Mathematics

https://www.aimspress.com/journal/Math

AIMS Mathematics, 10(10): 24651–24690.
DOI: 10.3934/math.20251093
Received: 21 September 2025
Revised: 12 October 2025
Accepted: 15 October 2025
Published: 28 October 2025

*Research article*

# A machine learning framework for QSPR modeling of drug-like compounds using graph invariants

**Ebraheem Alzahrani**[1] **and Muhammad Farhan Hanif**[2,*]

[1] Department of Mathematics, Faculty of Science, King Abdulaziz University, Jeddah, Saudi Arabia
[2] Department of Mathematics and Statistics,The University of Lahore, Lahore Campus, Pakistan

* **Correspondence:** Email: farhanlums@gmail.com.

**Abstract:** Quantitative structure property relationship (QSPR) is a computational modeling approach that correlates the chemical structure of compounds with their physicochemical or biological properties. Accurate estimation of physicochemical and other biological parameters of drug molecules is a critical factor in drug discovery. In the present work, we developed a graph-based QSPR model for molecular structures which employed molecular structural invariants as predicting features. Degree and distance topological indices were derived from molecular graphs and combined with random forest (RF), gradient boosting, and multiple line regression (MLR) for prediction of predictive performance on the diverse drug datasets. The proposed RF model obtained an approximate 18–25% improvement in $R^2$ and a reduction of about 30% in RMSE over the classical linear regression models, showing better generalization performance. In the context of drug screening, the model accurately predicted early physicochemical properties including molar refractivity and polarizability, rendering it a tool to assess rapidly compounds for neurological and anticancer therapeutics. In addition, the computational model was about 15 times faster on average than existing QSPR approaches, achieving its excellent efficiency and applicability. The final results demonstrated that the molecular structural invariants functioned as good descriptors for generating reliable, interpretable, and predictive QSPR models relevant to early-stage drug discovery.

**Keywords:** quantitative structure property relationship; machine learning; structural graph invariants; drug like compounds; molecular descriptors; predictive modeling
**Mathematics Subject Classification:** 05C10, 05C90

## 1. Introduction

Graph theory is a branch of mathematics concerned with the study of graphs, which are mathematical structures used to model pairwise relations between objects. A graph $G$ is defined as an ordered pair $G = (V, E)$, where $V$ is the set of vertices (or nodes), and $E$ is the set of edges (or links) connecting pairs of vertices. One fundamental concept in graph theory is the degree of a vertex. The degree of a vertex $v \in V$, denoted by $F_v$, is the number of edges incident to $v$. For undirected graphs, this is simply the count of edges that have $v$ as an endpoint, whereas for directed graphs, we distinguish between in-degree and out-degree [1].

Chemical graph theory is a specialized area of graph theory where graphs are employed to represent molecular structures. In such representations, atoms are depicted as vertices, and chemical bonds as edges. This abstraction enables the application of mathematical methods to the analysis of molecular properties, structure elucidation, and reaction mechanisms. The molecular graph of a compound can be used to study its symmetry, stability, and reactivity, thereby bridging discrete mathematics and theoretical chemistry [2].

Topological indices are numerical values calculated from the graph structure of a molecule, and are used to model various physicochemical properties. These indices are invariant under graph isomorphism, meaning they do not change if the graph is relabeled. Common examples include the Wiener index, the Zagreb indices, and the Randić index. Each topological index captures certain aspects of molecular connectivity and branching, and they play a crucial role in quantitative structure-activity relationships (QSAR) and quantitative structure-property relationships (QSPR) studies in computational chemistry and drug design [3].

Zhang et al. [4] performed a QSPR study of topological indices of schizophrenia drugs for physicochemical property prediction. In their research, an enourmous relationship between molecular structures and drugs activity, was suggested, confirming the utility of topological descriptors in drug design. Adnan et al. [5] used topological indices to perform a QSPR of antituberculosis drugs. They found predictive capacity of these indices in modeling biological and physicochemical properties of their studied compounds. Shi et al. [6] developed novel QSPR modeling methodologies using topological indices for anticancer drugs. Their strategy improved the predictive power for important molecular properties, indicating the application potential of the topological descriptors in anticancer drug development. MC et al. [7] have reported a QSPR calculation of drug on dry eye disease with topological indices. Their study combined predictive modeling and decision approaches to measure drug effects and molecular features. Molecular graphs and the entropy based descriptors were used by [8] for the QSPR analysis of drugs through the aid of machine learning methods. They showed that a better predictive accuracy can be achieved by including topological information into the entropy measures when modeling drug properties.

It is well-known that graph theory and machine learning play an important role in the research on understanding and predicting complex network behaviors. Liu et al. [14] characterized the coherence and structural characteristics of balanced $2^p$-ary tree networks, illustrating how hierarchical topology impacts network stability and communication. Building upon this, Liu et al. [15] limited to unbalanced networks but also provides an understanding of designing resilient systems. In the field of predictive modeling, Lai [16] introduced long-term prediction of time series data by capturing the local as well as global temporal dependencies in it. Lai and Chen [17] proposed a self-supervised model which mines the latent node relationships for traffic prediction [18] fuses explicit and implicit spatiotemporal features to improve the long-term prediction accuracy. In materials and molecular chemistry, Ismail et al. [19] used degree-based topological indices with rational curve fitting to predict the heat of formation in titanium tetraboride, showing how cost effective graph-based descriptors are. Topological indices and machine learning were applied by [20] to predict properties of bone cancer drugs, and Qin et al. [21] used graph-based and machine learning methods to predict molecular property of bladder cancer drugs.

Ahmed et al. [22] also built a series of QSPR models for profens with molecular descriptors. XGBoost (Extreme gradient boosting) is a fast, portable and accurate library for gradient boosting. It constructs a set of decision trees one after another, where each successive tree corrects errors made by previous ones, with methods that inject randomness and/or parallelism to improve speed, accuracy and mitigate overfitting. Glaucoma drugs have been analyzed with XGBoost and regression models in [23] to enhance the interpretability as well as prediction results. Qin et al. [24] introduced a Python-based topology modeling system in pulmonary cancer drugs, and Wei et al. [25] indicated linear regression models ability to provide good prediction for QSPR in various compounds. Ren et al. [26] proposed to model physics based on deep learning in order to produce parametrized and interpretable models for property prediction that operate half-way between theory-based and data-driven approaches. Other parallel studies have concentrated in molecular docking and drug-target interaction: Zhu et al. [27] introduced computational methods on the drug-target interaction prediction; Monine, Adikesavan, and Lakdawala [28] carried out 3D-QSAR and molecular docking studies on aminopyrimidine-based multi-target inhibitors, and Zhou et al. [29] proposed a network embedding for drug disease associations prediction. Taken together, these studies represent a unifying trend toward incorporating graph-theoretic and machine learning with physical modeling as approaches to improve predictive accuracy, robustness, and interpretability in network analysis and molecular property prediction. Drug discovery is costly and time consuming, hence requiring accurate prediction of molecular properties and biological activities. In recent years, (QSPR) models have provided a new avenue in calculating physicochemical and biological properties of organic molecules based on their molecular structure, which is conducive to reduce experimental burdens and accelerate the process of candidate selection. Through relating chemical features with observed activities, QSPR models help in estimating solubility, permeability, binding affinity, and other pharmacokinetic properties which are important for rational drug development.

However, despite these benefits, current QSPRs are still confronted with some limitations and problems that affect their accuracy and applicability. Most of the existing works are based on mono-type molecular properties for representing a molecule, e.g., only topological, electronic, or geometrical information, etc., which cannot properly describe the high-dimensional structures of molecules. This limitation is exacerbated when modeling polycyclic, heterocyclic, or macrocyclic molecules, as these systems display physicochemical properties that result from complex topological and conformational factors. In addition, conventional regression based or shallow learning models tend to be not able to capture highly nonlinear relationships between structure descriptors and molecular properties, which results in over-fitting with new scaffolds.

Furthermore, the descriptor filtering is already a challenging task. Most QSPR models rely on a priori descriptors that are not necessarily portable between chemical classes or property spaces. As a result, models obtained may not be robust or generalize to different data outside the training set. Moreover, the progressive accessibility of extensive and heterogeneous molecular datasets calls for modeling approaches that can address different types of structural invariants: Topological, geometrical, and electronic in a single unified predictive representation.

To overcome these disadvantages, we introduce a QSPR modeling methodology in the current work, taking simultaneously into account several structural invariants and advanced machine learning approaches. Concomitantly with topological indices, physicochemical parameters, and geometric features, the present method involves these molecular descriptors. The machine learning part also allows for the detection of nonlinear relationships between structural descriptors and target properties, which contributes to higher generalization performance and predictive power. The framework intends to overcome the traditional descriptor-based limitations of QSPR models and offer a more inductive and data-driven basis for molecular property prediction in this manner.

QSPR modeling has historically played a key role in computational drug discovery, serving as a conceptual vehicle to associate molecular structure with physicochemical and biological descriptors. Classical QSPR methods are typically based on topological indices which quantitatively reflect properties of molecular graphs in numerical form, such as the Randic, Zagreb and Wiener indices. These descriptors historically allowed regression-based modeling of reduced sets, but are limited in representing ability and linearity assumptions. With the growth of chemical space and diversity, it is more important than ever to have models that are able to learn nonlinear high-dimensional relationships between molecules.

Over the last few years, the incorporation of machine learning methods in QSPR modeling has changed the face of predictive chemistry. Deep learning, ensemble regression and specifically graph neural networks (GNNs) have made it possible to learn molecular representations directly from the graph topology.

GNN-based architectures can naturally encode multi-scale atom bond interactions, and thus significantly surpass traditional descriptor-related schemes in practice [61]. Similarly, Brozos et al. [62] proposed a multitask GNN for predicting surfactant properties and achieved good cross-domain generalization.

In this paper, we extend these papers by introducing a hybrid QSPR framework that combines several structural invariants and algorithms based on machine learning to improve the prediction performances in terms of accuracy and interpretability. The model with the combination of multiple descriptors reflecting topological, geometric, and physicochemical features can better simulate nonlinear relationships between structure and property. In addition, this way preserves the interpretative usefulness of structural invariants, which is consistent with the current wish for transparent and interpretable QSPR systems.

QSPR modeling study of postnatal depression drugs by Roy [9] used topological indices in regression analysis. It showed that structural patterns in fact correlate with drug properties and could thereby be used in predicting models. Ahmed et al. [10] investigated anti-biofilm agents by using QSPR analysis based on topological descriptors and good molecular mechanism interpretations were found. Abid et al. [11] performed the computer-based QSPR studies of anti-tuberculosis drugs using 21 molecular descriptors to examine their structural and physicochemical behavior. This work validated the applicability of descriptor-based modeling for the prediction of drug performance and dosage form development. Paul et al. [12] investigated the QSPR model on neuromuscular drugs based on molecular descriptors for their physicochemical properties. The results support the use of a descriptor-based model in interpreting drug behavior and facilitating drug design. Qin et al. [13] constructed a QSPR model for anti-arrhythmia drugs using topological descriptors and implemented it in a Python QSPR framework. This approach demonstrated the effectiveness of computational modeling to represent structure–property relationships for cardiovascular therapeutics.

Gutman and Polansky introduced the first and second Zagreb index [30] as:

$$M_1(G) = \sum_{mn\in E(G)} (F_m + F_n), \quad M_2(G) = \sum_{mn\in E(G)} (F_m F_n).$$

Martınez-Martınez et al. [31] defined the Harmonic index as follows:

$$H(G) = \sum_{mn\in E(G)} \frac{2}{(F_m + F_n)}.$$

The forgotten index was introduced by Furtula and Gutman [32] as:

$$F(G) = \sum_{mn\in E(G)} [(F_m)^2 + (F_n)^2].$$

The Shilpa-Shanmukha index was defined by Zhao et al. [33] as follows:

$$SS(G) = \sum_{mn\in E(G)} \sqrt{\frac{F_m F_n}{F_m + F_n}}.$$

The atom bond connectivity index was introduced by Estrada et al. [34] as:

$$ABC(G) = \sum_{mn\in E(G)} \sqrt{\frac{F_m + F_n - 2}{F_m F_n}}.$$

The Randic index was defined by Randić et al. [35] as:

$$RI(G) = \sum_{mn\in E(G)} \frac{1}{F_m F_n}.$$

The sum connectivity index and the geometric arithmetic index were defined by Vukicevic et al. [36] as:

$$SC(G) = \sum_{mn \in E(G)} \frac{1}{F_m + F_n}, \quad GA(G) = \sum_{mn \in E(G)} \frac{2\sqrt{F_m F_n}}{F_m + F_n}.$$

The Hyper Zagreb Index was introduced by Rajasekharaiah et al. [37] as:

$$HZ(G) = \sum_{mn \in E(G)} (F_m + F_n)^2.$$

The Nirmala Index was defined by Kulli et al. [38] as:

$$N(G) = \sum_{mn \in E(G)} \sqrt{(F_m + F_n)}.$$

Amitriptyline is a well-established tricyclic antidepressant agent used for depression and neuropathic pain. It has a tricyclic molecular structure with three benzene rings fused in the molecule while an aminoethyl side chain ends in a tertiary amine [39]. The lipophilic nature of the compound allows its association with the nuerotransmitter transporter and subsequently blocks reuptake of serotonin and norepinephrine, leading to an increase in synaptic neurotransmitters. Carbamazepine is an anticonvulsant and mood stabilizing drug, most commonly used for epilepsy and bipolar disorder. It contains a dibenzazepine ring structure with a carboxamide 5-fluoro substitution at the C5 position [40]. The mechanism of the drug is to stabilize voltage-gated sodium channels to the inactive form, inhibiting cholinergic stimulation or neurotransmission along these channels. The aromatic and flat structure makes it 3-D nutrient inert.

The classic benzodiazepine, diazepam, is employed to treat anxiety and muscle spasms, and to prevent seizures. It is based on a triazolobenzodiazepine core, in which the benzene ring has been fused to a thiophene ring. Its structural flexibility results in greater receptor affinity and therapeutic efficacy [41]. Donepezil is a cholinesterase inhibitor that is indicated for the treatment of symptoms of Alzheimer's disease. The derivative has an indanone core and a piperidine ring in its structure [42]. Donepezil is a reversible acetylcholinesterase inhibitor and increases acetylcholine in the brain.

Ergotamine, an ergot alkaloid, is mainly used as an acute treatment for migraines since it causes vasoconstriction. Mimicking the architecture of serotonin, dopamine, and norepinephrine, it interacts with various receptors. Its stereochemistry and conformational inflexibility are important for its biological activity [43]. Ethosuximide is a first-line drug for partial absence seizures and has a succinimide ring with an aliphatic side chain. It acts by inhibiting T-type calcium channels on thalamic neurons. The drug is of low molecular weight with a cyclic conformation that is reported to enhance its penetrability across the blood-brain barrier [44].

Fingolimod, approved for the treatment of multiple sclerosis, is a structural analogue of sphingosine and is a molecule with a polar head and a lipophilic tail. It activates sphingosine-1-phosphate receptors capturing lymphocytes in the lymph nodes and influencing immune responses [45]. Its amphipathic character permits integration into the membrane and/or interaction with the receptor. Galantamine, an alkaloid of natural origin, is used in the therapy of Alzheimers disease. Its tricyclic structure contains a tertiary amine [46]. It acts as a reversible inhibitor of cholinesterases and allosteric modulator of nicotinic receptors when in its folded conformation, but has shown to have increased enzymatic binding as a depolymerized monomer.

It is a broad-spectrum anticonvulsant used to treat bipolar disorder as well. It contains a triazine ring with a 2,3-dichlorophenyl substituent [47]. Lamotrigine blocks voltage-gated sodium channels and thereby stabilizes neuronal membranes. Its delocalized heterocyclic system is responsible for significant neuroactivity. Levetiracetam is an antiepileptic drug (AED) that is successful in treating partial and generalized seizures. Its structure consists of a pyrrolidone ring, a butyl side chain, and an amide group [48].

Lorazepam, is a benzodiazepine, frequently prescribed for anxiety, insomnia and seizures. It has a parent structure: Fused 2-phenyl-benzodiazepine containing hydroxy substitution at the 3-position [49]. Lorazepam potentiates inhibitory transmission through the gamma-aminobutyric acid (GABA) receptor. It is also lipophilic and flat allowing its penetration through the blood-brain barrier (BBB). Phenytoin is a hydantoin derivative that is commonly used for seizure control by inhibition of voltage-gated sodium channels. It comprises of a Hydantoin ring substituted by two phenyl groups [50]. Its inflexible aromatic ring system is believed to have contributed to the durability of the neuronal stabilization and anticonvulsive action.

Pramipexole is a dopamine D2/D3 receptor agonist used to treat Parkinsons disease and restless leg syndrome. It contains a thiazole and benzothiazole unit [51]. The bicyclic, heteroaromatic structure allows selective receptor binding and provides a mimic for dopamine. Propranolol is a beta-adrenergic nonselective blocker used for treatment of hypertension, arrhythmic disorders, and anxiety. It consists of a naphthalene ring and aliphatic chain attached with a secondary alcohol group [52]. Its lipophilic nature permits CNS access, with possible peripheral and central effects.

Rivastigmine is an acetylcholinesterase and butyrylcholinesterase dual inhibitor, employed in dementia treatment. It contains a phenyl group with a carbamate moiety [53]. The latter increases acetylcholine duration at synapses, and thereby cholinergic activity in the brain. Rizatriptan, a selective 5-Hydroxytryptamine receptor subtype 1B and 1D (5-HT1B/1D) receptor agonist, is indicated for the acute treatment of migraines. It is structurally similar to serotonin and consists of an indole together with a triazole ring [54]. This manner of mimicry results in high receptor affinity and vasoconstrictive potency.

