



PREDICTING THERMAL PROPERTIES OF CHEMICAL SUBSTANCES FROM MOLECULAR STRUCTURE AND COMPOSITION

Vinayak Gupta

Department of Chemistry, University of Rajasthan , Jaipur, 302004, Vinayakg2796@gmail.com, <https://orcid.org/0000-0003-1011-571X>

Received: - 11/06/2026, Revised: - 08/08/2026, Accepted: - 16/09/2026, Published: - 30/09/2026

ABSTRACT

Knowing the thermal properties, from molecular structure, can aid in chemical screening, process design and thermodynamic modelling, and help to minimize reliance on expensive experimental measurements. This study has created an interpretable machine-learning model to predict normal boiling points and critical temperatures from the molecular structure and composition. Molecular descriptors used were based on the SMILES representation and were: molecular weight; carbon count; heavy atom count; aromatic atom count; TPSA (topological polar surface area); and the count of rotatable bonds. Random Forest, Extra Trees and Gradient Boosting were trained and tested with 80/20 split respectively to get the separate predictive models. The performance of the model was tested in terms of the coefficient of determination, mean absolute error and root mean squared error, and the importance of the descriptors was determined by permutation importance. Extra Trees performed the best for both properties, $R^2=0.955$, MAE = 10.78 K and RMSE = 18.39 K for normal boiling point; as well as $R^2=0.908$, MAE = 18.41 K and RMSE = 34.43 K for critical temperature. Carbon count was the most influential descriptor for both properties, followed by descriptors representing molecular size, composition, aromaticity, polarity and flexibility. The results show that the descriptor-based ensemble learning is an accurate and efficient procedure to estimate the thermal properties with a high degree of chemical interpretability. This framework may thus be used to provide a quick screening and preliminary evaluation of chemical substances when direct measurement of the thermal properties is not available or is limited.

KEYWORDS

Green chemis Thermal property prediction; normal boiling point; critical temperature; molecular descriptors; ensemble learning

1. INTRODUCTION

Thermal properties such as normal boiling point and critical temperature are fundamental physicochemical characteristics governing the design, selection, and safe use of chemical substances in industrial, energy, refrigeration, separation, and process-engineering applications. Accurate knowledge of these properties is essential for vapour–liquid equilibrium calculations, equipment design, solvent selection, refrigerant screening, and thermodynamic modelling. However, direct experimental determination can be expensive, time-consuming, and technically challenging, particularly for unstable, hazardous, or scarcely available compounds. Consequently, computational approaches capable of predicting thermal properties from molecular structure and composition have gained considerable importance. Machine-learning and quantitative structure–property relationship (QSPR) methods have demonstrated strong predictive performance for physicochemical and thermophysical properties when appropriate molecular descriptors and robust algorithms are employed (Devotta & Chelani, 2022; Qu et al., 2022).

Traditional property-estimation approaches commonly rely on group-contribution methods, empirical correlations, and manually derived molecular descriptors. These methods are valuable because they associate molecular fragments with macroscopic thermodynamic behaviour, but their predictive performance may decline when molecules contain complex structural motifs or nonlinear fragment interactions. Villazón-León et al. (2022) demonstrated that first-order group-contribution methods can estimate normal boiling point and critical properties effectively, although their accuracy depends on molecular coverage and structural representation. More recently, hybrid strategies combining group-contribution principles with machine learning have been investigated to improve flexibility and predictive accuracy. Mann et al. (2022) developed an interpretable machine-learning framework for thermodynamic property estimation from molecular representations, while Lamprea et al. (2024) demonstrated that deep-learning-assisted group-contribution modelling can capture complex structure–property relationships.

QSPR modelling provides another important framework by statistically associating measurable physicochemical properties with molecular descriptors. Descriptor selection is particularly important because molecular weight, elemental composition, connectivity, polarity, aromaticity, and flexibility can influence intermolecular interactions and phase-transition behaviour. Comesana et al. (2022) emphasized the importance of systematic descriptor selection for physicochemical property prediction. Similarly, Fissa et al. (2023) demonstrated that QSPR models incorporating artificial neural networks can effectively estimate the critical properties of hydrocarbons. Niu et al. (2024) further demonstrated the broader applicability of descriptor-based QSPR approaches for estimating properties of organic compounds.

Machine-learning methods have expanded these capabilities by modelling highly nonlinear relationships between molecular structure and thermophysical properties. Random forests, ensemble trees, artificial neural networks, gradient boosting, and related algorithms have been investigated for predicting boiling point and critical properties. Que et al. (2022) predicted the critical temperature of organic refrigerants using molecular topology, while Liu et al. (2025) demonstrated the effectiveness of machine learning for boiling-point prediction of potential insulating gases. Rathod et al. (2024) similarly showed that data-driven algorithms can successfully predict thermophysical properties of alkanes and refrigerants. These findings demonstrate the potential of nonlinear and ensemble models for complex structure–property relationships.

Deep learning has further shifted molecular-property prediction from handcrafted descriptors toward structure-aware representations. Graph neural networks represent atoms as nodes and chemical bonds as edges, enabling molecular connectivity and topology to be learned directly. Wang and Hu (2023) applied graph neural networks and transfer learning to predict normal boiling point and critical temperature of refrigerants. Qu et al. (2024) likewise demonstrated the usefulness of graph neural networks for molecular-property prediction using high-quality experimental data. Geometry-enhanced molecular learning has additionally improved the representation of structural information relevant to property prediction (Fang et al., 2022). Molecular language modelling provides another strategy for extracting chemical information from sequence-based representations such as SMILES. Born and Manica (2023) demonstrated that transformer-based molecular language models can integrate property regression and molecular sequence generation, while Ross et al. (2022) showed that large-scale chemical language representations capture molecular structure and physicochemical characteristics.

