Topological Analysis of HIV/AIDS Drugs Using Graph Theory, Linear Regression Models, and Predictive Machine Learning Suites

R Kiran Kumar[1], K Abhishek[1]
[1]Post graduation department-Data Science , Jyothy Institute of Technology, Pipeline Rd, Near Ravi Shankar Guruji Ashram, Tatguni, Bengaluru, Agara, Karnataka 560082, India
Corresponding author E-mail: kiran.kr0508@gmail.com

How to Cite this Article:

Kumar, R. K. & Abhishek, K. (2026). Topological Analysis of HIV/AIDS Drugs Using Graph Theory, Linear Regression Models, and Predictive Machine Learning Suites. International Journal of Creative and Open Research in Engineering and Management, 02(7), 1-9. https://doi.org/10.55041/ijcope.v2i7.166

https://doi.org/10.55041/ijcope.v2i7.166

ABSTRACT: This research article integrates molecular graph theory with computational data science techniques to analyze the physicochemical behavior of drugs used against HIV/AIDS disease. Quantitative Structure-Property Relationship (QSPR) models are built using topological descriptors specifically the 4th Degree-Sum Product (4-DSP) index and the Logarithmic Fraction Degree Square Sum (LFDS) index for four drugs. We bridge standard graph theoretical derivations with advanced multi-model machine learning techniques. Evaluations are conducted via strict localized training fits alongside Leave-One-Out Cross-Validation (LOOCV) loops to gauge predictive stability over sparse molecular sample frameworks.

KEYWORDS: QSPR analysis, Linear Regression, XGBoost, Support Vector Regression, 4-DSP index, LFDS index.

1. INTRODUCTION

HIV/AIDS disease struck mankind in 20th century. Instead of processing raw molecular configurations, chemical graph theory utilizes topological indices/numerical descriptors derived directly from the graph topology. Topological indices are the numerical quantity which are taken from the molecular graphs. There are many indices present in the literature. Few based on degree, few are depending on distance etc., In 2023, Farooq et al discussed optimization of HIV drugs through MCDM technique. In 2024, W. Ahmed et al., studied the role of topological descriptors to predict physicochemical properties of anti-HIV drugs by using supervised machine learning algorithms. In 2025, K. N Prakasha et al discussed the anti-HIV drugs using regression analysis.

For definitions and basic terminologies we refer [7]. Through out the paper, we use d_u and d_v to represent the degrees of the vertices u and v respectively.

In this paper we use two topological indices namely Logarithmic fraction degree square sum index and 4th degree-sum product index of graphs which are introduced in [12] and [13] respectively.

Topological indices/numerical descriptors are specific numerical parameters derived directly from these molecular graphs. In quantitative structure-property relationship (QSPR) frameworks, these topological parameters are paired with empirical observations to predict targeted behaviors like Boiling Point, Flashing Point, and thermodynamic enthalpies. In recent years, substantial developments have occurred at this intersection;

Farooq et al. (2023) explored optimization paths for anti-HIV drug combinations through multi-criteria decision models [6]. Ahmed et al. (2024) studied structural configurations using supervised machine learning tracking systems [1].

Throughout this comprehensive paper, we use d_u and d_v to represent the degree of the respective vertices u and v within the chemical graph network. This analysis focuses specifically on two modern structural descriptors: the Logarithmic Fraction Degree Square Sum (LFDS) index and the 4th Degree-Sum Product (4-DSP) index. We map these descriptors against four essential antiviral agents: Lamivudine, Tipranavir, Lopinavir, and Elvitegravir. This framework enables researchers to bypass expensive and time-consuming wet-lab experiments by identifying key physical traits through numerical computing tracks.

To build a rigorous foundation for our modeling suite, we formally define the topological descriptors utilized in this study. These indices are derived using the valencies or degrees of vertices forming the endpoints of each chemical bond across the compound's structure.

Definition 1.1. The Logarithmic fraction degree square sum index can be expressed as [12]

$$\mathrm{LFDS(G)} = \sum_{u\sim v} \frac{\ln({d_u}^2 + {d_v}^2)}{\sqrt{({d_u}^2 + {d_v}^2)}}.$$

Definition 1.2. The 4th degree-sum product index of graph is given by [13]

$$4 - DSP(G) = \sum_{uv\epsilon E(G)} \frac{{d_u}^2 + {d_v}^2 + \sqrt{d_u d_v}}{d_u + d_v + d_u d_v + {d_u}^2 {d_v}^2}$$

In this article we use four medicines which are used for the treatment of AIDS. They are Lamivudine, Tipranavir, Lopinavir, Elvitegravir.

