Contemporary Mathematics
http://ojs.wiserpub.com/index.php/CM/

UNIVERSAL WISER PUBLISHER

Research Article

Machine Learning Regression Models for Predicting Anti-Cancer Drug Properties: Insights from Topological Indices in QSPR Analysis

Simran Kour, Ravi Sankar J.*

Department of Mathematics, School of Advanced Sciences, Vellore Institute of Technology, Vellore, Tamil Nadu, 632 014, India
E-mail: ravisankar.j@vit.ac.in

Received: 30 September 2024; **Revised:** 15 November 2024; **Accepted:** 12 December 2024

Abstract: This study advances the prediction of anti-cancer drug properties by integrating machine learning regression techniques with topological indices derived from hydrogen-depleted molecular graphs. Focusing on distance-based metrics, we investigate how atomic spatial arrangements influence molecular characteristics, enhancing Quantitative Structure-Property Relationship (QSPR) frameworks. Our analysis reveals significant correlations between various topological indices and key molecular properties, including polarizability, molar refractivity, and boiling point. The regression models, particularly Linear and Ridge Regression, achieved high predictive performance, with R^2 values exceeding 0.80, low Root Mean Square Error (RMSE), significant F-test, and p-values below 0.05 for highly correlated properties. The importance of careful index selection is underscored, as models using indices with strong correlations show superior predictive accuracy. This approach offers a more robust framework for predicting physicochemical properties, enhancing the efficiency of screening processes and optimizing lead compounds in oncological research. By bridging theoretical modeling and practical drug discovery, this work has the potential to accelerate the development of more effective and targeted anti-cancer therapies.

Keywords: anti-cancer drugs, molecular properties, topological indices, machine learning, regression models

MSC: 05C10, 05C12, 74E40

1. Introduction

In the 21st century, cancer persists as a leading cause of illness and death, presenting significant challenges to global health. This life-threatening condition arises from a breakdown in the body's cellular control systems. Instead of following their natural life cycle, damaged cells evade programmed death and grow uncontrollably. This process leads to the formation and spread of abnormal cells throughout the body, a phenomenon known as metastasis. The hallmark of cancer unregulated cellular proliferation presents an ongoing challenge to medical science and public health efforts worldwide [1]. With over 200 distinct types, cancers are typically named after the tissue where they are first identified. Carcinogens, such as those found in tobacco smoke, can trigger DNA mutations that lead to cancer development. Symptoms vary widely, including lumps, abnormal bleeding, persistent cough, and weight loss, while risk factors encompass tobacco use, obesity, poor diet, inactivity, and excessive alcohol consumption [2]. Early detection of tumors plays a critical role in reducing mortality rates. To address the complexity of cancer, various anti-cancer drugs have been developed, classified

DOI: https://doi.org/10.37256/cm.5420245826

into major categories such as alkylating agents, anti-metabolites, natural products, and hormones. While treatment remains challenging, prevention strategies and early detection are vital in reducing the global burden of the disease [3, 4].

Chemical graph theory combines graph theory and chemistry to model molecules and study their properties. Topological indices, which serve as numerical descriptors in these chemical graphs, strongly correlate with various biological and physicochemical properties of molecules. These indices are key to QSPR, linking molecular structure to properties. This study focuses on refining distance-based topological indices for anti-cancer drugs, particularly using hydrogen depleted graphs. In this study, ten anti-cancer drugs namely, altretamine [5], busulfan [6], carmustine [7], cyclophosphamide [8], dacarbazine [9], ifosfamide [10], mechlorethamine [11], procarbazine [12], temozolomide [13], and thiotepa [14] have been investigated. By eliminating hydrogen atoms that don't affect core connectivity, a clearer analysis of the molecule's topological properties is achieved, enhancing the predictive accuracy of indices related to the biological activity of anti-cancer drugs [15, 16].

A QSPR analysis explores the correlation between topological indices and the biological activity of anti-cancer drugs. In this study, machine learning advanced regression models are employed to find the best correlation between topological indices and physicochemical properties. Specifically, Linear Regression, as it is a straightforward yet powerful method to uncover basic relationships between variables. Ridge Regression, an extension of Linear Regression, enhances prediction accuracy by addressing multicollinearity through regularization, making it more robust in the presence of highly correlated predictors. Elastic Net Regression, which merges the strengths of both Lasso and Ridge, is particularly effective for datasets with numerous predictors, offering improved stability and better variable selection. Lastly, Bayesian Linear Regression incorporates a probabilistic approach, allowing us to quantify uncertainty in the predictions and provide a more comprehensive view of the model's performance. These approach captures non-linear relationships and handles multicollinearity, resulting in more robust predictions of biological activity. Combining graph theory and machine learning, this framework aims to advance the rational design of novel anti-cancer drugs with improved therapeutic potential [17].

