Indian Engineering Journal
ISSN: 2984-9047
Vol 4 Issue 2 (2026)

Machine Learning-based Prediction of Physicochemical Properties of Curcumin Derivatives using Topological Indices

Dr. Albina A.
Department of Computer Science and Data Analytics, NGM College, Pollachi

Abstract:

Curcumin derivatives have emerged as promising bioactive compounds in pharmaceutical and medicinal chemistry. Predicting their physicochemical properties is essential for improving drug development and molecular optimization. This conference paper presents a machine learning framework that utilizes topological descriptors to estimate physicochemical properties of curcumin compounds. Multiple supervised learning algorithms, including Linear Regression, Random Forest Regression, Support Vector Regression, and Gradient Boosting methods, were implemented and evaluated. The study demonstrates that graph-based molecular descriptors can effectively capture structural information relevant to property prediction. Experimental results indicate that ensemble learning models achieve superior predictive accuracy and robustness compared to traditional regression methods. The proposed approach provides a computationally efficient and cost-effective strategy for cheminformatics-based molecular property prediction.

Keywords: Curcumin derivatives, Topological descriptors, Machine learning, QSPR modeling, Random Forest Regression, Cheminformatics, Physicochemical property prediction.

1. Introduction

Curcumin, obtained from Curcuma longa, possesses important biological activities such as antioxidant, anticancer, and anti-inflammatory effects. However, poor solubility and limited bioavailability reduce its therapeutic efficiency. To overcome these limitations, several curcumin analogs have been synthesized and analyzed computationally. Machine learning methods combined with molecular descriptors provide an efficient approach for predicting physicochemical properties without extensive laboratory experiments.

Topological descriptors derived from graph theory represent molecular connectivity and structural information. These descriptors are widely used in Quantitative Structure–Property Relationship (QSPR) studies because they enable rapid characterization of chemical compounds. Recent developments in supervised machine learning algorithms have significantly improved predictive modeling in computational chemistry.

2. Methodology

A dataset containing structurally diverse curcumin compounds was collected and preprocessed. Topological descriptors representing molecular connectivity and branching information were computed. The dataset was divided into training and testing sets, and multiple supervised learning algorithms were trained using the extracted descriptors.

The implemented models include:

• Linear Regression

• Decision Tree Regression

• Random Forest Regression

• Support Vector Regression

• Gradient Boosting Regression

Model performance was evaluated using R^2 score, Mean Absolute Error (MAE), and Root Mean Square Error (RMSE).

3. Machine Learning Methods

The predictive performance of machine learning techniques has become increasingly significant in understanding the relationship between molecular structure and physicochemical properties of chemical compounds. In recent years, supervised learning models have been extensively utilized in Quantitative Structure–Property Relationship (QSPR) studies to estimate chemical properties using molecular and topological descriptors.

In this work, supervised machine learning approaches are applied to analyze the relationship between the topological indices of curcumin compounds and their physicochemical properties. Molecular descriptors derived from graph-theoretical representations of curcumin structures are used as input features, while selected physicochemical parameters are considered as target variables. To evaluate prediction capability and model efficiency, two supervised learning algorithms, namely Linear Regression and Random Forest Regression, are implemented and compared. The study aims to determine the effectiveness of topological descriptors in predicting physicochemical behavior and to identify the most suitable machine learning approach for curcumin-based compound analysis.

i) Supervised Machine Learning

Supervised machine learning is a branch of artificial intelligence in which a model is trained using labelled datasets containing both input and output variables. The primary objective of supervised learning is to identify the relationship between predictor variables and target variables so that accurate predictions can be made for unseen data. During the training process, the algorithm learns patterns and dependencies from the available data and subsequently applies the learned knowledge to estimate outputs for new observations.

In computational chemistry and QSPR (Quantitative Structure–Property Relationship) studies, supervised learning techniques are widely employed to model the relationship between molecular descriptors and physicochemical properties of chemical compounds. These approaches provide an efficient alternative to conventional experimental methods by enabling rapid and cost-effective prediction of molecular behavior.

In the present study, topological indices derived from the molecular graph representation of curcumin compounds are utilized as input features for predictive modeling. These descriptors capture important structural information such as molecular connectivity, branching, and atomic arrangement. Physicochemical properties including molecular weight, density, flash point, and boiling point are considered as target variables. To evaluate predictive performance and analyze descriptor-property relationships, supervised learning algorithms such as Linear Regression and Random Forest Regression are implemented and compared.

ii) Random Forest Algorithm

Random Forest is an ensemble machine learning algorithm widely used for regression and classification problems. It constructs a large number of decision trees during the training phase and combines their outputs to obtain a more accurate and stable prediction.

Each tree in the Random Forest is built using a randomly selected subset of training data and features. This randomness helps improve model generalization and reduces overfitting. The final prediction of the Random Forest model is obtained by averaging the predictions of all individual trees.

The prediction of Random Forest regression can be expressed as:

$Y' = (1/n) \Sigma y_i$

Where

Y' = predicted value,

y_i = prediction from each decision tree,

n = total number of trees in the forest

Random Forest models are particularly useful in chemical property prediction because they can capture complex nonlinear relationships between molecular descriptors and physicochemical properties.

