet aviation turbine fuels, including conventional and synthetic fuels largely from the work of Edwards et al.[115] These fuels are part of the National Jet Fuels Combustion Program (NJFCP) conducted by the US Air Force and NASA.[116] A small sample of six fuels from this program are stored in Trinity College Dublin, allowing for the quantification of the atom types as were shown in Figure 16. The database is not complete and due to the sample quantity available in Trinity College Dublin the tests cannot be performed meaning that some properties are underrepresented, i.e. surface tension. These fuels have an identification number which is prefaced with "POSF" as an indicator of the batch which it is derived from.

These fuels were chosen due to the nature of their existence in literature, meaning this work is now directly comparable with other predictive methods by comparing the capabilities using the NJFCP fuels which many institutions have access to. Additionally, due to the novel NMR methodology being used in Trinity College Dublin for the composition analysis of fuels the validation data set is limited by the samples available in the laboratory.

The surface tension model was validated using Jet A POSF 10325 and JP-5 POSF 10289 as these samples were the only NJFCP fuels with surface tension data. Future work will validate the surface tension model with synthetic fuel samples. Nonetheless, with two conventional aviation fuels the model is shown to have excellent fidelity. The performance of the validation set with the predicted and experimental surface tensions is shown in Figure 30. The CRC Handbook of Aviation Fuel Properties Jet A line, and the CRC World Fuel Survey average show that if these fuels were being analysed within the ASTM D4054 evaluation scope with only the predicted values they would receive a positive result. It is noted that the fuels do not have to fall within the average, only that they must conform to the standard characteristics seen for Jet fuels. The predictions at lower temperatures show that there is an increase in error likely due to the inconsistent amount of training points compared with the higher temperatures.

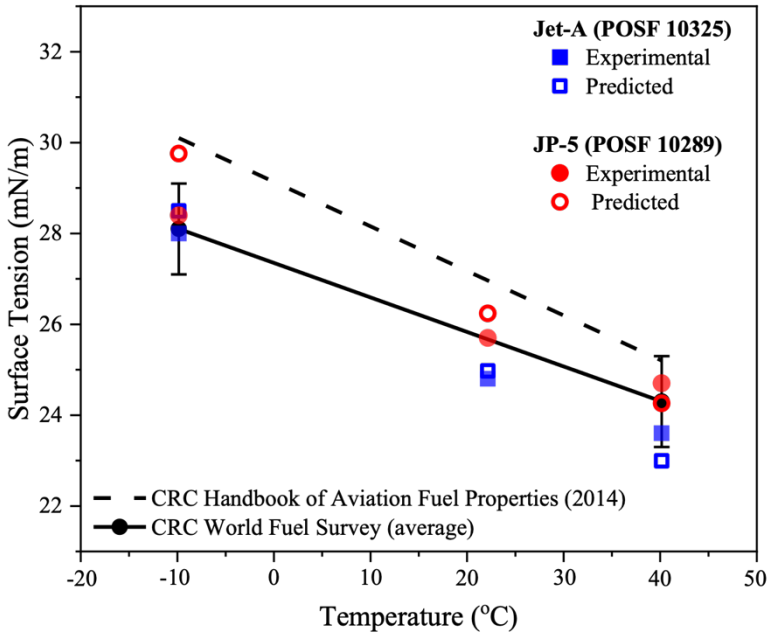


Figure 30. Surface tension versus temperature, literature experiment (full symbols) and unseen model predictions of this study (hollow symbols). For reference the ASTM fit-for-purpose properties are included as lines, which are derived from the CRC World Fuel Survey (2006). The bars represent the range of surface tensions measured for samples of Jet A, JP-5 and JP-8 in the survey.

The density model was validated using temperature-dependent data for Jet A POSF 10325 and JP-5 POSF 10289. Synthetic aviation turbine fuel measurements of density at 15 °C were available in literature and were used to validate the model. The temperature-dependent validation is shown in Figure 31, it is noted that there is a deviation in the temperature relationship between the predicted and measured properties. This could be caused by a temperature-dependent intermolecular interaction which is not accounted by the binary mixtures and pure components which the model is trained on. Future work will assess a density model which is train wholly on complex mixtures with atom type compositional data. Though, even with the temperature relationship the predicted values match the experimental values excellently demonstrating its applicability to prescreening. Four predicted densities for fuels at 15 °C are shown in Table 9 and show excellent agreement with the experimental values with the absolute error ranging from 3 – 0 kg/m^3.

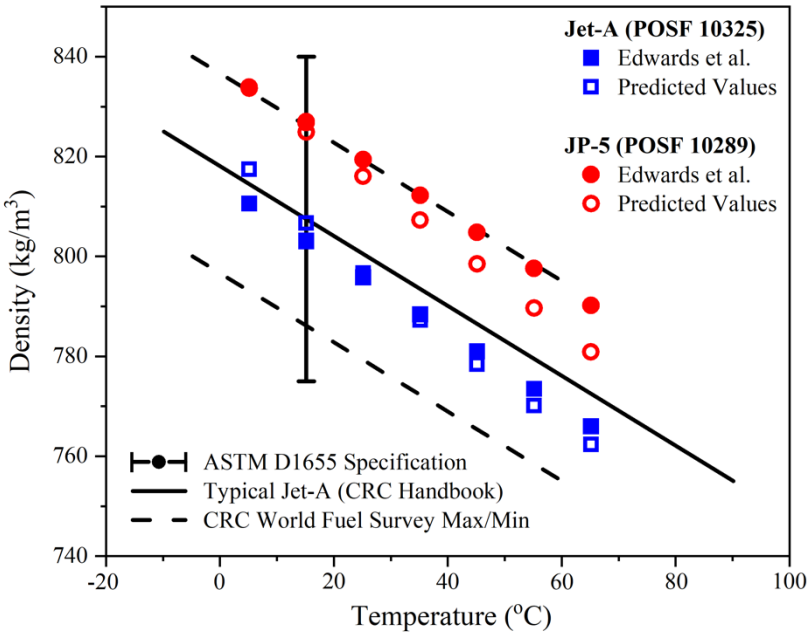


Figure 31. Liquid density versus temperature, literature experiment (full symbols) and unseen model predictions of this study (hollow symbols). The ASTM fit-for-purpose properties are included as lines for reference. The ASTM D1655 specification is used as a qualification (limits @ 15°C) metric in the initial ASTM D4054 evaluation.

Table 9. Liquid density at 15 °C as is required in Table A4.1 in ASTM D4054 for Fast Track Evaluation and required ASTM D7566 annexes for 2 conventional aviation turbine fuels and 2 synthetic aviation turbine fuels from literature.

Fuel	Density 15°C		
	Predicted (kg/m^3)	Experimental (kg/m^3)	Absolute Error (kg/m^3)
Jet-A POSF 10325	805	803	2
JP-5 POSF 10289	825	826	1
Shell-SPK POSF 5729	740	737	3
HRJ-Camelina POSF 7720	757	757	0

The flash point validation dataset is composed of the three conventional jet aviation fuels and three synthetic jet aviation fuels. Figure 32 shows two red regions, one as the minimum requirement for flash point as required by ASTM D7566 and the crossed region as the maximum requirement for flash point by ASTM D4054. It can be seen that while most of the predicted points fall within the range, Sasol-IPK POSF 7629 does not. Sasol-IPK is a highly paraffinic synthetic blending component. This is something which all of the models struggle

with as isomeric paraffins have the same atom types but may have different shapes (structural isomers). Structural isomers can exhibit different properties which cannot be captured by the analysis here. Though, the changes in the properties for structural isomers is usually small, this discrepancy with Sasol-IPK may be due to lack of compositional coverage in the training space. The model shows acceptable fidelity for a predictive method. Increased model complexity did not result in a better model. Instead, more data or better, more information such as molecular weight and initial boiling point would have allowed for greater predictive power. In Figure 33, a parity plot shows the property-property correlations between flash point, density, surface tension, initial boiling point and freezing point. Boiling point and flash point have a near perfect correlation, which can be deduced logically as both are a measure of the volatility of a sample at different temperatures but looking for different outcomes. Future work will develop a property-structure prediction model whereby atom types are used alongside physical characteristics would train a model, to allow for increased fidelity by capturing molecular interactions and characteristics not inherent to atom type descriptions. It is noted that other properties have a high correlation in figure 33 such as surface tension and density. This will be evaluated in the future property-property work as well.