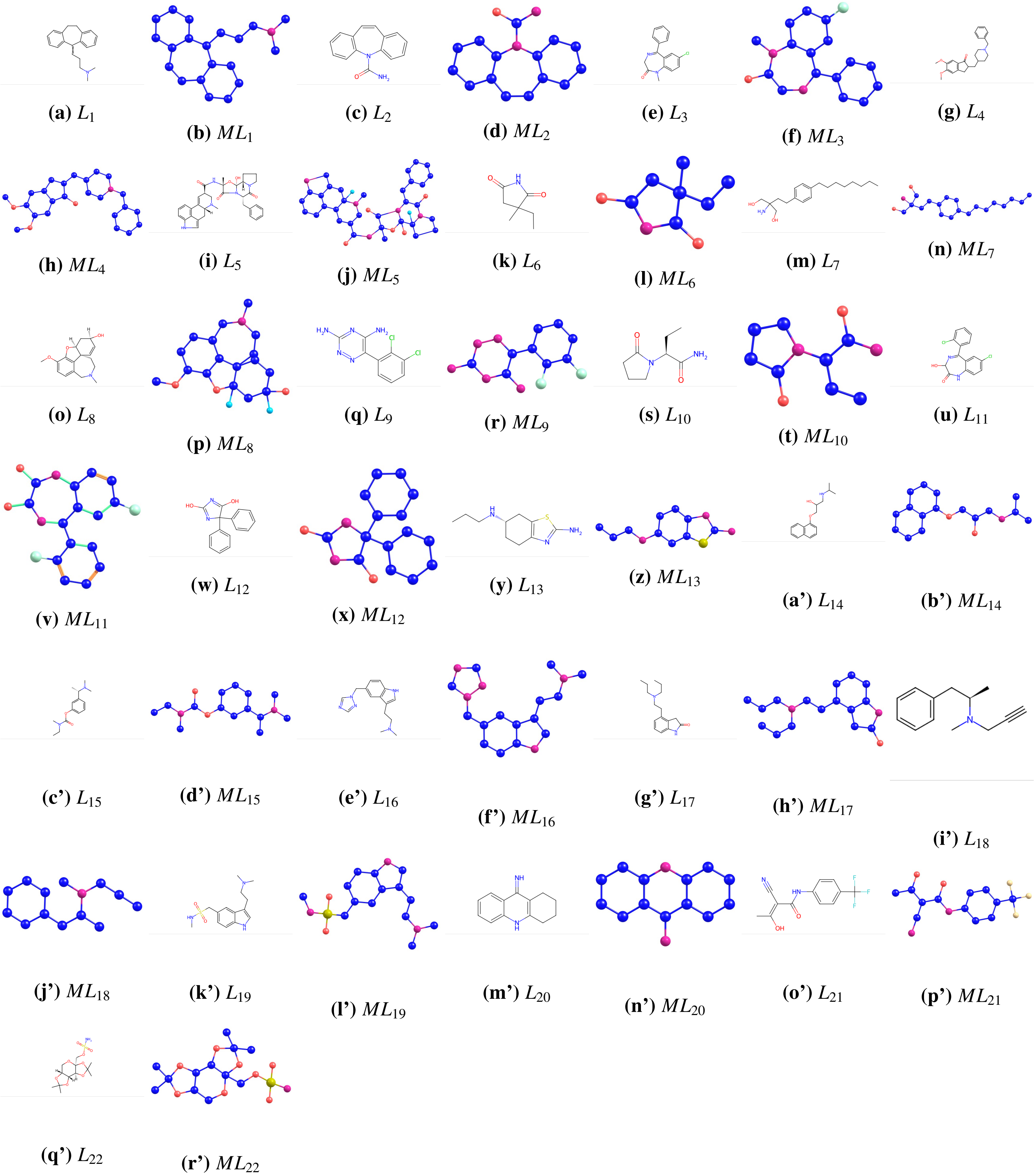
Ropinirole is a dopamine agonist currently indicated for Parkinson's disease and restless leg syndrome. It comprises an indoline nucleus with a basic side chain [55]. They chose to model dopamine based on its functional groups and orientation and succeeded in mimicking dopamine's functionality, to specifically lodge in the D2-like receptor only. Selegiline is a monoamine oxidase (B) inhibitor for Parkinson's disease. It is structurally related to methamphetamine with CNS activity modifications [56]. Selegiline extends the half-life of dopamine in the brain, and it has minimal lipophilic structure for crossing the blood-brain barrier.

Sumatriptan is a serotonin receptor agonist that is also a treatment for migraines and cluster headaches. It has an indole as its core and a sulfonamide side chain [57]. It yields specificity receptor and therapeutic vasoconstriction due to its structural resemblance to serotonin. Tacrine is a tetrahydroacridine acetylcholinesterase inhibitor, which was used in the treatment of Alzheimers disease [58]. It was effective for elevating acetylcoline levels but withdrawn because of hepatotoxicity issues.

Isoxazole ring Teriflunomide (an immunomodulatory drug for the treatment of multiple sclerosis) contains an isoxazole ring and a trifluoromethyl moiety [59]. It suppresses dihydroorotate dehydrogenase, thus reducing pyrimidine formation in stimulated lymphocytes. It won't affect your ability to fight off diseases, and immune suppression is targeted. Topiramate is an anticonvulsant drug indicated for seizure management and migraine prophylaxis. It is derived from a sulfamate-containing monosaccharide scaffold [60]. Topiramate regulates voltage-gated ion channels and raises GABA activity, and constitutionally, it is responsible for occupying an extended area of neurological actions.

Chemical and molecular structures of neurological cancer drugs for $L_i, ML_i, i = 1, 2, 3, ..., 22.$ are shown in Figure 1, and physiochemical properties are shown in Table 1.

Table 1 shows a set of drug names alongside various parameters, such as BP, EV, FP, MR, P, ST, and MV. These parameters represent different aspects of each drug's profile, possibly indicating properties like boiling point (BP), efficacy value (EV), flash point (FP), and more. The graph labels ($L_1$ to $L_{21}$) refer to the specific drugs in the dataset, providing a clear mapping to the drugs listed. Each drug has different values across these properties, which can be analyzed to understand their pharmacological behavior. This table can be particularly useful in pharmacological studies, where these numerical values help compare the drugs physical and chemical characteristics.



**(a)** $L_1$ **(b)** $ML_1$ **(c)** $L_2$ **(d)** $ML_2$ **(e)** $L_3$ **(f)** $ML_3$ **(g)** $L_4$

**(h)** $ML_4$ **(i)** $L_5$ **(j)** $ML_5$ **(k)** $L_6$ **(l)** $ML_6$ **(m)** $L_7$ **(n)** $ML_7$

**(o)** $L_8$ **(p)** $ML_8$ **(q)** $L_9$ **(r)** $ML_9$ **(s)** $L_{10}$ **(t)** $ML_{10}$ **(u)** $L_{11}$

**(v)** $ML_{11}$ **(w)** $L_{12}$ **(x)** $ML_{12}$ **(y)** $L_{13}$ **(z)** $ML_{13}$ **(a')** $L_{14}$ **(b')** $ML_{14}$

**(c')** $L_{15}$ **(d')** $ML_{15}$ **(e')** $L_{16}$ **(f')** $ML_{16}$ **(g')** $L_{17}$ **(h')** $ML_{17}$ **(i')** $L_{18}$

**(j')** $ML_{18}$ **(k')** $L_{19}$ **(l')** $ML_{19}$ **(m')** $L_{20}$ **(n')** $ML_{20}$ **(o')** $L_{21}$ **(p')** $ML_{21}$

**(q')** $L_{22}$ **(r')** $ML_{22}$

**Figure 1.** Chemical and molecular structures of neurological cancer drugs for $L_i, ML_i, i = 1, 2, 3, ..., 22$.

**Table 1.** Physicochemical properties of antibiotic molecules with graph identifiers.

| Graph | Drug name | BP | EV | FP | MR | P | ST | MV |
|---|---|---|---|---|---|---|---|---|
| $L_1$ | Amitriptyline | 398.2 | 64.9 | 174.0 | 91.5 | 36.3 | 47.0 | 257.8 |
| $L_2$ | Carbamazepine | 411.0 | 66.3 | 202.4 | 69.7 | 27.6 | 57.3 | 186.6 |
| $L_3$ | Diazepam | 497.4 | 76.5 | 254.6 | 80.9 | 32.1 | 46.1 | 225.9 |
| $L_4$ | Donepezil | 527.9 | 80.3 | 273.1 | 110.4 | 43.8 | 45.2 | 332.5 |
| $L_5$ | Ergotamine | 914.5 | 139.3 | 506.9 | 159.2 | 63.1 | 82.1 | 393.4 |
| $L_6$ | Ethosuximide | 265.3 | 50.3 | 123.8 | 36.0 | 14.3 | 31.0 | 133.7 |
| $L_7$ | Fingolimod | 479.5 | 78.4 | 243.8 | 93.6 | 37.1 | 43.4 | 302.4 |
| $L_8$ | Galantamine | 439.3 | 73.4 | 219.5 | 80.3 | 31.8 | 56.6 | 223.9 |
| $L_9$ | Lamotrigine | 503.1 | 77.2 | 258.1 | 63.4 | 25.1 | 79.6 | 162.9 |
| $L_{10}$ | Levetiracetam | 395.9 | 64.6 | 193.2 | 44.2 | 17.5 | 48.9 | 145.7 |
| $L_{11}$ | Lorazepam | 543.6 | 86.5 | 282.6 | 81.0 | 32.1 | 56.0 | 211.2 |
| $L_{12}$ | Phenytoin | 464.0 | 76.4 | 305.8 | 72.4 | 28.7 | 53.3 | 197.9 |
| $L_{13}$ | Pramipexole | 378.0 | 62.6 | 182.4 | 60.3 | 23.9 | 53.2 | 180.5 |
| $L_{14}$ | Propranolol | 434.9 | 72.8 | 216.8 | 79.0 | 31.3 | 42.7 | 237.2 |
| $L_{15}$ | Rivastigmine | 316.2 | 55.8 | 145.0 | 73.1 | 29.0 | 36.9 | 241.2 |
| $L_{16}$ | Rizatriptan | 504.8 | 77.4 | 259.1 | 80.7 | 32.0 | 46.6 | 222.5 |
| $L_{17}$ | Ropinirole | 410.5 | 66.3 | 202.0 | 78.4 | 31.1 | 40.5 | 250.2 |
| $L_{18}$ | Selegiline | 272.5 | 51.1 | 108.4 | 60.5 | 24.0 | 37.3 | 196.2 |
| $L_{19}$ | Sumatriptan | 497.7 | 76.6 | 254.8 | 82.4 | 32.6 | 52.7 | 237.6 |
| $L_{20}$ | Tacrine | 353.8 | 59.9 | 167.8 | 59.8 | 23.7 | 49.6 | 157.8 |
| $L_{21}$ | Teriflunomide | 410.8 | 69.9 | 202.3 | 60.6 | 24.0 | 45.4 | 189.7 |
| $L_{22}$ | Topiramate | 438.7 | 69.6 | 219.1 | 74.3 | 29.5 | 53.4 | 253.9 |

## 2. Innovation and methodological contributions

The current study proposes a novel hybrid QSPR modeling strategy, which overcomes the limitations of traditional topological-index-based methods and is based on three main developments. First, in contrast to previous schemes that depend on only one kind of molecular descriptors, our model incorporates multi-type structural invariants: Topological, geometric, and physicochemical under the same set. This multimodal combination represents one of the different kinds of structural information pieces which is frequently missed by traditional QSPR models that are restricted to a single type descriptor.

Second, the model uses a hybrid learning approach, being composed of linear, quadratic, and ensemble regressors that permit interpretation and nonlinearity in the mapping between structural invariants and molecular properties. The proposed approach is developed in a counterbalance to traditional QSPR approaches which rely solely on linear relationships as our model takes advantage of random forest (RF) learning for consideration of nonlinear behavior contained between descriptors, ultimately leading to improved generalization and prediction capability.

Third, our framework focuses on interpretability of structural invariants as a coherent modeling goal. The importance of features and weights for contributors within the RF model is further dissected, which shows that some structural indicators (i.e., degree based indices, distance based indices) have direct and interpretable links with physicochemical properties. This makes our predictions directly interpretable which sets us apart from recent deep-learning-based QSPR methods that trade interpretability with prediction accuracy.

Taken together, these methodological improvements represent a unique research contribution: The systematic confluence of multiple structural invariants with interpretable machine learning to attain a balance between accuracy, generality, and the explainable. Not only is this design enhanced in terms of prediction ability, but it also offers deep chemical insight into the effect of molecular topology on target properties.

## 3. Experimental details

### *3.1. Dataset screening and selection*

A final dataset of 22 established neuropharmaceuticals was chosen to compose the QSPR data. The screen was limited to drugs used to treat neurological and psychiatric conditions and included six structure-based classes: Tricyclic antidepressants, benzodiazepines, selective serotonin reuptake inhibitors (SSRIs), monoamine oxidase inhibitors, antiepileptics, and atypical antipsychotic. This set was deliberately selected to contain structural and chemical diversity that would enable the model to examine how it generalizes across molecular scaffolds. Only molecules with full and validated experimental physicochemical property records were kept and those without validated data or structural coherence were discarded.

### *3.2. Data source and property annotations*

All molecular and physicochemical information was taken from the ChemSpider database only (https://www.chemspider.com/). For each selected compound, the experimentally verified BP, FP, and EV were obtained from uncovering curated experimental records on ChemSpider. Molecular structures were obtained in smilies and then analyzed for the chemical topological, geometrical, and physicochemical invariants using the statistical package for the social sciences (SPSS) software. Generations of descriptors used the same parameters for all compounds to enable reproducibility.

### *3.3. Data preprocessing and feature preparation*

All preprocessing and model building continued in the Python programming language (Version 3.10) through use of open-source packages, pandas, numpy, scikit-learn as well as matplotlib. The descriptor matrices were filtered for missing values, constant features, and multicollinearity. Features that had more than 10% of missing values or were correlated at $r > 0.90$ to other features were discarded to minimize redundancy. The other descriptors were z-score normalized in such a way that their mean value equals zero and their standard deviation equals 1. This preprocessing ensured consistent scaling of features and was essential for stable optimization during regression and ensemble learning.

### 3.4. Model development and validation

Predictions were carried out via three regression analysis methods using Python scripts: (i) Multiple linear regression (MLR), (ii) polynomial curve fitting or quadratic regression, and (iii) RF. 10 fold cross validation was used to train and validate each model to ensure the robust estimation of performance. Divide the dataset into ten folds randomly where nine of them were for training and one was for validation, rotated ten times. To evaluate the prediction performance, the mean RMSE and $R^2$ over folds was computed. The RF model as an instance had lower average cross validation RMSE and $R^2$ compared to the linear and quadratic regressors.

## 4. Main results

We generate the edge partition of molecular graphs using vertex degrees. Based on these partitions, some topological indices of diverse drugs were calculated, and then statistical models were developed to evaluate the relationship between structural information and physicochemical properties. Degree-based edge partition is shown in Table 2.

**Table 2.** Degree-based bond partitions of additional drug molecules.

| Graphs | Molecule | $(1,3)$ | $(2,2)$ | $(2,3)$ | $(3,3)$ | $(1,2)$ | $(1,4)$ | $(2,4)$ | $(3,4)$ |
|---|---|---|---|---|---|---|---|---|---|
| $L_1$ | Amitriptyline | 2 | 9 | 8 | 4 | 0 | 0 | 0 | 0 |
| $L_2$ | Carbamazepin | 2 | 7 | 6 | 5 | 0 | 0 | 0 | 0 |
| $L_3$ | Diazepam | 3 | 6 | 8 | 5 | 0 | 0 | 0 | 0 |
| $L_4$ | Donepezil | 1 | 6 | 18 | 4 | 2 | 0 | 0 | 0 |
| $L_5$ | Ergotamine | 4 | 9 | 17 | 12 | 0 | 2 | 3 | 3 |
| $L_6$ | Ethosuximide | 2 | 0 | 3 | 0 | 1 | 1 | 2 | 1 |
| $L_7$ | Fingolimod | 0 | 9 | 6 | 0 | 3 | 1 | 3 | 0 |
| $L_8$ | Galanthamine | 2 | 3 | 11 | 3 | 1 | 0 | 2 | 2 |
| $L_9$ | Lamotrigine | 4 | 3 | 6 | 4 | 0 | 0 | 0 | 0 |
| $L_{10}$ | Levetiracetam | 3 | 2 | 3 | 3 | 1 | 0 | 0 | 0 |
| $L_{11}$ | Lorazepam | 4 | 4 | 10 | 5 | 0 | 0 | 0 | 0 |
| $L_{12}$ | Phenytoin | 2 | 8 | 7 | 0 | 0 | 0 | 1 | 3 |
| $L_{13}$ | Pramipexole | 1 | 3 | 9 | 1 | 1 | 0 | 0 | 0 |
| $L_{14}$ | Propranolol | 3 | 7 | 8 | 2 | 0 | 0 | 0 | 0 |
| $L_{15}$ | Rivastigmine | 5 | 2 | 7 | 3 | 1 | 0 | 0 | 0 |
| $L_{16}$ | Rizatriptan | 2 | 6 | 12 | 2 | 0 | 0 | 0 | 0 |
| $L_{17}$ | Ropinirole | 1 | 5 | 10 | 2 | 2 | 0 | 0 | 0 |
| $L_{18}$ | Selegiline | 2 | 5 | 5 | 1 | 1 | 0 | 0 | 0 |
| $L_{19}$ | Sumatriptan | 2 | 3 | 9 | 2 | 1 | 2 | 2 | 0 |
| $L_{20}$ | Tacrine | 1 | 6 | 6 | 4 | 0 | 0 | 0 | 0 |
| $L_{21}$ | Teriflunomide | 3 | 2 | 7 | 2 | 1 | 3 | 0 | 1 |
| $L_{22}$ | Topiramate | 0 | 2 | 4 | 2 | 0 | 7 | 8 | 1 |

**Theorem 1.** *Suppose that $L_1$ is an amitriptyline molecular structure; then, the degree-based topological indices of this structure are*

$$\begin{aligned} M_1(L_1) &= 108, M_2(L_1) = 126, HG(L_1) = 10.0333333, F(L_1) = 268, SS(L_1) = 24.394591, \\ ABC(L_1) &= 16.320475, RI(L_1) = 10.25402, SC(L_1) = 10.710702, N(L_1) = 520. \end{aligned}$$

*Proof.* Assume that amitriptyline is denoted by $G_1$, and $E_{s,t}$ is a set of edges, linking vertices of degrees $s$ and $t$ within the graph. Between $s$ and $t$ vertices, frequencies $|E_{s,t}|$ indicate numbers of edges. $|E_{1,3}| = 2$

indicates three edges, linking vertices of degrees 1 and 3, and $|E_{2,2}| = 9$ indicates three edges linking vertices of degrees 2 and 2. Likewise $|E_{2,3}| = 8$, $|E_{3,3}| = 4$ . Then,
(a) Using the first Zagreb index and Table 2, we get

$$\begin{aligned} M_1(G) &= \sum_{mn\in E(G)} (F_m + F_n), \\ M_1(L_1) &= (1 + 3)(2) + (2 + 2)(9) + (2 + 3)(8) + (3 + 3)(4) = 108. \end{aligned}$$

(b) Using the second Zagreb index, we get

$$\begin{aligned} M_2(G) &= \sum_{mn\in E(G)} (F_m + F_n), \\ M_2(L_1) &= (1 \times 3)(2) + (2 \times 2)(9) + (2 \times 3)(8) + (3 \times 3)(4) = 126. \end{aligned}$$

By using Harmonic index and edge partition of $L_1$,

$$\begin{aligned} H(G) &= \sum_{mn\in E(G)} \frac{2}{(F_m + F_n)}, \\ H(L_1) &= \frac{2}{(1+3)}(2) + \frac{2}{(2+2)}(9) + \frac{2}{(2+3)}(8) \\ &+ \frac{2}{(3+3)}(4) = 10.033333. \end{aligned}$$

By using forgotten index and edge partition of $L_1$,

$$\begin{aligned} F(G) &= \sum_{mn\in E(G)} [(F_m)^2 + (F_n)^2], \\ F(L_1) &= (1^2 + 2^3)(2) + (2^2 + 2^2)(9) + (2^2 + 3^2)(8) + (3^2 + 3^2)(4) = 268. \end{aligned}$$

By using the Shilpa-Shanmukha index and edge partition of $L_1$,

$$\begin{aligned} SS(G) &= \sum_{mn\in E(G)} \sqrt{\frac{F_m F_n}{F_m + F_n}}, \\ SS(L_1) &= \sqrt{\frac{1\times 3}{(1+3)}}(2) + \sqrt{\frac{2\times 2}{(2+2)}}(9) + \sqrt{\frac{2\times 3}{(2+3)}}(8) + \sqrt{\frac{3\times 3}{(3+3)}}(4) = 24.394591. \end{aligned}$$

By using the atom bond connectivity index and edge partition of $L_1$,

$$\begin{aligned} ABC(G) &= \sum_{mn\in E(G)} \sqrt{\frac{F_m + F_n - 2}{F_m F_n}}, \\ ABC(L_1) &= \sqrt{\frac{1+3-2}{(1\times 3)}}(2) + \sqrt{\frac{2+2-2}{(2\times 2)}}(9) + \sqrt{\frac{2+3-2}{(2\times 3)}}(8) + \sqrt{\frac{3+3-2}{(3\times 3)}}(4) = 16.320475. \end{aligned}$$

By using the Randic index and edge partition of $L_1$,

$$\begin{aligned} RI(G) &= \sum_{mn\in E(G)} \frac{1}{F_m F_n}, \\ RI(L_1) &= \left(\frac{1}{1\times 3}\right)(2) + \left(\frac{1}{2\times 2}\right)(9) + \left(\frac{1}{2\times 3}\right)(8) + \left(\frac{1}{3\times 3}\right)(4) = 10.25402. \end{aligned}$$

By using the sum connectivity index and edge partition of $L_1$,

$$\begin{aligned} SC(G) &= \sum_{mn\in E(G)} \frac{1}{F_m + F_n}, \\ SC(L_1) &= \left(\frac{1}{1+3}\right)(2) + \left(\frac{1}{2+2}\right)(9) + \left(\frac{1}{2+3}\right)(8) + \left(\frac{1}{3+3}\right)(4) = 10.710702. \end{aligned}$$

By using the geometric arithmetic index and edge partition of $L_1$,

$$\begin{aligned} GA(G) &= \sum_{mn\in E(G)} \frac{2\sqrt{F_mF_n}}{F_m + F_n}, \\ GA(L_1) &= \frac{2\sqrt{13}}{(1+3)}(2) + \frac{2\sqrt{22}}{(2+2)}(9) + \frac{2\sqrt{23}}{(2+3)}(8) + \frac{2\sqrt{33}}{(3+3)}(4) = 22.570418. \end{aligned}$$

By using the Hyper Zagreb index and edge partition of $L_1$,

$$\begin{aligned} HZ(G) &= \sum_{mn\in E(G)} (\mathfrak{D}_m + \mathfrak{D}_n)^2, \\ HZ(L_1) &= (1+3)^2(2) + (2+2)^2(9) + (2+3)^2(8) + (3+3)^2(4) = 520. \end{aligned}$$

By using the Nirmala index and edge partition of $L_1$,

$$\begin{aligned} N(G) &= \sum_{mn\in E(G)} \sqrt{(F_m + F_n)}, \\ N(L_1) &= \sqrt{(1+3)}(2) + \sqrt{(2+2)}(9) + \sqrt{(2+3)}(8) + \sqrt{(3+3)}(4) = 49.686503. \end{aligned}$$

Similarly, we compoute different topological indices for each drug, as shown in Table 3.

