



OPEN Machine and deep learning models for predicting high pressure density of heterocyclic thiophenic compounds based on critical properties

Amir Hossein Sheikhshoei¹ & Ali Khoshshima^{1,2}✉

The multifaceted effects of the presence of thiophenic compounds on the environment are significant and cannot be overlooked. As heterocyclic compounds, thiophene and its derivatives play a significant role in materials science, particularly in the design of organic semiconductors, pharmaceuticals, and advanced polymers. Accurate prediction of their thermophysical properties is critical due to its impact on structural, thermal, and transport properties. This study utilizes state-of-the-art machine learning and deep learning models to predict high-pressure density of seven thiophene derivatives, namely thiophene, 2-methylthiophene, 3-methylthiophene, 2,5-dimethylthiophene, 2-thiophenemethanol, 2-thiophenecarboxaldehyde, and 2-acetylthiophene. The critical properties including critical temperature (T_c), critical pressure (P_c), critical volume (V_c), and acentric factor (ω), together with boiling point (T_b), and molecular weight (M_w) were used as input parameters. Models employed include Decision Tree (DT), Adaptive Boosting Decision Tree (AdaBoost-DT), Light Gradient Boosting Machine (LightGBM), Gradient Boosting (GBoost), TabNet, and Deep Neural Network (DNN). The statistical error evaluation showed that the LightGBM model showed superior performance with an average absolute percent relative error (AAPRE) of 0.0231, a root mean square error of 0.3499, and coefficient of determination (R^2) of 0.9999. The leverage method showed that 99.10 percent of the data was valid. These findings highlight the effectiveness of using critical properties as inputs and underscore the potential of the LightGBM model for reliable high-pressure density prediction of thiophene derivatives. This provides a robust tool for advancing materials science applications, and offers valuable insights for material design under extreme conditions.

Keywords Machine learning, Deep learning, Density, Heterocyclic compounds, Thiophenes, Leverage method

Heterocyclic thiophenic compounds¹, which contain sulfur in a five-membered ring structure, are increasingly significant in environmental studies due to their widespread use in various industrial applications, including pharmaceuticals, agrochemicals, and organic electronics. Their environmental impact is multifaceted, as they can persist in ecosystems, leading to potential toxicity to aquatic life and soil microorganisms. In aquatic ecosystems, the presence of thiophenic compounds, which can enter water bodies through industrial discharge, runoff, and wastewater, often leads to detrimental effects on aquatic life². In soil, thiophenic compounds can inhibit microbial activity and alter soil chemistry. The stability of these compounds makes them resistant to typical biodegradation processes, accumulating in sediments and organic tissues of wildlife. In air, the presence of thiophenic compounds can significantly affect the air environment, and lead to both direct and indirect environmental consequences, primarily through industrial emissions, combustion processes, and the degradation of fossil fuels. Petroleum-derived transportation fuels contain considerable amounts of organic sulfur compounds, including benzothiophene, dibenzothiophene, and thiophene. Combustion of sulfur-rich fuels releases sulfur oxides (SOx), primarily sulfur dioxide (SO₂)³. This colorless, odorless, and corrosive gas poses serious environmental concerns, playing a key role in acid rain formation, the greenhouse effect,

¹School of Petroleum and Chemical Engineering, Hakim Sabzevari University, Sabzevar, Iran. ²Center for Atmospheric Research, University of Oulu, 90014 Oulu, Finland. ✉email: a.khoshshima@hsu.ac.ir

photochemical pollution, and eutrophication^{4–9}. Consequently, understanding and mitigating the properties of sulfur compounds is crucial for reducing their environmental and industrial impact.

Heterocyclic compounds represent a fascinating category of aromatic substances. These compounds represent one of the most extensive and structurally diverse families in organic chemistry, characterized by a broad spectrum of intermolecular interactions. Their remarkable diversity makes them a crucial and complex area of study. Alongside carbon and hydrogen, heterocyclic compounds commonly contain heteroatoms like sulfur, oxygen, and nitrogen^{1,10}. The most prevalent types of heterocyclic compounds include five-membered rings (such as furan, thiophene, dioxolane, imidazole, and pyrrole) and six-membered rings (like morpholine and pyridine), which are commonly found in a variety of sources including plants, herbs, animals, coal, and petroleum. Due to their diverse properties and applications across various fields, thiophene-based materials are found everywhere, including medicine^{11–16}, material sciences^{17–19}, or for use in organic electronic devices and molecular electronics^{20–22}. Thiophene-based compounds, particularly their derivatives, are widely utilized as chemo-sensors. They serve as effective fluorescence signaling promoters for detecting organic acids, metal ions, and cations^{23,24}. Their unique electronic properties and structural diversity render thiophene compounds vital in advanced technologies, such as optoelectronic devices OLEDs, OFETs, OTFTs, OSCs, OLFETs and various sensors^{25–36}. 2-thiophenecarboxaldehyde and 2-Thiophenemethanol find applications in material sciences, nanoparticles, and biotechnology^{37–47}. 2-Acetylthiophene finds applications in food flavoring and the synthesis of drugs for anxiety, inflammation, and parasitic infections, as well as in the production of metal complexes^{48–55}.

To maximize the performance of these chemicals in industrial processes, it is crucial to enhance our understanding of their various physicochemical properties. It is essential to emphasize volumetric properties like density and its related characteristics, including isobaric expansibility and isothermal compressibility, especially under high-temperature and high-pressure conditions. Density serves as a fundamental material property with significant implications for process mechanics and engineering in chemical plants. Furthermore, understanding a compound's density offers valuable insights into its molecular arrangement and packing behavior. Analyzing density variations with temperature or pressure allows for the determination of key parameters such as isothermal compressibility, isobaric expansibility, and internal pressure. However, traditional methods for determining density are often labor-intensive and susceptible to experimental errors.

Equations of state (EoS) and empirical relationships, while widely used for predicting thermophysical properties, often suffer from limitations such as the need for simplifying assumptions, poor accuracy under extreme conditions, and reliance on substance-specific constants that may not be available or accurate for all compounds. These models may also lack the flexibility to capture complex, nonlinear relationships inherent in experimental data, especially for structurally diverse compounds like thiophenes^{56–60}. In contrast, machine learning (ML) and deep learning (DL) models offer several advantages, including the ability to learn directly from data without predefined functional forms, adapt to nonlinear patterns, and generalize across a wide range of conditions and molecular structures. They also enable the integration of diverse input features and provide high predictive accuracy, making them powerful tools for property estimation tasks where traditional models

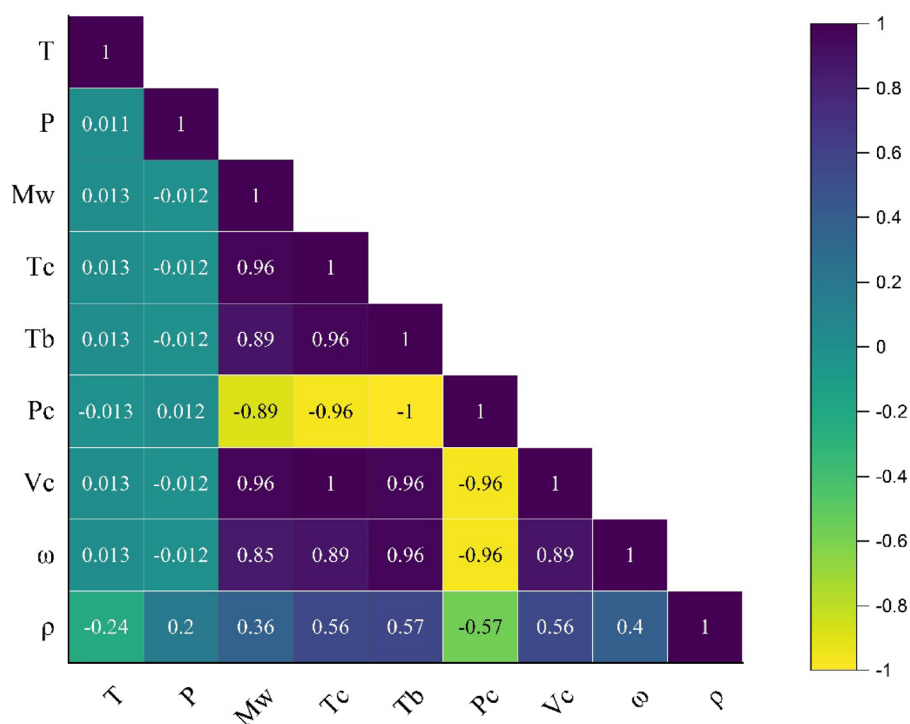


Fig. 1. Effect of input parameters on density.

Model name	Hyperparameter search range	Best hyperparameters
AdaBoost-DT	estimator__max_depth: [3, 5, 7] n_estimators: [50, 100, 200] learning_rate: [0.01, 0.1, 1]	estimator__max_depth: 7 n_estimators: 200 learning_rate: 1
Decision Tree (DT)	max_depth: [3, 5, 7, 10] min_samples_split: [2, 5, 10] min_samples_leaf: [1, 2, 4] max_features: [None, 'sqrt', 'log2']	max_depth: 10 min_samples_split: 2 min_samples_leaf: 1 max_features: None
Gradient Boosting (GBoost)	n_estimators: [50, 100, 200, 300, 400] max_depth: [2, 4, 8, 10, 12, 16] subsample: [0.25, 0.5, 0.75, 1] learning_rate: [0.01, 0.03, 0.05, 0.07, 0.1]	n_estimators: 300 max_depth: 8 subsample: 0.25 learning_rate: 0.03
LightGBM	n_estimators: [300, 500, 800] max_depth: [3, 5, 7] subsample: [0.7, 0.75] learning_rate: [0.05, 0.01] colsample_bytree: [0.3, 0.4] subsample_freq: [1, 2] num_leaves: [5, 8, 10]	n_estimators: 800 max_depth: 7 subsample: 0.75 learning_rate: 0.05 colsample_bytree: 0.4 subsample_freq: 2 num_leaves: 10
Deep Neural Network (DNN)	module__hidden_layers: [1–3] module__neurons: [16, 32, 64] module__activation: ['relu', 'tanh'] optimizer__lr: [0.001, 0.01, 0.1] batch_size: [10, 20, 40] max_epochs: [50, 100]	module__hidden_layers: 1 optimizer__lr: 0.001 module__activation: relu module__neurons: 64 batch_size: 10 max_epochs: 100
TabNet	n_d: [8, 16] n_a: [8, 16] n_steps: [3, 5] gamma: [1.0, 1.3, 1.5, 2.0] lambda_sparse: [1e-3, 1e-2] optimizer_params: ['lr': 1e-4, 'lr': 1e-3, 'lr': 1e-2, 'lr': 2e-2]	n_d: 16 n_a: 16 n_steps: 3 gamma: 1.0 lambda_sparse: 0.001 optimizer_params: 'lr': 0.02

Table 1. Hyperparameter search ranges and optimized values for each machine learning model.

fall short^{61–64}. Recently, Machine learning has been widely used for modelling of thermophysical properties, enabling faster predictions, enhanced accuracy, and exploration of complex systems efficiently^{63,65–77}.