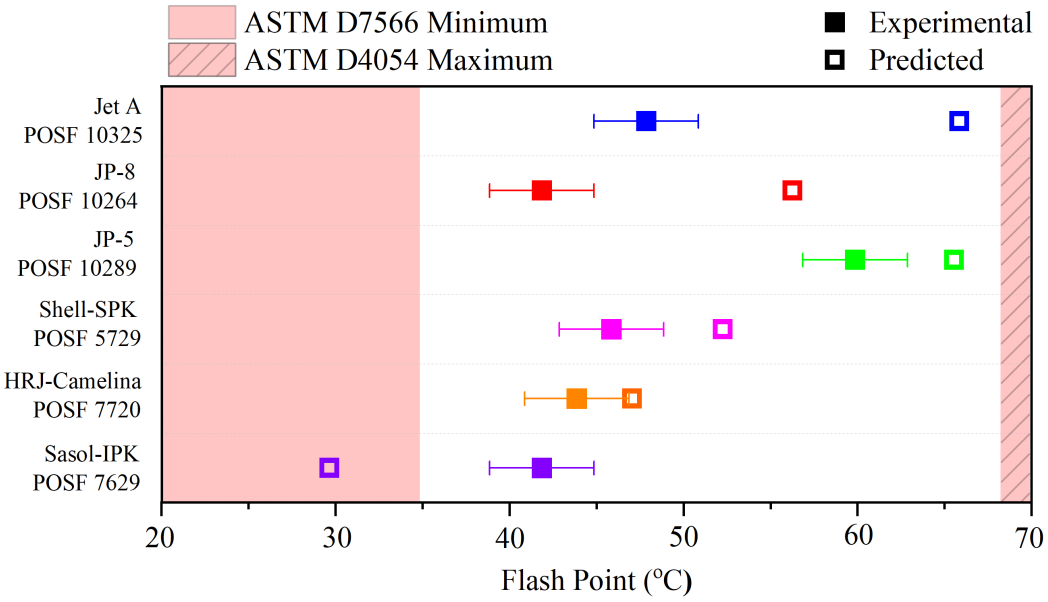


Figure 32. Flash point predictions of conventional and synthetic aviation turbine fuel. The error bars represent the ASTM D56 reproducibility of flash point via the closed cup method. The predicted error range is 1.74 – 5.64%. The experimental data was taken from Edwards et al.[115]

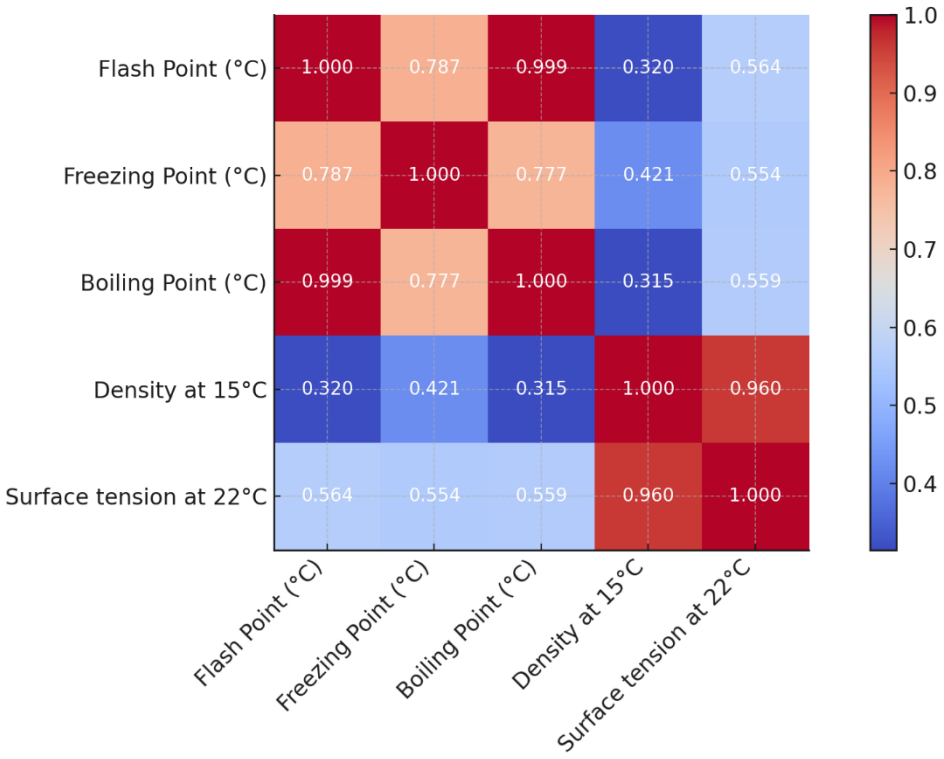


Figure 33. Property – Property correlation matrix for flash point, freezing point, boiling point, density and surface tension using physical property data gathered from DIPPR database.

The distillation model was verified with Jet A POSF 10325 and JP-8 POSF 10264 across the full range of distillation. The predictions shown in Figure 34 have high fidelity and be within the ASTM D4054 evaluation specification for distillation curve. This shows the models applicability to prescreening of new jet aviation fuels. Future work will expand the validation dataset as a lot of the fuel data gathered was used for the training of the distillation model as strictly mixtures of complex hydrocarbons are required. The use of a random forest regressor and fuel data could lead to overfitting and a restrictive predictive space. For this, a rudimentary distillation curve artificial neural network was constructed to quickly validate the results of the random forest.

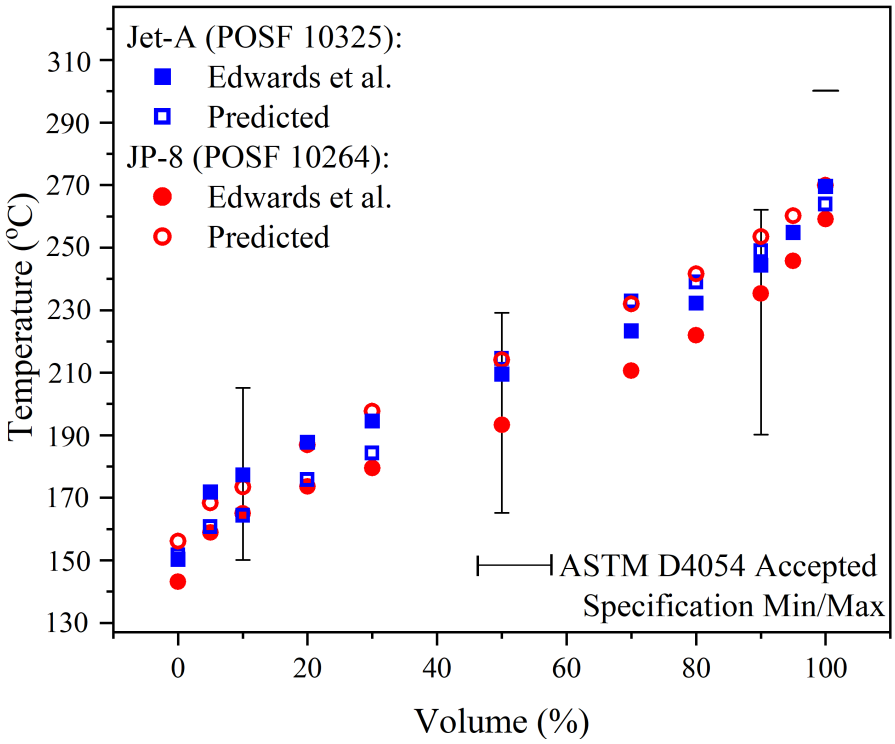


Figure 34. Temperature versus volume, literature experiment (full symbols) and unseen model predictions of this study (hollow symbols). The ASTM fit-for-purpose properties are included as lines for reference.

Derived cetane number did not have a high R^2 on the training or testing data used to form the linear regression equation. Though, it can be seen in Figure 35, that this does not affect the prediction results greatly. Dooley et al.[79] showed that the prediction of derived cetane number does not need many functional groups. More functional groups were included than in that study as there is the restriction of not knowing the molecular weight which was used as a proxy for molecular size. The model shows excellent reproducibility for the three conventional and three synthetic fuels which are used in the validation dataset. This performance shows that the model, as it is trained currently, is more than capable of being used for prescreening fuels using the limited atom types selected. This is because pCH_3 and pCH_2 and their ratio are the single largest contributor to ignition quality. As well, a grouped aromatic term allows the model to accurate predict the decrease in the DCN due to the aromatic content which has a known effect to decrease ignition quality.