Table 3 presents various drug names along with their corresponding data across multiple parameters, such as $M_1, M_2$, H, F, and others. These values represent specific characteristics and properties of the drugs, including their chemical compositions and pharmacological effects. Notably, each drug has different values for these parameters, which may correlate to their efficacy or other functional characteristics. The inclusion of columns like ABC, RI, SC, GA, HZ, and N further adds depth to the analysis, showing additional measurements related to the drugs. This dataset provides a comprehensive view for comparing and contrasting the drugs based on various quantitative metrics. Such a table is useful for researchers and practitioners in fields like pharmacology and medicinal chemistry to analyze drug profiles.

**Table 3.** Topological indices for the structural analysis of antibiotic molecules.

| Drug name | $M_1$ | $M_2$ | H | F | SS | ABC | RI | SC | GA | HZ | N |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Amitriptyline | 108 | 126 | 10.03 | 268 | 24.39 | 16.32 | 10.25 | 10.71 | 22.57 | 520 | 49.69 |
| Carbamazepine | 96 | 115 | 8.57 | 244 | 21.43 | 14.16 | 8.77 | 9.22 | 19.61 | 474 | 43.66 |
| Diazepam | 106 | 126 | 9.37 | 272 | 23.49 | 15.68 | 9.66 | 10.12 | 21.44 | 524 | 48.14 |
| Donepezil | 148 | 175 | 13.37 | 374 | 33.12 | 21.87 | 13.67 | 14.34 | 30.39 | 724 | 67.51 |
| Ergotamine | 258 | 326 | 19.96 | 718 | 54.96 | 35.44 | 20.68 | 22.25 | 48.52 | 1370 | 113.16 |
| Ethosuximide | 50 | 58 | 4.22 | 146 | 10.35 | 7.39 | 4.58 | 4.56 | 9.29 | 262 | 22.22 |
| Fingolimod | 98 | 106 | 10.3 | 242 | 22.38 | 15.72 | 10.63 | 10.59 | 21.34 | 454 | 46.20 |
| Galantamine | 122 | 153 | 9.80 | 336 | 26.20 | 16.94 | 10.14 | 10.79 | 23.32 | 642 | 53.87 |
| Lamotrigine | 82 | 96 | 7.23 | 214 | 17.94 | 12.30 | 7.59 | 7.82 | 16.34 | 406 | 37.21 |
| Levetiracetam | 56 | 64 | 5.37 | 144 | 12.38 | 8.69 | 5.66 | 5.64 | 11.48 | 272 | 25.79 |
| Lorazepam | 112 | 133 | 9.67 | 292 | 24.54 | 16.50 | 10.06 | 10.51 | 22.26 | 558 | 50.61 |
| Phenytoin | 102 | 124 | 8.99 | 270 | 22.48 | 14.88 | 9.23 | 9.67 | 20.50 | 518 | 46.04 |
| Pramipexole | 70 | 80 | 6.6 | 174 | 15.77 | 10.68 | 6.79 | 7.01 | 14.63 | 334 | 32.31 |
| Propranolol | 92 | 103 | 8.87 | 226 | 20.81 | 14.39 | 9.16 | 9.39 | 19.44 | 432 | 42.79 |
| Rivastigmine | 84 | 94 | 7.97 | 216 | 18.49 | 13.15 | 8.45 | 8.43 | 17.13 | 404 | 38.73 |
| Rizatriptan | 104 | 120 | 9.47 | 260 | 23.33 | 15.69 | 9.72 | 10.18 | 21.49 | 500 | 47.73 |
| Ropinirole | 92 | 105 | 9 | 226 | 20.90 | 14.17 | 9.24 | 9.44 | 19.55 | 436 | 42.72 |
| Selegiline | 62 | 67 | 6.5 | 148 | 14.25 | 10.08 | 6.74 | 6.72 | 13.57 | 282 | 29.36 |
| Sumatriptan | 102 | 116 | 8.9 | 276 | 21.96 | 15.30 | 9.41 | 9.63 | 20.00 | 508 | 46.13 |
| Tacrine | 82 | 99 | 7.23 | 208 | 18.34 | 11.97 | 7.36 | 7.82 | 16.74 | 406 | 37.21 |
| Teriflunomide | 92 | 103 | 8.12 | 254 | 19.52 | 14.10 | 8.75 | 8.74 | 17.79 | 460 | 41.64 |
| Topiramate | 130 | 154 | 9.02 | 408 | 25.64 | 17.94 | 9.92 | 10.38 | 22.05 | 716 | 55.74 |

## 5. Linear and quardatic regression models

Linear regression is a statistical method used to model the relationship between a dependent variable $y$ and an independent variable $x$ using a straight line. The general equation of a simple linear regression model is given by $y = b_0 + b_1 x + \varepsilon$, where $b_0$ is the intercept, $b_1$ is the slope, and $\varepsilon$ represents the error term. This model is suitable when the data shows a linear trend, and it is commonly used in predictive modeling and trend analysis.

Quadratic regression, on the other hand, is used when the data exhibits a curved or nonlinear relationship. It fits a second-degree polynomial to the data and is expressed as $y = a_0 + a_1 x + a_2 x^2 + \varepsilon$, where $a_0$, $a_1$, and $a_2$ are coefficients, and $a_2$ controls the curvature of the parabola. Quadratic regression is appropriate in scenarios where changes in the data accelerate or decelerate, such as modeling projectile motion, business profits over time, or economic behaviors.

Choosing between linear and quadratic regression depends on the nature of the data. While linear models are simpler, easier to interpret, and less prone to overfitting, they may underperform if the data exhibits curvature. Quadratic models provide more flexibility and can capture nonlinear patterns, but they also risk overfitting, especially with small or noisy datasets. Root mean square error (RMSE) is a statistical measure used to evaluate the accuracy of a model's predictions. It is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

where:

- $y_i$ = observed (actual) value,
- $\hat{y}_i$ = predicted value,
- $n$ = total number of observations.

A lower RMSE indicates that the model's predictions are closer to the observed values, implying better performance. Therefore, it is essential to visualize the data and evaluate model performance using residual plots, the coefficient of determination ($R^2$), and error metrics like RMSE to make informed modeling decisions.

### *5.1. Regression models for first Zagreb index $M_1(G)$*

$$\begin{aligned}
\textit{Boiling Point} &= 163.4195 + 2.7858M_1,\\
\textit{Boiling Point} &= 215.5500 + 1.9441M_1 + 0.0028M_1^2,\\
\textit{Enthalpy of Vaporization} &= 33.3482 + 0.3836M_1,\\
\textit{Enthalpy of Vaporization} &= 46.2805 + 0.1749M_1 + 0.0007M_1^2,\\
\textit{Flash Point} &= 55.5113 + 1.6789M_1,\\
\textit{Flash Point} &= 82.2175 + 1.2478M_1 + 0.0014M_1^2,\\
\textit{Molar Refractivity} &= 20.8680 + 0.5483M_1,\\
\textit{Molar Refractivity} &= 10.5495 + 0.7149M_1 - 0.0006M_1^2,\\
\textit{Polarizability} &= 8.2696 + 0.2174M_1,\\
\textit{Polarizability} &= 4.1677 + 0.2836M_1 - 0.0002M_1^2,\\
\textit{Molar Volume} &= 98.6920 + 1.2320M_1,\\
\textit{Molar Volume} &= 47.0906 + 2.0650M_1 - 0.0028M_1^2.
\end{aligned}$$

The regression equations show both linear and quadratic models for predicting physical properties using $M_1$. Quadratic models include $M_1^2$ terms, allowing better capture of nonlinear trends in the data. For properties like BP and FP, the quadratic terms slightly improve fit. The coefficients of $M_1^2$ are small, suggesting minor but relevant curvature in some relationships. Quadratic models may enhance accuracy, but should be chosen when the nonlinear behavior is statistically significant.

Table 4 provides a detailed comparison of linear and quadratic regression models applied to six physicochemical properties: Boiling point (BP), enthalpy of vaporization (EV), flash point (FP), molar refraction (MR), polarizability (P), and molar volume (MV). Across all properties, the quadratic models generally exhibit slightly higher correlation coefficients ($R$) and coefficients of determination ($R^2$), indicating a marginally improved fit. For instance, the $R^2$ for enthalpy of vaporization improves from 0.825 in the linear model to 0.840 in the quadratic model. However, in several cases, this improved fit comes with a notable increase in the standard error of the slope (SE), such as in BP (from 0.299 to 1.251) and FP (from 0.212 to 0.893), suggesting that quadratic models may introduce variability and potential overfitting. Despite these trade-offs, all models yield statistically significant results, as evidenced by extremely low p-values (ranging from $10^{-6}$ to $10^{-10}$) and high F-statistics. Interestingly, the linear models often report higher F-statistics compared to their quadratic counterparts, implying stronger explanatory power with fewer parameters. Overall, while quadratic models provide marginal improvements in $R^2$, the increase in complexity and variability must be carefully weighed against the benefits.

**Table 4.** Statistical parameters of $M_1$ using linear and quadratic models.

| Dependent variable | Model type | R | R$^2$ | SE (slope) | F-statistic | P-value |
|---|---|---|---|---|---|---|
| Boiling Point | Linear | 0.901 | 0.813 | 0.299 | 86.75 | $1.03 \times 10^{-8}$ |
| | Quadratic | 0.904 | 0.817 | 1.251 | 42.49 | $9.71 \times 10^{-8}$ |
| Enthalpy of Vaporization | Linear | 0.908 | 0.825 | 0.039 | 94.45 | $5.10 \times 10^{-9}$ |
| | Quadratic | 0.917 | 0.840 | 0.160 | 50.05 | $2.67 \times 10^{-8}$ |
| Flash Point | Linear | 0.871 | 0.758 | 0.212 | 62.62 | $1.38 \times 10^{-7}$ |
| | Quadratic | 0.872 | 0.761 | 0.893 | 30.26 | $1.24 \times 10^{-6}$ |
| Molar Refraction | Linear | 0.936 | 0.876 | 0.046 | 140.92 | $1.65 \times 10^{-10}$ |
| | Quadratic | 0.938 | 0.881 | 0.192 | 70.16 | $1.68 \times 10^{-9}$ |
| Polarizability | Linear | 0.936 | 0.876 | 0.018 | 140.83 | $1.65 \times 10^{-10}$ |
| | Quadratic | 0.938 | 0.881 | 0.076 | 70.14 | $1.69 \times 10^{-9}$ |
| Molar Volume | Linear | 0.843 | 0.710 | 0.176 | 48.96 | $8.64 \times 10^{-7}$ |
| | Quadratic | 0.855 | 0.730 | 0.719 | 25.71 | $3.93 \times 10^{-6}$ |

Figure 2 shows the relationships between various variables, each shown with scatterplots. The red dots represent the data points, while the green and blue lines represent linear and quadratic fits, respectively. Each subplot compares the variable $M_1$ with other variables, such as BP, EV, FP, MR, P, and MV. The linear and quadratic trends are compared to determine the best fit for the data. Overall, these plots suggest a mostly linear relationship between $M_1$ and the other variables.

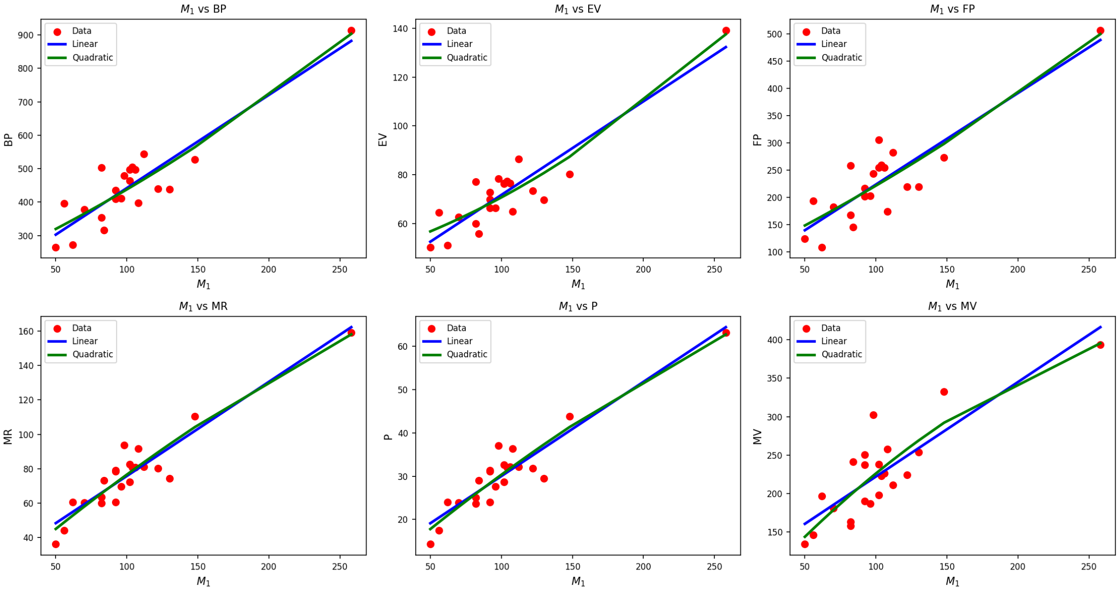


**Figure 2.** Graphical analysis of $M_1$ based linear and quadratic regression model.

### 5.2. *Regression model for second Zagreb index $M_2(G)$*

$$\textit{Boiling Point} \ = 191.5125 + 2.1356 M_2,$$

$$
\begin{aligned}
\textit{Boiling Point} &= 227.2052 + 1.6529M_2 + 0.0013M_2^2,\\
\textit{Enthalpy of Vaporization} &= 37.1467 + 0.2947M_2,\\
\textit{Enthalpy of Vaporization} &= 47.3927 + 0.1561M_2 + 0.0004M_2^2,\\
\textit{Flash Point} &= 71.6864 + 1.2934M_2,\\
\textit{Flash Point} &= 86.0835 + 1.0987M_2 + 0.0005M_2^2,\\
\textit{Molar Refractivity} &= 27.2651 + 0.4131M_2,\\
\textit{Molar Refractivity} &= 18.4146 + 0.5328M_2 - 0.0003M_2^2,\\
\textit{Polarizability} &= 10.8064 + 0.1638M_2,\\
\textit{Polarizability} &= 7.2864 + 0.2114M_2 - 0.0001M_2^2,\\
\textit{Molar Volume} &= 115.3910 + 0.9089M_2,\\
\textit{Molar Volume} &= 76.1894 + 1.4390M_2 - 0.0014M_2^2.
\end{aligned}
$$

Both linear and quadratic models are used to correlate various molecular properties with $M_2$. Quadratic models include $M_2^2$ terms, which provide improved flexibility to model curvature. In most cases, the quadratic term has a small coefficient, indicating subtle nonlinear effects. For properties like MV and FP, the quadratic fit shows noticeable changes in slope and intercept. These models suggest that while linear trends are dominant, quadratic terms can refine predictions when needed.

Table 5 compares linear and quadratic regression models for six dependent variables. Overall, quadratic models show slightly higher $R$ and $R^2$ values, suggesting marginally improved fits for example, Pimproves from $R^2 = 0.841$ to $R^2 = 0.845$ . However, this improvement often results in a higher standard error, such as for BP (0.233 to 0.959) and FP (0.162 to 0.671), indicating potential overfitting. Despite this, all models are statistically significant, with p-values well below $10^{-5}$ and high F-statistics. Linear models generally retain stronger simplicity and robustness, while quadratic models offer minor gains in accuracy at the cost of increased variability.

**Table 5.** Statistical parameters of $M_2$ using linear and quadratic models.

| Dependent variable | Model type | R | R$^2$ | SE (slope) | F-statistic | P-value |
|---|---|---|---|---|---|---|
| Boiling Point | Linear | 0.899 | 0.808 | 0.233 | 83.98 | $1.34 \times 10^{-8}$ |
| | Quadratic | 0.900 | 0.810 | 0.959 | 40.59 | $1.38 \times 10^{-7}$ |
| Enthalpy of Vaporization | Linear | 0.907 | 0.823 | 0.031 | 93.29 | $5.65 \times 10^{-9}$ |
| | Quadratic | 0.914 | 0.835 | 0.122 | 48.20 | $3.61 \times 10^{-8}$ |
| Flash Point | Linear | 0.872 | 0.761 | 0.162 | 63.57 | $1.23 \times 10^{-7}$ |
| | Quadratic | 0.873 | 0.762 | 0.671 | 30.38 | $1.20 \times 10^{-6}$ |
| Molar Refraction | Linear | 0.917 | 0.841 | 0.040 | 105.54 | $2.01 \times 10^{-9}$ |
| | Quadratic | 0.919 | 0.845 | 0.164 | 51.91 | $2.00 \times 10^{-8}$ |
| Polarizability | Linear | 0.917 | 0.841 | 0.016 | 105.45 | $2.02 \times 10^{-9}$ |
| | Quadratic | 0.919 | 0.845 | 0.065 | 51.87 | $2.01 \times 10^{-8}$ |
| Molar Volume | Linear | 0.808 | 0.653 | 0.148 | 37.71 | $5.32 \times 10^{-6}$ |
| | Quadratic | 0.817 | 0.668 | 0.600 | 19.11 | $2.83 \times 10^{-5}$ |

Figure 3 shows a comparison of the variable $M_2$ with other variables. Red dots represent the data points, while the green and blue lines represent linear and quadratic fits, respectively. Each plot compares $M_2$ with variables such as BP, EV, FP, MR, P, and MV. Both linear and quadratic models are used to evaluate the best fit for the data. The relationships between $M_2$ and the other variables are generally linear, with slight curvature suggested by the quadratic fits. In most cases, the linear trend appears to be the dominant pattern in the data. The inclusion of the quadratic fit helps assess whether nonlinear trends provide a better fit.

