Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Comparative Machine Learning Approaches for QSPR-Based Prediction of Thermochemical Properties in Rocket Fuel Candidate Screening

Yashwin G Devadiga
R.V. College of Engineering
Parth Sarathi Panigrahy
parthsarathip@rvce.edu.in

R.V. College of Engineering
Swarna M Patra
R.V. College of Engineering
Ganga Periyasamy
Bangalore University
K. N. Subramanya
R.V. College of Engineering

Article

Keywords: uantitative Structure–Property Relationship (QSPR), Artificial Neural Networks, Multi-Output Regression, Thermochemical Properties, Molecular Descriptors, Rocket Fuel Screening, Machine Learning

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11131297/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Comparative Machine Learning Approaches for QSPR-Based Prediction of Thermochemical Properties in Rocket Fuel Candidate Screening

Yashwin G Devadiga[1], Parth Sarathi Panigrahy[2*], Swarna M Patra[3*], Ganga Periyasamy[4], K. N. Subramanya[5]

[1]Department of Computer Science and Engineering, R.V. College of Engineering, Bangalore, Karnataka, India.
[2]Department of Chemistry, R.V. College of Engineering, Bangalore, Karnataka, India.
[3]Department of Electrical and Electronics Engineering, R.V. College of Engineering, Bangalore, Karnataka, India.
[4]Department of Chemistry, Bangalore University, Bangalore, Karnataka, India.
[5]Department of Industrial Engineering and Management, R.V. College of Engineering, Bangalore, Karnataka, India.

*Corresponding author(s). E-mail(s): parthsarathip@rvce.edu.in; swarnamp@rvce.edu.in;

Abstract

Experimental screening of compounds based on thermochemical properties is a time-consuming and expensive process. In this work, a multi-output Artificial Neural Network is developed using molecular descriptors to predict fuel-related thermochemical properties, such as flash point, vapor pressure and surface tension. The input features are selected using domain knowledge and information-gain metrics such as mutual information. Using cross-validation techniques and exhaustive feature engineering, the developed model produced competitive performance, with that of ensemble learning models like XGBRegressor and supervised machine learning models like SVR, for several target properties while enabling simultaneous prediction of multiple thermochemical properties. This architecture demonstrate the potential of using multi-output shared representation to accelerate the computational screening of candidate compounds for rocket fuel applications.

Keywords: Quantitative Structure–Property Relationship (QSPR), Artificial Neural Networks, Multi-Output Regression, Thermochemical Properties, Molecular Descriptors, Rocket Fuel Screening, Machine Learning

1 Introduction

The systematic study and screening of compounds evaluated through thermochemical properties is an important process in the development of rocket fuels. However, conducting experimental evaluations of compounds to determine these properties poses significant challenges such as a lack of scalability when large number of compounds are involved, cost, time constraints among others.

To address these challenges, Quantitative Structure–Property Relationship (QSPR) models are used to predict thermochemical properties using molecular descriptors. Such approaches enable scalable computational screening of candidate compounds while reducing the cost and time associated with experimental evaluation. However, several limitations remain in existing QSPR-based approaches.

Several studies have explored the use of Machine Learning based QSPR modeling to predict thermochemical properties. Runzhao et al. [1] developed a generalized ML-QSPR framework for predicting multiple physicochemical fuel properties across diverse fuel classes using molecular descriptors. In a subsequent work, the authors [2] proposed an integrated ML-QSPR and chemical kinetics framework for high-throughput virtual fuel screening. Many prior works have explored the use of topological indices and machine learning methods for predicting physicochemical fuel properties[3–5]. Several prior works have also employed descriptor-based QSPR techniques using models such as SVM and linear regression for predicting physicochemical properties. [6–12]. While gradient-boosted tree algorithms like XGBoost have demonstrated high predictive performance across tabular chemical data, their application as a high-capacity benchmark for evaluating complex multi-property thermochemical datasets remains limited.

Thermochemical properties often exhibit significant interdependence, due to the underlying molecular structure governing them [4]. For instance, the molecular characteristics that influence boiling point may also affect properties such as density, vapor pressure, and flash point. Capturing such underlying correlations may improve model generalization and provide a unified framework for simultaneous property prediction.[4]. However, many existing QSPR-based approaches predict these properties independently, thereby overlooking the shared relationships existing between the target properties. To explore these shared dynamics, this work investigates a multi-output ANN framework with shared hidden representations to evaluate its capacity for learning common molecular patterns across multiple thermochemical properties simultaneously.

The key thermochemical properties relevant to rocket fuel performance include boiling point (T_b), flash point (T_f), and safety parameters such as ignition temperature(T_i, among others. These properties collectively govern critical aspects of fuel behavior such as volatility, energy release, and safe handling during storage and

operation, and often exhibit interdependent behavior arising from shared molecular structure.

This work proposes a multi-output ANN framework that utilizes a shared layer architecture to predict thermochemical properties using molecular descriptors. The descriptors are calculated using a RDKit and Mordred-based algorithm that utilizes the SMILE notation of a molecule and graph theory. These descriptors are chosen based on their ability to capture molecular characteristics such as molecular branching, compactness, intermolecular forces of attraction and polarity .A dataset of 413 compounds was used in this study. Given the moderate size of the dataset, feature selection and mutual-information-based analysis were employed to identify informative descriptors and reduce model complexity.[13]. The model was evaluated on the aforementioned dataset. While the multi-output ANN successfully captured shared representation patterns across correlated target properties, experimental comparisons revealed that tree-based ensemble methods, specifically XGBRegressor, as well as SVM, achieved superior predictive performance on this dataset. This comparative evaluation highlights critical trade-offs between deep multi-task representation learning and tree-based gradient boosting for tabular thermochemical property prediction.

The remainder of this paper is organized as follows. Section 2 provides background on graph theory concepts and the molecular descriptors used in this study. Section 3 describes the proposed methodology, including the ANN architecture and training procedures employed. Section 4 discusses the dataset sources and pre-processing techniques applied. Section 5 presents the experimental results and provides a comparative performance analysis between the multi-output ANN and baseline regression models, including XGBRegressor and SVR models. Section 6 analyzes model interpretability and feature contributions. Finally, Section 7 concludes the paper and discusses potential future work.

2 Preliminary and Descriptor Engineering

2.1 Graph-Theoretic Preliminaries

In computational chemistry, chemical molecules are represented in the form of graphs. For a given molecule, a graph is a topological description, where the atoms are represented by vertices, and the bonds are represented by edges. In many graph-theoretic representation, the edges are treated uniformly, enabling structural analysis independent of geometric coordinates. Such treatment of molecules enables calculation of certain characteristics of the graph which can be used to predict various physicochemical and thermochemical properties of a molecule.

In order to calculate such descriptors, a chemical notation language called Simplified Molecular Input Line Entry System (SMILES) was used to represent molecular structures as a single-line text. These representations were subsequently converted into molecular graphs, where atoms correspond to vertices and bonds correspond to edges.

The aforementioned definitions form the basis for several graph- theoretic concepts used throughout this work.

Definition 1 The degree of an atom is defined as the number of edges that are connected to one vertex in the graph. For a vertex m in a chemical graph, the degree of the vertex is represented as d_m

Definition 2 The neighborhood degree of a vertex v is defined as the sum of degrees of all vertices adjacent to v. It is denoted by θ_v

$$\theta_v = \sum_{u \in N(v)} d_u \tag{1}$$

With these definitions, we move on to define some of the molecular descriptors used in our model

Definition 2.1 The Balaban Index is given by

$$J(G) = \frac{m}{\mu + 1} \sum_{(u,v) \in E(G)} \frac{1}{\sqrt{w(u)w(v)}} \tag{2}$$

where m is the no of edges in the graph, μ is the number of rings in the molecules, w(u) represents the sum of distances of vertex u from the other vertices.

