Advancing Sustainable Aviation Fuel Design: Machine Learning for High Energy Density Liquid Polycyclic Hydrocarbons

*Dilip Rijal[1], Vladislav Vasilyev[2] and Feng Wang[1]**

[1]Department of Chemistry and Biotechnology, School of Science, Computing and Engineering Technologies, Swinburne University of Technology, Melbourne, Victoria 3122, Australia

[2]National Computational Infrastructure, Australian National University, Canberra, ACT 0200, Australia

*Corresponding author: fwang@swin.edu.au

Abstract

Sustainable aviation fuels (SAFs) are crucial for addressing carbon emissions in the aviation industry. With a focus on SAFs, the research aims to establish a quantitative structure-property relationship for polycyclic hydrocarbons (PCHCs) and their net heat of combustion (NHOC) using the innovative approach of machine learning (ML). The model trained with support vector machine (SVM) algorithms in ML is selected as it demonstrates superior performance over other available algorithms with a high coefficient of determination (R^2) and low mean absolute error (MAE) of 27.821 KJ/mol for 20% test data. Using the optimum SVM model, thirty-five potential PCHCs are identified as SAF candidates from C6 to C15 sourced from reputable scientific literature and databases. Furthermore, structural analysis revealed that high-performance PCHCs typically consist of saturated alkanes with multiple 3, 4, and 5-

https://doi.org/10.26434/chemrxiv-2024-lsr77 **ORCID:** https://orcid.org/0000-0002-6584-0516

membered rings, suggesting that strained energy plays a role in their high energy density. The model obtained from ML can be employed to screen new hydrocarbons for their suitability as SAF candidates before costly experiments and ASTM evaluations.

Keywords: Sustainable aviation fuel; Machine learning; Polycyclic hydrocarbons; Net heat of combustion; Fuel efficiency

---

Nomenclature

ICAO	International Civil Aviation Organization
ML	Machine Learning
LTO	Landing and Take-Off
ρ	Density
SAFs	Sustainable Aviation Fuels
GUI	Graphical User Interface
PtL	Power-to-Liquid
HED	High Energy Density
PCHCs	Polycyclic Hydrocarbons
SVM	Support Vector Machines
NHOC	Net Heat of Combustion
KNN	K-Nearest Neighbor
$NHOC_G$	Gravimetric Net Heat of Combustion
RF	Random Forest
$NHOC_V$	Volumetric Net Heat of Combustion
DFT	Density Functional Theory
QSPR	Quantitative Structure-Property Relationship
AF	Aviation Fuel
R^2	Coefficient of Determination
MAE	Mean Absolute Error
RMSE	Root Mean Square Error
GC	Group Contribution

1. Introduction

The aviation industry is flourishing with the steady growth of air travel worldwide, driven by factors such as increasing global connectivity, limited time, and expanding tourism industries, all contributing to heightened demand for fuel to power commercial aircraft.

According to an International Civil Aviation Organization (ICAO) report, it is predicted that aviation fuel (AF) consumption will increase by 1.9 to 2.6 times the value of 2018 in 2050 [1]. Ultimately, heavy reliance on conventional AF such as Jet-A, Jet-A1, JP-4, and JP-5 poses a significant challenge in the fight against greenhouse gas (GHG) emissions. For instance, as per the forecast provided by ICAO, emissions from international aviation during both full-flight operations and landing and take-off (LTO) are anticipated to rise between 2 to 4 times by 2050 compared to the level observed in 2018 [1], highlighting the urgent need for an alternative source of AF. Despite the pressing need for alternative, cleaner energy, options like batteries and hydrogen still need to be viable for immediate implementation in aviation due to technical and infrastructure limitations [2].

A promising progress in developing alternative source sustainable AF (SAFs) derived from Power-to-Liquid (PtL) pathways. In the PtL process, H_2 is typically obtained through water electrolysis. The electricity used for electrolysis is often sourced from renewable energy systems such as solar, wind, or hydropower. As for the carbon source, CO_2 is commonly sourced from various industrial processes, such as power plants, cement production, and steel manufacturing, where it is emitted as a byproduct, as well as directly from the atmosphere. The PtL process can help mitigate GHG emissions by converting CO_2 into valuable liquid fuels, thereby contributing to carbon neutrality or even carbon negativity [3]. It was revealed that the system's electrical efficiency is higher when the solid oxide electrolyzer operates in co-electrolysis mode compared to the steam mode for many hydrocarbon-based fuel production systems [3]. The PtL pathways produce synthetic fuels known as eFuel, which offer a potential solution to decarbonize the aviation industry.

Aviation fuel is a complex mixture of hydrocarbons like alkanes, cycloalkanes, and aromatics, further complicating the transition to greener alternatives. To enhance the efficiency of eFuel and to increase the current 50% blending constraint of SAFs [4], there is a crucial need to optimize polycyclic hydrocarbons (PCHCs), the building blocks of these synthetic eFuels. Focusing on the composition and characteristics of PCHCs, one aims to develop more efficient and environmentally friendly alternatives to traditional AF, paving the way for a greener future in air travel [2, 5, 6]. eFuel may play a pivotal role in the aviation industry's imperative to reduce GHG emissions, offering an environmentally responsible alternative that holds the key to achieving long-term carbon-neutral growth and realizing net-zero targets.

Aviation fuel is subject to strict specifications. Evaluation of the physicochemical properties of each component in the AF mixture and the blended AF is a complex and expensive but essential process. First, fuels often comprise diverse components with unique chemical properties, requiring a comprehensive analysis of the mixture (blend) and individual components. Next, the physicochemical properties encompass a wide array of characteristics, including net heat of combustion (NHOC), density at 15°C, viscosity at -20°C and -40°C, flash point, surface tension at 22°C, cetane number, and octane number [4, 7]. Evaluating these properties for individual components in the blended fuel adds complexity. In addition, the compounds in a blended fuel may interact, leading to synergistic or antagonistic effects on the properties of the fuel. Third, aviation industries have stringent standards and specifications for fuel quality and performance. To meet these standards, a blended AF requires thoroughly evaluating its physicochemical properties. Finally, obtaining precise and accurate measurements and analysis of the fuel properties needs advanced testing equipment and methodologies to ensure the reliability of results. All contribute to the overall complexity and cost of the process.

The reliability and availability of AF depend on the quantity and composition of the components in the fuel. SAFs can be achieved by reducing or replacing unwanted components such as aromatics in conventional AF with strained PCHCs [5, 6] from PtL pathway [2] as blended fuel, and finally achieve 100% SAF. In the development phase of SAF, data obtained from measurements, testing, and computer modeling are employed to achieve insight into the desired properties from known structures of the compounds. For example, the net heat of combustion (NHOC) of AF is crucial for understanding a fuel's energy content and performance characteristics for energy efficiency, flight range, and payload capacity of jets [8]. The NHOC is categorized into gravimetric ($NHOC_G$) and volumetric ($NHOC_V$), and the former ($NHOC_G$) is suitable for weight-limited aircraft, such as rockets and spacecraft. Likewise, the latter ($NHOC_V$) helps to reduce the aircraft fuel tank volume. Therefore, for volume-limited aircraft such as missiles and military aircraft, high $NHOC_V$ fuel helps increase the payload without changing the tank size [9]. AF, such as RJ-4, RJ-5, and JP-10, have already been developed from PCHCs [10].

In this study, we developed a machine learning (ML) model from a training set of diverse hydrocarbons with known properties of NHOC and density. The model is then applied to pre-screen the chemical structures of PCHCs with high-performance for SAF applications. The present study is an initial pre-screen of a multilevel study aiming at rational design for high-performance SAF from the PtL pathway [2]. The selected high-performance PCHCs will be studied quantum mechanically using DFT calculations for their quantitative structure-property relationship (QSPR) to identify suitable SAF candidates, followed by synthesis reaction pathway and catalyst development, and finalized with a techno-economic assessment (TEA) for feasibility and costs associated.

