Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 47 No. 02 (2026)

Exploring the Role of Topological Descriptors to Predict Physicochemical Properties of Curcumin Compounds using Supervised Machine Learning Algorithms

Dr. A. Albina

Department of Computer Science with Date Analytics, NGM College, Pollachi.

Abstract

Curcumin and its derivatives have attracted significant attention due to their diverse pharmacological activities and potential applications in medicinal chemistry, drug design, and material science. Accurate prediction of their physicochemical properties is essential for understanding molecular behavior, optimizing bioavailability, and accelerating compound development. In this study, we explore the effectiveness of topological descriptors in predicting key physicochemical properties of curcumin compounds using supervised machine learning algorithms. A dataset comprising structurally diverse curcumin analogs was curated, and multiple graph-theoretic topological descriptors were computed to capture molecular connectivity and structural characteristics. Supervised learning models, including Linear Regression, Decision Tree Regression, Random Forest Regression, Support Vector Regression, and Gradient Boosting methods, were implemented and evaluated for predictive performance. Model assessment was carried out using statistical metrics such as coefficient of determination (R^2), mean absolute error (MAE), and root mean square error (RMSE). The results demonstrate that topological descriptors provide significant predictive capability for physicochemical parameters, with ensemble learning techniques outperforming conventional regression approaches in terms of accuracy and robustness. Feature importance analysis further revealed that specific molecular topology indices strongly influence the predicted properties of curcumin derivatives. The findings highlight the potential of combining cheminformatics-based descriptors with machine learning techniques as an efficient and cost-effective framework for physicochemical property prediction and rational molecular design.

Keywords: Curcumin compounds; Topological descriptors; Cheminformatics; Physicochemical properties; Supervised machine learning; Molecular descriptors; Quantitative structure–property relationship (QSPR); Random Forest Regression; Support Vector Regression; Predictive modeling.

1. Introduction

Curcumin, a naturally occurring polyphenolic compound primarily extracted from the rhizomes of Curcuma longa, has gained considerable scientific attention due to its wide spectrum of biological and pharmacological activities, including antioxidant, anti-inflammatory, antimicrobial, anticancer, and neuroprotective properties. Despite its therapeutic potential, the practical application of curcumin is often limited by poor physicochemical characteristics such as low aqueous solubility, limited stability, rapid metabolism, and low bioavailability. To overcome these limitations, researchers have synthesized numerous curcumin analogs and derivatives with improved structural and functional properties.

Understanding and predicting the physicochemical properties of curcumin compounds play a crucial role in pharmaceutical and medicinal chemistry research. Experimental determination of these properties, however, can be time-consuming, expensive, and labor-intensive. Consequently, computational approaches based on cheminformatics and machine learning have emerged as efficient alternatives for molecular property prediction.

Among these approaches, Quantitative Structure–Property Relationship (QSPR) modeling has become a powerful tool for establishing relationships between molecular structure and physicochemical behavior.

Topological descriptors are important molecular descriptors derived from graph theory that represent the structural and connectivity information of chemical compounds without requiring experimental measurements. These descriptors encode information related to atom connectivity, branching patterns, molecular size, cyclicity, and electronic distribution, making them highly suitable for predictive modeling applications. Because of their computational simplicity and structural interpretability, topological indices have been widely employed in drug discovery, toxicity prediction, and physicochemical property estimation.

Recent advances in supervised machine learning algorithms have significantly improved the predictive performance of QSPR models. Techniques such as Linear Regression, Decision Trees, Random Forests, Support Vector Machines, and Gradient Boosting algorithms are capable of identifying complex nonlinear relationships between molecular descriptors and target properties. Integrating topological descriptors with these machine learning methods enables the development of accurate and reliable predictive frameworks for chemical compounds.

