MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **95** (2026) 289–325
ISSN: 0340–6253
doi: 10.46793/match.95-2.18225

Machine Learning-Driven QSPR Analysis for Drug Property Prediction via Topological Indices

Mohammad Javad Nadjafi-Arani[a,b], **Sezer Sorgun**[a,*], **Mahsa Mirzargar**[a,b]

[a] *Department of Mathematics, Nevşehir Hacı Bektaş Veli University, 50300, Nevşehir, TÜRKİYE*

[b] *Faculty of Science, Mahallat Institute of Higher Education, Mahallat, Iran*

mjnajafiarani@gmail.com, srgnrzs@gmail.com, m.mirzargar@mahallat.ac.ir

(Received July 24, 2025)

Abstract

Quantitative Structure Property Relationship (QSPR) modeling is critical for predicting physicochemical properties of molecules, supporting drug discovery and material design. Traditional methods often use complex descriptors or opaque models, which limits interpretability. There is a need for QSPR approaches that balance accuracy and interpretability to elucidate structural influences on molecular properties. We introduce a QSPR method using degree-distance-based topological indices (TIs) derived from vertex edge weighted (VEW) molecular graphs, weighted by atomic number, mass, radius, density, electronegativity, and ionization energy. This approach captures detailed molecular connectivity and bonding while prioritizing interpretability. Using a 166-molecule dataset, our models -Ridge Regression, Random Forest, XGBoost, and Neural

[*] Corresponding author.

Networks- achieved high prediction accuracy for six physicochemical properties. Regularization ensured robust predictions. The performance metrics tables and the TI correlations clarified the structure-property relationships. This efficient and interpretable framework accelerates drug discovery by enabling virtual screening and informed molecular design.

1 Introduction

Studies on finding relationships between the physical properties of molecules and their topological indices are frequently found in the literature. In particular, QSPR analyses explaining such relationships have been extensively studied in drug design research [1, 27, 37]. Quantitative Structure-Property Relationship (QSPR) modeling is a powerful computational approach that correlates molecular structure with physicochemical and biological properties using mathematical descriptors. By transforming complex molecular graphs into numerical indices such as topological, geometric, or electronic descriptors QSPR enables the prediction of key characteristics like solubility, toxicity, and reactivity without extensive lab experimentation. This method accelerates drug discovery and materials science by prioritizing promising compounds before synthesis, reducing costs and time. Graph-theoretical descriptors, in particular, serve as molecular fingerprints, allowing machine learning models to identify structure-activity trends with high accuracy. As a bridge between theoretical chemistry and practical applications, QSPR provides data-driven insights for designing optimized pharmaceuticals and advanced materials, making it an indispensable tool in modern computational chemistry. The first QSPR study on vertex-edge weighted molecular graphs was conducted in [31]. The mentioned studies have utilized degree-based topological indices. In these studies, the topological indices of molecular graphs of drugs have been calculated, and some regression models have generally been used to relate them to physical properties. Some of the recent studies on QSPR can be found in [2, 22, 31, 33, 34].

In recent years, machine learning (ML) techniques have been extensively employed in chemistry to predict physicochemical properties, particularly when experimental data are limited [17, 24]. Studies such as [24] have demonstrated the effectiveness of ML approaches in handling small datasets by leveraging appropriate feature representations. Inspired by these advancements, we applied both linear and non-linear regression models to explore the relationship between physicochemical properties and topological indices. Specifically, linear regression, Lasso, and Ridge regression were utilized to capture simple linear dependencies, while Random Forest, XGBoost, and Neural Networks were employed to model more complex, non-linear interactions.

1.1 Motivation

Topological indices (TIs) are mathematical descriptors derived from molecular graphs, capturing structural features without requiring costly experimental measurements [7]. In contrast, determining physicochemical properties such as Boiling Point (BP), Molar Volume (MV), Molar Refractivity (MR), Flash Point (FP), Polarizability (Polar), and Enthalpy of Vaporization (EV) often involves resource-intensive laboratory procedures. These properties are critical in drug discovery, as they influence pharmacokinetic attributes like solubility (via MV, MR, Polar), stability (via BP, FP), and membrane permeability (via Polar, EV), which are essential for designing bioavailable and effective therapeutics [1]. Establishing reliable relationships between TIs and these properties enables preliminary insights into molecular behavior, optimizing resource utilization in cheminformatics research. While properties like the HOMO-LUMO gap (available in datasets like QM9) or toxicity (available in Tox21) are also relevant for modern drug discovery, the chosen properties provide a foundational understanding of molecular interactions, with potential extensions to quantum and toxicity predictions in future work. This study is motivated by the need for interpretable QSPR models that balance accuracy and structural insight to accelerate drug development.

1.2 Novelty and aim

This study introduces a novel QSPR framework using degree-distance-based TIs derived from vertex-edge weighted (VEW) molecular graphs, weighted by six atomic properties (Atomic Number, Atomic Mass, Atomic Radius, Density, Electronegativity, Ionization Energy). Unlike previous studies focusing on specific drug classes [23, 30, 38], our approach applies TIs to a diverse 166-molecule dataset, identifying universal descriptors (e.g., Harary, RDD) that generalize across multiple physicochemical properties. By integrating linear (Ridge, LASSO) and non-linear (Random Forest, XGBoost, Neural Networks) models with optimized hyperparameters (Tables 14, 15), we achieve high accuracy (e.g., $R^2 > 0.9$ for MR, Polar) while maintaining interpretability through TI correlations, contrasting with less transparent graph neural networks (GNNs) [25, 26]. Future extensions to advanced topological representations, such as simplicial complexes or hypergraphs, could further enhance predictive power. The aim is to develop an efficient, interpretable QSPR methodology that identifies key structural descriptors to guide molecular design in drug discovery, with open-access code and data to facilitate further research (`https://github.com/ssorgun/LNNR`).

This paper is structured as follows: Section 2 introduces the preliminaries and the graph-based molecular representation used throughout the study. Section 3 describes the materials and methods, including the dataset, topological indices (TIs), and machine learning models. Section 4 presents the results of predictive modeling using both linear and non-linear approaches across six physicochemical properties. Section 5 discusses the main findings and methodological limitations, emphasizing the interpretability of topological indices in contrast to graph neural networks. Finally, Section 6 concludes the study and outlines future research directions, including potential integration of TIs with more expressive representations and modern deep learning frameworks.

2 Preliminarily and graph model

A vertex and edge-weighted (VEW) molecular graph $\mathcal{G}$ is firstly defined in [9, 18] as

$$\mathcal{G} = \mathcal{G}(V, E, Sym, Bo, Vw, Ew, w)$$

such that the vertex and edge set is $V = V(\mathcal{G})$ and $E = E(\mathcal{G})$. respectively; Here a set of chemical symbols of the vertices $Sym = Sym(\mathcal{G})$, a set of topological bond orders (takes the value 1 for single bonds, 2 for double bonds, 3 for triple bonds and 1.5 for aromatic bonds) of the edges $Bo = Bo(\mathcal{G})$, a vertex weight set $Vw(w) = V_w(w, \mathcal{G})$, and an edge weight set $Ew(w) = E_w(\mathcal{G})$. Here w is the weighting scheme which is used to compute the $Vw(w)$ and $Ew(w)$. Generally, all schemes in a molecular graph are the properties of the atoms such as atomic number, atomic radius etc. [9].

$$Vw(w)_i = 1 - \frac{w_C}{w_i} \tag{1}$$

and

$$Ew(w)_{ij} = \frac{w_C w_C}{Bo_{ij} w_i w_j} \tag{2}$$

where $Vw(w)_i$ represents atom i from a molecule; $Ew(w)_{ij}$ represents the bonds between atom i and atom j and Bo_{ij} is the topological bonds order between i and j, respectively [18].

The adjacency matrix $A_w = A_w(\mathcal{G})$ of a vertex-edge-weighted molecular graph G with n vertices is the square $n \times n$ real symmetric matrix whose element $(A_w)_{uv}$ and $(D_w)_{uv}$ are defined in [18] (pg. 173-175) as:

$$(A_w)_{uv} = \begin{cases} V_w(w)_u, & \text{if } u = v. \\ E_w(w)_{uv}, & \text{if } uv \in E(G) \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

and

$$(D_w)_{uv} = \begin{cases} V_w(w)_u, & \text{if } u = v. \\ d_w(u, v), & \text{otherwise,} \end{cases} \tag{4}$$

respectively. Here $d_w(u, v)$ represents the distance between vertices u and

v where w denotes the weighting scheme employed to calculate the parameters V_w and E_w. In a VEW graph G, the length of a path p_{ij} between vertices v_i and v_j,

$$l(p_{ij}, w) = l(p_{ij}, w, G),$$

is equal to the sum of the edge parameters $Ew(w)_{ij}$ for all edges along the path.