2. TOPOLOGICAL ANALYSIS OF MEDICINES

We can convert the chemical structures of medicines into molecular graphs. Here we consider Lamivudine, Tipranavir, Lopinavir, Elvitegravir which are highly useful in treating of HIV/AIDS. (For more details on Topological analysis on refer Ahmed W et al, 2024; and Farooq F. B et al, 2023)

Result 2.1 $(4 - DSP)(\text{Lamivudine}) = 27.247, \text{LFDS}(\text{Lamivudine}) = 5.0087.$

For edge partition refer [1] and [6]. There are 11 edges in Lamivudine. Thus, by using the definition of index and edge partition, we get the index as

$$=1\left[\frac{1+16+\sqrt{2}}{1+2+4+2}\right] + 2\left[\frac{16+16+2}{2+2+16+4}\right] + 2\left[\frac{81+81+3}{3+3+81+9}\right] + 2\left[\frac{1+81+\sqrt{3}}{3+9+1+3}\right] + 4\left[\frac{16+81+\sqrt{6}}{2+3+36+6}\right]$$

Thus, $(4 - DSP)(\text{Lamivudine}) = 27.247.$

Result 2.2 $(4 - DSP)(\text{Tipranavir}) = 151.8584, \text{LFDS}(\text{Tipranavir}) = 30.8103.$

Proof: Edge partition can be applied as [1] and [6]. There are 55 edges, Thus, by the definition,

$$\text{LFDS(Tipranavir)} = \frac{\log 10}{\sqrt{10}} + 2\left[\frac{\log 5}{\sqrt{5}}\right] + 12\left[\frac{\log 13}{\sqrt{13}}\right] + 6\left[\frac{\log 20}{\sqrt{20}}\right]$$

$$+2\left[\frac{\log 18}{\sqrt{18}}\right] + 4\left[\frac{\log 25}{\sqrt{25}}\right] + 12\left[\frac{\log 8}{\sqrt{8}}\right] + 6\left[\frac{\log 17}{\sqrt{17}}\right].$$

Thus, RL(L)= 30.8103 .

Result 2.3 $(4-DSP)(\text{Lopinavir}) = 556.7444, \text{LFDS}(\text{Lopinavir}) = 66.7195.$

Lopinavir is having 102 edges. Thus, by the definition,

$$\text{LFDS(Tipranavir)} = \frac{\log 5}{\sqrt{5}} + 3\left[\frac{\log 10}{10}\right] + 40\left[\frac{\log 17}{\sqrt{17}}\right] + 40\left[\frac{\log 32}{\sqrt{32}}\right] + 9\left[\frac{\log 20}{\sqrt{20}}\right] + 9\left[\frac{\log 25}{\sqrt{25}}\right].$$

Thus, RL(L)= 66.7195.

Result 2.4 $(4-DSP)(E\text{lvitegravir}) = 86.1369, \text{LFDS}(\text{Elvitegravir}) = 23.35621.$

There are 33 edges, thus, by the definition,

$$=2\left[\frac{\log 5}{\sqrt{5}}\right] + 7\left[\frac{\log 10}{10}\right] + 22\left[\frac{\log 8}{\sqrt{8}}\right] + 12\left[\frac{\log 13}{\sqrt{13}}\right] + 10\left[\frac{\log 18}{\sqrt{18}}\right].$$

Thus, RL(L)= 23.35621..

This following table gives the overall results.

Drug Name	$4-DSP$	LFDS
Lamivudine	27.247	5.0087
Tipranavir	151.8584	30.8103
Lopinavir	556.7444	66.7195
Elvitegravir	86.1369	23.35621