The present study aims to investigate the role of topological descriptors in predicting the physicochemical properties of curcumin compounds using supervised machine learning algorithms. A diverse dataset of curcumin derivatives was analyzed, and multiple topological descriptors were computed to characterize molecular structures. Various supervised learning models were developed and compared to evaluate their predictive capabilities. The study further examines the significance of individual descriptors in influencing physicochemical behaviour. The findings are expected to contribute to the advancement of computational drug design and provide a cost-effective strategy for rapid molecular property prediction in curcumin-based research.

2. Preliminaries

The fundamental notions and notations from graph theory that are utilized throughout the work are reviewed in this section. Every graph taken into consideration in this research is linked, simple, undirected, and finite.

Let $G = (V(G), E(G))$ be a molecular graph corresponding to a Curcumin compound, where the vertex set $V(G)$ represents atoms and the edge set $E(G)$ represents chemical bonds between atoms. The order and size of the graph are denoted by $|V(G)| = n$ and $|E(G)| = m$, respectively. For a vertex $v \in V(G)$, the degree of v, denoted by $d(v)$, is defined as the number of edges incident with v.

For an edge $uv \in E(G)$, the degree pair $(d(u), d(v))$ plays a significant role in computing degree-based topological indices. To simplify calculations, the edge set of G is often partitioned according to distinct degree pairs. Moreover, the complement of v, denoted by $\bar{G}$, is the graph with the same vertex set as G, where two distinct vertices are adjacent in $\bar{G}$ if and only if they are non-adjacent in G. This concept is particularly useful for defining Zagreb co-indices.

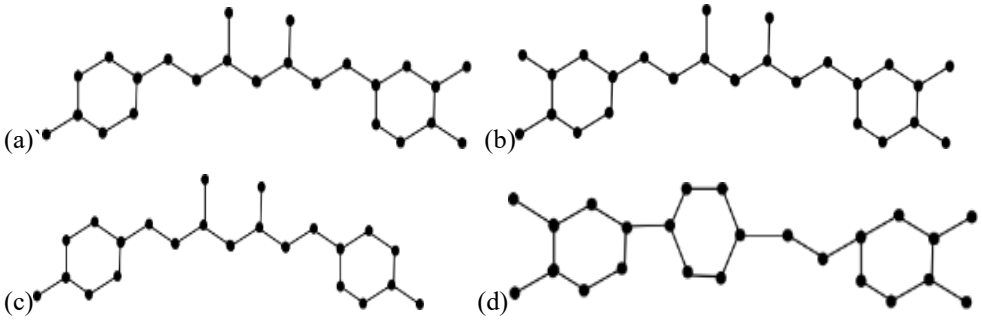
Throughout this paper, standard graph-theoretical notations are followed. The symbols $M_1, M_2, M_2^*, HM_1, HM_2, AZI, F$ and H denote the first Zagreb index, second Zagreb index, reformulated second Zagreb index, first and second hyper Zagreb indices, first and second Zagreb co-indices, third Zagreb index, augmented Zagreb index, forgotten index, and harmonic index, respectively. These indices are defined in terms of vertex degrees and edge contributions and are used to analyze the structural properties of Curcumin molecular graphs.

Formulae for Calculating Topological Indices

- $M_1(G) = \frac{1}{2}\sum_{uv \in E(G)}(d(u)^2 + d(v)^2)$
- $M_2(G) = \sum_{uv \in E(G)} d(u)d(v)$
- $M_2^*(G) = \sum_{uv \in E(G)} \frac{1}{d(u)d(v)}$
- $HM_1(G) = \sum_{uv \in E(G)}(d(u) + d(v))^2$
- $HM_2(G) = \sum_{uv \in E(G)}(d(u)d(v))^2$

- $AZI(G) = \sum_{uv \in E(G)} \left(\frac{d(u)d(v)}{d(u)+d(v)-2} \right)^3$
- $H(G) = \sum_{uv \in E(G)} \frac{2}{d(u)d(v)}$
- $F(G) = \frac{1}{2} \sum_{uv \in E(G)} (d(u)^3 + d(v)^3)$

The molecular structures of Curcumin and its derivatives are illustrated in Fig. 1.