Definition 2.2 For a connected graph G with the set of vertices $V(G)$, the distance between any two vertices u and v is denoted by $d(u, v)$. The Hyper-Wiener Index is mathematically defined as:

$$WW(G) = \frac{1}{2} \sum_{u,v \subseteq V(G)} (d(u, v) + d^2(u, v)) \tag{3}$$

Definition 2.3 The Bertz complexity index is a topological descriptor that quantifies molecular complexity based on atomic connectivity and bonding patterns. It is commonly expressed as

$$\text{BertzCT} = \sum_i \log_2(\delta_i) + \sum_{(i,j)} \log_2(b_{ij}) \tag{4}$$

where δ_i is the connectivity of atom i and b_{ij} represents the bond order between atoms i and j.

Definition 2.4 McGowan volume is a fragment-based descriptor that estimates the van der Waals volume of a molecule and serves as an indicator of molecular size. It is calculated as

$$V_x = \sum_i V_i - 6.56 \times 10^{-3} N_b \tag{5}$$

where V_i denotes the atomic fragment contribution and N_b is the total number of bonds.

Definition 2.5

Chi1v is a topological descriptor that characterizes molecular branching using valence connectivity information. It is defined as

$$^{1}\chi^{v} = \sum_{(i,j)} \frac{1}{\sqrt{\delta_i^v \delta_j^v}} \tag{6}$$

where δ_i^v and δ_j^v are the valence connectivities of adjacent atoms i and j.

Definition 2.6

FracSP3 represents the fraction of carbon atoms that are sp^3-hybridized and is commonly used as a measure of molecular saturation and three-dimensional character. It is calculated as

$$\text{FracSP3} = \frac{N_{\text{sp}^3}}{N_C} \tag{7}$$

where N_{sp^3} is the number of sp^3-hybridized carbon atoms and N_C is the total number of carbon atoms in the molecule.

Definition 2.7

The Radius of Gyration (RoG) is a geometrical descriptor that measures the spatial distribution of atomic mass about the molecular center of mass. It is given by

$$R_g = \sqrt{\frac{\sum_i m_i r_i^2}{\sum_i m_i}} \tag{8}$$

where m_i is the mass of atom i and r_i is its distance from the molecular center of mass.

Definition 2.8

Total Polar Surface Area (TPSA) is defined as the sum of the surface areas of all polar atoms in a molecule (primarily oxygen and nitrogen, including attached hydrogens).It is calculated as the additive sum of the tabulated surface area contributions of individual polar fragments:

$$\text{TPSA} = \sum_i n_i v_i \tag{9}$$

where:

- n_i is the occurrence count of the i-th polar fragment.
- v_i is the surface contribution value of that fragment in square angstroms (Å^2).

Definition 2.9

The Normalized Principal Moment of Inertia 2 (**NPR2**) is a 3D molecular shape descriptor derived from the principal moments of inertia of a molecule. It is used to characterize molecular geometry, distinguishing between rod-like, disc-like, and spherical shapes. It is given by

$$\text{NPR1} = \frac{I_1}{I_3}, \quad \text{NPR2} = \frac{I_2}{I_3} \tag{10}$$

where:

- I_1, I_2, and I_3 are the principal moments of inertia along the three orthogonal axes.

Definition 2.10

Mol Log P is formally defined as the logarithmic ratio of the solute's concentrations in the octanol and aqueous phases. It is given as:

$$\text{MolLogP} = \log_{10}\left(\frac{[\text{solute}]_{\text{octanol}}}{[\text{solute}]_{\text{water, unearned}}}\right) \tag{11}$$

where:

- $[\text{solute}]_{\text{octanol}}$ is the concentration of the chemical species in the octanol phase.
- $[\text{solute}]_{\text{water}}$ is the concentration of the neutral (un-ionized) species in the aqueous phase.

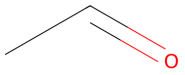
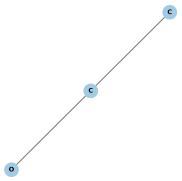
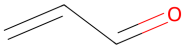
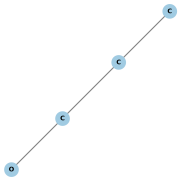
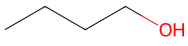
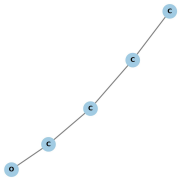
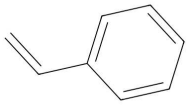
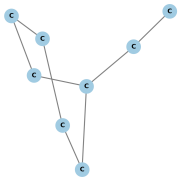
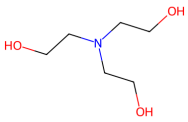
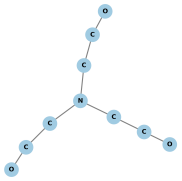
2.2 Molecular Descriptor Generation

In this work, molecular descriptors were calculated using the RDKit [14] and Mordred [15] libraries available in Python. RDKit was used for molecular structure handling, primarily using the SMILES notation system, as well as calculation of some of the molecular graph descriptors, while Mordred was used to calculate a variety of descriptors, including constitutional, electronic, and topological descriptors. These two libraries together help in efficient descriptor generation for large molecular datasets. The specific descriptors that were used in this study can be found in Table 2.

These descriptors were calculated using a RDKit-and-Mordred-based algorithm as given in Algorithm 1. This algorithm describes the process used to convert a molecule represented in SMILES format into a structured set of numerical features suitable for machine learning. The input SMILES string is converted into a suitable MOL object using RDKit. If the MOL object generated is invalid or null, it is discarded immediately. Once a valid MOL object is generated, an initial set of physicochemical descriptors is computed using RDKit, which captures fundamental properties such as molecular weight, hydrophobicity (LogP), polar surface area, and hydrogen bonding characteristics. These values are checked for null or invalid values, in which case the MOL object is discarded.

To enrich molecular representation, additional structural and topological descriptors are extracted using the Mordred library, which provides a broader and more detailed characterization of molecular structure. These descriptors are similarly checked for validity, with missing or non-numeric values handled appropriately to maintain consistency. Finally, the descriptors from RDKit and Mordred are combined together into a single feature vector. with any remaining invalid entries removed. This final vector serves as the input for downstream machine learning models used for predicting thermochemical properties.

Table 1 Molecular Structures and Corresponding Topological Graph Representations of Selected Target Compounds

Compound Name	Chemical Structure	Graph Data Structure
Acetaldehyde		
Prop-2-enal		
Butan-1-ol		
Styrene		
Triethanolamine		

These descriptors were selected using an embedded feature selection approach based on Random Forest approach. This approach enables us to select the best features, captures non-linear relationships between molecular descriptors and thermochemical properties and is computationally inexpensive, since the model is trained once and the features are selected. Further explanation of this approach is given by Algorithm 2

Table 2 Categorization of descriptors based on the information that they encode.

Descriptor Category	Descriptors
Constitutional Descriptors	C Count, H Count, N Count, O Count, H/C Ratio, N/C Ratio, FracSP3
Topological Descriptors	BertzCT, Balaban Index, Hyper-Wiener Index, Chi1v
Electronic Descriptors	Maximum Absolute Partial Charge, CHGHI, CHGLO
Physicochemical Descriptors	MolLogP, TPSA, NumHDonors, NumRotatableBonds, NumAromaticRings

Using the above algorithm, a total of 23 features have been selected for this work. Among these 23 features, 10 features as defined in the Graph Theoretic Preliminaries section (Section 2 Subsection A) were calculated and is presented in Table 3. The selected descriptors were chosen to represent the chemical and structural characteristics of the compounds, including topological, electronic, constitutional, and physicochemical information relevant to thermochemical property prediction.