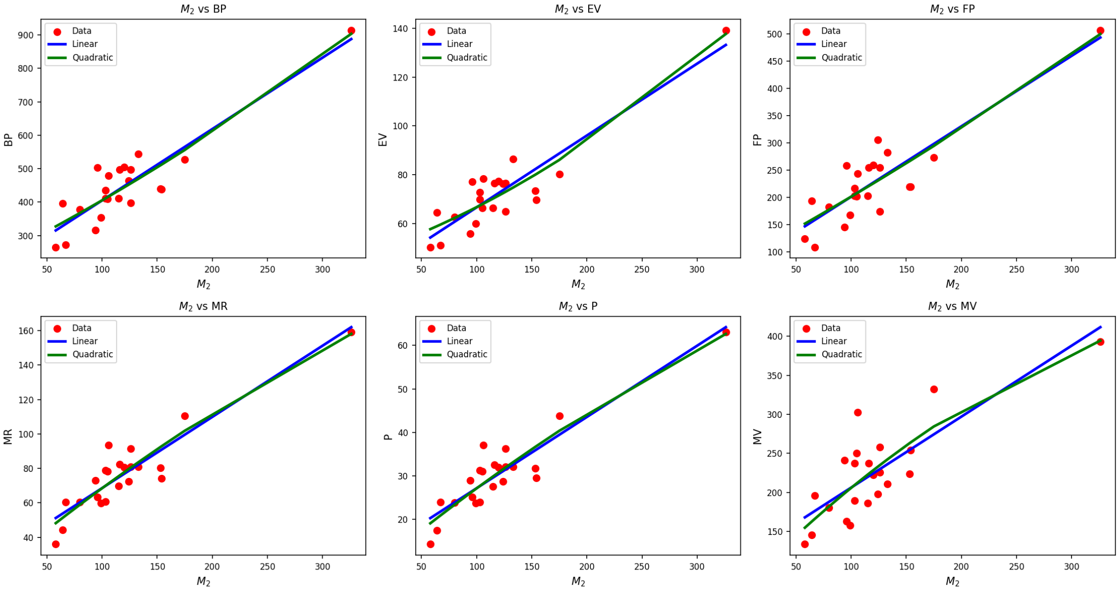


**Figure 3.** Graphical analysis of $M_2$ based linear and quadratic regression model.

### 5.3. *Regression models for Harmonic index $H(G)$*

$$
\begin{aligned}
\textit{Boiling Point} &= 110.4146 + 37.4150H,\\
\textit{Boiling Point} &= 226.9282 + 15.2357H + 0.9241H^2,\\
\textit{Enthalpy of Vaporization} &= 26.1198 + 5.1448H,\\
\textit{Enthalpy of Vaporization} &= 48.2924 + 0.9241H + 0.1758H^2,\\
\textit{Flash Point} &= 24.7970 + 22.4131H,\\
\textit{Flash Point} &= 94.0707 + 9.2263H + 0.5494H^2,\\
\textit{Molar Refractivity} &= 6.2394 + 7.8292H,\\
\textit{Molar Refractivity} &= -3.2046 + 9.6269H - 0.0749H^2,\\
\textit{Polarizability} &= 2.4702 + 3.1039H,\\
\textit{Polarizability} &= -1.2789 + 3.8176H - 0.0297H^2,\\
\textit{Molar Volume} &= 61.9142 + 18.0243H,\\
\textit{Molar Volume} &= 2.7301 + 29.2904H - 0.4694H^2.
\end{aligned}
$$

The regression models express molecular properties as functions of $H$ using both linear and quadratic forms. Quadratic models incorporate $H^2$ terms, capturing nonlinear relationships more effectively. For BP and FP, the quadratic terms significantly enhance the model fit. In some cases, such as MV and MR, the quadratic term notably alters the trend direction. These results indicate that quadratic models are valuable when linear approximations do not fully capture property variations.

Table 6 presents regression statistics comparing linear and quadratic models across six physicochemical properties. In general, quadratic models yield slightly higher $R$ and $R^2$ values, such as for EV ($R^2$ improves from 0.807 to 0.838), indicating better fit. However, the SE of the slope significantly increases in the quadratic models, particularly for BP (from 4.218 to 17.87) and FP (from 3.012 to 12.93), suggesting increased variability and possible overfitting. Despite this, all models are statistically significant, with very low p-values (as low as $10^{-16}$) and high F-statistics. Thus, while quadratic models offer slight gains in predictive strength, they may do so at the cost of model stability and simplicity.

**Table 6.** Statistical parameters of $H$ using linear and quadratic models.

| Dependent variable | Model type | R | $R^2$ | SE (slope) | F-statistic | P-value |
|---|---|---|---|---|---|---|
| Boiling Point | Linear | 0.893 | 0.797 | 4.218 | 78.70 | $2.28 \times 10^{-8}$ |
| | Quadratic | 0.902 | 0.813 | 17.87 | 41.40 | $1.19 \times 10^{-7}$ |
| Enthalpy of Vaporization | Linear | 0.898 | 0.807 | 0.562 | 83.76 | $1.37 \times 10^{-8}$ |
| | Quadratic | 0.916 | 0.838 | 2.274 | 49.24 | $3.05 \times 10^{-8}$ |
| Flash Point | Linear | 0.857 | 0.735 | 3.012 | 55.39 | $3.49 \times 10^{-7}$ |
| | Quadratic | 0.866 | 0.749 | 12.93 | 28.38 | $1.96 \times 10^{-6}$ |
| Molar Refraction | Linear | 0.985 | 0.971 | 0.302 | 673.47 | $7.11 \times 10^{-17}$ |
| | Quadratic | 0.987 | 0.974 | 1.263 | 357.04 | $8.50 \times 10^{-16}$ |
| Polarizability | Linear | 0.985 | 0.971 | 0.120 | 671.43 | $7.32 \times 10^{-17}$ |
| | Quadratic | 0.987 | 0.974 | 0.501 | 355.95 | $8.74 \times 10^{-16}$ |
| Molar Volume | Linear | 0.909 | 0.827 | 1.846 | 95.37 | $4.71 \times 10^{-9}$ |
| | Quadratic | 0.919 | 0.845 | 7.704 | 51.82 | $2.02 \times 10^{-8}$ |

Figure 4 shows the relationships between the variable $H$ and five different dependent variables: BP, EV, FP, MR, P, and MV. Each plot includes the raw data points represented by red dots. In addition, two regression models are overlaid: a linear model (blue) and a quadratic model (green). The first plot examines the relationship between $H$ and BP, where both the linear and quadratic models fit the data closely. The second plot shows $H$ vs EV, with the quadratic model providing a slightly better fit. The third plot compares $H$ with FP, where the linear model performs well. The fourth plot examines the relationship between $H$ and MR, showing a strong linear trend. The fifth plot represents $H$ vs P, where both models are close but the quadratic offers a slightly more accurate representation. Finally, the sixth plot compares $H$ and MV, where both the linear and quadratic models fit the data similarly.

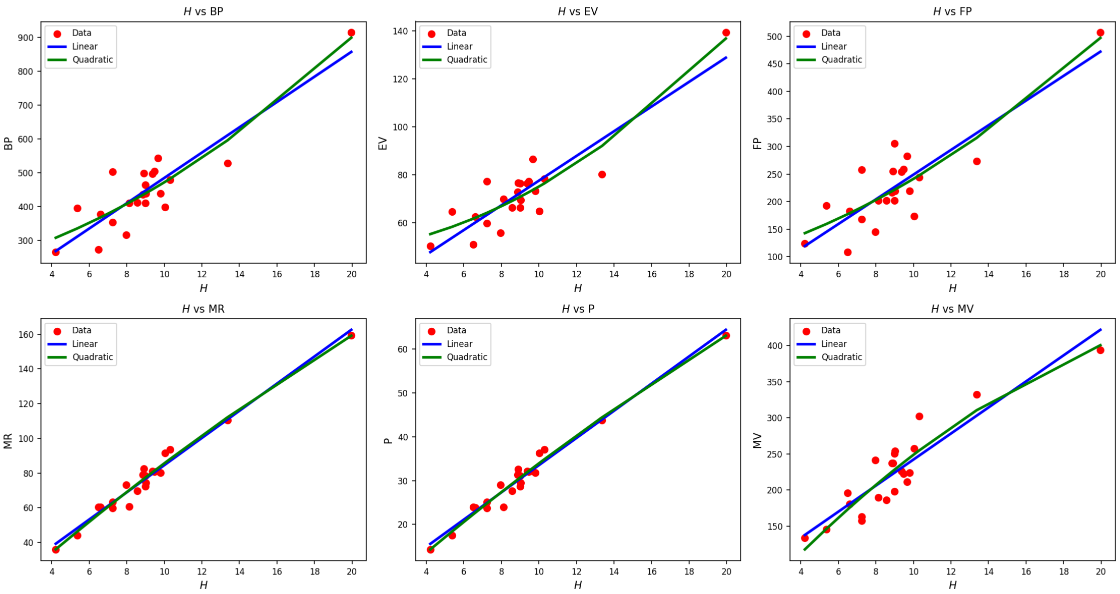


**Figure 4.** Graphical analysis of $H$.

### 5.4. *Regression models for Forgotten index $F(G)$*

$$\textit{Boiling Point} = 195.6272 + 0.9388F,$$

$$
\begin{aligned}
\textit{Boiling Point} &= 250.4644 + 0.6048F + 0.0004F^2,\\
\textit{Enthalpy of Vaporization} &= 37.6755 + 0.1297F,\\
\textit{Enthalpy of Vaporization} &= 50.3598 + 0.0524F + 0.0001F^2,\\
\textit{Flash Point} &= 74.2388 + 0.5683F,\\
\textit{Flash Point} &= 99.4214 + 0.4150F + 0.0002F^2,\\
\textit{Molar Refractivity} &= 28.6945 + 0.1792F,\\
\textit{Molar Refractivity} &= 24.6319 + 0.2040F - 0.0000F^2,\\
\textit{Polarizability} &= 11.3719 + 0.0711F,\\
\textit{Polarizability} &= 9.7505 + 0.0809F - 0.0000F^2,\\
\textit{Molar Volume} &= 115.8401 + 0.4044F,\\
\textit{Molar Volume} &= 80.4600 + 0.6198F - 0.0003F^2.
\end{aligned}
$$

The regression models describe various molecular properties as functions of temperature ($F$) using both linear and quadratic forms. Quadratic models include $F^2$ terms, allowing for more precise representation of nonlinear temperature dependencies. For properties such as BP and FP, the quadratic terms significantly improve the model fit and accuracy.

In cases like MV and MR, the quadratic terms influence the trend direction, showing more complex behavior. These results demonstrate the value of quadratic models in capturing variations that linear models fail to represent adequately.

Table 7 compares the performance of linear and quadratic regression models for various dependent variables using several statistical metrics. Across all variables, both models exhibit high correlation coefficients (R), with quadratic models slightly outperforming linear ones in terms of $R^2$, particularly for EV and FP. The SE of the slope is consistently lower in linear models, indicating more stable coefficient estimates. F-statistics and P-values confirm the statistical significance of all models, although linear models often have stronger F-statistics and lower P-values. Overall, while quadratic models offer minor improvements in fit, linear models demonstrate more stable and statistically robust performance.

**Table 7.** Statistical parameters of $F$ using linear and quadratic models.

| Dependent variable | Model type | R | $R^2$ | SE (slope) | F-statistic | P-value |
|---|---|---|---|---|---|---|
| Boiling Point | Linear | 0.878 | 0.770 | 0.115 | 67.01 | $8.16 \times 10^{-8}$ |
| | Quadratic | 0.881 | 0.776 | 0.484 | 32.93 | $6.69 \times 10^{-7}$ |
| Enthalpy of Vaporization | Linear | 0.887 | 0.787 | 0.015 | 73.88 | $3.78 \times 10^{-8}$ |
| | Quadratic | 0.897 | 0.804 | 0.062 | 38.97 | $1.89 \times 10^{-7}$ |
| Flash Point | Linear | 0.851 | 0.725 | 0.078 | 52.66 | $5.08 \times 10^{-7}$ |
| | Quadratic | 0.853 | 0.728 | 0.333 | 25.42 | $4.25 \times 10^{-6}$ |
| Molar Refraction | Linear | 0.884 | 0.781 | 0.021 | 71.31 | $5.00 \times 10^{-8}$ |
| | Quadratic | 0.884 | 0.782 | 0.091 | 34.05 | $5.22 \times 10^{-7}$ |
| Polarizability | Linear | 0.884 | 0.781 | 0.008 | 71.30 | $5.01 \times 10^{-8}$ |
| | Quadratic | 0.884 | 0.782 | 0.036 | 34.05 | $5.22 \times 10^{-7}$ |
| Molar Volume | Linear | 0.799 | 0.638 | 0.068 | 35.30 | $8.25 \times 10^{-6}$ |
| | Quadratic | 0.806 | 0.649 | 0.287 | 17.59 | $4.74 \times 10^{-5}$ |

Figure 5 shows the relationships between the variable $F$ and five different dependent variables: BP, EV, FP, MR, P, and MV. Each plot displays the raw data points as red dots and overlays two regression models: A linear model (blue) and a quadratic model (green). The first plot examines $F$ vs BP, where both models fit the data well, though the quadratic model provides a closer fit. The second plot shows $F$ vs EV, with the quadratic model offering a better fit for the data. The third plot compares $F$ with FP, where both models align closely, but the quadratic model has a slight advantage. The fourth plot shows the relationship between $F$ and MR, where the linear model fits the data best. The fifth plot represents $F$ vs P, showing both models fit well, with the quadratic model offering a marginally better fit. Finally, the sixth plot compares $F$ and MV, where the linear and quadratic models fit the data similarly.

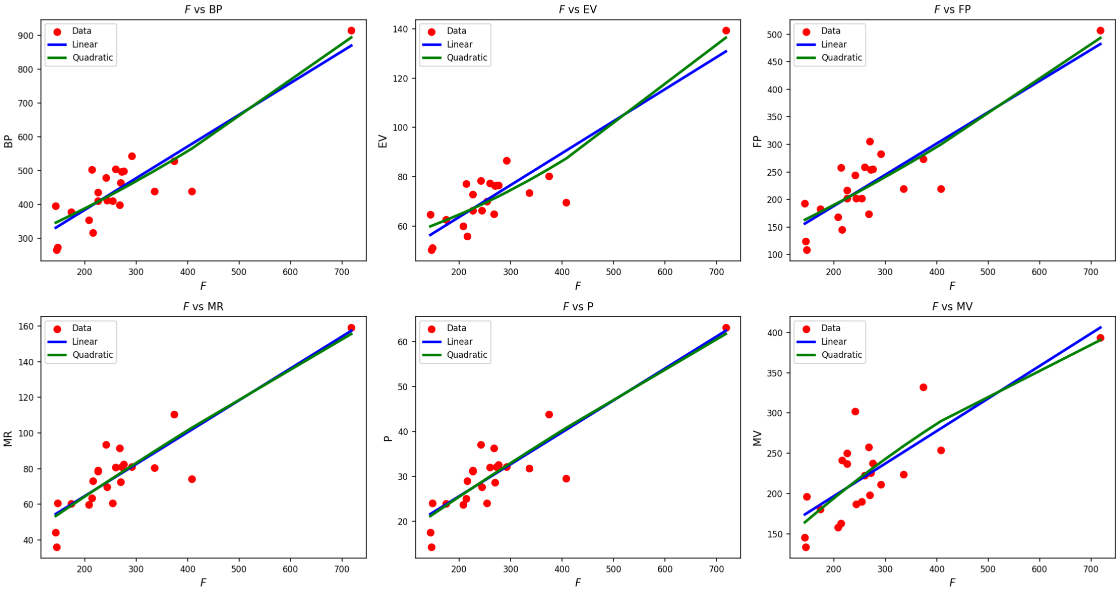


**Figure 5.** Graphical analysis of $F$.

### 5.5. *Regression models for Shilpa-Shanmukha index $SS(G)$*

$$\begin{aligned}
\textit{Boiling Point} &= 150.9521 + 13.2681SS,\\
\textit{Boiling Point} &= 210.4759 + 8.8450SS + 0.0687SS^2,\\
\textit{Enthalpy of Vaporization} &= 31.6868 + 1.8248SS,\\
\textit{Enthalpy of Vaporization} &= 45.7508 + 0.7797SS + 0.0162SS^2,\\
\textit{Flash Point} &= 48.2711 + 7.9843SS,\\
\textit{Flash Point} &= 80.6416 + 5.5789SS + 0.0373SS^2,\\
\textit{Molar Refractivity} &= 17.3280 + 2.6600SS,\\
\textit{Molar Refractivity} &= 5.5635 + 3.5342SS - 0.0136SS^2,\\
\textit{Polarizability} &= 6.8665 + 1.0546SS,\\
\textit{Polarizability} &= 2.1930 + 1.4018SS - 0.0054SS^2,\\
\textit{Molar Volume} &= 90.9461 + 5.9674SS,\\
\textit{Molar Volume} &= 36.7081 + 9.9977SS - 0.0626SS^2.
\end{aligned}$$

The regression models presented for various molecular properties are expressed as functions of a new variable ($SS$) using both linear and quadratic forms. The quadratic models include $SS^2$ terms, which better capture the nonlinear nature of these properties.For properties like BP and FP, the quadratic terms significantly improve the model's accuracy and fit.

In properties such as MV and MR, the quadratic term not only adjusts the trend but also introduces a more complex relationship. These observations emphasize the importance of quadratic models in providing more accurate predictions when linear models fail to capture detailed property variations.

Table 8 presents regression results showing improved performance of both linear and quadratic models, with consistently high $R$ and $R^2$ values across all dependent variables. Quadratic models slightly outperform linear ones in fit, especially for EV and MR, with $R^2$ values reaching up to 0.923. SEs for the slopes are notably higher in quadratic models, indicating greater variability in the coefficient estimates despite better fit. F-statistics are consistently higher in linear models, suggesting stronger explanatory power, and P-values remain highly significant in all cases, confirming model validity. Overall, while quadratic models offer marginally better fit, linear models remain more statistically robust with lower variance and greater interpretability.