To address these challenges, this study leverages machine learning (ML) and deep learning (DL) methods for prediction of the density of seven thiophene-based heterocyclic compounds. An important innovation followed in this work is the use of critical properties including critical temperature (T_c), critical pressure (P_c), critical volume (V_c), and acentric factor (ω), together with boiling point (T_b), and molecular weight (Mw) as input parameters to predict the density of thiophene derivatives. These parameters inherently reflect the molecular structure, intermolecular interactions, and phase transition characteristics, which are essential for accurately predicting density under various conditions. Choosing the right input parameters is crucial for developing accurate and reliable predictive models, as they directly influence the model's ability to capture the underlying physical and chemical relationships. Using such physically meaningful and experimentally accessible inputs not only improves model performance and generalizability, but also ensures that predictions remain grounded in real-world chemical behaviour. In this work, in addition to machine learning models (DT, AdaBoost-DT, LightGBM and GBoost), we also used two deep learning models (TabNet and DNN) for high-pressure density prediction. By modelling the complex relationships between molecular structure and density, these computational approaches provide an efficient and scalable alternative to experimental methods. The findings not only enhance our understanding of thiophene derivatives but also demonstrate the potential of ML and DL in advancing predictive materials science, particularly for applications in pharmaceuticals, organic electronics, and sustainable energy solutions.

Theory and methodology

Dataset construction and description

This study delves into the density predictions for seven compounds from the thiophene family containing different functional groups (see Fig. S1 in Supplementary Material). Density prediction was studied in a wide temperature range (283.15–338.15 K) and pressure range (0.1–65 MPa), including 1336 data points¹. The experimental data for these compounds (thiophene, 2-methylthiophene, 3-methylthiophene, 2,5-dimethylthiophene, 2-thiophenemethanol, 2-thiophenecarboxaldehyde, and 2-acetylthiophene) were obtained from literature^{78–80}. Also, the critical properties of the compounds were extracted directly from the experimental data reported in sources⁸¹. As the ML/DL models were used in this work for thiophenic materials, at specified temperature and pressure ranges (283.15–338.15 K and 0.1–65 MPa), the potential reduced applicability of these models for molecules with different functional groups or different temperature and pressure ranges not present in the current dataset should be considered.

The thermal map presented in Fig. 1 shows a clear relationship between the density of thiophenes and temperature (T), pressure (P), critical temperature (T_c), critical pressure (P_c), critical volume (V_c), acentric factor (ω), boiling point (T_b), and molecular weight (Mw). This thermal map shows that P , Mw , T_c , T_b , V_c , ω have a direct relationship with the density of thiophene. Meanwhile, T and P_c have an inverse relationship with density. Fig. 1 shows that strong correlations between some descriptors may lead to overfitting by introducing

redundancy and multicollinearity into the model. To address this issue, we applied data normalization to ensure all features contribute equally during training, preventing those with larger scales from dominating the learning process. Additionally, we employed k-fold cross-validation to evaluate model performance across multiple data splits, which helps in selecting models that generalize well rather than fitting noise in the training data. Together, these techniques enhance the robustness and reliability of the model by reducing the risk of overfitting and improving predictive performance on unseen data.

Box plots offer valuable insights into outliers, median values, as well as minimum and maximum data points. The dataset comprises five key features: minimum, Q1 (the median of the lower half of the dataset), median (the middle value of the dataset), Q3 (the median of the upper half of the dataset), and maximum values. This figure consists of two main components: a pair of whiskers and a box. The lower whisker represents the minimum value, while the upper whisker indicates the maximum. The box itself spans from Q1 to Q3, illustrating data distribution. Additionally, the horizontal line in the center marks the median value. The box plots representing the input and target variables for temperature (T), pressure (P), critical temperature (T_c), critical pressure (P_c), critical volume (V_c), acentric factor (ω), boiling point (T_b), molecular weight (Mw) and density (ρ) are presented in Fig. S2 in Supplementary Material.

Predictive analytics

Enhancing accuracy can be achieved through grid search cross-validation⁸². This method systematically explores various models and hyperparameter combinations by testing each one individually and validating the results. The goal of grid search is to identify the optimal combination that yields the best model performance for prediction tasks⁸³. Typically, grid search is integrated with k-fold cross-validation to establish a reliable evaluation metric for classification models^{82,84}. In scikit-learn⁸⁵ the 'GridSearchCV' function can be utilized to implement the grid search algorithm for identifying the optimal hyperparameters⁸⁶. In this study, we used GridSearchCV to tune hyperparameters. Table 1 shows the hyperparameter search ranges used for machine learning and deep learning models, along with the optimal values identified through Grid Search.

Machine learning models

Decision tree (DT)

The decision tree (DT) method is a widely recognized machine learning approach for both classification and regression tasks⁸⁷. It derives its name from its hierarchical, tree-like structure, which operates similarly to a flowchart and is constructed using a partitioning process. Over time, various decision tree algorithms have been introduced, including ID3, C4.5, CART, CHAID, and MARS. The primary aim of DT learning is to establish a framework capable of effectively predicting variations in a response variable or categorizing data within a test dataset. To accomplish this, DT employs a branching structure where internal nodes represent decision points based on attributes, and leaf nodes indicate predicted output label^{88,89}. One of the strengths of the DT algorithm is its robustness to missing data and outliers, making it well-suited for both categorical and continuous variables. To prevent overfitting, key hyperparameters such as the minimum number of samples per leaf node and the maximum depth of the tree can be adjusted. Additionally, DT regression provides an intuitive way to examine the relationships between input and output variables, with its graphical representation serving as a practical tool for predicting continuous target values⁹⁰. Fig. S3 presents a schematic representation of the DT model.

Adaptive boosting decision tree (AdaBoost-DT)

Freund and Schapire introduced the adaptive boosting method (Adaboost) in 1997⁹¹ to develop a classifier. An adaptive resampling technique selects training samples, with classifiers being trained iteratively. During each iteration, misclassified samples are assigned more weight. Therefore, the final classifier is derived from a weighted aggregation of predictions from all trained models in the ensemble⁹². When paired with the AdaBoost algorithm, the DT, typically considered a weak classifier, is expected to achieve notably improved performance. The AdaBoost-DT model is implemented in Python 3.7 using the AdaBoost class from the scikit-learn library.

Gradient boosting (GBoost)

The Gradient Boosting Regressor (GBoost) is an ensemble learning technique that builds a series of decision trees in a sequential manner, where each successive tree is trained to minimize the errors made by the previous one. GBoost is an iterative learning algorithm designed to enhance predictive performance by combining multiple weak learners into a more robust model⁹³. As the number of weak models increases, the model's error progressively reduces⁹⁴. Furthermore, boosting addresses the bias-variance trade-off by initially constructing a weak learner and progressively enhancing its performance by sequentially adding new trees. Each newly added tree focuses on correcting the errors made by its predecessor by prioritizing the training instances with the highest prediction errors⁹⁵. In essence, the new tree assigns greater importance to the misclassified rows from the previous iteration. A schematic representation of the GBoost concept is shown in Fig. S4.

Deep neural network (DNN)

A Deep Neural Network (DNN) is a type of neural network that consists of multiple hidden layers. In recent years, DNNs have gained widespread popularity, largely due to advancements in computational resources and increased accessibility to high-performance computing^{96,97}. An appropriate network architecture is essential for ensuring the effective performance of a neural network. A standard DNN comprises an input layer, one or more hidden layers, and an output layer. The input and output layers define the model's inputs and expected outputs, while the hidden layers play a key role in extracting meaningful features from the given dataset. Each layer consists of numerous neurons that apply mathematical operations to the input data. Throughout the training

process, the model refines its performance by adjusting neuron-associated weights (w) and biases (b), a process guided by optimization techniques like gradient descent⁹⁸.

Light gradient boosting machine (LightGBM)

In 2016, Guolin Ke et al.⁹⁹ presented a new machine learning model, LightGBM, based on gradient boosting theory. Unlike other machine learning approaches, LightGBM requires less memory. LightGBM and XGBoost support parallel computations, but LightGBM outperforms the previous XGBoost model with faster training speed and lower memory usage. This reduction in memory occupation results in decreased communication costs during parallel learning. LightGBM stands out due to its decision tree-based architecture, which leverages gradient-based one-side sampling (GOSS), exclusive feature bundling (EFB), and a histogram-based learning strategy with a depth-constrained, leaf-wise growth mechanism¹⁰⁰. Gradually based one-sided sampling (GOSS) can strike a desirable balance between the sample size and the accuracy of LightGBM's decision tree. In LightGBM, an efficient algorithm called EFB is employed to group those parameters that rarely have nonzero values simultaneously (see Fig. S5 in Supplementary Material for the leaf-wise tree growth strategy). As decision trees deepen, overfitting tendencies increase, leading to more undesirable leaf directions. LightGBM's crucial parameters enable it to handle large volumes of data, perform at high speed, and achieve higher accuracy in predictions¹⁰¹. However, when LightGBM leads to overfitting, setting a maximum depth limit for the leaf nodes can result in higher efficiency^{99,102}. Concerning the construction of a LightGBM model, parameters and computations can be described as follows^{103,104}.

$$X = \{(x_j, y_j)\}_{j=1}^N \tag{1}$$

After minimizing the loss function L , the value of $f(x)$ was predicted:

$$L(y, f(x)) : \hat{f}(x) = \operatorname{argmin}_{f(x)} E_{x,y}.L(y, f(x)) \tag{2}$$

In conclusion, the training process of each tree can be described as follows:

$$W_{q(x)}, q \in (1, 2, 3, \dots, N) \tag{3}$$

In the given equation, N represents the leaf count in a tree, q indicates the decision rules employed in a single tree, and W signifies the weight term of each leaf node. To minimize the objective function using Newton's method, the outcome of each stage's training is adjusted as follows:

Models	Statistical parameters				
	AAPRE	APRE	RMSE	SD	R ²
LightGBM					
Train	0.02126	0.00040	0.32131	0.00031	0.99998
Test	0.03034	−0.00231	0.44628	0.00043	0.99997
Total	0.02308	−0.00014	0.34998	0.00033	0.99998
AdaBoost-DT					
Train	0.12346	0.00041	1.78314	0.00163	0.99961
Test	0.18487	0.00811	2.61380	0.00240	0.99921
Total	0.13578	0.00196	1.97794	0.00181	0.99953
GBoost					
Train	0.01967	0.00001	0.29777	0.00029	0.99998
Test	0.05051	0.00212	0.70362	0.00067	0.99994
Total	0.02585	0.00044	0.41254	0.00039	0.99997
DT					
Train	0.05648	−0.00096	1.02467	0.00098	0.99987
Test	0.25361	0.01116	3.03449	0.00282	0.99894
Total	0.09602	0.00216	1.63904	0.00154	0.99968
TabNet					
Train	0.18276	−0.06131	2.81115	0.00258	0.99905
Test	0.17718	−0.07547	2.64649	0.00245	0.99919
Total	0.18164	−0.06415	2.77890	0.00255	0.99908
DNN					
Train	0.09832	0.07863	1.44123	0.00125	0.99975
Test	0.09970	0.07394	1.46719	0.00127	0.99975
Total	0.09860	0.07769	1.44647	0.00126	0.99975

Table 2. Statistical error analysis for the models developed in this work. The best results are in bold.

$$G_t \cong \sum_{i=1}^N L[y_i, F_{t-1}(x_i) + f_t(x_i)] \quad (4)$$

TabNet

TabNet is a deep learning model specifically designed for tabular data¹⁰⁵. Unlike traditional deep learning models, it directly processes raw data without requiring manual feature engineering. TabNet employs a sparse attention mechanism to dynamically select relevant features, enhancing both interpretability and efficiency. Its core components include¹⁰⁶:

Feature Transformer: Processes input data and generates complex feature representations.