Fig 1 presents the Pearson correlation heatmap between the various descriptors and targets. Several descriptors, such as Bertz CT, Mol Log P, McGowan Volume and Radius of Gyration (ROG) show a strong linear relationship with several of the targets. Many other features show a moderate linear relationship with the targets. The Pearson correlation heatmap reveals that multiple descriptors exhibit non-negligible correlations with the target properties, indicating the existence of underlying descriptor-property relationships that can be exploited by machine learning models.

Fig 1 also presents the Mutual Information heatmap between the various descriptors, The main advantage of Mutual Information over the Pearson Correlation is that it can capture non-linear relationship between the molecular descriptors and the targets. Descriptors such as C count, Chi1v, MinEState and MaxEState, which showed moderate Pearson correlation scores across all targets, show relatively high Mutual Information scores across all targets, indicating the presence of various non-linear relationships in the dataset. This observation justifies the usage of an ANN to extract such non-linear trends for improved predictive performance.

In order to validate the feature selection that was derived by the Random Forest based Embedded Feature Selection algorithm, SHapely Additive Explainers (SHAP) plots were generated. The results and their interpretation are presented in Section 6.

Table 3 Features calculated using RDKit and Mordred Algorithm

Compound Name	Balaban Index	Hyper Wiener Index	Bertz CT	McGowan Vol.	Chi1v	FracSP3	ROG	TPSA	NPR2	Mol Log P
methanamine	1.0000	1	0.00	36.86	0.5774	1.00	0.8919	26.02	0.9601	-0.4251
oxirane	2.2500	3	4.75	39.59	1.0774	1.00	0.9506	12.53	0.6421	0.0166
ethanol	1.6330	5	0.00	51.95	1.0233	1.00	1.1829	20.23	0.8722	-0.0014
N-methylmethanamine	1.6330	5	0.00	54.16	1.0000	1.00	1.1961	12.03	0.8792	-0.1644
methoxymethane	1.6330	5	0.00	51.95	0.8165	1.00	1.1463	9.23	0.8862	0.2626
acetaldehyde	2.1875	5	7.51	49.31	0.8131	0.50	1.1386	17.07	0.8941	0.2052
2-methyloxirane	2.1711	10	19.22	56.89	1.5100	1.00	1.2716	12.53	0.8927	0.4051
N,N-dimethylmethanamine	2.3238	12	4.75	71.46	1.3416	1.00	1.3624	3.24	0.5681	0.1778
propan-2-ol	2.3238	12	4.75	69.25	1.4129	1.00	1.3685	20.23	0.6057	0.3871
2-aminoethanol	1.9747	15	2.00	62.95	1.2245	1.00	1.3908	46.25	0.8313	-1.0626
propanal	2.2968	15	14.00	66.61	1.3511	0.67	1.3923	17.07	0.7783	0.5953
prop-2-en-1-ol	2.2968	15	14.00	66.61	1.1327	0.33	1.4656	20.23	0.9738	0.1647
prop-2-enal	2.7321	15	21.02	63.97	0.9773	0.00	1.3538	17.07	0.7811	0.3713
prop-2-yn-1-ol	2.4331	15	29.79	63.97	0.9585	0.33	1.4567	20.23	0.8975	-0.3881
propan-1-ol	1.9747	15	2.00	69.25	1.5233	1.00	1.4396	20.23	0.8547	0.3887
furan	3.1250	20	56.52	58.01	1.4714	0.00	1.2602	13.14	0.5113	1.2796
oxolane	2.0833	20	11.61	74.19	2.0774	1.00	1.3595	9.23	0.5749	0.7968
pyrrolidine	2.0833	20	11.61	76.40	2.2071	1.00	1.3836	12.03	0.5840	0.3698
2-methylpropan-2-ol	3.0237	22	15.51	86.55	1.7236	1.00	1.4854	20.23	0.9578	0.7772
2-ethyloxirane	1.9989	26	27.02	74.19	2.0480	1.00	1.5045	12.53	0.9187	0.7952
2-ethenyloxirane	2.2145	26	42.00	71.55	1.6742	0.50	1.5751	12.53	0.9061	0.5712
2-methylprop-2-enal	3.2048	28	51.06	81.27	1.3779	0.25	1.5667	17.07	0.6777	0.7614
but-3-en-2-one	3.2048	28	51.06	81.27	1.4010	0.25	1.5770	17.07	0.7335	0.7614
2-methylpropanal	2.8257	28	27.02	83.91	1.7237	0.75	1.5802	17.07	0.7318	0.8413
butan-2-one	2.8474	28	35.30	83.91	1.7648	0.75	1.5921	17.07	0.7837	0.9854
butan-2-ol	2.5395	28	14.00	86.55	1.9509	1.00	1.5732	20.23	0.8155	0.7772
2-methylpropan-1-ol	2.5395	28	14.00	86.55	1.8792	1.00	1.6473	20.23	0.7591	0.6347
(E)-but-2-enal	2.9244	35	42.00	81.27	1.4797	0.25	1.8648	17.07	0.9463	0.7614
butanenitrile	2.4837	35	38.49	83.48	1.7843	0.75	1.7033	23.79	0.8511	1.3101
butanal	2.4017	35	21.22	83.91	1.8511	0.75	1.7539	17.07	0.9315	0.9854

Algorithm 1 Molecular Feature Extraction using RDKit and Mordred

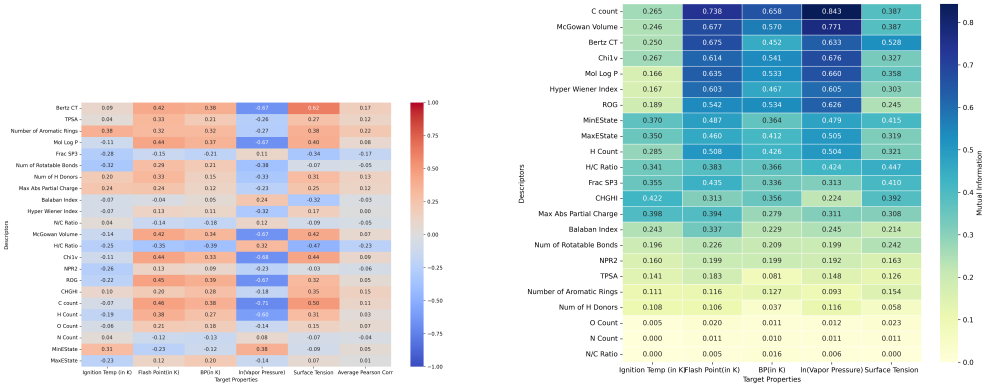
Input: SMILES string S
Output: Feature vector F
Convert SMILES string S into RDKit molecular object M
if M is invalid **then**
    Discard molecule and exit procedure
end if
Initialize feature vector $F \leftarrow [\,]$
Compute physicochemical descriptors of M using RDKit
(e.g., molecular weight, LogP, TPSA, H-bond donors, H-bond acceptors)
if any descriptor value is missing or invalid **then**
    Discard molecule and exit procedure
end if
Append RDKit descriptor values to F
Compute extended molecular descriptors of M using Mordred library
(e.g., structural and topological descriptors such as McGowan volume, CHGHI, etc.)
if any descriptor value is missing or invalid **then**
    Handle missing values or discard molecule based on threshold criteria
end if
Append Mordred descriptor values to F
Remove any remaining non-numeric or invalid values from F
return F

Algorithm 2 Random Forest-Based Embedded Feature Selection

Require: Dataset X, target vector y
Ensure: Selected feature set X_{sel}
Train Random Forest Regressor on (X, y)
Compute feature importance scores $I(x_i)$ for all features
Compute threshold $T = \text{median}(I)$
Initialize $X_{sel} = \emptyset$
for each feature $x_i \in X$ **do**
    if $I(x_i) \geq T$ **then**
        Add x_i to X_{sel}
    end if
end for
return X_{sel}

2.3 Illustration of Descriptor Computation using Ethanol

To better understand how molecular descriptors are derived from SMILES representations, ethanol is taken as a simple illustrative example. The SMILES notation for ethanol is given by `CCO`, which corresponds to a linear molecular structure consisting of two carbon atoms and one oxygen atom.