Topological indices are numerical descriptors derived from graphs, often calculated based on the elements of the graph. These indices frequently depend on properties such as vertex degrees or other structural characteristics, making them valuable tools for analyzing and interpreting molecular structures in various scientific fields. The application of topological indices in drug discovery has been well documented in the literature, with numerous studies highlighting their effectiveness.

Unlike the classical degree-distance-based topological indices, the topological definitions for VEW graphs are derived from equations in (1) and (2) as shown in the table below.

Table 1. VEW-based degree distance topological indices for $\mathcal{G}$

TIs names	**vew-based description**
Wiener Index	$W(\mathcal{G}) = \sum_{u<v} (D_w)_{uv}$
Harary Index	$H(\mathcal{G}) = \frac{1}{2} \sum_{u<v} \frac{1}{(D_w)_{uv}}$
Balaban Index	$J(\mathcal{G}) = \frac{m}{m-n+2} \sum_{uv \in E(\mathcal{G})} \frac{1}{\sqrt{(D_w)_u (D_w)_v}}$
Total Eccentric Index	$TEI(\mathcal{G}) = \sum_{u \in V} \epsilon(u)$
Eccentric Connectivity Index	$ECI(\mathcal{G}) = \sum_{u \in V} \epsilon(u)(A_w^2)_{uu}$
Degree Distance Index	$DD(\mathcal{G}) = \sum_{uv \in E} [(A_w^2)_{uu} + (A_w^2)_{vv}](D_w)_{uv}$
Gutman Index	$G(\mathcal{G}) = \sum_{uv \in E} [(A_w^2)_{uu}(A_w^2)_{vv}](D_w)_{uv}$
Reciprocal Degree Distance Index	$RDD(\mathcal{G}) = \sum_{uv \in E} [(A_w^2)_{uu} + (A_w^2)_{vv}]/(D_w)_{uv}$

In above table, notations of $(D_w)_u$ and $\epsilon(u)$ are the sum of all entries in the uth row of VEW distance matrix of graph $\mathcal{G}$ and the maximum value of the uth row in the $D_w(\mathcal{G})$ matrix, respectively.

For unweighted graphs, the classical definitions (Wiener [10]; Balaban [3]; Harary [28]; Total Eccentric and Eccentric Connectivity index [29];

Degree Distance Index [6]; Gutman Index [11]; Reciprocal Degree Distance Index [16]) of distance and degree-distance-based topological indices, applications in chemistry and related between them can be found in, [8, 12, 13, 19, 21, 32, 33] as listed in Table 1.

3 Material and method

Molecular graphs and physicochemical properties of 166 drug molecules were sourced from the Chemspyder database using SMILES codes and topological indices were calculated using vertex and edge weightings based on atomic properties of the molecules, including atomic radius, atomic mass, density, ionization, electronegativity, and atomic number (see Supplementary Data). While larger datasets are often utilized in QSPR modeling to enhance generalizability, our study employed a carefully curated dataset of 166 drug molecules sourced from Chemspyder. This selection ensures a diverse range of molecular structures and physicochemical properties, critical for validating the effectiveness of our degree-distance-based topological indices. Previous studies, such as those predicting logKoc for persistent organic pollutants with similar dataset sizes Fuzzy QSARs for logKoc, have demonstrated that well-selected smaller datasets can yield robust QSPR models, particularly when focused on specific classes of compounds or properties. Our dataset's diversity, as evidenced by statistical analysis of property variability (e.g., standard deviation in boiling point: 45°C), supports its suitability for this analysis [36].

In the following, briefly we introduce both linear and nonlinear modeling approaches to establish the relationship between six physical properties—Boiling Point (BP), Molar Volume (MV), Molar Refraction (MR), Flash Point (FP), Polarizability (Polar), and Enthalpy of Vaporization (EV)—and eight topological indices mentioned in Table 1. Six atomic properties (Atomic Number, Atomic Mass, Atomic Radius, Density, Electronegativity, Ionization Energy) were used as inputs. Six machine learning models, Linear Regression (LR), Ridge Regression (RR), LASSO Regression, Random Forest (RF), XGBoost (XGB), and Neural Networks (NN), were trained and evaluated using R^2, RMSE, and MAE metrics via 5-fold cross-validation to ensure robustness. Hyperparameters for RF and XGB were optimized via grid search (Tables 14, 15). TI importance was assessed through correlation coefficients for linear models and feature importance scores for RF and XGB, providing interpretable structural insights. This methodology emphasizes computational efficiency and reproducibility, with all code and data publicly available on GitHub at https://github.com/ssorgun/LNNR. The results were validated through cross-validation to ensure robustness, and comparisons with graph neural networks (GNNs) demonstrated a strong balance between the interpretability of topological indices (TIs) and predictive performance [25,26].

Linear Regression assumes a linear relationship between indices and properties [14]. Lasso Regression uses L_1 regularization for feature selection, reducing complexity [35]. Ridge Regression applies L_2 regularization to mitigate multicollinearity [15]. Beyond linear models, we applied nonlinear techniques to capture complex relationships. Random Forest (RF) aggregates decision trees for improved generalization [4]. XGBoost sequentially builds trees to correct errors, enhancing accuracy [5]. Neural Networks (ANNs) uses layered neurons to model intricate patterns, with careful tuning to avoid overfitting on smaller datasets [20]. Models were assessed using R^2 (explanatory power), RMSE (error magnitude), and MAE (average accuracy), selected for their complementary evaluation of performance [14].

The TIs in Table 1 were chosen for their proven ability to encode degree-distance structural information in weighted graphs, drawing from established QSPR literature [7,29]. High correlations among certain indices were managed through Ridge regularization to prevent overfitting and enhance model stability, as evidenced by different tables in the results.

To address the variability of topological indices across molecular graphs of different orders, all TI values and target properties were standardized using the ‘StandardScaler‘ from scikit-learn, transforming them to a mean of 0 and standard deviation of 1. This preprocessing step, combined with VEW graph weighting by atomic properties, effectively balances the impact of molecular size and complexity.

The 166-molecule dataset was divided into training (80%,133 molecules) and test ((20%,33 molecules) sets using stratified sampling with a fixed random seed (42) to ensure reproducibility and maintain property diversity. The models were trained on the training set and evaluated on the test set using R^2, MAE, and RMSE metrics.

4 Results

This section analyzes the predictive performance of machine learning models: linear regression (LR), ridge regression (RR), LASSO regression, random forest (RF), XGBoost (XGB), and neural networks (NN) for six physicochemical properties: Boiling point (BP), molecular volume (MV), molecular refractivity (MR), flash point (FP), polarizability (polar), and enthalpy of vaporization (EV). For each property, two tables present prediction metrics (R^2, RMSE, MAE) across atomic properties (Atomic Number, Atomic Mass, Atomic Radius, Density, Electronegativity, Ionization Energy) and the importance of topological indices (TIs) such as Harary, RDD, TEI, ECI, Wiener, DD, Gutman, and Balaban. Additionally, hyperparameter optimization results for RF and XGB models are provided to contextualize model performance. The analysis is organized by property, followed by a cross-property comparison to guide Quantitative Structure-Property Relationship (QSPR) modeling for drug design.

4.1 Boiling point (BP)

4.1.1 Prediction performance

As shown in Table 2, both linear and non-linear models exhibit limited predictive performance for the boiling point (BP). Among linear models, R^2 values range from 0.4545 (LASSO with Atomic Radius) to a maximum of 0.4999 (Linear Regression with Electronegativity), indicating modest predictive power. Non-linear models offer only marginal improvements: the Random Forest (RF) model achieves $\mathrm{R}^2 = 0.51$, RMSE = 108.49, and MAE = 58.20 using Atomic Number, while XGBoost (XGB) follows closely with $\mathrm{R}^2 = 0.47$, RMSE = 113.04, and MAE = 61.92. These results suggest that, despite optimized hyperparameters (e.g., `max_depth` = None and `n_estimators` = 200 for RF; `max_depth` = 7 and `n_estimators` = 50 for XGB; see Tables 14 and 15), non-linear models do not substantially outperform their linear counterparts for BP prediction. Among the atomic features evaluated, Ionization Energy appears to be the most informative, possibly due to its relevance in governing intermolecular interactions. Neural Networks (NN) achieve slightly higher performance (R^2 = 0.613), though their results may be constrained by dataset size and inherent complexity of the BP property.