2. Methods development

2.1 General process of ML model

Developing predictive models for fuel properties, particularly focusing on net heat of combustion (NHOC) and density in alternative fuels, involves systematically integrating experimental or computational data and machine learning algorithms. Training data collection is a crucial first step, which depends on experimentally measured values. Experimental measurements utilize different devices. For instance, calorimeters are the major measurement devices for accurately and reliably determining the NHOC of alternative fuels. A bomb calorimeter determines the NHOC by quantifying the energy released as the fuel undergoes combustion, a straightforward and popular method [11]. Other calorimeters include adiabatic flame calorimeters and oxygen bomb calorimeters [12]. However, practical constraints such as cost, fuel volume, and time impede the acquisition of experimental data [4].

In addition to experimental data, one can also use density functional theory (DFT) [13, 14] computed data for training models, as DFT is a reliable and accurate computational method. For example, Alibakhshi recently estimated the NHOC of up to 40 organic molecules using the more expensive quantum mechanical CCSD-F12b and DSD-PBEB86 methods, achieving correlation coefficients of 0.9999 and 0.9998, respectively [13]. Similarly, another recent study calculated the NHOC for as many as 295 sesquiterpenoid high-energy density (HED) fuels and reported an average absolute error as small as 2.6% [14]. While these methods ensure high accuracy, they can be resource-intensive and time-consuming.

To address the limitations of experimental and computationally demanding methods, predictive methodologies such as group contribution (GC) methods [15, 16] and machine learning (ML) algorithms [17, 18] offer efficient alternatives. Albahri [19] introduced a more accurate GC method for computing net heat of combustion (NHOC). This method computes

32 atom-type structural groups for NHOC in up to 452 hydrocarbons, yielding NHOC predictions with an average absolute error of 0.71% and a correlation coefficient (R) of 0.9982 [19]. However, this method may be time-consuming for many hydrocarbons and may be unsuitable for newly synthesized complex molecules. GC methods are empirical [15, 16], relying on empirical data and expert knowledge to assign contributions to functional groups.

Machine learning (ML), in particular, emerges as a cost-effective option for analyzing large datasets and predicting complex chemical properties by following the general steps in Fig. 1. By training models on molecular structures and corresponding descriptors, ML extracts patterns and relationships, facilitating the rapid screening of numerous fuel candidates. This predictive capability complements experimental measurements and computationally demanding quantum chemical calculations, enabling the virtual construction of desired molecular structures and significantly reducing the time and resources required for fuel screening. This rapid screening identifies the most promising candidates, saving significant time and resources before future investigation can be done for NHOC (ASTM D4809) and density (ρ) at 15° C (ASTM D4052) and other alternative fuel properties [4]. The general process of ML is presented in Fig. 1.

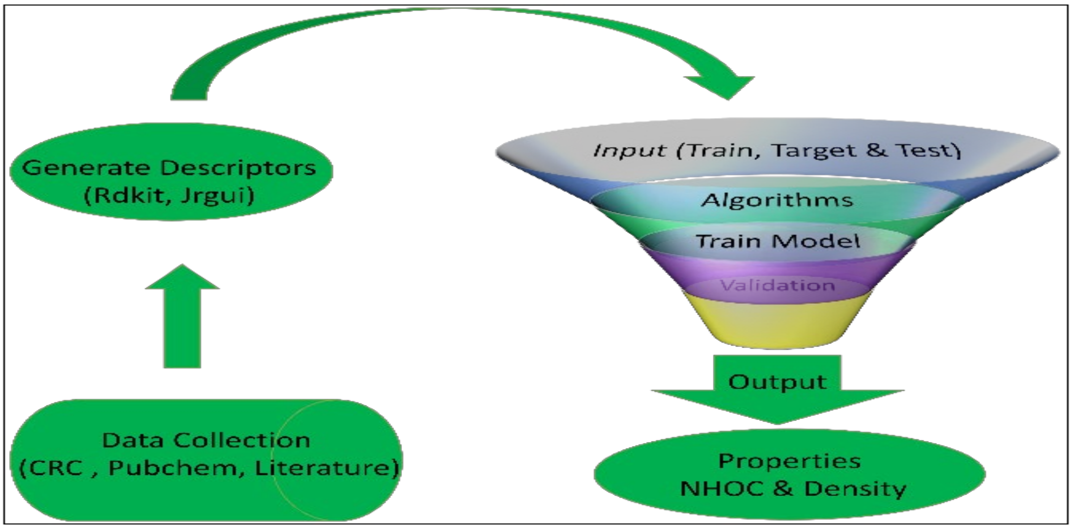

Fig. 1. Flowchart of model development in ML process in the present study.

2.2. Data collection and feature engineering

A comprehensive dataset, AFProp(N, M), is utilized in this study as a training dataset, where N represents the number of fuel properties, and M denotes the number of organic compounds with known properties. Focusing on two fundamental properties (N=2), NHOC and density (ρ). For NHOC properties, the dataset AFProp(1, M=452) compiles data on up to 452 pure hydrocarbons, encompassing paraffins, olefins, naphthenes, and aromatics, sourced from reference [19]. This dataset incorporates NHOC values obtained from experimental measurements and calculations drawn from the American Petroleum Institute - Technical Data Book (API-TDB) [20]. Additionally, the dataset AFProp(2, M=486) comprises density (ρ) properties at temperatures ranging from 15°C to 30°C for up to 486 distinct hydrocarbons, sourced from the CRC Handbook of Chemistry and Physics [21] and [18]. Up to 17 hydrocarbons were excluded from the NHOC dataset AFProp(1, 452) because they were identified as duplicates due to sharing identical Simplified Molecular Input Line Entry System (SMILES) notation [22] for being isomeric hydrocarbons. Furthermore, certain data points were computed values that displayed significant deviations, as highlighted by Albahri [19] and were excluded. Hence, the dataset for NHOC becomes AFProp(1, 435), which is the final training dataset for NHOC.

The chemical structures of the candidate hydrocarbons undergoing screening are also compiled into a dataset, SAFCan(P, Q), where they are systematically represented using SMILES notation. The SMILES notations of dataset AFProp(N, M) and 30 existing PCHCs were directly sourced from Pubchem [23]. Additionally, the SMILES notation for five novel PCHCs circled in Fig. 5 was generated using the Open Babel toolbox [24], ensuring a compact and unambiguous representation of each compound's structure. We utilize a Python program called GUIDEMOL [25, 26] to derive molecular descriptors from the SMILES representations. GUIDEMOL leverages the RDKit toolkit for cheminformatics [25, 26], offering a range of functionalities, including molecular structure handling, substructure searching, molecular similarity calculation, chemical reaction handling, and descriptor calculation.

2.3. ML configuration and molecular descriptors

Machine learning (ML) relies on hyperparameters, which are external configuration settings that cannot be learned directly from the data. These parameters are predetermined and remain fixed throughout training, influencing the model's behavior and performance. Precisely adjusting hyperparameters is essential for accurate model predictions and optimizing predictive precision. To achieve this, hyperparameter tuning techniques such as GridSearchCV [27] are employed. The optimal parameter obtained in different algorithms for NHOC and density (ρ) using GridSearchCV refer to Table S1 in supplementary. GridSearchCV systematically explores a subset of hyperparameters, evaluating the model's performance through cross-validation of the training data. This approach simplifies the tuning process and improves predictive accuracy and optimized model performance [27].

Molecular descriptors or features representing measurable data points' properties were generated using the JRgui graphical user interface (GUI) [25, 28] powered by the Tkinter package. This tool computes descriptors integrated into RDKit [26] and generates grid representations of 3D molecular structures [29]. Leveraging the JRgui and RDKit toolkit [25, 28, 29], we extracted a comprehensive set of approximately 200 descriptors. However, after a thorough examination, only 6 descriptors for NHOC and 40 for density (ρ) were meticulously selected from this set to train the models. In other words, the hydrocarbons' properties can be expressed as a function of these descriptors,

$$AFProp(j,i) = \sum_{i=1}^{n_j} C_{ji}\, x_i + PConst_j \quad (1)$$

Where AFProp(j, i) presents the aviation fuel compound property j and its descriptor i. In the present study, j=1, 2 as only two properties, NHOC and density (ρ), are investigated. As a result, AFProp(1, i=1,2,…,6) for NHOC and AFProp(2, i=1,2,…,40) for density (ρ). $PConst_j$ represents the intercept (constant term) of property j. While x_i is descriptors, the index i runs from 1 to n_j (the number of descriptors) for the property j of the compound under study. For example, for NHOC *(AFProp(1, i=1,2,…,6)*) property of a compound contains six descriptors (n_1=6), and 40 descriptors (n_2=40) for density (ρ) ($Prop_2$). That is, AFProp(1, 6) for NHOC and AFProp(2, 40) for density. Here, C_{ji} (i=1, 2,…,n_j) are the obtained coefficients for property j obtained from the ML model through the training dataset. Table 1 reports the information of the six descriptors x_i (i=1, 2,…,6) of NHOC, whereas the 40 descriptors and corresponding coefficients obtained for the density (ρ) property (j=2, n_2=40) in the model are given in Table S2 in the supplementary materials.