Attentive Transformer: Determines which features should be selected at each decision step using a sparse attention mechanism.

Masking Mechanism: Guides the feature selection process to improve model transparency and efficiency.

Aggregation: Combines the selected features from multiple steps to produce the final output.

TabNet is built on a multi-step decision-making process, refining its feature selection iteratively¹⁰⁷. Its architecture integrates a feature transformer, an attentive transformer, and a masking mechanism, making it a powerful model for structured data tasks. Empirical studies demonstrate its high performance and strong generalization capabilities across various datasets^{108,109}.

Statistical error analysis

The following statistical parameters were used to compare the performance of the model used in this study. (ρ_{pred}) represents the density predicted by the deep learning and machine learning models, and (ρ_{exp}) represents the experimental values of the density.

average percent relative error (APRE)

$$APRE = \frac{1}{n} \sum_{i=1}^n \frac{(\rho_{iexp} - \rho_{ipred})}{(\rho_{iexp})} \quad (5)$$

average absolute percent relative error (AAPRE)

$$AAPRE = \frac{1}{n} \sum_{i=1}^n \left| \frac{(\rho_{iexp} - \rho_{ipred})}{(\rho_{iexp})} \right| \quad (6)$$

root mean square error (RMSE)

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (\rho_{iexp} - \rho_{ipred})^2}{n}} \quad (7)$$

standard deviation (SD)

$$SD = \sqrt{\frac{\sum_{i=1}^n \frac{(\rho_{iexp} - \rho_{ipred})^2}{\rho_{iexp}}}{n - 1}} \quad (8)$$

coefficient of determination (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^n (\rho_{iexp} - \rho_{ipred})^2}{\sum_{i=1}^n (\rho_{iexp} - \bar{\rho}_{iexp})^2} \quad (9)$$

Results and discussion

Table 2 provides an overview of the statistical performance metrics for six models: LightGBM, AdaBoost-DT, GBoost, DT, TabNet, and DNN. The assessment was conducted on training (1068 data points), testing (268 data points), and the complete dataset (1336 data points). The performance metrics in the table clearly demonstrate that LightGBM outperforms both deep learning models TabNet and DNN as well as other traditional machine learning models in predicting the density of seven thiophene-based compounds. LightGBM achieves the lowest errors across all key metrics, including AAPRE (0.03034 test), RMSE (0.44628 test), and SD (0.00043 test), while maintaining an exceptionally high R^2 of 0.99997 on the test set. Although the LightGBM model demonstrates extremely high accuracy and low error metrics, suggesting excellent predictive performance, we acknowledge the importance of assessing the risk of overfitting. To address this, we employed k-fold cross-validation during model training, which ensures the model's performance is consistent across multiple data subsets and not just the training set. The close alignment between training and test performance, along with low standard deviation in error metrics, indicates that the model generalizes well and is not overfitting. Nonetheless, we remain cautious and have included model validation measures to confirm its robustness and reliability. In contrast, TabNet and DNN show significantly higher prediction errors, with TabNet yielding an RMSE of 2.64649 and DNN 1.46719 on the test set, indicating weaker generalization. This superior performance of LightGBM is primarily due to the

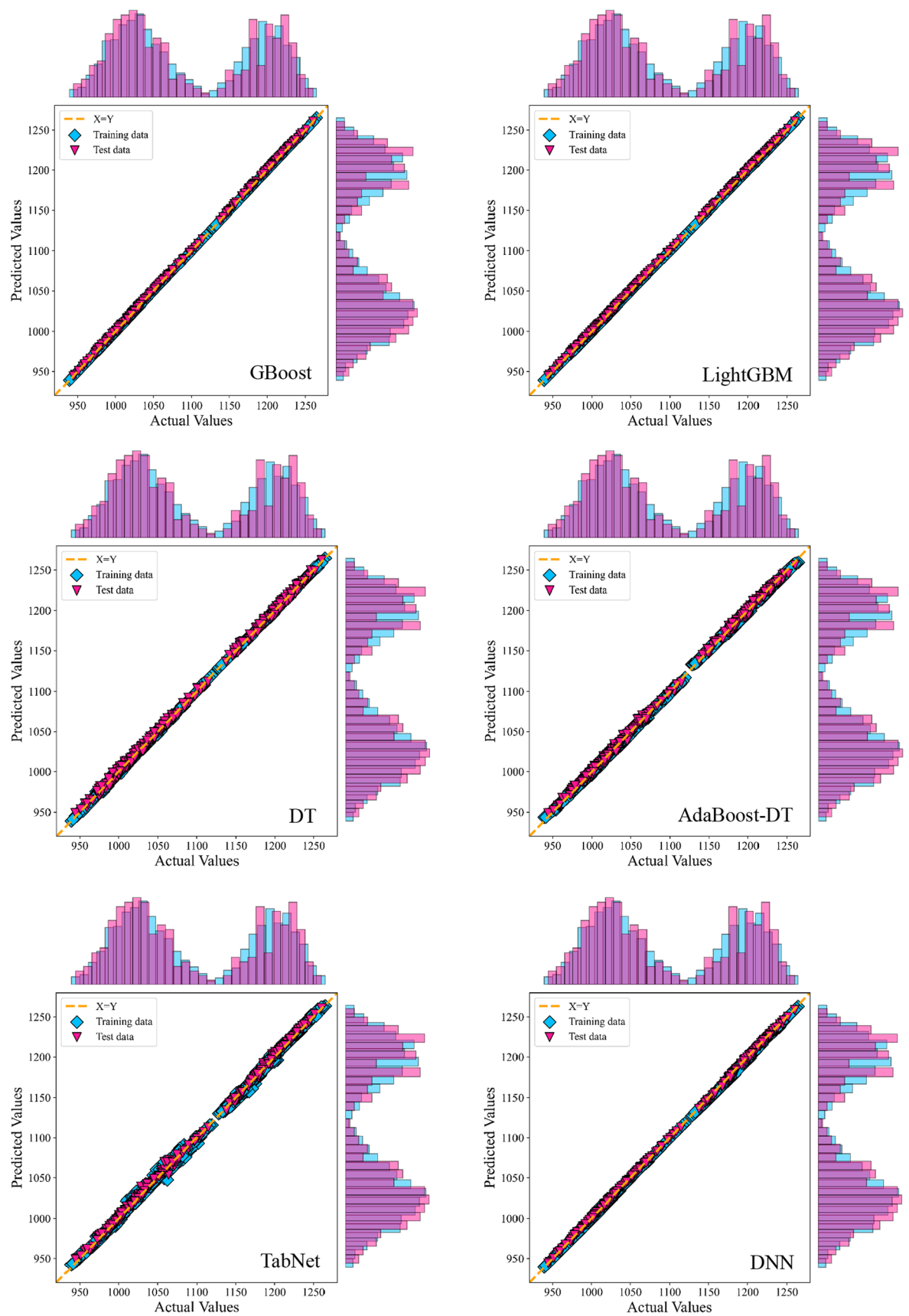


Fig. 2. Cross-plot of the developed models for density prediction.

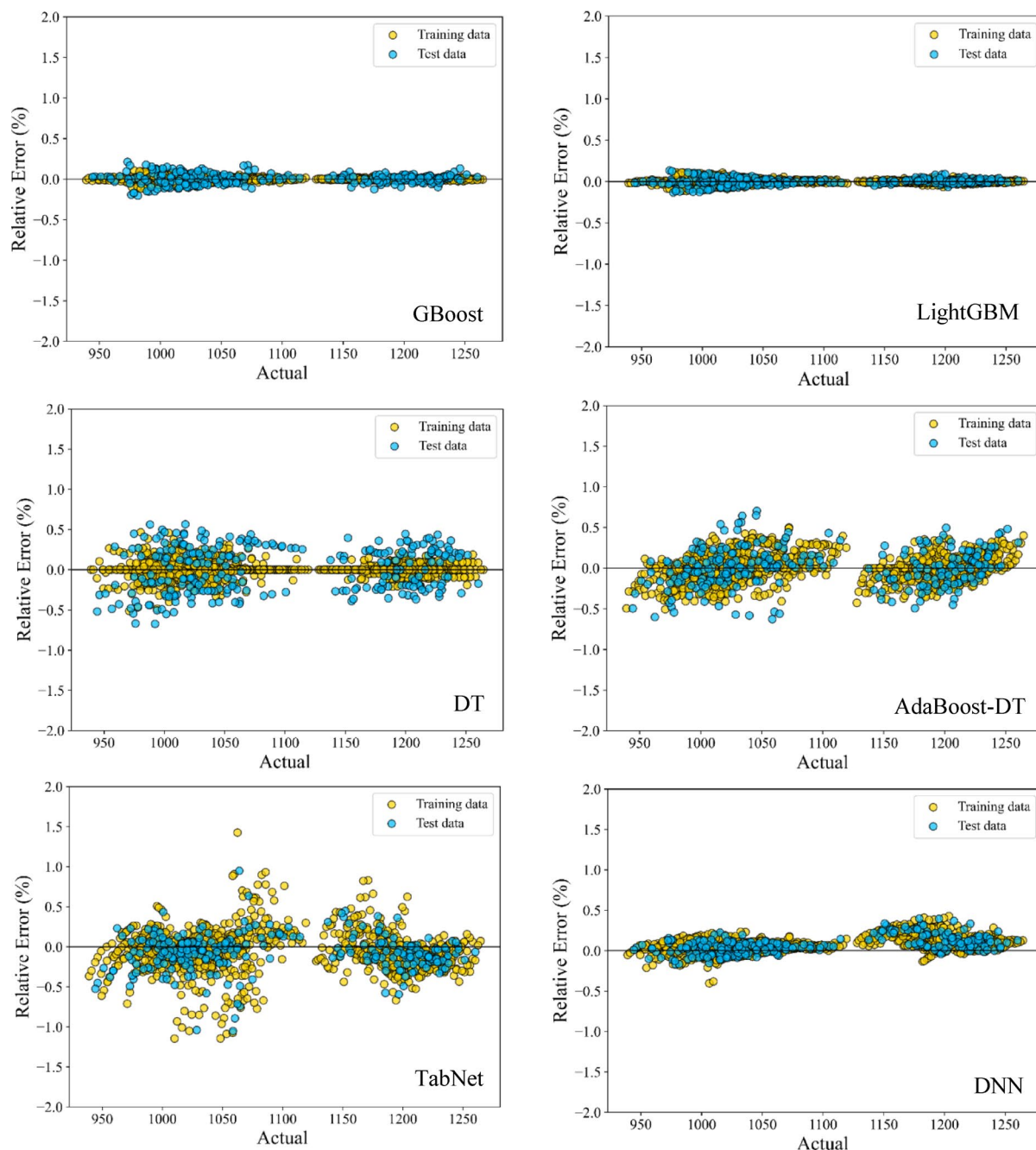


Fig. 3. Error distribution diagrams of the models.

inherent suitability of tree-based models for structured, tabular data such as the molecular descriptors used in this study. Tree-based models like LightGBM can naturally model non-linear feature interactions and manage small to medium datasets more efficiently, without requiring extensive tuning. Meanwhile, deep learning models like TabNet and DNN face architectural limitations in tabular contexts they often struggle to generalize well without large datasets, are prone to overfitting, and require complex hyperparameter optimization. These findings highlight LightGBM's superior accuracy, as further illustrated in Fig. S6 (Supplementary Material).