Table 2. Performance metrics and influential TIs for boiling point (BP) prediction

Atomic Property	**Model**	**R^2**	**RMSE**	**MAE**
Atomic Number	LR	0.4792	115.4367	69.6274
	RR	0.4792	115.4367	69.6274
	LASSO	0.4792	115.4312	69.6230
	RF	0.51	108.49	58.20
	XGB	0.42	117.98	70.26
	NN	0.613	92.860	70.767
Atomic Mass	LR	0.4798	115.3330	69.2126
	RR	0.4798	115.3330	69.2126
	LASSO	0.4799	115.3273	69.2079
	RF	0.51	108.79	58.25
	XGB	0.47	113.04	61.92
	NN	0.624	91.070	66.983
Atomic Radius	LR	0.4558	117.0522	75.5980
	RR	0.4558	117.0522	75.5980
	LASSO	0.4545	117.1947	75.7172
	RF	0.41	118.50	65.55
	XGB	0.39	120.71	70.92
	NN	0.596	95.114	71.516
Density	LR	0.4875	113.8972	68.3838
	RR	0.4875	113.8972	68.3838
	LASSO	0.4875	113.8971	68.3839

Continued on next page

Table 2 (continued)

Atomic Property	**Model**	**R²**	**RMSE**	**MAE**
	RF	0.52	107.28	58.90
	XGB	0.45	115.19	61.80
	NN	0.672	86.019	63.921
Electronegativity	LR	0.4999	113.5759	68.2110
	RR	0.4999	113.5759	68.2110
	LASSO	0.4980	113.7959	68.1240
	RF	0.38	121.75	70.82
	XGB	0.34	126.07	73.23
	NN	0.630	91.864	68.577
Ionization Energy	LR	0.4830	114.4173	66.6389
	RR	0.4830	114.4173	66.6390
	LASSO	0.4816	114.5740	67.1658
	RF	0.38	122.38	71.06
	XGB	0.37	123.02	69.83
	NN	0.622	91.698	67.655

4.1.2 Importance of topological indices

Table 3 highlights the dominance of Harary and RDD indices across linear models, with correlation coefficients near 0.82 and 0.81 respectively for Atomic Number, reflecting their ability to capture molecular connectivity relevant to BP. RF and XGB emphasize Harary and Wiener indices, albeit with slightly different importance scores, underscoring the complementary nature of these descriptors in non-linear modeling. The consistent importance of Gutman and DD indices in non-linear models further suggests complex graph-theoretical features underpin BP prediction.

Table 3. Top topological indices (TIs) for boiling point (BP) prediction across atomic properties and models

Atomic Property	**Correlation-Based (LR/LASSO/RR)**	**Random Forest (RF)**	**XGBoost (XGB)**
Atomic Number	Harary: 0.8183 RDD: 0.8054 TEI: 0.7655 ECI: 0.7532 Wiener: 0.7454	Harary: 0.5834 Wiener: 0.1138 RDD: 0.0738 Gutman: 0.0674 Balaban: 0.0540	Harary: 0.5749 RDD: 0.1287 Wiener: 0.1273 DD: 0.0464 Balaban: 0.0463
Atomic Mass	Harary: 0.8183 RDD: 0.8057 TEI: 0.7656 ECI: 0.7535 Wiener: 0.7453	Harary: 0.5512 Wiener: 0.1161 RDD: 0.0958 Gutman: 0.0656 Balaban: 0.0554	Harary: 0.5258 RDD: 0.1558 Wiener: 0.1391 DD: 0.0573 Balaban: 0.0527
Atomic Radius	RDD: 0.8090 Harary: 0.8041 TEI: 0.7602 ECI: 0.7527 DD: 0.7506	Harary: 0.3140 RDD: 0.2389 Wiener: 0.1870 Gutman: 0.0613 DD: 0.0572	Harary: 0.3629 RDD: 0.1757 Wiener: 0.1679 DD: 0.0974 ECI: 0.0780
Density	RDD: 0.8388 Harary: 0.7722 DD: 0.7562 Wiener: 0.7529 TEI: 0.7526	RDD: 0.5262 DD: 0.1130 Wiener: 0.1038 Harary: 0.0829 Gutman: 0.0629	RDD: 0.3603 Gutman: 0.1750 Wiener: 0.1338 DD: 0.1191 Harary: 0.0752
Electronegativity	Harary: 0.8147 RDD: 0.7980 TEI: 0.7641 ECI: 0.7455 Wiener: 0.7408	Harary: 0.5523 Wiener: 0.1089 RDD: 0.0739 DD: 0.0698 Gutman: 0.0687	Harary: 0.4370 Wiener: 0.1950 Gutman: 0.1248 DD: 0.0871 TEI: 0.0512
Ionization Energy	Harary: 0.8170 RDD: 0.7967 TEI: 0.7646 ECI: 0.7496 Wiener: 0.7381	Harary: 0.5112 Wiener: 0.1361 DD: 0.1039 RDD: 0.0856 Balaban: 0.0673	Harary: 0.4239 Wiener: 0.1813 RDD: 0.1070 DD: 0.0913 Gutman: 0.0640

4.2 Molar volume (MV)

4.2.1 Prediction performance

Table 4 reveals strong predictive accuracy for MV, with linear models achieving R^2 above 0.84 and RF and XGB excelling (R^2 up to 0.93). RF's n_estimators = 200 and XGB' s learning_rate = 0.3 (Tables 14 and 15) were critical in capturing the molecular volume complexity. Density in linear and Electronegativity in non-linear models consistently emerge as the top atomic properties, highlighting its fundamental role in molecular packing and volume. NN performance is respectable but slightly lower, possibly due to overfitting risks in limited data contexts.

Table 4. Performance metrics and influential TIs for molar volume (MV) prediction

Atomic Property	**Model**	**R^2**	**RMSE**	**MAE**
Atomic Number	LR	0.8455	33.0305	28.8814
	RR	0.8455	33.0305	28.8814
	LASSO	0.8457	33.0104	28.8597
	RF	0.90	25.96	19.84
	XGB	0.92	23.99	18.32
	NN	0.848	41.541	32.925
Atomic Mass	LR	0.8455	33.0297	28.8454
	RR	0.8455	33.0297	28.8454
	LASSO	0.8457	33.0100	28.8243
	RF	0.91	25.77	20.07
	XGB	0.91	25.25	19.58
	NN	0.875	37.108	29.766
Atomic Radius	LR	0.8451	33.0754	26.9180
	RR	0.8451	33.0754	26.9180
	LASSO	0.8471	32.8609	26.7316
	RF	0.90	26.63	21.00
	XGB	0.90	27.21	20.49
	NN	0.849	39.675	32.007

Continued on next page

Table 4 (continued)

Atomic Property	**Model**	**R²**	**RMSE**	**MAE**
Density	LR	0.8869	28.2627	23.9887
	RR	0.8869	28.2627	23.9887
	LASSO	0.8869	28.2628	23.9891
	RF	0.85	32.46	23.97
	XGB	0.86	31.98	23.08
	NN	0.879	35.811	28.928
Electronegativity	LR	0.8627	31.1373	26.5694
	RR	0.8627	31.1374	26.5694
	LASSO	0.8612	31.3019	26.5889
	RF	0.91	25.69	20.06
	XGB	0.93	21.99	16.56
	NN	0.874	37.259	28.970
Ionization Energy	LR	0.8656	30.8072	26.8388
	RR	0.8656	30.8072	26.8388
	LASSO	0.8658	30.7855	26.8080
	RF	0.89	28.23	21.52
	XGB	0.91	25.45	19.35
	NN	0.879	35.938	29.061

4.2.2 Importance of topological indices

From Table 5, ECI and TEI dominate linear models with correlations above 0.92 for Atomic Number, confirming their relevance to volume-related properties. RF and XGB prioritize Gutman and DD indices, reflecting their sensitivity to subtle structural variations impacting molecular volume. The agreement across atomic properties reinforces these indices' utility.