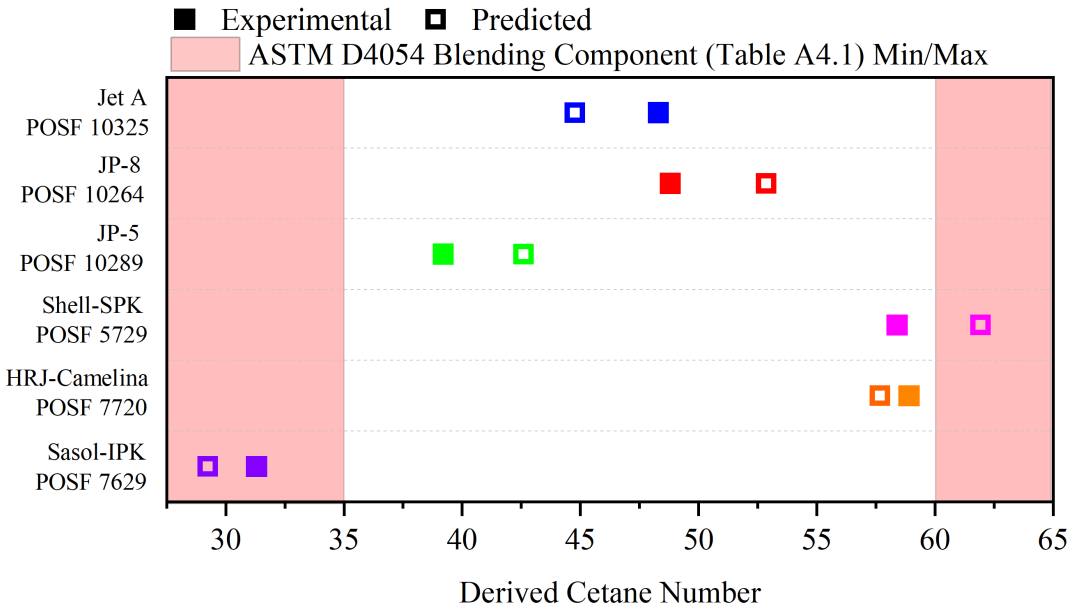


Figure 35. Derived Cetane Number validation data set of real fossil and synthetic aviation turbine fuels using atom types quantified from ^{1}H ^{13}C HSQC NMR spectroscopy. The experimental data was taken from Edwards et al.[115]

The validation dataset for kinematic viscosity at 20 °C is shown in Figure 36. The maximum specification allowed is included from ASTM D7566 which specifies kinematic viscosity cannot exceed 8 mm^2/s at 20 °C. The model shows that for the three conventional fuels the performance is good, with prediction JP-8 being nearly identical to the experimental result, but less so for JP-5 which has an overprediction. The synthetic aviation fuels have a wider prediction range, again due to the increased number of iso-paraffins which was unable to be covered in the dataset for the kinematic viscosity model. Kinematic viscosity is not quantity measured by a device, instead it is a result of the dynamic viscosity normalised by the density. This means, that the training dataset, and the validation dataset, both need the dynamic viscosity and the density of the fuel at X °C, Future result will extend the dataset and provide experimental measurements for fuels and complex mixture to allow the model, which is currently data limited, to reach higher fidelity.

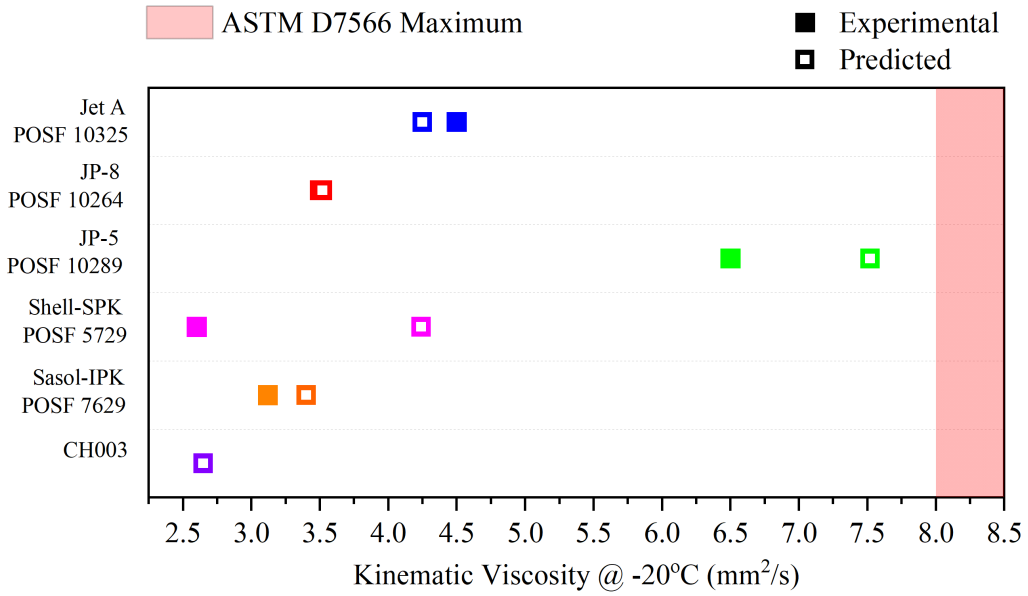


Figure 36. Kinematic viscosity validation data set of real fossil and synthetic aviation turbine fuels using atom types quantified from ^{1}H ^{13}C HSQC NMR spectroscopy. The experimental data was taken from Edwards et al.[115]

4.5 Abolition Study of Density and Surface Tension

Figures 30 and 31 indicate that while both models do achieve a useful predictive capability, the models do not perfectly capture the temperature dependent density and temperature dependent surface tension behaviours of the JP-5 in particular. Though uncertainty in the experimental measurement should not be discounted as a source of this discrepancy, it is expected that a more complex training database will result in a better capture of this behaviour.

The surface tension model shows less accuracy, which can be indicative of an inherently greater complexity in these phenomena compared to density. Despite these shortcomings, the predictions for both properties are sufficiently accurate to classify the analysed fuels as "fit-for-purpose", thus fulfilling the primary objective of this study.

The reference real fuels, which serve as the benchmark for evaluating the predictive ability of the models, are highly complex fluids composed of thousands of discrete molecules. This complexity highlights the need for a more diverse set of complex mixtures in the training data to capture the full range of compositional variability. Ideally, a model would be trained on a dataset that comprehensively and evenly represents the compositional space of the fluids being

analysed. However, this is challenging in practice due to the vast compositional diversity found in conventional and synthetic aviation fuels, as well as the significant time and cost associated with conducting experimental determinations for reference pure component mixtures.

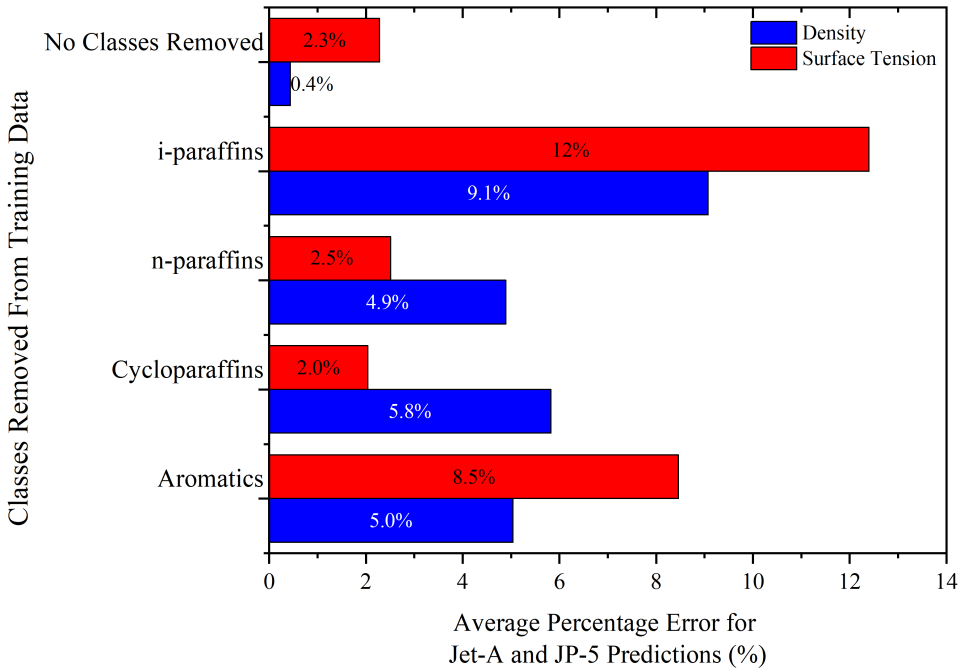


Figure 37. Average percentage error in the model prediction of temperature dependent density and temperature dependent surface tension of Jet-A POSF 10325 and JP-5 POSF 10289, relative to experiment, for different descriptions of the compositional space. The y-axis denotes the singular absence of each molecular class from the training data.

From Figure 16, the large variabilities in the composition space occupied by the real reference fuels becomes obvious, particularly the lower prevalence of aromatic and cycloparaffinic atom types in the synthetic aviation fuels. To study the effect of compositional space on model performance, the influence of this factor is analysed using the original datasets. In this analysis, the training set comprises only pure components, as mixtures may include multiple classes. Model training by multiple linear regression was performed again, but with individual molecular classes removed from the training set, and the performance of the resulting model assessed, all other variables held constant. Figure 37 shows the results against density and surface tension data for Jet-A POSF 10325 and JP-5 POSF 10289.

In particular, for both density and surface tension, the removal of isomerized paraffins and aromatics have a considerable degrading effect on the predictive ability, emphasizing their importance.

With respect to density, the removal of any hydrocarbon class leads to a considerable, but more consistent, increase in the model error, of the order of 5%. The extent of performance degradation for the surface tension model is more varied depending on which molecular class training data is removed. This feature further points to the more complex nature of the surface tension physical phenomenon, over that of liquid density.

The analysis evidences the importance of having a comprehensive coverage in molecular class/atom type training data in order to accurately predict the properties of real complex fuels, by this, and similar methodologies. Furthermore, and of fundamental scientific value, though density and surface tension are usually labelled as “physical” properties as no covalent bonds are broken or formed, the analysis presented here makes clear that these phenomena are dependent on molecular structure.