Table 1: The descriptors of property NHOC *(AFProp(1, i=1, 2,…,6)*) in the ML model.*

Symbol	***AFProp(1, i=1, 2,…,6)***	**Description**
X_1	NC	Number of carbons
X_2	NH	Number of hydrogens
X_3	Num_of Atoms	Total number of atoms
X_4	BalabanJ	Topological index
X_5	NumAromaticRings	Number of aromatic rings
X_6	Kappa3	Coefficient of characteristics

*Descriptors with a strong correlation to NHOC are highlighted in grey.

As can be seen in Table 1, the three descriptors, NC, NH, and Num_of _Atoms, are dependent as NC+NH=Num_of_Atoms for hydrocarbons. Theoretically, orthogonal (independent) descriptors in Equ (1) are preferred because they simplify the interpretation of the model and reduce multicollinearity, which can lead to instability in the estimates of model coefficients. However, in practice, it's not always possible to have completely orthogonal descriptors. Sometimes, including non-orthogonal variables can improve the fitting for several reasons. For example, non-orthogonal variables might provide additional information that improves the model's predictive performance and reduces bias; they can also handle nonlinear relationships between the descriptors and the target variable, leading to more accurate predictions. Sometimes, domain-specific tasks need non-orthogonal variables that are known to be relevant to the prediction task, even if they are correlated with other variables. Moreover, Equ (1) can also be extended to other organic compounds rather than hydrocarbons. We are working on developing a new set of orthogonal descriptors for AF candidates, which are closely related to molecular structures in 3D and energies.

2.4. Machine learning model training

After collecting and performing feature engineering on the dataset AFProp(N, M) and SAFCand(P, Q), the subsequent step entails partitioning the dataset AFProp(N, M) into (80%) for training and (20%) for testing. Following this partitioning, the next step involves selecting a suitable algorithm and training the model for analysis. This study chooses a supervised machine learning (ML) training model, as it can accurately predict target properties [30]. Several commonly used algorithms within the supervised ML training model are considered, including the support vector machines (SVM) [31], random forest (RF) [32], and k-nearest neighbors (KNN) [33]. These algorithms are evaluated to determine the most appropriate one for the task. The model training process is conducted using Google Collaboratory [34], based on Python 3.10 [35]. This platform offers data exploration and visualization flexibility through libraries such as Pandas, NumPy, and Matplotlib. Additionally, it seamlessly integrates with ML libraries like Scikit-Learn, TensorFlow, and PyTorch, providing a comprehensive toolkit for model development [36].

In summary, as indicated earlier, the ML training process depicted in Fig. 1 of the present study begins with data collection from multiple databases and resources. Descriptor generation follows, utilizing tools like Rdkit and Jugui [25, 28, 29]. Subsequently, a supervised training model is chosen, employing an appropriate algorithm such as SVM. The output properties are then validated against a set of properties with available experimental values. This iterative process continues until the output aligns with the target fuel properties.

3. Results and discussion

3.1. Performance of the algorithms

The performance of three major supervised ML algorithms is considered for training. That is, the support vector machines (SVM) [31], random forest (RF) [32], and k-nearest neighbors (KNN) [33]. Evaluation metrics such as mean absolute error (MAE), root mean square error (RMSE), and coefficient of determination (R^2) are employed for the performance of these algorithms. MEA and RMSE are both metrics used to evaluate the accuracy of predictions made by an algorithm or model, whereas R^2 quantifies the extent to which the model's predicted values align with the observed ones [18, 37]. The optimal algorithm performance is characterized by achieving the maximum value of R^2 (0-1) while minimizing the values of MAE and RMSE.

The 5-fold cross validation (5-fold CV) is a widely used approach for assessing prediction accuracy and validating machine learning models applied to evaluate their efficacy [38]. In a 5-fold CV, the data are randomly divided into 5 folds or groups, and the model's ability is summarized using the sample of model evaluation scores. Moreover, as mentioned earlier, the dataset was randomly divided into training (80%) and testing (20%) subsets to ensure robust model development. Additionally, descriptors were standardized to bring different units onto a common scale without altering their original units, thus enhancing the model's performance. Combining these techniques can avoid issues such as overfitting and underfitting and obtain a sense of how the model will transfer to a different dataset [39]. This rigorous approach aimed to improve the accuracy of the training data, resulting in more precise predictions. Fig. 2 compares the performance of the three algorithms, SVM, RF, and KNN, in the prediction of NHOC at 20% random test data (Fig. 2a) and at 5-fold cross-validation (Fig. 2b). Detailed results can be found in Tables S3 and S4 in the supplementary materials.

Fig. 2 illustrates the coefficient of determination (R^2) of NHOC produced by three algorithms, SVM, RF, and KNN, all of which are close to 1.0, indicating a high level of agreement between predicted and observed values. In terms of RMSE values for 20% random test data (green in Fig. 2a), SVM performs the best NHOC value of 47.237 kJ/mol, followed by RF with the NHOC value of 97.493 kJ/mol, and KNN with 134.753 kJ/mol. Similarly, the MAE values (orange in Fig. 2a) are the smallest for SVM at 27.821 kJ/mol again, while RF and KNN have higher MAE values of 54.058 kJ/mol and 70.472 kJ/mol, respectively. These trends are consistent in Fig. 2b as well. Overall, the SVM algorithm demonstrates superior performance compared to RF and KNN, making it the preferred choice for further calculations. The SVM algorithm is known for its computational efficiency and robust predictive capabilities, particularly when dealing with limited data availability [40].

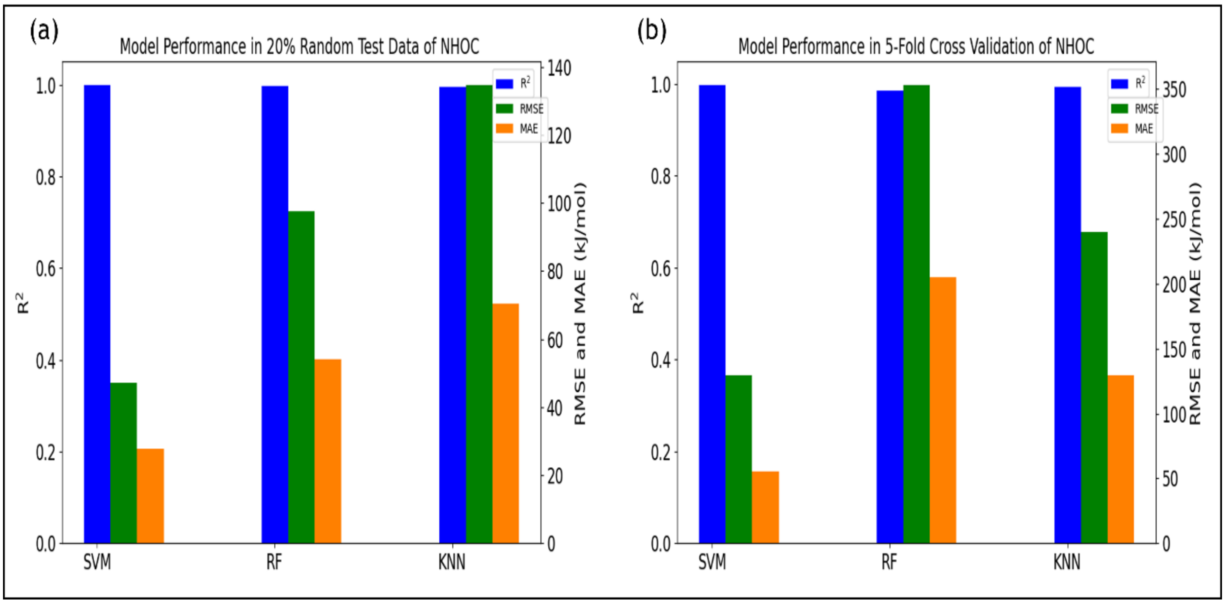

Fig. 2. Comparison of the performance parameters R^2 (Blue), RMSE (green), and MAE (orange) of three algorithms, SVM, RF, and KNN prediction of NHOC. (a) 20% random test data, and (b) 5-fold cross-validation.