Table 5. Top topological indices (TIs) for molar volume (MV) prediction across atomic properties and models

Atomic Property	**Correlation-Based (LR/LASSO/RR)**	**Random Forest (RF)**	**XGBoost (XGB)**
Atomic Number	ECI: 0.9273 TEI: 0.9259 RDD: 0.9161 Harary: 0.9058 DD: 0.9007	Gutman: 0.2971 DD: 0.1790 Wiener: 0.1398 Harary: 0.1114 RDD: 0.0958	Gutman: 0.5301 DD: 0.1674 RDD: 0.1596 Wiener: 0.0448 TEI: 0.0409
Atomic Mass	ECI: 0.9273 TEI: 0.9259 RDD: 0.9162 Harary: 0.9059 DD: 0.9005	Gutman: 0.3003 DD: 0.1629 Wiener: 0.1492 Harary: 0.1086 RDD: 0.0963	Gutman: 0.2907 DD: 0.2489 Wiener: 0.1731 RDD: 0.1164 TEI: 0.1138
Atomic Radius	Harary: 0.9175 TEI: 0.9136 RDD: 0.9067 Wiener: 0.8914 ECI: 0.8911	Harary: 0.2783 Wiener: 0.2636 DD: 0.1759 TEI: 0.1230 RDD: 0.0669	DD: 0.2392 TEI: 0.2153 Harary: 0.2095 Wiener: 0.1468 RDD: 0.0889
Density	Harary: 0.9115 Wiener: 0.8883 TEI: 0.8796 RDD: 0.8763 DD: 0.8339	Harary: 0.4160 Wiener: 0.3468 TEI: 0.1069 DD: 0.0738 RDD: 0.0247	Harary: 0.3396 Wiener: 0.2447 TEI: 0.1979 DD: 0.1471 RDD: 0.0332
Electronegativity	ECI: 0.9332 TEI: 0.9281 RDD: 0.9173 Gutman: 0.9074 Harary: 0.9062	Gutman: 0.3444 DD: 0.2577 Wiener: 0.1005 ECI: 0.0877 RDD: 0.0691	Gutman: 0.6600 ECI: 0.1136 Wiener: 0.0555 RDD: 0.0472 TEI: 0.0413

Continued on next page

Table 5 (continued)

Atomic Property	Correlation-Based (LR/LASSO/RR)	Random Forest (RF)	XGBoost (XGB)
Ionization Energy	ECI: 0.9364 TEI: 0.9283 RDD: 0.9165 Harary: 0.9105 DD: 0.8983	Gutman: 0.3313 DD: 0.2936 ECI: 0.0974 TEI: 0.0833 Harary: 0.0803	DD: 0.3324 Gutman: 0.1989 ECI: 0.1407 RDD: 0.1109 Wiener: 0.1011

4.3 Molar refractivity (MR)

4.3.1 Prediction performance

According to Table 6, MR exhibits excellent predictability, with R^2 exceeding 0.94 for most models. RF and XGB maintain high accuracy ($R^2 \approx 0.96$), benefiting from tuned hyperparameters (Tables 14 and 15). Electronegativity again ranks highest among atomic properties, consistent with MR's dependence on molecular polarizability and size. NN shows competitive but slightly lower performance.

Table 6. Performance metrics and influential TIs for molar refraction (MR) prediction

Atomic Property	Model	R^2	RMSE	MAE
Atomic Number	LR	0.9541	6.7408	5.6431
	RR	0.9541	6.7408	5.6431
	LASSO	0.9541	6.7352	5.6451
	RF	0.95	6.76	4.99
	XGB	0.95	6.98	5.11
	NN	0.927	9.738	7.389
Atomic Mass	LR	0.9543	6.7245	5.6100
	RR	0.9543	6.7245	5.6100
	LASSO	0.9544	6.7183	5.6122

Continued on next page

Table 6 (continued)

Atomic Property	Model	R^2	RMSE	MAE
	RF	0.95	6.79	4.98
	XGB	0.95	7.04	5.37
	NN	0.951	8.012	6.311
Atomic Radius	LR	0.9559	6.6053	5.4746
	RR	0.9559	6.6053	5.4746
	LASSO	0.9551	6.6664	5.4967
	RF	0.95	7.36	5.27
	XGB	0.94	7.52	5.46
	NN	0.928	9.418	7.258
Density	LR	0.9460	7.3057	6.0635
	RR	0.9460	7.3057	6.0635
	LASSO	0.9460	7.3060	6.0636
	RF	0.93	8.32	6.12
	XGB	0.94	7.53	5.51
	NN	0.934	9.196	7.256
Electronegativity	LR	0.9572	6.5094	5.4507
	RR	0.9572	6.5094	5.4507
	LASSO	0.9571	6.5120	5.4576
	RF	0.97	5.55	4.25
	XGB	0.97	5.21	4.09
	NN	0.954	7.897	6.261
Ionization Energy	LR	0.9557	6.6212	5.4401
	RR	0.9557	6.6212	5.4401
	LASSO	0.9560	6.5985	5.4762
	RF	0.96	5.96	4.69
	XGB	0.97	5.79	4.41
	NN	0.945	8.427	6.701

4.3.2 Importance of topological indices

Table 7 highlights RDD and ECI as key indices, with correlations exceeding 0.95 for linear models. RF and XGB models stress Gutman and DD, emphasizing their role in capturing the nuanced electronic environment influencing MR.

Table 7. Top topological indices (TIs) for molar refraction (MR) prediction across atomic properties and models

Atomic Property	**Correlation-Based (LR/LASSO/RR)**	**Random Forest (RF)**	**XGBoost (XGB)**
Atomic Number	RDD: 0.9514 ECI: 0.9514 TEI: 0.9468 Harary: 0.9385 Gutman: 0.9172	Gutman: 0.2847 DD: 0.2058 Wiener: 0.1913 RDD: 0.1764 Harary: 0.0521	DD: 0.3434 Gutman: 0.3171 Wiener: 0.1520 RDD: 0.1505 Harary: 0.0251
Atomic Mass	RDD: 0.9516 ECI: 0.9515 TEI: 0.9468 Harary: 0.9386 Gutman: 0.9171	Gutman: 0.2926 DD: 0.1979 Wiener: 0.1940 RDD: 0.1899 TEI: 0.0520	DD: 0.3682 Gutman: 0.2560 Wiener: 0.2075 RDD: 0.1060 Harary: 0.0296
Atomic Radius	Harary: 0.9481 RDD: 0.9425 TEI: 0.9371 ECI: 0.9190 DD: 0.9060	Wiener: 0.4249 Harary: 0.2744 DD: 0.1122 RDD: 0.0659 Gutman: 0.0565	DD: 0.4225 Wiener: 0.3264 TEI: 0.1084 Harary: 0.1075 RDD: 0.0181
Density	RDD: 0.9328 Harary: 0.9314 TEI: 0.9133 Wiener: 0.9085 DD: 0.8764	Wiener: 0.5266 Harary: 0.2805 TEI: 0.0674 RDD: 0.0629 DD: 0.0465	Wiener: 0.3254 TEI: 0.3092 DD: 0.1599 Harary: 0.1278 RDD: 0.0579
Electronegativity	ECI: 0.9552 RDD: 0.9509 TEI: 0.9485 Harary: 0.9393 Gutman: 0.9228	DD: 0.4022 Gutman: 0.2990 Wiener: 0.0946 RDD: 0.0823 ECI: 0.0579	DD: 0.4650 Gutman: 0.2845 Wiener: 0.1347 RDD: 0.0818 ECI: 0.0140

Continued on next page

Table 7 (continued)

Atomic Property	Correlation-Based (LR/LASSO/RR)	Random Forest (RF)	XGBoost (XGB)
Ionization Energy	ECI: 0.9536 RDD: 0.9489 TEI: 0.9470 Harary: 0.9462 DD: 0.9085	Gutman: 0.3235 DD: 0.3180 Wiener: 0.1337 RDD: 0.0971 ECI: 0.0493	DD: 0.3874 Gutman: 0.3511 Wiener: 0.0964 RDD: 0.0505 Harary: 0.0444

4.4 Flash point (FP)

4.4.1 Prediction performance

Table 8 indicates moderate predictive performance for linear models, with R^2 near 0.71 and slightly lower for ensemble models. Optimized hyperparameters such as n_estimators = 50 (RF) and learning_rate = 0.1 (XGB) facilitated modeling of this more complex property. Density's predictive strength persists, suggesting molecular packing influences volatility. NN models achieve moderate results, highlighting FP's challenging prediction landscape.