Despite these advances, descriptor-based ensemble models remain attractive because of their computational efficiency, interpretability, and suitability for medium-sized chemical datasets. Such approaches enable direct evaluation of chemically meaningful variables while maintaining strong predictive capability. Modified group-contribution descriptors have demonstrated utility for predicting boiling point and critical properties of organic compounds (Guo et al., 2026), while machine-learning-assisted QSPR modelling has shown strong potential for normal boiling-point prediction from molecular information (Yu et al., 2026).

Accordingly, the present study investigates the prediction of normal boiling point and critical temperature from molecular structure and composition. Molecular descriptors derived from chemical structures are evaluated using multiple ensemble regression algorithms, with predictive performance compared using standard regression metrics. The study also examines the relative contribution of molecular descriptors to thermal-property prediction, providing an interpretable framework for understanding the molecular characteristics associated with boiling and critical behaviour.

2. METHODOLOGY

2.1 Data Source and Extraction

The dataset of thermal properties supplied with the study included the data for chemical substances encoded as molecular structures in the form of Simplified Molecular Input Line Entry System (SMILES) strings (Burkhardt et al., 2026). The analysis was limited to the two thermal properties for which there were data available that could be broken down into structures and subsequently reported in the results: normal boiling point (T_b) and critical temperature (T_c).

2.2 Data Preparation and Thermal-Property Reconstruction

Normal boiling point and critical temperature were calculated using the group-contribution routines and model parameters provided in the data set. Only materials for which the provided routines could provide positive values for the thermal properties were kept. According to this process, the last analytical datasets were 5,591 substances for normal boiling point and 1,454 substances for critical temperature. CAS identifiers were used to find the overlap between the two sets of properties. The resulting thermal-property distributions and molecular-weight ranges were then described.

2.3 Molecular Structure and Composition Descriptors

The SMILES representation of each retained substance was used to calculate the molecular descriptors directly, for the characterization of the structural and compositional properties relevant to the prediction of thermal properties. Molecular weight, carbon number, number of heavy atoms, number of aromatic atoms, topological polar surface area (TPSA) and number of rotatable bonds were assessed as the descriptors for the predictive analysis. The following variables were used as predictor variables for normal boiling point and critical temperature: molecular size, elemental composition, aromaticity, polarity, and structural flexibility.

2.4 Predictive Modelling and Performance Evaluation

Two separate regression models were prepared for normal boiling point and for critical temperature. The three ensemble regression algorithms were tested: Random Forest, Extra Trees and Gradient Boosting. When comparing the three algorithms, for each thermal property, the substances were randomly split into 80% training set and 20% held-out test set, and the same random split strategy was used for each of the three different algorithms. The coefficient of determination (R^2), mean absolute error (MAE) and root mean squared error (RMSE) were used as metrics to quantify predictive performance on the held-out test set. The highest predictive performance held-out model was employed in the following prediction-error and descriptor-importance analyses.

2.5 Prediction Error and Molecular Descriptor Importance

The held-out test observations from the selected Extra Trees models were compared with the reference values for the thermal properties. The mean error, median error, mean absolute error, root mean square error and the 90th percentile of absolute prediction error at the normal boiling point and critical temperature were used to summarise the prediction error. The agreement of the predicted values with the corresponding reference values was also assessed within the ranges of temperatures observed.

The relative contribution of molecular characteristics to prediction was calculated using the fitted Extra Trees models. The importance of the descriptors was reported as the drop in model (R^2) when permuting one individual descriptor. The molecular descriptors with the highest level of decrease in the (R^2) value were thought to be most important in the prediction of normal boiling point or critical temperature.

2.6 Ethical Considerations

This study was a secondary analysis of an existing chemical-property data set, so no human subjects, animals, personal information, or identifiable clinical information was involved. The analysis was restricted to the

chemical structure information, model information and thermal-property data provided in the dataset and the source data was applied solely for scientific analysis and reporting.

3. RESULTS

3.1 Dataset Coverage and Thermal-Property Space

Structure-resolved substance lists were obtained for the normal boiling point (Tb) and the critical temperature (Tc) by extracting them from the archive, and molecular structures were encoded in the form of SMILES. Once duplicates were removed and molecules for which the CAS-provided group-contribution routines generated valid, positive reference values were retained, there were 5,591 molecules to be used in the creation of Tb and 1,454 molecules to be used in the creation of Tc. The two property sets had 1,260 common CAS identifiers, while the combined datasets contained 5,785 unique CAS substances. The Tb reference values covered 197.59–918.82 K (mean 440.93 K), whereas Tc covered 228.35–963.81 K (mean 623.55 K). The chemical range of molecular weight was also wide, between 40.07 and 891.50 g/mol for Tb and 34.03 and 838.11 g/mol for Tc (Table 1).

Table 1. Structural and thermal-property coverage of the analyzed substance sets

Property	n	Mean (K)	SD (K)	Min (K)	Median (K)	Max (K)	Mean MW (g/mol)
Tb	5,591	440.93	86.64	197.59	434.34	918.82	157.53
Tc	1,454	623.55	119.08	228.35	622.88	963.81	152.57

Figure 1 illustrates that both prediction tasks utilize the same temperature region, but different and relevant parts. The Tb values were relatively grouped in region of intermediate temperatures, whereas the Tc was concentrated at higher temperatures with a large scatter. This large test set is used to evaluate how well the nonlinear structure–property relationships of the models supplied for thermal properties can be captured by a compact molecular description.