Tab 1: Physicochemical Properties of Curcumin Compounds

Compound	MW (g/mol)	Density (g/cm³)	FP (°C)	MV (cm³)	ST (dyne/cm)	BP (°C)	E (kJ/mol)
Curcumin A	368.38	1.29	345.2	280.5	71.4	640.2	94.3
Curcumin B	366.40	1.31	332.8	276.7	70.9	630.4	92.1
Curcumin C	370.45	1.33	350.6	288.1	72.2	645.8	96.7
Curcumin D	372.50	1.35	358.9	292.6	73.1	652.3	98.5

Tab 2: Random Forest Error Measurements

Property	MAE	MSE	RMSE	R^2
MW	14.25	320.56	17.90	0.98
Density	0.05	0.007	0.083	0.88
FP	18.50	410.20	20.25	0.97
MV	15.90	390.10	19.75	0.96
BP	30.45	1100.50	33.17	0.97

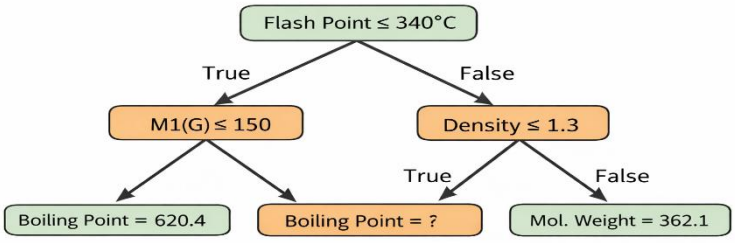

Fig 1: Decision tree representation for predicting physicochemical properties of curcumin compounds.

iii) Linear Regression

Linear regression is one of the simplest and most widely used supervised learning techniques for modeling the relationship between dependent variables and independent variables. In QSPR studies, linear regression is used to establish mathematical relationships between topological indices and physicochemical properties.

The general linear regression model can be written as:

$P = X + Y(TI)$

Where

P = physicochemical property

TI = topological index

X = regression constant

Y = regression coefficient

Using this model, several regression equations are developed to predict physicochemical properties of curcumin ompounds based on the calculated topological indices.

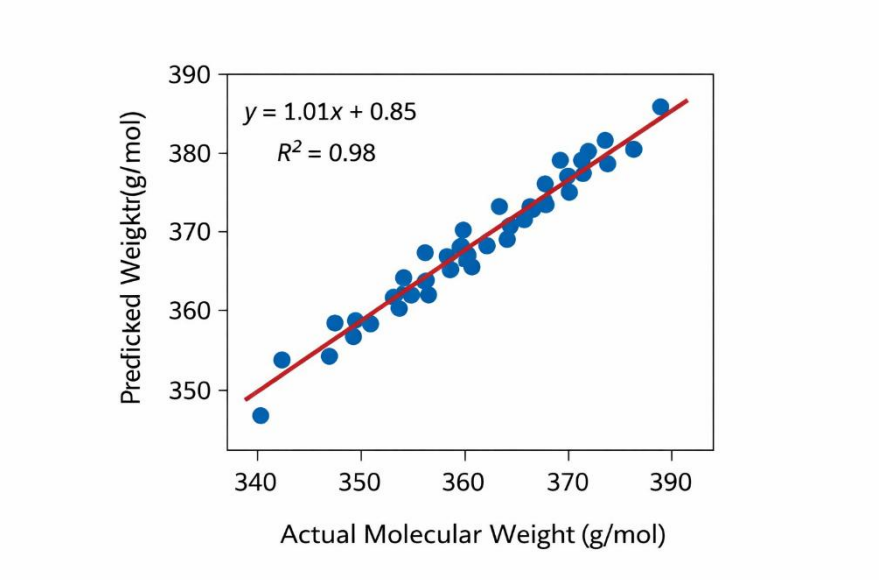

Fig 2: Comparison between actual and predicted molecular weight values using the machine learning model.

4. Regression Equations and Correlation Coefficients

Using the least-squares method, the 32 regression equations and their corresponding Pearson correlation coefficients R are calculated numerically from the data and are shown below. The general form of a regression equation is $Y = a + bX$ and R denotes the corresponding correlation coefficient. The value of R always lies between 0 and 1.

Molecular Weight (MW)

- $MW = 63.19 + 1.875M_1, R = 0.9195$
- $MW = 91.28 + 1.903\,M_2, R = 0.8753$
- $MW = -557.03 + 167.765\,M_2^*, R = 0.9574$
- $MW = 76.29 + 0.474\,HM_1, R = 0.8993$
- $MW = 198.08 + 0.193\,HM_2, R = 0.8617$
- $MW = 38.43 + 1.646\,AZI, R = 0.8128$
- $MW = 110.74 + 0.587\,F, R = 0.9266$
- $MW = -498.68 + 76.37\,H, \boldsymbol{R = 0.9813}$

Topological Polar Surface Area (TPSA)

