BMC Chemistry

RESEARCH

Open Access

Exploring the role of topological descriptors to predict physicochemical properties of anti-HIV drugs by using supervised machine learning algorithms

Wakeel Ahmed[1,2*], Shahid Zaman[1,5], Eizzah Asif[1], Kashif Ali[2], Emad E. Mahmoud[3] and Mamo Abebe Asheboss[4]

Abstract

In order to explore the role of topological indices for predicting physio-chemical properties of anti-HIV drugs, this research uses python program-based algorithms to compute topological indices as well as machine learning algorithms. Degree-based topological indices are calculated using Python algorithm, providing important information about the structural behavior of drugs that are essential to their anti-HIV effectiveness. Furthermore, machine learning algorithms analyze the physio-chemical properties that correspond to anti-HIV activities, making use of their ability to identify complex trends in large, convoluted datasets. In addition to improving our comprehension of the links between molecular structure and effectiveness, the collaboration between machine learning and QSPR research further highlights the potential of computational approaches in drug discovery. This work reveals the mechanisms underlying anti-HIV effectiveness, which paves the way for the development of more potent anti-HIV drugs. This work reveals the mechanisms underlying anti-HIV efficiency, which paves the way for the development of more potent anti-HIV drugs which demonstrates the invaluable advantages of machine learning in assessing drug properties by clarifying the biological processes underlying anti-HIV behavior, which paves the way for the design and development of more effective anti-HIV drugs.

Keywords Anti-HIV-1 drugs, Topological indices, Python algorithm, Machine learning algorithm, QSPR analysis

Introduction

Human Immunodeficiency Virus (HIV) was firstly identified in the early 1980s as a consequence of the appearance of an immune system-damaging disease [1]. Later on, the illness was identified as Acquired Immunodeficiency Syndrome (AIDS). In 1983–1984, French scientists Francoise Barre-Sinoussi and Luc Montagnier became essential in discovering the virus. HIV caused a global pandemic that has killed countless people and infected millions of people globally. Its impact on global health is immense, as it not only threatens human health but also affects economies and healthcare systems around the globe [2]. There are two primary types of HIV: HIV-1, which is common surrounding the world, and HIV-2, which is primarily linked to West Africa. Here we focus on HIV-1, HIV-1 target CD4 cells by engaging to their surface receptors, which starts the process of the virus entering and taking control of the cell's functions as shown in Fig. 1. New

*Correspondence:
Wakeel Ahmed
wakeelahmed784@gmail.com
[1] Department of Mathematics, University of Sialkot, Sialkot 51310, Pakistan
[2] Department of Mathematics, COMSATS University, Islamabad Lahore Campus, Lahore 51000, Pakistan
[3] Department of Mathematics and Statistics, Collage of Science, Taif University, P.O. Box 11099, 21944 Taif, Saudi Arabia
[4] Department of Mathematics, Wollega University, 395, Nekemte, Ethiopia
[5] Department of Mathematical and Physical Sciences, University of Nizwa, Nizwa, Oman

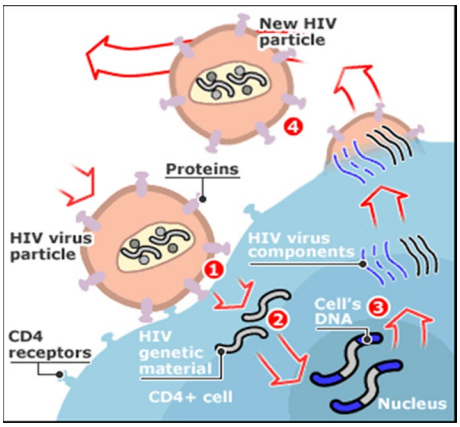


Fig. 1 Virus entering and functioning

viruses are created as a result, ultimately the CD4 cells are destroyed HIV causes the immune system to become extremely weakened by destroying CD4 cells, which sets off a series of immune issues. Gradually, this causes CD4 cell depletion. The immune system's capacity to mount effective defenses against infections is weakened by a decrease in CD4 cells [1, 3–5]. Breast milk, vaginal fluids, rectal fluids, semen and blood represent some of the bodily fluids that may transmit the virus. These bodily fluids can spread HIV when persons engage in risky sexual behavior, share needles with injecting drug users, or are pregnant, giving birth, or nursing a kid [6]. The goal of antiviral therapy is to stop HIV-1 replication in order to protect CD4 cell levels and immune system health [7]. An extensive variety of drugs, including Rilpivirine, Nevirapine, Emtricitabine, Delavirdine, Elvitegravir, Ritonavir, Saquinavir, Indinavir, and Bictegravir (these drugs are referred to as a, b, c,...., i respectively, as shown in Fig. 2 and their molecular graphs represented in Fig. 3) are required to cure HIV-1. These drugs are used to treat HIV-1 infection and stop the HIV virus from growing and from spreading throughout the body by a number of distinct mechanisms. By doing this, these drugs contribute to the regulation of HIV levels in the blood, which protects CD4 cells. In the area of HIV-1 analysis, graph theory provides a fundamental statistical application particularly in the field of chemistry and drugs development. Some embedding's of drugs and diseases through the dual-channel network are characterized in [8–11]. On the other hand, the bridges between largest herbal medicines, chemical ingredients, target proteins, and associated diseases with respect to the neural network and deep learning-based invariants are discussed in [12–17].

Graph theory is essential to the analysis of biochemical networks in medicine, including drug-target relationships and protein–protein interactions [18–22]. To aid in the identification of possible drug candidates and the

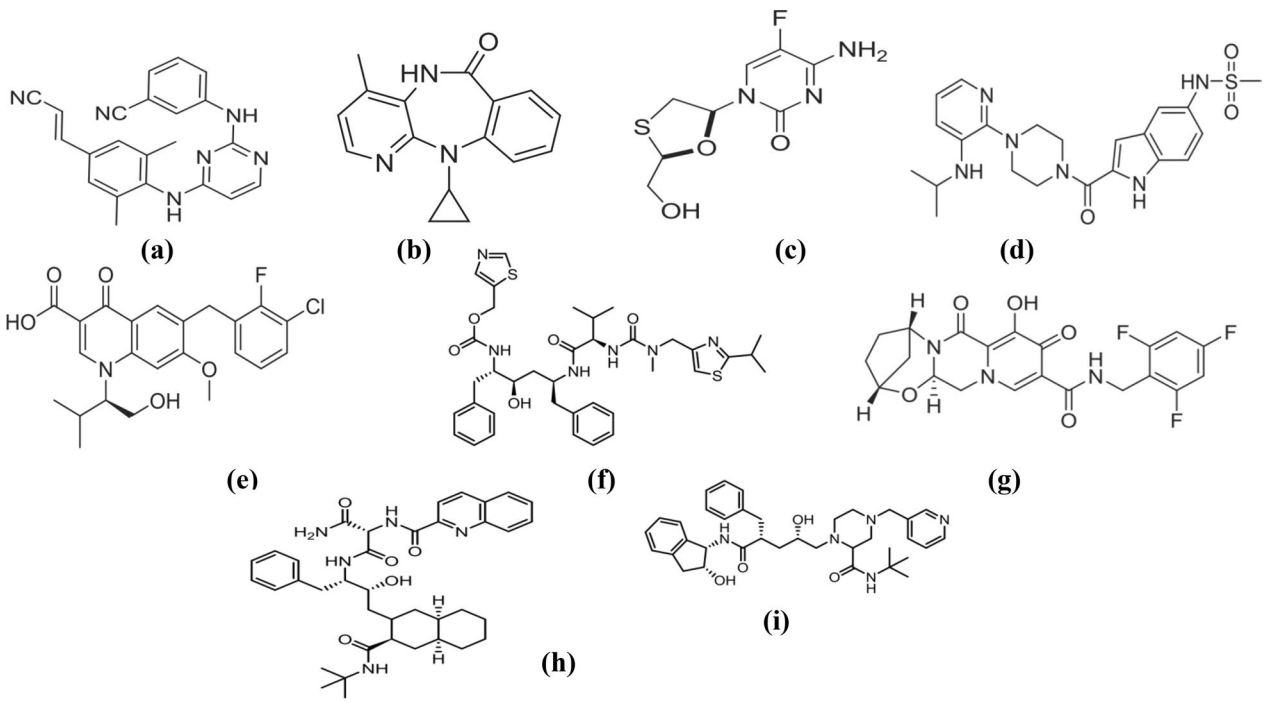


Fig. 2 Molecular structure of antiviral HIV-1 drugs

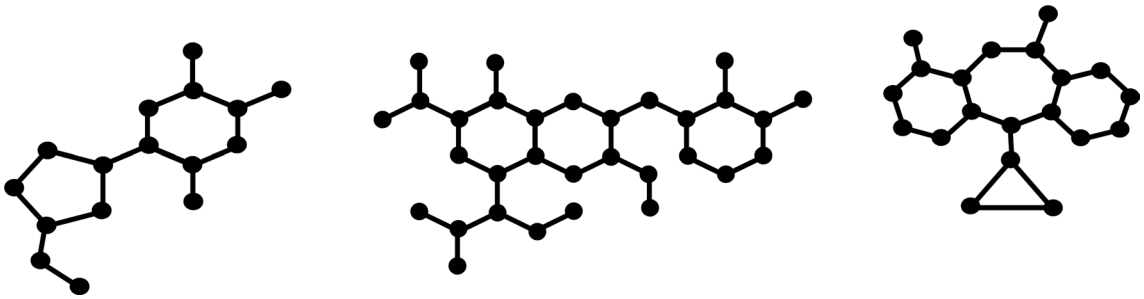


Fig. 3 Molecular graph of emtricitabine, nevirapine and elviteravir

Table 1 Topological indices with notations and formula

	Notation	Formula
First Zagreb Index [34]	M_1 (G)	$\sum_{rj\in E(G)} (dr + dj)$
Second Zagreb Index [34]	M_2 (G)	$\sum_{rj\in E(G)} (dr \times dj)$
Harmonic Index [35]	H (G)	$\sum_{rj\in E(G)} \frac{2}{(dr+dj)}$
Forgotton Index [36]	F (G)	$\sum_{rj\in E(G)} [(dr)^2 + (dj)^2]$
Shilpa-Shanmukha Index [37]	SS (G)	$\sum_{rj\in E(G)} \sqrt{\frac{dr\times dj}{dr+dj}}$
Atom Bond Connectivity Index [38]	ABC (G)	$\sum_{rj\in E(G)} \sqrt{\frac{dr+dj-2}{dr\times dj}}$
Randic Index [39]	RI (G)	$\sum_{rj\in E(G)} \sqrt{\frac{1}{dr\times dj}}$
Sum Connectivity Index [40]	SC (G)	$\sum_{rj\in E(G)} \sqrt{\frac{1}{dr+dj}}$
Geometric Arithmetic Index[38]	GA (G)	$\sum_{rj\in E(G)} 2\frac{\sqrt{dr\times dj}}{dr+dj}$
Hyper Zagreb Index [41]	HZ (G)	$\sum_{rj\in E(G)} (dr + dj)^2$
Redefined First Zagreb Index [42]	ReZ_1 (G)	$\sum_{rj\in E(G)} \frac{(dr\times dj)}{(dr+dj)}$
Redefined Second Zagreb Index [42]	ReZ_2 (G)	$\sum_{rj\in E(G)} (dr \times dj)(dr + dj)$

optimization of drug design, graphs depict pharmaceuticals as nodes and their interactions with targets as edges. Furthermore, proteins are shown as nodes in graphs that represent protein–protein interactions as edges. This makes it possible to identify important protein hubs and pathways that are connected to disease causes and potential treatment approaches. Topological indices (TIs) from graph theory are essential for drugs discovery [23–25].

Our main goal is to conduct an extensive review of nine selected antiviral drugs for HIV-1. Using Python algorithm, which involves finding their degree base TIs such as (Randic, Sum Connectivity, First Zagreb, Second Zagreb) Indices which shown in Table 1 by developing python algorithm based on graph theory. Python programs are essential resources for researchers examining the chemical properties of drugs and computing topological indices. In addition to improving analytical efficiency by automating repetitive processes and quickly processing enormous data sets, the computational approach offers substantial benefits in the simultaneous research of many drugs. By revealing complex links between molecular descriptors and biological activities, the integration of physio-chemical characteristics such as molecular weight (MW), complexity (Comp), density (Den), flash point (FP), molar volume (MV), surface tension (ST), polarizability (Pol), boiling point (BP) and enthalpy of vaporization (EV) into the study through machine learning algorithms, contributes to our understanding of the potential efficacy and safety profiles of drugs against HIV. In order to provide a thorough understanding of the molecular characteristics of HIV drugs and to provide insights into their modes of action and potential side effects, it is imperative to combine topological indices with physio-chemical parameters. It is essential to combine topological indices with physio-chemical parameters to provide a comprehensive understanding of the molecular properties of HIV drugs, as well as insights into their modes of action and potential adverse effects. In order to predict drug efficacy based on molecular features, researchers utilize supervised machine learning models to establish quantitative correlations between calculated molecular descriptors and observed biological activity. Supervised machine learning predictive models offer valuable

insights into the potential efficacy of anti-HIV drug by analyzing their molecular properties and estimating their effectiveness against the illness. The utilization of Quantitative Structure–Property Relationship (QSPR) analysis is becoming increasingly important in understanding the relationships between drug structures and biological behavior [26–30]. QSPR analysis provides a rational framework for drug design and optimization [31–33]. By combining computational methods and QSPR analysis, researchers hope to obtain a deeper understanding of the molecular mechanisms underlying anti-HIV drugs, which will help in the development of more focused and efficient treatment options.