Recent studies have explored the use of topological indices to predict the therapeutic efficacy of anti-cancer drugs. Zhang et al. investigated the use of topological indices derived from molecular graphs to predict the effectiveness of anti-cancer drugs, focusing on drugs such as pamidronic acid and olaparib for blood cancer. Their work delves into the connections between the physical and chemical properties of these drugs, leveraging topological indices to predict attributes such as molecular weight and structural complexity. This work laid the groundwork for understanding how molecular structures can be linked to drug efficacy, which directly informs our approach to using topological indices to predict drug properties in our study [18]. Balasubramaniyan et al. applied QSPR analysis to investigate prostate cancer drugs, utilizing degree-based and innovative neighborhood-degree-based indices to uncover correlations with physicochemical properties. Their study of nineteen drugs has enhanced our understanding of how molecular descriptors can be used to predict drug efficacy, opening the door to more targeted treatments. Our study extends this work by incorporating machine learning regression techniques to improve the predictive accuracy of topological indices, specifically for anti-cancer drugs [19]. Sethumadhavan et al. investigated the use of topological indices to predict the physicochemical properties of lung cancer drugs, revealing a robust correlation between these indices and the drugs' physical properties. This framework has proven useful in minimizing experimental trials during drug development. Our study advances this work by refining predictive methods through the application of distance-based metrics from molecular graphs, incorporating additional molecular properties for more comprehensive predictions [20]. Zabidi et al. pioneered the use of degree-based topological indices for predicting HOMO and LUMO energies, leveraging machine learning algorithms to refine the predictions. Their study highlighted how linear regression with moment Balaban indices could outperform traditional methods, demonstrating the integration of machine learning with topological indices. We extend these ideas by focusing on a distinct set of molecular properties, applying similar techniques to improve the prediction of anti-cancer drug properties [21]. Yu et al. devised an algorithm based on k-eccentricity-based topological indices to predict anti-HIV activity, emphasizing the significance of molecular descriptors through their work on 3-eccentricity. This approach highlights the value of graph-theoretical methods in drug discovery. Our research takes this idea further by applying distance-based metrics, demonstrating their utility in predicting the properties of anti-cancer drugs [22]. Dhanajayamurthy et al. explored physicochemical properties like enthalpy and molar volume to the molecular structures

of anti-cancer drugs, with QSPR analysis serving as a key tool in drug design. However, while traditional topological indices have been widely explored, reduced neighborhood topological indices remain underutilized. This study addresses this gap by introducing these indices and demonstrating their strong correlation with drug properties, contributing a novel approach to predictive modeling and enhancing the accuracy of drug design beyond existing methods [23]. Shanmukha et al. introduced distance-based topological indices, demonstrating strong correlations with the physicochemical properties of anti-cancer drugs. By applying QSPR analysis, it provides a novel approach to predictive modeling, contributing to a more refined understanding of how these indices can enhance drug design and development. However, much of the research has focused on traditional molecular descriptors, leaving a gap in the exploration of distance-based topological indices [24].

Building on the work of these foundational studies, our research integrates machine learning regression techniques with topological indices from hydrogen-depleted molecular graphs, focusing on distance-based metrics. Through the investigation of atomic spatial arrangements and their impact on molecular properties, we propose a stronger framework for predicting essential physicochemical properties of anti-cancer drugs. The strong performance of our models, particularly with Linear and Ridge Regression, emphasizes the significance of selecting the right indices to enhance prediction accuracy and optimize the drug discovery process.

In this study, we used regression models to establish baseline relationships between topological indices and drug properties. While linear models are useful for initial analysis, we recognize that these models may not fully capture the complex non-linear relationships present in the data. In future research, we plan to incorporate more sophisticated machine learning models, such as support vector machines (SVMs) and artificial neural networks (ANNs), which are better suited for modeling non-linear correlations.

The outline of the paper is as follows: Section 2 presents the materials and methods used in the study, detailing the process of data collection and the analysis of topological indices. Section 3 showcases the key findings, emphasizing the models' performance and their accuracy in predicting key outcomes. Section 4 provides a detailed discussion of the findings, including a comprehensive analysis of model performance and implications for predictive modeling. Finally, Section 5 concludes the paper by highlighting the major insights and addressing the limitations of the study.

2. Material and method

This section will examine ten commonly prescribed anti-cancer drugs, focusing on their primary uses in cancer treatment. These drugs have been chosen for their proven effectiveness, widespread use, and significant impact in oncology. For each drug, we will present its name, chemical structural formula, and key applications, providing insights into their chemical properties and molecular structures. The chemical formulas will illustrate the arrangement of atoms within each molecule, highlighting functional groups and attributes crucial to their biological activity. By understanding these chemical details, we can better grasp how these drugs work, their interactions with cellular mechanisms, and the rationale behind their use in various cancer therapies. Information on these drugs, including their names, structural formulas, and uses, is sourced from PubChem, a database curated by National Institutes of Health (NIH) [25]. The Table 1 provides a detailed overview of these drugs and their characteristics.

Table 1. Anti-cancer drugs with corresponding chemical structures and therapeutic uses