Table 8. Performance metrics and influential TIs for flash point (FP) prediction

Atomic Property	Model	R^2	RMSE	MAE
Atomic Number	LR	0.7048	48.3148	37.3700
	RR	0.7048	48.3148	37.3700
	LASSO	0.7048	48.3181	37.3702
	RF	0.68	48.21	32.85
	XGB	0.64	51.24	33.87
	NN	0.618	56.570	43.494
Atomic Mass	LR	0.7044	48.3123	37.1798
	RR	0.7044	48.3124	37.1798

Continued on next page

Table 8 (continued)

Atomic Property	Model	R^2	RMSE	MAE
	LASSO	0.7044	48.3152	37.1800
	RF	0.67	48.88	33.26
	XGB	0.64	51.26	36.12
	NN	0.607	57.352	43.582
Atomic Radius	LR	0.6453	53.2214	42.0346
	RR	0.6453	53.2214	42.0346
	LASSO	0.6455	53.2043	42.0268
	RF	0.61	53.22	37.30
	XGB	0.51	59.59	41.38
	NN	0.591	58.424	45.077
Density	LR	0.7131	47.5233	37.4076
	RR	0.7131	47.5233	37.4076
	LASSO	0.7167	47.2273	36.8919
	RF	0.65	50.21	34.18
	XGB	0.62	52.24	35.14
	NN	0.681	51.603	39.681
Electronegativity	LR	0.7134	47.6816	36.8679
	RR	0.7134	47.6816	36.8679
	LASSO	0.7099	47.9670	36.7952
	RF	0.54	57.68	40.18
	XGB	0.51	59.91	41.73
	NN	0.610	56.881	43.735
Ionization Energy	LR	0.7146	46.9982	35.0963
	RR	0.7146	46.9983	35.0963
	LASSO	0.7115	47.2478	35.6502
	RF	0.54	57.67	40.29
	XGB	0.50	60.34	45.87
	NN	0.611	56.499	42.776

4.4.2 Importance of topological indices

As seen in Table 9, Harary and RDD remain dominant for linear methods. RF and XGB highlight Gutman and Harary, indicating their ability to encode molecular features affecting FP. The recurrent importance of Harary across properties supports its robustness.

Table 9. Top topological indices (TIs) for flash point (FP) prediction across atomic properties and models

Atomic Property	**Correlation-Based (LR/LASSO/RR)**	**Random Forest (RF)**	**XGBoost (XGB)**
Atomic Number	Harary: 0.8101 RDD: 0.7903 TEI: 0.7694 ECI: 0.7544 Wiener: 0.7458	Harary: 0.4885 Gutman: 0.1532 RDD: 0.0910 Wiener: 0.0827 Balaban: 0.0642	Harary: 0.3400 Gutman: 0.2033 Wiener: 0.1533 TEI: 0.0750 RDD: 0.0689
Atomic Mass	Harary: 0.8102 RDD: 0.7907 TEI: 0.7694 ECI: 0.7547 Wiener: 0.7457	Harary: 0.4768 Gutman: 0.1536 RDD: 0.0982 Wiener: 0.0833 Balaban: 0.0671	Harary: 0.3569 Gutman: 0.2213 TEI: 0.0899 RDD: 0.0801 Wiener: 0.0780
Atomic Radius	RDD: 0.7917 Harary: 0.7905 TEI: 0.7754 ECI: 0.7641 DD: 0.7518	Harary: 0.2513 RDD: 0.2249 Wiener: 0.1865 ECI: 0.0827 Gutman: 0.0747	Harary: 0.2995 RDD: 0.2661 ECI: 0.1029 Wiener: 0.0867 DD: 0.0695
Density	RDD: 0.8365 DD: 0.7791 TEI: 0.7764 Wiener: 0.7636 Harary: 0.7469	RDD: 0.3216 Gutman: 0.2349 Wiener: 0.1168 Harary: 0.0818 DD: 0.0795	Gutman: 0.3313 RDD: 0.2566 Wiener: 0.1079 DD: 0.0894 Harary: 0.0791

Continued on next page

Table 9 (continued)

Atomic Property	**Correlation-Based (LR/LASSO/RR)**	**Random Forest (RF)**	**XGBoost (XGB)**
Electronegativity	Harary: 0.8053 RDD: 0.7860 TEI: 0.7704 ECI: 0.7534 Wiener: 0.7446	Harary: 0.5229 Gutman: 0.1459 Wiener: 0.0843 RDD: 0.0711 Balaban: 0.0591	Harary: 0.3897 Gutman: 0.1785 Wiener: 0.1348 TEI: 0.0796 DD: 0.0669
Ionization Energy	Harary: 0.8103 RDD: 0.7841 TEI: 0.7679 ECI: 0.7518 Wiener: 0.7410	Harary: 0.4952 DD: 0.1037 Wiener: 0.1032 RDD: 0.0876 Gutman: 0.0725	Harary: 0.6456 ECI: 0.1191 Balaban: 0.0583 RDD: 0.0447 Wiener: 0.0407

4.5 Polarizability (polar)

4.5.1 Prediction performance

Table 10 reports high accuracy for Polar, with RF and XGB achieving R^2 values around 0.94. Consistent hyperparameter optimization, including n_estimators = 200 and learning_rate = 0.3, played a significant role. Electronegativity and Ionization Energy stands out, reflecting polarizability's strong dependence on molecular size and electronic environment. NN results are comparable, underscoring the potential of deep learning in this domain.

Table 10. Performance metrics and influential TIs for polarizability (polar) prediction

Atomic Property	**Model**	**R^2**	**RMSE**	**MAE**
Atomic Number	LR	0.9262	3.3253	2.4715
	RR	0.9262	3.3253	2.4715
	LASSO	0.9264	3.3196	2.4707

Continued on next page

Table 10 (continued)

Atomic Property	Model	R^2	RMSE	MAE
	RF	0.93	3.22	2.19
	XGB	0.93	3.17	2.15
	NN	0.919	3.959	3.036
Atomic Mass	LR	0.9259	3.3317	2.4631
	RR	0.9259	3.3317	2.4631
	LASSO	0.9262	3.3258	2.4624
	RF	0.93	3.24	2.18
	XGB	0.93	3.13	2.25
	NN	0.941	3.409	2.597
Atomic Radius	LR	0.9257	3.3362	2.4244
	RR	0.9257	3.3362	2.4244
	LASSO	0.9245	3.3638	2.4367
	RF	0.93	3.35	2.24
	XGB	0.92	3.54	2.39
	NN	0.923	3.966	3.068
Density	LR	0.9219	3.4195	2.6004
	RR	0.9219	3.4195	2.6004
	LASSO	0.9219	3.4197	2.6005
	RF	0.91	3.58	2.48
	XGB	0.92	3.52	2.44
	NN	0.921	4.020	3.167
Electronegativity	LR	0.9286	3.2706	2.4024
	RR	0.9286	3.2706	2.4024
	LASSO	0.9286	3.2709	2.4046
	RF	0.94	2.96	1.89
	XGB	0.95	2.68	1.66
	NN	0.945	3.275	2.560
Ionization Energy	LR	0.9289	3.2639	2.3817
	RR	0.9289	3.2639	2.3817
	LASSO	0.9293	3.2530	2.3938
	RF	0.94	3.12	2.09

Continued on next page

Table 10 (continued)

Atomic Property	Model	R^2	RMSE	MAE
	XGB	0.94	3.02	1.96
	NN	0.932	3.719	2.842

4.5.2 Importance of topological indices

From Table 11, ECI and RDD show the highest correlations for linear models, while RF and XGB assign considerable importance to Gutman and DD indices. This alignment across models and properties underscores these descriptors' effectiveness for Polar prediction.