The Taylor diagram [78] provides a visual representation of key statistical metrics R^2 , RMSE, and standard deviation (SD) to assess how well the predicted density aligns with experimental data. In this diagram, models with higher accuracy appear closer to the reference measurement point, while those with greater error deviate further. Among the evaluated models, LightGBM demonstrates the closest alignment with experimental data for both training and test sets, confirming its superior predictive accuracy (see Fig. S7 in Supplementary Material).

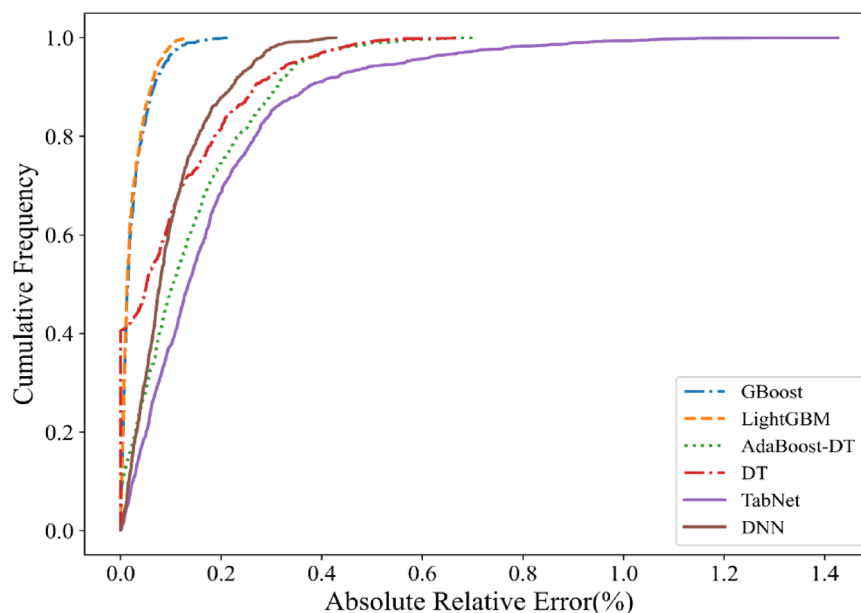


Fig. 4. Cumulative frequency distribution of the models developed in this study.

Graphical analysis

Graphical error analysis is a method for evaluating a model's performance. This graphical tool is handy for comparing the performance of multiple models. Various schematic analyses were conducted in this study to demonstrate the effectiveness of the developed model. Graphical curves, including cross-plots, error distributions, group errors, and cumulative frequencies, were used to illustrate the reliability of the developed models.

Cross-plot

A cross plot is a type of scatter plot that visualizes the relationship between actual and predicted values by aligning them along a 45° line that passes through the origin. Fig. 2 plots the predicted values of the models against the experimental data. The greater the concentration of points on the $Y = X$ line, the greater the accuracy of the model. As can be seen in Fig. 2, all models perform well and the points on the ideal line are aligned.

Error distribution plot

Fig. 3 shows the distribution of relative errors of the proposed models in the training and testing processes. The lower the data density near the line $Y = 0$ the greater the model error and the lower the accuracy for predicting density. As a result, the GBoost and LightGBM models have lower relative error than the proposed models for the training and testing data, thus they have higher accuracy for density prediction.

Cumulative frequency graph

The cumulative frequency plot of absolute relative error (%) for the models used in this study is shown in Fig. 4. This figure clearly shows the higher accuracy of the GBoost and LightGBM models than other proposed models for density prediction. In addition, the TabNet model has a higher error than other models.

This study also explores error frequency by creating histograms of relative error. Fig. 5 displays histograms of relative error for six developed models. In the LightGBM model, most data points have errors between -0.25 and 0.25 , centering around zero relative error. Data with errors outside the range of -0.25 to 0.25 for the DT, AdaBoost-DT, TabNet and DNN models indicates that these models have less coverage than LightGBM for both the training and testing data.

The obtained results for the error values (see Fig. S8 in Supplementary Material) show that LightGBM model produces the smallest error distribution range, from -0.1337 to 0.1321 . The GBoost model ranges from -0.2130 to 0.2028 , while other models show a higher error distribution range than these two models.

Fig. 6 compares the effect of input parameters (critical temperature, critical pressure, critical volume, molecular weight, boiling temperature, together with operational determination temperature and pressure), on absolute relative error (%) for all models. As can be observed, in all ranges of molecular weight, boiling temperature, critical temperature, critical pressure, critical volume, temperature, and pressure, the LightGBM model has the least error compared to other models, which confirms the high accuracy of this model.

A comparative analysis of the relative error among the proposed models offers valuable insights into identifying the most accurate predictive approach. This visual assessment demonstrates the strong alignment between experimental data and the predictions generated by the LightGBM model, as depicted in Fig. S9 of the Supplementary Materials.

Fig. 7 illustrates the percentage of relative error for the LightGBM model across the studied materials. The consistently low relative error across all materials underscores the model's high precision in predicting density.

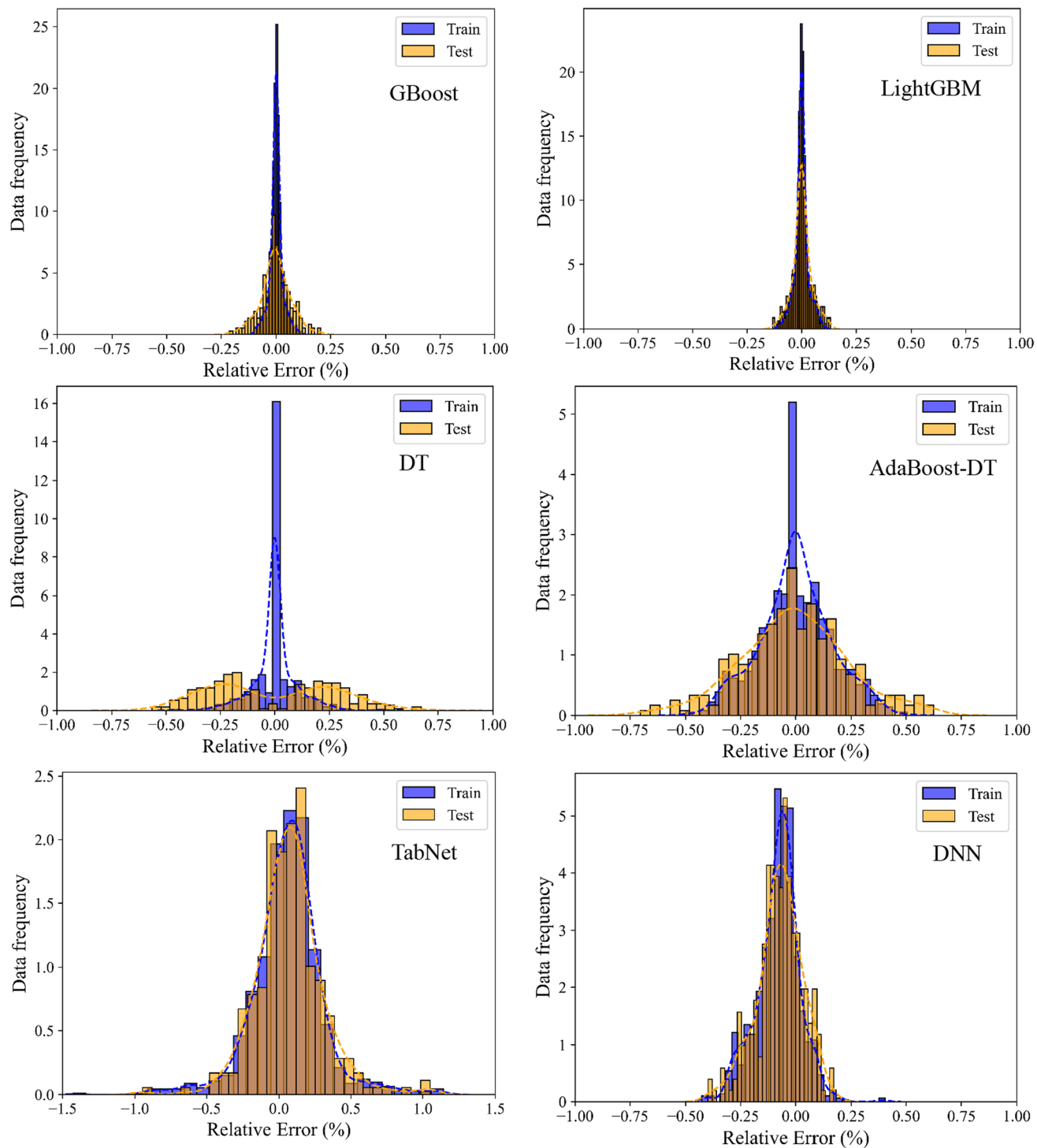


Fig. 5. Histograms of relative error for the proposed models in density prediction.

This minimal deviation further confirms the reliability and effectiveness of the LightGBM model in accurately estimating density values.

Model trend analysis

To assess how well the developed models capture the expected density trends, Fig. 8 presents the LightGBM model's predicted values as a function of temperature and pressure. The plots illustrate that at fixed pressures of 7.0 and 65 MPa, density decreases with rising temperature. Conversely, at constant temperatures of 303.15 K and 338.15 K, increasing pressure results in higher density.