5. Conclusions

Predictive models for surface tension, liquid density, flash point, distillation curve, derived cetane number, and kinematic viscosity have been developed to assess their applicability for the prescreening of sustainable aviation fuel (SAF) candidates. These models were constructed using multiple linear regression, binomial regression, random forest regression, and artificial neural network techniques, with atom-type mole fractions serving as predictors. Atom types were quantified using ^{1}H ^{13}C HSQC NMR spectroscopy, which requires millilitres of fuel samples.

Comprehensive datasets were assembled for each property: 1,241 data points for density, 1,260 for surface tension, 184 for flash point, 1,074 for distillation curve, 405 for derived cetane number, and 1,373 for kinematic viscosity. Atom-type compositions were determined for 19 conventional and synthetic aviation turbine fuels, and kinematic viscosity and density were measured for nine samples at −20, 0, 20, and 40 °C for model training and testing.

Each dataset was used to train and evaluate numerical models describing the relevant property in terms of atom-type composition. The surface tension model demonstrated excellent predictive performance, achieving an average error of 2.22 % for unseen pure components and mixtures, and predicting the surface tension of real aviation fuels within 4.41 %. The liquid density model displayed similarly strong predictive accuracy, with an average error of 0.47 % for unseen pure components and mixtures, and predicted the density of real fossil, synthetic aviation fuels, and gasolines to within 5.35 %.

The flash point model, based on binomial regression, showed high fidelity to the training and testing datasets, with a mean absolute percentage error of 0.81 %. However, due to its polynomial form, it is more susceptible to overfitting, and its performance could be further improved through the inclusion of additional complex mixture data. The distillation curve model, trained using a random forest regressor, demonstrated excellent agreement with both training and unseen datasets at 10 %, 50 %, and 90 % volume recovered, achieving $R^2 > 0.97$. Validation against two conventional jet fuels confirmed the robustness of this model.

The derived cetane number model performed less effectively overall ($R^2 < 0.95$), but despite its simplicity, it demonstrated strong fidelity when validated against real fuel data, indicating sound physical consistency. The kinematic viscosity model, developed using an artificial neural network and incorporating both experimental and calculated data, achieved strong performance for training and unseen datasets, with a mean average percentage error of 3.61 %.

Overall, the models developed in this study show promise for the prescreening of SAF candidates based on atom type composition. Further improvement in predictive performance can be achieved through the inclusion of more compositionally complex mixtures and broader coverage of the compositional space in the training datasets, which will require additional experimental measurements.

Future Recommendations

Preliminary applications of ^{1}H ^{13}C HSQC quantified atom types have been explored in this work through predictive models for fit-for-purpose fuel properties. Although complex hydrocarbon mixtures (aviation turbine fuels) were analysed using NMR spectroscopy, the predictive models were trained primarily on property databases consisting of pure components or simple mixtures. Future research would benefit from evaluating how model fidelity changes when trained on datasets composed of complex hydrocarbon mixtures. Such datasets may better capture intermolecular interactions that are not represented in pure-component systems.

More advanced machine learning methodologies could also be explored to better account for non-linear intermolecular effects, as these interactions are not explicitly represented in linear regression models. However, given that the intended purpose of the models is fuel prescreening, the trade-off between increased predictive accuracy and computational or methodological complexity should be carefully evaluated. Quantitative Property Structure Relationships could improve the fidelity of the more complex physical characteristics (flash point, distillation curve, kinetic viscosity), by allowing the models to learn slight molecular interactions and characteristics from other physical properties of a fuel. Though, the viability of this would need to be investigated as it increases the work and sample quantity needed for prescreening.

Beyond predictive modelling, atom type quantification demonstrates strong potential for sensitive chemical characterisation of fuels, particularly in areas where conventional analytical techniques may lack resolution. Further development and refinement of atom-type classification using higher-sensitivity NMR instrumentation would likely provide significant benefits to the field. Increased signal sensitivity, achieved through higher-field spectrometers, additional scans, or larger sample volumes, could improve peak detection and enable more detailed examination of low-ppm hydrocarbon class content.

Bibliography

(1) ASTM. *D4054: Standard Practice for Evaluation of New Aviation Turbine Fuels and Fuel Additives*; Conshohocken, PA: ASTM International, 2023. DOI: 10.1520/D4054-23.
(2) Heyne, J.; Rauch, B.; Le Clercq, P.; Colket, M. Sustainable aviation fuel prescreening tools and procedures. *Fuel* 2021, *290*, 120004. DOI: https://doi.org/10.1016/j.fuel.2020.120004.
(3) Petrucci, R. H. *General chemistry : principles and modern applications*; Pearson Canada Toronto, Ont., 2011.
(4) Wei, Y.-J.; Zhang, Y.-J.; Zhu, X.-D.; Gu, H.-M.; Zhu, Z.-Q.; Liu, S.-H.; Sun, X.-Y.; Jiang, X.-L. Effects of Diesel Hydrocarbon Components on Cetane Number and Engine Combustion and Emission Characteristics. In *Applied Sciences*, 2022; Vol. 12.
(5) Yang, Z.; Bell, D. C.; Cronin, D. J.; Boehm, R.; Heyne, J.; Ramasamy, K. K. Volatility Measurements of Sustainable Aviation Fuels: A Comparative Study of D86 and D2887 Methods. *ACS Sustainable Chemistry & Engineering* 2024, *12* (19), 7414-7420. DOI: 10.1021/acssuschemeng.4c00678.
(6) Amine, M.; Mohammed, H. A.; Barakat, Y. Volatility criteria and physicochemical properties of the promising dimethyl carbonate-gasoline blends. *Scientific Reports* 2022, *12* (1), 17183. DOI: 10.1038/s41598-022-21303-3.
(7) ASTM. *D56: Standard Test Method for Flash Point by Tag Closed Cup Tester*; 2022. DOI: 10.1520/D0056-21A.
(8) ASTM. *D7566: Standard Specification for Aviation Turbine Fuel Containing Synthesized Hydrocarbons*; 2022.
(9) Cooper, J. B.; Larkin, C. M.; Schmitigal, J.; Morris, R. E.; Abdelkader, M. F. Rapid Analysis of Jet Fuel Using a Handheld Near-Infrared (NIR) Analyzer. *Applied Spectroscopy* 2011, *65* (2), 187-192. DOI: 10.1366/10-06076 (accessed 2024/07/23).
(10) Heyne, J.; Bell, D.; Feldhausen, J.; Yang, Z.; Boehm, R. Towards fuel composition and properties from Two-dimensional gas chromatography with flame ionization and vacuum ultraviolet spectroscopy. *Fuel* 2022, *312*, Article. DOI: 10.1016/j.fuel.2021.122709 Scopus.
(11) Johnson, K. J.; Morris, R. E.; Rose-Pehrsson, S. L. Evaluating the Predictive Powers of Spectroscopy and Chromatography for Fuel Quality Assessment. *Energy & Fuels* 2006, *20* (2), 727-733. DOI: 10.1021/ef050347t.
(12) Kim, D.; Martz, J.; Abdul-Nour, A.; Yu, X.; Jansons, M.; Violi, A. A six-component surrogate for emulating the physical and chemical characteristics of conventional and alternative jet fuels and their blends. *Combustion and Flame* 2017, *179*, 86-94. DOI: https://doi.org/10.1016/j.combustflame.2017.01.025.
(13) Vozka, P.; Kilaz, G. A review of aviation turbine fuel chemical composition-property relations. *Fuel* 2020, *268*, 117391. DOI: https://doi.org/10.1016/j.fuel.2020.117391.
(14) Wang, Y.; Ding, Y.; Wei, W.; Cao, Y.; Davidson, D. F.; Hanson, R. K. On estimating physical and chemical properties of hydrocarbon fuels using mid-infrared FTIR spectra and regularized linear models. *Fuel* 2019, *255*, 115715. DOI: https://doi.org/10.1016/j.fuel.2019.115715.
(15) Yang, Z.; Kosir, S.; Stachler, R.; Shafer, L.; Anderson, C.; Heyne, J. S. A GC × GC Tier α combustor operability prescreening method for sustainable aviation fuel candidates. *Fuel* 2021, *292*, 120345. DOI: https://doi.org/10.1016/j.fuel.2021.120345.
(16) Ambrose, D. *Correlation and estimation of vapour-liquid critical properties : I. critical temperatures of organic compounds / by D. Ambrose*; Teddington : National Physical Laboratory, 1978.