Drug name	Chemical structures	Therapeutic uses
Altretamine	$C_9H_{18}N_6$	Palliative treatment for recurrent or persistent ovarian cancer, following cisplatin and/or alkylating agents like cyclophosphamide.
Busulfan	$C_6H_{14}O_6S_2$	Used in conditioning regimens before allogeneic hematopoietic progenitor cell transplantation for chronic myelogenous leukemia (CML) and bone marrow transplantation.
Carmustine	$C_5H_9Cl_2N_3O_2$	Treatment of brain tumors, due to its ability to penetrate the blood-brain barrier.
Cyclophosphamide	$C_7H_{15}Cl_2N_2O_2P$	Widely used in various cancers for initial treatment, relapse, and advanced stages in combination therapies.
Dacarbazine	$C_6H_{10}N_6O$	Primarily treats metastatic malignant melanoma and Hodgkin's disease when other therapies fail.
Ifosfamide	$C_7H_{15}Cl_2N_2O_2P$	Used for testicular cancer, cervical cancer, soft tissue sarcomas, bladder cancer, osteosarcoma, small cell lung cancer, and non-Hodgkin's lymphoma.
Mechlorethamine	$C_5H_{11}Cl_2N$	Palliative treatment for advanced Hodgkin's disease, chronic leukemias, mycosis fungoides, and metastatic carcinomas with effusion.
Procarbazine	$C_{12}H_{19}N_3O$	Used in combination therapy for advanced Hodgkin's disease (Stage III and IV).
Temozolomide	$C_6H_6N_6O_2$	Treats glioblastoma and refractory anaplastic astrocytoma, especially in CNS malignancies.
Thiotepa	$C_6H_{12}N_3PS$	Pre-transplant conditioning for hematopoietic progenitor cell transplantation (HPCT) in cancers like breast, ovarian, and bladder cancers.

In our upcoming analysis, we will use physicochemical properties which are fundamental to our QSPR approach. By integrating eight key properties, as detailed in Table 2, our aim is to enhance the molecular interactions and improve predictive modeling, ultimately refining the precision of drug design. Following this, Figure 1 presents the molecular graphs of the drugs, which illustrate their structural configurations and form the basis for the calculation of topological indices in our analysis.

Table 2. Anti-cancer drugs with their physicochemical properties

Drug	Boiling point	Melting point	Enthalpy	Flash point	Molar refractivity	Polarizability	Surface tension	Molar volume
Altretamine	339.4	168	58.3	159.1	63.5	25.2	53.8	183.1
Busulfan	464.0	118	69.8	234.4	50.9	20.2	46.6	182.4
Carmustine	309.6	30	63.8	141.0	46.6	18.5	50.4	146.4
Cyclophosphamide	336.1	51	57.9	157.1	58.1	23.0	44.3	195.7
Dacarbazine	456.3	205	71.6	229.7	46.2	18.3	60.7	122.6
Ifosfamide	336.1	40	57.9	157.1	58.1	23.0	44.3	195.7
Mechlorethamine	110.3	108	34.9	20.5	38.6	15.3	31.3	141.1
Procarbazine	384.6	223	63.3	148.9	65.8	26.1	37.9	213.6
Temozolomide	526.6	212	80.1	272.3	45.6	18.1	105.1	98.4
Thiotepa	270.2	51.5	50.8	117.2	49.1	19.5	77.8	125.8