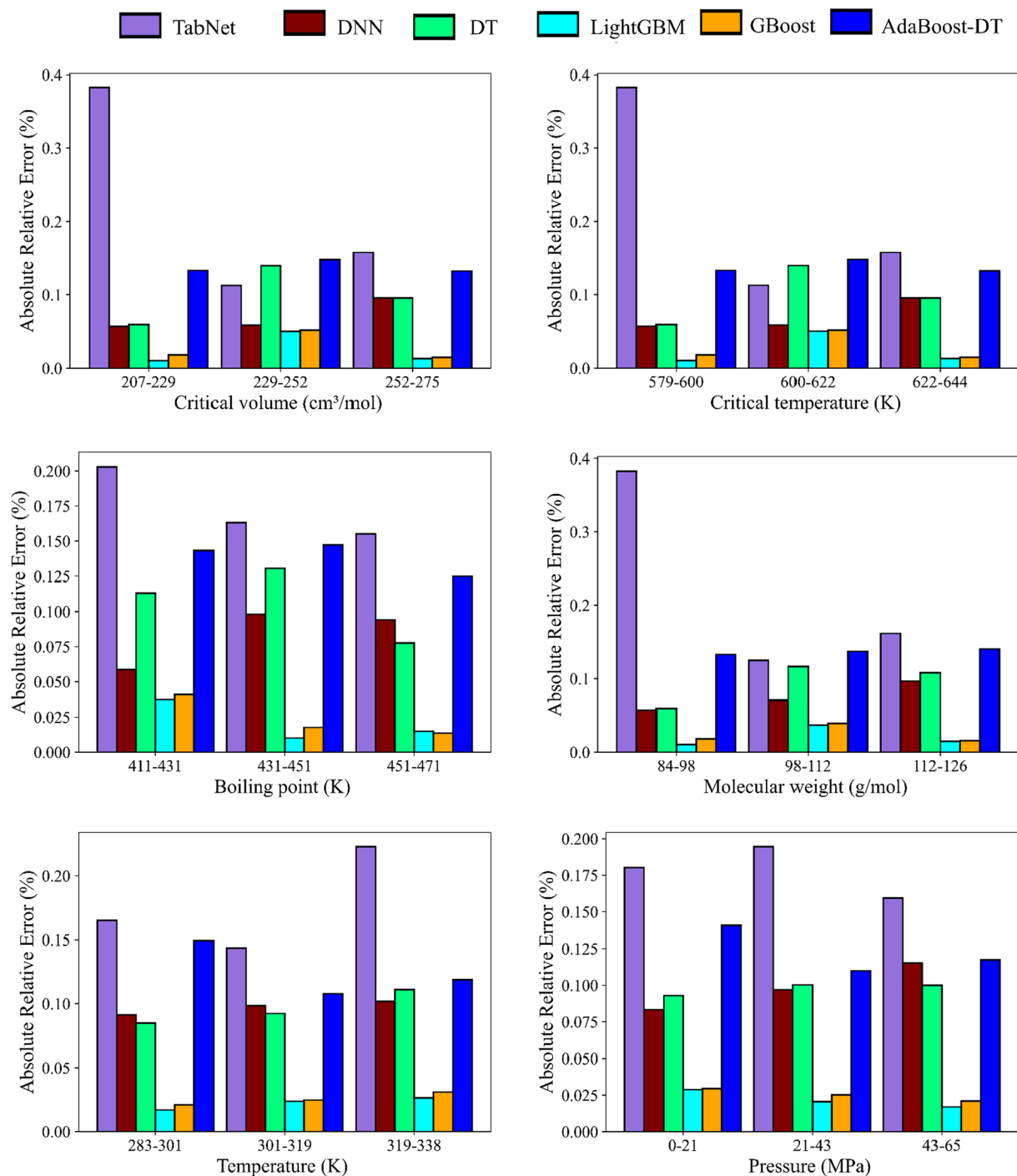


Fig. 6. AARE of all models proposed in this work for different input parameter ranges.

Sensitivity analysis

The relevancy factor (r) and the output of the LightGBM model are employed to assess the relative significance of input variables in predicting density. The correlation coefficient for each input parameter is determined using the following formula [63, 64]:

$$r(I_k, y) = \frac{\sum_{i=1}^n (I_{i,k} - \bar{I}_k)(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (I_{i,k} - \bar{I}_k)^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (10)$$

$I_{i,k}$ and \bar{I}_k represent the i th average values of the k th input, respectively. K represents pressure temperature or other input parameters. y_i and \bar{y} represent the i th predictive value and average. The parameter (r) varies between

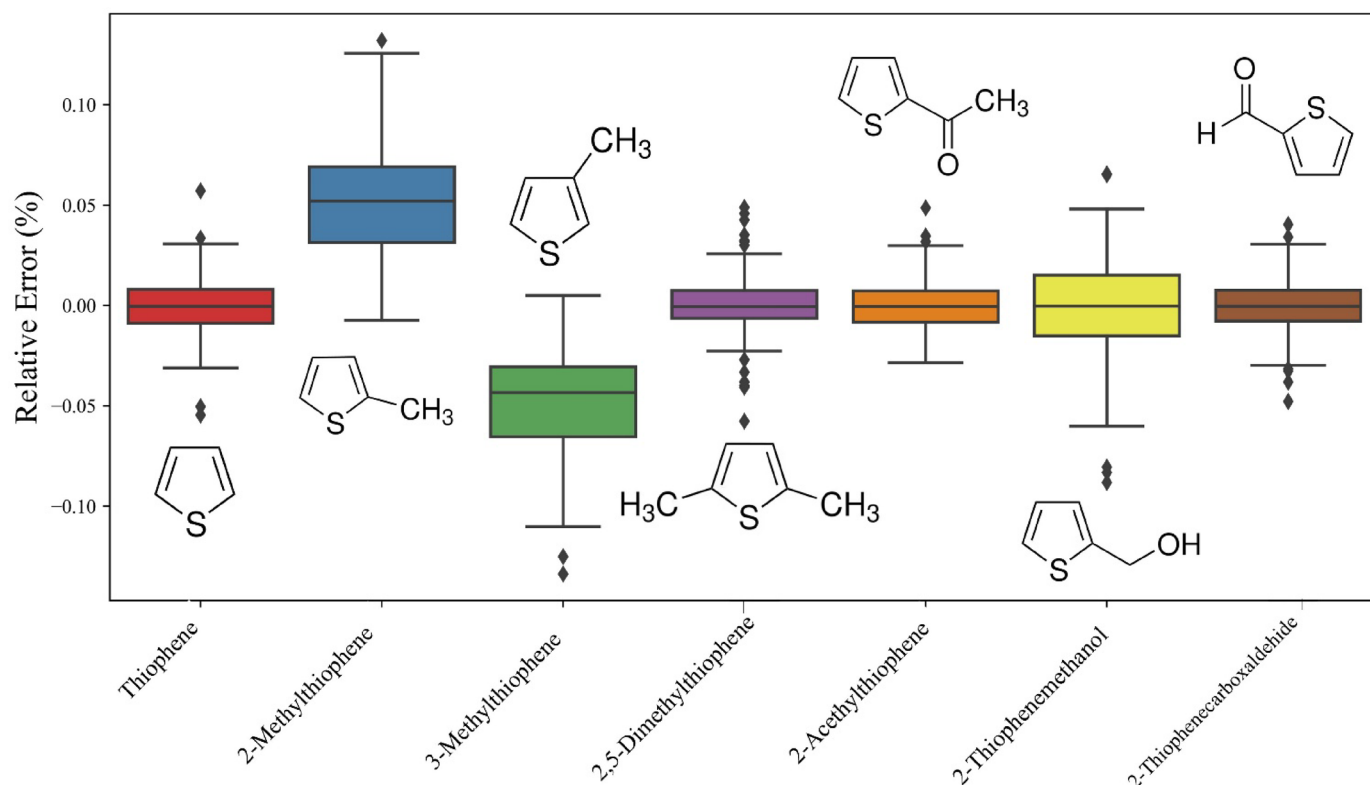


Fig. 7. Box plots displaying the relative error distribution for various thiophene compounds.

–1 and 1, reflecting the correlation between independent and dependent variables. A positive (r) suggests that as the input variable increases, the output also rises, whereas a negative (r) implies an inverse correlation. And the closer (r) are to 1, the stronger the association between the model's input and output values. The findings of the sensitivity analysis on the results of the LightGBM model, as the best-obtained model, are presented visually in Fig. 9. The relevancy factor plot clearly shows how each input parameter influences the model's prediction of density, with boiling point (T_b), critical volume (V_c), and critical temperature (T_c) having the highest positive relevancy, indicating they are the most influential features. This aligns well with established physicochemical principles of thiophenes, where thermophysical properties such as density are strongly governed by phase behavior and intermolecular interactions both of which are reflected in critical and boiling point properties. For example, the strong correlation of T_b (0.7302) suggests that vaporization characteristics significantly impact the density profile. Similarly, the contributions of T_c (0.5683), V_c (0.5857), and ω (0.5131) highlight the importance of molecular structure and dispersion forces, which are central to understanding thiophene derivatives due to their aromatic and heterocyclic nature. The negative relevancy of P_c (–0.4949) and T (–0.1675) further supports the idea that increased external pressure or system temperature can reduce the predictability of density if not properly accounted for by structural properties. Overall, this plot confirms that the selected features not only enhance model performance but also reflect fundamental chemical behavior.

Implementation of the Leverage method

After following statistical and graphical analyses that confirmed the superiority of the LightGBM model over other approaches, an additional outlier detection method was applied to identify data points that could adversely affect model predictions and to validate the reliability domain of the proposed model. The Williams plot visualizes standardized residuals (R) against hat values (H), providing insights into potential outliers. The key parameters for constructing this plot are determined using the following calculations^{77,110,111}:

Hat matrix (H):

$$H = X(X^T X)^{-1} X^T \quad (11)$$

Here, X^T represents the transpose of the matrix X , which is a ($y \times z$) matrix. In this case, y refers to the number of data points, and z refers to the number of input variables used by the model.

• Leverage limit (H^*):

$$H^* = \frac{3 \times (z + 1)}{y} \quad (12)$$

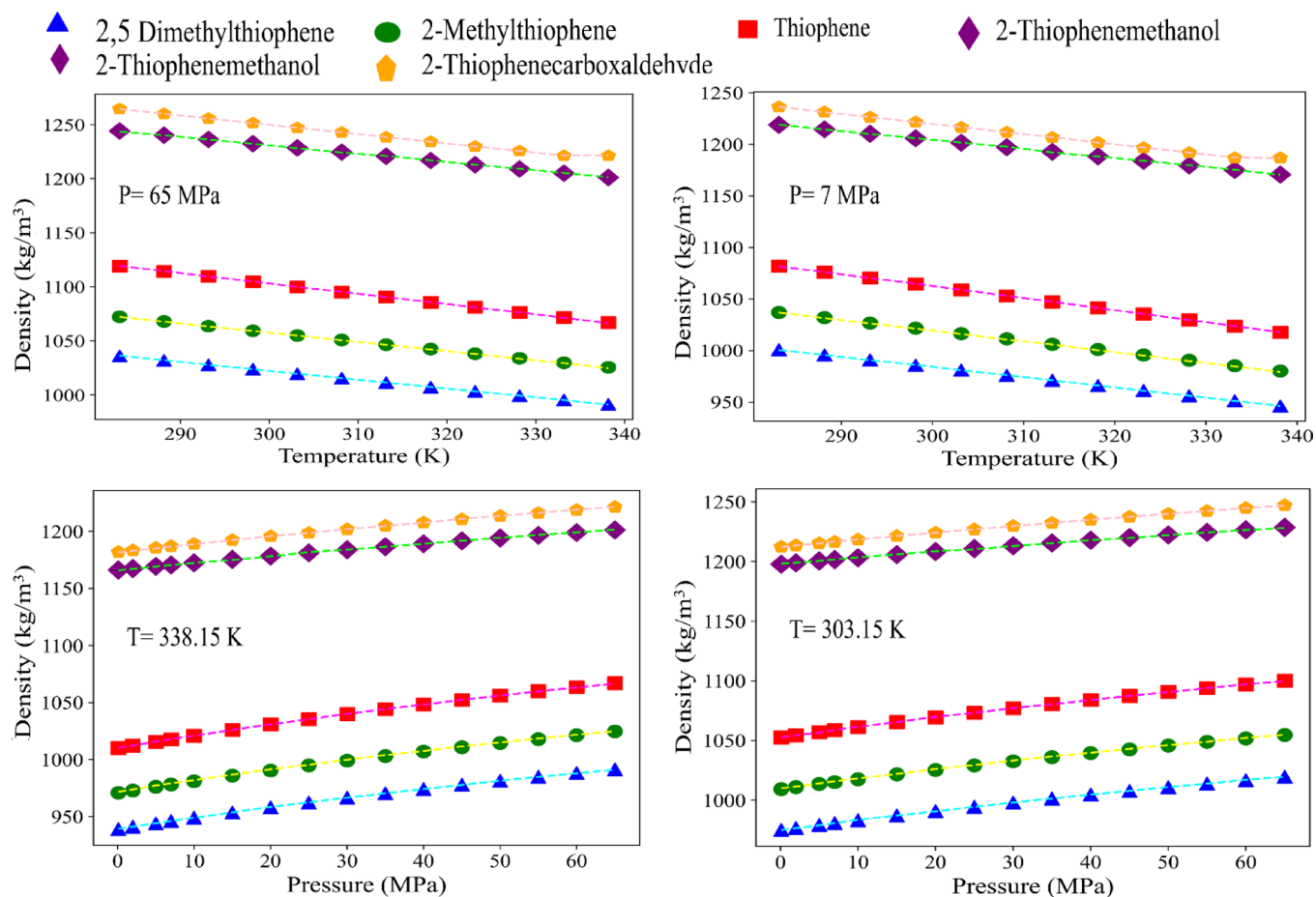


Fig. 8. Top: Effect of temperature change on density at constant pressures 7 and 65 MPa; Bottom: Effect of pressure change on density at constant temperatures 303.15, and 338.15 K.