Material and method

We initially determined the edge partition based on graph connectivity was adopted to define molecular graphs, which is an important step in recognizing the structural properties. Then, degree-based TIs were calculated through analyzing the molecular graph's node degree variation. To make this process easier, a unique Python algorithm was developed. After that, Python programs were used to develop machine learning methods for the analysis of physiochemical properties. Furthermore, using Statistical Package for the Social Sciences (SPSS) software to analyze relationships between the computed indices and experimental features, we also performed graphical comparison analysis between actual and computed drug property, ensuring the accuracy and credibility of our results.

Data acquisition and preparation

- We utilized the latest version of python 3.12 to compute topological indices and sourced physiochemical properties from online database Chemspider (https://www.chemspider.com) and Pubchem (https://pubchem.ncbi.nlm.nih.gov). The topological descriptors were employed as feature variables (input variables), while the physiochemical properties served as target variables. Our analysis covered a dataset composed of multiple feature variables and target variables, representing a considerable amount of data points.
- Given that our dataset is labeled, we opted for supervised machine learning algorithms, specifically Random Forest and XGBoost, to analyze the data and derive insights. RF is chosen for its proficiency in handling overfitting through its ensemble approach, where multiple decision trees contribute to a more stable and accurate prediction while XGBoost is based on the gradient boosting framework, which builds one tree at a time. Each new tree helps to correct errors made by previously trained tree models. By averaging several trees, Random Forest reduces the risk of overfitting, which is common with single decision trees while XGBoost is based on the gradient boosting framework, which builds one tree at a time. Each new tree helps to correct errors made by previously trained tree models.
- The primary libraries utilized for Random Forest and XGBoost are:
  - "pandas" for data manipulation,
  - "numpy" for numerical operations,
  - "scikit-learn" for machine learning algorithms, including Random Forest and XGBoost,
  - "matplotlib" and "seaborn" for data visualization,
  - Computational resources: the computations were performed on a machine with an Intel core i7 processor and 16 GB of RAM.

Results and discussion

Theorem 1 *Let G be a graph and G_1 denotes the elvitegravir, then the following axioms holds for the graph G_1:*

(*a*) $M_1(G_1)=162$; (*b*) $M_2(G_1)=195$; (*c*) $H(G_1)=13.966$; (*d*) $F(G_1)=432$; (*e*) $SS(G_1)=35.088$; (*f*) $ABC(G_1)=23.695$; (*g*) $RI(G_1)=14.688$; (*h*) $SC(G_1)=15.1037$; (*i*) $GA(G_1)=131.705$; (*j*) $HZ(G_1)=822$; (*k*) $ReZG1(G_1)=37.983$; (*l*) $ReZG2(G_1)=1028$.

Proof Suppose that Gramicidin S is represented by *G*1, where $E_{r,s}$ is the set of edges connecting vertices in the graph with corresponding degrees *r* and *s*. Between vertices of degrees *r* and *s*, the frequencies $|E_{r,s}|$ show the number of edges. The expression $|E_{1,2}|=2$ denotes two edges present between the vertices of degree 1 and 2, while the expression $|E_{1,3}|=7$ denotes eighteen edges present between the vertices of degree 1 and 3. Similarly, $|E_{2,2}|=2$, $|E_{2,3}|=12$ $|E_{3,3}|=10$. Then,

a) By using First Zagreb Index

$$M_1\ (G) = \sum_{rs\in E(G)} (dr + ds),$$

$$\begin{aligned} M_1\ (G_1) &= 2(1+2) + 7(1+3) + 2(2+2) \\ &\quad + 12(2+3) + 10(3+3) = 2\times 3 \\ &\quad + 7\times 4 + 2\times 4 + 12\times 5 + 10\times 6 = 162. \end{aligned}$$

b) By using Second Zagreb Index

$$M_2\ (G) = \sum_{rs\in E(G)} (dr \times ds),$$

$$\begin{aligned} M_2\ (G_1) &= 2(1\times 2) + 7(1\times 3) + 2(2\times 2) \\ &\quad + 12(2\times 3) + 10(3\times 3) \\ &= 2\times 2 + 7\times 3 + 2\times 4 + 12\times 6 \\ &\quad + 10\times 9 = 195. \end{aligned}$$

c) By using Forgotten Index

$$H\ (G) = \sum_{rs\in E(G)} \frac{2}{(dr+ds)},$$

$$\begin{aligned} H\ (G) &= 2\frac{2}{1+2} + 7\frac{2}{1+3} + 2\frac{2}{2+2} \\ &\quad + 12\frac{2}{2+3} + 10\frac{2}{3+3} \\ &= 2\frac{2}{3} + 7\frac{2}{4} + 2\frac{2}{4} + 12\frac{2}{5} + 10\frac{2}{6} = 13.966. \end{aligned}$$

d) By using Forgotten Index

$$F\ (G) = \sum_{rs\in E(G)} \left[(dr)^2 + (ds)^2\right],$$

$$\begin{aligned} \mathrm{F\ (G1)} &= [2(12+22) + 7(12+32) \\ &\quad + 2(22+22) + 12(22+32) \\ &\quad + 10(32+32) \\ &= 2\times 5 + 7\times 10 + 2\times 8 \\ &\quad + 12\times 13 + 10\times 18 = 432. \end{aligned}$$

e) By using Shilpa-Shanmukha Index

$$\mathrm{SS\ (G1)} = \sum_{\mathrm{rs\in E(G)}} \sqrt{\frac{\mathrm{dr}\times \mathrm{ds}}{\mathrm{dr}+\mathrm{ds}}},$$

$$\begin{aligned} \mathrm{SS\ (G1)} &= 2\sqrt{\frac{1\times 2}{1+2}} + 7\sqrt{\frac{1\times 3}{1+3}} + 2\sqrt{\frac{2\times 2}{2+2}} \\ &\quad + 12\sqrt{\frac{2\times 3}{2+3}} + 10\sqrt{\frac{3\times 3}{3+3}} \\ &= 2\sqrt{\frac{2}{3}} + 7\sqrt{\frac{3}{4}} + 2\sqrt{\frac{4}{4}} \\ &\quad + 12\sqrt{\frac{6}{5}} + 10\sqrt{\frac{9}{6}} = 35.088. \end{aligned}$$

f) By using Randic Index

$$RI\ (G_1) = \sum_{rs\in E(G)} \sqrt{\frac{1}{dr\times ds}},$$

$$\begin{aligned} RI\ (G_1) &= 2\sqrt{\frac{1}{1\times 2}} + 7\sqrt{\frac{1}{1\times 3}} + 2\sqrt{\frac{1}{2\times 2}} \\ &\quad + 12\sqrt{\frac{1}{2\times 3}} + 10\sqrt{\frac{1}{3\times 3}} \\ &= 2\sqrt{\frac{1}{2}} + 7\sqrt{\frac{1}{3}} + 2\sqrt{\frac{1}{4}} \\ &\quad + 12\sqrt{\frac{1}{6}} + 10\sqrt{\frac{1}{9}} = 14.688. \end{aligned}$$

g) By using Sum Connectivity Index

- $\mathrm{SC\ (G_1)} = \sum_{rs\in E(G)} \sqrt{\frac{1}{dr+ds}},$

$$\begin{aligned} SC\ (G_1) &= 2\sqrt{\frac{1}{1+2}} + 7\sqrt{\frac{1}{1+3}} + 2\sqrt{\frac{1}{2+2}} \\ &\quad + 12\sqrt{\frac{1}{2+3}} + 10\sqrt{\frac{1}{3+3}} \\ &= 2\sqrt{\frac{1}{3}} + 7\sqrt{\frac{1}{4}} + 2\sqrt{\frac{1}{4}} \\ &\quad + 12\sqrt{\frac{1}{5}} + 10\sqrt{\frac{1}{6}} = 15.1037. \end{aligned}$$

h) By using Geometric Arithmetic Index

- $\mathrm{GA\ (G_1)} = \sum_{rs\in E(G)} 2\frac{\sqrt{dr\times ds}}{dr+ds},$

$$GA\ (G_1) = 2 \times 2\frac{\sqrt{1\times 2}}{1+2} + 2 \times 7\frac{\sqrt{1\times 3}}{1+3} + 2 \times 2\frac{\sqrt{2\times 2}}{2+2} + 2 \times 12\frac{\sqrt{2\times 3}}{2+3} + 2 \times 10\frac{\sqrt{3\times 3}}{3+3} = 4\frac{\sqrt{2}}{2} + 14\frac{\sqrt{3}}{4} + 4\frac{\sqrt{4}}{4} + 24\frac{\sqrt{6}}{5} + 20\frac{\sqrt{9}}{6} = 31.7053.$$

i) By using Hyper Zagreb Index

- $\mathrm{HZ}\ (\mathrm{G}_1) = \sum_{rs\in E(G)} (dr + ds)^2,$

$$HZ\ (G_1) = \left[2(1+2)^2 + 7(1+3)^2 + 2(2+2)^2 + 12(2+3)^2 + 10(3+3)^2\right] = 2(3)^2 + 7(4)^2 + 2(4)^2 + 12(5)^2 + 10(6)^2 = 822.$$

j) By using Redefined First Zagreb Index

$$ReZ_1\ (G_1) = \sum_{rs\in E(G)} \frac{(dr \times ds)}{(dr + ds)},$$

$$ReZ_1\ (G_1) = 2\frac{1\times 2}{1+2} + 7\frac{1\times 3}{1+3} + 2\frac{2\times 2}{2+2} + 12\frac{2\times 3}{2+3} + 10\frac{3\times 3}{3+3} = 2\frac{2}{3} + 7\frac{3}{4} + 2\frac{4}{4} + 12\frac{6}{5} + 10\frac{9}{6} = 37.9833.$$

k) By using Redefined Second Zagreb Index

- $\mathrm{ReZ}_2(\mathrm{G}_1) = \sum_{rs\in E(G)} (dr \times ds)(dr + ds)$

$$= 2(1\times 2)\ (1+2) + 7(1\times 3)\ (1+3) + 2(2\times 2)\ (2+2) + 12(2\times 3)\ (2+3) + 10(3\times 3)\ (3+3) = 2\times 2\times 3 + 7\times 3\times 4 + 2\times 4\times 4 + 12\times 6\times 5 + 10\times 9\times 6 = 1028,$$

Remark 3.2 The topological indices of other drugs can be obtained using a similar technique as that used in Theorem 1 and their output is provided in Table 2.

Although a lot of scholars are already calculating topological indices [43–46], we contribute by creating an efficient Python program (see Algorithm 1) to compute these indices. Especially, our technique can quickly compute through integrating edge partition values for every molecular graph in an elegant and seamless manner. This Python method advances the field with its efficiency by providing simplified procedures, improved accuracy and time saving for computing topological indices.