3.2. Performance of the SVM trained model on NHOC

The SVM algorithm is applied to the training dataset so that the model (i.e., the coefficients in Equ (1)) is trained and obtained. Table 2 represents the coefficients obtained from the SVM algorithm used in the machine learning model (see Equ (1) and Table 1).

Table 2: The model (coefficients of Equ (1)) obtained for NHOC property.

Descriptors	Coefficients symbol	Coefficients Values
NC	C_{11}	509.9569
NH	C_{12}	78.8326
Num_of Atoms	C_{13}	-21.0102
BalabanJ	C_{14}	13.3489
NumAromaticRings	C_{15}	-205.7653
Kappa3	C_{16}	5.9531
Intercept	$PConst_1$	0.6981

As reported in Table 2, the 6+1 coefficients of the multiple descriptor linear equation (Equ (1)) are positive except for the total number of atoms (Num_of_Atoms) and the number of aromatic rings (NumAromaticRings), which are negative. A positive coefficient suggests that an increase in the descriptor is associated with an increase in the target property (NHOC), whereas a negative coefficient suggests that a decrease in the descriptor is associated with an increase in the target property (NHOC). As a result, if one wishes to enhance the target property NHOC, the descriptors with positive coefficients in Table 2, including NC, NH, BalabanJ, and Kappa3, need to be enhanced. For example, the largest coefficient of the multiple descriptor linear equation (Equ (1)) is NC, the number of carbons, with a coefficient value as large as 509.9569. Fig. 3 displays the relationship of the NHOC property with the number of atoms.

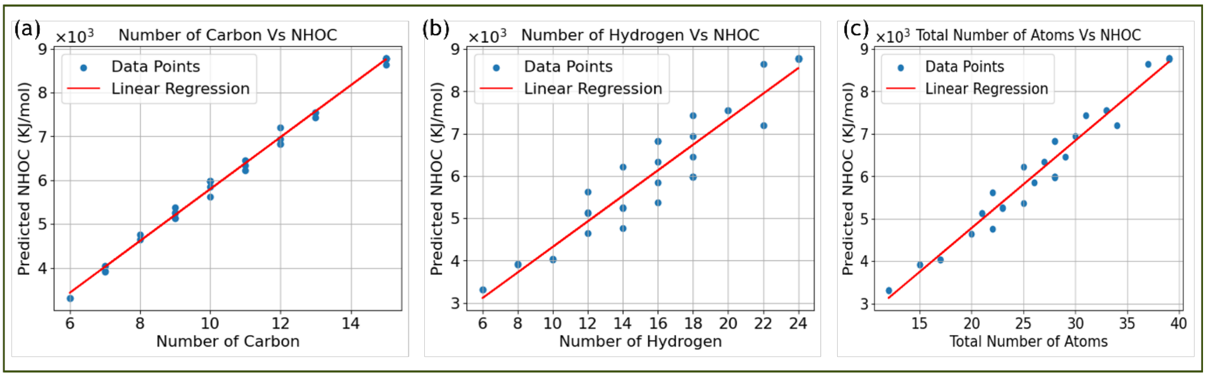

Fig. 3. Impact of positive descriptors on NHOC. (a) NC (Number of carbons), (b) NH (Number of hydrogens), and (c) Total number of atoms in PCHCs (NC+NH).

However, for AF fuel candidate hydrocarbons, the NC of hydrocarbons can only increase within a boundary of approximately 6 < NC < 17, although this boundary varies due to structures and is possible for higher NC hydrocarbons for liquid. If the NC number of a hydrocarbon compound is up to 17, the maximum number of hydrogens is no more than NH < NC×2+2 (36) (for n-alkanes, any unsaturated carbons and rings will reduce the number of hydrogens). As a result, to design novel hydrocarbons with high NHOC, one needs to increase the structure and topological descriptors, BalabanJ and Kappa3, by designing novel hydrocarbon structures. Here, BalabanJ provides essential topological insights [41], and Kappa3 is a third-order molecular connectivity index that captures crucial information regarding a molecule's structural topology and connectivity [41].

In addition, one of these negative coefficients in Table 2, Num_of Atoms, is not orthogonal, as indicated earlier, which restricts the increase of NC and NH. The other negative coefficient, NumAromaticRings, is unwanted, as aromatic compounds produced 88% more soot formation than cycloalkanes due to incomplete combustion [42]. Consequently, the coefficients reported in Table 2 of the ML model provide rich information for the future development of the ML model (Equ (1)) with more efficient descriptors and molecular structures for new candidates.

The accuracy of the SVM-trained model was evaluated by applying it to compute the NHOC of hydrocarbons in various datasets, including training, testing, cross-validation (5-fold), and combined datasets (training + testing). Fig. 4 illustrates the correlation between the predicted NHOC values and the measured NHOC values in different datasets, such as the training dataset 80% of (AFProp(N, M), the cross-validation (5-fold) dataset, the 20% test dataset of (AFProp(N, M), and the combined dataset (train+test or AFProp(N, M)). The consistently high R^2 values near unity with not less than 0.997 for NHOC indicate the model's high accuracy in NHOC property prediction. In addition, the SVM algorithm also demonstrates comparable proficiency in density (ρ) prediction. For more detailed information on the performance of the SVM model in density (ρ) estimation, please refer to Fig. S1 and S2 in the Supplementary Materials. These findings highlight the effectiveness of the ML model in precisely predicting the NHOC and density (ρ) of SAF candidates.

The agreement between the predicted and the literature NHOC values of the compounds is excellent. Most of the compounds are along a straight line except for a small number of compounds in Fig. 4 (a, b, and d), with minor discrepancies as indicated in the oval. Specifically, the NHOC of the hydrocarbons in the vicinity of 0.50 - 0.70×10^4 KJ/mol exhibits larger discrepancies, which are not seen in the test dataset (Fig. 4c). Further examination of the datasets reveals that the molecules in training data sets such as 2,3 pentadiene, 2-Methyl-2,4-Hexadiene, 2-Methyl-1,5-Hexadiene, 2,3-Hexadiene, 2,4-dimethylhexane, and 2,3-dimethyl-1-hexene were either with the computed NHOC or large errors in reference data as highlighted by Albahri [19]. Despite discrepancies from reference data, our SVM-trained model consistently produces accurate results across various datasets.