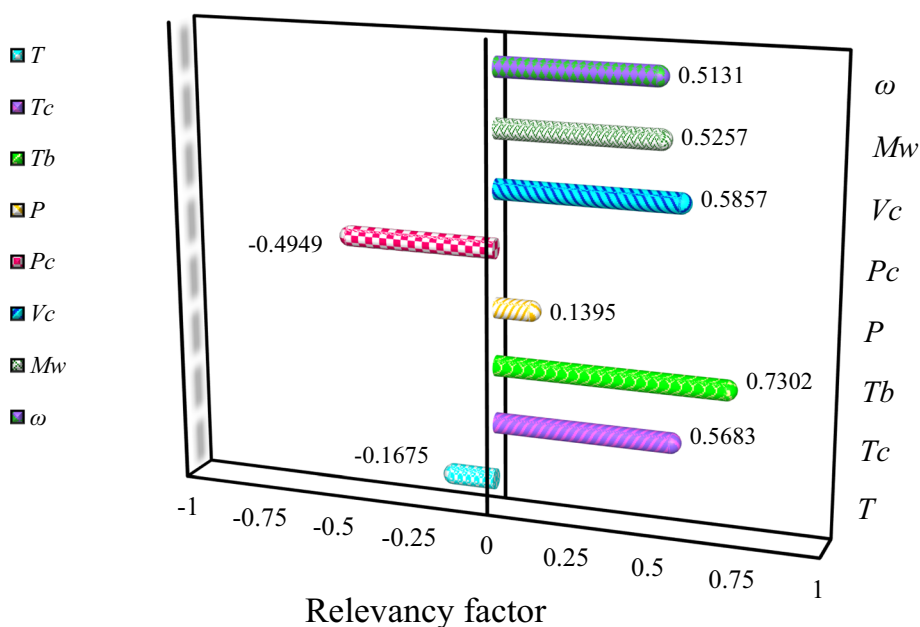


Fig. 9. Sensitivity analysis on the LightGBM model.

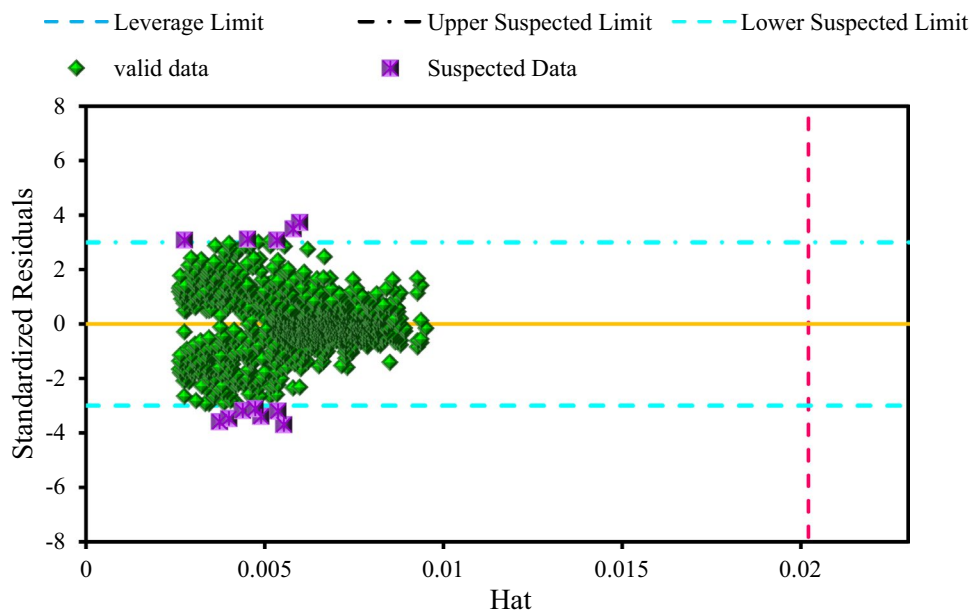


Fig. 10. William's design is based on the LightGBM model.

- standardized residuals (SR):

$$R_i = \frac{e_j}{\sqrt{MSE(1 - H_j)}} \quad (13)$$

where e_j is the ordinary residual of the j th index, MSE is the mean square error, and H_j is the j th Leverage value. H values greater than H^* are outside the range of applicability of the model. In addition, data points with H values less than H^* and R values greater than 3 or less than -3 are considered suspected data. Data points with H values less than H^* and R values between -3 and 3 are considered valid data¹¹². As illustrated in Fig. 10, over 99% of the dataset is deemed valid, with only 12 out of 1336 data points identified as potential anomalies. Williams chart analysis shows that 99.10% of the data falls within the acceptable range.

Conclusion

In this study, the critical properties including critical temperature (T_c), critical pressure (P_c), critical volume (V_c), and acentric factor (ω), together with boiling point (T_b), and molecular weight (Mw) were used as input parameters for machine / deep learning models to predict the density of thiophenes. Accurate density prediction is vital for understanding and mitigating the environmental and industrial impact of sulfur compounds in fuels. In this work, in addition to four machine learning models (DT, AdaBoost-DT, LightGBM, and GBoost) we also used two deep learning models (TabNet and DNN) for density prediction. Results revealed that the LightGBM model outperformed the others, with the lowest errors in statistical evaluations (AAPRE=0.02308, APRE=-0.00014, RMSE=0.34998, and an $R^2=0.99998$). Graphical evaluations further confirmed the LightGBM model's high accuracy in predicting thiophene density across training and test datasets. In addition, the comparison of the experimental data and predicted values by the LightGBM model at constant temperatures of 303.15 and 338.15 K and constant pressures of 7 and 65 MPa proved the accuracy of the prediction. Using the relevance factor, the impact of input characteristics on the model's target parameter was also investigated. The Leverage technique revealed that all data points appeared trustworthy and valid, except for a few that fell into the suspected data region. In summary, applying the Leverage method confirmed the data integrity and effectiveness of the proposed LightGBM model. This study distinguishes itself through its comprehensive dataset, a broader range of thiophene derivatives, and the incorporation of advanced machine/deep learning models. The findings provide a robust foundation for optimizing the properties of thiophene derivatives, supporting innovations in fuel refinement, environmental sustainability, and advanced material applications.

Data availability

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Received: 28 April 2025; Accepted: 30 June 2025