Table 2 The topological indices values for the candidate drugs

Drugs	M_1 (G)	M_2 (G)	H (G)	F (G)	SS (G)	ABC (G)	RI (G)	SC (G)	GA (G)	HZ (G)	ReZ_1 (G)	ReZ_2 (G)
a	132	151	12.13	332	29.48	20.04	12.54	12.99	27.20	334	31.31	740
b	114	141	9.533	298	25.06	16.19	9.754	10.43	22.57	580	27.60	738
c	82	96	7.266	214	17.94	12.22	7.613	7.827	16.35	406	19.21	490
d	172	200	14.50	460	37.18	25.35	15.19	15.89	33.59	860	39.88	1020
e	162	195	13.96	432	35.08	23.69	14.68	15.10	31.70	822	37.98	1028
f	248	280	23.06	622	55.39	38.25	23.97	24.66	51.26	1182	58.45	1366
g	272	318	22.81	746	58.20	40.00	24.05	24.97	52.48	1382	62.36	1670
h	236	272	20.73	614	51.84	35.31	21.55	22.48	47.34	1158	55.33	1362
i	206	256	15.20	606	42.32	28.33	16.26	17.15	36.86	1118	46.71	1436

Algorithm 1 .

```python
#python algorithm to compute topological indices for anti-HIV-1 drugs
import numpy as np
P = np.array([1,1,2,2,2,3])
Q = np.array([3,4,2,3,4,3])
R = np.array([3,3,6,17,1,5])

G1 = np.zeros(6)
G2 = np.zeros(6)
G3 = np.zeros(6)
G4 = np.zeros(6)
G5 = np.zeros(6)
G6 = np.zeros(6)
G7 = np.zeros(6)
G8 = np.zeros(6)
G9 = np.zeros(6)
G10 = np.zeros(6)
G11 = np.zeros(6)
G12 = np.zeros(6)

for m in range(6):
    G1[m] = R[m] * (P[m] + Q[m])
    G2[m] = R[m] * (P[m] * Q[m])
    G3[m] = R[m] * (2 / (P[m] + Q[m]))
    G4[m] = R[m] * (P[m]**2 + Q[m]**2)
    G5[m] = R[m] * (np.sqrt(P[m] * Q[m] / (P[m] + Q[m])))
    G6[m] = R[m] * (np.sqrt((P[m] + Q[m] - 2) / (P[m] * Q[m])))
    G7[m] = R[m] * (np.sqrt(1 / (P[m] * Q[m])))
    G8[m] = R[m] * (np.sqrt(1 / (P[m] + Q[m])))
    G9[m] = R[m] * (2 * np.sqrt(P[m] * Q[m]) / (P[m] + Q[m]))
    G10[m] = R[m] * (P[m] + Q[m])**2
    G11[m] = R[m] * (P[m] * Q[m] / (P[m] + Q[m]))
    G12[m] = R[m] * (P[m] * Q[m] * (P[m] + Q[m]))

G11 = np.sum(G1)
G22 = np.sum(G2)
G33 = np.sum(G3)
G44 = np.sum(G4)
G55 = np.sum(G5)
G66 = np.sum(G6)
G77 = np.sum(G7)
G88 = np.sum(G8)
G99 = np.sum(G9)
C11 = np.sum(G10)
C22 = np.sum(G11)
C33 = np.sum(G12)

print("G11 =", G11)
print("G22 =", G22)
print("G33 =", G33)
print("G44 =", G44)
print("G55 =", G55)
print("G66 =", G66)
print("G77 =", G77)
print("G88 =", G88)
print("G99 =", G99)
print("C11 =", C11)
print("C22 =", C22)
print("C33 =", C33)
```

Theorem 1 and Algorithm 1 can both be used to compute topological indices; however algorithmic approach is more effective and beneficial in this respect. Moreover, Table 3 shows the physio-chemical properties of selected drugs collected from ChemSpider [47] and PubChem [48] and the computed TIs obtained from their molecular structures by developing python algorithm respectively as seen above.

Supervised machine learning

Within the field of artificial intelligence, machine learning focuses on creating statistical models and algorithms that allow computers to learn and make decisions without explicit programming. The development of drugs usually involves machine learning techniques like Random Forest Algorithm (RFA), Extreme Gradient Boosting (XGB), and linear analysis. Linear analysis techniques like linear regression are helpful for simpler, easier-to-understand models, ensemble learning techniques like XGB and RFA are capable of managing complex nonlinear correlations and interactions in data.

Random forest

For machine learning tasks including regression, RFA is a potent ensemble learning technique. During training, it builds a large number of decision trees, and it produces the mean prediction (regression) of each individual tree. In order to begin, RF bootstraps a technique many random sections of the training set. A decision tree is trained using each subset, also referred to as a bootstrap sample. At every split point, a decision tree is built for every bootstrap sample using a random subset of features. The model performs better overall because of this randomness, which aids in decorrelation between the trees. Without any pruning, each tree is grown to its fullest depth. When every tree is constructed, its predictions are combined using the Random Forest algorithm. The following is a mathematical representation of the prediction formula for regression:

$$Y' = \frac{1}{n}\sum_{i=1}^{n} y(i),$$

where Y' is the predicted output, y_1, y_2,.....y_n are the predicted outputs from individual decision trees, and n is the total number of trees in the Random Forest. Figure 4 represent the feature importance of some physiochemical properties w.r.t topological indices; also Figs. 5 and 6 illustrate the decision trees.

Violin plots highlight gaps in the data distribution and help evaluate the accuracy of predictions against actual values graphically as shown in Figs. 7, 8 and 9. RFA

Table 3 The properties of drugs related to their physical characteristics

Drugs	MW (g/mol)	Comp	Den (g/cm^3)	FP (°C)	MV (cm^3)	ST (dyne/cm)	P (cm^3)	BP (°C)	E (kJ/mol)
a	366.4	607	1.3	337.3	287	72.3	42.3	634.1	93.7
b	266.30	397	1.4	205	197	66.3	29.2	415.4	66.8
c	247.25	374	1.8	221.9	135.2	72.8	21.4	443.3	80.9
d	456.6	749	1.4	396.5	328.8	72.3	49.3	732.0	106.8
e	447.9	702	1.4	330.9	329.9	55.0	44.9	623.6	97.1
f	720.9	1040	1.2	526.6	581.7	53.7	78.9	947.0	144.4
g	670.8	1140	1.2	567.7	553.9	54.2	75	1015	155.3
h	613.8	952	1.3	484.7	491.0	63.7	69.8	877.9	133.7
i	449.4	912	1.6	366.6	276.2	70.6	40.6	682.5	105.2

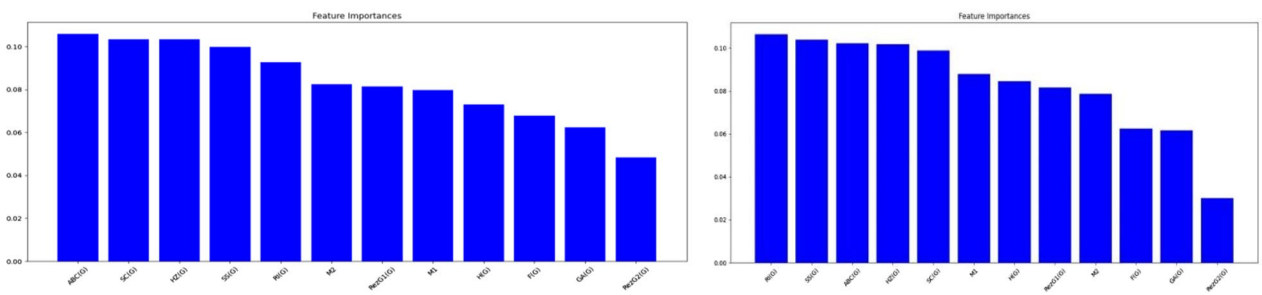


Fig. 4 Graphical representation of feature importance of MW and Den w.r.t TIs

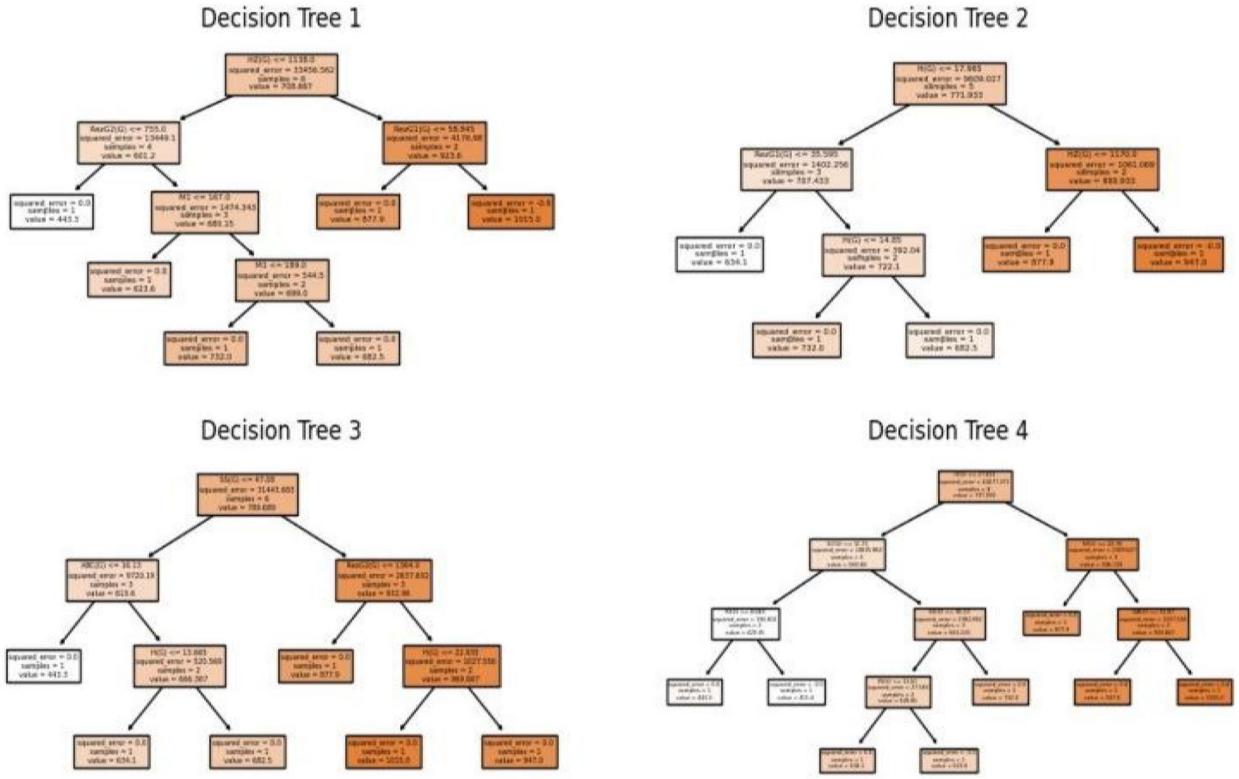
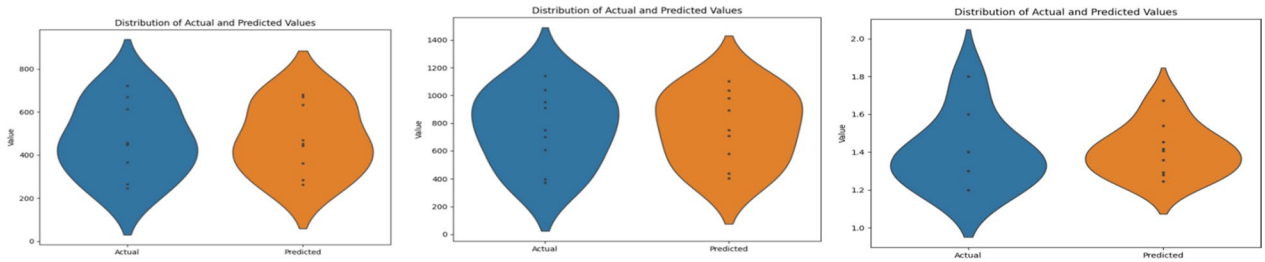


Fig. 5 Decision trees for BP



(a) Violin plot for MW (b) Violin plot for Comp (c) Violin plot for Den

Fig. 6 Random forest algorithm based violin distribution plot

output error measures are shown in Table 4 and include specific parameters like Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). The following formulas can be used to determine MAE, MSE, and RMSE:

- $\mathrm{MAE} = \frac{1}{\mathrm{n}} \sum \left|\mathrm{actual} - \mathrm{predicted}\right|$,
- $\mathrm{MSE} = \frac{1}{\mathrm{n}} \sum (\mathrm{actual} - \mathrm{predicted})^2$,
- $\mathrm{RMSE} = \frac{1}{\mathrm{n}} \sqrt{\left(\sum (\mathrm{actual} - \mathrm{predicted})^2\right)}$.

The random forest algorithm's performance and prediction accuracy were examined through information gained from both the violin plots and tables.

Linear regression

Linear regression is a fundamental supervised machine learning technique that predicts the connection between dependent variable and one or more independent variables. These models quantify the relationship between drug structures and their medical impacts through the use of various components, such