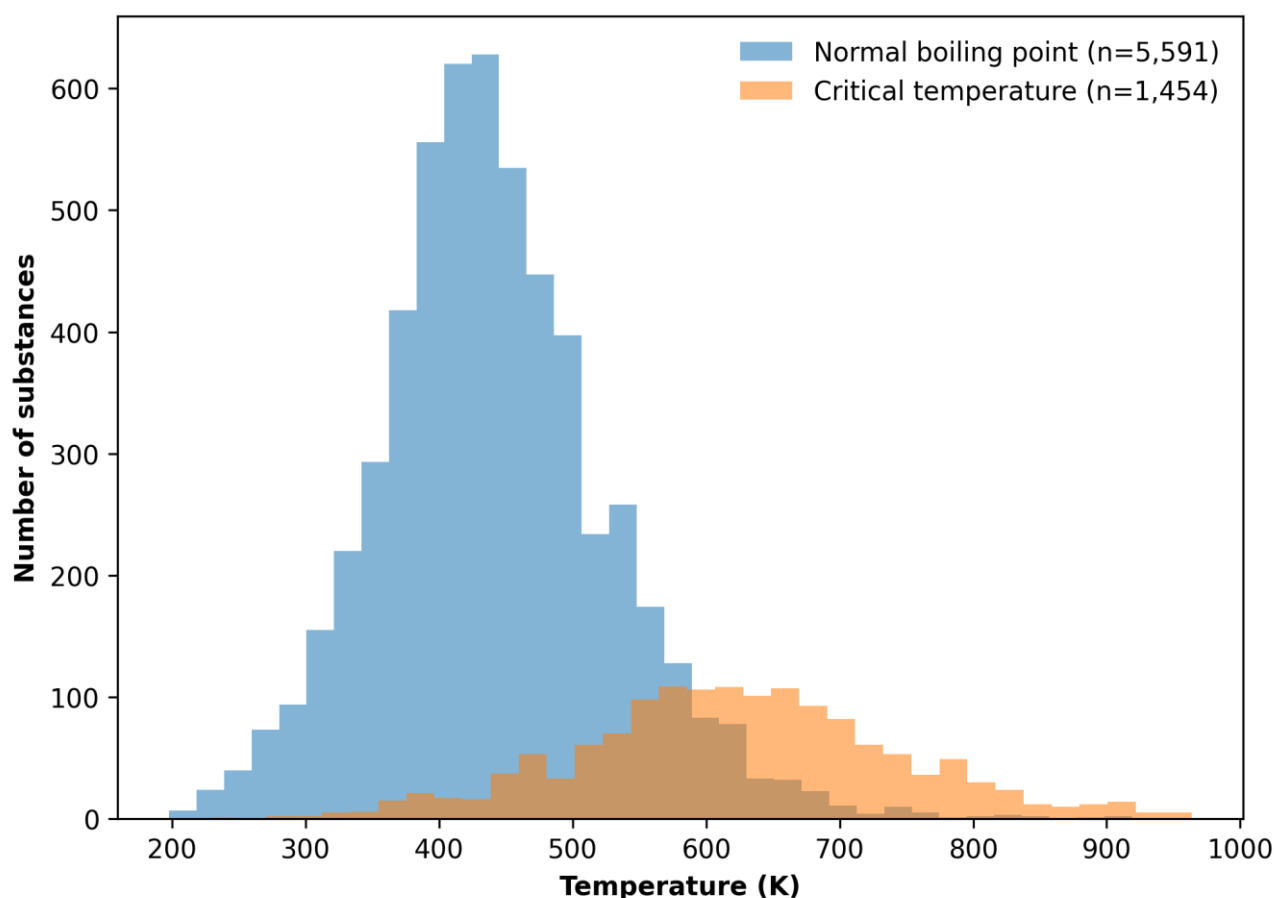


Figure 1. Distribution of reconstructed reference normal boiling point (Tb) and critical temperature (Tc) values for substances successfully processed from the supplied archive

3.2 Predictive Performance of Molecular-Descriptor Models

Three ensemble regression algorithms were tested on the same random 80:20 split and molecular descriptors were calculated directly from SMILES. Extra Trees performed best at the held-out test sets of both properties. For Tb, Extra Trees achieved $R^2 = 0.955$, MAE = 10.78 K, and RMSE = 18.39 K; Random Forest was close ($R^2 = 0.946$, MAE = 12.22 K), whereas Gradient Boosting showed lower accuracy ($R^2 = 0.891$). For Tc, Random Forest achieved $R^2 = 0.894$, whereas Gradient Boosting achieved $R^2 = 0.852$ (Table 2; Figure 2).