Table 11. Top topological indices (TIs) for polarizability (polar) prediction across atomic properties and models

Atomic Property	Correlation-Based (LR/LASSO/RR)	Random Forest (RF)	XGBoost (XGB)
Atomic Number	ECI: 0.9502 RDD: 0.9499 TEI: 0.9458 Harary: 0.9369 Gutman: 0.9166	Gutman: 0.2852 DD: 0.1997 Wiener: 0.1930 RDD: 0.1756 Harary: 0.0506	DD: 0.3354 Gutman: 0.3273 RDD: 0.1776 Wiener: 0.1146 Harary: 0.0346
Atomic Mass	ECI: 0.9503 RDD: 0.9501 TEI: 0.9458 Harary: 0.9370 Gutman: 0.9165	Gutman: 0.2893 Wiener: 0.2012 DD: 0.1991 RDD: 0.1881 TEI: 0.0496	DD: 0.3806 Gutman: 0.2279 Wiener: 0.2112 RDD: 0.1119 Harary: 0.0408
Atomic Radius	Harary: 0.9466 RDD: 0.9415 TEI: 0.9360 ECI: 0.9186 DD: 0.9057	Wiener: 0.4245 Harary: 0.2940 DD: 0.1053 RDD: 0.0681 Gutman: 0.0453	DD: 0.4384 Wiener: 0.2982 TEI: 0.1272 Harary: 0.1056 RDD: 0.0177

Continued on next page

Table 11 (continued)

Atomic Property	**Correlation-Based (LR/LASSO/RR)**	**Random Forest (RF)**	**XGBoost (XGB)**
Density	RDD: 0.9315 Harary: 0.9303 TEI: 0.9120 Wiener: 0.9079 DD: 0.8756	Wiener: 0.5472 Harary: 0.2561 DD: 0.0654 RDD: 0.0646 TEI: 0.0522	Wiener: 0.3585 TEI: 0.2841 DD: 0.1368 Harary: 0.1258 RDD: 0.0711
Electronegativity	ECI: 0.9538 RDD: 0.9495 TEI: 0.9474 Harary: 0.9382 Gutman: 0.9220	DD: 0.4096 Gutman: 0.3051 Wiener: 0.1095 RDD: 0.0756 ECI: 0.0453	DD: 0.4178 Gutman: 0.4112 Wiener: 0.0845 RDD: 0.0614 ECI: 0.0108
Ionization Energy	ECI: 0.9522 RDD: 0.9474 TEI: 0.9458 Harary: 0.9450 DD: 0.9080	DD: 0.3444 Gutman: 0.3077 Wiener: 0.1293 RDD: 0.0989 Harary: 0.0474	DD: 0.4133 Gutman: 0.2798 Wiener: 0.1264 TEI: 0.0637 RDD: 0.0444

4.6 Enthalpy of vaporization (EV)

4.6.1 Prediction performance

Table 12 demonstrates moderate prediction results for EV, with R^2 around 0.72 for RF and slightly lower for XGB. Hyperparameters such as n_estimators = 50 for RF and learning_rate = 0.1 for XGB (Tables 14 and 15) were critical. Atomic Mass obtains a little better results than other atomic properties in predictability, possibly reflecting electron cloud effects. NN yields moderate results, indicating room for model improvement.

Table 12. Performance metrics and influential TIs for enthalpy of vaporization (EV) prediction

Atomic Property	**Model**	**R^2**	**RMSE**	**MAE**
Atomic Number	LR	0.7226	9.5651	7.6559
	RR	0.7226	9.5651	7.6559
	LASSO	0.7248	9.5270	7.6282
	RF	0.72	9.23	6.27
	XGB	0.67	10.07	6.95
	NN	0.674	11.574	8.895
Atomic Mass	LR	0.7229	9.5567	7.5519
	RR	0.7229	9.5567	7.5519
	LASSO	0.7250	9.5198	7.5296
	RF	0.72	9.27	6.22
	XGB	0.71	9.47	6.31
	NN	0.636	11.978	9.213
Atomic Radius	LR	0.6743	10.4026	8.1233
	RR	0.6743	10.4026	8.1233
	LASSO	0.6770	10.3587	8.1050
	RF	0.61	10.94	7.79
	XGB	0.62	10.84	7.94
	NN	0.668	11.657	9.022
Density	LR	0.7275	9.7586	7.6991
	RR	0.7275	9.7586	7.6991
	LASSO	0.7275	9.7588	7.6991
	RF	0.64	10.51	6.78
	XGB	0.58	11.34	7.06
	NN	0.697	11.304	8.777
Electronegativity	LR	0.7142	9.7818	7.9742
	RR	0.7142	9.7818	7.9742
	LASSO	0.7136	9.7925	7.9903
	RF	0.53	11.98	8.17
	XGB	0.50	12.42	8.59
	NN	0.667	11.578	8.681

Continued on next page

Table 12 (continued)

Atomic Property	Model	R^2	RMSE	MAE
Ionization Energy	LR	0.7110	9.6739	7.6957
	RR	0.7110	9.6739	7.6957
	LASSO	0.7121	9.6552	7.6743
	RF	0.54	11.93	7.82
	XGB	0.51	12.21	8.19
	NN	0.661	11.692	8.664

4.6.2 Importance of topological indices

Table 13 confirms Harary and RDD as the most influential TIs across linear and ensemble models, highlighting their capability to capture essential structural features influencing Enthalpy of Vaporization.

Table 13. Top topological indices (TIs) for enthalpy of vaporization (EV) prediction across atomic properties and models

Atomic Property	Correlation-Based (LR/LASSO/RR)	Random Forest (RF)	XGBoost (XGB)
Atomic Number	Harary: 0.8524 RDD: 0.8356 Wiener: 0.7791 TEI: 0.7742 DD: 0.7719	Harary: 0.5856 RDD: 0.1537 Balaban: 0.0859 Wiener: 0.0608 DD: 0.0325	RDD: 0.4301 Harary: 0.3440 Balaban: 0.0562 Wiener: 0.0486 Gutman: 0.0434
Atomic Mass	Harary: 0.8524 RDD: 0.8358 Wiener: 0.7790 TEI: 0.7741 DD: 0.7718	Harary: 0.5248 RDD: 0.1998 Balaban: 0.0901 Wiener: 0.0624 Gutman: 0.0392	Harary: 0.4966 RDD: 0.1673 DD: 0.0859 Balaban: 0.0693 Wiener: 0.0584

Continued on next page

Table 13 (continued)

Atomic Property	**Correlation-Based (LR/LASSO/RR)**	**Random Forest (RF)**	**XGBoost (XGB)**
Atomic Radius	RDD: 0.8391 Harary: 0.8371 DD: 0.7824 Wiener: 0.7815 TEI: 0.7802	RDD: 0.3291 Harary: 0.3096 ECI: 0.0911 Wiener: 0.0699 Balaban: 0.0587	Harary: 0.3250 RDD: 0.2313 Gutman: 0.1067 ECI: 0.1046 Wiener: 0.0744
Density	RDD: 0.8555 Harary: 0.8060 Wiener: 0.7832 DD: 0.7749 TEI: 0.7692	RDD: 0.4782 Harary: 0.2669 TEI: 0.0600 Wiener: 0.0564 Balaban: 0.0477	RDD: 0.3603 Harary: 0.3085 DD: 0.0932 TEI: 0.0747 Gutman: 0.0486
Electronegativity	Harary: 0.8504 RDD: 0.8307 Wiener: 0.7759 TEI: 0.7722 DD: 0.7676	Harary: 0.6368 RDD: 0.1496 Balaban: 0.0664 Wiener: 0.0467 TEI: 0.0275	Harary: 0.5336 RDD: 0.1710 Wiener: 0.1092 Balaban: 0.0467 TEI: 0.0424
Ionization Energy	Harary: 0.8488 RDD: 0.8309 TEI: 0.7793 Wiener: 0.7765 DD: 0.7689	Harary: 0.5334 RDD: 0.2132 Balaban: 0.0714 Wiener: 0.0556 DD: 0.0418	Harary: 0.3628 RDD: 0.3110 DD: 0.0807 Wiener: 0.0666 Balaban: 0.0638

Table 14. Best hyperparameters for random forest models across physicochemical properties