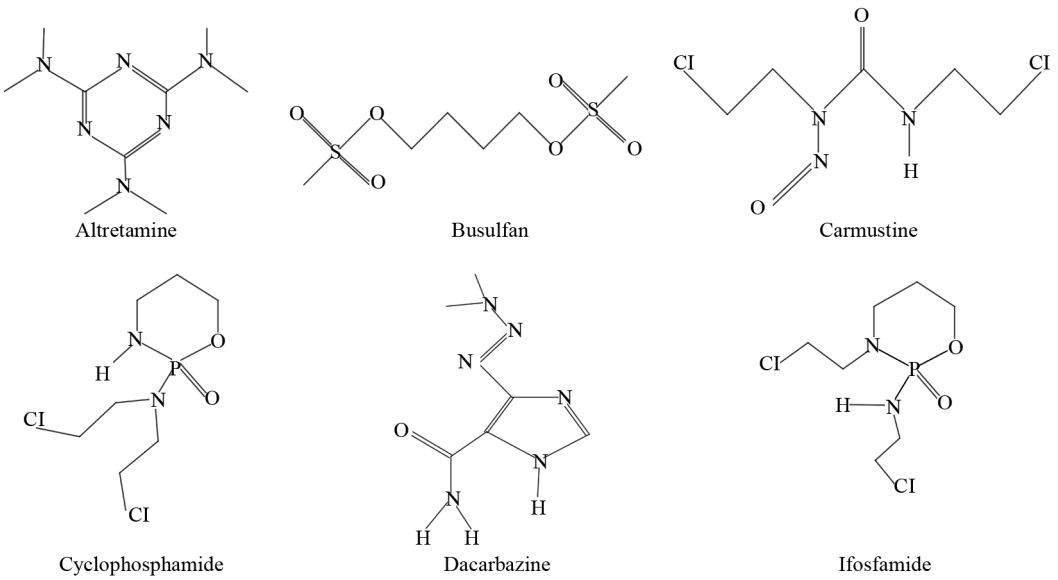

Figure 1. Molecular graphs of anti-cancer drugs

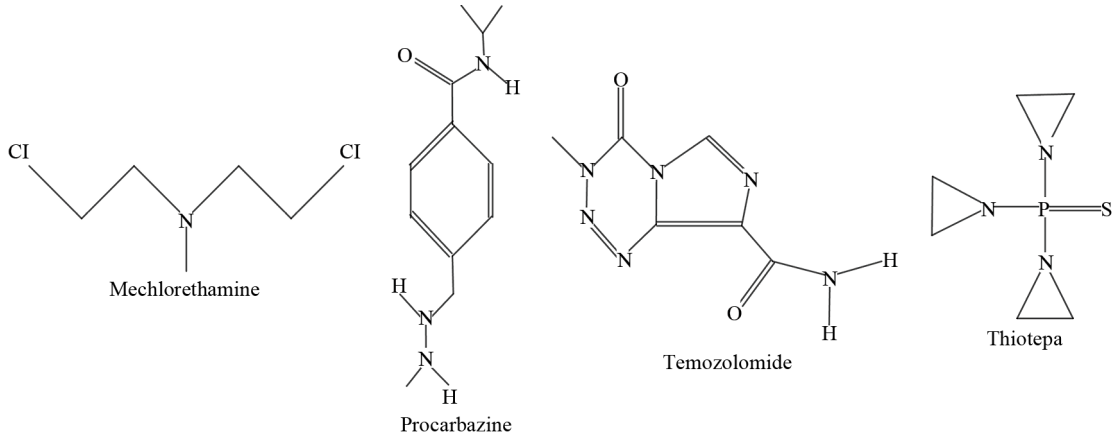

Figure 2. Molecular graphs of anti-cancer drugs (contd.)

2.1 *Distance-based topological indices*

This subsection highlights five important distance-based topological indices namely, Wiener Index, Hyper-Wiener Index, Harary Index, Detour Index, and Detour Harary Index. These indices are utilized to analyze the molecular structures of anti-cancer drugs. These indices provide essential quantitative insights into the size, connectivity, and structural features of molecules, which are crucial for predictive modeling in drug design. We specifically use hydrogen-depleted graphs, excluding hydrogen atoms to emphasize the core molecular framework that more directly impacts molecular properties and biological activities. The computed topological indices for these drugs are presented in Table 3, offering a detailed perspective on their structural characteristics and potential impacts on biological function.

Definition 1 The Wiener Index [26] of the molecular graph G is defined as

$$W(G) = \sum_{1 \le x < y \le n} d(v_x, v_y) \tag{1}$$

where, $d(v_x, v_y)$ be the distance between the vertices v_x and v_y.

Definition 2 The Hyper-Wiener Index [27] of the molecular graph G is defined as

$$WW(G) = \sum_{1 \le x < y \le n} \frac{d(v_x, v_y) + d^2(v_x, v_y)}{2} \tag{2}$$

Definition 3 The Harary Index [28] of the molecular graph G is defined as

$$H(G) = \sum_{1 \le x < y \le n} \frac{1}{d(v_x, v_y)} \tag{3}$$

Definition 4 The Detour Index [29] of the molecular graph G is defined as

$$D(G) = \sum_{1 \le x < y \le n} D(v_x, v_y) \tag{4}$$

where, $D(v_x, v_y)$ be the length of the longest path between the vertices v_x and v_y.

Definition 5 The Detour Harary [30] Index of the molecular graph G is defined as

$$DH(G) = \sum_{1 \le x < y \le n} \frac{1}{D(v_x, v_y)} \tag{5}$$

Table 3. Topological indices of anti-cancer drugs

Name of the drug	Wiener index	Hyper-wiener index	Harary index	Detour index	Detour harary index
Altretamine	354	906	42.6	552	28.7786
Busulfan	393	1,353	34.05	393	34.05
Carmustine	226	605	27.5202	226	27.5202
Cyclophosphamide	301	765	37.8357	433	27.3230
Dacarbazine	262	693	33.1214	391	23.9218
Ifosfamide	307	804	37.6536	499	25.0116
Mechlorethamine	75	165	14.3161	75	14.3161
Procarbazine	511	1,735	44.004	667	32.6826
Temozolomide	277	664	40.3024	668	17.3294
Thiotepa	139	277	28	196	20

3. Main results

This study conducted an extensive QSPR analysis on ten anti-cancer drugs, examining their physicochemical attributes using hydrogen-depleted molecular structures. Essential properties mentioned in Table 2 were meticulously computed from reliable databases. Our objective was to investigate how five selected distance-based topological indices correlate with key physicochemical properties crucial for influencing drug design and behavior. Machine learning models were employed to further enhance predictive accuracy, integrating these topological indices with physicochemical properties to refine QSPR analysis. Each distance-based topological index was computed to capture distinct aspects of molecular topology that may impact the properties of the drugs.

3.1 *Correlation analysis*

Within our QSPR study, correlation analysis was conducted and calculated Pearson coefficient to rigorously assess the strength and direction of these relationships. This analytical approach helps identify meaningful associations within our dataset, highlighting the impact of molecular topology on the observed physicochemical properties.

We present Pearson correlation coefficients in a heatmap of Figure 3, illustrating the relationships between the selected topological indices and the aforementioned properties of the anti-cancer drugs based on hydrogen-depleted structures. These coefficients indicate varying degrees of correlation: strong correlations are represented by values greater than 0.7 or less than -0.7, moderate correlations fall between 0.5 and 0.7 or between -0.5 and -0.7, and weak correlations are below 0.5 or above -0.5. These correlations offer insights into how specific topological features influence drug properties. This visual representation offers an intuitive understanding of how molecular topology affects drug properties, enhancing the depth of our findings beyond the numerical data.