Table 2. Held-out predictive performance for normal boiling point and critical temperature

Property	Model	R^2	MAE (K)	RMSE (K)
Tb	Random Forest	0.946	12.22	20.17
Tb	Extra Trees	0.955	10.78	18.39
Tb	Gradient Boosting	0.891	20.16	28.59
Tc	Random Forest	0.894	22.12	37.12
Tc	Extra Trees	0.908	18.41	34.43
Tc	Gradient Boosting	0.852	29.21	43.80

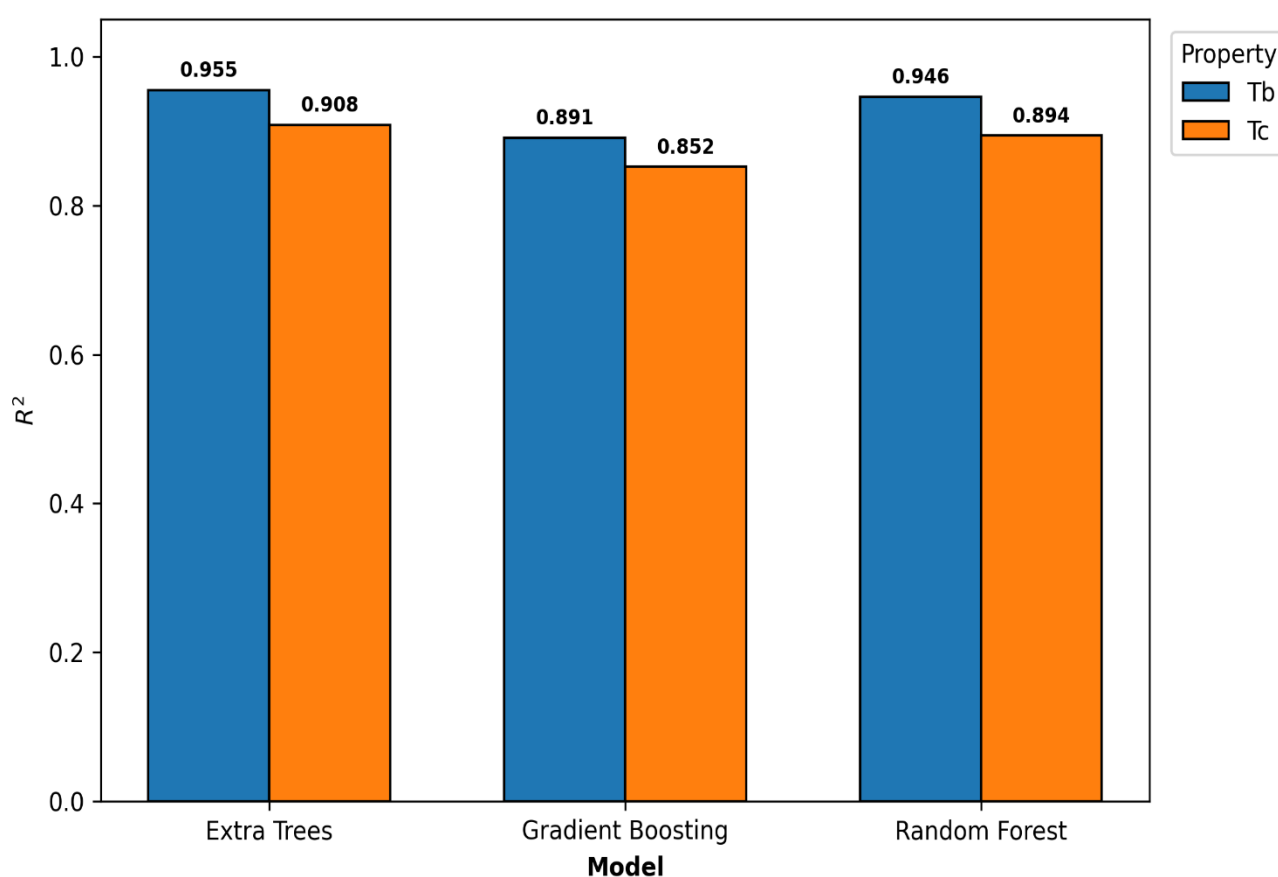


Figure 2. Comparison of held-out R^2 across ensemble regression models for normal boiling point (Tb) and critical temperature (Tc)

3.3 Agreement Between Reference and Predicted Thermal Properties

Agreement between the reference and predicted thermal properties was evaluated using the held-out test sets. Good agreement was obtained in the predicted-versus-reference plot over most of the temperature-range observed (Figure 3). The extra trees predictions for Tb were also grouped around the identity line and the MAE and RMSE obtained for that endpoint was less. The residuals for the Tc predictions were also generally along the identity line for the central range but were larger for a few high temperature substances. There was little systematic directional error for the two endpoints, as evidenced by the near-zero mean prediction error (Table 3).

Table 3. Error characteristics of the Extra Trees model on the held-out test sets

Property	Test n	Mean error (K)	Median error (K)	MAE (K)	RMSE (K)	90th percentile absolute error (K)
Tb	1119	-2.41	-0.00	10.78	18.39	28.72
Tc	291	-4.56	-0.00	18.41	34.43	48.91