Atomic Property	**BP**	**MV**	**MR**	**FP**	**Polar**	**EV**
	Random Forest Hyperparameters					
	max depth, min samples leaf, min samples split, n_estimators					
Atomic Number	None, 1, 2, 200	None, 1, 2, 200	None, 1, 2, 200	10, 1, 5, 50	10, 1, 2, 200	10, 1, 2, 200
Atomic Mass	None, 1, 2, 200	None, 1, 2, 200	None, 1, 2, 200	10, 1, 5, 50	None, 1, 2, 200	None, 1, 2, 200
Atomic Radius	None, 1, 2, 200	10, 1, 2, 100	10, 1, 2, 200	None, 1, 2, 200	None, 1, 2, 200	None, 2, 5, 50
Density	None, 1, 5, 200	10, 1, 2, 50	None, 1, 2, 50	10, 2, 2, 100	10, 1, 2, 200	10, 2, 5, 100
Electronegativity	None, 1, 5, 100	10, 1, 2, 200	10, 1, 2, 200	None, 2, 2, 200	None, 1, 2, 200	None, 2, 5, 200
Ionization Energy	None, 1, 2, 100	10, 1, 2, 100	None, 1, 2, 200	None, 1, 2, 200	10, 1, 2, 200	10, 1, 2, 200

Table 15. Best hyperparameters for XGBoost models across physicochemical properties

Atomic Property	**BP**	**MV**	**MR**	**FP**	**Polar**	**EV**
	XGBoost Hyperparameters					
	max depth, min child weight, n-estimators, learning rate, colsample bytree, subsample					
Atomic Number	7, 3, 50, 0.1, 0.8, 1.0	3, 3, 200, 0.3, 0.8, 0.8	3, 3, 200, 0.3, 0.8, 0.9	3, 5, 100, 0.1, 0.9, 0.8	3, 1, 200, 0.3, 0.8, 0.9	7, 5, 50, 0.1, 0.8, 0.9

Continued on next page

Table 15 – continued from previous page

Atomic Property	**BP**	**MV**	**MR**	**FP**	**Polar**	**EV**
Atomic Mass	5, 1, 200, 0.3, 0.8, 1.0	3, 1, 200, 0.1, 1.0, 0.9	5, 5, 200, 0.1, 0.8, 0.8	3, 5, 50, 0.3, 0.9, 0.8	5, 5, 200, 0.1, 0.8, 0.8	7, 3, 50, 0.1, 0.9, 0.8
Atomic Radius	7, 5, 50, 0.1, 0.8, 0.8	3, 5, 200, 0.1, 0.8, 0.8	7, 5, 50, 0.1, 0.8, 1.0	7, 1, 200, 0.1, 0.8, 0.9	7, 5, 50, 0.1, 0.8, 1.0	5, 5, 50, 0.1, 0.9, 0.8
Density	3, 5, 100, 0.1, 0.8, 0.8	7, 5, 50, 0.1, 0.9, 0.9	7, 5, 50, 0.1, 0.8, 0.9	5, 5, 50, 0.1, 0.8, 0.9	5, 5, 50, 0.1, 0.8, 0.9	7, 5, 50, 0.1, 0.8, 0.8
Electronegativity	7, 5, 50, 0.1, 1.0, 0.8	3, 1, 200, 0.3, 0.9, 0.8	5, 5, 50, 0.1, 0.8, 0.8	7, 5, 50, 0.1, 0.8, 0.9	5, 3, 200, 0.1, 0.8, 0.8	7, 5, 50, 0.1, 1.0, 0.8
Ionization Energy	7, 3, 50, 0.1, 0.8, 0.9	3, 1, 200, 0.1, 1.0, 0.8	3, 5, 100, 0.1, 0.8, 0.8	7, 1, 200, 0.01, 1.0, 1.0	3, 5, 50, 0.1, 0.8, 0.8	7, 3, 50, 0.1, 0.9, 0.9

5 Discussion and limitations

Our approach, utilizing degree-distance-based topological indices (TIs), offers an interpretable and computationally efficient alternative to complex machine learning frameworks such as graph neural networks (GNNs). TIs are derived from the structural features of molecules encoded in vertex-edge weighted (VEW) graphs and have well-defined mathematical formulations. This transparency allows researchers to directly associate individual indices with specific physicochemical properties. For example, the high correlation of the Reciprocal Distance Degree (RDD) index with Molar Refractivity (MR) ($r = 0.95$) provides a mechanistically interpretable connection between atomic connectivity and electronic polarizability, offering valuable guidance for molecular design.

Unlike GNNs, which learn abstract representations through iterative message passing between atoms and bonds, TIs aggregate global structural characteristics. Although GNNs can model complex non-linear dependencies and capture local environments (e.g., ring systems, functional groups), they are often regarded as "black boxes" due to their lack of interpretability. In contrast, TIs such as Harary or Gutman encode chemically meaningful quantities, enabling domain experts to identify how specific structural motifs influence a target property. This distinction becomes crucial in drug discovery, where the explainability of predictions is essential for regulatory approval and rational optimization.

Our comparative analysis highlights the nuanced role of atomic properties across different modeling paradigms and property complexities. Among the atomic descriptors used to weight vertex-edge weighted (VEW) molecular graphs, density emerges as a particularly effective input for predicting moderately complex properties such as the boiling point (BP) and the flash point (FP). Its integration with topological indices (TIs) like Harary and RDD enhances performance across both linear and non-linear models, likely due to its capacity to capture mass distribution effects that influence thermal and volatility-related behaviors. While overall predictive power for BP and FP remains modest ($R^2 < 0.7$), Density consistently yields superior results relative to other atomic features in this category.

In contrast, Electronegativity and Ionization Energy emerge as consistently effective atomic descriptors across both linear and non-linear models. Their ability to capture electron distribution and bonding behavior proves especially valuable for predicting properties influenced by subtle electronic interactions, including Boiling Point (BP), Flash Point (FP), and Enthalpy of Vaporization (EV). These features enhance the performance of topological indices such as Gutman and Degree Distance (DD), particularly when paired with ensemble methods like Random Forest (RF) and XGBoost (XGB), which can better exploit non-linear relationships.

Most notably, the proposed QSPR framework achieves high predictive accuracy for properties such as Molar Volume (MV), Molar Refractivity (MR), and Polarizability (Polar), where R^2 values exceed 0.9 using optimized linear and ensemble models. These results underscore the strong synergy between degree-distance-based TIs (especially Harary and RDD), well-chosen atomic weightings (e.g., density and electronegativity), and suitable learning algorithms. This combination not only ensures robust and interpretable predictions but also positions our method as a practical and transparent alternative to less interpretable approaches such as graph neural networks making it highly applicable to drug discovery and molecular design tasks.

Importantly, the application of our approach to a general set of 166 drug-like molecules—rather than a single chemical class—enabled the identification of universal descriptors applicable across multiple physicochemical properties. For instance, Harary's wide applicability underscores its central role in QSPR modeling. These findings offer a structured and interpretable framework for virtual screening, reducing reliance on costly experimental procedures and accelerating the drug development process.

While the results are promising, the study is not without limitations. The relatively small dataset of 166 molecules, although chemically diverse, may not capture the full variability of real-world chemical space. Larger datasets, such as QM9 or Tox21, would allow for more generalizable conclusions. Additionally, while VEW molecular graphs account for pairwise atomic interactions, they are inherently limited in representing higher-order interactions and three-dimensional conformational effects.

Another limitation is that global TIs may not sufficiently encode local structural features that significantly affect certain properties (e.g., reactive sites influencing FP or toxicophores affecting toxicity). Substructure-aware features (e.g., SMARTS patterns) could be incorporated to address this limitation. Although GNNs offer improved accuracy by learning such features, they were not adopted here due to their black-box nature and computational complexity. Our choice reflects a prioritization of interpretability, which is often more actionable in the context of drug discovery.

6 Conclusion and future work

This study demonstrates that degree-distance-based TIs can effectively model key physicochemical properties of drug-like molecules through machine learning. By integrating atomic-level features with structural indices, we developed interpretable models capable of providing insight into structure-property relationships. TIs such as RDD and Harary consistently emerged as strong predictors, particularly when paired with appropriately selected linear or non-linear models.

In future work, we propose the development of hybrid models that combine the interpretability of TIs with the representational power of GNNs. For instance, TIs could be used as additional input features within GNN architectures or employed to guide attention mechanisms, enhancing both transparency and accuracy. Expanding the dataset to include thousands of molecules from benchmark sets like QM9 or Tox21 will also improve model robustness and facilitate validation across diverse chemical domains.

Further enhancement could come from extending classical graph representations to higher-dimensional structures such as simplicial complexes or hypergraphs, which capture multi-atom interactions and complex structural hierarchies. Additionally, the use of explainability techniques such as SHAP (SHapley Additive exPlanations) could quantitatively reveal the contribution of each TI to model predictions, enhancing interpretability.