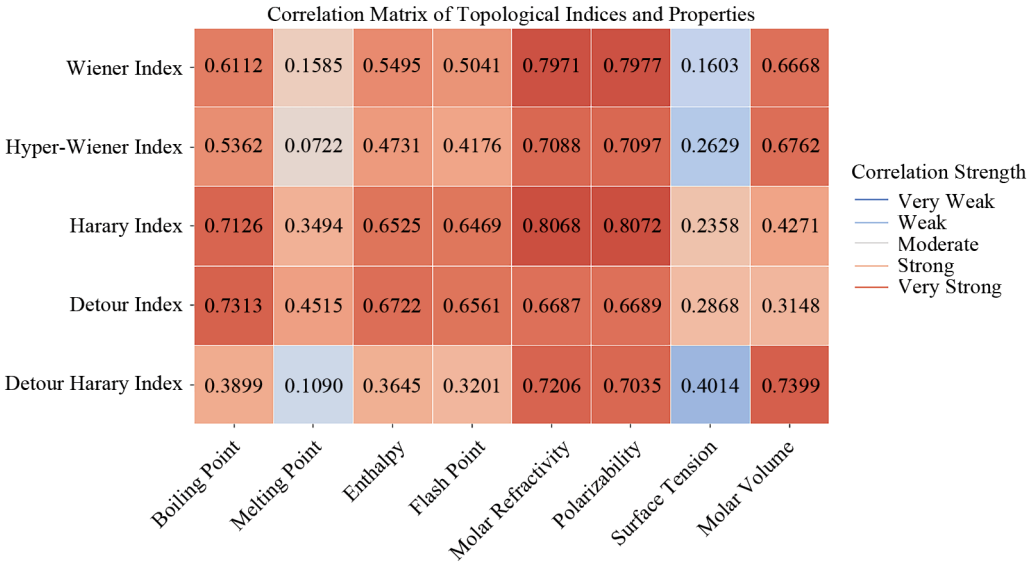

Figure 3. Correlation heatmap for topological indices and physicochemical properties of drugs

3.2 *Machine learning models for QSPR*

In this section, we applied machine learning to examine the QSPR of anti-cancer drugs. Four regression models were utilized: Linear Regression, Ridge Regression, Elastic Net Regression, and Bayesian Linear Regression. The models were trained using the default hyperparameters provided by scikit-learn. No additional hyperparameter tuning or optimization was performed to maintain consistency across models and to assess their performance under standard conditions. These

models, ranging from simple linear to more complex regularized and probabilistic approaches, were trained on our dataset to predict molecular properties from topological indices.

The models were trained on a dataset comprising molecular topological indices, which were derived from the chemical structures of the drugs. These indices were used as features to predict various molecular properties. We focused on strong correlations $|r| > 0.7$, ensuring a focus on relationships that are likely meaningful in the context of drug properties.

To evaluate the models, we used R^2, F-test, p-value, and RMSE. R^2 reveals how well the model explains the variance in the target property, while the F-test and p-value assess the statistical significance of the relationships between features and target. RMSE was used to determine the closeness of the model's predictions to the actual values.

The detailed performance metrics of each regression model, as summarized in Tables 4-8, illustrate their efficacy and accuracy in predicting the molecular properties.

Table 4. Regression analysis for wiener index

Property	Regression model	Regression equation	R^2	F-test	p-value	RMSE
Molar refractivity	Linear regression	36.2057 + 0.0564 (W (G))	0.6353	13.9388	0.0058	5.0140
	Ridge regression	36.2058 + 0.0564 (W (G))	0.6353	13.9388	0.0058	5.0140
	Elastic net regression	36.2166 + 0.0564 (W (G))	0.6353	13.9388	0.0058	5.0140
	Bayesian Linear Regression	37.2278 + 0.0528 (W (G))	0.6328	13.9388	0.0058	5.0317
Polarizability	Linear regression	14.3590 + 0.0224 (W (G))	0.6364	13.9993	0.0057	1.9836
	Ridge regression	14.3591 + 0.0224 (W (G))	0.6364	13.9993	0.0057	1.9836
	Elastic net regression	14.3696 + 0.0223 (W (G))	0.6363	13.9993	0.0057	1.9836
	Bayesian linear regression	14.7609 + 0.0209 (W (G))	0.6338	13.9993	0.0057	1.9905

Table 5. Regression analysis for hyper-wiener index

Property	Regression model	Regression equation	R^2	F-test	p-value	RMSE
Molar refractivity	Linear regression	41.5968 + 0.0134 (WW (G))	0.5024	8.0768	0.0217	5.8572
	Ridge regression	41.5968 + 0.0134 (WW (G))	0.5024	8.0768	0.0217	5.8572
	Elastic net regression	41.5989 + 0.0134 (WW (G))	0.5024	8.0768	0.0217	5.8572
	Bayesian linear regression	42.7513 + 0.0119 (WW (G))	0.4965	8.0768	0.0217	5.8918
Polarizability	Linear regression	16.4942 + 0.0053 (WW (G))	0.5037	8.1197	0.0215	2.3172
	Ridge regression	16.4942 + 0.0053 (WW (G))	0.5037	8.1197	0.0215	2.3172
	Elastic net regression	16.4962 + 0.0053 (WW (G))	0.5037	8.1197	0.0215	2.3172
	Bayesian linear regression	16.9160 + 0.0048 (WW (G))	0.4987	8.1197	0.0215	2.3289