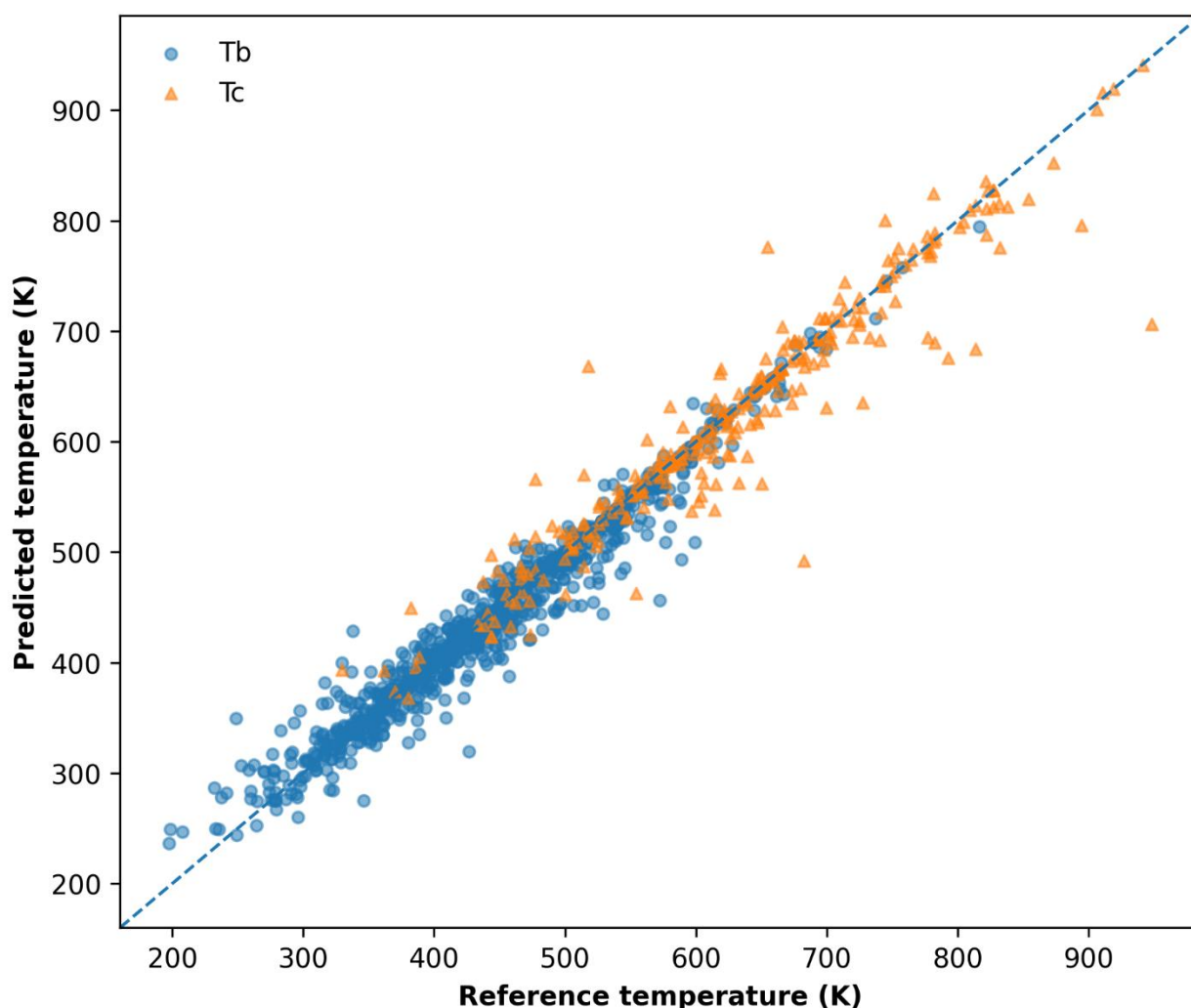


Figure 3. Extra Trees predicted versus reference temperatures for the held-out normal boiling point (Tb) and critical temperature (Tc) sets. The dashed line represents perfect agreement

3.4 Molecular Drivers of Thermal-Property Prediction

The dominant signals identified in the Extra Trees models that were fitted by permutation analysis were molecular size and composition. The most significant descriptors with a decrease in R^2 were carbon count for Tb (0.481) and Tc (0.331). The next most significant group consisted of heavy-atom count, aromatic-atom count, molecular weight and topological polar surface area (TPSA). The contribution of polarity and structural flexibility was a little higher for Tc than any of the other descriptors: TPSA was second followed by heavy-atom count, aromatic-atom count, rotatable bonds, and molecular weight (Table 4; Figure 4). Overall, molecular size was influential for both properties, while polarity and structural flexibility contributed relatively more strongly to critical-temperature prediction.

Table 4. Leading molecular descriptors from permutation importance of the Extra Trees models

Property	Descriptor	Permutation importance (ΔR^2)
Tb	Carbon count	0.481
Tb	Heavy-atom count	0.165
Tb	Aromatic-atom count	0.117
Tb	Molecular weight	0.111
Tb	TPSA	0.093
Tb	Rotatable bonds	0.063
Tc	Carbon count	0.331
Tc	TPSA	0.175
Tc	Heavy-atom count	0.155
Tc	Aromatic-atom count	0.145
Tc	Rotatable bonds	0.103
Tc	Molecular weight	0.099