Our publicly available codebase (https://github.com/ssorgun/LNNR) lays the groundwork for reproducibility and future exploration. Overall, this work provides a foundation for interpretable, scalable, and efficient QSPR modeling and opens promising directions for the integration of topological reasoning with modern machine learning.

Data availability

The dataset and code used in this study are available at `https://github.com/ssorgun/LNNR`https://github.com/ssorgun/LNNR. It is recommended that readers look at the README file in Github for information on how the codes and analysis work.

***Acknowledgment*:** The authors gratefully acknowledge the financial support from the Scientific and Technological Research Council of Turkey (TUBITAK) under the BİDEB 2221 program.

References

[1] M. Arockiaraj, J. J. Godlin, S. Radha, T. Aziz, M. Al-Harbi, Comparative study of degree, neighborhood and reverse degree based indices for drugs used in lung cancer treatment through qspr analysis, *Sci. Reports* **15** (2025) #3639.

[2] N. Awan, A. Ghaffar, F. M. Tawfiq, G. Mustafa, M. Bilal, M. Inc, QSPR analysis for physiochemical properties of new potential antimalarial compounds involving topological indices, *Int. J. Quantum Chem.* **124** (2024) #e27391.

[3] A. T. Balaban, Highly discriminating distance-based topological index, *Chem. Phys. Lett.* **89** (1982) 399–404.

[4] C. M. Bishop, N. M. Nasrabadi, Pattern recognition and machine learning, *J. El. Imag.* **16** (2007) #049901.

[5] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: B. Krishnapuram, M. Shah (Eds.), *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Assoc. Comput. Machinery, New York, 2016, pp. 785–794.

[6] A. Dobrynin, A. A. Kochetova, Degree distance of a graph: A degree analog of the Wiener index, *J. Chem. Inf. Comput. Sci.* **34** (1994) 1082–1086.

[7] E. Estrada, E. Uriarte, Recent advances on the role of topological indices in drug discovery research, *Curr. Med. Chem.* **8** (2001) 1573–1588.

[8] G. H. Fath-Tabar, M. J. Nadjafi-Arani, M. Mogharrab, A. R. Ashrafi, Some inequalities for Szeged-like topological indices of graphs, *MATCH Commun. Math. Comput. Chem.* **63** (2010) 145–150.

[9] J. Gasteiger, *Handbook of Chemoinformatics: From Data to Knowledge*, Wiley, Weinheim, 2003.

[10] I. Gutman, O. E. Polansky, *Mathematical Concepts in Organic Chemistry*, Springer, Berlin, 1986.

[11] I. Gutman, Selected properties of the Schultz molecular topological index, *J. Chem. Inf. Comput. Sci.* **34** (1994) 1087–1089.

[12] I. Gutman, B. Furtula, I. Redžepović, On topological indices and their reciprocals, *MATCH Commun. Math. Comput. Chem* **91** (2024) 287–297.

[13] I. Gutman, Geometric approach to degree-based topological indices: Sombor indices, *MATCH Commun. Math. Comput. Chem.* **86** (2021) 11–16.

[14] P. Bickel, P. Diggle, S. Fienberg, U. Gather, I. Olkin, S. Zeger, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York, 2009.

[15] A. E. Hoerl, R. W. Kennard, Ridge regression: Biased estimation for non orthogonal problems, *Technometrics* **12** (1970) 55–67.

[16] H. Hua, S. Zhang, On the reciprocal degree distance of graphs, *Discr. Appl. Math.* **160** (2012) 1152–1163.

[17] C. Huang, W. Gao, Y. Zheng, W. Wang, Y. Zhang, K. Liu, Universal machine-learning algorithm for predicting adsorption performance of organic molecules based on limited data set: Importance of feature description, *Sci. Total Environ.* **859** (2023) #160228.

[18] O. Ivanciuc, T. Ivanciuc, A. Balaban, Vertex-and edge-weighted molecular graphs and derived structural descriptors, in: J. Devillers, A. T. Balaban (Eds.), *Topological Indices and Related Descriptors in QSAR and QSPR*, Gordon & Breach, Amsterdam, 1999, pp. 169–220.

[19] H. Khodashenas, M. J. Nadjafi-Arani, A. R. Ashrafi, I. Gutman, A new proof of the Szeged-Wiener theorem, *Kragujevac J. Math.* **35** (2011) 165–172.

[20] Y. Lecun, Y. Bengio, G. Hinton, Deep learning, *Nature* **521** (2015) 436–444.

[21] M. J. Nadjafi-Arani, H. Khodashenas, A. R. Ashrafi, Relationship between edge Szeged and edge Wiener indices of graphs, *Glasnik matematički* **47** (2012) 21–29.

[22] S. Nasir, Topological descriptors of colorectal cancer drugs and characterizing physical properties via QSPR analysis, *Int. J. Anal. Chem.* **1** (2025) #5512172.

[23] S. Nasir, F. B. Farooq, S. Parveen, Topological indices of novel drugs used in blood cancer treatment and its QSPR modeling, *AIMS Math.* **7** (2022) 11829–11850.

[24] Z. Qi, S. Zhong, X. Huang, Y. Xu, H. Zhang, B. Shi, Concentration division for adsorption coefficient prediction using machine learning with Abraham descriptors: Data-splitting approach comparison and critical factors identification, *Carbon* **230** (2024) #119573.

[25] C. Qu, A. J. Kearsley, B. I. Schneider, W. Keyrouz, T. C. Allison, Graph convolutional neural network applied to the prediction of normal boiling point, *J. Mol. Graphics Model.* **112** (2022) #108149.

[26] C. Qu, B. I. Schneider, A. J. Kearsley, W. Keyrouz, T. C. Allison, Predicting Kováts retention indices using graph neural networks, *J. Chromatography A* **1646** (2021) #462100.

[27] S. Parveen, N. H. Awan, F. Farooq, R. Fanja, Q. Anjum, Topological indices of novel drugs used in autoimmune disease vitiligo treatment and its QSPR modeling, *BioMed Res. Int.* **2022** (2022) #6045066.

[28] D. Plavsić, S. Nikolić, N. Trinajstić, Z. Mihalić, On the Harary index for the characterization of chemical graphs, *J. Math. Chem.* **12** (1993) 235–250.

[29] S. Sardana, A. Madan, Application of graph theory: Relationship of molecular connectivity index, Wiener's index and eccentric connectivity index with diuretic activity, *MATCH Commun. Math. Comput. Chem.* **43** (2001) 85–98.

[30] M. Shanmukha, N. Basavarajappa, K. Shilpa, A. Usha, Degree-based topological indices on anticancer drugs with QSPR analysis, *Heliyon* **6** (2020) #e04235.

[31] S. Sorgun, K. Birgin, Vertex-edge-weighted molecular graphs: A study on topological indices and their relevance to physicochemical properties of drugs used in cancer treatment, *J. Chem. Inf. Model.* **65** (4) (2025) 2093—2106.

[32] S. Sorgun, H. Küçük, K. Birgin, Some distance-based topological indices of certain polysaccharides, *J. Mol. Struct.* **1250** (2022) #131716.

[33] S. Sorgun, A. Ullah, A python-based novel vertex–edge-weighted modeling framework for enhanced QSPR analysis of cardiovascular and diabetes drug molecules, *Eur. Phys. J. E* **48** (2025) #36.

[34] S. Sultana, Chemical application of topological indices in infertility treatment drugs and QSPR analysis, *Int. J. Anal. Chem.* **1** (2023) #6928167.

[35] R. Tibshirani, Regression shrinkage and selection via the lasso, *J. Royal Stat. Soc. Ser. B Stat. Meth.* **58** (1996) 267–288.

[36] V. Uddameri, M. Kuchanur, Fuzzy QSARs for predicting logKoc of persistent organic pollutants, *Chemosphere* **54** (2004) 771–776.

[37] X. Zhang, M. J. Saif, N. Idrees, S. Kanwal, S. Parveen, F. Saeed, QSPR analysis of drugs for treatment of schizophrenia using topological indices, *ACS Omega* **8** (2023) 41417–41426.

[38] X. Zhang, Z. S. Bajwa, S. Zaman, S. Munawar, D. Li, The study of curve fitting models to analyze some degree-based topological indices of certain anti-cancer treatment, *Chem. Papers* **78** (2024) 1055—1068.

.