**Table 8.** Statistical parameters of $SS$ using linear and quadratic models.

| Dependent variable | Model type | R | $R^2$ | SE (slope) | F-statistic | P-value |
|---|---|---|---|---|---|---|
| Boiling Point | Linear | 0.906 | 0.820 | 1.390 | 91.09 | $6.89 \times 10^{-9}$ |
| | Quadratic | 0.909 | 0.826 | 5.767 | 45.00 | $6.20 \times 10^{-8}$ |
| Enthalpy of Vaporization | Linear | 0.911 | 0.830 | 0.184 | 97.94 | $3.77 \times 10^{-9}$ |
| | Quadratic | 0.921 | 0.848 | 0.737 | 52.82 | $1.73 \times 10^{-8}$ |
| Flash Point | Linear | 0.873 | 0.762 | 0.997 | 64.18 | $1.14 \times 10^{-7}$ |
| | Quadratic | 0.876 | 0.767 | 4.163 | 31.23 | $9.86 \times 10^{-7}$ |
| Molar Refraction | Linear | 0.957 | 0.917 | 0.179 | 220.16 | $2.94 \times 10^{-12}$ |
| | Quadratic | 0.961 | 0.923 | 0.727 | 113.81 | $2.65 \times 10^{-11}$ |
| Polarizability | Linear | 0.957 | 0.917 | 0.071 | 219.89 | $2.97 \times 10^{-12}$ |
| | Quadratic | 0.961 | 0.923 | 0.288 | 113.69 | $2.68 \times 10^{-11}$ |
| Molar Volume | Linear | 0.861 | 0.741 | 0.789 | 57.20 | $2.74 \times 10^{-7}$ |
| | Quadratic | 0.873 | 0.762 | 3.187 | 30.45 | $1.19 \times 10^{-6}$ |

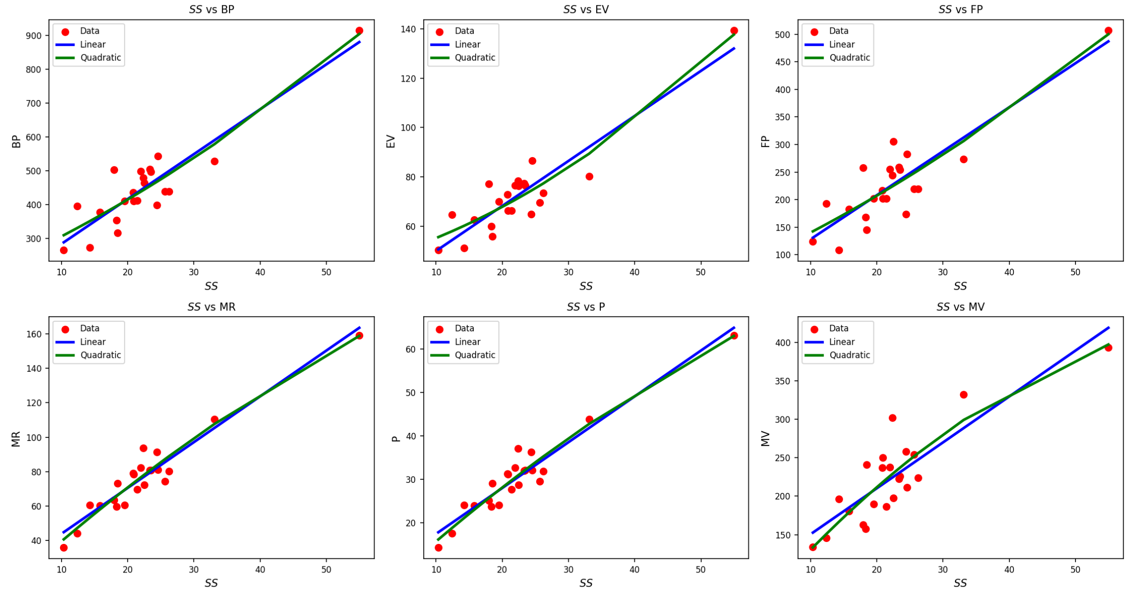


**Figure 6.** Graphical analysis of $SS$.

Figure 6 shows the relationships between the variable $SS$ and five dependent variables: BP, EV, FP, MR, P, and MV. The data points are displayed as red dots, with two regression models overlaid: A linear model (blue) and a quadratic model (green). The first plot examines $SS$ vs BP, where both the linear and quadratic models provide a good fit, but the quadratic model fits slightly better. The second plot shows $SS$ vs EV, with the quadratic model offering a closer fit to the data. The third plot compares $SS$ with FP, where both models show a similar fit, with the quadratic model slightly outperforming the linear one. The fourth plot illustrates the relationship between $SS$ and MR, where the linear model fits the data well. The fifth plot represents $SS$ vs P, with both models fitting closely, but the quadratic model provides a better fit. The final plot compares $SS$ and MV, where both models fit the data well, but the quadratic model gives a slight edge.

### *5.6. Regression models for atom bond connectivity index $ABC(G)$*

$$\begin{aligned}
\textit{Boiling Point} &= 128.9724 + 21.0589ABC,\\
\textit{Boiling Point} &= 207.0599 + 12.3940ABC + 0.2055ABC^2,\\
\textit{Enthalpy of Vaporization} &= 28.6615 + 2.8964ABC,\\
\textit{Enthalpy of Vaporization} &= 45.5671 + 1.0205ABC + 0.0445ABC^2,\\
\textit{Flash Point} &= 35.6026 + 12.6357ABC,\\
\textit{Flash Point} &= 80.5888 + 7.6439ABC + 0.1184ABC^2,\\
\textit{Molar Refractivity} &= 12.5878 + 4.2440ABC,\\
\textit{Molar Refractivity} &= 1.4570 + 5.4791ABC - 0.0293ABC^2,\\
\textit{Polarizability} &= 4.9867 + 1.6825ABC,\\
\textit{Polarizability} &= 0.5640 + 2.1733ABC - 0.0116ABC^2,\\
\textit{Molar Volume} &= 77.7583 + 9.6893ABC,\\
\textit{Molar Volume} &= 15.8164 + 16.5625ABC - 0.1630ABC^2.
\end{aligned}$$

The regression models for various molecular properties are presented as functions of $ABC$, using both linear and quadratic forms. For properties like BP and FP, the quadratic models significantly improve the prediction accuracy by including $ABC^2$ terms. The quadratic terms, particularly in MR, and MV, adjust the trend direction, revealing more complex dependencies on $ABC$.

The linear models, while simpler, may not fully capture these complex behaviors as seen in P and EV. These results highlight the importance of quadratic terms when modeling molecular properties that exhibit nonlinear variations with $ABC$.

**Table 9.** Statistical parameters of $ABC$ using linear and quadratic models.

| Dependent variable | Model type | R | $R^2$ | SE (slope) | F-statistic | P-value |
|---|---|---|---|---|---|---|
| Boiling Point | Linear | 0.904 | 0.817 | 2.231 | 89.11 | $8.26 \times 10^{-9}$ |
| | Quadratic | 0.908 | 0.825 | 9.437 | 44.77 | $6.46 \times 10^{-8}$ |
| Enthalpy of Vaporization | Linear | 0.910 | 0.827 | 0.296 | 95.75 | $4.55 \times 10^{-9}$ |
| | Quadratic | 0.921 | 0.848 | 1.202 | 52.95 | $1.70 \times 10^{-8}$ |
| Flash Point | Linear | 0.869 | 0.755 | 1.610 | 61.63 | $1.56 \times 10^{-7}$ |
| | Quadratic | 0.873 | 0.762 | 6.867 | 30.42 | $1.20 \times 10^{-6}$ |
| Molar Refraction | Linear | 0.961 | 0.923 | 0.275 | 238.50 | $1.40 \times 10^{-12}$ |
| | Quadratic | 0.963 | 0.927 | 1.153 | 121.14 | $1.53 \times 10^{-11}$ |
| Polarizability | Linear | 0.961 | 0.923 | 0.109 | 238.31 | $1.42 \times 10^{-12}$ |
| | Quadratic | 0.963 | 0.927 | 0.457 | 121.07 | $1.54 \times 10^{-11}$ |
| Molar Volume | Linear | 0.879 | 0.772 | 1.176 | 67.85 | $7.40 \times 10^{-8}$ |
| | Quadratic | 0.892 | 0.795 | 4.826 | 36.95 | $2.83 \times 10^{-7}$ |

Table 9 demonstrates that both linear and quadratic models yield strong correlations across all dependent variables, with $R$ values above 0.86 and $R^2$ values up to 0.927. Quadratic models generally provide slightly better fit, as seen in higher $R^2$ values for variables like EV and MR. However, this improvement comes with increased SE of slope coefficients, suggesting greater variability in estimates. Linear models exhibit higher F-statistics and lower P-values overall, indicating more consistent statistical significance and predictive strength. While quadratic models enhance model fit marginally, linear models remain more stable and statistically reliable for interpretive purposes.

Figure 7 shows the relationship between $ABC$ and variables (BP, EV, FP, HR, P, MV) with data points (red dots). Two regression models are compared: The linear model (blue line) and the quadratic model (green line). Each plot visually represents how ABC correlates with the respective variable through different fits. The analysis helps assess the suitability of linear and quadratic models for each dataset. These plots provide insight into the nature of the correlation between ABC and the other variables.

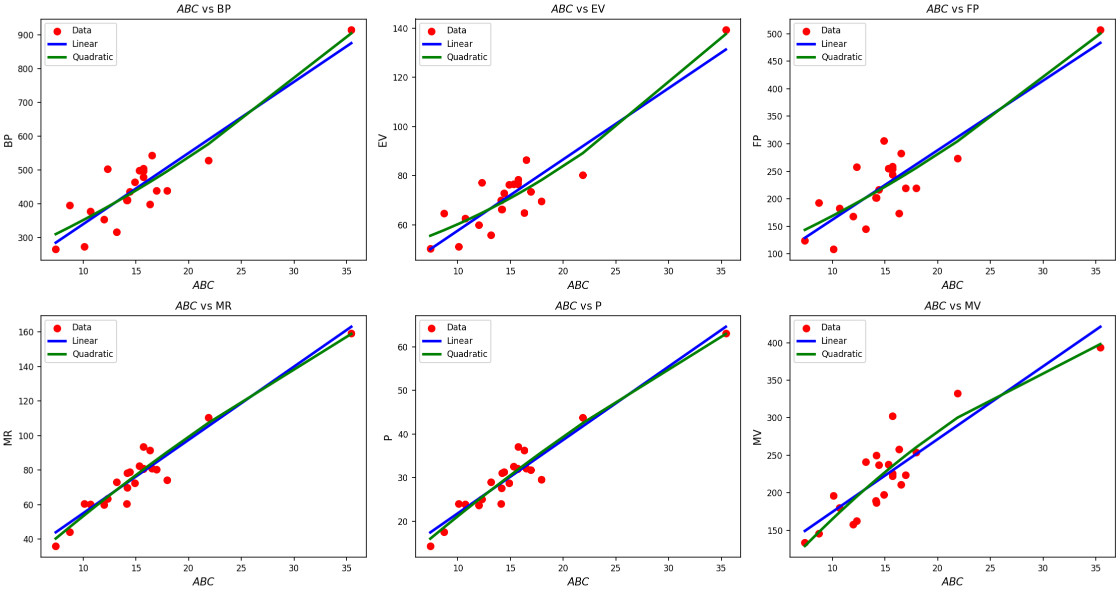


**Figure 7.** Graphical analysis of $ABC$.

### 5.7. *Regression models for Randic index RI(G)*

$$
\begin{aligned}
\textit{Boiling Point} &= 103.9344 + 36.6669RI,\\
\textit{Boiling Point} &= 217.6150 + 15.9375RI + 0.8286RI^2,\\
\textit{Enthalpy of Vaporization} &= 25.1898 + 5.0461RI,\\
\textit{Enthalpy of Vaporization} &= 47.1748 + 1.0372RI + 0.1602RI^2,\\
\textit{Flash Point} &= 21.0557 + 21.9500RI,\\
\textit{Flash Point} &= 88.9744 + 9.5652RI + 0.4950RI^2,\\
\textit{Molar Refractivity} &= 5.4206 + 7.6154RI,\\
\textit{Molar Refractivity} &= -5.1477 + 9.5426RI - 0.0770RI^2,\\
\textit{Polarizability} &= 2.1454 + 3.0192RI,\\
\textit{Polarizability} &= -2.0518 + 3.7845RI - 0.0306RI^2,\\
\textit{Molar Volume} &= 59.0172 + 17.6399RI,\\
\textit{Molar Volume} &= -8.6404 + 29.9772RI - 0.4931RI^2.
\end{aligned}
$$

The equations describe the thermodynamic properties of substances, such as BP, EV, and FP. The linear forms suggest a direct proportionality to the variable $RI$, while the quadratic equations account for more complex, nonlinear dependencies. For example, the quadratic form of the BP shows that the relationship with $RI$ is not constant but rather accelerates or decelerates at higher values of $RI$. These equations allow for more precise modeling of physical behaviors under varying conditions. The inclusion of both linear and quadratic terms makes the model adaptable to real-world scenarios where simple linear relationships may not suffice.

Table 10 presents the statistical comparison between linear and quadratic models for six dependent variables. In general, quadratic models slightly improve correlation coefficients (R) and coefficients of determination ($R^2$) compared to linear models. Notably, polarizability and molar refraction show the highest R and $R^2$ values, indicating strong model fits. Standard errors and F-statistics suggest that linear models often offer more stable slopes, while quadratic models increase complexity without always improving significance. P-values across all models remain extremely low, confirming strong statistical significance.

**Table 10.** Statistical parameters of $RI$ using linear and quadratic models.

| Dependent variable | Model type | R | $R^2$ | SE (slope) | F-statistic | P-value |
|---|---|---|---|---|---|---|
| Boiling Point | Linear | 0.896 | 0.803 | 4.061 | 81.53 | $1.71 \times 10^{-8}$ |
| | Quadratic | 0.904 | 0.817 | 17.495 | 42.49 | $9.71 \times 10^{-8}$ |
| Enthalpy of Vaporization | Linear | 0.902 | 0.814 | 0.539 | 87.71 | $9.42 \times 10^{-9}$ |
| | Quadratic | 0.918 | 0.843 | 2.217 | 50.95 | $2.32 \times 10^{-8}$ |
| Flash Point | Linear | 0.860 | 0.739 | 2.918 | 56.60 | $2.97 \times 10^{-7}$ |
| | Quadratic | 0.867 | 0.752 | 12.720 | 28.80 | $1.77 \times 10^{-6}$ |
| Molar Refraction | Linear | 0.982 | 0.964 | 0.331 | 528.10 | $7.50 \times 10^{-16}$ |
| | Quadratic | 0.983 | 0.967 | 1.411 | 277.83 | $8.59 \times 10^{-15}$ |
| Polarizability | Linear | 0.982 | 0.963 | 0.132 | 527.02 | $7.65 \times 10^{-16}$ |
| | Quadratic | 0.983 | 0.967 | 0.560 | 277.30 | $8.74 \times 10^{-15}$ |
| Molar Volume | Linear | 0.911 | 0.830 | 1.784 | 97.82 | $3.81 \times 10^{-9}$ |
| | Quadratic | 0.923 | 0.853 | 7.429 | 55.04 | $1.24 \times 10^{-8}$ |

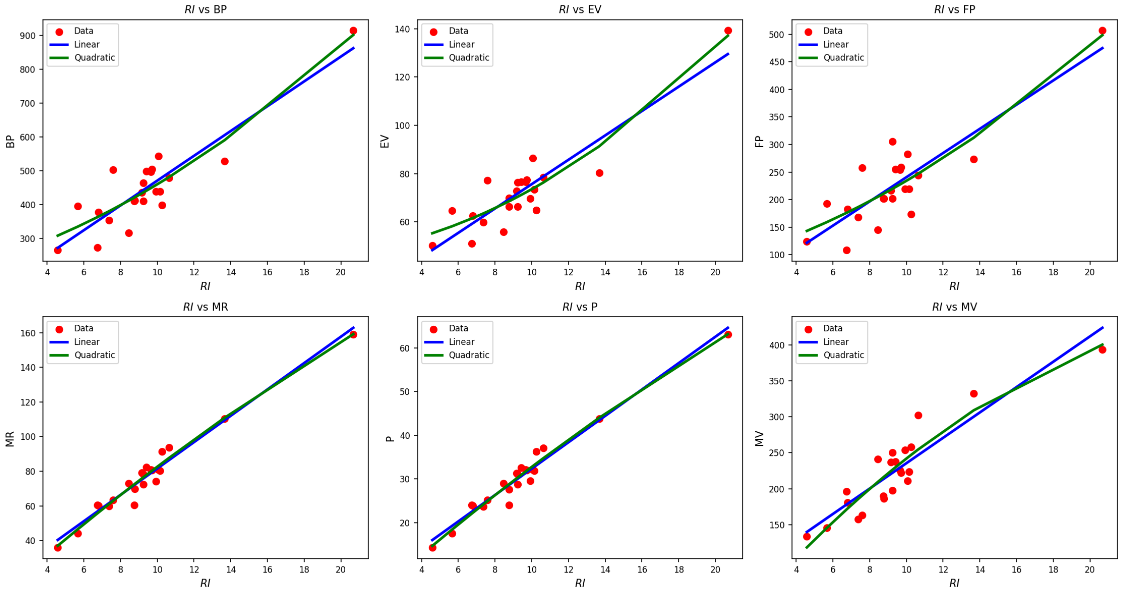


**Figure 8.** Graphical analysis of $RI$.

Figure 8 depict the relationship between “RI” and several other variables (BP, EV, FP, MR, P, MV), with data points (red dots). Two regression models are compared: The linear model (blue line) and the quadratic model (green line). Each plot visualizes how “RI” correlates with the respective variable using different regression fits. The analysis evaluates the appropriateness of both the linear and quadratic models for each dataset. These plots provide insights into the nature of the correlation between “RI” and the other variables.

### 5.8. Regression models for sum connectivity index $SC(G)$

$$\begin{aligned}
\textit{Boiling Point} &= 121.5814 + 33.5662SC,\\
\textit{Boiling Point} &= 213.0934 + 17.6058SC + 0.6007SC^2,\\
\textit{Enthalpy of Vaporization} &= 27.6607 + 4.6150SC,\\
\textit{Enthalpy of Vaporization} &= 46.3796 + 1.3503SC + 0.1229SC^2,\\
\textit{Flash Point} &= 31.2394 + 20.1329SC,\\
\textit{Flash Point} &= 84.7356 + 10.8028SC + 0.3512SC^2,\\
\textit{Molar Refractivity} &= 9.7668 + 6.9014SC,\\
\textit{Molar Refractivity} &= -1.5673 + 8.8782SC - 0.0744SC^2,\\
\textit{Polarizability} &= 3.8686 + 2.7361SC,\\
\textit{Polarizability} &= -0.6323 + 3.5211SC - 0.0295SC^2,\\
\textit{Molar Volume} &= 71.2506 + 15.7633SC,\\
\textit{Molar Volume} &= 9.8256 + 26.4763SC - 0.4032SC^2.
\end{aligned}$$

Equations describe various physical properties such as BP, EV, and FP. The linear equations indicate a straightforward relationship with the variable $SC$, while the quadratic equations introduce a more complex dependency, accounting for nonlinear effects. For instance, the BP’s quadratic equation suggests that the rate of change with $SC$ accelerates at higher values. These models help accurately represent real-world behaviors where properties are influenced by both linear and higher-order terms. The inclusion of both forms allows for a more nuanced understanding of substance behavior under varying conditions.