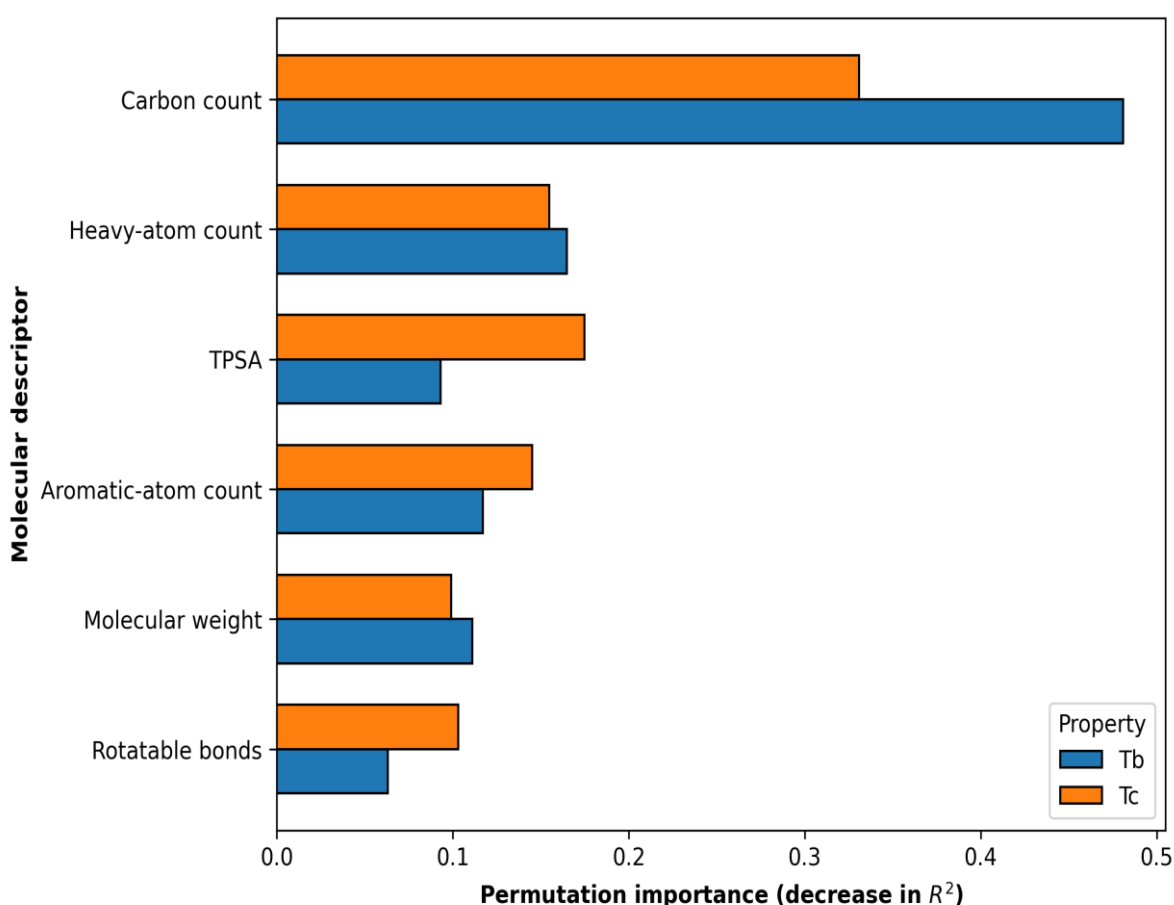


Figure 4. Permutation importance of the molecular descriptors for Extra Trees prediction of normal boiling point (Tb) and critical temperature (Tc)

4. Discussion

The present study demonstrated that molecular structure and composition provide strong predictive information for estimating normal boiling point (Tb) and critical temperature (Tc). Among the three ensemble algorithms evaluated, Extra Trees achieved the best performance, with $R^2 = 0.955$, MAE = 10.78 K, and RMSE = 18.39 K for Tb and $R^2 = 0.908$, MAE = 18.41 K, and RMSE = 34.43 K for Tc. These findings are consistent with previous studies showing that nonlinear machine-learning methods can effectively map molecular information to thermophysical properties (He et al., 2024; Rathod et al., 2024).

The stronger performance of Extra Trees compared with Gradient Boosting suggests that randomized ensemble partitioning was effective in capturing complex interactions among the molecular descriptors. Random Forest

also performed well, with R^2 values of approximately 0.946 for Tb and 0.894 for Tc, indicating that tree-based ensembles were generally suitable for the investigated structure–property relationships. Similar benefits of nonlinear modelling have been reported for boiling-point estimation using machine learning and QSPR approaches (Kim et al., 2024; Sun et al., 2023). The present findings also agree with recent studies demonstrating the use of molecular topology and learned molecular representations for thermophysical-property prediction (Buterez et al., 2024; Wang & Hu, 2023).

Boiling-point prediction was more accurate than critical-temperature prediction. This difference may reflect the greater structural complexity of critical-state behaviour, where intermolecular forces, molecular geometry, polarity, and molecular size collectively influence the temperature at which liquid and vapour phases become indistinguishable. Previous QSPR studies have similarly shown that critical temperature is governed by multiple nonlinear structural factors (Pan et al., 2023; Huoyu et al., 2022a). Critical temperature has also been linked to molecular constitution and connectivity in ketones, aldehydes, and ethers (Huoyu et al., 2022b).

Carbon count was the most influential descriptor for both properties, with permutation-importance values of approximately 0.481 for Tb and 0.331 for Tc. Molecular weight, topological polar surface area (TPSA), heavy-atom count, aromatic-atom count, and rotatable bonds also contributed to prediction. These findings support the expectation that molecular size, polarity, aromaticity, and flexibility influence intermolecular interactions and phase-transition behaviour. Descriptor-based QSPR studies have likewise identified molecular topology and composition as informative predictors of critical properties (Fissa et al., 2023; Que et al., 2022). The usefulness of chemically meaningful descriptors is also consistent with systematic descriptor-selection approaches for physicochemical modelling (Comesana et al., 2022).

TPSA showed greater relative importance for Tc than for Tb, suggesting a stronger contribution of molecular polarity to variation in critical temperature. This observation is consistent with studies showing that structurally informed molecular representations can improve the modelling of complex chemical properties (Kong et al., 2022; Fang et al., 2022). Graph-based methods can additionally identify substructural contributions that may be difficult to capture using conventional descriptors (Wu et al., 2023). Such interpretability remains important because strong predictive performance alone does not necessarily provide chemically meaningful explanations of structure–property relationships (Proietti et al., 2024; Yang & Li, 2023).