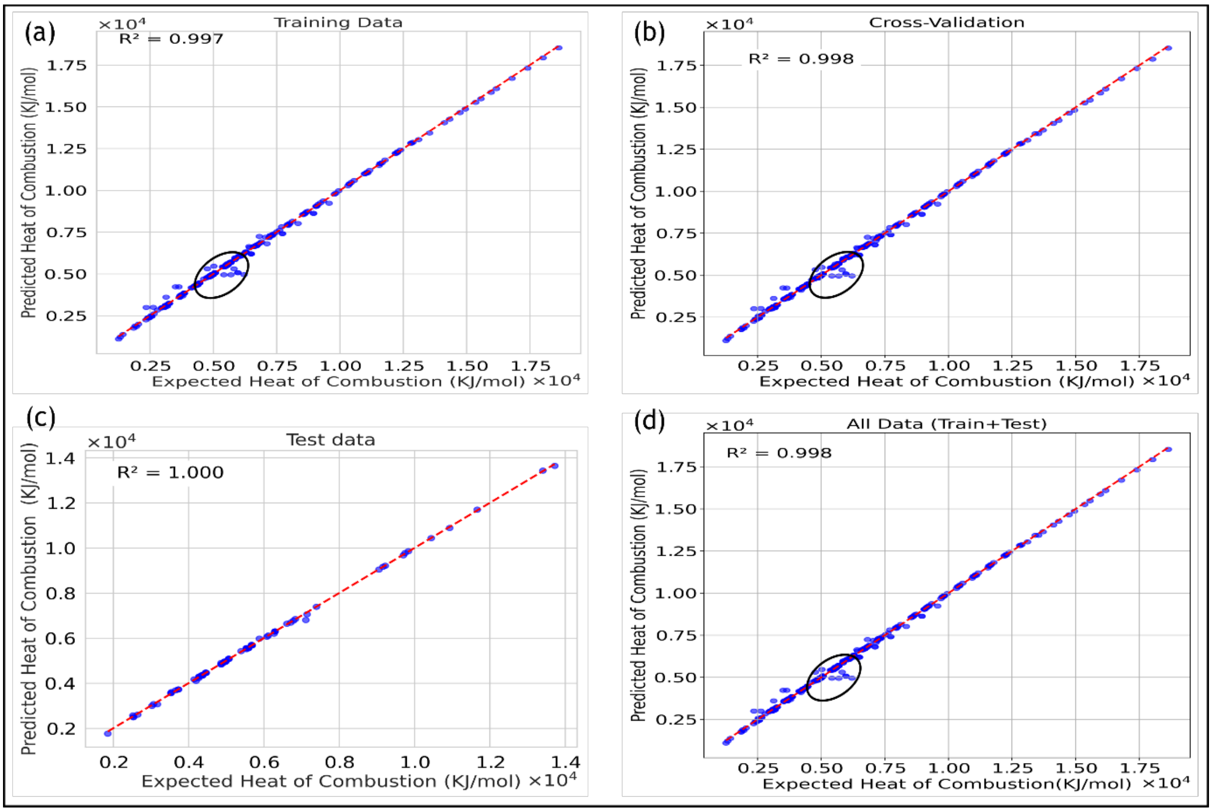

Fig. 4. Predicted versus expected NHOC using the SVM model. (a) Training data (b) Cross-validation (5-fold) data (c) 20% test data (d) All data (Train+Test)

3.3 Screen of liquid hydrocarbons with high NHOC

The coefficient of the SVM-trained model given in Table 2 is employed to examine the hydrocarbons in SAFCand(P, N), which shows the structural properties of various hydrocarbons in Supplementary Fig. S3. The majority of PCHCs in SAFCand(P, Q) align with currently available hydrotreated esters and fatty acids (HEFA) SAFs, which usually contain C9-C16 carbons [43], and SAFs hydrocarbons produced from biomass or other waste in the range of C8-C18 [44]. Although squalane ($C_{30}H_{62}$), a saturated hydrocarbon with an IUPAC name of 2,6,10,15,19,23-hexamethyltetracosane, is a liquid hydrocarbon with up to 30 carbon atoms, the majority of liquid hydrocarbons under ambient temperature do not exceed 17 carbon atoms, whereas the number of hydrogens of the potential PCHCs ranges from H6 to H24, as shown in Fig. 3 (b). Such the numbers of carbons and hydrogens in the compounds in the SAFCand(P, Q) dataset suggest these hydrocarbons are likely polycyclic with saturated C-C bonds and polycyclic hydrocarbons (PCHCs). Leveraging the structural information will help to design novel PCHCs with preferred fuel properties, such as high NHOC and high energy density [45]. Utilizing the SVM-trained model (Equ (1) and Table 2), up to 35 PCHCs are selected as suitable SAF candidates. Fig. 5 reports the chemical structures of these identified candidates using Google Collaboratory [34]. Most selected hydrocarbons correspond to existing compounds, except five PCHCs structures numbered 1 (6377), 2 (250609), 7 (268141), 8 (82630), and 9 (33744) (circled) in Fig. 5, which were designed for HED fuel applications in a previous study [45]. The remaining 30 polycyclic hydrocarbons (PCHCs) were sourced from the PubChem database [23]. As shown in Fig. 5, these PCHCs predominantly comprise saturated cycloalkanes except for ten compounds (28.571%) containing unsaturated C=C bonds. This observation agrees with the fact that saturated hydrocarbons are often preferred for SAFs compared to unsaturated ones [4]. The majority (77.143%) of compounds in Fig. 5 exhibit pentagon ring configuration, and nearly half (42.857%) contain triangular rings, consistent with the outcome reported earlier [45].

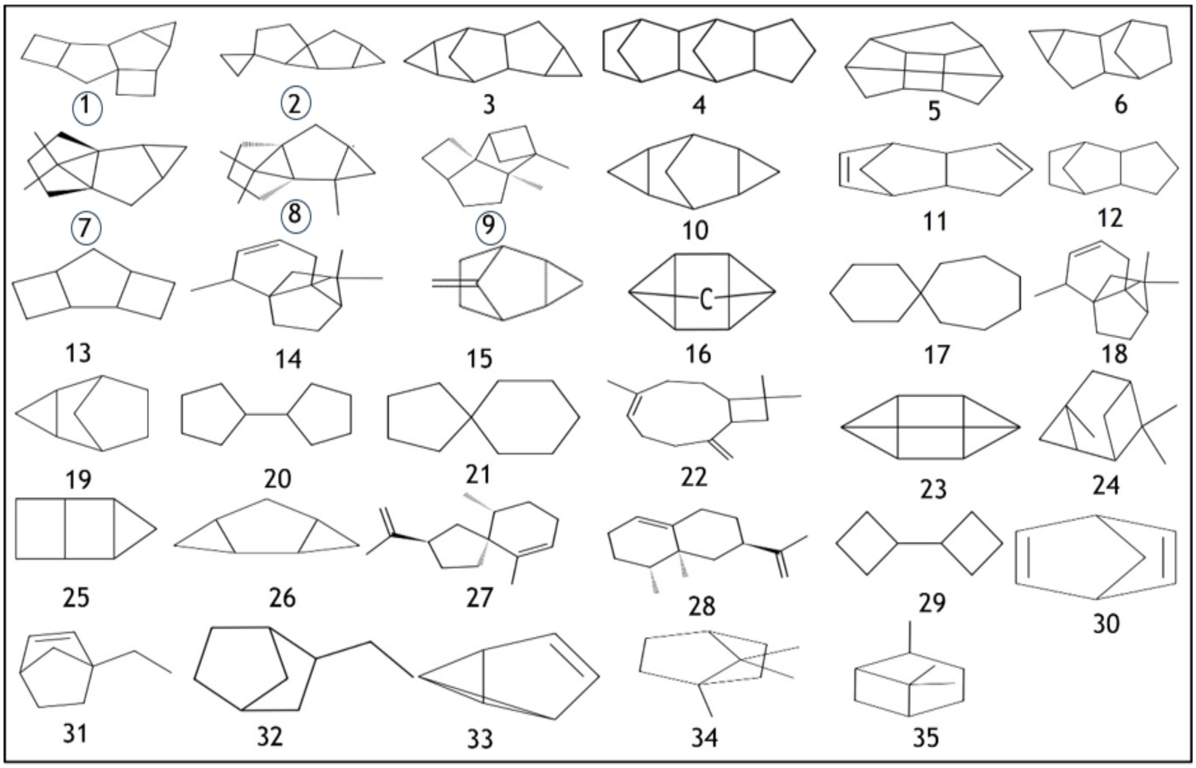

Fig. 5. Chemical structures of 35 PCHCs obtained from ML screen in the present study. Structure 12 (exo-Tetrahydrodicyclopentadiene) is the dominant component of aviation fuel JP-10.

The fuel properties such as gravimetric NHOC ($NHOC_G$), volumetric NHOC ($NHOC_V$), H/C ratio, and density (ρ) of these liquid PCHCs were obtained using the present ML are summarised in Table 3. Note that gravimetric NHOC ($NHOC_G$) and density (ρ) are obtained from ML, and other properties/descriptors such as volumetric NHOC ($NHOC_V$), H/C ratio, and the total number of rings (N_{ring}) are derived. The total number of rings (N_{ring}) in Table 3 can be obtained from RDKit [28, 29] or counted manually from the structures. As seen in the table, almost all these compounds contain either triangular rings or rectangular rings with acute angles or pentagon rings, indicating that they are strained with possibly higher internal energy.

The PCHCs in the table exhibit required ranges of $NHOC_G$ of 42.366-43.277 MJ/kg and $NHOC_V$ of 35.849-52.039 MJ/L.