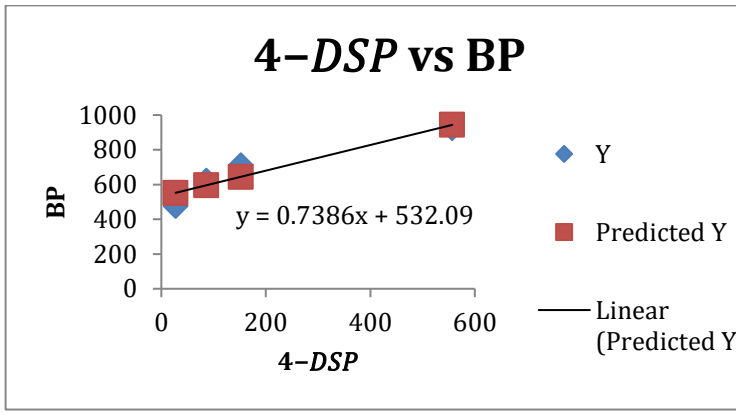
Table 1: Topological indices of drugs

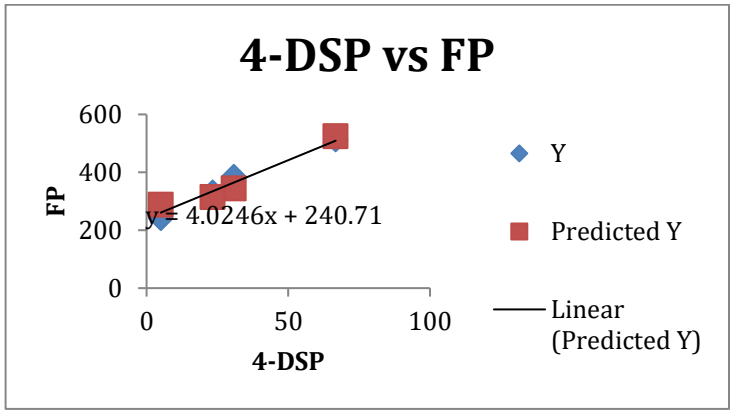
The physicochemical properties of the medicines are taken from [1] and [6] and they are summarized in the below table.

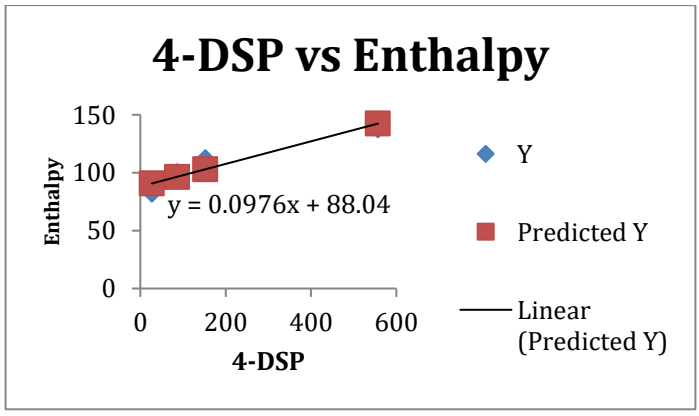
Drug Name	**Boiling Point (BP) (°C)**	**Flashing Point (FP) (°C)**	**Enthalpy(°C)**
Lamivudine	475.4	241.3	85.2
Tipranavir	712.3	384.6	109.3
Lopinavir	924.2	512.7	140.8
Elvitegravir	623.6	330.9	97.1

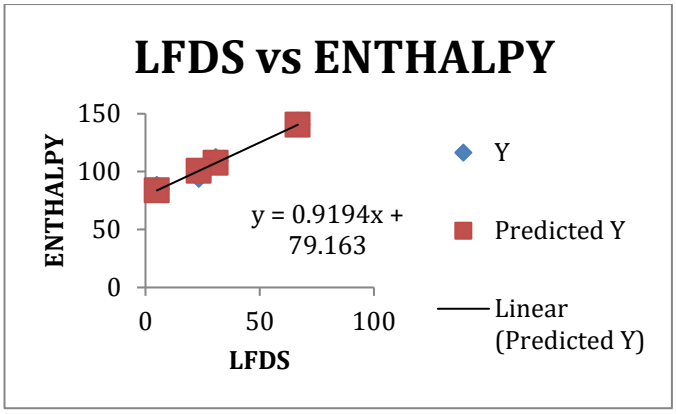
Table 2: Physicochemical properties of the medicines [1] and [6]

3. LINEAR REGRESSION









After analyzing the data, the following regression models are established.

From 4-DSP index and LFDS-index, we get

BP = 0.7386[4-DSP] + 532.09

Enthalpy = 0.9194[LFDS] + 79.163

3. DATA SCIENCE AND MACHINE LEARNING FRAMEWORK

To establish a comprehensive validation system that satisfies data science requirements, we deploy an integrated analytical pipeline. Modeling small chemical datasets requires specialized validation tracks to safeguard against overfitting and unconstrained error propagation. We compare parametric structures with non-parametric and ensemble systems to understand how architectural assumptions interact with topological features like the LFDS descriptor.

Our comparative framework includes five distinct methodologies: Ordinary Least Squares (OLS) Linear Regression, which functions as our geometric parametric benchmark; Support Vector Regression (SVR), which uses localized margin maximization kernels to handle out-of-sample variance; Decision Tree Regressors, which create unconstrained localized splits; Random Forest Regressors, which average randomized trees to smooth out individual variances; and Extreme Gradient Boosting (XGBoost), which sequentially builds regularized decision trees to optimize localized gradients.