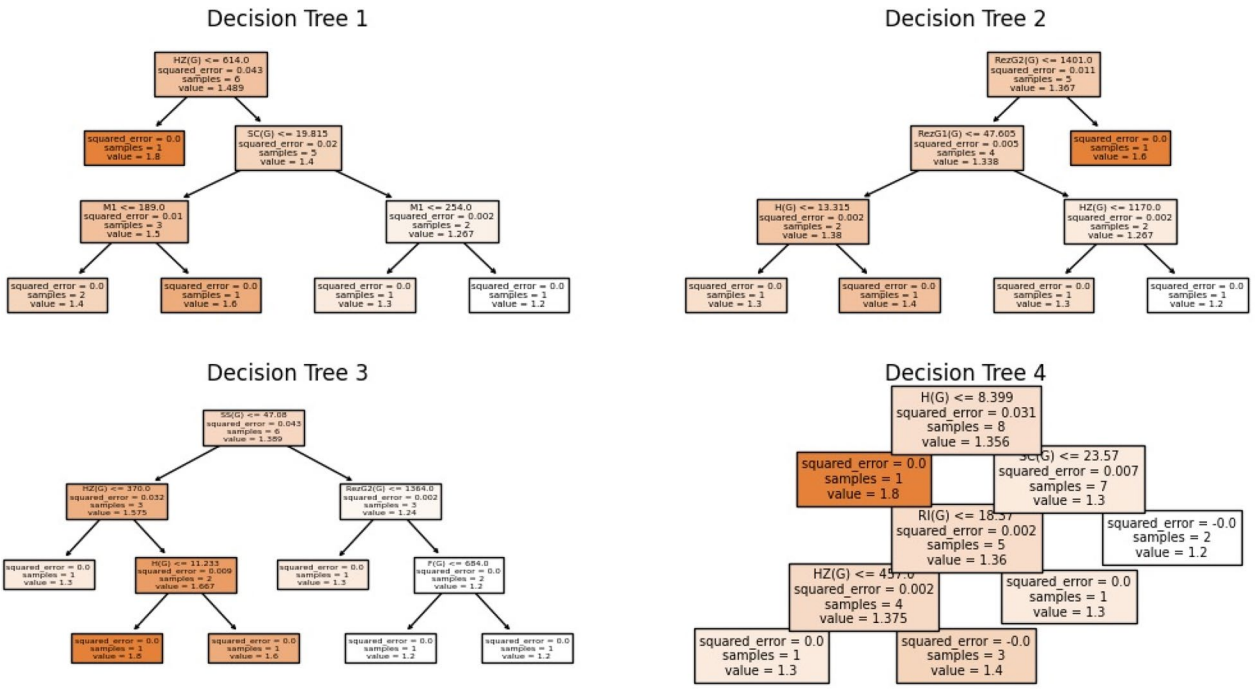
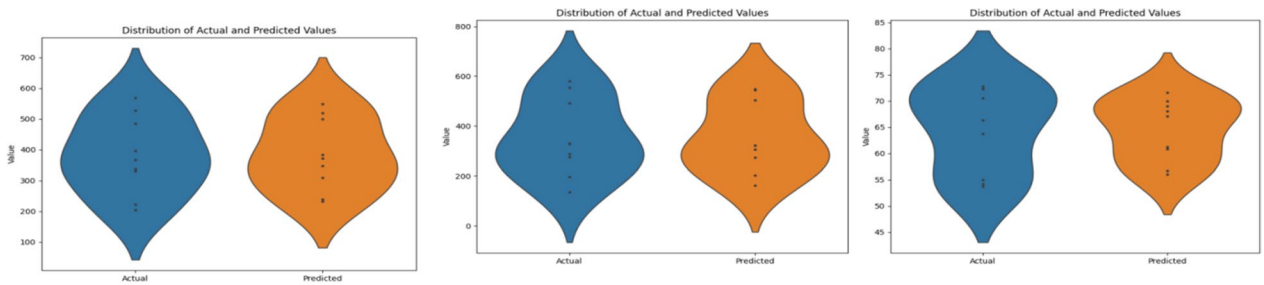
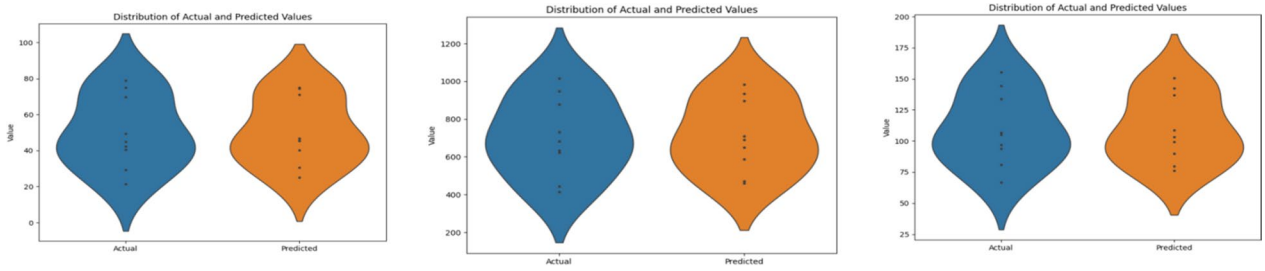


Fig. 7 Decision trees for Den



(d) Violin plot for FP (e) Violin plot for MV (f) Violin plot for ST

Fig. 8 Random forest algorithm based violin distribution plot



(g) Violin plot for Pol (h) Violin plot for BP (i) Violin plot for EV

Fig. 9 Random forest algorithm based violin distribution plot

as TIs. The QSPR results are constantly provided by the regression equation, which makes it an invaluable formula that is P = X + Y (TI). Here, P is the physiochemical parameter of a potential drug. Topological index, constant and regression coefficient are indicated by the symbols TI, X and Y respectively. The correlation coefficients between each topological indicator and the nine physio-chemical parameters are calculated and shown in Table 5 while, bar graph representing the correlation coefficients of all physio-chemical properties across different topological indices is shown in Fig. 10. Linear regression equations and physio-chemical properties w.r.t TIs derived below.

Table 4 Random forest error measurement

PP	MW	Comp	Den	FP	MV	ST	Pol	BP	EV
MAE	14.3971	21.690	0.0506	18.9814	16.6246	3.01311	2.2503	30.8837	3.6944
MSE	341.1717	674.165	0.0079	418.5225	399.098	11.1851	7.2899	1127.198	18.5042
RMSE	18.47083	25.9646	0.06277	20.4578	19.977	3.3444	2.6999	33.573	4.3016
R^2	0.9864	0.9897	0.8817	0.9705	0.9815	0.8152	0.9799	0.9709	0.9761

Table 5 Correlation coefficients of TI w.r.t to different physiochemical properties

TIs	MW	Comp	Den	FP	MV	ST	Pol	BP	E
M_1 (G)	0.9605	0.9874	0.6658	0.9553	0.9298	0.6472	0.9320	0.9553	0.9452
M_2 (G)	0.9326	0.9831	0.6290	0.9266	0.8937	0.6338	0.8947	0.9266	0.9158
H (G)	0.9929	0.9609	0.7430	0.9750	0.9835	0.7000	0.9842	0.9750	0.9667
F (G)	0.9124	0.9831	0.5856	0.9184	0.8662	0.5962	0.8683	0.9184	0.9093
SS (G)	0.9765	0.9801	0.7031	0.9655	0.9547	0.6682	0.9568	0.9655	0.9550
ABC (G)	0.9820	0.9807	0.7031	0.9729	0.9609	0.6698	0.9629	0.9729	0.9640
RI (G)	0.9923	0.9686	0.7292	0.9765	0.9789	0.6971	0.9793	0.9764	0.9690
SC (G)	0.9890	0.9702	0.7283	0.9744	0.9748	0.6877	0.9761	0.97440	0.9655
GA (G)	0.9846	0.9727	0.7237	0.9708	0.9687	0.6800	0.9706	0.9708	0.9607
HZ (G)	0.8766	0.9264	0.4820	0.8499	0.8189	0.6368	0.8166	0.8499	0.8590
ReZ_1 (G)	0.9680	0.982	0.6903	0.9570	0.9426	0.6625	0.9445	0.9570	0.9458
ReZ_2 (G)	0.8769	0.9653	0.5521	0.8762	0.8249	0.5957	0.8243	0.8762	0.8668

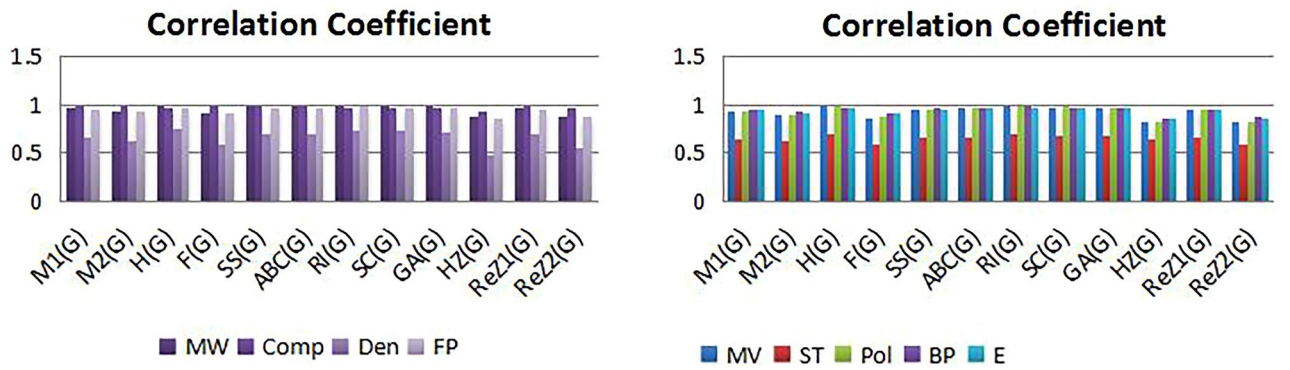


Fig. 10 Correlation coefficients of all physio-chemical properties w.r.t Tis

Linear regression models w.r.t TIs

Regression model for [M2(G)]	Regression model for M_1(G)
Molecular weight = 22.1100 + 2.1165[M_2(G)]	Molecular weight = 20.3377 + 2.4977[M_1(G)]
Complexity = 0.2808 + 3.5990[M_2(G)]	Complexity = 16.339 + 4.1416[M_1(G)]
Density = 1.7482 − 0.0016[M_2(G)]	Density = 1.7593 − 0.0020[M_1(G)]
Flash point = 46.9325 + 1.5793[M_2(G)]	Flash point = 45.2833 + 1.8655[M_1(G)]
Molar volume = − 45.305 + 1.8798[M_1(G)]	Molar volume = − 50.941 + 2.2409[M_1(G)]
Surface tension = 79.4937 − 0.0705[M_2(G)]	Surface tension = 79.4230 − 0.0825[M_1(G)]
Polarizability = − 1.6054 + 0.2440[M_2(G)]	Polarizability = − 2.3993 + 0.2913[M_1(G)]
Boiling point = 154.073 + 2.6109[M_2(G)]	Boiling point = 151.346 + 3.0842[M_1(G)]
Enthalpy of variation = 32.0454 + 0.3643[M_2(G)]	Enthalpy of variation = 31.5778 + 0.4309[M_1(G)]

Regression model for F(G)	Regression model for H(G)
Molecular weight = 56.1270 + 0.8636[F(G)]	Molecular weight = 14.0675 + 29.5447[H(G)]
Complexity = 42.5485 + 1.5009[F(G)]	Complexity = 50.3265 + 46.1198[H(G)]
Density = 1.7062 − 0.0006[F(G)]	Density = 1.79322 − 0.0254[H(G)]
Flash point = 68.2462 + 0.6529[F(G)]	Flash point = 44.9218 + 21.7875[H(G)]
Molar volume = − 11.6411 + 0.7598[F(G)]	Molar volume = − 66.0938 + 27.1224[H(G)]
Surface tension = 77.8286 − 0.0277[F(G)]	Surface tension = 80.3293 − 1.0205[H(G)]
Polarizability = 2.6997 + 0.0987[F(G)]	Polarizability = − 4.2840 + 3.5197[H(G)]
Boiling point = 189.3117 + 1.0793[F(G)]	Boiling point = 150.7451 + 36.0198[H(G)]
Enthalpy of variation = 36.8359 + 0.1509[F(G)]	Enthalpy of variation = 31.3317 + 5.0423[H(G)]

Regression model for ABC	Regression model for SS(G)
Molecular weight = 20.3827 + 16.9418[**ABC(G)]**	Molecular weight = 9.7915 + 11.7757[SS(G)]
Complexity = 37.6722 + 27.2928[**ABC(G)]**	Complexity = 16.8938 + 19.0652[SS(G)]
Density = 1.7710 − 0.0140[**ABC(G)]**	Density = 1.7819 − 0.0098[SS(G)]
Flash point = 46.5920 + 12.6059[**ABC(G)]**	Flash point = 39.4157 + 8.7440[SS(G)]
Molar volume = − 55.2781 + 15.3641[**ABC(G)]**	Molar volume = − 64.516 + 10.6698[SS(G)]
Surface tension = 79.6061 − 0.5662[**ABC(G)]**	Surface tension = 80.0087 − 0.3948[SS(G)]
Polarizability = − 2.9516 + 1.9965[**ABC(G)]**	Polarizability = − 4.1564 + 1.3866[SS(G)]
Boiling point = 153.5087 + 20.8404[**ABC(G)]**	Boiling point = 141.6433 + 14.4558[SS(G)]
Enthalpy of variation = 31.769 + 2.915[**ABC(G)]**	Enthalpy of variation = 30.2478 + 2.0188[SS(G)]

Regression model for SC	Regression model for RI
Molecular weight = 13.3437 + 27.1882[**SC(G)]**	Molecular weight = 16.2195 + 28.1100[**RI(G)]**
Complexity = 39.4253 + 43.0217[**SC(G)]**	Complexity = 47.5607 + 44.2587[**RI(G)]**
Density = 1.7876 − 0.0230[**SC(G)]**	Density = 1.7843 − 0.0238[**RI(G)]**
Flash point = 43.2542 + 20.1170[**SC(G)]**	Flash point = 45.7931 + 20.7737[**RI(G)]**
Molar volume = − 64.6865 + 24.836[**SC(G)]**	Molar volume = − 62.3854 + 25.6982[**RI(G)]**
Surface tension = 80.1367 − 0.9262[**SC(G)]**	Surface tension = 80.1996 − 0.9676[**RI(G)]**
Polarizability = − 4.1379 + 3.2252[**SC(G)]**	Polarizability = − 3.7882 + 3.3340[**RI(G)]**
Boiling point = 147.989 + 33.2581[**SC(G)]**	Boiling point = 152.1868 + 34.3436[**RI(G)]**
Enthalpy of variation = 30.9983 + 4.6526[**SC(G)]**	Enthalpy of variation = 31.4694 + 4.8117[**RI(G)]**