The present descriptor-based ensemble framework is computationally simpler than graph neural networks, molecular transformers, and large-scale self-supervised representations. Nevertheless, recent developments in contrastive learning and graph transformers indicate further opportunities for improving thermal-property prediction beyond manually calculated descriptors (Wang et al., 2022; Rong et al., 2020). Molecular language representations can also encode structural information not directly captured by traditional descriptors (Ross et al., 2022). Reviews of graph-based molecular prediction further emphasize the importance of sufficiently large and chemically diverse datasets for effective representation learning (Wieder et al., 2020).

The findings also support the continuing integration of conventional thermodynamic estimation methods with modern machine learning. Group-contribution approaches remain useful because they preserve chemical interpretability, while data-driven extensions can improve flexibility and predictive accuracy (Villazón-León et al., 2025; Devotta & Chelani, 2022). Recent studies of boiling point and critical properties further support combining structural descriptors with advanced computational modelling (Guo et al., 2026). Likewise, assessment of normal boiling-point estimation methods highlights the importance of selecting predictive approaches according to molecular class and available structural information (Yu et al., 2024). Overall, the results support ensemble learning as an accurate and interpretable approach for predicting thermal properties directly from molecular structure and composition.

5. Conclusion

The results of this study showed that molecular structure and composition can be used to predict certain thermal properties such as normal boiling point and critical temperature of chemical substances with good reliability. The best predictive performance for the ensemble regression models has been achieved by Extra Trees with an R^2 of 0.955 and an MAE of 10.78 K for the normal boiling point and an R^2 of 0.908 and an MAE of 18.41 K for the critical temperature. The relatively poorer results for critical temperature may suggest that this property is not as easily described from the selected structural and compositional properties. Additionally, permutation importance analysis revealed that the carbon count, heavy atom count, aromatic atom count, molecular weight, TPSA, and rotatable bond values were important in prediction. For both thermal properties, carbon count was the most important descriptor, and for critical temperature, polarity-related information was more important. The results presented here show that interpretable molecular descriptors, combined with ensemble learning, can be a viable alternative to more computationally expensive representation-learning methods. The developed

framework provides a practical approach for quick estimation of thermal properties in cases of lack of experimental data or when it is expensive. The methodology could be expanded in future research to include wider classes of chemicals, other thermophysical data, and more sophisticated molecular structures to increase the generalizability and predict robustness.

References

1. Born, J., & Manica, M. (2023). Regression transformers enable concurrent sequence regression and generation for molecular language modelling. *Nature Machine Intelligence*, 5(4), 432-444.
2. Buterez, D., Janet, J. P., Kiddle, S. J., Oglic, D., & Lió, P. (2024). Transfer learning with graph neural networks for improved molecular property prediction in the multi-fidelity setting. *Nature communications*, 15(1), 1517.
3. Comesana, A. E., Huntington, T. T., Scown, C. D., Niemeyer, K. E., & Rapp, V. H. (2022). A systematic method for selecting molecular descriptors as features when training models for predicting physiochemical properties. *Fuel*, 321, 123836.
4. Devotta, S., & Chelani, A. (2022). Unified artificial neural network-group contribution method for predictions of normal boiling point and critical temperature of refrigerants and related compounds. *International Journal of Refrigeration*, 140, 112-124.
5. Fang, X., Liu, L., Lei, J., He, D., Zhang, S., Zhou, J., ... & Wang, H. (2022). Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence*, 4(2), 127-134.
6. Fissa, M. R., Lahiouel, Y., Khaouane, L., & Hanini, S. (2023). Development of QSPR-ANN models for the estimation of critical properties of pure hydrocarbons. *Journal of Molecular Graphics and Modelling*, 121, 108450.
7. Guo, Y., Li, J., Tang, J., Zhu, M., & Li, J. (2026). Prediction of Boiling Point and Critical Properties of Acyclic Carbon-, Hydrogen-, and Oxygen-Containing Organic Compounds Using an Eigenvector-Centrality-Corrected Group Contribution Descriptor. *Chinese Journal of Chemical Engineering*.
8. He, Y., Feng, Y., Qiu, L., & Tang, D. (2024). Data-driven approach augmented by attention mechanism in critical and boiling thermophysical properties prediction of fluorine/chlorine-based refrigerants. *Energy*, 306, 132490.
9. Huoyu, R., Zhiqiang, Z., Zhanggao, L., & Zhenzhen, X. (2022a). QSPR models for the critical temperature and pressure of cycloalkanes. *Chemical Physics Letters*, 808, 140088.
10. Huoyu, R., Zhiqiang, Z., Zhanggao, L., & Zhenzhen, X. (2022b). Quantitative structure–property relationship for the critical temperature of saturated monobasic ketones, aldehydes, and ethers with molecular descriptors. *International Journal of Quantum Chemistry*, 122(18), e26950.
11. Kim, S. Y., Jeon, I., & Kang, S. J. (2024). Integrating data science and machine learning to chemistry education: Predicting classification and boiling point of compounds. *Journal of Chemical Education*, 101(4), 1771-1776.
12. Kong, Y., Zhao, X., Liu, R., Yang, Z., Yin, H., Zhao, B., ... & Yan, A. (2022). Integrating concept of pharmacophore with graph neural networks for chemical property prediction and interpretation. *Journal of Cheminformatics*, 14(1), 52.
13. Lamprea, T., Orozco, G. A., & Hortua, H. J. (2024). A bayesian deep learning-based approach to group contribution methods: a study on cannabinoids and terpenes. *Fluid Phase Equilibria*, 583, 114116.
14. Liu, W., Zha, J., Ling, M., Li, D., Shen, K., & Cheng, L. (2025). A boiling point prediction method based on machine learning for potential insulating gases. *Chemical Physics*, 588, 112447.
15. Mann, V., Brito, K., Gani, R., & Venkatasubramanian, V. (2022). Hybrid, interpretable machine learning for thermodynamic property estimation using grammar2vec for molecular representation. *Fluid Phase Equilibria*, 561, 113531.
16. Niu, H., Zhang, Y., Jia, Q., Wang, Q., & Yan, F. (2024). Property estimation of organic compounds based on QSPR models with norm indices. *Chemical Engineering Science*, 288, 119835.
17. Pan, Y., Yang, F., Zhang, H., Yan, Y., Ping, X., Yu, M., & Yang, A. (2023). New QSPR models for predicting critical temperature of binary organic mixtures using linear and nonlinear methods. *Fluid Phase Equilibria*, 575, 113916.
18. Proietti, M., Ragno, A., Rosa, B. L., Ragno, R., & Capobianco, R. (2024). Explainable AI in drug discovery: self-interpretable graph neural network for molecular property prediction using concept whitening. *Machine Learning*, 113(4), 2013-2044.
19. Qu, C., Kearsley, A. J., Schneider, B. I., Keyrouz, W., & Allison, T. C. (2022). Graph convolutional neural network applied to the prediction of normal boiling point. *Journal of Molecular Graphics and Modelling*, 112, 108149.
20. Qu, C., Schneider, B. I., Kearsley, A. J., Keyrouz, W., & Allison, T. C. (2024). Applying graph neural