(a) Pearson Correlation Heatmap (b) Mutual Information Heatmap

Fig. 1 Comparison of Pearson correlation and Mutual Information heatmaps.

Once the SMILES string is converted into a molecular graph using RDKit, several physicochemical and topological descriptors can be computed. Among the selected descriptors, the calculations of some of these descriptors are given below:

2.3.1 Topological Polar Surface Area (TPSA)

TPSA is calculated based on the contribution of polar fragments such as oxygen atoms. For ethanol, only one hydroxyl group contributes to the polar surface area.

$$TPSA = \sum_i A_i \tag{12}$$

For ethanol:

$$TPSA = A_{(OH)} \approx 20.23\,\text{Å}^2$$

This indicates that ethanol has a relatively low polar surface area, which is consistent with its small size and moderate polarity.

2.3.2 McGowan Volume

McGowan volume is computed using atomic fragment contributions, representing the effective molecular size.

$$V_w = \sum_i v_i \tag{13}$$

For ethanol:

$$V_w = v_C + v_C + v_O + v_H$$

Approximating atomic contributions:

$$V_w \approx (16.35 \times 2) + 12.43 + 8.0 \approx 53.1\,\text{(arbitrary units)}$$

This value reflects the compact size of the ethanol molecule.

2.3.3 Hyper Wiener Index

The Hyper Wiener Index is a graph-based descriptor derived from shortest path distances between atoms in the molecular graph.

For ethanol, the distance matrix is:

$$D = \begin{bmatrix} 0 & 1 & 2 \\ 1 & 0 & 1 \\ 2 & 1 & 0 \end{bmatrix}$$

The Wiener-type contribution is computed as:

$$W = \frac{1}{2} \sum_{i,j} d(i,j) \tag{14}$$

$$W = \frac{1}{2}(1+2+1+1+2+1) = \frac{8}{2} = 4$$

The Hyper Wiener Index extends this idea by including squared distances:

$$HW = \frac{1}{2} \sum_{i,j} (d(i,j) + d(i,j)^2) \tag{15}$$

For ethanol:

$$HW = \frac{1}{2}[(1+1) + (2+4) + (1+1) + (1+1) + (2+4) + (1+1)] = 9$$

Thus, ethanol exhibits a low Hyper Wiener Index, consistent with its simple linear structure.

2.3.4 Discussion

This example illustrates how a simple SMILES string is transformed into a molecular graph and further converted into meaningful numerical descriptors. RDKit provides the tools necessary for physicochemical descriptor computation, while Mordred extends this with advanced structural and topological indices. These descriptors collectively form the feature representation, which is used for machine learning-based prediction of thermochemical properties.

3 Architecture and Methods

3.1 Problem Formulation

The prediction of thermochemical properties is formulated as a multi-output regression problem.Let X denote the molecular descriptor space of the chemical compounds and let Y denote the target space corresponding to the thermochemical properties to be predicted.The objective of the model is to find a mapping function f such that

$$f : X \to Y$$

which approximates the relationship between the descriptors and the properties.

3.2 Model Architecture

To model the complex relations between molecular descriptors and the thermochemical properties, a multi-output ANN model is used. The architecture of such a network is as shown in Figure 2.

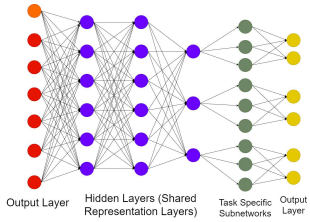


Fig. 2 An illustration of the proposed ANN with a shared representation layer to predict the target variables.

The network consists of an input layer and multiple hidden layers that represent a common learning space for all the properties. This is then followed by a shared hidden layer that acts as a shared representation layer, whose outputs are passed onto task-specific sub-networks for each property. Table 4 shows the hyper-parameters of the Neural Network used. A brief description of each part is given in the following subsections.

Table 4 Architecture of the ANN

Architecture and Hyper-parameters	**Value**
No of inputs (features)	22
No of hidden layers	2
No of neurons in first hidden layer	20
No of neuron in second hidden layer	12
No of heads	3
Learning Rate (α)	0.001
Beta (used in momentum gradient descent) (β)	0.9
Huber Delta (δ)	1.30

3.2.1 Input Layer

The input layer consists of 22 different molecular and chemical descriptors of each compound. These descriptors are selected based on their relevance to the target thermochemical properties to be predicted,

3.2.2 Hidden Layers

Each hidden layer performs a linear transformation, which is followed by a non-linear activation. Given an input X, the transformation is defined as:

$$Z = W \cdot X + b \tag{16}$$

where W denotes the weight matrix and b represents the bias vector.

The output H of the hidden layer is obtained by applying a non-linear activation function $A(\cdot)$:

$$H = A(Z) \tag{17}$$

For this work, we utilize a hybrid hyperbolic tangent (tanh) activation function, defined as:

$$A(z) = \begin{cases} -a + (z+a)(1-\tanh(a))^2 & \text{if } z < -a \\ \tanh(z) & \text{if } |z| \leq a \\ a + (z-a)(1-\tanh(a))^2 & \text{if } z > a \end{cases} \tag{18}$$

where a is a tunable hyperparameter.

This activation function behaves as a standard tanh function within the interval $[-a, a]$, while transitioning to a linear approximation outside this range. This hybrid behaviour helps to mitigate near-zero gradients(commonly referred to as the vanishing gradient problem), while simultaneously allowing for a non-linear representation.

3.2.3 Shared Representation

The final hidden layer of the network acts as a shared representation layer which helps to capture underlying correlations between the various target properties, allowing the model to understand common patterns that exist across multiple outputs. This is beneficial for multi-output models that contain several interdependent targets. Following this, the outputs are passed to separate task specific sub-networks, each sub-network being dedicated to either a single target or a group of thermochemical targets, which are interdependent on each other.

3.2.4 Output Heads

Each task-specific head in the network behaves as a task-specific sub-network, that processes the shared representation to either predict one target property or a group of target properties that are interdependent. Since our study tries to predict different thermochemical properties, some of which are easier to learn than others, this architecture mitigates negative transfer, where easily learnable targets dominate the learning process and adversely affect the prediction of more complex outputs. The results of this output is a set of continuous predictions of different targets.

3.3 Training Procedure

In this section, we discuss the procedures used to train the model, including the loss function used and how the network back-propagates.

3.3.1 Loss Function

The model is trained using a Huber loss function [16], defined for a single prediction as follows:

$$L_\delta(y, \hat{y}) = \begin{cases} \frac{1}{2}(y - \hat{y})^2 & \text{if } |y - \hat{y}| \leq \delta \\ \delta(|y - \hat{y}| - \frac{1}{2}\delta) & \text{otherwise} \end{cases} \tag{19}$$

where y is the true value, $\hat{y}$ is the predicted value and δ is a threshold hyperparameter. Depending on the magnitude of the prediction error,two behaviors are exhibited by this function:

- If $|y - \hat{y}| \leq \delta$, then the function behaves like Mean Squared Error(MSE).
- If $|y - \hat{y}| > \delta$, then the function behaves like Mean Absolute Error(MAE).