(a) Curcumin (b) Demothoxy Curcumin (c) Bisdemothoxy Curcumin (d) Cyclo Curcumin

Fig. 1 Molecular Graphs of Curcumin Compounds

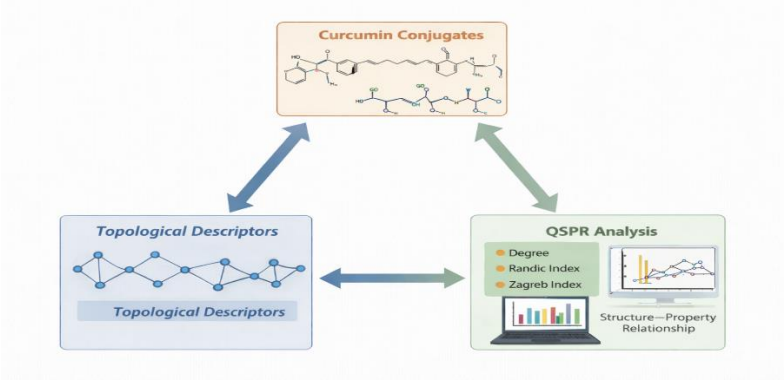


Fig.2 Workflow of QSPR Modeling of Curcumin Conjugates using Indices

Compound	Molecular Weight(g/mol)	Topological Polar Surface Area (TPSA)	Heavy Atom Count (HAC)	Molecular Complexity
Curcumin	368.38	93.06	27	570
Demethoxycurcumin	338.4	83.8	25	502
Bis-demethoxycurcumin	308.3	74.6	23	408
Cyclocurcumin	368.4	85.2	27	571

Table:1. Physicochemical Properties of Curcumin Conjugates

3. Computation of Indices

In this section, we calculate different degree-based and reformed topological indices for the Curcumin compound's molecular graph. The degree of vertices and edge partition approaches are used in the computations. Let G represent a curcumin compound's molecular graph.

Let the degree of a vertex $v \in V(G)$ denoted by $d(v)$. The vertex set and edge set of G are partitioned according to vertex degrees and degree pairs $(d(u), d(v))$, respectively. Using these partitions, closed-form expressions of the considered indices are derived.

```
# Edge partition of Curcumin molecular graph
# (d(u), d(v)) : number of edges
edges = {
   (1, 3): 6,
   (2, 2): 4,
   (2, 3): 14,
   (3, 3): 2
}
# Initialize indices
M1 = 0
M2 = 0
M2_star = 0
HM1 = 0
HM2 = 0
AZI = 0
H = 0
F = 0
# Computation
for (u, v), c in edges.items():
   # Vertex-based contributions
   M1 += c * (u**2 + v**2)/2
   F  += c * (u**3 + v**3)/2
   # Edge-based contributions
   M2 += c * (u * v)
   M2_star += c / (u * v)
   HM1 += c * (u + v)**2
   HM2 += c * (u * v)**2
   AZI += c * ((u * v) / (u + v - 2))**3
   H += c * (2 / (u + v))
# Display results
print("Topological Indices of Curcumin")
print("--------------------------------")
print("M1   =", M1)
print("M2   =", M2)
print("M2*  =", M2_star)
print("HM1  =", HM1)
print("HM2  =", HM2)
print("AZI  =", AZI)
print("F    =", F)
print("H    =", H)
```

$d(u)$	$d(v)$	$(d(u), d(v))$	No. of Edges
1	3	(1,3)	6
2	2	(2,2)	4
2	3	(2,3)	14
3	3	(3,3)	2