Table 3. Properties of selected PCHCs for SAF using ML. #

No	Hydrocarbon	Formula CAS No	H/C	ρ (g/ml)	$NHOC_G$ (MJ/Kg)	$NHOC_V$ (MJ/L)	N_{ring}
1	6377*	C13H18	1.385	1.221	42.620	52.039	5
2	250609*	C12H16	1.333	1.221	42.565	51.972	5
3	Pentacyclo (6.3.1.0(2,7).0(3,5).0(9,11)) dodecane	C12H16 82110-70-1	1.333	1.200	42.573	51.087	5
4	THTCPD pentacycyclo (6.5.1.13,6.02,7.09,13) pentadecane	C15H22 75172-85-9	1.467	1.192	42.701	50.900	5
5	Pentacyclo (5.4.0.02,6.03,10.05,9) undecane	C11H14 4421-32-3	1.273	1.165	42.529	49.547	5
6	Tetracyclo (6.2.1.0(2,7).0(3,5)) undecane	C11H16 1 777-44-2	1.455	1.093	42.750	46.726	4
7	268141*	C12H18	1.500	1.083	42.780	46.331	4
8	82630*	C13H20	1.538	1.075	42.813	46.024	4
9	33744*	C13H20	1.538	1.058	42.819	45.303	5
10	Tetracyclo (3.3.1.02,4.06,8) nonane	C9H12 187-49-5	1.333	1.062	42.646	45.290	4
11	Dicyclopentadiene	C10H12 77-73-6	1.200	1.018	42.517	43.282	3
12	Exo-THDCPD (JP-10)	C10H16 2825-82-3	1.600	0.990	42.968	42.539	3
13	Tricyclo (5.2.0.02,5) nonane	C9H14	1.555	0.982	42.928	42.156	3
14	Alpha neoclovene	C15H24 45-45-68-0	1.600	0.972	42.913	41.712	3
15	Tricyclo (3.2.1.0(2,4)) octane	C9H12 38310-48-4	1.333	0.970	42.900	41.613	3
16	Quadricyclane (QC)	C7H8 278-06-8	1.143	0.978	42.492	41.557	5
17	Spiro (5,6) dodecane	C12H22 181-15-7	1.833	0.957	43.277	41.416	2
18	Gamma neoclovene	C15H24	1.600	0.957	42.913	41.068	3
19	Tricyclo (3.2.1.02,4) octane	C8H12 13377-46-3	1.500	0.952	42.900	40.841	3
20	Bicyclopentane	C10H18 1636-39-1	1.800	0.933	43.252	40.354	3
21	Spiro (4,5) decane	C10H18 176-63-6	1.800	0.931	43.259	40.274	2
22	Caryophyllene	C15H24	1.600	0.928	42.976	39.882	2

		87-44-5					
23	Prismane	C6H6 650-42-0	1.000	0.941	42.366	39.866	5
24	4,7,7-Trimethyltricyclo (4.1.1.02,4) octane	C11H18	1.636	0.927	42.981	39.843	3
25	Tricyclo (3.2.0.02,4) heptane	C7H10 28102-61-6	1.429	0.929	42.834	39.793	3
26	Tricyclo (4.1.0.02,4) heptane	C7H10 187-26-8	1.429	0.927	42.845	39.717	3
27	Premnaspirodiene	C15H24 82189-85-3	1.600	0.915	42.970	39.318	2
28	Valencene	C15H24 3-07-4630	1.600	0.912	42.975	39.193	2
29	Bicyclobutyl	C8H14 7051-52-7	1.750	0.897	43.214	38.763	2
30	Norbornadiene	C7H8 16422-76-7	1.143	0.895	42.608	38.134	2
31	Ethylnorbornene	C9H14 2146-41-0	1.556	0.884	43.005	38.016	2
32	5-Ethylnorbornane	C9H16 2146-41-0	1.778	0.860	43.254	37.191	2
33	Benzvalene	C6H6 659-85-8	1.000	0.875	42.404	37.104	4
34	Camphane	C10H18 464-15-3	1.800	0.851	43.256	36.811	2
35	Pinane	C10H18 473-55-2	1.800	0.829	43.243	35.849	3

#The $NHOC_G$ of jet fuels is 42.20-43.98 MJ/kg and $NHOC_V$ is 32.26-39.64 MJ/L [10]. *Ref [45] (no CAS numbers available).

There exists a linear positive relationship between $NHOC_G$ and the H/C ratio of the PCHCs. That is, the larger the H/C ratio, the larger the $NHOC_G$. For example, Spiro (5,6) dodecane (No 17) exhibits the largest $NHOC_G$ of 43.277 MJ/Kg among the set of PCHCs with an H/C ratio of 1.833, whereas Prismane (No 22) displays the smallest $NHOC_G$ of 42.366 KJ/kg, as detailed in Table 3, is characterized by an H/C ratio of 1. A larger H/C ratio suggests a preference for saturated hydrocarbons with more C-H bonds, aligning with the fact that composition standards of SAFs are characterized by a higher H/C ratio [46]. However, the H/C ratio should not be overstated, as molecular electronic configuration and chemical bonding play an essential role in molecular properties.

SAF candidates also need to exhibit balanced properties. The optimal SAF candidates in Table 3 are not necessarily those showing the largest $NHOC_G$ nor the largest $NHOC_{V,}$ as it is often unlikely that the hydrocarbons with large $NHOC_G$ also have large $NHOC_V$ or vice versa. Achieving the optimal balance between $NHOC_G$ and $NHOC_V$ requires a holistic approach considering specific aircraft requirements, such as operational conditions and technological advancements. The present study employs a high energy density (HED) aviation fuel JP-10 as the fuel reference. The density of JP-10 is 0.940 g/ml, with a high $NHOC_G$ of 42.200 MJ/kg and an $NHOC_V$ of 39.640 MJ/L [10].

The $NHOC_G$ of SAF candidates in Table 3 is above the reference (i.e., the $NHOC_G$ of JP-10 42.200 MJ/kg). As a result, it is important to examine the density and $NHOC_V$ properties of these SAF candidates. Fig. 6 plots $NHOC_V$ and density of these PCHCs, which varies significantly from as low as 35.849 MJ/L to as high as 52.039 MJ/L due to the density variation. For new SAF candidates with higher density (>0.94 g/ml) and larger $NHOC_V$ (39.64 MJ/L), the preferred PCHCs need to be on the right-hand side of the vertical orange dash line and above the horizontal orange dash line (north-east or phase I). Up to 20 PCHCs in this region in Fig. 6 fit these criteria and, therefore, can be excellent candidates with superior density and $NHOC_v$ than JP-10 to proceed with the development.

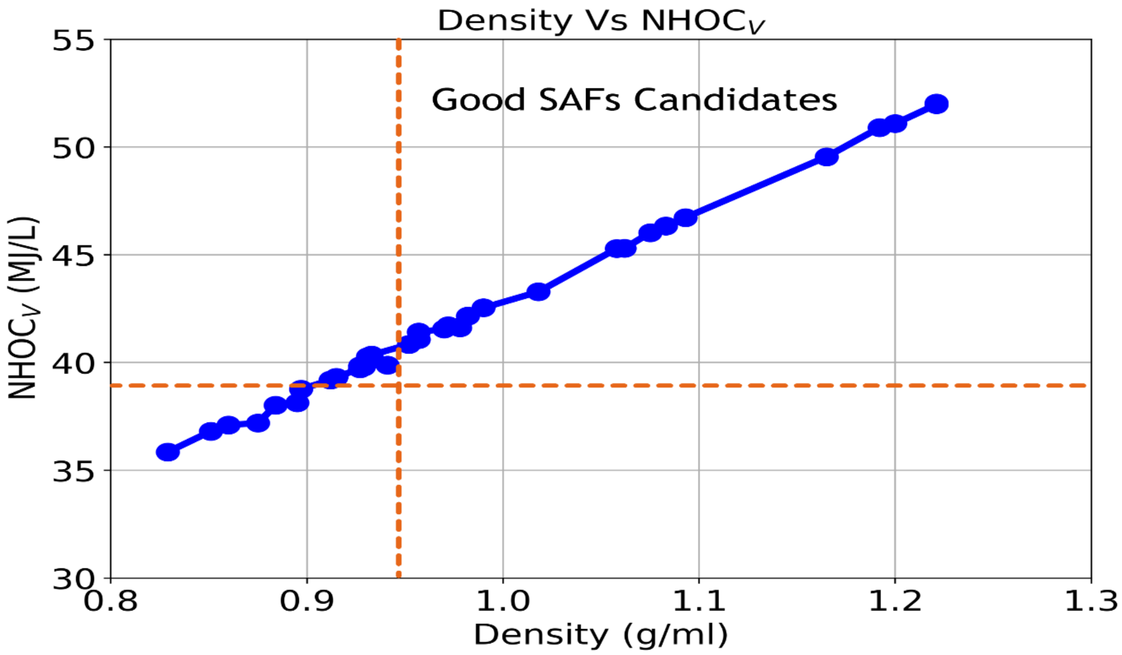

Fig. 6. Relationship between NHOC$_V$ and density of SAF candidates in Table 3. The 20 preferred candidates are No 1-No 19 and No 23 PCHCs.