Regression model for HZ	Regression model for GA
Molecular weight = 123.0392 + 0.3994[**HZ(G)]**	Molecular weight = 9.4468 + 13.0085[**GA(G)]**
Complexity = 170.448 + 0.6808[**HZ(G)]**	Complexity = 28.0960 + 20.7296[**GA(G)]**
Density = 1.6200 − 0.0003[**HZ(G)]**	Density = 1.7901 − 0.0110[**GA(G)]**
Flash point = 128.5374 + 0.2908[**HZ(G)]**	Flash point = 40.1194 + 9.6328[**GA(G)]**
Molar volume = 52.1288 + 0.3458[**HZ(G)]**	Molar volume = − 67.435 + 11.8602[**GA(G)]**
Surface tension = 76.9305 − 0.0142[**HZ(G)]**	Surface tension = 80.1620 − 0.4401[**GA(G)]**
Polarizability = 11.1992 + 0.0447[**HZ(G)]**	Polarizability = − 4.5259 + 1.5410[**GA(G)]**
Boiling point = 288.9813 + 0.4807[**HZ(G)]**	Boiling point = 142.8059 + 15.9243[**GA(G)]**
Enthalpy of variation = 49.5430 + 0.0686[**HZ(G)]**	Enthalpy of variation = 30.377 + 2.2248[**GA(G)]**

Regression model for ReZ_2	**Regression model for ReZ_1**
Molecular weight = 57.3870 + 0.3780[**ReZ_2(G)**]	Molecular weight = 6.6707 + 11.0317[**ReZ_1(G)**]
Complexity = 29.1286 + 0.6712[**ReZ_2(G)**]	Complexity = 3.7703 + 18.0524[**ReZ_1(G)**]
Density = 1.6995 − 0.0003[**ReZ_2(G)**]	Density = 1.7808 − 0.0091[**ReZ_1(G)**]
Flash point = 71.4891 + 0.2836[**ReZ_2(G)**]	Flash point = 37.1349 + 8.1906[**ReZ_1(G)**]
Molar volume = − 7.2151 + 0.3295[**ReZ_2(G)**]	Molar volume = − 65.6712 + 9.9559[**ReZ_1(G)**]
Surface tension = 78.3134 − 0.0126[**ReZ_2(G)**]	Surface tension = 80.1156 − 0.3699[**ReZ_1(G)**]
Polarizability = 3.4246 + 0.0427[**ReZ_2(G)**]	Polarizability = − 4.2984 + 1.2936[**ReZ_1(G)**]
Boiling point = 194.6729 + 0.4689[**ReZ_2(G)**]	Boiling point = 137.8725 + 13.5410[**ReZ_1(G)**]
Enthalpy of variation = 37.6498 + 0.0655[**ReZ_2(G)**]	Enthalpy of variation = 29.7872 + 1.8895[**ReZ_1(G)**]

Computation of statistical parameters

The use of statistical parameters to compare Topological Indices (TIs) with characteristic of correlation coefficients is useful in model analysis. In a regression model, the standard error (SE) of the estimate measures the mean variance of expected outcomes from actual values, Tables 6, 7 and 8 shows the SE, F-statistics and significance p values. Furthermore, comparison graphs through Figs. 11, 12, 13, 14, 15, 16, 17, 18 and 19 include both actually acquired and mathematically derived physiochemical property values from regression models.

Additionally, the majority of p-values are less than 0.05 a specific value, and mostly r exceeds 0.6 on a consistent basis as seen in Table 4.

Extreme gradient boosting

Extreme Gradient Boosting, is a powerful machine learning method that is well-known for its efficiency in predictive mathematical modeling, here we provided Pseudo-code namely Algorithm-2, provides useful information about XGB, including information about its flexibility and adaptability. The distributions plot of the actual and predicted values are shown in Figs. 20, 21 and 22, which are essential for evaluating the effectiveness of the model and detecting any variations. Furthermore aiding in our understanding is the violin plot, which displays the data distribution graphically while highlighting the peculiarities specific to XGB. Table 9 also offers error estimates, which helps towards a comprehensive

Table 6 Statistical parameter SE of selected TI w.r.t to different physiochemical properties

PP	M_1 (G)	M_2 (G)	H (G)	F (G)	SS (G)	ABC (G)	RI (G)	SC (G)	GA (G)	HZ (G)	ReZ_1 (G)	ReZ_2 (G)
MW	50.098	64.993	21.4217	73.6889	38.810	34.0248	22.3184	26.6421	31.4443	86.6267	45.1974	86.5371
Comp	46.022	53.155	80.4328	53.1921	57.622	56.7448	72.2319	70.3806	67.3475	109.336	54.8500	75.7839
Den	0.1545	0.1609	0.1386	0.1678	0.1472	0.1472	0.1417	0.1419	0.1429	0.1814	0.1498	0.1726
FP	39.990	50.854	30.0609	53.4796	35.212	31.2538	29.1692	30.3989	32.4264	71.2495	39.2341	65.1496
MV	61.389	74.847	30.1454	83.3679	49.639	46.1817	34.1257	37.1889	41.4316	95.7486	55.6984	94.3075
ST	6.7244	6.8226	6.2991	7.0814	6.5623	6.5493	6.3240	6.4041	6.4678	6.8008	6.6074	7.0845
Pol	7.8448	9.6645	3.8262	10.7309	6.2941	5.8387	4.3806	4.6934	5.2106	12.4898	7.1077	12.2490
BP	66.111	84.071	49.6893	88.4150	58.209	51.6654	48.2170	50.2493	53.6010	117.788	64.8577	107.707
E	10.305	12.677	8.0765	13.1314	9.3606	8.3882	7.8016	8.2202	8.7636	16.1544	10.2498	15.7389

Table 7 Statistical parameter F of selected TI w.r.t to different physiochemical properties

PP	M_1 (G)	M_2 (G)	H (G)	F (G)	SS (G)	ABC (G)	RI (G)	SC (G)	GA (G)	HZ (G)	ReZ_1 (G)	ReZ_2 (G)
MW	83.395	46.709	487.405	34.7817	143.621	188.974	448.475	312.633	222.460	23.2333	104.061	23.2960
Comp	271.70	201.92	84.2454	201.633	170.782	176.326	106.140	112.171	123.147	42.3803	189.211	95.7834
Den	5.5750	4.5815	8.6267	3.6527	6.8434	6.8431	7.9472	7.9071	7.6985	2.118	6.3713	3.0698
FP	73.011	42.478	134.601	37.7397	96.1971	123.998	143.390	131.469	114.694	18.2061	76.1270	23.1472
MV	44.703	27.782	207.420	21.03	72.0792	84.3624	160.318	133.890	106.512	14.2541	55.8091	14.9086
ST	5.0446	4.7003	6.7258	3.860	5.6467	5.6972	6.6178	6.2794	6.0191	4.7753	5.4749	3.8512
Pol	46.245	28.082	216.826	21.45	75.7137	89.1208	163.754	141.758	113.688	14.0056	57.8610	14.8396
BP	73.017	42.480	134.645	37.73	96.2157	124.017	143.428	131.506	114.726	18.2072	76.1393	23.1466
E	58.639	36.375	99.8741	33.42	72.5636	92.0793	107.540	96.1721	83.7735	19.7141	59.3577	21.1432

Table 8 Statistical parameter P of selected TI w.r.t to different physiochemical properties

PP	M_1 (G)	M_2 (G)	H (G)	F (G)	SS (G)	ABC (G)	RI (G)	SC (G)	GA (G)	HZ (G)	ReZ_1 (G)	ReZ_2 (G)
MW	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.001
Comp	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Den	0.04	0.04	0.02	0.03	0.03	0.03	0.02	0.02	0.027	0.189	0.039	0.123
FP	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.003	0.000	0.001
MV	0.000	0.001	0.000	0.002	0.000	0.000	0.000	0.000	0.000	0.006	0.000	0.006
ST	0.05	0.06	0.03	0.09	0.04	0.04	0.03	0.040	0.043	0.065	0.051	0.090
Pol	0.000	0.001	0.000	0.002	0.000	0.000	0.000	0.000	0.000	0.007	0.000	0.006
BP	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.003	0.000	0.001
E	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.003	0.001	0.002