(17) Constantinou, L.; Gani, R. New group contribution method for estimating properties of pure compounds. *AIChE Journal* 1994, *40* (10), 1697-1710. DOI: https://doi.org/10.1002/aic.690401011.
(18) Joback, K. G.; Reid, R. C. Estimation of Pure-Component Properties From Group-Contributions. *Chemical Engineering Communications* 1982. DOI: 10.1080/00986448708960487.
(19) Lydersen, A. L.; Station, E. E.; Greenkorn, R. A.; Hougen, O. A. *Estimation of Critical Properties of Organic Compounds by the Method of Group Contibutions*; University of Wisconsin, 1955.
(20) Carreón-Calderón, B.; Uribe-Vargas, V.; Ramírez-Jaramillo, E.; Ramírez-de-Santiago, M. Thermodynamic Characterization of Undefined Petroleum Fractions using Group Contribution Methods. *Industrial & Engineering Chemistry Research* 2012, *51* (43), 14188-14198. DOI: 10.1021/ie3016076.
(21) Mishra, S.; Bukkarapu, K. R.; Krishnasamy, A. A composition based approach to predict density, viscosity and surface tension of biodiesel fuels. *Fuel* 2021, *285*, 119056. DOI: https://doi.org/10.1016/j.fuel.2020.119056.
(22) Paduszyński, K. Extensive Databases and Group Contribution QSPRs of Ionic Liquids Properties. 1. Density. *Industrial & Engineering Chemistry Research* 2019, *58* (13), 5322-5338. DOI: 10.1021/acs.iecr.9b00130.
(23) Skander, N.; Chitour, C. E. A New Group-Contribution Method for the Estimation of Physical Properties of Hydrocarbons. *Oil & Gas Science and Technology - Revue d'IFP Energies nouvelles* 2002, *57* (4), 369-376. DOI: 10.2516/ogst:2002024 Ogst.
(24) Stefanis, E.; Constantinou, L.; Tsivintzelis, I.; Panayiotou, C. New Group-Contribution Method for Predicting Temperature-Dependent Properties of Pure Organic Compounds. *International Journal of Thermophysics* 2005, *26* (5), 1369-1388. DOI: 10.1007/s10765-005-8092-7.
(25) Wakeham, W. A.; Cholakov, G. S.; Stateva, R. P. Liquid Density and Critical Properties of Hydrocarbons Estimated from Molecular Structure. *Journal of Chemical & Engineering Data* 2002, *47* (3), 559-570. DOI: 10.1021/je010308l.
(26) Deng, J.; Zhang, Y.; Jia, G. Surface tension of binary and ternary mixtures mapping with ASP and UNIFAC models based on machine learning. *Physics of Fluids* 2023, *35* (6). DOI: 10.1063/5.0152893.
(27) Gharagheizi, F.; Ilani-Kashkouli, P.; Mohammadi, A. H. Group contribution model for estimation of surface tension of ionic liquids. *Chemical Engineering Science* 2012, *78*, 204-208. DOI: https://doi.org/10.1016/j.ces.2012.05.008.
(28) Mulero, A.; Cachadiña, I.; Cardona, L. F.; Valderrama, J. O. Pressure–Surface Tension–Temperature Equation of State for n-Alkanes. *Industrial & Engineering Chemistry Research* 2022, *61* (9), 3457-3473. DOI: 10.1021/acs.iecr.1c04979.
(29) Naef, R.; Acree, W. E. Calculation of the Surface Tension of Ordinary Organic and Ionic Liquids by Means of a Generally Applicable Computer Algorithm Based on the Group-Additivity Method. *Molecules* 2018, *23* (5), 1224.
(30) Randová, A.; Bartovská, L. Group contribution method: Surface tension of linear and branched alkanes. *Fluid Phase Equilibria* 2016, *429*, 166-176. DOI: https://doi.org/10.1016/j.fluid.2016.09.007.
(31) Shahsavari, S.; Mesbah, M.; Soroush, E.; Farhangian, H.; Alizadeh, S.; Soltanali, S. A simple group contribution correlation for modeling the surface tension of pure ionic liquids.

Journal of Molecular Liquids 2018, *265*, 292-298, Article. DOI: 10.1016/j.molliq.2018.06.004 Scopus.
(32) Shi, X.; Li, H.; Song, Z.; Zhang, X.; Liu, G. Quantitative composition-property relationship of aviation hydrocarbon fuel based on comprehensive two-dimensional gas chromatography with mass spectrometry and flame ionization detector. *Fuel* 2017, *200*, 395-406. DOI: https://doi.org/10.1016/j.fuel.2017.03.073.
(33) Shao, Y.; Yu, M.; Zhao, M.; Xue, K.; Zhang, X.; Zou, J.-J.; Pan, L. Comprehensive accurate prediction of critical jet fuel properties with multiple machine learning models. *Chemical Engineering Science* 2025, *304*, 121018. DOI: https://doi.org/10.1016/j.ces.2024.121018.
(34) Saldana, D. A.; Starck, L.; Mougin, P.; Rousseau, B.; Creton, B. Prediction of Flash Points for Fuel Mixtures Using Machine Learning and a Novel Equation. *Energy & Fuels* 2013, *27* (7), 3811-3820. DOI: 10.1021/ef4005362.
(35) Patel, S. J.; Ng, D.; Mannan, M. S. QSPR Flash Point Prediction of Solvents Using Topological Indices for Application in Computer Aided Molecular Design. *Industrial & Engineering Chemistry Research* 2009, *48* (15), 7378-7387. DOI: 10.1021/ie9000794.
(36) Pan, Y.; Jiang, J.; Wang, Z. Quantitative structure–property relationship studies for predicting flash points of alkanes using group bond contribution method with back-propagation neural network. *Journal of Hazardous Materials* 2007, *147* (1), 424-430. DOI: https://doi.org/10.1016/j.jhazmat.2007.01.025.
(37) Lee, C. J.; Ko, J. W.; Lee, G. Flash point prediction of organic compounds using a group contribution and support vector machine. *Korean Journal of Chemical Engineering* 2012, *29* (2), 145-153. DOI: 10.1007/s11814-011-0164-8.
(38) Gad, E.; Al-Fahemi, J.; Albis, N. QSPR Model for Predicting Flash Point of Some Organic Hydrocarbons. *Chemical Science Review and Letters* 2015, *2015*, 583-590.
(39) Dai, Y.-m.; Liu, H.; Chen, X.-q.; Liu, Y.-n.; Li, X.; Zhu, Z.-p.; Zhang, Y.-f.; Cao, Z. A new group contribution-based method for estimation of flash point temperature of alkanes. *Journal of Central South University* 2015, *22* (1), 30-36. DOI: 10.1007/s11771-015-2491-0.
(40) Boddapati, V.; Ferris, A. M.; Hanson, R. K. 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. DOI: https://doi.org/10.1016/j.fuel.2023.129557.
(41) Atabati, M.; Emamalizadeh, R. A Quantitative Structure Property Relationship for Prediction of Flash Point of Alkanes Using Molecular Connectivity Indices. *Chinese Journal of Chemical Engineering* 2013, *21* (4), 420-426. DOI: https://doi.org/10.1016/S1004-9541(13)60483-8.
(42) Albahri, T. A.; Esmael, N. A. M. SGC based prediction of the flash point temperature of pure compounds. *Journal of Loss Prevention in the Process Industries* 2018, *54*, 303-311. DOI: https://doi.org/10.1016/j.jlp.2018.05.005.
(43) Albahri, T. A. MNLR and ANN structural group contribution methods for predicting the flash point temperature of pure compounds in the transportation fuels range. *Process Safety and Environmental Protection* 2015, *93*, 182-191. DOI: https://doi.org/10.1016/j.psep.2014.03.005.
(44) Flora, G.; Karimzadeh, F.; Kahandawala, M. S. P.; DeWitt, M. J.; Corporan, E. Prediction of hydrocarbons ignition performances using machine learning modeling. *Fuel* 2024, *368*, 131619. DOI: https://doi.org/10.1016/j.fuel.2024.131619.