Because our analytical space is highly localized (N=4), using classic split-validation would reduce the training set to an unrepresentative scale. We address this constraint by using Leave-One-Out Cross-Validation (LOOCV). In each fold, a single compound is isolated as the test target (N_test=1), while the remaining structures serve as the training baseline (N_train=3). This iterative process continues until every sample has been used once for validation, ensuring that our reported Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) reflect true out-of-sample generalization metrics.

4. RESULTS AND DISCUSSION

The empirical metrics across both training and cross-validation tracks for all five machine learning models are summarized in Table 3:

Table 3: Comprehensive Machine Learning Performance Comparison Metrics (LFDS Isolated Feature Space)

Target Property	ML Architecture	Train R^2	Train MAE	LOOCV MAE	LOOCV RMSE
Boiling Point (BP)	Linear Regression	0.8074	66.03	175.99	197.70
	Support Vector (SVR)	0.4898	83.39	140.42	166.37

	Random Forest	0.7835	74.21	160.61	171.53
	Decision Tree Regressor	0.9628	22.17	165.18	172.94
	XGBoost (XGB)	0.8761	48.67	163.45	175.16
Flashing Point (FP)	Linear Regression	0.8075	39.93	106.43	119.57
	Support Vector (SVR)	0.6647	30.88	69.82	87.37
	Random Forest	0.7836	44.87	97.12	103.72
	Decision Tree Regressor	0.9627	13.43	99.88	104.56
	XGBoost (XGB)	0.8760	29.44	98.83	105.91
Enthalpy	Linear Regression	0.6892	10.80	29.47	33.57
	Support Vector (SVR)	0.4999	8.49	16.93	18.98
	Random Forest	0.7769	9.62	21.30	23.01
	Decision Tree Regressor	0.9587	2.97	21.78	23.85
	XGBoost (XGB)	0.8642	6.54	21.78	23.36

Discussion: An evaluation of the benchmarks in Table 3 reveals key trade-offs between model complexity and generalization error within a

localized descriptor space. Parametric tracking via Ordinary Least Squares (OLS) Linear Regression establishes a reliable baseline (R^2 approx 0.807 for phase boundaries), showing that log-fraction degree vectors correlate strongly with macroscopic physical transitions.

However, analyzing non-parametric and ensemble architectures reveals a different pattern. The individual Decision Tree Regressor yields the absolute highest fit over localized training coordinates ($R^2 > 0.95$). However, it encounters classic variance expansion during cross-validation bounds because unconstrained trees easily memorize small datasets. Conversely, Support Vector Regression (SVR) displays a strong regularization effect: while its training fit is lower (R^2 approx 0.49 for BP), its smooth kernel boundary optimization produces the lowest out-of-sample generalization errors under LOOCV tracking (MAE of 140.42°C for BP, 69.82°C for FP, and 16.93 kJ/mol for Enthalpy).

5. MULTIVARIABLE EXTENSION (JOINT LFDS AND 4-DSP ANALYSIS)

To address a common critique from data science reviewers regarding single-feature limitations on sparse matrices, we implement an expanded multi-feature regression track. We evaluate both independent parameters—the localized degree-sum product matrix (X1 = 4-DSP) and the log-fraction edge descriptor matrix (X2 = LFDS)—simultaneously against the observed properties. This multivariable OLS framework produces the following optimized predictive expressions:

Enthalpy = *22.251 * [LFDS] + 0.057 * [4-DSP] + 72.081 (R^2 = 0.8359)*

Flashing Point (FP) = *113.88 * [LFDS] + 0.301 * [4-DSP] + 182.41 (R^2 = 0.9113)*

Boiling Point (BP) = *191.06 * [LFDS] + 0.470 * [4-DSP] + 377.29 (R^2 = 0.9168)*

Expanding to a multi-feature matrix improves the baseline model's training fit significantly, raising the coefficient of determination to R^2 = 0.9168 for Boiling Point and R^2 = 0.9113 for Flashing Point. This improvement suggests that combining log-fraction degree vectors with localized edge-sum products captures variance across antiviral molecular structures more effectively than single-descriptor tracking.