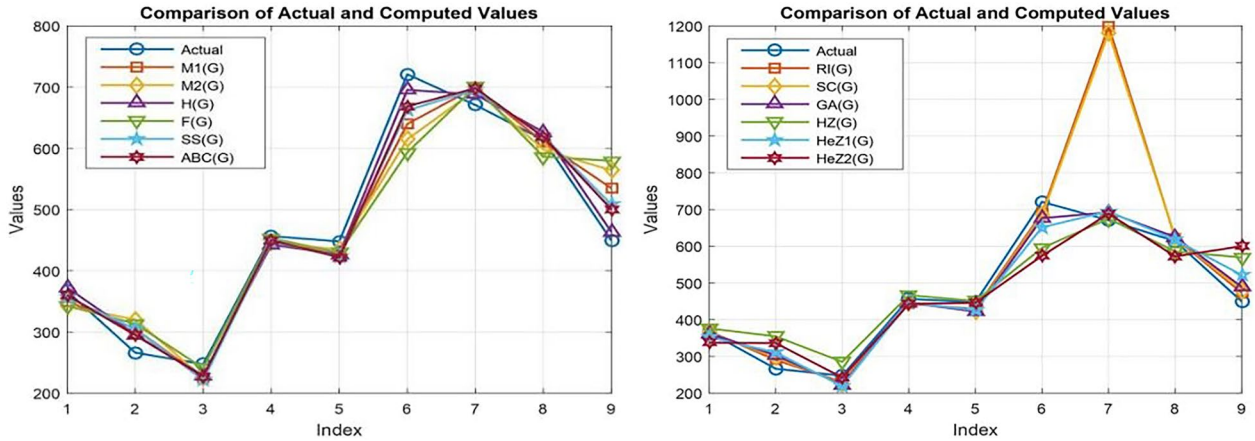


Fig. 11 Graphical comparison w.r.t linear regression for MW

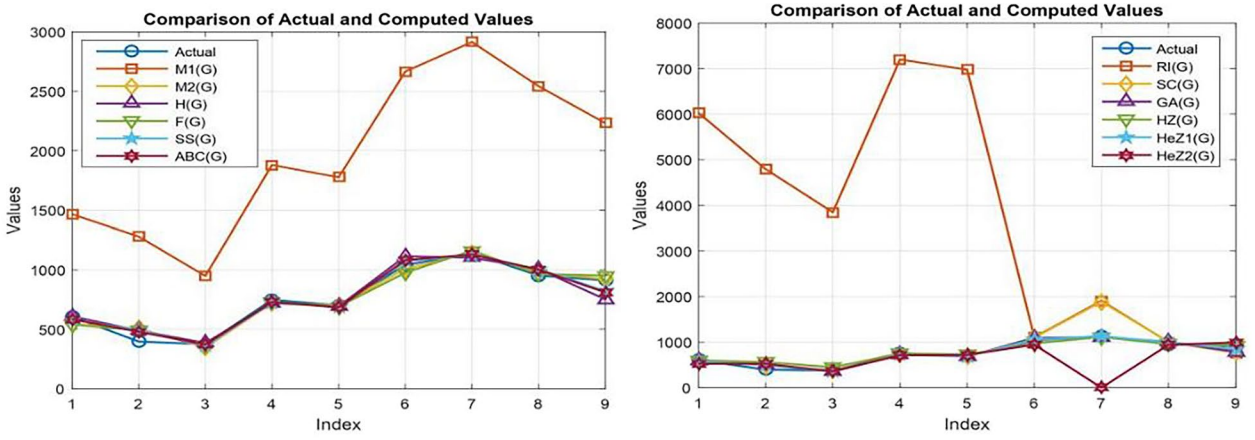


Fig. 12 Graphical comparison w.r.t linear regression for Comp

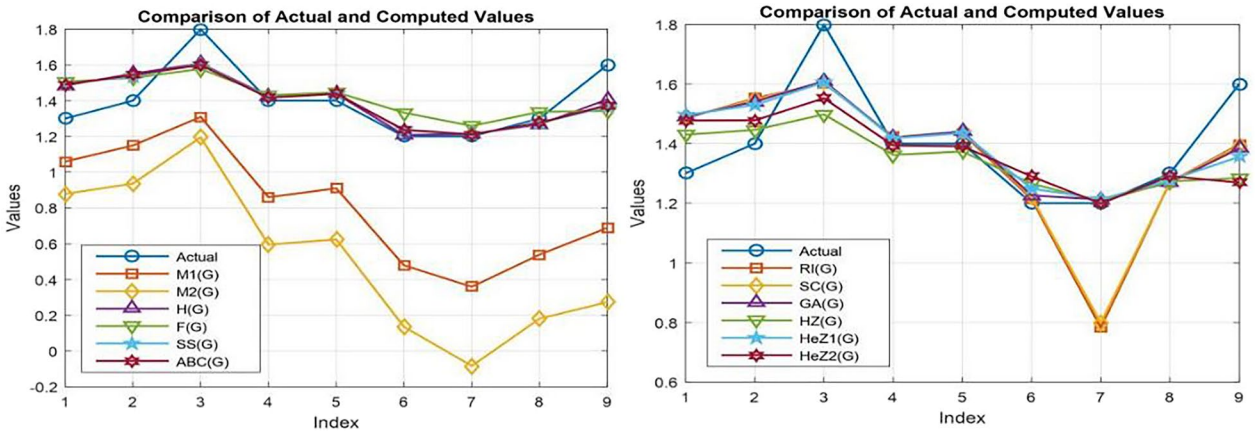


Fig. 13 Graphical comparison w.r.t linear regression for Den

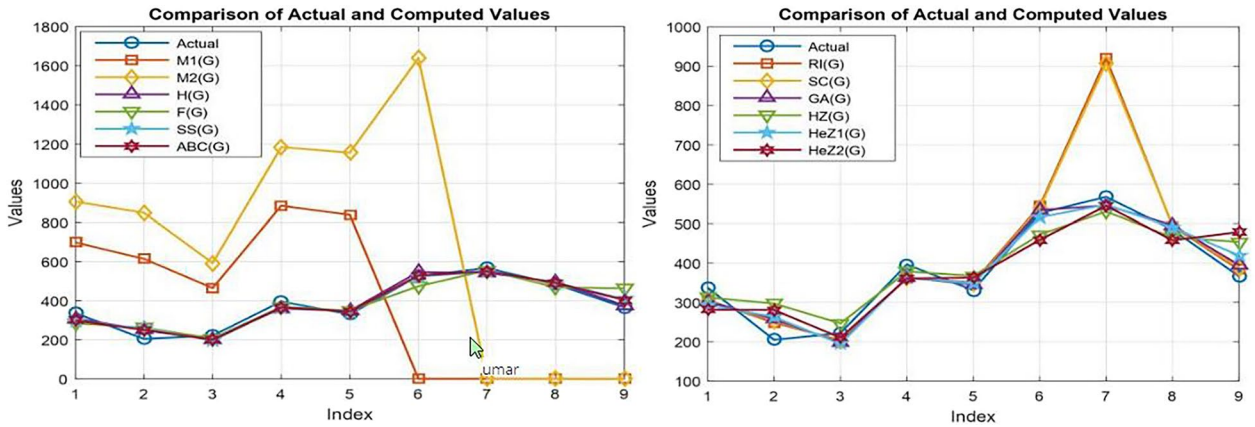


Fig. 14 Graphical comparison w.r.t linear regression for FP

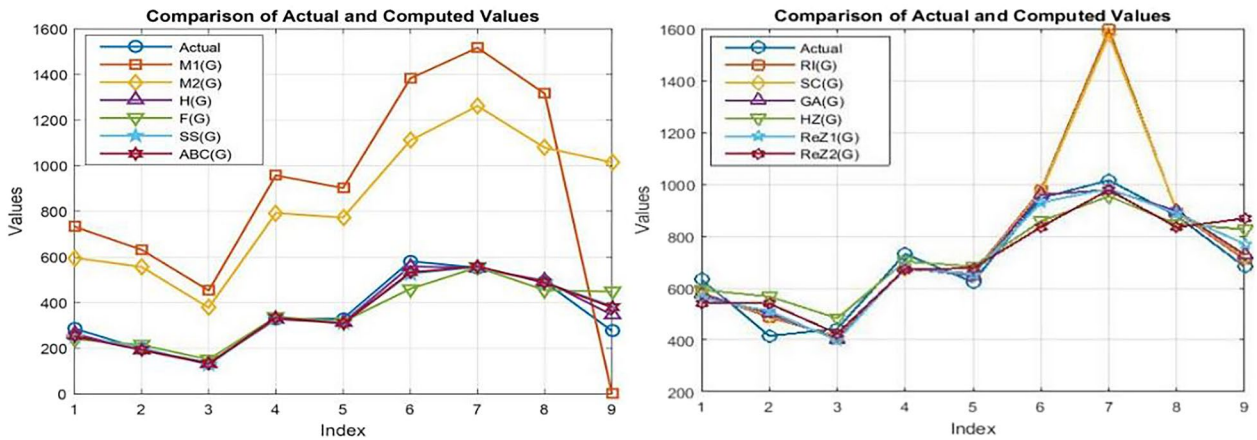


Fig. 15 Graphical comparison w.r.t linear regression for MV

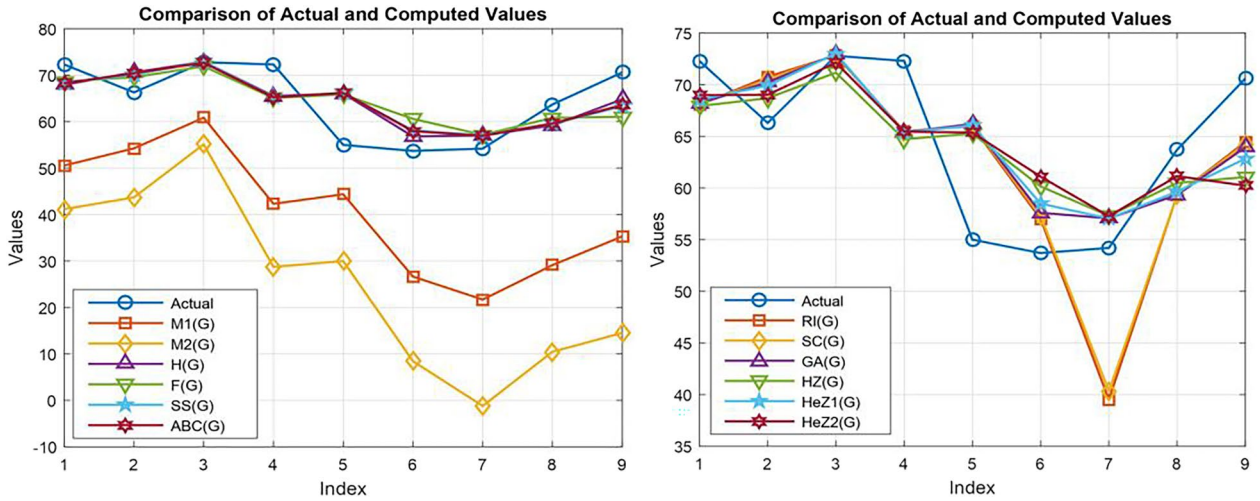


Fig. 16 Graphical comparison w.r.t linear regression for ST

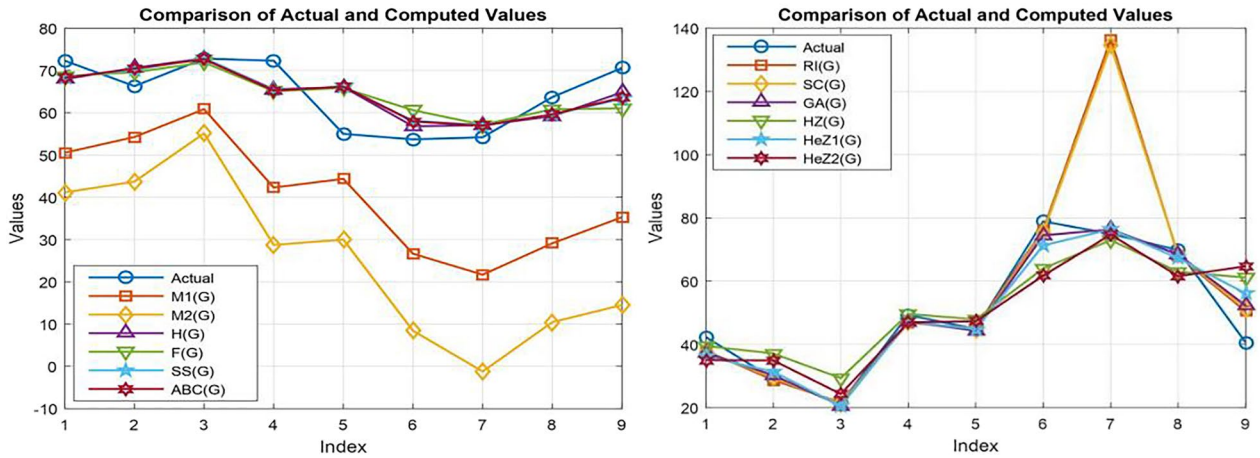


Fig. 17 Graphical comparison w.r.t linear regression for Pol

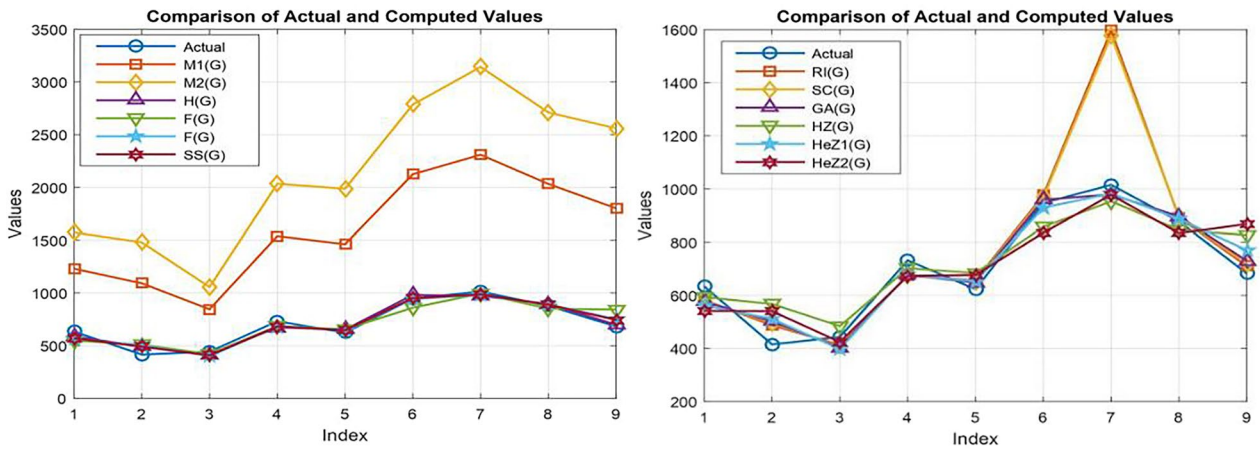


Fig. 18 Graphical comparison w.r.t linear regression for BP

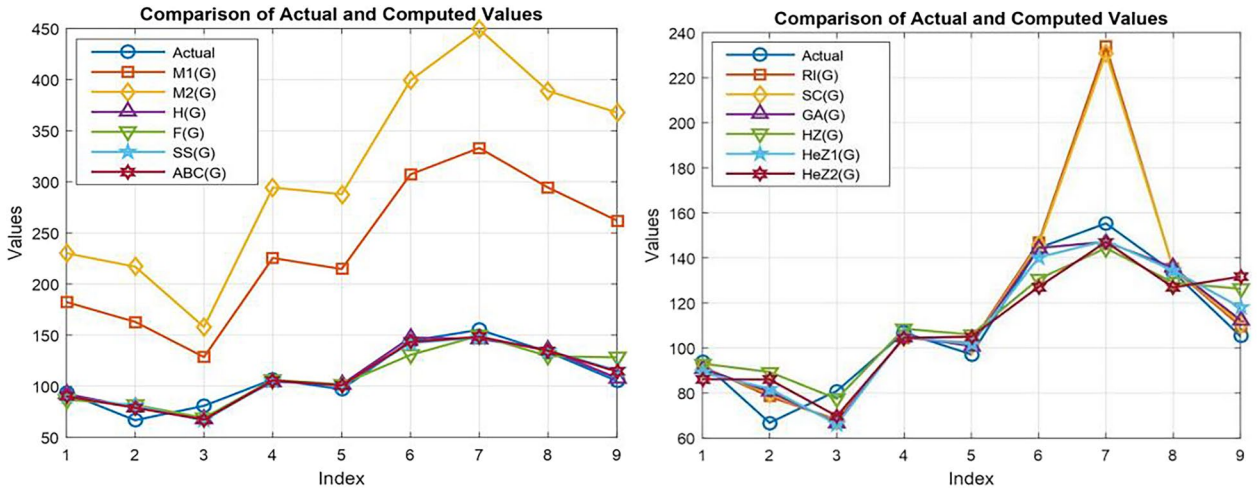
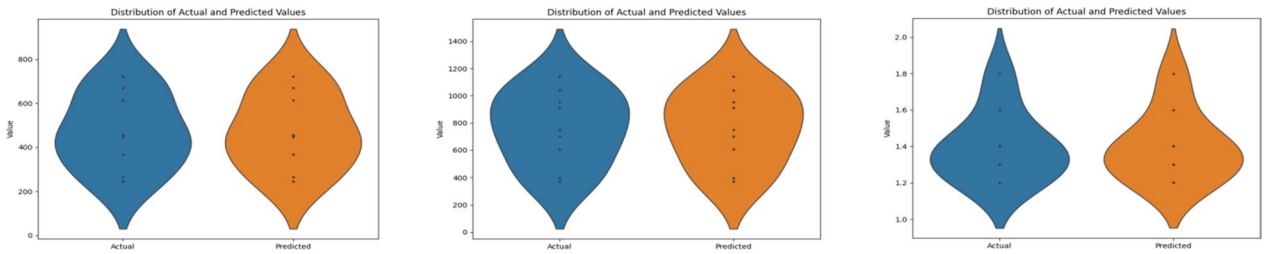
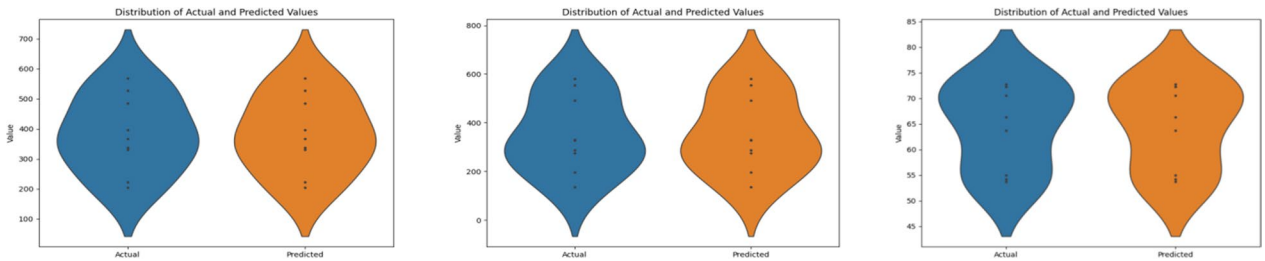