The results show that properties such as NHOC$_V$ and density of PCHCs are more sensitive to the molecular structure of the PCHCs than NHOC$_G$. Further structure analysis of these PCHCs reveals that the preferred PCHCs (25 out of 35) possess 3-5 rings. NHOC$_V$ and density are likely related to the number of rings in the structures. For example, compound (No 1, C13H18) in Table 3 with 5 rings has the largest NHOC$_V$ of 52.039 MJ/L, whereas the majority of the PCHCs possessing 2 rings are at the bottom of the list. JP-10 fuel (dominated by exo-THDCPD C10H16) possesses 3 pentagon rings (See Structure 12 in Fig. 5) with a density of 0.990 g/ml, NHOC$_G$ 42.968 MJ/kg, and NHOC$_V$ 42.539 MJ/L, likewise THTCPD pentacyclic (6.5.1.13,6.02,7.09,13) pentadecane (C15H22) possesses 5 pentagon rings analogy to the structure of JP-10 (See Structure 4 in Fig. 5). The predicted properties for this compound 4 are density 1.192 g/ml, NHOC$_G$ 42.701 MJ/kg, and NHOC$_V$ 50.900 MJ/L, which indicate that THTCPD is a highly promising compound for SAF with a potential as an HED fuel for military aircraft as well as the substitute for aromatic components in conventional aviation fuels.

3.4. Structure-property relationship for hydrocarbons in SAF

Understanding the structural characteristics of these PCHCs provides insight for the QSPR for future design and development of SAFs. Among the 35 PCHCs screened from the ML model, Structures 1- 11 in Table 3 exhibit superior $NHOC_v$ and density properties than the HED JP-10 aviation fuel. Examination of the characteristics of these hydrocarbons in Table 3 reveals that they share the following features:

1. Compact molecular structure: the obtained PCHCs have a compact molecular structure with multiple fused rings. This compactness allows for efficient packing of molecules, leading to higher energy density per unit volume.
2. Ratio of H/C: NHOC of PCHCs is predominantly determined by the ratio of H/C. Selection and design of new PCHCs for SAFs can be advanced by prioritizing these influential factors. However, the number of total atoms descriptors, which is not independent, may be removed for new descriptor development for SAF.
3. Multiple C-C bonds: PCHCs contain several C-C bonds within their fused ring structures. These bonds store large amounts of energy during combustion reactions, producing high heat release rates and enhanced energy output.
4. Ring strain: The presence of fused rings in PCHCs often leads to significant ring strain, which arises from the forced bending or distortion of C-C bonds to accommodate ring fusion. This strain imparts high reactivity to the molecule, facilitating rapid combustion and efficient energy release.
5. Saturation vs unsaturation: PCHCs may be both saturated and unsaturated. Saturation dominates the PCHCs and contributes to the stability and thermal resistance of the molecule, while unsaturation enhances reactivity and combustion efficiency.

6. Substituents: Functional groups or substituents attached to the polycyclic ring system can further modulate the properties of a hydrocarbon, such as polarity, solubility, and reactivity. For example, alkyl groups may increase the hydrophobicity and stability of the molecule, while polar functional groups may enhance interactions with other molecules or surfaces.
7. Steric hindrance: The three-dimensional (3D) arrangement of atoms of the compounds can introduce steric hindrance, affecting the molecule's interactions with surrounding molecules, surfaces, or catalysts. This can influence factors such as combustion kinetics, reaction rates, and product distributions.

4. Conclusions

The study focused on developing a machine learning (ML) model that efficiently estimates critical fuel properties like net heat of combustion (NHOC) and hydrocarbon density. Using the supervised support vector machines (SVM) algorithm, models with six and forty descriptors were trained for NHOC and density, respectively, ensuring accuracy and reliability. These models were then applied to screen molecules from literature and database, identifying 35 high-energy density polycyclic hydrocarbon (PCHCs) molecules suitable for sustainable aviation fuel (SAF) applications. Interestingly, around 70% of these PCHCs exhibited NHOCv values comparable to or better than JP-10 jet fuel. Notably, the optimal PCHCs favored multiple rings with C-C single bonds and a high H/C ratio. However, this pre-screening step is just the beginning, as further steps involve developing quantitative structure-property relationships (QSPR), selecting or developing suitable catalysts for PCHCs synthesis, blending the PCHCs into aviation fuel, and assessing the impact on fuel properties according to ASTM specifications. Feasibility studies, techno-economic analyses, and environmental impact assessments are also crucial aspects of fuel development.

CRediT authorship contribution statement

Dilip Rijal: Writing – original draft, Data collection, Methodology, Editing. **Feng Wang:** Conceptualization, Supervision, Visualization, Writing – review, Editing. **Vladislav Vasilyev:** Review, Supervision, Methodology

Acknowledgements

DR acknowledges the Tuition Fee Scholarship of the Swinburne University of Technology. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

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

Appendix A. Supplementary materials

References

[1] International Civil Aviation Organization (ICAO). *Environmental Trends in Aviation to 2050.*https://www.icao.int/environmental-protection/Documents/EnvironmentalReports/2022/ENVReport2022_Art7.pdf (accessed March 14, 2024).

[2] Wang, F.; Rijal, D. Sustainable Aviation Fuels for Clean Skies: Exploring the Potential and Perspectives of Strained Hydrocarbons. *Energy & Fuels* **2024**. DOI: https://doi.org/10.1021/acs.energyfuels.3c04935

[3] Ince, A. C.; Colpan, C. O.; Hagen, A.; Serincan, M. F. Modeling and simulation of Power-to-X systems: A review. *Fuel* **2021**, *304*, 121354. DOI: https://doi.org/10.1016/j.fuel.2021.121354

[4] 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

[5] Landera, A.; Bambha, R. P.; Hao, N.; Desai, S. P.; Moore, C. M.; Sutton, A. D.; George, A. Building structure-property relationships of cycloalkanes in support of their use in sustainable aviation fuels. *Frontiers in Energy Research* **2022**, *9*, 771697. DOI: https://doi.org/10.3389/fenrg.2021.771697

[6] Kosir, S.; Heyne, J.; Graham, J. A machine learning framework for drop-in volume swell characteristics of sustainable aviation fuel. *Fuel* **2020**, *274*, 117832. DOI: https://doi.org/10.1016/j.fuel.2020.117832

[7] TB, G. H. Aviation Fuels Technical Review. **2007**. https://www.chevron.com/-/media/chevron/operations/documents/aviation-tech-review.pdf (accessed Jan 5, 2024).

[8] Zou, J.-J.; Zhang, X.; Pan, L. *High-energy-density fuels for advanced propulsion: Design and synthesis*; John Wiley & Sons, 2020.

[9] Nie, J.; Jia, T.; Pan, L.; Zhang, X.; Zou, J.-J. Development of high-energy-density liquid aerospace fuel: a perspective. *Transactions of Tianjin University* **2022**, *28* (1), 1-5. DOI: https://doi.org/https://doi.org/10.1007/s12209-021-00302-x

[10] Woodroffe, J.-D.; Harvey, B. G. High-performance, biobased, jet fuel blends containing hydrogenated monoterpenes and synthetic paraffinic kerosene's. *Energy & Fuels* **2020**, *34* (5), 5929-5937. DOI: https://doi.org/10.1021/acs.energyfuels.0c00274

[11] Jessup, R. S. *Precise measurement of heat of combustion with a bomb calorimeter*; US Department of Commerce, National Bureau of Standards, 1960.