Table 6. Regression analysis for harary index

Property	Regression model	Regression equation	R^2	F-test	p-value	RMSE
Boiling point	Linear regression	35.1561 + 9.3742 (H (G))	0.5078	8.2531	0.0207	77.6587
	Ridge regression	35.6048 + 9.3610 (H (G))	0.5078	8.2531	0.0207	77.6587
	Elastic net regression	37.6252 + 9.3015 (H (G))	0.5078	8.2531	0.0207	77.6611
	Bayesian linear regression	69.4238 + 8.3646 (H (G))	0.5019	8.2531	0.0207	77.6611
Molar refractivity	Linear regression	25.2297 + 0.7961 (H (G))	0.6509	14.9143	0.0048	4.9061
	Ridge regression	25.2678 + 0.7950 (H (G))	0.6509	14.9143	0.0048	4.9061
	Elastic net regression	25.6572 + 0.7835 (H (G))	0.6507	14.9143	0.0048	4.9072
	Bayesian linear regression	26.8402 + 0.7487 (H (G))	0.6486	14.9143	0.0048	4.9223
Polarizability	Linear regression	10.0104 + 0.3155 (H (G))	0.6515	14.9579	0.0048	1.9417
	Ridge regression	10.0255 + 0.3151 (H (G))	0.6515	14.9579	0.0048	1.9417
	Elastic net regression	10.3235 + 0.3063 (H (G))	0.6510	14.9579	0.0048	1.9433
	Bayesian linear regression	10.6467 + 0.2968 (H (G))	0.6492	14.9579	0.0048	1.9481

Table 7. Regression analysis for detour index

Property	Regression model	Regression equation	R^2	F-test	p-value	RMSE
Boiling point	Linear regression	176.8677 + 0.4304 (D (G))	0.5348	9.1956	0.0162	75.5004
	Ridge regression	176.8682 + 0.4304 (D (G))	0.5348	9.1956	0.0162	75.5004
	Elastic net regression	176.8760 + 0.4304 (D (G))	0.5348	9.1956	0.0162	75.5004
	Bayesian linear regression	193.9230 + 0.3888 (D (G))	0.5298	9.1956	0.0162	75.9047

Table 8. Regression analysis for detour harary index

Property	Regression model	Regression equation	R^2	F-test	p-value	RMSE
Molar refractivity	Linear regression	28.0536 + 0.9643 (DH (G))	0.4936	7.7983	0.0235	5.9086
	Ridge regression	28.1195 + 0.9616 (DH (G))	0.4936	7.7983	0.0235	5.9086
	Elastic net regression	28.7179 + 0.9378 (DH (G))	0.4932	7.7983	0.0235	5.9107
	Bayesian linear regression	30.8119 + 0.8543 (DH (G))	0.4872	7.7983	0.0235	5.9459
Polarizability	Linear regression	11.1220 + 0.3825 (DH (G))	0.4949	7.8387	0.0232	2.3377
	Ridge regression	11.1482 + 0.3814 (DH (G))	0.4949	7.8387	0.0232	2.3377
	Elastic net regression	11.5895 + 0.3639 (DH (G))	0.4937	7.8387	0.0232	2.3404
	Bayesian linear regression	12.2102 + 0.3391 (DH (G))	0.4885	7.8387	0.0232	2.3524
Molar volume	Linear regression	48.3052 + 4.4703 (DH (G))	0.5475	9.6782	0.0144	24.5885
	Ridge regression	48.6108 + 4.4581 (DH (G))	0.5475	9.6782	0.0144	24.5886
	Elastic net regression	50.1551 + 4.3966 (DH (G))	0.5473	9.6782	0.0144	24.5925
	Bayesian linear regression	58.6080 + 4.0597 (DH (G))	0.5428	9.6782	0.0144	24.7136

4. Discussion

In evaluating the performance of various regression models for predicting molecular properties using topological indices, a thorough analysis was conducted to determine their effectiveness and accuracy. The assessment revealed that Harary Index demonstrated the highest predictive capability (R^2 = 0.6509-0.6515) across different properties, while maintaining low RMSE values. This evaluation provides a foundation for understanding which indices and models offer the best performance, setting the stage for a detailed discussion of the specific results.

The Wiener Index showed the strongest correlations with polarizability (0.7977) and molar refractivity (0.7971), indicating its effectiveness in predicting these optical properties. In contrast, the Hyper-Wiener Index demonstrated good correlations with molar refractivity (0.7088) and polarizability (0.7097), but lower correlations with melting point (0.0722) and flash point (0.4176), suggesting limited applicability for these latter properties. The Harary Index exhibited high correlations with boiling point (0.7126), molar refractivity (0.8068), and polarizability (0.8072), showcasing its versatility across different property types. The Detour Index correlated well with molar refractivity (0.6687), polarizability (0.6689), boiling point (0.7313), and enthalpy (0.6561), demonstrating broad applicability. The Detour Harary Index showed strong correlations with molar refractivity (0.7206), molar volume (0.7399), and polarizability (0.7035), indicating its effectiveness for these specific properties. The heatmap in Figure 3 visually represents the correlation matrix, with the color gradient indicating the strength and direction of relationships between the topological indices and physicochemical properties. Darker shades represent stronger correlations, with positive correlations shown in warm colors like red and negative correlations in cool colors like blue. For example, the strong positive correlation between the Wiener Index and molar refractivity is highlighted in dark red, while weaker correlations, such as those between the Hyper-Wiener Index and melting point, are reflected in lighter or cooler colors. This visual aid facilitates the identification of key trends and outliers, enhancing the interpretability of the correlation analysis.