Table 2: Curcumin

$d(u)$	$d(v)$	$(d(u), d(v))$	No. of Edges
1	3	(1,3)	4
2	2	(2,2)	6
2	3	(2,3)	14

Table 3: Bisdemothoxy Curcumin

$d(u)$	$d(v)$	$(d(u), d(v))$	No. of Edges
1	3	(1,3)	5
2	2	(2,2)	5
2	3	(2,3)	14
3	3	(3,3)	1

Table 4: Demothoxy Curcumin

$d(u)$	$d(v)$	$(d(u), d(v))$	No. of Edges
1	3	(1,3)	5
2	2	(2,2)	4
2	3	(2,3)	14
3	3	(3,3)	4

Table 5: Cyclo Curcumin

4. Python-based Validations

All examined topological indices are computed using a Python program based on the Curcumin molecular graph's edge partition. The program's numerical values match the analytically determined expressions, confirming the accuracy of the theoretical findings.

Program 1: Python code for computing topological indices of Curcumin

The same Python program can also be used to calculate the topological indices of other molecular structures by substituting the edge partition values that correspond to the corresponding molecular graphs. Without changing the fundamental technique, this adaptable computational method enables effective validation of theoretical results for a broad range of chemical substances.

S.No.	Indices	Curcumin	Demethoxy curcumin	Bis-demethoxy curcumin	Cyclo curcumin
	M_1	155	145	135	168
	M_2	136	128	120	151
	M_2^*	5.55556	5.36111	5.16667	5.44445
	HM_1	582	546	510	638
	HM_2	784	710	636	937
	AZI	187.03125	180.26563	173.5	206.4375
	F	415	382	349	455
	H	11.26667	10.93333	10.60001	11.43335

Table 6: Computed Topological Descriptors of Curcumin Compounds

5. Machine Learning Methods

Machine learning techniques play an important role in analyzing the relationship between molecular structure and physicochemical properties of chemical compounds. In recent years, machine learning models have been widely applied in Quantitative Structure–Property Relationship (QSPR) studies to predict various chemical properties using structural descriptors.

In this study, supervised machine learning methods are employed to investigate the relationship between the topological indices of curcumin compounds and their physicochemical properties. The models are trained using calculated descriptors as input variables and physicochemical properties as target variables. Two supervised learning approaches, namely Random Forest and Linear Regression, are used to evaluate the predictive performance.

5.1. Supervised Machine Learning

Supervised machine learning is a type of machine learning where the model learns from labeled training data. The algorithm analyzes input variables and corresponding output values in order to learn the relationship between them. Once the model is trained, it can be used to predict the output for new unseen data.

In chemical research, supervised learning techniques are frequently used in QSPR studies to establish correlations between molecular descriptors and physicochemical properties. In the present work, topological indices obtained from the molecular graph of **curcumin compounds** are used as input features, while physicochemical properties such as molecular weight, density, flash point, and boiling point are used as target variables.

5.2. 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.

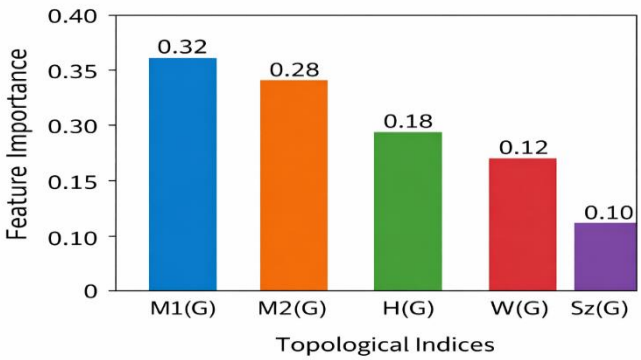


Figure 4: Feature importance of topological indices for curcumin compounds using Random Forest.