Published online: 15 July 2025

References

- Razavi, E., Khoshshima, A. & Shahriari, R. Phase behavior modeling of mixtures containing N-, S-, and O-heterocyclic compounds using PC-SAFT equation of state. *Ind. Eng. Chem. Res.* **58**, 11038–11059. <https://doi.org/10.1021/acs.iecr.9b01429> (2019).
- Bai, Y. et al. Research on ultrafast catalytic degradation of heterocyclic compounds by the novel high entropy catalyst. *J. Environ. Chem. Eng.* **11**, 110703. <https://doi.org/10.1016/j.jece.2023.110703> (2023).
- Abro, R. et al. Extractive desulfurization of fuel oils using deep eutectic solvents—A comprehensive review. *J. Environ. Chem. Eng.* **10**, 107369. <https://doi.org/10.1016/j.jece.2022.107369> (2022).
- Granger, P. & Parvulescu, V. I. Catalytic NO_x abatement systems for mobile sources: From three-way to lean burn after-treatment technologies. *Chem. Rev.* **111**, 3155–3207 (2011).
- Zipper, C. E. & Gilroy, L. Sulfur dioxide emissions and market effects under the clean air act acid rain program. *J. Air Waste Manag. Assoc.* **48**, 829–837 (1998).
- Peng, Y., Liu, C., Zhang, X. & Li, J. The effect of SiO₂ on a novel CeO₂–WO₃/TiO₂ catalyst for the selective catalytic reduction of NO with NH₃. *Appl. Catal. B* **140**, 276–282 (2013).
- Singh, A. & Agrawal, M. Acid rain and its ecological consequences. *J. Environ. Biol.* **29**, 15 (2007).
- Chen, R. et al. Short-term exposure to sulfur dioxide and daily mortality in 17 Chinese cities: The China air pollution and health effects study (CAPES). *Environ. Res.* **118**, 101–106 (2012).
- Chiang, T.-Y., Yuan, T.-H., Shie, R.-H., Chen, C.-F. & Chan, C.-C. Increased incidence of allergic rhinitis, bronchitis and asthma, in children living near a petrochemical complex with SO₂ pollution. *Environ. Int.* **96**, 1–7 (2016).
- Alshareef, F. M. et al. Recent advancement in organic fluorescent and colorimetric chemosensors for the detection of Al³⁺ ions: A review (2019–2024). *J. Environ. Chem. Eng.* **12**, 114110. <https://doi.org/10.1016/j.jece.2024.114110> (2024).
- Chen, H.-J. et al. Rational design and synthesis of 2, 2-bisheterocycle tandem derivatives as non-nucleoside hepatitis B virus inhibitors. *ChemMedChem* **3**, 1316–1321 (2008).
- Monge Vega, A., Aldana, I., Rabbani, M. & Fernandez-Alvarez, E. The synthesis of 11H-1, 2, 4-triazolo [4, 3-b] pyridazino [4, 5-b] indoles, 11H-tetrazolo [4, 5-b] pyridazino [4, 5-b] indoles and 1, 2, 4-triazolo [3, 4-f]-1, 2, 4-triazino [4, 5-a] indoles. *J. Heterocycl. Chem.* **17**, 77–80 (1980).
- Burnett, D. A., Caplen, M. A., Davis, H. R. Jr., Burrier, R. E. & Clader, J. W. 2-Azetidinones as inhibitors of cholesterol absorption. *J. Med. Chem.* **37**, 1733–1736 (1994).
- Russell, R. K. et al. Thiophene system. 9. Thienopyrimidinedione derivatives as potential antihypertensive agents. *J. Med. Chem.* **31**, 1786–1793 (1988).
- Mohareb, R. M., Abdallah, A. E. & Abdelaziz, M. A. New approaches for the synthesis of pyrazole, thiophene, thieno [2, 3-b] pyridine, and thiazole derivatives together with their anti-tumor evaluations. *Med. Chem. Res.* **23**, 564–579 (2014).
- Chen, Z., Ku, T. C. & Seley-Radtke, K. L. Thiophene-expanded guanosine analogues of gemcitabine. *Bioorg. Med. Chem. Lett.* **25**, 4274–4276 (2015).
- Chan, H. S. O. & Ng, S. C. Synthesis, characterization and applications of thiophene-based functional polymers. *Prog. Polym. Sci.* **23**, 1167–1231 (1998).
- Song, X. & Parish, C. A. Pyrolysis mechanisms of thiophene and methylthiophene in asphaltenes. *J. Phys. Chem. A* **115**, 2882–2891 (2011).
- Abdou, M. M. Thiophene-based azo dyes and their applications in dyes chemistry. *Am. J. Chem.* **3**, 126–135 (2013).
- Perepichka, I. F., Perepichka, D. F., Meng, H. & Wudl, F. Light-emitting polythiophenes. *Adv. Mater.* **17**, 2281–2305 (2005).
- Mi, S., Wu, J., Liu, J., Zheng, J. & Xu, C. Donor– π -bridge–acceptor fluorescent polymers based on thiophene and triphenylamine derivatives as solution processable electrochromic materials. *Org. Electron.* **23**, 116–123 (2015).
- Kim, C. et al. The effects of octylthiophene ratio on the performance of thiophene based polymer light-emitting diodes. *Sci. Adv. Mater.* **7**, 2401–2409 (2015).
- Jeong, H. Y., Lee, S. Y., Han, J., Lim, M. H. & Kim, C. Thiophene and diethylaminophenol-based “turn-on” fluorescence chemosensor for detection of Al³⁺ and F[–] in a near-perfect aqueous solution. *Tetrahedron* **73**, 2690–2697 (2017).
- Fernandes, R. S., Shetty, N. S., Mahesha, P. & Gaonkar, S. L. A Comprehensive review on thiophene based chemosensors. *J. Fluoresc.* **32**, 19–56. <https://doi.org/10.1007/s10895-021-02833-x> (2022).
- Mishra, A., Ma, C.-Q. & Bauerle, P. Functional oligothiophenes: Molecular design for multidimensional nanoarchitectures and their applications. *Chem. Rev.* **109**, 1141–1276 (2009).
- Thompson, B. C. & Fréchet, J. M. Polymer–fullerene composite solar cells. *Angew. Chem. Int. Ed.* **47**, 58–77 (2008).
- Allard, S., Forster, M., Souharce, B., Thiem, H. & Scherf, U. Organic semiconductors for solution-processable field-effect transistors (OFETs). *Angew. Chem. Int. Ed.* **47**, 4070–4098 (2008).
- Gather, M. C., Köhnen, A. & Meerholz, K. White organic light-emitting diodes. *Adv. Mater.* **23**, 233–248 (2011).
- Raposo, M. M. M. et al. Synthesis and characterization of dicyanovinyl-substituted thienylpyrroles as new nonlinear optical chromophores. *Org. Lett.* **8**, 3681–3684 (2006).
- Tang, C. W. & VanSlyke, S. A. Organic electroluminescent diodes. *Appl. Phys. Lett.* **51**, 913–915 (1987).
- Xu, L., Tang, C. W. & Rothberg, L. J. High efficiency phosphorescent white organic light-emitting diodes with an ultra-thin red and green co-doped layer and dual blue emitting layers. *Org. Electron.* **32**, 54–58 (2016).
- Piyakulawat, P. et al. Effect of thiophene donor units on the optical and photovoltaic behavior of fluorene-based copolymers. *Sol. Energy Mater. Sol. Cells* **95**, 2167–2172 (2011).
- Klauck, H. Organic thin-film transistors. *Chem. Soc. Rev.* **39**, 2643–2666 (2010).
- Lin, Y., Li, Y. & Zhan, X. Small molecule semiconductors for high-efficiency organic photovoltaics. *Chem. Soc. Rev.* **41**, 4245–4272 (2012).
- O'Neill, M. & Kelly, S. M. Ordered materials for organic electronics and photonics. *Adv. Mater.* **23**, 566–584 (2011).
- Usta, H., Facchetti, A. & Marks, T. J. n-Channel semiconductor materials design for organic complementary circuits. *Acc. Chem. Res.* **44**, 501–510 (2011).
- Bae, S.-E., Kim, K.-J., Hwang, Y.-K. & Huh, S. Simple preparation of Pd-NP/polythiophene nanospheres for heterogeneous catalysis. *J. Colloid Interface Sci.* **456**, 93–99 (2015).
- Lee, H., Shin, M., Lee, M. & Hwang, Y. J. Photo-oxidation activities on Pd-doped TiO₂ nanoparticles: critical PdO formation effect. *Appl. Catal. B* **165**, 20–26 (2015).
- Megahed, A. S., Al-Amoudi, M. & Refat, M. A modern technique for preparation of zinc (II) and nickel (II) nanometric oxides using Schiff base compounds: Synthesis, characterization, and antibacterial properties. *Res. Chem. Intermed.* **40**, 1425–1439 (2014).
- Xie, J., Dong, H., Yu, Y. & Cao, S. Inhibitory effect of synthetic aromatic heterocycle thiosemicarbazone derivatives on mushroom tyrosinase: Insights from fluorescence, 1H NMR titration and molecular docking studies. *Food Chem.* **190**, 709–716 (2016).
- Roy, S. et al. Structural basis for molecular recognition, theoretical studies and anti-bacterial properties of three bis-uracil derivatives. *Tetrahedron* **70**, 6931–6937 (2014).
- Caboni, P. et al. Nematicidal activity of 2-thiophenecarboxaldehyde and methylisothiocyanate from caper (*Capparis spinosa*) against *Meloidogyne incognita*. *J. Agric. Food Chem.* **60**, 7345–7351 (2012).
- Pereira, G. A. et al. A broad study of two new promising antimycobacterial drugs: Ag (I) and Au (I) complexes with 2-(2-thienyl) benzothiazole. *Polyhedron* **38**, 291–296 (2012).