This formulation ensures that the loss function is not influenced by extreme outliers, while also keeping the function differentiable everywhere.

3.3.2 Optimizer

The network parameters are optimized using gradient descent with momentum. This ensures faster convergence and stabilizes the optimization process by incorporating past gradient information.

3.3.3 Early stopping

The model adopts an early stopping approach, training on the dataset until the validation loss ceased improving. This approach allows for the model to find the optimum amount of training epochs, without overfitting to the training set.

3.3.4 Cross-Validation

To evaluate the generalization capability of the proposed model, K-fold cross-validation was employed with $K = 5$. Alternative cross-validation strategies such as Bemis-Murcko scaffold splitting and Butina clustering using Tanimoto distances were considered and empirically evaluated, but were considered unsuitable for this particular dataset.

As seen in Table 5, around 41.26 % of molecules returned null scaffolds due to acyclic structures.This forced nearly half the dataset into a single partition, creating extreme fold imbalance during training.

To overcome this limitation, distance-based Butina clustering was evaluated as an alternative partitioning scheme to group compounds by structural similarity rather than ring frameworks. In order to evaluate Butina clustering, the pairwise Tanimoto similarity was evaluated using Morgan fingerprints of radius 2 and 2048 bits.From this, Tanimoto distances between two compounds A and B was calculated by using the formula :

$$d(A, B) = 1 - S(A, B) \tag{20}$$

where S is the Tanimoto Similarity between the compounds A and B.

From Table 5, the mean pair-wise Tanimoto distance of the dataset was found to be 0.90. This suggests that the target compounds are structurally diverse and widely distributed in chemical space. As a result, Butina distance clustering resulted in severe over-fragmentation into folds containing single compounds, causing fold-size disparities and mini-batch instability under GroupKFold. Consequently, standard 5-fold cross-validation with inside-the-loop feature scaling was selected to maintain uniform fold balance and optimization stability across the chemical space."

Table 5 Diagnostic Evaluation of Candidate Dataset Partitioning Schemes ($N = 412$)

Metric / Parameter	**Value**
Bemis-Murcko Scaffold Partitioning	
Total Compounds	412
Empty Scaffolds (Acyclic)	170 (41.26%)
Valid Ring Scaffolds	242 (58.74%)
Butina Distance Clustering	
Distance Cutoff (d_{cutoff})	0.40
Pairwise Tanimoto Distance	0.90 ± 0.10
Total Clusters Formed	260
Singletons (1 mol / cluster)	204 (78.50%)
Average Cluster Size	1.58 molecules
Largest Cluster Size	17 molecules
Final Validation Protocol	
Selected Strategy	Standard 5-Fold CV
Fold Sizing	~ 82 mol / fold

3.4 Evaluation Metrics

The model's predictions are evaluated using standard regression metrics, including the coefficient of determination R^2 and Mean Absolute Error(MAE)

The coefficient of determination is defined as:

$$R^2 = 1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \bar{y})^2} \tag{21}$$

where y is the true values, $\hat{y}$ represents the predicted values and $\bar{y}$ represents the mean of the true values.The R^2 score indicates how well the model can explain the variance in the data.

The Mean Absolute Error (MAE) is defined as:

$$\text{MAE} = \frac{1}{n} \sum |y - \hat{y}| \tag{22}$$

where n is the number of samples. MAE measures the average magnitude of prediction errors.

4 Dataset

4.1 Dataset Collection

The dataset used in this study is compiled from multiple publicly available sources including PubChem [17], NIST Chemistry Webbook [18] and ChemSpider [19], along with the `thermos` library [20] in Python to obtain target information for density, viscosity, surface tension and vapor pressure. A subset of the compiled data from these sources is presented in Table 6

The target variables considered in this study include vapor pressure, surface tension, flash point, boiling point, and ignition temperature.

The dataset used for training the model consists of 413 compounds. Each compound is represented by 23 molecular descriptors, together with five thermochemical properties including vapor pressure, flash point, boiling point, surface tension and ignition temperature.

The dataset consists of diverse classes of compounds, including alcohols, aldehydes, ketones, aromatic compounds, aliphatic cyclic compounds and alkanes.

4.2 Data Preprocessing

Prior to training, the dataset was preprocessed to remove inconsistencies and handle missing values. Additionally, descriptors containing incomplete data or exhibiting high collinearity were eliminated to improve model stability and reduce redundancy. For targets which showed high skewness and variance, such as Vapor Pressure, a natural logarithmic transformation is applied to improve numerical stability and facilitate more effective model training.

Subsequently, both input features and target variables are scaled using Robust Scaling, which is defined as follows:

$$X_{\text{scaled}} = \frac{X - \text{median}(X)}{\text{IQR}(X)} \tag{23}$$

where X is the original data point, X_{scaled} is the scaled data point, and IQR is the Inter-quartile Range defined as the difference between 75th percentile and 25th percentile in the dataset. The main advantage of Robust Scaling is its invariance to outlier data-points, making the training proecss easier as well as avoid overfitting.

5 Results and Discussions

In this section, the results obtained from the model are presented and evaluated. Table 7 shows some of the actual values of the physical properties against the predicted properties by our ANN model.

5.1 Experimental Setup

The proposed model has been trained on the dataset described in the previous section. The data is divided into training, testing and validation sets, with the validation

Table 6 Subset of the data used in this work from the listed sources

Compound Name	ln(Vapor Pressure)	Surface Tension	Flash Point (K)	Boiling Point (K)	Ignition Temp (K)	Density	Viscosity
methanamine	12.7609	0.0190	263.15	266.65	703.15	1.2694	0.0000
oxirane	11.8913	0.0236	255.37	283.87	702.04	0.8222	0.0095
ethanol	8.5816	0.0220	285.93	351.65	638.15	0.7900	0.0011
N-methylmethanamine	5.3111	0.0162	266.48	280.54	675.37	1.8428	0.0000
methoxymethane	5.6479	0.0160	232.15	249.55	623.15	0.7240	0.0001
acetaldehyde	10.8842	0.0212	233.15	294.26	458.15	0.7800	0.0003
2-methyloxirane	10.9908	0.0245	235.93	307.38	738.15	0.8590	0.0003
N,N-dimethylmethanamine	5.3405	0.0135	260.93	276.04	463.15	2.4161	0.0000
propan-2-ol	8.6621	0.0209	284.82	355.65	672.04	0.7819	0.0021
2-aminoethanol	1.7566	0.0485	366.48	443.15	683.15	1.0124	0.0184
propanal	10.3524	0.0234	263.71	322.04	480.37	0.8050	0.0003
prop-2-en-1-ol	8.0625	0.0257	295.15	369.82	715.93	0.8540	0.0012
prop-2-enal	9.8033	0.0240	255.37	325.37	507.04	0.8389	0.0003
prop-2-yn-1-ol	7.3777	0.0354	305.93	386.76	638.15	0.9485	0.0017
propan-1-ol	8.6990	0.0238	298.15	370.37	644.26	0.8030	0.0023
furan	11.0933	0.0238	237.59	305.37	663.15	0.9371	0.0004
oxolane	9.6290	0.0264	258.71	339.26	594.26	0.8880	0.0005
pyrrolidine	9.0521	0.0292	275.93	359.71	618.00	0.8389	0.0007
2-methylpropan-2-ol	8.3268	0.0200	284.26	355.37	753.15	0.7800	0.0035
2-ethyloxirane	9.9679	0.0235	260.93	335.93	788.15	0.8260	0.0004
2-ethenyloxirane	9.9297	0.0291	227.59	422.15	703.15	0.8700	0.0004
2-methylprop-2-enal	9.9362	0.0770	274.82	341.55	568.15	0.8490	0.0005
but-3-en-2-one	9.3471	0.0290	266.48	354.82	763.71	0.8407	0.0008
2-methylpropanal	10.0286	0.0220	293.15	337.04	469.26	0.7910	0.0005
butan-2-one	9.2430	0.0240	269.82	352.76	789.26	0.8060	0.0004
butan-2-ol	7.3777	0.0230	297.04	372.59	1036.15	0.8100	0.0042
2-methylpropan-1-ol	7.1954	0.0230	300.93	380.37	699.82	0.8020	0.0040
(E)-but-2-enal	7.8372	0.0366	285.93	377.04	505.37	0.8530	0.0004
butanenitrile	7.8632	0.0266	289.82	390.93	760.93	0.7863	0.0005
butanal	9.4091	0.0299	266.48	348.15	503.15	0.8030	0.0005