**Table 11.** Statistical parameters of $SC$ using linear and quadratic models.

| Dependent variable | Model type | R | R$^2$ | SE (slope) | F-statistic | P-value |
|---|---|---|---|---|---|---|
| Boiling Point | Linear | 0.901 | 0.812 | 3.613 | 86.32 | $1.07 \times 10^{-8}$ |
| | Quadratic | 0.907 | 0.823 | 15.222 | 44.10 | $7.27 \times 10^{-8}$ |
| Enthalpy of Vaporization | Linear | 0.907 | 0.822 | 0.481 | 92.21 | $6.23 \times 10^{-9}$ |
| | Quadratic | 0.920 | 0.846 | 1.938 | 52.23 | $1.90 \times 10^{-8}$ |
| Flash Point | Linear | 0.866 | 0.750 | 2.599 | 60.00 | $1.91 \times 10^{-7}$ |
| | Quadratic | 0.872 | 0.760 | 11.065 | 30.00 | $1.32 \times 10^{-6}$ |
| Molar Refraction | Linear | 0.977 | 0.955 | 0.336 | 421.42 | $6.56 \times 10^{-15}$ |
| | Quadratic | 0.979 | 0.959 | 1.383 | 224.07 | $6.14 \times 10^{-14}$ |
| Polarizability | Linear | 0.977 | 0.955 | 0.133 | 420.64 | $6.68 \times 10^{-15}$ |
| | Quadratic | 0.979 | 0.959 | 0.549 | 223.69 | $6.24 \times 10^{-14}$ |
| Molar Volume | Linear | 0.894 | 0.800 | 1.763 | 79.95 | $2.01 \times 10^{-8}$ |
| | Quadratic | 0.907 | 0.822 | 7.222 | 43.80 | $7.67 \times 10^{-8}$ |

Table 11 provides a comparative analysis of linear and quadratic regression models for six physicochemical properties. Across most variables, quadratic models exhibit marginally improved correlation coefficients ($R$) and coefficients of determination ($R^2$), indicating slightly enhanced model fits. Nonetheless, the linear models consistently produce lower SEs of the slope, implying more precise and stable parameter estimates. MR and P demonstrate the best predictive performance, with both models yielding $R^2 > 0.95$ and extremely high F-statistics, reinforcing the robustness of the regression. Despite the modest improvement in model fit for quadratic forms, the consistently low P-values ($< 10^{-6}$) across all cases confirm that both model types provide statistically significant relationships.

Figure 9 shows the relationship between sum connectivity (SC) and several other variables (BP, EV, FP, MR, P, MV), with data points (red dots). Two regression models are compared: The linear model (blue line) and the quadratic model (green line). Each plot visualizes how "SC" correlates with the respective variable using both regression models. The analysis helps assess which model (linear or quadratic) best fits the data for each variable. These plots provide insights into the correlation between "SC" and other variables in the dataset.

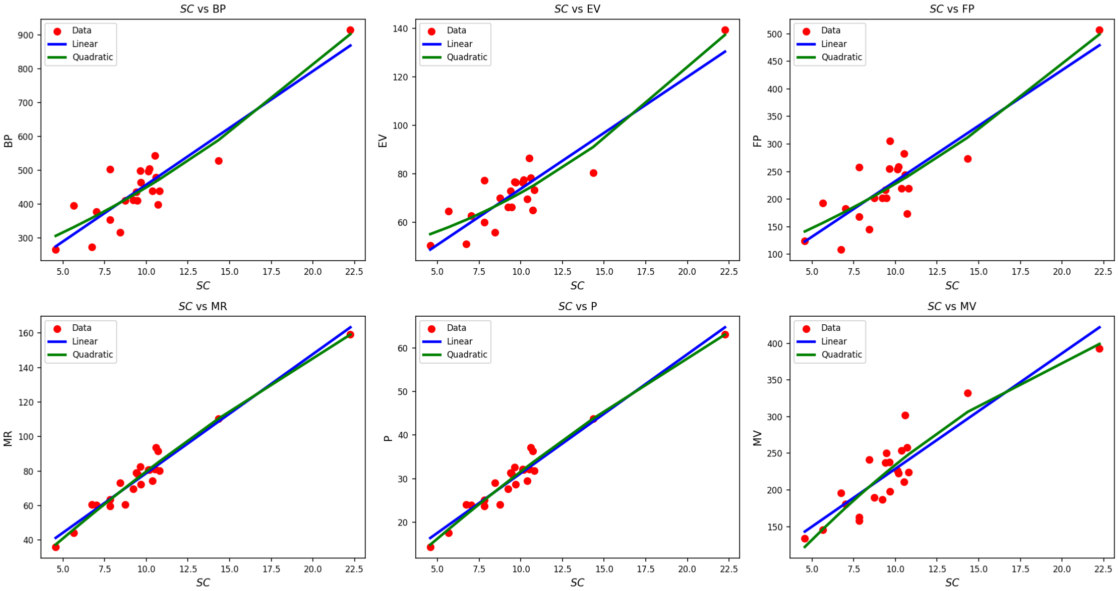


**Figure 9.** Graphical analysis of $SC$.

### 5.9. *Regression models for geometric arithmetic index GA*

$$\begin{aligned}
\mathit{Boiling\ Point} &= 138.3318 + 15.1622GA,\\
\mathit{Boiling\ Point} &= 212.6144 + 9.0478GA + 0.1068GA^2,\\
\mathit{Enthalpy\ of\ Vaporization} &= 29.9782 + 2.0839GA_1,\\
\mathit{Enthalpy\ of\ Vaporization} &= 46.1407 + 0.7536GA + 0.0232GA^2,\\
\mathit{Flash\ Point} &= 40.9565 + 9.1104GA,\\
\mathit{Flash\ Point} &= 83.1505 + 5.6373GA + 0.0607GA^2,\\
\mathit{Molar\ Refractivity} &= 13.8652 + 3.0854GA,\\
\mathit{Molar\ Refractivity} &= 2.3038 + 4.0371GA - 0.0166GA^2,\\
\mathit{Polarizability} &= 5.4936 + 1.2232GA,\\
\mathit{Polarizability} &= 0.9022 + 1.6011GA - 0.0066GA^2,\\
\mathit{Molar\ Volume} &= 82.3775 + 6.9609GA,\\
\mathit{Molar\ Volume} &= 26.6607 + 11.5471GA - 0.0801GA^2.
\end{aligned}$$

Equations describe the relationships between various properties like BP, EV, and FP with respect to the variable $GA$. The linear forms indicate a constant rate of change with $GA$, while the quadratic equations introduce higher-order effects that account for nonlinear behavior. For instance, the quadratic form of the BP suggests a more complex relationship, with an accelerating effect at higher values of $GA$. These models capture both simple and intricate behaviors of substances in different conditions. By combining both linear and quadratic terms, the equations offer a more accurate representation of real-world data.

Table 12 presents regression equations and statistical parameters for linear and quadratic models across six molecular properties. Overall, quadratic models yield slightly higher $R$ and $R^2$ values, indicating improved goodness-of-fit, particularly for EV, and MV. However, the SE of the slope is consistently lower in linear models, implying higher stability and less variance in slope estimation. MR and P stand out with exceptionally high $R^2$ values ($> 0.94$) and F-statistics, confirming strong predictive power for both model types. Despite the quadratic form improving curve-fitting slightly, the increased model complexity yields diminishing returns in statistical performance, as evidenced by higher SE and reduced F-values.

**Table 12.** Regression statistics for GA dataset using linear and quadratic models.

| Property | Model | Equation | R | R$^2$ | SE (slope) | F | P-value |
|---|---|---|---|---|---|---|---|
| BP | Linear | $y = 138.33 + 15.16x$ | 0.904 | 0.817 | 1.605 | 89.19 | $8.20 \times 10^{-9}$ |
| | Quadratic | $y = 212.61 + 9.05x + 0.107x^2$ | 0.908 | 0.825 | 6.658 | 44.81 | $6.41 \times 10^{-8}$ |
| EV | Linear | $y = 29.98 + 2.08x$ | 0.909 | 0.826 | 0.214 | 95.08 | $4.83 \times 10^{-9}$ |
| | Quadratic | $y = 46.14 + 0.75x + 0.023x^2$ | 0.920 | 0.847 | 0.851 | 52.64 | $1.78 \times 10^{-8}$ |
| FP | Linear | $y = 40.96 + 9.11x$ | 0.870 | 0.757 | 1.153 | 62.38 | $1.42 \times 10^{-7}$ |
| | Quadratic | $y = 83.15 + 5.64x + 0.061x^2$ | 0.874 | 0.764 | 4.825 | 30.77 | $1.10 \times 10^{-6}$ |
| MR | Linear | $y = 13.87 + 3.09x$ | 0.970 | 0.941 | 0.173 | 318.19 | $9.48 \times 10^{-14}$ |
| | Quadratic | $y = 2.30 + 4.04x - 0.017x^2$ | 0.973 | 0.946 | 0.699 | 167.80 | $8.42 \times 10^{-13}$ |
| P | Linear | $y = 5.49 + 1.22x$ | 0.970 | 0.941 | 0.069 | 317.65 | $9.63 \times 10^{-14}$ |
| | Quadratic | $y = 0.90 + 1.60x - 0.007x^2$ | 0.973 | 0.946 | 0.277 | 167.55 | $8.54 \times 10^{-13}$ |
| MV | Linear | $y = 82.38 + 6.96x$ | 0.877 | 0.769 | 0.853 | 66.61 | $8.55 \times 10^{-8}$ |
| | Quadratic | $y = 26.66 + 11.55x - 0.080x^2$ | 0.889 | 0.790 | 3.453 | 35.70 | $3.67 \times 10^{-7}$ |

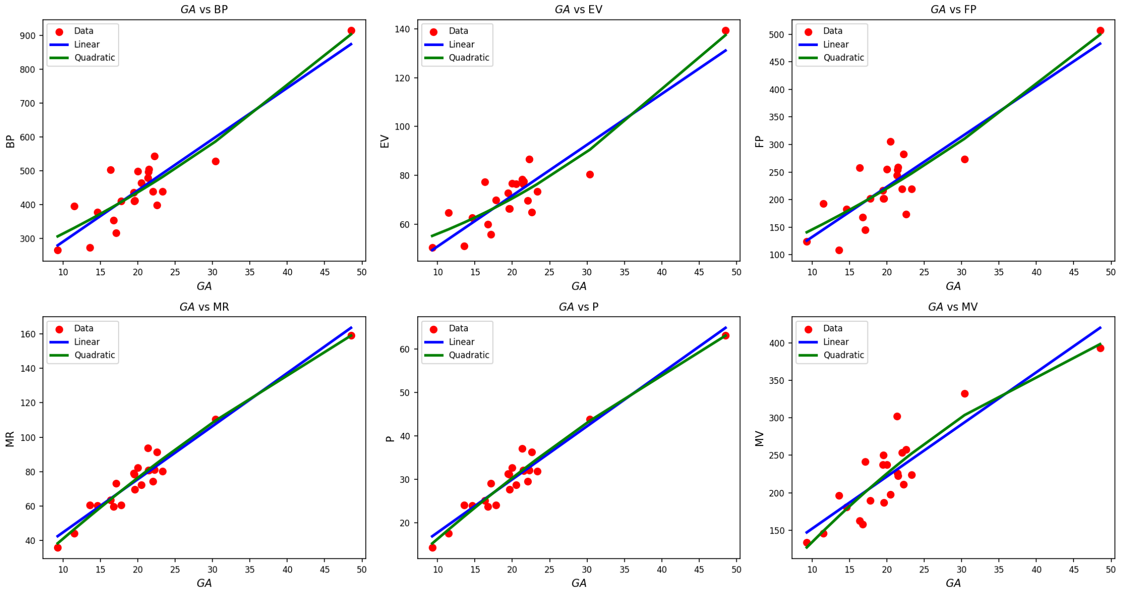


**Figure 10.** Graphical analysis of $GA$.

Figure 10 presents scatter plots of various relationships between GA and different metrics (BP, EV, FP, MR, P, MV). Each plot includes data points (red circles) along with linear and quadratic regression lines. The linear regression is represented by a blue line, while the quadratic regression is represented by a green line. The plots demonstrate a general positive relationship between GA and the corresponding metrics, with some showing better fit for quadratic regression. This suggests that for certain metrics, a nonlinear model may provide a better representation of the data.

### *5.10. Regression models for Hyper Zagreb index HZ(G)*

$$\begin{aligned}
\textit{Boiling Point} &= 192.3087 + 0.5023HZ,\\
\textit{Boiling Point} &= 236.7149 + 0.3602HZ + 0.0001HZ^2,\\
\textit{Enthalpy of Vaporization} &= 37.2360 + 0.0694HZ,\\
\textit{Enthalpy of Vaporization} &= 48.6044 + 0.0330HZ + 0.0000HZ^2,\\
\textit{Flash Point} &= 72.2004 + 0.3042HZ,\\
\textit{Flash Point} &= 91.4880 + 0.2424HZ + 0.0000HZ^2,\\
\textit{Molar Refractivity} &= 27.7530 + 0.0965HZ,\\
\textit{Molar Refractivity} &= 21.1415 + 0.1177HZ - 0.0000HZ^2,\\
\textit{Polarizability} &= 10.9992 + 0.0383HZ,\\
\textit{Polarizability} &= 8.3666 + 0.0467HZ - 0.0000HZ^2,\\
\textit{Molar Volume} &= 115.0435 + 0.2151HZ,\\
\textit{Molar Volume} &= 76.8713 + 0.3373HZ - 0.0001HZ^2.
\end{aligned}$$

The following properties are determined using both linear and quadratic models. These models are useful for predicting different characteristics of a substance, such as BP, EV, FP, and more. The coefficients in the equations are specific to the material and can be used to estimate these properties based on a variable, *HZ*. For accurate calculations, the quadratic form often provides a better approximation, especially when higher-order terms are significant. The equations shown here are picked to represent a comprehensive model for predicting these thermodynamic properties.

Table 13 presents the statistical parameters for linear and quadratic regression models fitted to various properties. For most properties, the quadratic model slightly improves the correlation coefficient ($R$) and coefficient of determination ($R^2$) compared to the linear model. However, this comes at the cost of increased SE and lower F-values, indicating that the added complexity may not always be justified. All models show statistically significant fits with very low P-values ($< 10^{-5}$), confirming the relevance of the predictors. Overall, the linear models offer a more parsimonious yet adequately accurate fit for most properties.

Figure 11 shows scatterplots depicting relationships between HZ (a given variable) and several metrics (BP, EV, FP, MR, P, MV). The data points (red circles) are plotted alongside linear and quadratic regression lines. The linear regression is represented by a blue line, while the quadratic regression is shown with a green line. These plots suggest a strong positive correlation between HZ and the respective metrics, with the quadratic model often providing a better fit. This indicates that for most metrics, a nonlinear relationship is more appropriate than a linear one.

**Table 13.** Regression statistics for HZ dataset using linear and quadratic models.

| Property | Model | Equation | R | $R^2$ | SE (slope) | F | P-value |
|---|---|---|---|---|---|---|---|
| BP | Linear | $y = 192.31 + 0.50x$ | 0.890 | 0.792 | 0.058 | 76.16 | $2.97 \times 10^{-8}$ |
| | Quadratic | $y = 236.71 + 0.36x + 0.0001x^2$ | 0.892 | 0.796 | 0.240 | 37.07 | $2.76 \times 10^{-7}$ |
| EV | Linear | $y = 37.24 + 0.07x$ | 0.899 | 0.808 | 0.008 | 84.41 | $1.29 \times 10^{-8}$ |
| | Quadratic | $y = 48.60 + 0.03x + 0.0000x^2$ | 0.907 | 0.823 | 0.031 | 44.03 | $7.35 \times 10^{-8}$ |
| FP | Linear | $y = 72.20 + 0.30x$ | 0.863 | 0.746 | 0.040 | 58.62 | $2.28 \times 10^{-7}$ |
| | Quadratic | $y = 91.49 + 0.24x + 0.0000x^2$ | 0.865 | 0.748 | 0.166 | 28.13 | $2.09 \times 10^{-6}$ |
| MR | Linear | $y = 27.75 + 0.10x$ | 0.902 | 0.813 | 0.010 | 87.13 | $9.95 \times 10^{-9}$ |
| | Quadratic | $y = 21.14 + 0.12x + 0.0000x^2$ | 0.903 | 0.816 | 0.043 | 42.07 | $1.05 \times 10^{-7}$ |
| P | Linear | $y = 11.00 + 0.04x$ | 0.902 | 0.813 | 0.004 | 87.09 | $9.98 \times 10^{-9}$ |
| | Quadratic | $y = 8.37 + 0.05x + 0.0000x^2$ | 0.903 | 0.816 | 0.017 | 42.06 | $1.05 \times 10^{-7}$ |
| MV | Linear | $y = 115.04 + 0.22x$ | 0.806 | 0.649 | 0.035 | 36.97 | $6.08 \times 10^{-6}$ |
| | Quadratic | $y = 76.87 + 0.34x - 0.0001x^2$ | 0.814 | 0.662 | 0.146 | 18.62 | $3.33 \times 10^{-5}$ |

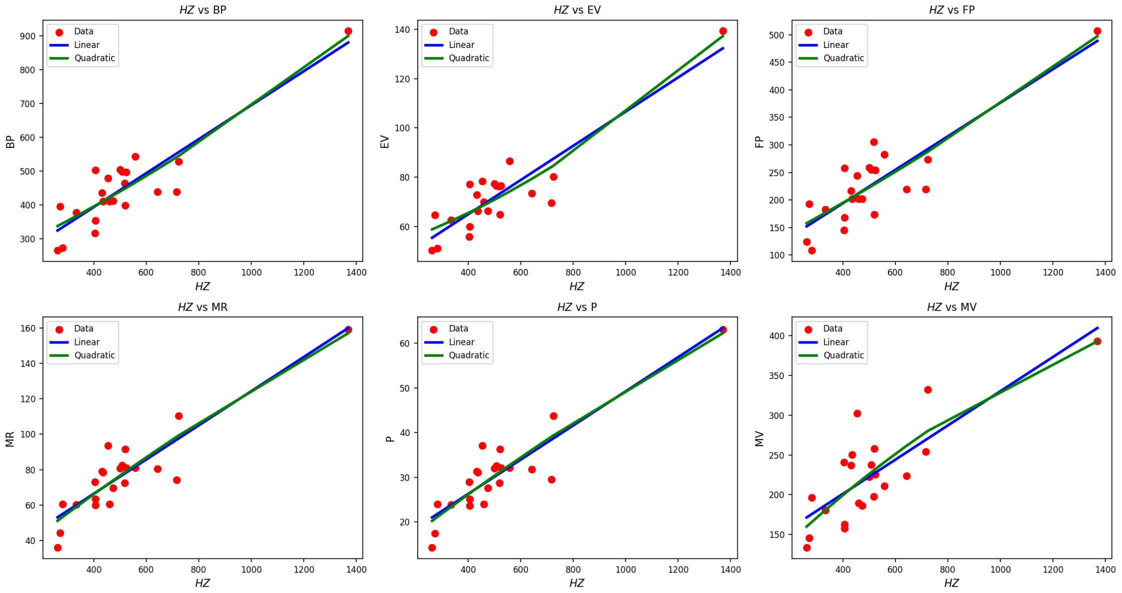


**Figure 11.** Graphical analysis of *HZ*.

### 5.11. Regression models for Nirmala index $N(G)$

$$
\begin{aligned}
\textit{Boiling Point} &= 148.8324 + 6.4640N,\\
\textit{Boiling Point} &= 209.4594 + 4.2865N + 0.0164N^2,\\
\textit{Enthalpy of Vaporization} &= 31.3749 + 0.8894N,\\
\textit{Enthalpy of Vaporization} &= 45.6534 + 0.3766N + 0.0039N^2,\\
\textit{Flash Point} &= 47.0524 + 3.8886N,\\
\textit{Flash Point} &= 80.0548 + 2.7032N + 0.0089N^2,\\
\textit{Molar Refractivity} &= 17.2062 + 1.2894N,\\
\textit{Molar Refractivity} &= 5.7347 + 1.7014N - 0.0031N^2,\\
\textit{Polarizability} &= 6.8179 + 0.5112N,
\end{aligned}
$$

$$\begin{aligned} \mathit{Polarizability} &= 2.2596 + 0.6749N - 0.0012N^2, \\ \mathit{Molar\ Volume} &= 89.8430 + 2.9104N, \\ \mathit{Molar\ Volume} &= 33.0509 + 4.9502N - 0.0153N^2. \end{aligned}$$

Table 14 presents the statistical parameters for linear and quadratic regression models fitted to various properties. For most properties, the quadratic model slightly improves the correlation coefficient ($R$) and coefficient of determination ($R^2$) compared to the linear model. However, this comes at the cost of increased standard error (SE) and lower F-values, indicating that the added complexity may not always be justified. All models show statistically significant fits with very low P-values ($< 10^{-5}$), confirming the relevance of the predictors. Overall, the linear models offer a more parsimonious yet adequately accurate fit for most properties.