The performance analysis of various regression models demonstrated minimal variations in their predictive capabilities. Linear and Ridge Regression models showed nearly identical performance metrics (e.g., R^2 = 0.6353 for Wiener Index), with Elastic Net following closely. Bayesian linear regression models typically showed slightly lower R^2 values compared to other approaches, particularly for properties like polarizability and molar refractivity.

The significant correlations found with certain topological indices, particularly the Harary Index achieving the highest R^2 values (0.6509-0.6515), emphasize the critical role of index selection in predictive modeling. These findings suggest that in the context of QSPR analysis, the choice of topological index has a greater impact on model performance than the selection of regression method, providing a more robust framework for predicting physicochemical properties.

The predictive models developed in this study show substantial promise for use in the drug discovery and development pipeline, especially in drug screening, property optimization, and precision medicine. The inputs to these models consist of topological indices, which were derived from the hydrogen-depleted molecular graphs of anti-cancer drugs. These indices quantify molecular structure and connectivity and are used to predict key physicochemical properties of the drugs, such as polarizability, molar refractivity, and boiling point. These outputs, the predicted molecular properties, are crucial for early-stage candidate identification, as they are closely related to the drug's biological activity, stability, and interactions with cellular mechanisms.

By accurately predicting these properties, the models facilitate efficient screening processes and minimize the need for extensive experimental validation, thus accelerating the development of more effective anti-cancer therapies. These models can play a pivotal role in drug screening and filtering by aiding in the identification of compounds with desirable optical and electronic properties, such as those indicated by the Harary and Wiener indices. This is particularly advantageous for developing anti-cancer drugs, where specific molecular properties can serve as indicators of effectiveness and safety. By incorporating these models into virtual screening, researchers can streamline the selection process, prioritizing compounds with ideal characteristics, which could potentially enhance the success rates in drug discovery. Furthermore, these models provide valuable insights for optimizing drug properties crucial to stability, solubility, and bioavailability, including parameters like boiling point and molar volume. Researchers can leverage these accurate predictions to enhance pharmacokinetic profiles and optimize formulations, thereby reducing the need for exhaustive experimental procedures and expediting drug development timelines. Finally, these models offer a means

to reduce the experimental load by identifying compounds with the highest potential for synthesis and testing, particularly in budget-constrained environments. This computational approach enables resource prioritization and reduces both the time and costs typically associated with experimental trials.

5. Conclusion

The application of machine learning regression models for predicting molecular properties with topological indices demonstrates significant success, particularly with Linear and Ridge Regression models. These models achieved high R^2 values and low RMSE across properties like polarizability and molar refractivity, demonstrating their effectiveness and suggesting potential applicability in cheminformatics. This study highlights the practical benefits of machine learning in QSPR applications, with implications for fields such as drug discovery and materials science where accurate molecular insights are essential. The superior performance of these models highlights machine learning's capacity to streamline and enhance predictive modeling, which could accelerate drug discovery by offering insights into molecular interactions and optimizing lead compounds. Future studies should work to refine these models further, considering non-linear indices and hybrid methods to enhance precision. Altogether, this work underscores the value of advanced regression techniques for molecular property prediction and promotes broader adoption of machine learning in cheminformatics.

However, the effectiveness of Linear and Ridge Regression models in this study is limited by their reliance on a restricted set of topological indices, which may overlook more complex molecular interactions. These linear models often fail to capture the intricate, non-linear relationships present in chemical data, potentially leading to suboptimal predictions, especially for diverse chemical datasets. Additionally, the performance of these models can vary significantly across different chemical datasets, raising concerns about their generalizability.

To overcome these limitations, future research may benefit from integrating non-linear indices and exploring hybrid modeling techniques. Non-linear indices might provide deeper insights into complex molecular structures, improving predictions for compounds with intricate spatial or electronic properties. Moreover, hybrid models that merge linear and non-linear methods, including machine learning-based solutions like neural networks, could provide a more comprehensive modeling solution. These approaches could enhance the interpretive clarity and accuracy of predictions, advancing their practical utility in drug discovery.

Conflict of interest

The authors declare no competing financial interest.

References

[1] Trichopoulos D, Li FP, Hunter DJ. What causes cancer? *Scientific American*. 1996; 275(3): 80-87.

[2] Dallavalle S, Dobričić V, Lazzarato L, Gazzano E, Machuqueiro M, Pajeva I, et al. Improvement of conventional anti-cancer drugs as new tools against multidrug resistant tumors. *Drug Resistance Updates*. 2020; 50(10): 100682.

[3] Isoldi MC, Visconti MA, Castrucci AMdL. Anti-cancer drugs: Molecular mechanisms of action. *Mini Reviews in Medicinal Chemistry*. 2005; 5(7): 685-695.

[4] Arrondeau J, Gan HK, Razak AR, Paoletti X, Tourneau CL. Development of anti-cancer drugs. *Discovery Medicine*. 2010; 10(53): 355-362.

[5] National Center for Biotechnology Information. *PubChem Compound Summary for CID 2123, Altretamine; 2024*. 2024. Available from: https://pubchem.ncbi.nlm.nih.gov/compound/Altretamine [Accessed 1 July 2024].