Table 7 Actual vs Predicted Values of the ANN Model for a few compounds

Physical Prop	**Exp Value**	**3,7-dimethyloct-6-enal**	**4-ethenylphenol**	**3-methylbut-2-enal**	**(Z)-hex-3-en-1-ol**
ln(Vapor Pressure)	Actual	7.837	0.0927	3.7497	8.2759
	Predicted	6.4034	1.9817	5.1106	6.7948
Surface Tension	Actual	0.0365	0.0337	0.0289	0.0287
	Predicted	0.0292	0.0344	0.0282	0.0447
Flash Point	Actual	285.92	373.78	352.32	324.50
	Predicted	289.55	392.25	351.77	304.73
Boiling Point	Actual	377.03	547.15	491.04	311.15
	Predicted	389.70	503.24	456.52	392.85
Ignition Temp	Actual	505.37	743.00	635.23	644.59
	Predicted	522.73	666.65	634.47	594.98

set being used in early stopping. The network is trained using gradient descent with momentum as the optimization strategy. In order to avoid overfitting to the training data, early stopping has been used, which terminates training when the validation error starts increasing.

A second experiment was conducted to investigate if inclusion of related but distinct target properties introduces multi-task interference. The model was trained on the same dataset and the same features,along with additional targets such as density, viscosity and latent heat of vaporization. These properties are related to the primary target properties and share several underlying molecular determinants; however, their prediction may require information that is not fully captured by the selected descriptor set. For example, accurate prediction of the density of a molecule, benefits from descriptors that can encode the three dimensional folding and intramolecular bonding within the molecule,along with the electronic and molecular constitution. The model is trained to predict these targets, along with our key targets, and the results of the prediction of the key targets, with and without the aforementioned targets was obtained and shown in Table 10

5.2 Results

The following section presents the results obtained from the evaluation of the ANN Model, along with comparative testing of the ANN with XGBRegresor and SVR. Table 8 presents the performance of the model across all target properties.

Table 8 Model performance across various thermochemical properties

Property	**Test R^2**	**Std(R^2)**	**MAE**	**Std(MAE)**
ln (Vapor Pressure)	0.6438	0.1120	1.8317	0.2259
Surface Tension	0.4466	0.2236	0.0048	0.0009
Boiling Point	0.3938	0.2458	40.1436	5.7333
Ignition Temperature	0.2872	0.1405	66.2614	5.5993
Flash Point	0.6068	0.1150	19.8399	1.9583

As part of the model evaluation, the predictive performance of the proposed ANN model and XGBRegressor are compared, using identical train-test splits for each target property. The results are presented in Table 10

As part of the model performance evaluation, a graph of actual vs predicted values for each of the 5 target properties has been plotted and is presented as shown in Figure 3

5.3 Discussion

The comparative results across the single output ANN, multi-output ANN, XGBRegressor, and SVR show that no single model consistently performs best for all thermochemical properties. Instead, the prediction accuracy depends strongly on

Table 9 Comparison of performance between a model with a multi-target interference and one without interference

Target Property	**ANN R^2 with Multi-Target Interference**	**ANN R^2 without Multi-Target**
ln(Vapor Pressure)	0.6525	0.6726
Surface Tension	0.4426	0.4466
Boiling Point	0.4240	0.3938
Ignition Temperature	0.3026	0.2872
Flash Point	0.6257	0.6068

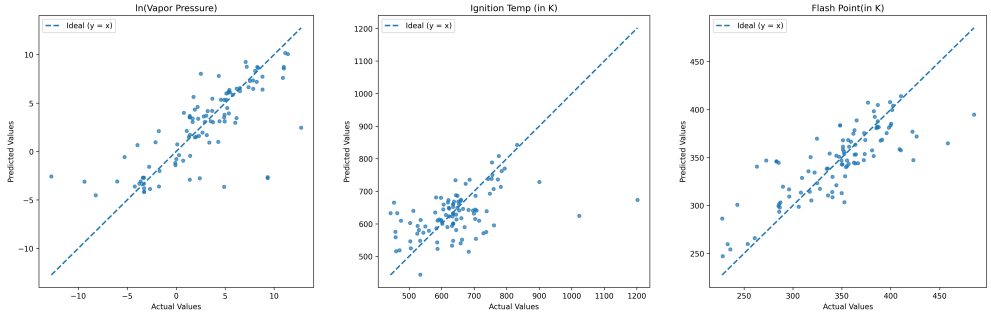
Table 10 Comparison of Test R^2 values of model and XGBRegressor. Note that this experiment has been conducted with identical train-test splits for each target property.

Target Property	**ANN Single Test R^2**	**ANN Multi Test R^2**	**XGBRegressor Test R^2**	**SVR Test R^2**
ln(Vapor Pressure)	0.6388	0.6438	0.6857	0.4443
Surface Tension	0.4240	0.4466	0.5167	0.9624
Boiling Point	0.4030	0.3938	0.5905	0.6022
Ignition Temperature	0.3152	0.2872	0.4706	0.1674
Flash Point	0.5734	0.6068	0.8314	0.6755

the nature of the target property and how effectively the underlying descriptor relationships are captured by each learning algorithm.

For Surface Tension, the SVR model significantly outperforms both ANN and XGBRegressor, achieving a test R^2 value of 0.9624. In comparison, the ANN and XGBRegressor achieve lower values of 0.4466 and 0.5167 respectively. This indicates that the relationship between the selected molecular descriptors and surface tension may follow a smoother nonlinear trend that is more effectively captured by kernel-based methods such as SVR. The comparatively lower ANN performance also suggests that the shared multi-task interference may not let the ANN to fully capture the trend for Surface Tension.