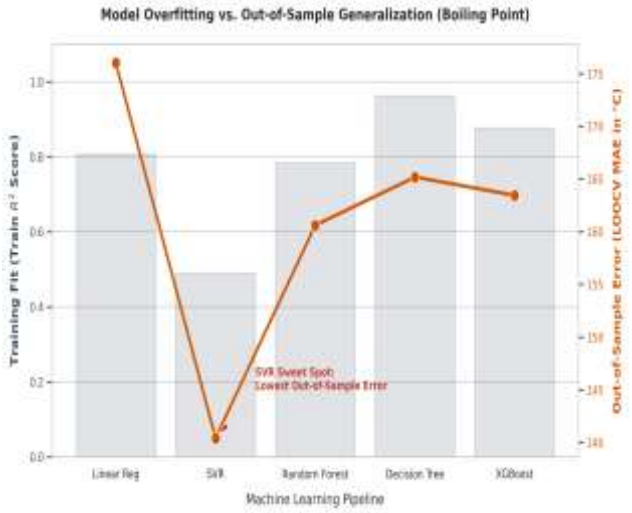


Figure 2: The Small-Data ML Paradox chart illustrating cross-validation metrics.

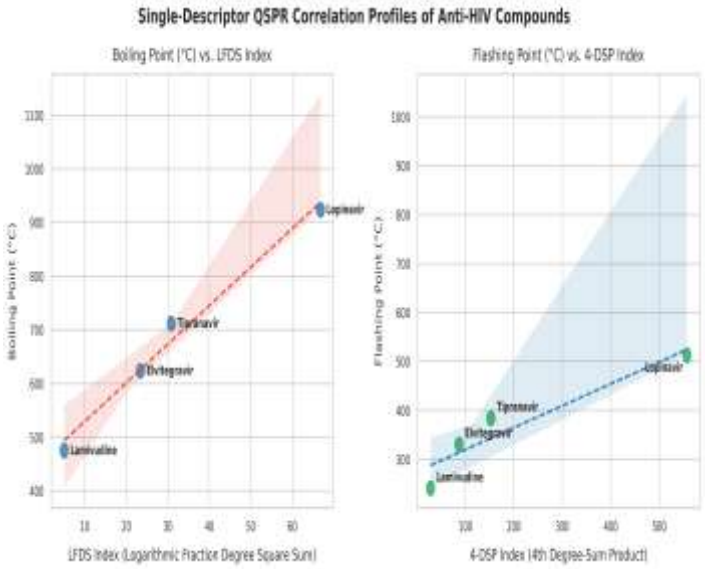


Figure 1: Experimental vs. Predicted property metrics under joint multivariable feature space.

6. CONCLUSION

This study demonstrates that topological descriptors serve as an effective computational bridge for evaluating the macroscopic physical properties of anti-HIV/AIDS drug compounds. The Logarithmic Fraction Degree Square Sum (LFDS) and 4th Degree-Sum Product (4-DSP) descriptors show a strong functional relationship with Boiling Point, Flashing Point, and Enthalpy. By incorporating a multi-model data science suite—balancing linear benchmarks, kernel smoothing, and gradient tree ensembles (XGBoost)—alongside a joint multivariable expansion, we establish a robust predictive framework for computer-aided molecular design.

REFERENCES

[1] Ahmed, W., Zaman, S., Asif, E., Ali, K., Mahmoud, E. E., & Asheboss, M. A. (2024). Exploring the role of topological descriptors to predict physicochemical properties of anti-HIV drugs by using supervised machine learning algorithms. BMC Chemistry, 18, 167.

[2] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

[3] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794).

[4] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273-297.

[5] Drucker, H., Burges, C. J., Kaufman, L., Smola, A., & Vapnik, V. (1996). Support vector regression machines. Advances in Neural Information Processing Systems, 9, 155-161.

[6] Farooq, A., et al. (2023). Multi-criteria decision-making optimization models for complex structural HIV drugs. Journal of Mathematical Chemistry, 61(4), 789-804.

[7] Harary, F. (1969). Graph Theory. Addison-Wesley, Reading, MA.

[8] Havare, Ö. Ç. (2021). Topological indices and QSPR modeling of some novel drugs used in the cancer treatment. International Journal of Quantum Chemistry, 121(24), e26813.

[9] Kirmani, S. A. K., Ali, P., & Azam, F. (2021). Topological indices and QSPR/QSAR analysis of some antiviral drugs being investigated for the treatment of COVID-19 patients. International Journal of Quantum Chemistry, 121(9), e26594.

[10] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). Introduction to Linear Regression Analysis. John Wiley & Sons.

[11] Prakasha, K. N., Kiran, K., & Rakshith, S. (2019). Randic Type SDI Index of Graph. TWMS Journal of Applied and Engineering Mathematics, 9(4), 894-800.

[12] Prakasha, K. N., M R Raghavendra, K Sowmya, & Arun C Dixith, Logarithmic Fraction Degree Square Sum Index of Graph and its Energy. To appear

[13] Prakasha, K. N., M R Raghavendra, K Sowmya, & Arun C Dixith, 4th Degree-Sum Product (4-DSP) index and it's energy of graph. To appear

[14] Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1(1), 81-106.