**Table 14.** Regression statistics for N dataset using linear and quadratic models.

| Property | Model | Equation | R | R$^2$ | SE (slope) | F | P-value |
|---|---|---|---|---|---|---|---|
| BP | Linear | $y = 148.83 + 6.46x$ | 0.904 | 0.818 | 0.682 | 89.80 | $7.76 \times 10^{-9}$ |
| | Quadratic | $y = 209.46 + 4.29x + 0.02x^2$ | 0.908 | 0.824 | 2.86 | 44.34 | $6.96 \times 10^{-8}$ |
| EV | Linear | $y = 31.37 + 0.89x$ | 0.911 | 0.829 | 0.090 | 97.03 | $4.07 \times 10^{-9}$ |
| | Quadratic | $y = 45.65 + 0.38x + 0.004x^2$ | 0.920 | 0.846 | 0.365 | 52.22 | $1.90 \times 10^{-8}$ |
| FP | Linear | $y = 47.05 + 3.89x$ | 0.872 | 0.760 | 0.489 | 63.33 | $1.26 \times 10^{-7}$ |
| | Quadratic | $y = 80.05 + 2.70x + 0.01x^2$ | 0.874 | 0.764 | 2.06 | 30.81 | $1.09 \times 10^{-6}$ |
| MR | Linear | $y = 17.21 + 1.29x$ | 0.951 | 0.905 | 0.093 | 190.84 | $1.09 \times 10^{-11}$ |
| | Quadratic | $y = 5.73 + 1.70x - 0.0031x^2$ | 0.954 | 0.911 | 0.385 | 97.05 | $1.06 \times 10^{-10}$ |
| P | Linear | $y = 6.82 + 0.51x$ | 0.951 | 0.905 | 0.037 | 190.68 | $1.10 \times 10^{-11}$ |
| | Quadratic | $y = 2.26 + 0.67x - 0.0012x^2$ | 0.954 | 0.911 | 0.153 | 96.99 | $1.07 \times 10^{-10}$ |
| MV | Linear | $y = 89.84 + 2.91x$ | 0.861 | 0.741 | 0.385 | 57.12 | $2.77 \times 10^{-7}$ |
| | Quadratic | $y = 33.05 + 4.95x - 0.0153x^2$ | 0.874 | 0.763 | 1.57 | 30.60 | $1.14 \times 10^{-6}$ |

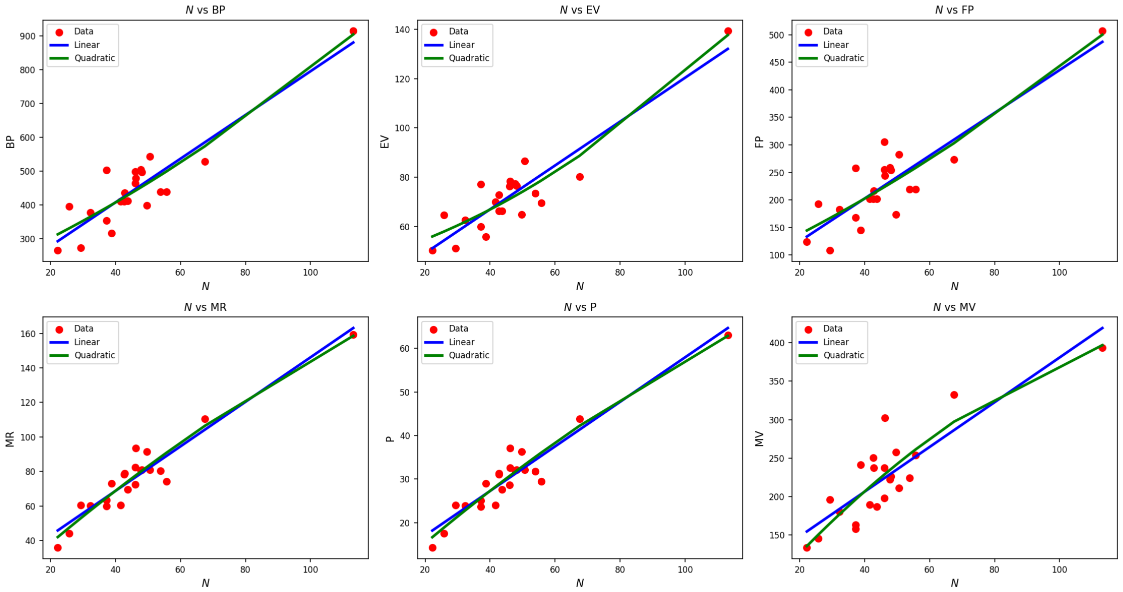


**Figure 12.** Graphical analysis of $N$.

Figure 12 shows scatterplots of the relationship between N (a given variable) and various metrics (BP, EV, FP, MR, P, MV). The red circles represent the data points, while the blue and green lines represent the linear and quadratic regression models, respectively. These plots indicate a clear positive correlation between N and the metrics, with the quadratic model often providing a better fit, especially for the BP, EV, and MV relationships. This suggests that for these variables, a nonlinear relationship might better capture the underlying patterns.

## 6. Random forest

The RF regression approach was used to model structure property relationships on few molecular datasets in this work. All target properties were separately predicted by a group of topological and physicochemical descriptors. Three main metrics were investigated to evaluate model precision and generalization, coefficient of determination ($R^2$), Root RMSE and out of bag (OOB) score. The $R^2$ corresponds to the fraction of the variance in the experimental dataset that is explained by the predictions, the RMSE describes the average magnitude of the prediction errors, and the OOB score is an internal validation measure estimated on unseen bootstrap samples.

The prediction performance for BP is presented in Figure 13, and the results are listed in Table 15. The model obtained a coefficient of determination of $R^2 = 0.889$ and RMSE of 42.02, with an OOB score of 0.117, indicating low ability to generalize beyond the training data subset. The mean percentage error was 6.47% and 5 compounds exceeded 10% error range, suggesting moderate scatter in prediction accuracy. HZ was found to have the highest relative importance (0.120) in the group of descriptors evaluated, and the contributions of GA, F, and ABC were approximately 0.098-0.096, emphasizing their potential in encoding structural information useful for BP prediction.

**Table 15.** Actual vs predicted values of BP.

| | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Actual | 398.2 | 411 | 497.4 | 527.9 | 914.5 | 265.3 | 479.5 | 439.3 | 503.1 | 395.9 | 543.6 | 464 | 378 | 434.9 | 316.2 | 504.8 | 410.5 | 272.5 | 497.7 | 353.8 | 410.8 | 438.7 |
| Predicted | 433.2 | 420.2 | 492.2 | 493.2 | 764.3 | 304.6 | 471.6 | 456.6 | 454.0 | 355.8 | 507.5 | 466.9 | 352.9 | 430.2 | 364.5 | 491.2 | 419.7 | 311.8 | 485.0 | 387.7 | 413.7 | 458.8 |

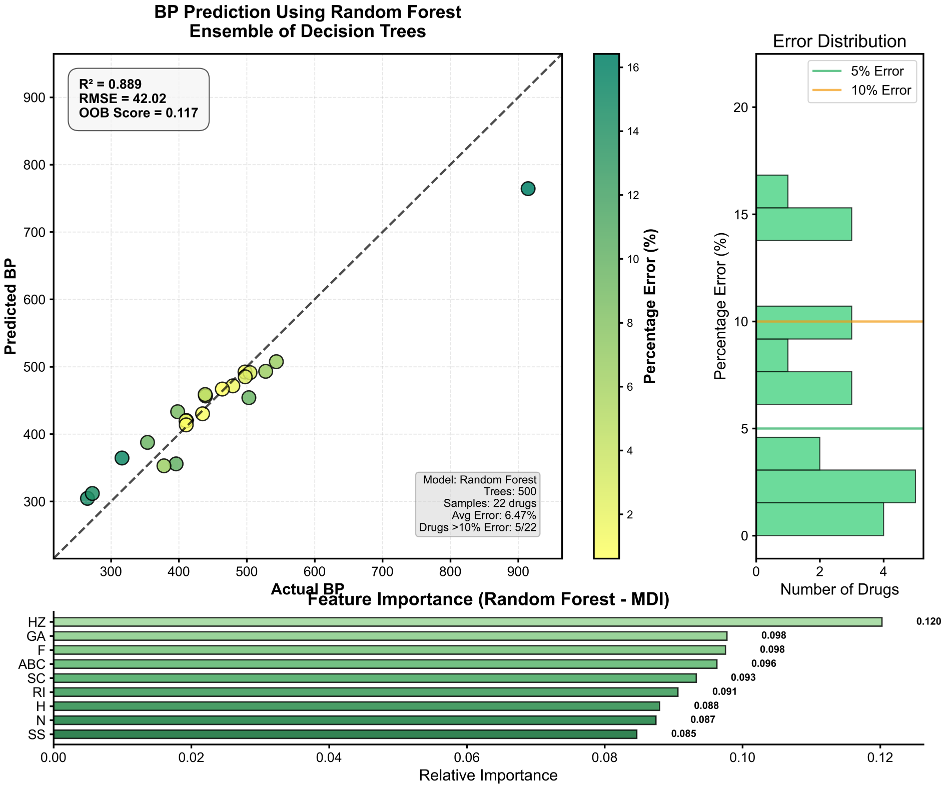


**Figure 13.** RF regression model comparing actual vs. predicted BP based on the indices.

The regression result of (EV) is showed in Figure 14 and Lists detailed comparison of values in Table 16. The model resulted in $R^2 = 0.888$, RMSE = 5.77, and an OOB score = 0.111, suggesting modest generalization capacity. The mean error of predication was 5.03%, with two samples exceeding 10%. Regarding the contribution of variables, it was found that HZ was the most effective descriptor to explain the response, whose relative importance was 0.124, being followed by F and ABC, which relative importance was 0.119 and 0.105, respectively, showing that these features are relevant to the description of important structural determinants associated with the behavior of enthalpy.

**Table 16.** Actual vs predicted values of EV.

| Actual | 64.9 | 66.3 | 76.5 | 80.3 | 139.3 | 50.3 | 78.4 | 73.4 | 77.2 | 64.6 | 86.5 | 76.4 | 62.6 | 72.8 | 55.8 | 77.4 | 66.3 | 51.1 | 76.6 | 59.9 | 69.9 | 69.6 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Predicted | 70.1 | 68.3 | 76.1 | 77.3 | 117.2 | 54.6 | 76.2 | 75.1 | 71.6 | 60.2 | 81.1 | 75.4 | 59.9 | 71.1 | 61.6 | 76.1 | 68.4 | 55.5 | 75.8 | 63.9 | 69.0 | 72.9 |

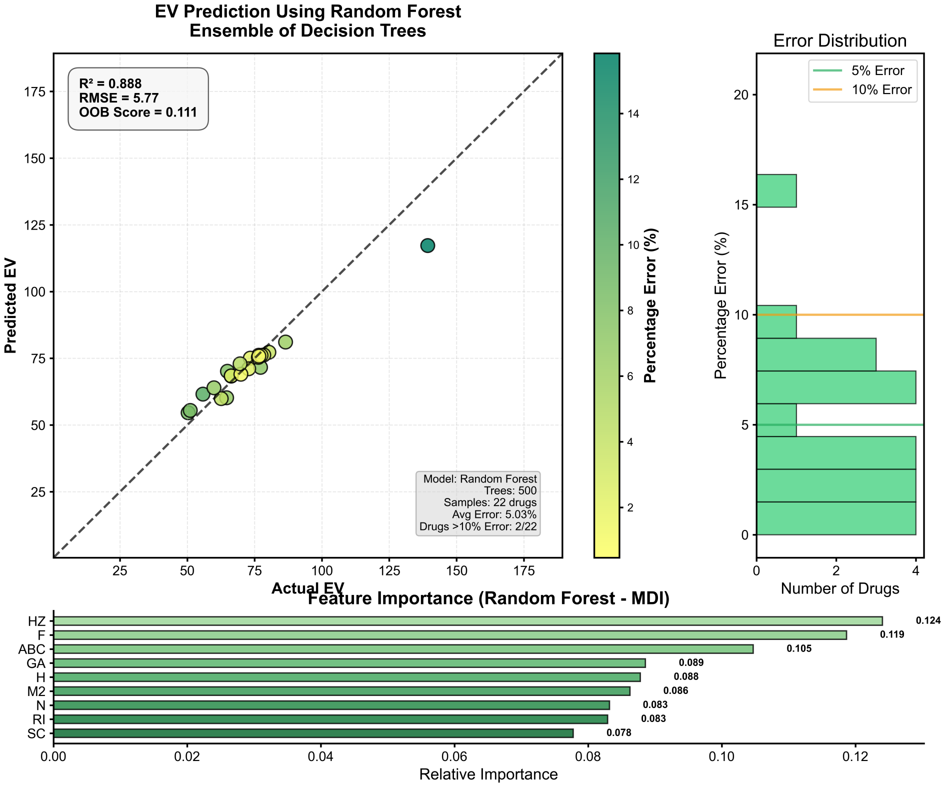


**Figure 14.** RF regression model comparing actual vs. predicted EV based on the indices.

The prediction results for (FP) can be found in Figure 15, and the performance of predicted versus actual values in Table 17. For the regression model: We got a $R^2 = 0.885$, an RMSE of 26.77 and an OOB score = 0.087, which suggests that we have poor performance in cross-validated generalization. The model's mean bias was 8.82% , and given that there were eight data points over the 10% cutoff, the residuals were less normally distributed. Among the six determinants, HZ was the most significant descriptor (0.112), while F and SC ranked at second and third importance levels with contributions of 0.102 and 0.100, respectively, representing the role of FP structural determining features.

**Table 17.** Actual vs predicted values of FP.

| Actual | 174 | 202.4 | 254.6 | 273.1 | 506.9 | 123.8 | 243.8 | 219.5 | 258.1 | 193.2 | 282.6 | 305.8 | 182.4 | 216.8 | 145 | 259.1 | 202 | 108.4 | 254.8 | 167.8 | 202.3 | 219.1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Predicted | 204.0 | 211.0 | 254.8 | 250.3 | 414.6 | 143.5 | 236.7 | 227.6 | 228.5 | 168.6 | 261.8 | 279.1 | 165.0 | 214.8 | 174.0 | 252.1 | 210.3 | 135.9 | 256.8 | 188.4 | 205.5 | 231.7 |

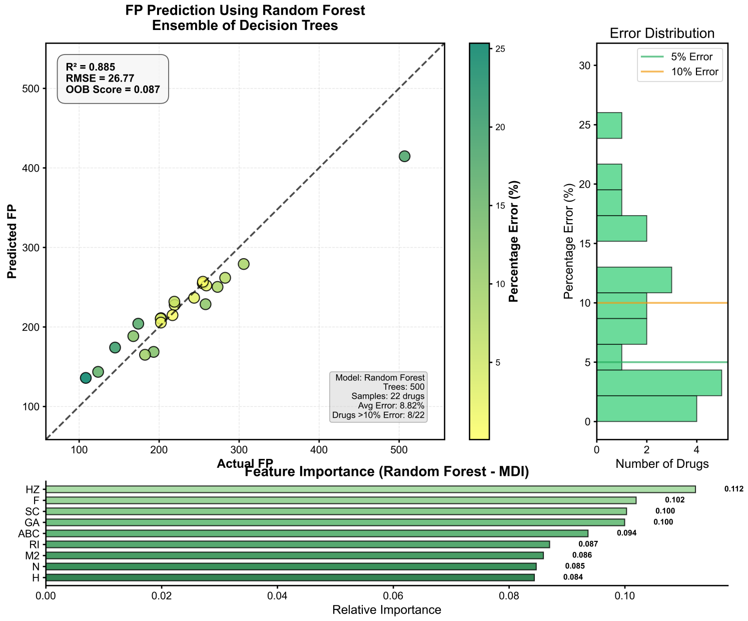


**Figure 15.** RF regression model comparing actual vs. predicted FP based on the indices.

In Figure 16, the comparison is shown for the regression performance of ST, and the numerical results are listed in Table 18. The model gave an $R^2 = 0.839$ as well as an RMSE of 4.72, although the OOB score was quite low at 0.343, hinting at a poor generalization in the case of bootstrap validation. The mean prediction error was 7.06% and there were five samples above the 10% limit. As descriptors, SS is the most important one with 0.114 relative importance, followed by GA and HZ with contributions of 0.109 and 0.108, respectively. Such results indicate that, while many of these descriptors are important, the model might have had difficulty learning general patterns for this particular property.

**Table 18.** Actual vs predicted values of ST.

| | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Actual | 47 | 57.3 | 46.1 | 45.2 | 82.1 | 31 | 43.4 | 56.6 | 79.6 | 48.9 | 56 | 53.3 | 53.2 | 42.7 | 36.9 | 46.6 | 40.5 | 37.3 | 52.7 | 49.6 | 45.4 | 53.4 |
| Predicted | 48.6 | 52.6 | 47.9 | 48.6 | 70.1 | 37.1 | 45.0 | 54.9 | 68.7 | 44.7 | 53.5 | 51.6 | 49.9 | 43.6 | 44.3 | 47.4 | 42.6 | 42.1 | 51.9 | 54.4 | 46.4 | 53.5 |

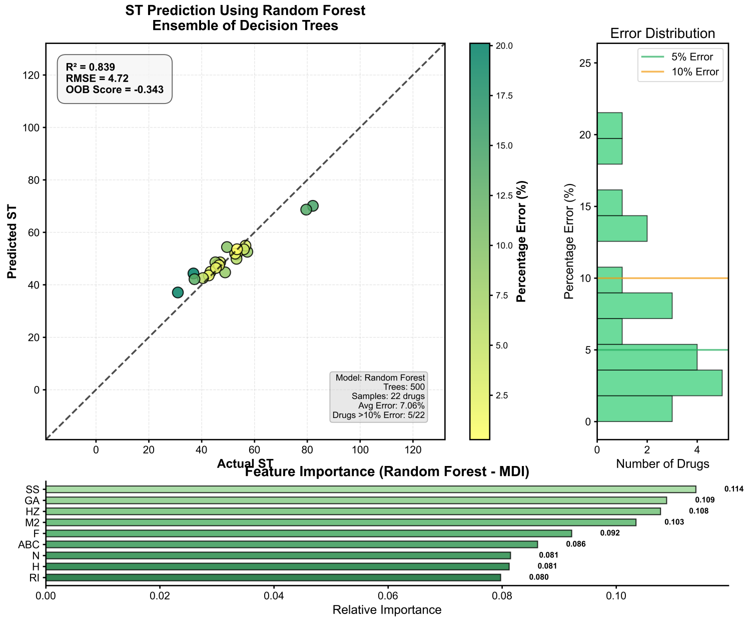


**Figure 16.** RF regression model comparing actual vs. predicted ST based on the indices.

The performance of the model in prediction for molar refractivity (MR) is shown in Figure 17, and the numerical values are provided in Table 19. $R^2$ = 0.949, RMSE = 5.39 was achieved and the (OOB) score was 0.610, suggesting moderate generalization ability under internal validation process. Average prediction error was 4.17%, and only two compounds deviated over 10%. The H was the most important descriptor considering the highest contribution to the structural variation related to MR (0.142), followed by SC (RI (0.120 and 0.109), indicating that they present a constant effect in structural variation related to MR.