- network models to molecular property prediction using high-quality experimental data. *Artificial Intelligence Chemistry*, 2(1), 100050.
21. Que, Y., Ren, S., Hu, Z., & Ren, J. (2022). Machine learning prediction of critical temperature of organic refrigerants by molecular topology. *Processes*, 10(3), 577.
 22. Rathod, K., Ravula, S. C., Kommireddi, P. S. C., Thangeda, R., Kikugawa, G., & Chilukoti, H. K. (2024). Predicting thermophysical properties of alkanes and refrigerants using machine learning algorithms. *Fluid Phase Equilibria*, 578, 114016.
 23. Rong, Y., Bian, Y., Xu, T., Xie, W., Wei, Y., Huang, W., & Huang, J. (2020). Self-supervised graph transformer on large-scale molecular data. *Advances in neural information processing systems*, 33, 12559-12571.
 24. Ross, J., Belgodere, B., Chenthamarakshan, V., Padhi, I., Mroueh, Y., & Das, P. (2022). Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12), 1256-1264.
 25. Sun, L., Zhang, M., Xie, L., Gao, Q., Xu, X., & Xu, L. (2023). In silico prediction of boiling point, octanol–water partition coefficient, and retention time index of polycyclic aromatic hydrocarbons through machine learning. *Chemical Biology & Drug Design*, 101(1), 52-68.
 26. Villazón-León, V., Bonilla-Petriciolet, A., & Tapia-Picazo, J. C. (2022). On the Analysis and Assessment of First-Order Group Contribution Models for the Calculation of Normal Boiling Point and Critical Properties of Pure Compounds. *International Journal of Chemical Engineering*, 2022(1), 5943833.
 27. Villazón-León, V., Suárez, R. R., Bonilla-Petriciolet, A., & Tapia-Picazo, J. C. (2025). A group contribution-based machine learning model to estimate the triple-point temperature. *Fluid Phase Equilibria*, 595, 114395.
 28. Wang, G., & Hu, P. (2023). Prediction of normal boiling point and critical temperature of refrigerants by graph neural network and transfer learning. *International Journal of Refrigeration*, 151, 97-104.
 29. Wang, Y., Wang, J., Cao, Z., & Barati Farimani, A. (2022). Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3), 279-287.
 30. Wieder, O., Kohlbacher, S., Kuenemann, M., Garon, A., Ducrot, P., Seidel, T., & Langer, T. (2020). A compact review of molecular property prediction with graph neural networks. *Drug Discovery Today: Technologies*, 37, 1-12.
 31. Wu, Z., Wang, J., Du, H., Jiang, D., Kang, Y., Li, D., ... & Hou, T. (2023). Chemistry-intuitive explanation of graph neural networks for molecular property prediction with substructure masking. *Nature communications*, 14(1), 2585.
 32. Yang, C. I., & Li, Y. P. (2023). Explainable uncertainty quantifications for deep learning-based molecular property prediction. *Journal of Cheminformatics*, 15(1), 13.
 33. Yu, P., Liu, H., Xue, J., Li, P., Zhang, Y., & Che, D. (2024). Assessment of the estimation methods for the normal boiling point of liquid organic hydrogen carriers. *International Journal of Hydrogen Energy*, 62, 904-912.
 34. Yu, X., Zhou, C., Yuan, S., Zhu, J., Shang, J., Teng, J., ... & Xiang, S. (2026). Machine Learning-assisted QSPR modeling for accurate prediction of normal boiling points in organic fluorides. *Chinese Journal of Chemical Engineering*.
 35. Burkhardt, J., Bauer, G., Stierle, R., & Gross, J. (2026). Supplementary material for: A new group-contribution approach for ideal gas heat capacity, critical temperature and normal boiling point (Version 1.0) [Data set]. DaRUS. <https://doi.org/10.18419/DARUS-5298>