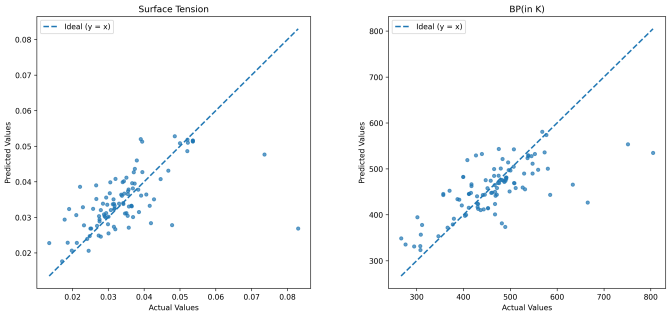
For Flash Point and Ignition Temperature, XGBRegressor achieves a much superior predictive performance. In particular, XGBRegressor achieves a high test R^2 value of 0.8314, outperforming both ANN and SVR. Similar trends are observed for Ignition Temperature, although the prediction accuracy across all three models was not



(a) ln(Vapor Pressure Actual vs Predicted)

(b) Ignition Temperature Actual vs Predicted

(c) Flash Point Actual vs Predicted



(d) Surface Tension Actual vs Predicted

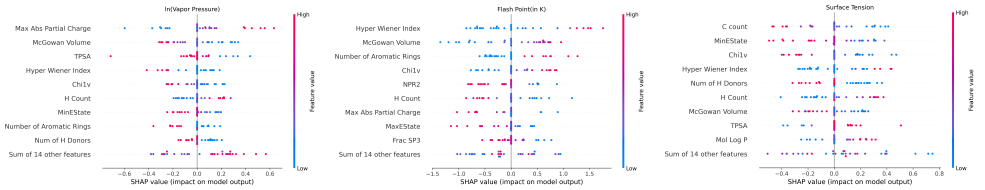
(e) Boiling Point Actual vs Predicted

Fig. 3 Predicted vs Actual values for selected thermochemical properties.

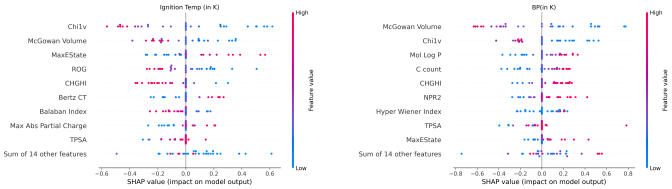
as strong as compared to Flash Point and Surface Tension. These results suggest that tree-based ensemble methods are more effective at capturing the localized and potentially threshold-dependent relationships present in these properties. It is also worth noting that Ignition Temperature depends on chemical kinetics, molecular constitution, electronic interactions and oxygen content. As such, a generalized prediction of such a property is very difficult.

For Vapor Pressure, the ANN and XGBRegressor models produce comparable results, with XGBRegressor showing a slight advantage. This suggests that the selected descriptor set contains meaningful information relevant to volatility-related behavior, although the relationship remains moderately complex. In the case of Boiling Point prediction, both SVR and XGBRegressor outperform the ANN, indicating that the current neural network architecture may not fully capture the descriptor interactions associated with this property.

The comparison between the single-output and multi-output ANN models shows that the differences in predictive performance are generally small across all target properties. The multi-output model achieved slightly higher R^2 values for vapor pressure, surface tension, and flash point, while showing a modest decrease in performance for boiling point and ignition temperature.



(a) SHAP plot of Vapor Pressure (b) SHAP plot of Flash Point (c) SHAP plot of Surface Tension



(d) SHAP plot of Ignition Temperature (e) SHAP plot of Boiling Point

Fig. 4 SHAP plots of various thermochemical properties

These results demonstrate that while the shared hidden representations within the multi-output ANN successfully captured underlying molecular patterns common across multiple thermochemical targets, the joint optimization task introduced trade-offs. The presence of minor negative transfer effects, combined with the inherent sample complexity of deep neural networks, limited the ANN's overall predictive ceiling on a dataset of this scale ($N = 413$). In contrast, SVR and XGBRegressor effectively leveraged tree-based recursive partitioning to capture non-linear, high-order descriptor interactions without requiring large sample sizes or suffering from gradient competition across targets. Overall, these findings indicate that while multi-output ANNs offer an elegant framework for unified property representation, SVR models and tree-based ensemble models like XGBoost provide superior predictive accuracy for moderate-sized tabular QSPR datasets.

Overall, the results demonstrate that thermochemical property prediction is highly property-dependent, and different machine learning approaches exhibit strengths under different modeling conditions. While the ANN benefits from shared feature learning across multiple targets, the results also indicate that such shared representations are not uniformly effective for every property. Depending on the physical constraints upon which the target properties are dependent on, shared learning may lead to multi-task interference, leading to less generalized predictions. This observation leads to the conclusion that multi-target shared representation is a powerful method of predicting key targets, provided that those targets use the same feature space and representations. The strong performance of SVR for Surface Tension and XGBRegressor for several combustion-related properties suggests that property-specific modeling strategies may be necessary for achieving consistently high predictive accuracy.

Additionally, the variation in predictive performance across models suggests that the current descriptor space may not fully represent all relevant chemical and physical effects governing the investigated properties. Therefore, the incorporation of more physically meaningful descriptors, such as quantum chemical or three-dimensional molecular descriptors, may further improve model generalization and predictive capability across diverse chemical classes.

Along with the aforementioned improvement in feature engineering, the following observations also suggest that the utilization of larger datasets may improve the generalization ability of the model and enable the learning of stable and consistent relationships between the molecular descriptors and thermochemical properties.

As detailed in Section 5 part A, a second experiment to identify multi-target interference was conducted and the results are shown in Table 9.As can be seen from the table, the results of this experiment are mixed. While some targets' predictive performance improved, some others' degraded, albeit the magnitude in the change is not substantial.This suggests that the shared representation layer was able to capture molecular patterns that are beneficial across multiple prediction tasks, while introducing only limited interference between them. The absence of significant performance degradation indicates that the inclusion of additional thermochemical properties did not substantially affect the model's ability to predict the primary target properties.This observation supports the premise that a shared representation model can extract information about different targets, even when features explicit to those targets are absent.

6 Interpretation

To better understand how the model makes its predictions, SHAP (SHapley Additive exPlanations) analysis is used to examine the contribution of individual molecular descriptors to each target property.

Figures 4 represents the SHAP summary (beeswarm) plots for selected thermochemical properties. Each point in the plot represents the contribution of a feature for a given sample, allowing both the magnitude and distribution of feature effects to be observed.

For properties such as Vapor Pressure, a few descriptors, such as Carbon Count, Hyper Wiener Index, TPSA and H/C Ratio consistently exhibit larger SHAP values across various samples. In particular, the Carbon count shows the highest overall influence, where larger carbon counts generally contribute negatively toward vapor pressure prediction. This trend is chemically consistent, as larger molecules typically exhibit lower volatility and reduced vapor pressure due to stronger intermolecular interactions. Additionally, polarity-related descriptors such as TPSA exhibit moderate influence on the prediction behavior, indicating that intermolecular interaction effects are also partially captured by the model. Vapor Pressure is highly dependent on both the molecular composition as well as the intermolecular forces that act between them.

In contrast, for properties such as Surface Tension, the SHAP values are largely dominated by features like Carbon Count, Hyper Wiener Index and H/C Ratio. The dependence of Surface Tension on the carbon count is consistent with the increased

cohesive intermolecular forces typically observed in larger organic molecules. In addition, higher values of the H/C ratio and Mol Log P tend to produce more negative SHAP contributions in several samples, suggesting that increased hydrocarbon-like character and hydrophobicity may reduce predicted surface tension within the studied chemical space. The SHAP plot indicates that the target property is dominated by molecular size, polarity, and intermolecular interaction characteristics.

In addition, the SHAP plot for Flash Point indicates that descriptors related to molecular size, polarity, and structural complexity contribute most strongly to the model predictions. The most influential descriptor is C count, where higher carbon counts (red points) generally correspond to positive SHAP values, indicating an increase in predicted flash point. This trend is chemically reasonable, since molecules with longer carbon chains typically exhibit lower volatility and therefore require higher temperatures for ignition. The H/C ratio exhibits an opposite trend, where higher values tend to contribute negatively toward flash point prediction. This may reflect the influence of increased hydrogen saturation and lower molecular complexity, which are often associated with more volatile hydrocarbon structures.