44. Zhou, Y. et al. Sulfur-rich carbon cryogels for supercapacitors with improved conductivity and wettability. *J. Mater. Chem. A* **2**, 8472–8482 (2014).
45. Kumar, S., Dutta, P. & Sen, P. Preparation and characterization of optical property of crosslinkable film of chitosan with 2-thiophenecarboxaldehyde. *Carbohydr. Polym.* **80**, 563–569 (2010).
46. Wang, H., Wang, J.-L., Yuan, S.-C., Pei, J. & Pei, W.-W. Novel highly fluorescent dendritic chiral amines: Synthesis and optical properties. *Tetrahedron* **61**, 8465–8474 (2005).
47. Yannai, S. *Dictionary of Food Compounds with CD-ROM* (CRC Press, 2012).
48. Bredie, W. L., Mottram, D. S. & Guy, R. C. Effect of temperature and pH on the generation of flavor volatiles in extrusion cooking of wheat flour. *J. Agric. Food Chem.* **50**, 1118–1125 (2002).
49. Mahadevan, K. & Farmer, L. Key odor impact compounds in three yeast extract pastes. *J. Agric. Food Chem.* **54**, 7242–7250 (2006).
50. Fujima, Y., Ikunaka, M., Inoue, T. & Matsumoto, J. Synthesis of (S)-3-(N-Methylamino)-1-(2-thienyl) propan-1-ol: Revisiting Eli Lilly's resolution—racemization—recycle synthesis of duloxetine for its robust processes. *Org. Process Res. Dev.* **10**, 905–913 (2006).
51. White, J. D., Juniku, R., Huang, K., Yang, J. & Wong, D. T. Synthesis of 1, 1-[1-Naphthyloxy-2-thiophenyl]-2-methylaminomethylcyclopropanes and their evaluation as inhibitors of serotonin, norepinephrine, and dopamine transporters. *J. Med. Chem.* **52**, 5872–5879 (2009).
52. Lazer, E. S., Wong, H. C., Possanza, G. J., Graham, A. G. & Farina, P. R. Antiinflammatory 2, 6-di-tert-butyl-4-(2-arylethenyl) phenols. *J. Med. Chem.* **32**, 100–104 (1989).
53. Hussain, H., Al-Harrasi, A., Al-Rawahi, A., Green, I. R. & Gibbons, S. Fruitful decade for antileishmanial compounds from 2002 to late 2011. *Chem. Rev.* **114**, 10369–10428 (2014).
54. Esteruelas, M. A., Lopez, A. M. & Olivan, M. Polyhydrides of platinum group metals: Nonclassical interactions and σ -bond activation reactions. *Chem. Rev.* **116**, 8770–8847 (2016).
55. Allais, C., Grassot, J.-M., Rodriguez, J. & Constantieux, T. Metal-free multicomponent syntheses of pyridines. *Chem. Rev.* **114**, 10829–10868 (2014).
56. Ahmadi, M., Chen, Z., Clarke, M. & Fedutenko, E. Comparison of kriging, machine learning algorithms and classical thermodynamics for correlating the formation conditions for CO₂ gas hydrates and semi-clathrates. *J. Nat. Gas Sci. Eng.* **84**, 103659. <https://doi.org/10.1016/j.jngse.2020.103659> (2020).
57. von Solms, N., Kouskoumvekaki, I. A., Michelsen, M. L. & Kontogeorgis, G. M. Capabilities, limitations and challenges of a simplified PC-SAFT equation of state. *Fluid Phase Equilib.* **241**, 344–353. <https://doi.org/10.1016/j.fluid.2006.01.001> (2006).
58. Sun, Y., Zuo, Z., Laaksonen, A., Lu, X. & Ji, X. How to detect possible pitfalls in ePC-SAFT modelling: Extension to ionic liquids. *Fluid Phase Equilib.* **519**, 112641. <https://doi.org/10.1016/j.fluid.2020.112641> (2020).
59. Sun, Y., Laaksonen, A., Lu, X. & Ji, X. How to detect possible pitfalls in ePC-SAFT modelling. 2. Extension to binary mixtures of 96 ionic liquids with CO₂, H₂S, CO, O₂, CH₄, N₂, and H₂. *Ind. Eng. Chem. Res.* **59**, 21579–21591. <https://doi.org/10.1021/acs.iecr.0c04485> (2020).
60. Chaparro, G. & Müller, E. A. Development of thermodynamically consistent machine-learning equations of state: Application to the Mie fluid. *J. Chem. Phys.* <https://doi.org/10.1063/5.0146634> (2023).
61. Ding, J. et al. Machine learning for molecular thermodynamics. *Chin. J. Chem. Eng.* **31**, 227–239. <https://doi.org/10.1016/j.cjche.2020.10.044> (2021).
62. Luo, W. et al. Bridging machine learning and thermodynamics for accurate pKa prediction. *JACS Au* **4**, 3451–3465. <https://doi.org/10.1021/jacsau.4c00271> (2024).
63. Cersonsky, R. K., Cheng, B., Kofke, D. & Müller, E. A. Machine learning for generating and analyzing thermophysical data: Where we are and where we're going. *J. Chem. Eng. Data* **69**, 2041–2043. <https://doi.org/10.1021/acs.jced.4c00207> (2024).
64. Kevrekidis, G. A. et al. Neural network representations of multiphase equations of State. *Sci. Rep.* **14**, 30288. <https://doi.org/10.1038/s41598-024-81445-4> (2024).
65. Jirasek, F. & Hasse, H. Perspective: Machine learning of thermophysical properties. *Fluid Phase Equilib.* **549**, 113206. <https://doi.org/10.1016/j.fluid.2021.113206> (2021).
66. Saha, A., Basu, A. & Banerjee, S. Enhancing thermal management systems: A machine learning and metaheuristic approach for predicting thermophysical properties of nanofluids. *Eng. Res. Express* **6**, 045537. <https://doi.org/10.1088/2631-8695/ad8536> (2024).
67. Shateri, A., Yang, Z. & Xie, J. Machine learning-based molecular dynamics studies on predicting thermophysical properties of ethanol–octane blends. *Energy Fuels* **39**, 1070–1090. <https://doi.org/10.1021/acs.energyfuels.4c05653> (2025).
68. Liang, W., Lu, G. & Yu, J. Machine-learning-driven simulations on microstructure and thermophysical properties of MgCl₂–KCl eutectic. *ACS Appl. Mater. Interfaces* **13**, 4034–4042. <https://doi.org/10.1021/acsami.0c20665> (2021).
69. Wang, H. & Chen, X. A comprehensive review of predicting the thermophysical properties of nanofluids using machine learning methods. *Ind. Eng. Chem. Res.* **61**, 14711–14730. <https://doi.org/10.1021/acs.iecr.2c02059> (2022).
70. Akilu, S., Sharma, K. V., Baheta, A. T., Kanti, P. K. & Paramasivam, P. Machine learning analysis of thermophysical and thermohydraulic properties in ethylene glycol- and glycerol-based SiO₂ nanofluids. *Sci. Rep.* **14**, 14829. <https://doi.org/10.1038/s41598-024-65411-8> (2024).
71. Kanti, P. K., Paramasivam, P., Wanatasanappan, V. V., Dhanasekaran, S. & Sharma, P. Experimental and explainable machine learning approach on thermal conductivity and viscosity of water based graphene oxide based mono and hybrid nanofluids. *Sci. Rep.* **14**, 30967. <https://doi.org/10.1038/s41598-024-81955-1> (2024).
72. Ahmad, H. et al. Numerical and machine learning based evaluation of ethylene glycol based hybrid nano-structured (TiO₂-SWCNTs) fluid flow. *Sci. Rep.* **15**, 6084. <https://doi.org/10.1038/s41598-025-88789-5> (2025).
73. Gomaa, S. et al. Machine learning prediction of methane, nitrogen, and natural gas mixture viscosities under normal and harsh conditions. *Sci. Rep.* **14**, 15155. <https://doi.org/10.1038/s41598-024-64752-8> (2024).
74. Abdollahzadeh, M. et al. Estimating the density of deep eutectic solvents applying supervised machine learning techniques. *Sci. Rep.* **12**, 4954. <https://doi.org/10.1038/s41598-022-08842-5> (2022).
75. Yarahmadi, A., Rashedi, A. & Bemani, A. Machine learning based estimation of density of binary blends of cyclohexanes in normal alkanes. *Sci. Rep.* **15**, 8469. <https://doi.org/10.1038/s41598-025-92608-2> (2025).
76. Huwaimel, B., Alanazi, J., Alanazi, M., Alharby, T. N. & Alshammari, F. Computational models based on machine learning and validation for predicting ionic liquids viscosity in mixtures. *Sci. Rep.* **14**, 31857. <https://doi.org/10.1038/s41598-024-82989-1> (2024).
77. Kiani, S. et al. Modeling of ionic liquids viscosity via advanced white-box machine learning. *Sci. Rep.* **14**, 8666. <https://doi.org/10.1038/s41598-024-55147-w> (2024).
78. Antón, V., Muñoz-Embid, J., Artigas, H., Artal, M. & Lafuente, C. Thermophysical properties of oxygenated thiophene derivatives: Experimental data and modelling. *J. Chem. Thermodyn.* **113**, 330–339. <https://doi.org/10.1016/j.jct.2017.07.008> (2017).
79. Antón, V., Lomba, L., Cea, P., Giner, B. & Lafuente, C. Densities at high pressures and derived properties of thiophenes. *J. Chem. Thermodyn.* **109**, 16–22. <https://doi.org/10.1016/j.jct.2016.08.005> (2017).
80. Antón, V., Artigas, H., Muñoz-Embid, J., Artal, M. & Lafuente, C. Thermophysical study of 2-acetylthiophene: Experimental and modelled results. *Fluid Phase Equilib.* **433**, 126–134. <https://doi.org/10.1016/j.fluid.2016.10.026> (2017).

81. Piña-Martínez, A., Privat, R. & Jaubert, J.-N. Use of 300,000 pseudo-experimental data over 1800 pure fluids to assess the performance of four cubic equations of state: SRK, PR, tc-RK, and tc-PR. *AIChE J.* **68**, e17518. <https://doi.org/10.1002/aic.17518> (2022).
82. Yan, T., Shen, S.-L., Zhou, A. & Chen, X. Prediction of geological characteristics from shield operational parameters by integrating grid search and K-fold cross validation into stacking classification algorithm. *J. Rock Mech. Geotech. Eng.* **14**, 1292–1303 (2022).
83. Ranjan, G., Verma, A. K. & Radhika, S. In *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)* 1–5 (IEEE).
84. Mangkunegara, I. S. & Purwono, P. In *2022 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom)* 427–432.
85. Pedregosa, F. et al. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
86. Belete, D. M. & Huchaiah, M. D. Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results. *Int. J. Comput. Appl.* **44**, 875–886 (2022).
87. Loh, W. Y. Classification and regression trees. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **1**, 14–23 (2011).
88. Rokach, L. & Maimon, O. Decision trees. *Data Min. Knowl. Discov. Handbook* 165–192 (2005).
89. Ding, Y. & Jin, Y. Development of advanced hybrid computational model for description of molecular separation in liquid phase via polymeric membranes. *J. Mol. Liq.* **396**, 123999. <https://doi.org/10.1016/j.molliq.2024.123999> (2024).
90. Song, H., Shao, H., Zhang, Y. & Wang, X. Advancing nanomedicine production via green method: Modeling and simulation of pharmaceutical solubility at different temperatures and pressures. *J. Mol. Liq.* **411**, 125806. <https://doi.org/10.1016/j.molliq.2024.125806> (2024).
91. Freund, Y. & Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **55**, 119–139. <https://doi.org/10.1006/jcss.1997.1504> (1997).
92. Hong, H. et al. Landslide susceptibility mapping using J48 decision tree with AdaBoost, bagging and rotation forest ensembles in the guangchang area (China). *CATENA* **163**, 399–413 (2018).
93. Kumar, R., Rai, B. & Samui, P. Machine learning techniques for prediction of failure loads and fracture characteristics of high and ultra-high strength concrete beams. *Innov. Infrastruct. Solut.* **8**, 219 (2023).
94. Elith, J., Leathwick, J. R. & Hastie, T. A working guide to boosted regression trees. *J. Anim. Ecol.* **77**, 802–813 (2008).
95. James, G., Witten, D., Hastie, T. & Tibshirani, R. *An Introduction to Statistical Learning* Vol. 112 (Springer, 2013).
96. Chong, E., Han, C. & Park, F. C. Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies. *Expert Syst. Appl.* **83**, 187–205 (2017).
97. Asim, M., Rashid, A. & Ahmad, T. Scour modeling using deep neural networks based on hyperparameter optimization. *ICT Express* **8**, 357–362 (2022).
98. Krauss, C., Do, X. A. & Huck, N. Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *Eur. J. Oper. Res.* **259**, 689–702 (2017).
99. Ke, G. et al. Lightgbm: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **30** (2017).
100. Ma, X. et al. Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning. *Electron. Commer. Res. Appl.* **31**, 24–39 (2018).
101. Fan, J. et al. Light gradient boosting machine: An efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data. *Agric. Water Manag.* **225**, 105758 (2019).
102. Yang, X., Dindoruk, B. & Lu, L. A comparative analysis of bubble point pressure prediction using advanced machine learning algorithms and classical correlations. *J. Petrol. Sci. Eng.* **185**, 106598 (2020).
103. Sun, X., Liu, M. & Sima, Z. A novel cryptocurrency price trend forecasting model based on LightGBM. *Financ. Res. Lett.* **32**, 101084 (2020).
104. Zheng, H., Mahmoudzadeh, A., Amiri-Ramshah, B. & Hemmati-Sarapardeh, A. Modeling viscosity of CO₂-N₂ gaseous mixtures using robust tree-based techniques: Extra tree, random forest, GBoost, and LightGBM. *ACS Omega* **8**, 13863–13875 (2023).
105. Arik, S. Ö. & Pfister, T. In *Proceedings of the AAAI Conference on Artificial Intelligence* 6679–6687.
106. Mao, S., Liang, Y., Sun, W. & Li, Q. Interpretable transfer learning for small sample coal and gas outburst risk identification using TabNet. *Energy Sci. Eng.* **13**, 909–925. <https://doi.org/10.1002/ese3.2049> (2025).
107. McDonnell, K., Murphy, F., Sheehan, B., Masello, L. & Castignani, G. Deep learning in insurance: Accuracy and model interpretability using TabNet. *Expert Syst. Appl.* **217**, 119543 (2023).
108. Razavi-Termeh, S. V., Sadeghi-Niaraki, A., Sorooshian, A., Abuhmed, T. & Choi, S.-M. Spatial mapping of land susceptibility to dust emissions using optimization of attentive Interpretable Tabular Learning (TabNet) model. *J. Environ. Manag.* **358**, 120682. <https://doi.org/10.1016/j.jenvman.2024.120682> (2024).
109. Wang, H. et al. Enhancing predictive accuracy for urinary tract infections post-pediatric pyeloplasty with explainable AI: a#n ensemble TabNet approach. *Sci. Rep.* **15**, 2455. <https://doi.org/10.1038/s41598-024-82282-1> (2025).
110. Rousseeuw, P. J. & Leroy, A. M. *Robust Regression and Outlier Detection* (Wiley, 2005).
111. Goodall, C. R. 13 Computation using the QR decomposition (1993).
112. Sheikhshoei, A. H., Khoshshima, A. & Zabihzadeh, D. Predicting the heat capacity of strontium-praseodymium oxyisilicate SrPr₄(SiO₄)₃O using machine learning, deep learning, and hybrid models. *Chem. Thermodyn. Thermal Anal.* **17**, 100154. <https://doi.org/10.1016/j.cta.2024.100154> (2025).

Author contributions

Amirhossein Sheikhshoei: Investigation, Writing—original draft, Software, Validation, Data curation, Formal analysis. Ali Khoshshima: Supervision, Project administration, Conceptualization, Methodology, Visualization, Writing—review & editing.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-025-09600-z>.

Correspondence and requests for materials should be addressed to A.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025