Fig. 19 Graphical comparison w.r.t linear regression for EV



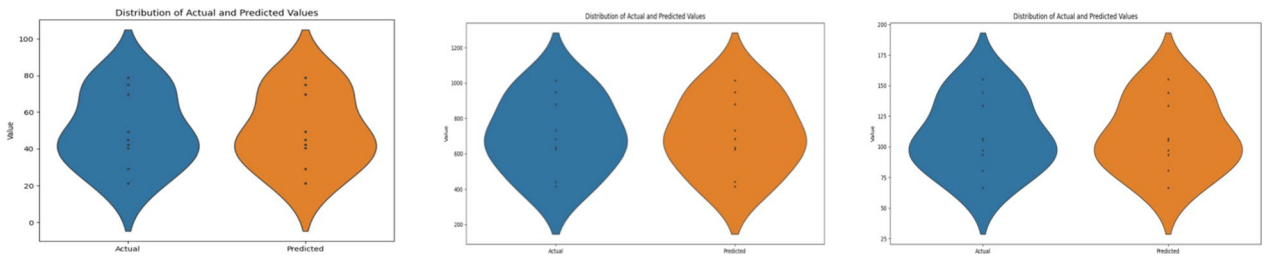
(a) Violin plot for MW (b) Violin plot for Comp (c) Violin plot for Den

Fig. 20 XGB algorithm based violin distribution plot of MW, Comp and Den



(a) Violin plot for FP (b) Violin plot for MV (c) Violin plot for ST

Fig. 21 XGB algorithm based violin distribution plot of FP, MV and ST



(a) Violin plot for Pol (b) Violin plot for BP (c) Violin plot for EV

Fig. 22 XGB algorithm based violin distribution plot of Pol, BP and EV

Table 9 XGB error measurement

PP	MW	Comp	Den	FP	MV	ST	Pol	BP	EV
MAE	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
MSE	4.0729	3.005	5.7742	4.4125	4.3283	7.3177	7.0550	4.7962	5.4326
RMSE	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
R^2	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999

Table 10 Drug properties predicted by the RFA

Drugs	MW (g/mol)	Comp	Den (g/cm^3)	FP (°C)	MV (cm^3)	ST (dyne/cm)	P (cm^3)	BP (°C)	E (kJ/mol)
a	361.98	578.90	1.36	309.61	272.60	69.99	40.18	587.25	89.71
b	284.66	438.18	1.45	238.35	203.62	69.05	30.46	470.53	75.99
c	263.50	404.96	1.67	232.31	161.92	71.64	24.96	460.52	79.77
d	451.40	749.63	1.42	372.34	321.59	67.12	46.84	692.06	103.38
e	443.29	708.86	1.41	347.55	322.41	60.90	45.32	651.10	99.31
f	680.60	1037.04	1.25	518.40	544.12	56.72	74.92	933.48	142.44
g	669.76	1101.52	1.28	549.22	547.41	56.01	74.51	984.45	150.50
h	632.84	978.68	1.29	499.90	504.34	61.20	70.93	897.61	136.86
i	469.76	892.64	1.54	383.23	305.96	68.10	45.42	710.03	108.59

Table 11 Drug properties predicted by the XGB

Drugs	MW (g/mol)	Comp	Den (g/cm^3)	FP (°C)	MV (cm^3)	ST (dyne/cm)	P (cm^3)	BP (°C)	E (kJ/mol)
a	366.4	606.99	1.30	337.29	287	72.29	42.30	634.09	93.70
b	266.30	397	1.40	205.0	197	66.30	29.20	415.40	66.80
c	247.25	374	1.79	221.89	135.20	72.79	21.49	443.29	80.89
d	456.59	748.99	1.39	396.49	328.79	72.29	49.29	731.99	106.79
e	447.90	702	1.39	330.90	329.89	55	44.89	623.60	97.10
f	720.89	1040	1.20	526.59	581.69	53.70	78.89	946.99	144.39
g	670.79	1139.99	1.20	567.69	553.89	54.20	74.99	1014.99	155.29
h	613.79	951.99	1.30	484.69	491	63.69	69.80	877.89	133.69
i	449.40	911.99	1.59	366.0	276.2	70.59	40.60	682.50	105.20

review of the model's predictive power and general accuracy while using XGB algorithm, having a well-organized overview of implementation procedures like the one provided by pseudo-code proves invaluable for expediting the process and improving understanding of its complexities.

Algorithm 2 XGB for QSPR model of anti-HIV

Step 1:

i. Start working with the dataset, import the necessary libraries, such as numpy, pandas, xgboost, matplotlib and plot-tree into Python

ii. In Python, define a dataset as a dictionary. Include key-value pairs, where a collection of data points for a given feature correlates to each key, which shows an attribute

Step 2:

iii. Get the data ready for analysis after defining the data-set dictionary

iv. Use the pd.DataFrame(data) function to convert a dictionary into a pandas DataFrame, making data management and analysis easier

v. Separate the features (X) and the target variable (y)

Step 3:

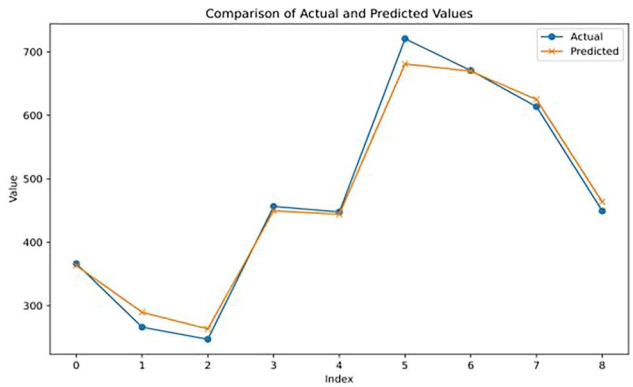
vi. Train an XGB regression model

vii. After the model has been trained, evaluate its efficiency and prediction using unknown data

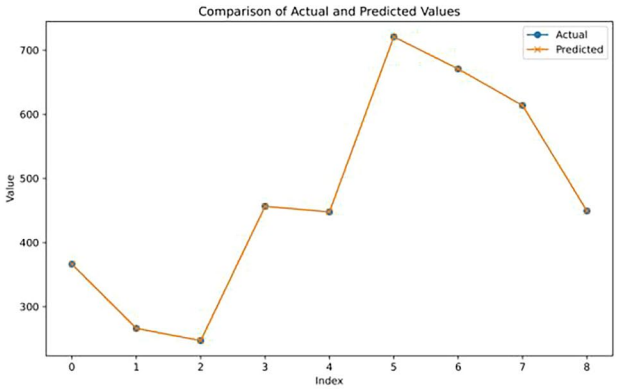
Step 4:

viii. To illustrate the difference between predicted and actual values, make a scatter plot with the actual values on the y-axis and expected values on the x-axis

ix. Provide predicted values in a tabular style, like a DataFrame or structured array, so that they may be easily compared to actual values