(45) Dalmiya, A.; Sheyyab, M.; Mehta, J. M.; Brezinsky, K.; Lynch, P. T. Derived cetane number prediction of jet fuels and their functional group surrogates using liquid phase infrared absorption. *Proceedings of the Combustion Institute* 2023, *39* (1), 1495-1504. DOI: https://doi.org/10.1016/j.proci.2022.08.104.
(46) Boddapati, V.; Ferris, A. M.; Hanson, R. K. On the use of extended-wavelength FTIR spectra for the prediction of combustion properties of jet fuels and their constituent species. *Proceedings of the Combustion Institute* 2023, *39* (1), 1347-1355. DOI: https://doi.org/10.1016/j.proci.2022.08.041.
(47) Al Ibrahim, E.; Farooq, A. Prediction of the Derived Cetane Number and Carbon/Hydrogen Ratio from Infrared Spectroscopic Data. *Energy & Fuels* 2021, *35* (9), 8141-8152. DOI: 10.1021/acs.energyfuels.0c03899.
(48) Li, R.; Herreros, J. M.; Tsolakis, A.; Yang, W. Machine learning regression based group contribution method for cetane and octane numbers prediction of pure fuel compounds and mixtures. *Fuel* 2020, *280*, 118589. DOI: https://doi.org/10.1016/j.fuel.2020.118589.
(49) Yarranton, H. W.; Satyro, M. A. Expanded Fluid-Based Viscosity Correlation for Hydrocarbons. *Industrial & Engineering Chemistry Research* 2009, *48* (7), 3640-3648. DOI: 10.1021/ie801698h.
(50) van Velzen, D.; Cardozo, R. L.; Langenkamp, H. A Liquid Viscosity-Temperature-Chemical Constitution Relation for Organic Compounds. *Industrial & Engineering Chemistry Fundamentals* 1972, *11* (1), 20-25. DOI: 10.1021/i160041a004.
(51) Watson-Murphy, T.; Ure, A. D.; Ghaani, M. R.; Dooley, S. Machine Learning and HSQC-NMR for viscosity prediction of sustainable aviation fuel. In *American Chemical Society*, Indianapolis, United States of America,, 2023.
(52) Saldana, D. A.; Starck, L.; Mougin, P.; Rousseau, B.; Ferrando, N.; Creton, B. Prediction of Density and Viscosity of Biofuel Compounds Using Machine Learning Methods. *Energy & Fuels* 2012, *26* (4), 2416-2426. DOI: 10.1021/ef3001339.
(53) Panwar, P.; Yang, Q.; Martini, A. Temperature-Dependent Density and Viscosity Prediction for Hydrocarbons: Machine Learning and Molecular Dynamics Simulations. *Journal of Chemical Information and Modeling* 2024, *64* (7), 2760-2774. DOI: 10.1021/acs.jcim.3c00231.
(54) Mickoleit, E.; Jäger, A.; Grau Turuelo, C.; Thol, M.; Bell, I. H.; Breitkopf, C. Group Contribution Method for the Residual Entropy Scaling Model for Viscosities of Branched Alkanes. *International Journal of Thermophysics* 2023, *44* (12), 176. DOI: 10.1007/s10765-023-03289-w.
(55) Kotzakoulakis, K.; George, S. C. A simple and flexible correlation for predicting the viscosity of crude oils. *Journal of Petroleum Science and Engineering* 2017, *158*, 416-423. DOI: https://doi.org/10.1016/j.petrol.2017.08.058.
(56) Klein, T.; Yan, S.; Cui, J.; Magee, J. W.; Kroenlein, K.; Rausch, M. H.; Koller, T. M.; Fröba, A. P. Liquid Viscosity and Surface Tension of n-Hexane, n-Octane, n-Decane, and n-Hexadecane up to 573 K by Surface Light Scattering. *Journal of Chemical & Engineering Data* 2019, *64* (9), 4116-4131. DOI: 10.1021/acs.jced.9b00525.
(57) Jovanović, J. D.; Grozdanić, N. D.; Radović, I. R.; Lj. Kijevčanin, M. A new group contribution model for prediction liquid hydrocarbon viscosity based on free-volume theory. *Journal of Molecular Liquids* 2023, *376*, 121452. DOI: https://doi.org/10.1016/j.molliq.2023.121452.

(58) Hall, C.; Bell, D. C.; Feldhausen, J.; Rauch, B.; Heyne, J. Quantifying isomeric effects: A key factor in aviation fuel assessment and design. *Fuel* 2024, *357*, 129912. DOI: https://doi.org/10.1016/j.fuel.2023.129912.
(59) Feldhausen, J.; Bell, D. C.; Yang, Z.; Faulhaber, C.; Boehm, R.; Heyne, J. Synthetic aromatic kerosene property prediction improvements with isomer specific characterization via GCxGC and vacuum ultraviolet spectroscopy. *Fuel* 2022, *326*, 125002. DOI: https://doi.org/10.1016/j.fuel.2022.125002.
(60) Das, S.; Roy, R. K.; Bezboruah, T. A computational approach for prediction of viscosity of chemical compounds based on molecular structures. *Results in Chemistry* 2025, *13*, 102039. DOI: https://doi.org/10.1016/j.rechem.2025.102039.
(61) Caponi, M.; Cox, A.; Misra, S. Viscosity prediction using image processing and supervised learning. *Fuel* 2023, *339*, 127320. DOI: https://doi.org/10.1016/j.fuel.2022.127320.
(62) Cai, G.; Liu, Z.; Zhang, L.; Zhao, S.; Xu, C. Quantitative Structure–Property Relationship Model for Hydrocarbon Liquid Viscosity Prediction. *Energy & Fuels* 2018, *32* (3), 3290-3298. DOI: 10.1021/acs.energyfuels.7b04075.
(63) Caceres-Martinez, L. E.; Kilaz, G. Kinematic viscosity prediction of jet fuels and alternative blending components via comprehensive two-dimensional gas chromatography, partial least squares, and Yeo-Johnson transformation. *Journal of Separation Science* 2024, *47* (5), 2300816. DOI: https://doi.org/10.1002/jssc.202300816 (accessed 2025/08/25).
(64) Aralaguppi, M. I.; Jadar, C. V.; Aminabhavi, T. M. Density, Refractive Index, Viscosity, and Speed of Sound in Binary Mixtures of Cyclohexanone with Hexane, Heptane, Octane, Nonane, Decane, Dodecane, and 2,2,4-Trimethylpentane. *Journal of Chemical & Engineering Data* 1999, *44* (3), 435-440. DOI: 10.1021/je9802266.
(65) Lötgering-Lin, O.; Gross, J. Group Contribution Method for Viscosities Based on Entropy Scaling Using the Perturbed-Chain Polar Statistical Associating Fluid Theory. *Industrial & Engineering Chemistry Research* 2015, *54* (32), 7942-7952. DOI: 10.1021/acs.iecr.5b01698.
(66) Pasadakis, N.; Sourligas, S.; Foteinopoulos, C. Prediction of the distillation profile and cold properties of diesel fuels using mid-IR spectroscopy and neural networks. *Fuel* 2006, *85* (7), 1131-1137. DOI: https://doi.org/10.1016/j.fuel.2005.09.016.
(67) Mohan, B.; Alsaleh, F.; AlRamadan, A. S.; Almbdal, S.; Al-Ammari, A.; Saihati, Z.; Voice, A.; Badra, J. In silico estimation of density, vapor pressure, distillation curve, and octane numbers of gasoline fuels using machine learning. *Fuel* 2025, *399*, 135640. DOI: https://doi.org/10.1016/j.fuel.2025.135640.
(68) Karonis, D.; Zahos-Siagos, I.; Filimon, D.; Vasileiou, F. A multivariate statistical analysis to evaluate and predict ignition quality of marine diesel fuel distillates from their physical properties. *Fuel Processing Technology* 2017, *166*, 299-311. DOI: https://doi.org/10.1016/j.fuproc.2017.06.021.
(69) Hosseinifar, P.; Shahverdi, H. A predictive method for constructing the distillation curve of petroleum fluids using their physical bulk properties. *Journal of Petroleum Science and Engineering* 2021, *200*, 108403. DOI: https://doi.org/10.1016/j.petrol.2021.108403.
(70) Govindaraju, P. B.; Ihme, M. Group contribution method for multicomponent evaporation with application to transportation fuels. *International Journal of Heat and Mass Transfer* 2016, *102*, 833-845. DOI: https://doi.org/10.1016/j.ijheatmasstransfer.2016.06.079.
(71) Geber, E.; Rodriguez, C.; Karathanassis, I. K.; Lopez-Pintor, D.; Manin, J.; Pickett, L.; Gavaises, M. A general predictive methodology for fuel-mixture properties up to supercritical