Overall, the SHAP beeswarm plots provide a more detailed view of feature importance and highlight the role of descriptor relevance in determining model performance.

7 Conclusion

In this work, a comparative evaluation of machine learning models—including a multi-output Artificial Neural Network (ANN), XGBRegressor, and Support Vector Regression (SVR)—was conducted to predict key thermochemical properties of rocket fuels using RDKit-derived molecular descriptors. The multi-output ANN incorporated a shared-layer architecture designed to capture common underlying structure–property relationships across interdependent thermochemical targets simultaneously.The experimental results demonstrated that predictive performance is strongly target-dependent, with SVR and tree-based ensemble learning via XGBRegressor providing superior overall accuracy across several key properties. While the multi-output ANN successfully learned joint representations, its predictive capacity was constrained by sample complexity limits on the moderate-sized dataset ($N = 413$) and minor multi-target interference (negative transfer). Interpretability analysis using SHAP values further confirmed that topological indices, molecular size, and polarity-related descriptors serve as the primary drivers of model predictions across all architectures.

Overall, this study highlights that while multi-output neural frameworks offer an efficient paradigm for simultaneous parameter learning, non-parametric and tree-based ensemble models remain the most effective approach for moderate-sized tabular QSPR datasets. Future work will focus on expanding the dataset size, exploring domain-specific regularized loss functions to mitigate negative transfer in multi-task networks, and utilizing complex descriptors that capture more complex behaviors of chemical molecules.

References

[1] 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 **304**, 121437 (2021) https://doi.org/10.1016/j.fuel.2021.121437

[2] Li, R., Herreros, J.M., Tsolakis, A., Yang, W.: Integrated machine learning-quantitative structure property relationship (ml-qspr) and chemical kinetics for high throughput fuel screening toward internal combustion engine. Fuel **307**, 121908 (2022) https://doi.org/10.1016/j.fuel.2021.121908

[3] Kehimkar, B., Hoggard, J.C., Marney, L.C., Billingsley, M.C., Fraga, C.G., Bruno, T.J., Synovec, R.E.: Correlation of rocket propulsion fuel properties with chemical composition using comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry followed by partial least squares regression analysis. Journal of Chromatography A **1327**, 132–140 (2014) https://doi.org/10.1016/j.chroma.2013.12.060

[4] Dubiniak, A., Egazar'yants, S., Maximov, A.: Application of machine learning methods to the prediction of physicochemical properties of potential bio-based oxygenated fuel additives. Chemical Papers (2026) https://doi.org/10.1007/s11696-026-04914-2

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

[6] Shafii Abubakar, M., Ojonugwa, E., Sanusi, R.A., Hassan Ibrahim, A., Olalekan Aremu, K.: Predicting antibacterial drugs properties using graph topological indices and machine learning. IEEE Access **12**, 181420–181435 (2024) https://doi.org/10.1109/ACCESS.2024.3503760

[7] Almohanna, N., Alotaibi, A.M., Ahmad, A., Alhulwah, K., Azeem, M.: Regression models in qspr analysis for blood cancer treatment: A connection number approach. IEEE Access **13**, 153420–153435 (2025) https://doi.org/10.1109/ACCESS.2025.3602919

[8] Hayat, S., Mahadi, H., Alanazi, S.J.F., Wang, S.: Predictive potential of eigenvalues-based graphical indices for determining thermodynamic properties of polycyclic aromatic hydrocarbons with applications to polyacenes. Computational Materials Science **238**, 112944 (2024) https://doi.org/10.1016/j.commatsci.2024.112944

[9] Bojović, V., Lučić, B., Bešlo, D., Skala, K., Trinajstić, N.: Calculation of

topological molecular descriptors based on degrees of vertices. In: 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 266–269 (2019). https://doi.org/10.23919/MIPRO.2019.8757128

[10] Almasi, M., Neyband, R.S.: A thermodynamically-guided svr framework for density of propyl ethanoate + 2-alkanol mixtures: Beyond traditional equations of state. Results in Engineering, 111172 (2026) https://doi.org/10.1016/j.rineng.2026.111172

[11] Wang, D., Yuan, Y., Duan, S., Liu, R., Gu, S., Zhao, S., Liu, L., Xu, J.: Qspr study on melting point of carbocyclic nitroaromatic compounds by multiple linear regression and artificial neural network. Chemometrics and Intelligent Laboratory Systems **143**, 7–15 (2015) https://doi.org/10.1016/j.chemolab.2015.02.009

[12] Ghezelahmad, S.K., Rouhani, M.: A python-powered qspr framework for benzodiazepine analysis using topological indices. Results in Chemistry **26**, 103280 (2026) https://doi.org/10.1016/j.rechem.2026.103280

[13] Eraqi, B.A., Khizbullin, D., Nagaraja, S.S., Sarathy, S.M.: Molecular property prediction in the ultra-low data regime. Communications Chemistry **8**(1), 201 (2025) https://doi.org/10.1038/s42004-025-01592-1

[14] Landrum, G.: RDKit: Open-source cheminformatics. https://www.rdkit.org. Accessed: 2026-08-08. https://doi.org/10.5281/zenodo.591637

[15] Moriwaki, H., Tian, Y.-S., Kawashita, N., Takagi, T.: Mordred: a molecular descriptor calculator. Journal of Cheminformatics **10**(1), 4 (2018) https://doi.org/10.1186/s13321-018-0258-y

[16] Huber, P.J.: Robust estimation of a location parameter. The Annals of Mathematical Statistics **35**(1), 73–101 (1964)

[17] Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B.A., Thiessen, P.A., Yu, B., Zaslavsky, L., Zhang, J., Bolton, E.E.: PubChem 2025 update. Nucleic Acids Research **53**(D1), 1516–1525 (2025) https://doi.org/10.1093/nar/gkae1059

[18] Linstrom, P.J., Mallard, W.G.: NIST Chemistry WebBook, NIST Standard Reference Database Number 69. National Institute of Standards and Technology, Gaithersburg, MD (2025). https://doi.org/10.18434/T4D303 . https://webbook.nist.gov

[19] Pence, H.E., Williams, A.: ChemSpider: An Online Chemical Information Resource. Journal of Chemical Education **87**(11), 1123–1124 (2010) https://doi.org/10.1021/ed100697w

[20] Bell, C., Contributors: Thermo: Chemical properties component of Chemical Engineering Design Library (ChEDL). https://github.com/CalebBell/thermo (2016–2024)

Author contributions

Conceptualization: Yashwin G Devadiga, Swarna M Patra and Parth Sarathi Panigrahy; Methodology: Yashwin G Devadiga, Swarna M Patra, Parth Sarathi Panigrahy and Ganga Periyasamy; Writing—original draft preparation: Yashwin G Devadiga and Swarna M Patra; Writing—final review and editing: Yashwin G Devadiga, Swarna M Patra and Parth Sarathi Panigrahy; Project administration: Swarna M Patra, Parth Sarathi Panigrahy and K.N. Subramanya;

Funding

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Declarations

Competing interests

The authors have no relevant financial or non-financial interests to disclose.

Data Availability

The datasets used and analyzed during the current study were compiled from publicly available repositories, specifically PubChem (https://pubchem.ncbi.nlm.nih.gov/), NIST Chemistry WebBook(https://webbook.nist.gov/chemistry/) and ChemSpider(https://www.chemspider.com/)

Human and animal rights

This article does not contain any studies with human participants or animals performed by any of the authors.

Additional information

Correspondence and requests for materials should be addressed to Parth Sarathi Panigrahy or Swarna M Patra.