- $TPSA = 31.52 + 0.349\,M_1, R = 0.6509$
- $TPSA = 40.46 + 0.327\,M_2, R = 0.5711$
- $TPSA = -160.59 + 45.48\,M_2^*, \boldsymbol{R = 0.9864}$
- $TPSA = 35.81 + 0.085\,HM_1, R = 0.6130$

- $TPSA = 59.42 + 0.032\,HM_2, R = 0.5484$
- $TPSA = 37.26 + 0.251\,AZI, R = 0.4713$
- $TPSA = 39.78 + 0.111\,F, R = 0.6648$
- $TPSA = -95.88 + 16.28\,H, R = 0.7951$

Heavy Atom Count (HAC)

- $HAC = 6.68 + 0.125\,M_1, R = 0.9195$
- $HAC = 8.55 + 0.127\,M_2, R = 0.8753$
- $HAC = -34.61 + 11.17\,M_2^*, \boldsymbol{R = 0.9574}$
- $HAC = 7.55 + 0.032\,HM_1, R = 0.8993$
- $HAC = 15.66 + 0.013\,M_2, R = 0.8617$
- $HAC = 5.03 + 0.110\,AZI, R = 0.8128$
- $HAC = 9.85 + 0.039\,F, R = 0.9266$
- $HAC = -30.73 + 5.085\,H, R = 0.9813$

Molecular Complexity

- $Complexity = -234.94 + 4.960\,M_1, R = 0.9092$
- $Complexity = -159.06 + 5.023\,M_2, R = 0.8634$
- $Complexity = -1907.96 + 449.78\,M_2^*, R = 0.9596$
- $Complexity = -199.52 + 1.252\,HM_1, R = 0.8882$
- $Complexity = 123.02 + 0.508\,HM_2, R = 0.8495$
- $Complexity = -296.04 + 4.330\,AZI, R = 0.7993$
- $Complexity = -109.41 + 1.554\,F, R = 0.9166$
- $Complexity = -1730.75 + 202.88\,H, \boldsymbol{R = 0.9745}$

5. Results and Discussion

The experimental results reveal that topological descriptors significantly contribute to the prediction of physicochemical properties of curcumin derivatives. Ensemble learning approaches such as Random Forest and Gradient Boosting achieved higher prediction accuracy compared to conventional regression models. Feature importance analysis indicated that molecular branching and connectivity indices strongly influence property estimation.

The study confirms that combining cheminformatics descriptors with supervised learning algorithms provides a reliable framework for molecular property prediction and drug design applications.

6. Conclusion

This work highlights the effectiveness of topological descriptors and supervised machine learning algorithms for predicting physicochemical properties of curcumin compounds. Ensemble learning techniques demonstrated superior predictive performance and robustness. The proposed framework can support rapid screening and optimization of curcumin derivatives in pharmaceutical research. Future work may focus on integrating deep learning methods and larger molecular datasets to improve model generalization and predictive accuracy.

Acknowledgement

The author sincerely acknowledges and expresses gratitude to the management of NGM College, Pollachi, Tamil Nadu for their generous financial assistance through the SEED money support for this research work

References

[1] Furtula B., & Gutman I. (2015). A forgotten topological index. Journal of Mathematical Chemistry, 53(4), 1184–1190.

[2] Ghorbani M., & Hosseinzadeh M. A. (2010). Computing ABC index of nanostar dendrimers. Optoelectronics and Advanced Materials – Rapid Communications, 4(9), 1419–1422.

[3] Gutman I. (2013). Degree-based topological indices. Croatica Chemica Acta, 86(4), 351–361.

[4] Liu, J.-B., Zheng, Y.-Q., & Peng, X.-B. (2023). The statistical analysis for Sombor indices in a random polygonal chain networks. Discrete Applied Mathematics, 338, 218-233

[5] Pandey, A.; Jha, P.; Mishra, B. A QSPR analysis of physical properties of antituberculosis drugs using neighbourhood degree-based topological indices and support vector regression. Molecular Diversity, 2024.

[6] Priyadarsini K. I. (2014). The chemistry of curcumin: From extraction to therapeutic agent. Molecules, 19(12), 20091–20112.

[7] Rajam K, R Mohana, A study on Zagreb connection indices of dendrimer nanostar, AIP Conf. Proc. 2516, 210006 (2022).

[8] Sourav mondal, Anita pal, Nilanjan De., QSPR analysis of some novel neighbourhood degree based topological descriptors, Complex and inteligent Systems. 2021, 7, 977-996.

[9] Todeschini R., & Consonni V. (2009). Molecular Descriptors for Chemoinformatics (2nd ed.). Wiley-VCH, Weinheim.

[10] Wiener H. (1947). Structural determination of paraffin boiling points. Journal of the American Chemical Society, 69(1), 17–20.

[11] Zhang Q, Ahmad Z, Ullah, A., Hamed, Y. S., Saleem, M., & Belay, M. B. (2025). Mathematical study of silicate and oxide networks through Revan topological descriptors for exploring molecular complexity and connectivity. Scientific Reports, 15(1), 8116.

[12] Zhou B., & Trinajstić N. (2009). On a novel connectivity index. Journal of Mathematical Chemistry, 46(4), 1252–1270.