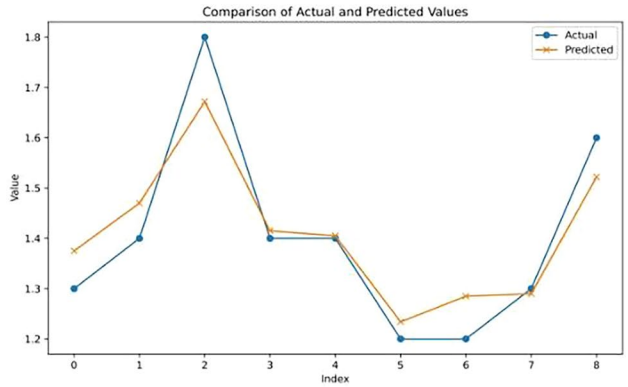


RFA line plot for MW

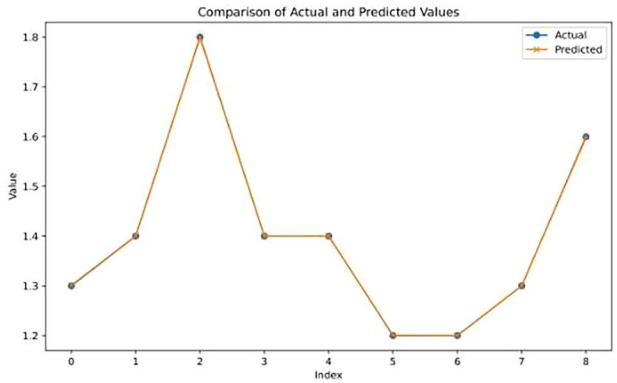


XGB line plot for MW

Fig. 23 Graphical comparison of MW

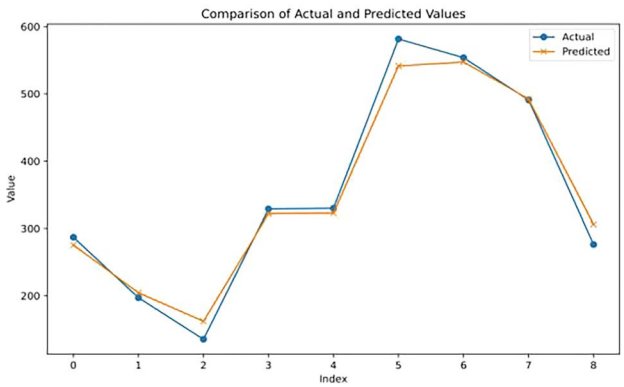


RFA line plot for Den

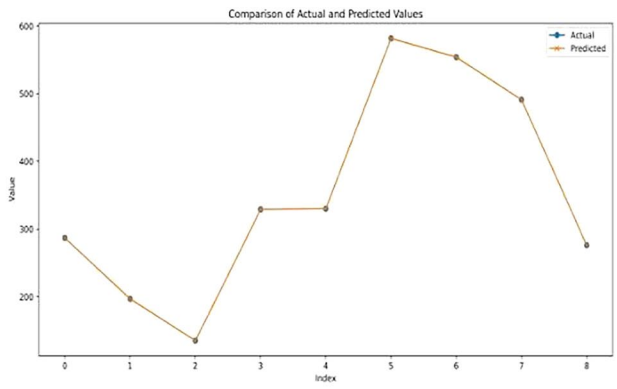


XGB line plot for Den

Fig. 24 Graphical comparison of Den



RFA line plot for MV



XGB line plot for MV

Fig. 25 Graphical comparison of MV

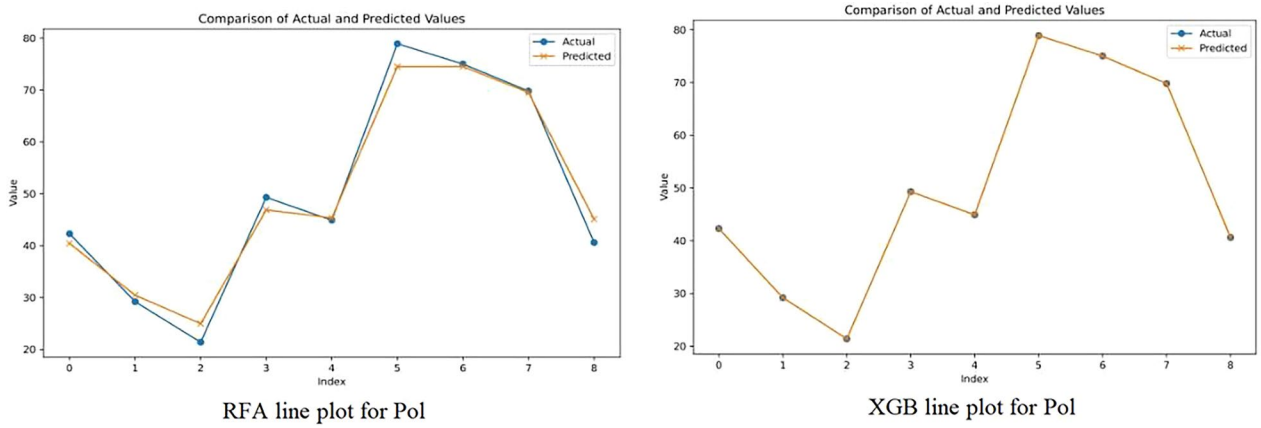


Fig. 26 Graphical comparison of Pol

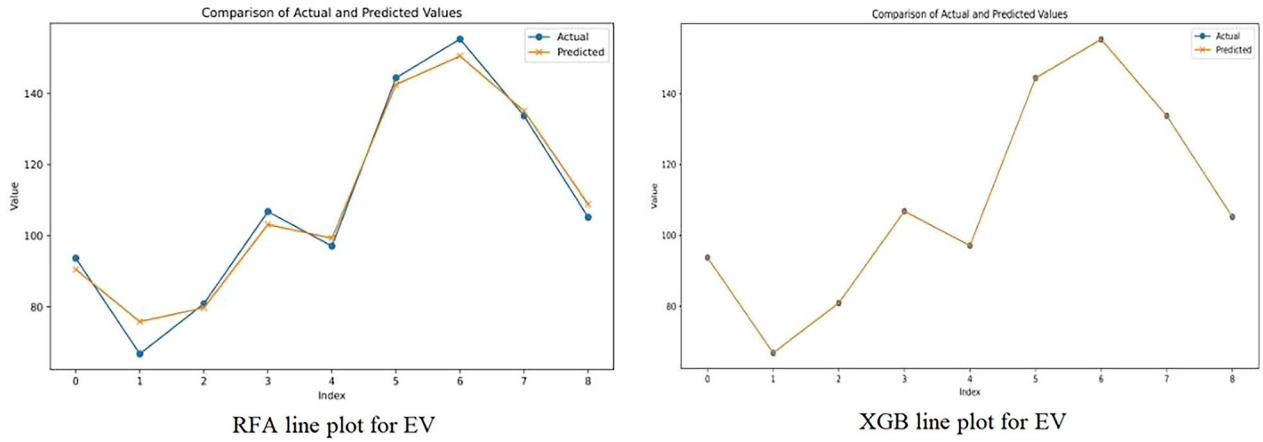


Fig. 27 Graphical comparison of EV

Physio-chemical parameters comparison analysis

When XGB and RFA were used to forecast the physio-chemical properties of anti-HIV medicines, the results showed that XGB predictions consistently produced higher values than RFA. This implies that when it comes to the physio-chemical characteristics of anti-HIV drugs, the XGB algorithm typically yields more optimistic forecasts.

Even though these two machine learning models provide insightful information about the structure–activity relationship of associated drugs, the difference in predicted values emphasizes how crucial it is to take into account a variety of computational strategies and validation methods in order to guarantee the precision and dependability of predictions made during the drug discovery and development process. Tables 10 and 11 are the Experimental and actual data for prediction of RFA and XGB w.r.t physical properties as well as through Figs. 23, 24, 25, 26 and 27 shown the graphical comparison between XGB and RFA listed below.

Standard errors measurements like MAE, MSE, and RMSE are used to evaluate the performance of predictive models like RFA and XGB. To evaluate the relative efficiency of the models and compare the error indicators, visualizations such as tables and graphs were used. In terms of prediction accuracy, XGB performed better than RFA, as seen by lower MAE, MSE, and RMSE values. Furthermore, compared to RFA, greater R^2 values for XGB demonstrated a better fit of the model to the data. It was easier to comprehend why XGB is such a strong algorithm for predictive modeling problems compared to the graphical representations and error tables.

Conclusions

The conclusion of our analysis gives information on the potential efficacy of the drugs under examination in treating HIV-1 disease. In order to predict physiochemical properties, we compared ability to forecast of RFA, Linear Regression, and XGB in this work. Metrics including MAE, MSE, RMSE, and R^2 values were used to assess their effectiveness. With substantially lower error rates and higher R^2 values than the other models, XGB performed better. The efficacy of XGB was further demonstrated by graphical representations. Particularly in the treatment of HIV, the findings have important implications for drug development. Using machine learning algorithms such as XGB can improve drug property prediction efficiency. The superiority of XGB is derived from its iterative prediction refining. Some more techniques and data-set optimization may be investigated in future studies. The research contributes to larger-scale predictive modeling efforts in the pharmaceutical industry. The possibilities of predictive modeling will grow with further development of machine learning techniques. Overall, this work shows that advanced algorithms can be used to improve the drug development process.

Acknowledgements
The authors extend their appreciation to Taif University, Saudi Arabia, for supporting this work through project number (TU-DSPP-2024-94).

Author contributions
All the authors Wakeel Ahmed, Shahid Zaman, Eizzah Asif, Kashif Ali, Emad E. Mahmoud and Mamo Abebe Asheboss have equally contributed to this manuscript in all stages, from conceptualization to the write-up of final draft.

Funding
This research was funded by Taif University, Saudi Arabia, Project No. TU-DSPP-2024-94).

Data availability
No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
The manuscript has been approved by all authors and consent for publication has been granted.

Competing interests
The authors declare no competing interests.

Received: 31 May 2024 Accepted: 12 August 2024
Published online: 12 September 2024

References

1. Khan MM, Khan MM. Acquired immune deficiency syndrome. In: Immunopharmacology. Cham: Springer; 2016. p. 293–330.
2. Sellier P, et al. Updated mortality and causes of death in 2020–2021 in people with HIV: a multicenter study in France. AIDS. 2023;37(13):2007–13.
3. Okoye AA, Picker LJ. CD 4+ T-cell depletion in HIV infection: mechanisms of immunological failure. Immunol Rev. 2013;254(1):54–64.
4. Paiardini M, Müller-Trutwin M. HIV-associated chronic immune activation. Immunol Rev. 2013;254(1):78–101.
5. Veazey RS. Intestinal CD4 depletion in HIV/SIV infection. Curr Immunol Rev. 2019;15(1):76–91.
6. Wilson NL, et al. Identifying symptom patterns in people living with HIV disease. J Assoc Nurses AIDS Care. 2016;27(2):121–32.
7. Joseph SB, et al. HIV-1 target cells in the CNS. J Neurovirol. 2015;21:276–89.
8. Hu L, et al. Dual-channel hypergraph convolutional network for predicting herb–disease associations. Brief Bioinform. 2024;25(2): bbae067.
9. Zhao B-W, et al. Motif-aware miRNA-disease association prediction via hierarchical attention network. IEEE J Biomed Health Inform. 2024;28(7):4281–94.
10. Zhao B-W, et al. iGRLDTI: an improved graph representation learning method for predicting drug–target interactions over heterogeneous biological information network. Bioinformatics. 2023;39(8): btad451.
11. Zhao B-W, et al. A geometric deep learning framework for drug repositioning over heterogeneous information networks. Brief Bioinform. 2022;23(6): bbac384.
12. Lv Q, et al. TCMBank: bridges between the largest herbal medicines, chemical ingredients, target proteins, and associated diseases with intelligence text mining. Chem Sci. 2023;14(39):10684–701.
13. Lv Q, et al. TCMBank-the largest TCM database provides deep learning-based Chinese-Western medicine exclusion prediction. Signal Transduct Target Ther. 2023;8(1):127.
14. Lv Q, et al. Meta learning with graph attention networks for low-data drug discovery. IEEE Trans Neural Netw Learn Syst. 2023;35(8):11218–30.
15. Lv Q, et al. Meta-molnet: a cross-domain benchmark for few examples drug discovery. IEEE Trans Neural Netw Learn Syst. 2024. https://doi.org/10.1109/TNNLS.2024.335965.
16. Lv Q, et al. Mol2Context-vec: learning molecular representation from context awareness for drug discovery. Brief Bioinform. 2021;22(6): bbab317.
17. Lv Q, et al. 3D graph neural network with few-shot learning for predicting drug–drug interactions in scaffold-based cold start scenario. Neural Netw. 2023;165:94–105.
18. Ahmed W, et al. A python based algorithmic approach to optimize sulfonamide drugs via mathematical modeling. Sci Rep. 2024;14(1):12264.
19. Zaman S, et al. On neighborhood eccentricity-based topological indices with QSPR analysis of PAHs drugs. Meas Interdiscip Res Perspect. 2024. https://doi.org/10.1080/15366367.2024.2329950.
20. Ahmed W, et al. Molecular insights into anti-Alzheimer's drugs through predictive modeling using linear regression and QSPR analysis. Modern Phys Lett B. 2024. https://doi.org/10.1142/S0217984924502609.
21. Zaman S, et al. Mathematical modeling and topological graph description of dominating David derived networks based on edge partitions. Sci Rep. 2023;13(1):15159.
22. Zaman S, et al. Mathematical analysis and molecular descriptors of two novel metal–organic models with chemical applications. Sci Rep. 2023;13(1):5314.
23. Aqib M, et al. On topological indices of some chemical graphs. Mol Phys. 2023. https://doi.org/10.1080/00268976.2023.2276386.
24. Bhatia KS, Gupta AK, Saxena AK. Physicochemical significance of topological indices: importance in drug discovery research. Curr Top Med Chem. 2023;23(29):2735–42.
25. Zanni R, et al. What place does molecular topology have in today's drug discovery? Expert Opin Drug Discov. 2020;15(10):1133–44.
26. Ullah A, Bano Z, Zaman S. Computational aspects of two important biochemical networks with respect to some novel molecular descriptors. J Biomol Struct Dyn. 2024;42(2):791–805.
27. Ullah A, et al. Predictive potential of K-Banhatti and Zagreb type molecular descriptors in structure–property relationship analysis of some novel drug molecules. J Chin Chem Soc. 2024;71(3):250–76.
28. Zaman S, et al. Three-dimensional structural modelling and characterization of sodalite material network concerning the irregularity topological indices. J Math. 2023;2023(1):5441426.

29. Zhang X, et al. The study of curve fitting models to analyze some degree-based topological indices of certain anti-cancer treatment. Chem Pap. 2024;78(2):1055–68.
30. Meharban S, et al. Molecular structural modeling and physical characteristics of anti-breast cancer drugs via some novel topological descriptors and regression models. Curr Res Struct Biol. 2024;7: 100134.
31. Patel HM, et al. Quantitative structure–activity relationship (QSAR) studies as strategic approach in drug discovery. Med Chem Res. 2014;23:4991–5007.
32. Zaman S, et al. QSPR analysis of some novel drugs used in blood cancer treatment via degree based topological indices and regression models. Polycycl Aromat Compd. 2023;44:1–17.
33. Hakeem A. et al. QSPR analysis of some novel drugs used for cardiovascular diseases through degree-based topological indices and regression models. 2023.
34. Gutman I, Polansky OE. Mathematical concepts in organic chemistry. Berlin: Springer Science & Business Media; 2012.
35. Fajtlowicz S. On conjectures of Graffiti-II. Congr Numer. 1987;60:187–97.
36. Furtula B, Gutman I. A forgotten topological index. J Math Chem. 2015;53(4):1184–90.
37. Zhao W, et al. Computing SS index of certain dendrimers. J Math. 2021;2021:1–14.
38. Ashraful Alam M, et al. Degree-based entropy for a non-kekulean benzenoid graph. J Math. 2022;2022:1–12.
39. Gutman I, Furtula B, Katanić V. Randić index and information. AKCE Int J Graphs Comb. 2018;15(3):307–12.
40. Farahani MR. On the Randic and sum-connectivity index of nanotubes. Ann West Univ Timisoara-Math Comput Sci. 2013;51(2):39–46.
41. Shirdel GH, Rezapour H, Sayadi AM. The hyper-zagreb index of graph operations. Iran J Math Chem. 2013;4(2):213–20.
42. Ranjini P, Lokesha V, Usha A. Relation between phenylene and hexagonal squeeze using harmonic index. Int J Graph Theory. 2013;1(4):116–21.
43. Havare ÖÇ. Topological indices and QSPR modeling of some novel drugs used in the cancer treatment. Int J Quantum Chem. 2021;121(24): e26813.
44. Kirmani SAK, Ali P, Azam F. Topological indices and QSPR/QSAR analysis of some antiviral drugs being investigated for the treatment of COVID-19 patients. Int J Quantum Chem. 2021;121(9): e26594.
45. Gnanaraj LRM, Ganesan D, Siddiqui MK. Topological indices and QSPR analysis of NSAID drugs. Polycycl Aromat Compd. 2023;43(10):9479–95.
46. Huang L, et al. Topological indices and QSPR modeling of new antiviral drugs for cancer treatment. Polycycl Aromat Compd. 2023;43(9):8147–70.
47. Pence HE, Williams A. ChemSpider: an online chemical information resource. Washington, DC: ACS Publications; 2010.
48. Kim S, et al. PubChem 2019 update: improved access to chemical data. Nucleic Acids Res. 2019;47(D1):D1102–9.

Publisher's Note