conditions. *Fluid Phase Equilibria* 2023, *574*, 113888. DOI: https://doi.org/10.1016/j.fluid.2023.113888.
(72) Bell, D.; Heyne, J. S.; Won, S. H.; Dryer, F.; Haas, F. M.; Dooley, S. On the Development of General Surrogate Composition Calculations for Chemical and Physical Properties. In *55th AIAA Aerospace Sciences Meeting*, AIAA SciTech Forum, American Institute of Aeronautics and Astronautics, 2017.
(73) Chen, Y.; Zheng, Z.; Lu, Z.; Wang, H.; Wang, C.; Sun, X.; Xu, L.; Yao, M. Machine learning-based screening of fuel properties for SI and CI engines using a hybrid group extraction method. *Applied Energy* 2024, *366*, 123257. DOI: https://doi.org/10.1016/j.apenergy.2024.123257.
(74) Marinović, S.; Krištović, M.; Špehar, B.; Rukavina, V.; Jukić, A. Prediction of diesel fuel properties by vibrational spectroscopy using multivariate analysis. *Journal of Analytical Chemistry* 2012, *67*. DOI: 10.1134/S1061934812120039.
(75) Dobbelaere, M. R.; Ureel, Y.; Vermeire, F. H.; Tomme, L.; Stevens, C. V.; Van Geem, K. M. Machine Learning for Physicochemical Property Prediction of Complex Hydrocarbon Mixtures. *Industrial & Engineering Chemistry Research* 2022, *61* (24), 8581-8594. DOI: 10.1021/acs.iecr.2c00442.
(76) Bolanča, T.; Marinović, S.; Ukić, S.; Jukić, A.; Rukavina, V. Development of Artificial Neural Network Model for Diesel Fuel Properties Prediction using Vibrational Spectroscopy. *Acta Chim Slov* 2012, *59* (2), 249-257. From NLM.
(77) Abdul Jameel, A. G.; Naser, N.; Emwas, A.-H.; Dooley, S.; Sarathy, S. M. Predicting Fuel Ignition Quality Using 1H NMR Spectroscopy and Multiple Linear Regression. *Energy & Fuels* 2016, *30* (11), 9819-9835. DOI: 10.1021/acs.energyfuels.6b01690.
(78) Dussan, K.; Won, S. H.; Ure, A. D.; Dryer, F. L.; Dooley, S. Chemical functional group descriptor for ignition propensity of large hydrocarbon liquid fuels. *Proceedings of the Combustion Institute* 2019, *37* (4), 5083-5093. DOI: https://doi.org/10.1016/j.proci.2018.05.079.
(79) Dooley, S.; Won, S. H.; Dryer, F. L. Chapter 10 - Surrogate fuels and combustion characteristics of liquid transportation fuels. In *Computer Aided Chemical Engineering*, Faravelli, T., Manenti, F., Ranzi, E. Eds.; Vol. 45; Elsevier, 2019; pp 513-602.
(80) Ure, A. D.; O'Brien, J. E.; Dooley, S. Quantitative NMR Spectroscopy for the Analysis of Fuels: A Case Study of FACE Gasoline F. *Energy and Fuels* 2019, Article. DOI: 10.1021/acs.energyfuels.9b01019 Scopus.
(81) Won, S. H.; Lim, S. J.; Nates, S.; Alwahaibi, A. K.; Dryer, F. L.; Farid, F.; Hase, M. Combustion characteristics of crude oils for gas turbine applications by DCN measurements and NMR spectroscopy. *Proceedings of the Combustion Institute* 2021, *38* (4), 5463-5473. DOI: https://doi.org/10.1016/j.proci.2020.06.184.
(82) Ure, A. D.; O'Brien, J. E.; Dooley, S. Quantitative NMR Spectroscopy for the Analysis of Fuels: A Case Study of FACE Gasoline F. *Energy & Fuels* 2019, *33* (11), 11741-11756. DOI: 10.1021/acs.energyfuels.9b01019.
(83) Sarpal, A. S.; Kapur, G. S.; Mukherjee, S.; Tiwari, A. K. PONA analyses of cracked gasoline by 1H NMR spectroscopy. Part II. *Fuel* 2001, *80* (4), 521-528. DOI: https://doi.org/10.1016/S0016-2361(00)00123-X.
(84) Sanchez-Minero, F.; Ancheyta, J.; Silva-Oliver, G.; Flores-Valle, S. Predicting SARA composition of crude oil by means of NMR. *Fuel* 2013, *110*, 318-321. DOI: https://doi.org/10.1016/j.fuel.2012.10.027.

(85) Robinson, A.; Lim, S. J.; Alwahaibi, A. K.; Zdanowicz, A.; Török, D.; Windom, B.; Dryer, F. L.; Won, S. H. Elucidating NO coupling effects on ignition of toluene reference fuels by chemical functional group analysis. *Proceedings of the Combustion Institute* 2023, *39* (4), 4919-4928. DOI: https://doi.org/10.1016/j.proci.2022.08.100.
(86) Rakhmatullin, I.; Efimov, S.; Varfolomeev, M.; Klochkov, V. High-resolution NMR study of light and heavy crude oils: "structure-property" analysis. *IOP Conference Series: Earth and Environmental Science* 2018, *155* (1), 012014. DOI: 10.1088/1755-1315/155/1/012014.
(87) Mukherjee, S.; Kapur, G. S.; Chopra, A.; Sarpal, A. S. 1H NMR Spectroscopic-Based Method for the Estimation of Total Aromatic Content of Aviation Turbine Fuels (ATF): Comparison with Liquid Chromatographic Methods. *Energy & Fuels* 2004, *18* (1), 30-36. DOI: 10.1021/ef020275v.
(88) Mondal, S.; Chattopadhayay, K.; Kumar, R.; Christopher, J.; Kapur, G. S. A Rapid 1H NMR-Based Estimation of PONA for Light and Narrow Cut Naphtha Samples of Refinery Streams toward BS-VI Gasoline. *Energy & Fuels* 2021, *35* (9), 7883-7892. DOI: 10.1021/acs.energyfuels.1c00413.
(89) Mondal, S.; Chattopadhayay, K.; Srivastav, B.; Garg, K.; Kumar, R.; Chopra, A.; Christopher, J.; Kapur, G. S. Dynamic Approach for the Estimation of Olefins in Cracked Fuel Range Products of Variable Nature and Composition by 1H NMR Spectroscopy. *Energy & Fuels* 2019, *33* (2), 1114-1122. DOI: 10.1021/acs.energyfuels.8b04392.
(90) Mondal, S.; Yadav, A.; Kumar, R.; Bansal, V.; Das, S. K.; Christopher, J.; Kapur, G. S. Molecular-Level Structural Insight into Clarified Oil by Nuclear Magnetic Resonance (NMR) Spectroscopy: Estimation of Hydrocarbon Types and Average Structural Parameters. *Energy & Fuels* 2017, *31* (7), 7682-7692. DOI: 10.1021/acs.energyfuels.7b00994.
(91) Mondal, S.; Kumar, R.; Bansal, V.; Patel, M. B. A 1H NMR method for the estimation of hydrogen content for all petroleum products. *Journal of Analytical Science and Technology* 2015, *6* (1), 24. DOI: 10.1186/s40543-015-0064-3.
(92) Molina V, D.; Uribe, U. N.; Murgich, J. Correlations between SARA fractions and physicochemical properties with 1H NMR spectra of vacuum residues from Colombian crude oils. *Fuel* 2010, *89* (1), 185-192. DOI: https://doi.org/10.1016/j.fuel.2009.07.021.
(93) Lee, S. W.; Glavincevski, B. NMR method for determination of aromatics in middle distillate oils. *Fuel Processing Technology* 1999, *60* (1), 81-86. DOI: https://doi.org/10.1016/S0378-3820(99)00013-2.
(94) Kapur, G. S.; Chopra, A.; Sarpal, A. S. Estimation of Total Aromatic Content of Vacuum Gas Oil (VGO) Fractions (370−560 °C) by 1H NMR Spectroscopy. *Energy & Fuels* 2005, *19* (3), 1065-1071. DOI: 10.1021/ef040069i.
(95) Kapur, G. S.; Ecker, A.; Meusinger, R. Establishing Quantitative Structure−Property Relationships (QSPR) of Diesel Samples by Proton-NMR & Multiple Linear Regression (MLR) Analysis. *Energy & Fuels* 2001, *15* (4), 943-948. DOI: 10.1021/ef010021u.
(96) Kapur, G. S.; Singh, A. P.; Sarpal, A. S. Determination of aromatics and naphthenes in straight run gasoline by 1H NMR spectroscopy. Part I. *Fuel* 2000, *79* (9), 1023-1029. DOI: https://doi.org/10.1016/S0016-2361(99)00238-0.
(97) Kaiser, C. R.; Borges, J. L.; dos Santos, A. R.; Azevedo, D. A.; D′ Avila, L. A. Quality control of gasoline by 1H NMR: Aromatics, olefinics, paraffinics, and oxygenated and benzene contents. *Fuel* 2010, *89* (1), 99-104. DOI: https://doi.org/10.1016/j.fuel.2009.06.023.