[12] Rojas-Aguilar, A. An isoperibol micro-bomb combustion calorimeter for measurement of the enthalpy of combustion. Application to the study of fullerene C60. *The Journal of Chemical*

Thermodynamics **2002**, *34* (10), 1729-1743. DOI: https://doi.org/10.1016/S0021-9614(02)00257-4

[13] Alibakhshi, A. High precision evaluation of the combustion enthalpy by ab-intio computations. **2021**. DOI: https://doi.org/10.26434/chemrxiv-2021-fvcph

[14] Yang, H.; Yang, Z.-J.; Yang, Q.-F.; Wei, X.-M.; Yuan, Y.-Q.; Wang, L.-L.; Hu, Y.-F.; Ding, J.-J. Simple and high-precision DFT-QSPR prediction of enthalpy of combustion for sesquiterpenoid high-energy–density fuels. *Fuel* **2023**, *332*, 126157. DOI: https://doi.org/10.1016/j.fuel.2022.126157

[15] Frutiger, J.; Marcarie, C.; Abildskov, J.; Sin, G. r. A Comprehensive Methodology for Development, Parameter Estimation, and Uncertainty Analysis of Group Contribution Based Property Models☐An Application to the Heat of Combustion. *Journal of Chemical & Engineering Data* **2016**, *61* (1), 602-613. DOI: https://doi.org/10.1021/acs.jced.5b00750

[16] Marrero, J.; Gani, R. Group-contribution based estimation of pure component properties. *Fluid phase equilibria* **2001**, *183*, 183-208. DOI: https://doi.org/10.1016/S0378-3812(01)00431-9

[17] Abdul Jameel, A. G.; Al-Muslem, A.; Ahmad, N.; Alquaity, A. B.; Zahid, U.; Ahmed, U. Predicting enthalpy of combustion using machine learning. *Processes* **2022**, *10* (11), 2384. DOI: https://doi.org/10.3390/pr10112384

[18] Li, R.; Herreros, J. M.; Tsolakis, A.; Yang, W. Machine learning-quantitative structure property relationship (ML-QSPR) method for fuel physicochemical properties prediction of multiple fuel types. *Fuel* **2021**, *304*, 121437. DOI: https://doi.org/10.1016/j.fuel.2021.121437

[19] Albahri, T. A. Method for predicting the standard net heat of combustion for pure hydrocarbons from their molecular structure. *Energy conversion and management* **2013**, *76*, 1143-1149. DOI: https://doi.org/10.1016/j.enconman.2013.09.019

[20] API. Technical data book—petroleum refining. *Metric Edition, Vols* **1997**, *1* (2).

[21] Rumble, J. CRC handbook of chemistry and physics. **2017**.

[22] DAYLIGHT. *SMILES*.https://www.daylight.com/dayhtml/doc/theory/theory.smiles.html (accessed May 24, 2024).

[23] PubChem. *Explore Chemistry*. https://pubchem.ncbi.nlm.nih.gov/ (accessed Jan 26, 2024).

[24] O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *Journal of cheminformatics* **2011**, *3* (1), 1-14. DOI: https://doi.org/10.1186/1758-2946-3-33

[25] Aires de Sousa, J. GUIDEMOL: a Python graphical user interface for molecular descriptors based on RDKit. *Molecular Informatics* **2023**. DOI: https://doi.org/10.1002/minf.202300190

[26] The RDKit documentation. https://www.rdkit.org/docs/index.html (accessed Feb 25, 2024).

[27] Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* **2011**, *12*, 2825-2830.

[28] Shi, C.; Borchardt, T. B. JRgui: A Python program of Joback and Reid method. *ACS omega* **2017**, *2* (12), 8682-8688. DOI: https://doi.org/10.1021/acsomega.7b01464

[29] Duarte Ramos Matos, G.; Pak, S.; Rizzo, R. C. Descriptor-Driven de Novo Design Algorithms for DOCK6 Using RDKit. *Journal of Chemical Information and Modeling* **2023**, *63* (18), 5803-5822. DOI: https://doi.org/10.1021/acs.jcim.3c01031

[30] Moubayed, A.; Injadat, M.; Nassif, A. B.; Lutfiyya, H.; Shami, A. E-learning: Challenges and research opportunities using machine learning & data analytics. *IEEE Access* **2018**, *6*, 39117-39138. DOI: https://doi.org/10.1109/ACCESS.2018.2851790

[31] Soman, K.; Loganathan, R.; Ajay, V. *Machine learning with SVM and other kernel methods*; PHI Learning Pvt. Ltd., 2009.

[32] Shaik, A. B.; Srinivasan, S. A brief survey on random forest ensembles in classification model. In *International Conference on Innovative Computing and Communications: Proceedings of ICICC 2018, Volume 2*, 2019; Springer: pp 253-260. DOI: https://doi.org/10.1007/978-981-13-2354-6_27

[33] Steinbach, M.; Tan, P.-N. *kNN: k-nearest neighbors*; 2009.

[34] Sukhdeve, D. S. R.; Sukhdeve, S. S. Google Colaboratory. In *Google Cloud Platform for Data Science: A Crash Course on Big Data, Machine Learning, and Data Analytics Services*, Springer, 2023; pp 11-34.

[35] Martelli, A.; Ravenscroft, A. M.; Holden, S.; McGuire, P. *Python in a Nutshell*; " O'Reilly Media, Inc.", 2023.

[36] 10 Essential Data Science Package for Python-Kite Blog. https://www.kite.com/blog/python/data-science-packages-python/ (accessed Feb 25, 2024).

[37] Montgomery, D. C. *Design and analysis of experiments*; John Wiley & sons, 2017.

[38] Baykan, N. A.; Yılmaz, N. A mineral classification system with multiple artificial neural network using k-fold cross validation. *Mathematical and Computational Applications* **2011**, *16* (1), 22-30. DOI: https://doi.org/10.3390/mca16010022

[39] Montesinos López, O. A.; Montesinos López, A.; Crossa, J. Overfitting, model tuning, and evaluation of prediction performance. In *Multivariate statistical machine learning methods for genomic prediction*, Springer, 2022; pp 109-139.

[40] Otchere, D. A.; Ganat, T. O. A.; Gholami, R.; Ridha, S. Application of supervised machine learning paradigms in the prediction of petroleum reservoir properties: Comparative analysis

of ANN and SVM models. *Journal of Petroleum Science and Engineering* **2021**, *200*, 108182. DOI: https://doi.org/10.1016/j.petrol.2020.108182

[41] CHEMICAL COMPUTING GROUP. *QuaSAR-Descriptor* https://cadaster.eu/sites/cadaster.eu/files/challenge/descr.htm (accessed Feb 28, 2024).

[42] Ladommatos, N.; Rubenstein, P.; Bennett, P. Some effects of molecular structure of single hydrocarbons on sooting tendency. *Fuel* **1996**, *75* (2), 114-124. DOI: https://doi.org/10.1016/0016-2361(94)00251-7

[43] Kittel, H.; Horský, J.; Šimáček, P. Properties of Selected Alternative Petroleum Fractions and Sustainable Aviation Fuels. *Processes* **2023**, *11* (3), 935. DOI: https://doi.org/10.3390/pr11030935

[44] Huq, N. A.; Hafenstine, G. R.; Huo, X.; Nguyen, H.; Tifft, S. M.; Conklin, D. R.; Stück, D.; Stunkel, J.; Yang, Z.; Heyne, J. S. Toward net-zero sustainable aviation fuel with wet waste–derived volatile fatty acids. *Proceedings of the National Academy of Sciences* **2021**, *118* (13), e2023008118. DOI: https://doi.org/10.1073/pnas.2023008118

[45] Li, G.; Hu, Z.; Hou, F.; Li, X.; Wang, L.; Zhang, X. Machine learning enabled high-throughput screening of hydrocarbon molecules for the design of next generation fuels. *Fuel* **2020**, *265*, 116968. DOI: https://doi.org/10.1016/j.fuel.2019.116968

[46] Lin, J.-K.; Abi Nurazaq, W.; Wang, W.-C. The properties of sustainable aviation fuel I: Spray characteristics. *Energy* **2023**, *283*, 129125. DOI: https://doi.org/10.1016/j.energy.2023.129125