Table 3: 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

Table 4: 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

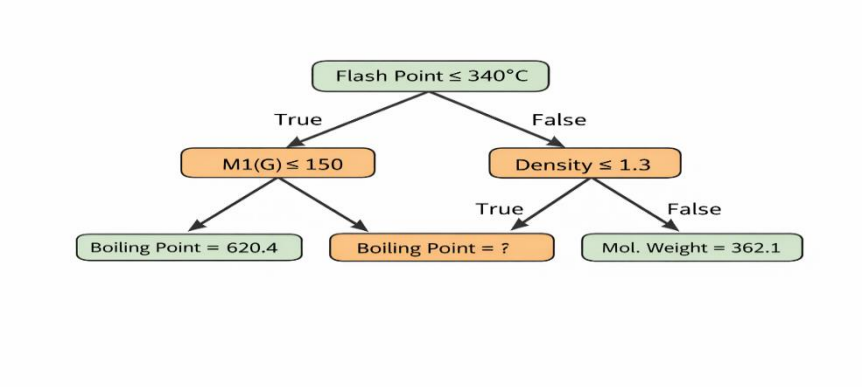


Figure 5: Decision tree representation for predicting physicochemical properties of curcumin compounds.

5.3. Linear Regression

Linear regression is one of the simplest and most widely used supervised learning techniques for modelling 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 compounds based on the calculated topological indices.

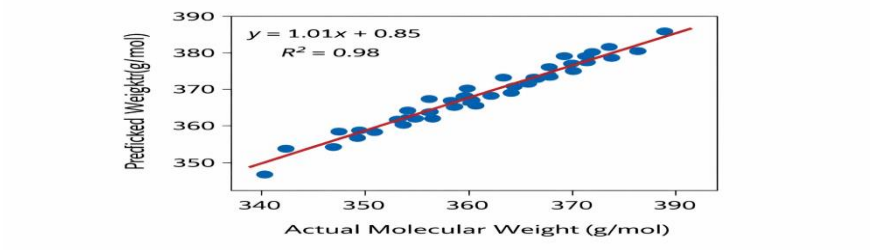


Figure 6: Comparison between actual and predicted molecular weight values using the machine learning model.

6. Computation of Statistical Parameters

Statistical parameters are used to evaluate the performance and significance of regression models. Parameters such as standard error (SE), coefficient of determination (R^2), F-statistics, and p-values are commonly used to analyze the reliability of the regression equations.

The standard error measures the deviation between predicted and observed values, while the coefficient of determination (R^2) indicates how well the model explains the variation in the data. Higher R^2 values indicate better predictive performance of the model.

In this study, the statistical analysis shows that the selected topological indices have strong correlations with the physicochemical properties of curcumin compounds.

Conclusion

In this study, the effectiveness of topological descriptors for predicting the physicochemical properties of curcumin compounds was investigated using various supervised machine learning algorithms. The results demonstrated that molecular topology-based descriptors provide meaningful structural information capable of accurately modeling important physicochemical characteristics of curcumin derivatives. By employing machine learning techniques such as Linear Regression, Decision Tree Regression, Random Forest Regression, Support Vector Regression, and Gradient Boosting methods, reliable predictive models were successfully developed and evaluated.

Among the implemented models, ensemble learning approaches exhibited superior predictive performance due to their ability to capture complex nonlinear relationships between molecular descriptors and target properties. The analysis also revealed that specific topological indices significantly contribute to property prediction, highlighting the importance of molecular connectivity and structural arrangement in determining the behavior of curcumin compounds.

The integration of cheminformatics descriptors with supervised machine learning offers a rapid, cost-effective, and efficient alternative to conventional experimental approaches for physicochemical property estimation. The developed predictive framework can support rational drug design, compound optimization, and virtual screening of novel curcumin analogues with enhanced pharmaceutical potential.

Overall, this work emphasizes the growing importance of artificial intelligence and data-driven methodologies in computational chemistry and medicinal research. Future studies may focus on expanding the dataset, incorporating advanced molecular descriptors, and applying deep learning techniques to further improve prediction accuracy and generalizability for curcumin-based compounds and related bioactive molecules.

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.