(98) Hertzog, J.; Garnier, C.; Mase, C.; Mariette, S.; Serve, O.; Hubert-Roux, M.; Afonso, C.; Giusti, P.; Barrère-Mangote, C. Fractionation by flash chromatography and molecular characterization of bio-oil by ultra-high-resolution mass spectrometry and NMR spectroscopy. *Journal of Analytical and Applied Pyrolysis* 2022, *166*, 105611. DOI: https://doi.org/10.1016/j.jaap.2022.105611.
(99) Heredia-Langner, A.; Cort, J. R.; Grubel, K.; O'Hagan, M. J.; Jarman, K. H.; Linehan, J. C.; Albrecht, K. O.; Polikarpov, E.; King, D. L.; Smurthwaite, T. D.; et al. Methodology for the Development of Empirical Models Relating 13C NMR Spectral Features to Fuel Properties. *Energy & Fuels* 2020, *34* (10), 12556-12572. DOI: 10.1021/acs.energyfuels.0c00883.
(100) Gulder, O. L.; Glavincevski, B. Prediction of cetane number of diesel fuels from carbon type structural composition determined by proton NMR spectroscopy. *Industrial & Engineering Chemistry Product Research and Development* 1986, *25* (2), 153-156. DOI: 10.1021/i300022a005.
(101) Cookson, D. J.; Smith, B. E. Calculation of jet and diesel fuel properties using carbon-13 NMR spectroscopy. *Energy & Fuels* 1990, *4* (2), 152-156. DOI: 10.1021/ef00020a004.
(102) Cookson, D. J.; Smith, B. E. One and two-dimensional NMR methods for elucidating structural characteristics of aromatic fractions from petroleum and synthetic fuels. *Energy & Fuels* 1987, *1* (1), 111-120. DOI: 10.1021/ef00001a021.
(103) Burri, J.; Crockett, R.; Hany, R.; Rentsch, D. Gasoline composition determined by 1H NMR spectroscopy. *Fuel* 2004, *83* (2), 187-193. DOI: https://doi.org/10.1016/S0016-2361(03)00261-8.
(104) Burger, J. L.; Widegren, J. A.; Lovestead, T. M.; Bruno, T. J. 1H and 13C NMR Analysis of Gas Turbine Fuels As Applied to the Advanced Distillation Curve Method. *Energy & Fuels* 2015, *29* (8), 4874-4885. DOI: 10.1021/acs.energyfuels.5b01035.
(105) Basu, B.; Kapur, G. S.; Sarpal, A. S.; Meusinger, R. A Neural Network Approach to the Prediction of Cetane Number of Diesel Fuels Using Nuclear Magnetic Resonance (NMR) Spectroscopy. *Energy & Fuels* 2003, *17* (6), 1570-1575. DOI: 10.1021/ef030083f.
(106) Bansal, V.; Krishna, G. J.; Chopra, A.; Sarpal, A. S. Detailed Hydrocarbon Characterization of RFCC Feed Stocks by NMR Spectroscopic Techniques. *Energy & Fuels* 2007, *21* (2), 1024-1029. DOI: 10.1021/ef060268x.
(107) Bansal, V.; Kapur, G. S.; Sarpal, A. S.; Kagdiyal, V.; Jain, S. K.; Srivastava, S. P.; Bhatnagar, A. K. Estimation of Total Aromatics and Their Distribution as Mono and Global Di-Plus Aromatics in Diesel-Range Products by NMR Spectroscopy. *Energy & Fuels* 1998, *12* (6), 1223-1227. DOI: 10.1021/ef980052y.
(108) Abdul Jameel, A. G. Identification and Quantification of Hydrocarbon Functional Groups in Gasoline Using 1H-NMR Spectroscopy for Property Prediction. In *Molecules*, 2021; Vol. 26.
(109) Abdul Jameel, A. G.; Naser, N.; Issayev, G.; Touitou, J.; Ghosh, M. K.; Emwas, A.-H.; Farooq, A.; Dooley, S.; Sarathy, S. M. A minimalist functional group (MFG) approach for surrogate fuel formulation. *Combustion and Flame* 2018, *192*, 250-271. DOI: https://doi.org/10.1016/j.combustflame.2018.01.036.
(110) Determination of saturates, aromatics and ethanol in Brazilian commercial gasoline with low olefin content using 1 H NMR spectroscopy. 2010.
(111) Bansal, V.; Krishna, G. J.; Singh, A. P.; Gupta, A. K.; Sarpal, A. S. Determination of Hydrocarbons Types and Oxygenates in Motor Gasoline: A Comparative Study by Different Analytical Techniques. *Energy & Fuels* 2008, *22* (1), 410-415. DOI: 10.1021/ef0701211.

(112) Daly, S. R.; Niemeyer, K. E.; Cannella, W. J.; Hagen, C. L. FACE Gasoline Surrogates Formulated by an Enhanced Multivariate Optimization Framework. *Energy & Fuels* 2018, *32* (7), 7916-7932. DOI: 10.1021/acs.energyfuels.8b01313.
(113) Bell, D. C.; Boehm, R.; Heyne, J. S. Freezing Point of Hydrocarbon Fuels from Single Species Concentrations. *Energy & Fuels* 2025, *39* (9), 4221-4226. DOI: 10.1021/acs.energyfuels.4c06091.
(114) Boehm, R. C.; Yang, Z.; Bell, D. C.; Faulhaber, C.; Mayhew, E.; Bauder, U.; Eckel, G.; Heyne, J. S. Perspective on Fuel Property Blending Rules for Design and Qualification of Aviation Fuels: A Review. *Energy & Fuels* 2024, *38* (18), 17128-17145. DOI: 10.1021/acs.energyfuels.4c02457.
(115) Edwards, J. T. Reference Jet Fuels for Combustion Testing. In *55th AIAA Aerospace Sciences Meeting*.
(116) Colket, M.; Heyne, J.; Rumizen, M.; Gupta, M.; Edwards, T.; Roquemore, W. M.; Andac, G.; Boehm, R.; Lovett, J.; Williams, R.; et al. Overview of the National Jet Fuels Combustion Program. *AIAA Journal* 2017, *55* (4), 1087-1104. DOI: 10.2514/1.J055361 (accessed 2025/09/29).
(117) Dooley, S.; Won, S. H.; Chaos, M.; Heyne, J.; Ju, Y.; Dryer, F. L.; Kumar, K.; Sung, C.-J.; Wang, H.; Oehlschlaeger, M. A.; et al. A jet fuel surrogate formulated by real fuel properties. *Combustion and Flame* 2010, *157* (12), 2333-2339. DOI: https://doi.org/10.1016/j.combustflame.2010.07.001.
(118) Corporan, E.; Edwards, T.; Shafer, L.; DeWitt, M. J.; Klingshirn, C.; Zabarnick, S.; West, Z.; Striebich, R.; Graham, J.; Klein, J. Chemical, Thermal Stability, Seal Swell, and Emissions Studies of Alternative Jet Fuels. *Energy & Fuels* 2011, *25* (3), 955-966. DOI: 10.1021/ef101520v.
(119) Authority, F. A. National Alternative Jet Fuel Database.
(120) Design Institute for Physical, P. Design Institute for Physical Properties (DIPPR) Database. 2022.
(121) Yaws, C. L. Yaws' Handbook of Physical Properties for Hydrocarbons and Chemicals. Knovel.
(122) Sheyyab, M.; Lynch, P. T.; Mayhew, E. K.; Brezinsky, K. Optimized synthetic data and semi-supervised learning for Derived Cetane Number prediction. *Combustion and Flame* 2024, *259*, 113184. DOI: https://doi.org/10.1016/j.combustflame.2023.113184.
(123) Naser, N.; Sarathy, S. M.; Chung, S. H. Ignition delay time sensitivity in ignition quality tester (IQT) and its relation to octane sensitivity. *Fuel* 2018, *233*, 412-419. DOI: https://doi.org/10.1016/j.fuel.2018.05.131.
(124) Fabian Pedregosa; Gael Varoquaux; Alexandre Gramfort; Vincent Michel; Bertrand Thirion; Olivier Grisel; Mathieu Blondel; Peter Prettenhofer; Ron Weiss; Vincent Dubourg; et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 2011, *12*.
(125) *H2O: Open Source Machine Learning for Python*; H2O.ai: 2024. https://docs.h2o.ai/h2o/latest-stable/h2o-docs/index.html
(126) Parker, R. P.; Kelly, M.; Watson-Murphy, T.; Ghaani, M. R.; Dooley, S. Predictive Modeling of Liquid Density and Surface Tension for Sustainable Aviation Fuels Using Nuclear Magnetic Resonance Atom Types. *Energy & Fuels* 2025, *39* (7), 3690-3702. DOI: 10.1021/acs.energyfuels.4c05601.

Appendix 1. Flash Point Model Coefficients and P-values

Table A.1. Flash Point Coefficients for Binomial Model showing the possible combinations of the atom types used in the refined dataset. All red p-values indicate that they are not statistically significant, reducing the amount of coefficients to 21.

Atom Type	Coefficient	p-value
pCH3 pCH2	-53.96	0.000
pCH3 pCH	-57.54	0.000
pCH3 pC	-58.92	0.000
pCH^2	50.29	0.000
pC^2	55.13	0.000
pCH2	37.05	0.000
cyCH^2	-30.40	0.000
pCH2 pCH	29.48	0.000
ArC	-28.92	0.000
pCH2 pC	28.83	0.000
pCH pC	25.42	0.000
cyCH	21.18	0.000
pC	-18.64	0.000
ArCH	-17.53	0.000
ArCH ArC	16.45	0.001
ArC^2	16.12	0.001
cyCH2^2	-15.30	0.002
cyCH2	13.94	0.005
pCH3^2	13.59	0.006
pCH2^2	11.52	0.020
pCH3	-10.29	0.037
pCH2 cyCH	-9.63	0.05
pCH2 cyCH2	8.18	0.10
pCH cyCH2	7.82	0.11
cyCH2 cyCH	7.75	0.11
pCH2 AlCH2	7.38	0.13
pCH	-7.20	0.14
ArCH^2	6.30	0.20

Atom Type	Coefficient	p-value
ArCH rC	6.04	0.22
pCH cyCH	-5.36	0.27
AlCH2 rC	-5.11	0.30
ArC AlCH2	4.23	0.39
rC	3.98	0.42
AlCH2^2	2.90	0.55
pCH3 ArCH	-2.58	0.60
pCH2 ArCH	2.34	0.63
pCH ArC	2.33	0.63
pCH ArCH	2.33	0.63
AlCH2	2.19	0.65
rC^2	1.88	0.70
pCH AlCH2	1.69	0.73
pC ArCH	1.57	0.75
pC ArC	1.57	0.75
pC cyCH2	1.36	0.78
pC cyCH	1.36	0.78
ArC rC	1.34	0.78
pCH3 AlCH2	0.63	0.90
pCH3 ArC	-0.56	0.91
ArCH AlCH2	0.51	0.92
pCH3 cyCH2	0.45	0.93
cyCH ArC	-0.35	0.94
cyCH ArCH	-0.35	0.94
cyCH2 ArCH	-0.35	0.94
cyCH2 ArC	-0.35	0.94
pCH2 ArC	-0.04	0.99
pCH3 cyCH	-0.01	1.00