[6] National Center for Biotechnology Information. *PubChem Compound Summary for CID 2478, Busulfan; 2024*. 2024. Available from: https://pubchem.ncbi.nlm.nih.gov/compound/Busulfan [Accessed 1 July 2024].

[7] National Center for Biotechnology Information. *PubChem Compound Summary for CID 2578, Carmustine; 2024*. 2024. Available from: https://pubchem.ncbi.nlm.nih.gov/compound/Carmustine [Accessed 1 July 2024].

[8] National Center for Biotechnology Information. *PubChem Compound Summary for CID 2907, Cyclophosphamide; 2024*. 2024. Available from: https://pubchem.ncbi.nlm.nih.gov/compound/Cyclophosphamide [Accessed 1 July 2024].
[9] National Center for Biotechnology Information. *PubChem Compound Summary for CID 135398738, Dacarbazine; 2024*. 2024. Available from: https://pubchem.ncbi.nlm.nih.gov/compound/Dacarbazine [Accessed 1 July 2024].
[10] National Center for Biotechnology Information. *PubChem Compound Summary for CID 3690, Ifosfamide; 2024*. 2024. Available from: https://pubchem.ncbi.nlm.nih.gov/compound/Ifosfamide [Accessed 1 July 2024].
[11] National Center for Biotechnology Information. *PubChem Compound Summary for CID 4033, Mechlorethamine; 2024*. 2024. Available from: https://pubchem.ncbi.nlm.nih.gov/compound/Mechlorethamine [Accessed 1 July 2024].
[12] National Center for Biotechnology Information. *PubChem Compound Summary for CID 4915, Procarbazine; 2024*. 2024. Available from: https://pubchem.ncbi.nlm.nih.gov/compound/Procarbazine [Accessed 1 July 2024].
[13] National Center for Biotechnology Information. *PubChem Compound Summary for CID 5394, Temozolomide; 2024*. 2024. Available from: https://pubchem.ncbi.nlm.nih.gov/compound/Temozolomide [Accessed 1 July 2024].
[14] National Center for Biotechnology Information. *PubChem Compound Summary for CID 5453, Thiotepa; 2024*. 2024. Available from: https://pubchem.ncbi.nlm.nih.gov/compound/Thiotepa [Accessed 1 July 2024].
[15] Gutman I, Polansky OE. *Mathematical Concepts in Organic Chemistry*. New York: Springer Science and Business Media; 2012.
[16] Gutman I. Chemical graph theory-the mathematical connection. *Advances in Quantum Chemistry*. 2006; 51: 125-138. Available from: https://doi.org/10.1016/S0065-3276(06)51003-2.
[17] Huang R, Naeem M, Siddiqui MK, Rauf A, Rashid MU, Ali MA. Statistical analysis of topological indices in linear phenylenes for predicting physicochemical properties using algorithms. *Scientific Reports*. 2024; 14(1): 19282.
[18] Zhang X, Bajwa ZS, Zaman S, Munawar S, Li D. The study of curve fitting models to analyze some degree-based topological indices of certain anti-cancer treatment. *Chemical Papers*. 2024; 78(2): 1055-1068.
[19] Balasubramaniyan D, Chidambaram N, Ravi V. Estimating physico-chemical properties of drugs for prostate cancer using degree-based and neighbourhood degree-based topological descriptors. *Physica Scripta*. 2024; 99(6): 065233.
[20] Sethumadavan N, Durga M. Degree-based topological indices and QSPR analysis of lung cancer Drugs. *Global Journal of Engineering and Technology Advances*. 2024; 19(3): 79-102.
[21] Zabidi ZM, Alias AN, Zakaria NA, Mahmud ZS, Ali R, Yaakob MK, et al. Machine learning predictor models in the electronic properties of alkanes based on degree-topology indices. *International Journal of Emerging Technology and Advanced Engineering*. 2021; 11(22): 1-14.
[22] Yu G, Li X, He D. Topological indices based on 2-or 3-eccentricity to predict anti-HIV activity. *Applied Mathematics and Computation*. 2022; 416(2): 126748.
[23] Dhanajayamurthy B, Shalini GS. Reduced neighborhood degree-based topological indices on anti-cancer drugs with QSPR analysis. *Materials Today: Proceedings*. 2022; 54(03): 608-614.
[24] Shanmukha M, Basavarajappa N, Shilpa K, Usha A. Degree-based topological indices on anticancer drugs with QSPR analysis. *Heliyon*. 2020; 6(6): e04235.
[25] National Center for Biotechnology Information. *PubChem Database; n.d. 2024*. 2024. https://pubchem. ncbi.nlm.nih.gov/ [Accessed 11 November 2024].
[26] Wiener H. Structural determination of paraffin boiling points. *Journal of the American Chemical Society*. 1947; 69(1): 17-20.
[27] Randić M. Novel molecular descriptor for structure-property studies. *Chemical Physics Letters*. 1993; 211(4-5): 478-483.
[28] Plavšić D, Nikolić S, Trinajstić N, Mihalić Z. On the harary index for the characterization of chemical graphs. *Journal of Mathematical Chemistry*. 1993; 12(1): 235-250.
[29] Lukovits I. The detour index. *Croatica Chemica Acta*. 1996; 69(3): 873-882.
[30] Fang W, Liu WH, Liu JB, Chen FY, Hong ZM, Xia ZJ. Maximum detour-harary index for some graph classes. *Symmetry*. 2018; 10(11): 608.