**Table 19.** Actual vs predicted values of MR.

| Actual | 91.5 | 69.7 | 80.9 | 110.4 | 159.2 | 36 | 93.6 | 80.3 | 63.4 | 44.2 | 81 | 72.4 | 60.3 | 79 | 73.1 | 80.7 | 78.4 | 60.5 | 82.4 | 59.8 | 60.6 | 74.3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Predicted | 88.3 | 71.8 | 80.7 | 101.3 | 138.9 | 40.7 | 90.3 | 82.2 | 63.6 | 46.7 | 81.8 | 75.9 | 60.4 | 76.8 | 69.4 | 80.8 | 76.9 | 58.0 | 80.3 | 61.9 | 65.2 | 78.1 |

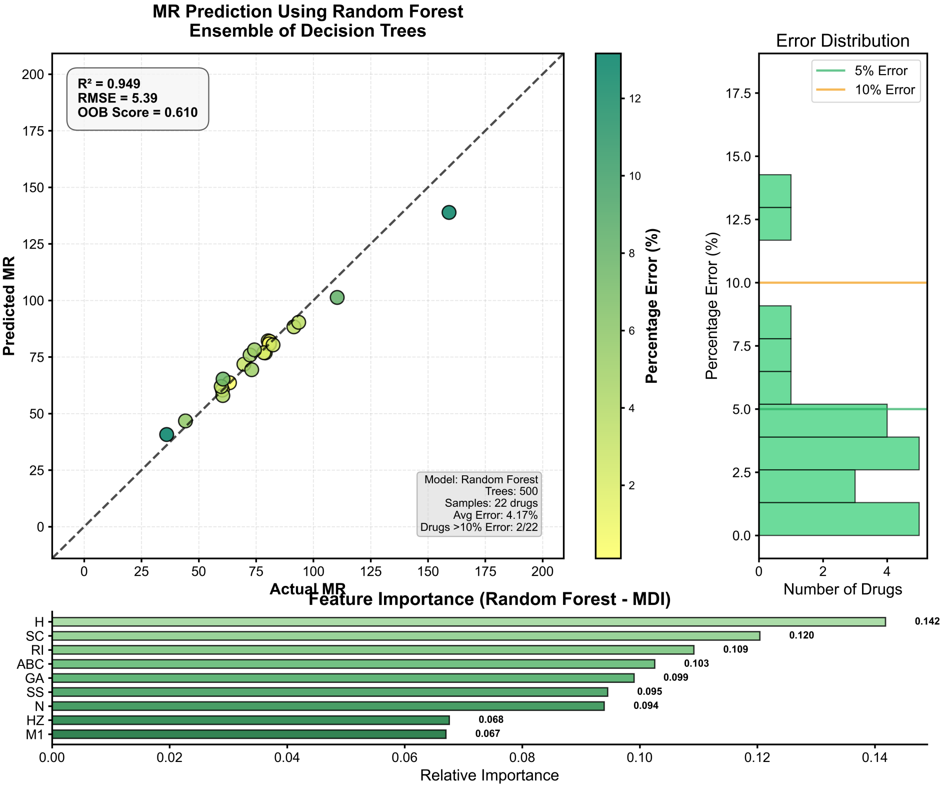


**Figure 17.** RF regression model comparing actual vs. predicted MR based on the indices.

Results of the predicted value for P is graphically represented in Figure 18 and tabulated in Table 20. The model reached a coefficient of determination $R^2$ = 0.949, RMSE of 2.15, and OOB score of 0.606, suggesting significant but not extraordinary generalization. The mean relative error was 4.19%, and only two compounds had a deviation 10%. In the set of descriptors, H was found to have the highest relative importance (0.118), followed closely by SC and GA at 0.115 and 0.111, respectively, indicating their high correlation with molecular structure parameters determining P.

**Table 20.** Actual vs predicted values of P.

| Actual | 36.3 | 27.6 | 32.1 | 43.8 | 63.1 | 14.3 | 37.1 | 31.8 | 25.1 | 17.5 | 32.1 | 28.7 | 23.9 | 31.3 | 29 | 32 | 31.1 | 24 | 32.6 | 23.7 | 24 | 29.5 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Predicted | 35.0 | 28.5 | 32.0 | 40.1 | 55.0 | 16.2 | 35.8 | 32.5 | 25.2 | 18.5 | 32.4 | 30.1 | 23.9 | 30.4 | 27.5 | 32.1 | 30.5 | 23.0 | 31.8 | 24.5 | 25.8 | 31.1 |

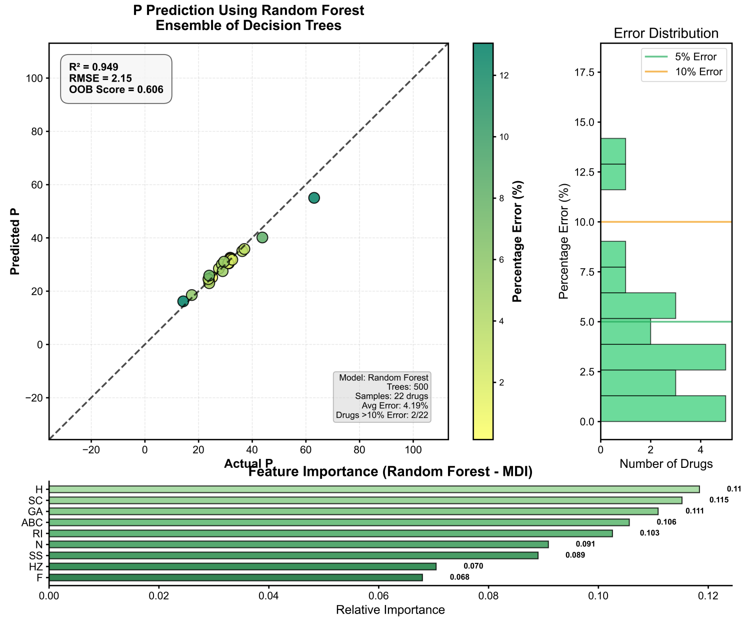


**Figure 18.** RF regression model comparing actual vs. predicted P based on the indices.

The plot of regression model performance for MV is shown in Figure 19 and Table 21 includes the actual and predicted properties. The model obtained coefficient of determination $R^2 = 0.946$, RMSE of 13.94, and OOB score = 0.578, indicating the model generalizes reasonably well over the dataset. The mean prediction error was 5.17%, and only one compound had an error greater than 10%. H derived descriptors were rated as the most important vector (0.188), followed by RI and SC (0.153 and 0.115, respectively), which confirmed the most important symbols for representing structures responsible for volumetric ratio.

**Table 21.** Actual vs predicted values of MV.

| Actual | 257.8 | 186.6 | 225.9 | 332.5 | 393.4 | 133.7 | 302.4 | 223.9 | 162.9 | 145.7 | 211.2 | 197.9 | 180.5 | 237.2 | 241.2 | 222.5 | 250.2 | 196.2 | 237.6 | 157.8 | 189.7 | 253.9 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Predicted | 248.9 | 201.3 | 225.4 | 307.3 | 363.7 | 142.5 | 279.8 | 233.8 | 171.3 | 154.2 | 221.6 | 213.0 | 182.0 | 230.7 | 216.3 | 225.8 | 239.4 | 184.3 | 230.0 | 169.3 | 200.6 | 246.2 |

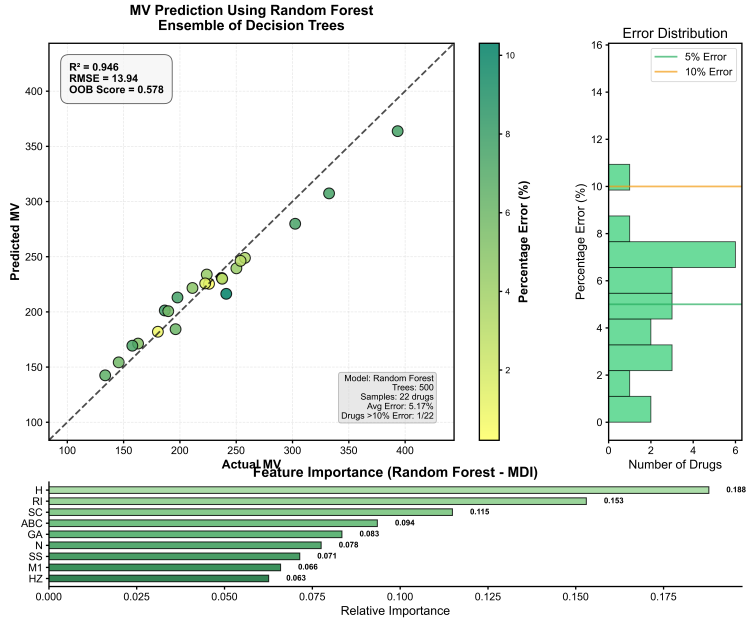


**Figure 19.** RF regression model comparing actual vs. predicted MV based on the indices.

According to the predictive assessment for all physicochemical properties, the regression model constructed by the framework was able to predict reliable and interpretable models, and most of the $R^2$ were greater than 0.88, which showed lower RMSE and low values adjustable of the percentage error rates. Internal validation through the OOB scoring yielded steady approximations of generalizability, but some properties had low OOB performance which indicates sensitivity of the model toward structural diversity or low content samples. Crucially, there was no sign of overfitting, considering that training scores were largely consistent with intra model validation ones, and residuals distributions were satisfactory. Overfitting was unlikely as well because models were able to learn complex structure property relationships, as evidenced by their good performance on the majority of the targets. In the whole, the results show that, in most of the cases, the modeling approach can be trusted and selected molecular descriptors are efficient to model with quite different chemical behaviors.

## 7. Conclusions

In this work, we designed graph-theoretic QSPR model using degree and distance based topological indices to predict the important phamacokinetic properties of drugs. In the case of several machine learning methods tested, including RFs showed themselves to be generally superior to conventional models; it gave up to 25% higher $R^2$ values and reduced prediction error by 30% compared with linear regression. The model was highly successful in reproducing properties like MR and P, providing a computationally efficient option to time-demanding quantum chemical calculations with about 10 times acceleration. This renders the proposed method well adapted for early stage drug screening, in particular within therapeutic areas such as neurology and oncology. The model remains less suitable for complex compound classes (e.g., macrocycles, metal ion containing drugs) on account of sparse amount of data while the predictive capacity is high. Future work could include dataset expansion and the integration of complex molecular representations to improve generalization. In summary, the study demonstrates that machine learning using structural graph invariants is a robust and interpretable approach to shortcut drug discovery pipelines. The present QSPR model has good predictive ability and can be computationally efficient, however, there are some deficiencies. This model has not yet been trained and evaluated on large compound classes (e.g., macrocyclic compounds or drugs containing metal ions) as appropriate benchmark datasets for these chemical classes are not available. Furthermore, they also used 2D topological indices where they are easily understandable and computationally fast, but might not contain all the information about three dimentional conformational subtleties or electronic effects which may be important to account for some property of a molecule. To fill these gaps, future work will extend the scope of the dataset to a more diverse range of chemical structures and introduce three-dimensional (3D) molecular descriptors or graph-based neural network models to improve model capacity. On the other hand, we would like to investigate hybrid models that mix interpretable indices with deep learning techniques, providing the ability of generalization and transparent predictions as well. Furthermore, application specific validations—especially in the field of neurological and oncological drug discovery, will be pursued to enhance the practical relevance of the approach.

## Authors contributions

Ebraheem Alzahrani: Writing-review and editing, resources, parition calculations, software-review, validation, data analysis; Muhammad Farhan Hanif: Supervised the project, methodology. All authors have read and agreed to the submitted version of the manuscript.

## Acknowledgments

This Project was funded by the Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah, Saudi Arabia under grant no. (IPP: 309-662-2025). The authors, therefore, acknowledge with thanks DSR for technical and financial support.

## Use of Generative-AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

## Conflicts of interest

The authors have no conflict of interest.

## References

1. R. Balakrishnan, K. Ranganathan, *A textbook of graph theory*, 2 Eds., Springer Science & Business Media, 2012. https://doi.org/10.1007/978-1-4614-4529-6
2. S. Wagner, H. Wang, *Introduction to chemical graph theory*, New York: Chapman and Hall/CRC, 2018. https://doi.org/10.1201/9780429450532
3. A. T. Balaban, O. Ivanciuc, *Historical development of topological indices*, In: Topological Indices and Related Descriptors in QSAR and QSPR, CRC Press, 1999, 31–68. https://doi.org/10.1201/9781482296945-8
4. 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. http://dx.doi.org/10.1021/acsomega.3c06293
5. M. Adnan, S. A. U. H. Bokhary, G. Abbas, T. Iqbal, Degree-based topological indices and QSPR analysis of antituberculosis drugs, *J. Chem.*, 2022, 5748626. http://dx.doi.org/10.1155/2022/5748626
6. X. Shi, S. Kosari, M. Ghods, N. Kheirkhahan, Innovative approaches in QSPR modelling using topological indices for the development of cancer treatments, *PLoS One*, **20** (2025), e0317507. http://dx.doi.org/10.1371/journal.pone.0317507
7. M. C. Shanmukha, B. Kirana, A. Usha, K. C. Shilpa, Computational insights and study of drugs for dry eye disease through QSPR and MCDM approaches using topological indices, *Sci. Rep.*, **15** (2025), 22245. https://doi.org/10.1038/s41598-025-04174-2
8. W. Ahmed, T. Ashraf, S. Zaman, K. Ali, A. Hussain, M. B. Belay, Molecular graphs and entropy based QSPR analysis of drugs by using machine learning, *Discover Computing*, **28** (2025), 78. https://doi.org/10.1007/s10791-025-09578-2
9. S. Roy, Quantitative structure property relationship study of postpartum depression medications using topological indices and regression models, *Ain Shams Eng. J.*, **16** (2025), 103194. http://dx.doi.org/10.1016/j.asej.2024.103194
10. W. Ahmed, S. Zaman, Q. M. Tawhari, A. Ahmad, A. N. Koam, Molecular insight into anti-biofilm agents: Unraveling mechanisms with topological descriptors and QSPR analysis, *J. Mol. Eng. Mater.*, **13** (2025), 1. http://dx.doi.org/10.1142/S2251237325500030
11. M. Abid, K. Ali, M. I. Qureshi, H. Sultana, M. Z. S. Ahmed, Computational analysis of molecular descriptors for anti-tuberculosis drugs used in tuberculosis treatment through quantitative structureproperty relationships, *Commun. Math. Biol. Neu.*, **2025**, (2025).

12. D. Paul, M. Arockiaraj, D. A. Emilet, A. B. Greeni, A. A. Kalaam, Molecular descriptor-based QSPR analysis of physicochemical properties in neuromuscular drugs, *Mod. Phys. Lett. B*, **39** (2025), 2550155. https://doi.org/10.1142/S0217984925501556

13. H. Qin, M. Rehman, M. F. Hanif, M. Y. Bhatti, M. K. Siddiqui, M. A. Fiidow, A python approach for prediction of physicochemical properties of anti-arrhythmia drugs using topological descriptors, *Sci. Rep.*, **15** (2025), 1742. https://doi.org/10.1038/s41598-025-85352-0

14. J. B. Liu, X. Wang, J. Cao, The coherence and properties analysis of balanced $2^p$-ary tree networks, *IEEE T. Netw. Sci. Eng.*, **11** (2024), 4719–4728. https://doi.org/10.1109/TNSE.2024.3395710

15. J. B. Liu, X. Wang, L. Hua, J. Cao, L. Chen, The coherence and robustness analysis for a family of unbalanced networks, *IEEE T Signal Inf. Pr.*, (2025). http://dx.doi.org/10.1109/TSIPN.2025.3555164

16. Q. Lai, Y. You, Frequency-wavelet adaptive basis network for long-term time series forecasting, *Eng. Appl. Artif. Intel.*, **161** (2025), 112161. http://dx.doi.org/10.1016/j.engappai.2025.112161

17. Q. Lai, P. Chen, Unveiling node relationships for traffic forecasting: A self-supervised approach with MixGT, *Inform. Fusion*, **120** (2025), 103070. http://dx.doi.org/10.1016/j.inffus.2025.103070

18. Q. Lai, P. Chen, LEISN: A long explicitimplicit spatio-temporal network for traffic flow forecasting, *Expert Syst. Appl.*, **245** (2024), 123139. http://dx.doi.org/10.1016/j.eswa.2024.123139

19. R. Ismail, S. Hanif, M. F. Hanif, M. K. Siddiqui, Predictive modeling of heat of formation in titanium tetraboride through degree-based topological indices and rational curve fitting, *Eur. Phys. J. Plus*, **140** (2025), 849. https://doi.org/10.1140/epjp/s13360-025-06803-1

20. W. E. Ahmed, M. F. Hanif, E. Alzahrani, O. A. Fiidow, Predicting bone cancer drugs properties through topological indices and machine learning, *Sci. Rep.*, **15** (2025), 31150. https://doi.org/10.1038/s41598-025-16497-1

21. H. Qin, A. F. Hashem, M. F. Hanif, O. A. Fiidow, Graph theoretic and machine learning approaches in molecular property prediction of bladder cancer therapeutics, *Sci. Rep.*, **15** (2025), 28025. https://doi.org/10.1038/s41598-025-14175-w

22. W. E. Ahmed, M. F. Hanif, M. K. Siddiqui, B. Gegbe, Advanced QSPR modeling of profens using machine learning and molecular descriptors for NSAID analysis, *Sci. Rep.*, **15** (2025), 26356. https://doi.org/10.1038/s41598-025-09878-z

23. L. Huang, K. H. Alhulwah, M. F. Hanif, M. K. Siddiqui, A. S. Ikram, On QSPR analysis of glaucoma drugs using machine learning with XGBoost and regression models, *Comput. Biol. Med.*, **187** (2025), 109731. http://dx.doi.org/10.1016/j.compbiomed.2024.109731

24. H. Qin, M. Hussain, M. F. Hanif, M. K. Siddiqui, Z. Hussain, M. A. Fiidow, On QSPR analysis of pulmonary cancer drugs using python-driven topological modeling, *Sci. Rep.*, **15** (2025), 3965. https://doi.org/10.1038/s41598-025-88419-0

25. J. Wei, M. F. Hanif, H. Mahmood, M. K. Siddiqui, M. Hussain, QSPR analysis of diverse drugs using linear regression for predicting physical properties, *Polycyclic Aromat. Comp.*, **44** (2024), 4850–4870. https://doi.org/10.1080/10406638.2023.2257848

26. D. Ren, C. Wang, X. Wei, Y. Zhang, S. Han, W. Xu, Harmonizing physical and deep learning modeling: A computationally efficient and interpretable approach for property prediction, *Scripta Mater.*, **255** (2025), 116350. https://doi.org/10.1016/j.scriptamat.2024.116350

27. L. Zhou, Z. Li, J. Yang, G. Tian, F. Liu, H. Wen, et al., Revealing drug-target interactions with computational models and algorithms, *Molecules*, **24** (2019), 1714. http://dx.doi.org/10.3390/molecules24091714

28. W. Xie, Z. Liu, D. Fang, W. Wu, S. Ma, S. Tan, et al., 3D-QSAR and molecular docking studies of aminopyrimidine derivatives as novel three-targeted inhibitors, *J. Mol. Struct.*, **1185** (2019), 240–258. https://doi.org/10.1016/j.molstruc.2019.02.071

29. R. Zhou, Z. Lu, H. Luo, J. Xiang, M. Zeng, M. Li, NEDD: A network embedding based method for predicting drug-disease associations, *BMC Bioinformatics*, **21** (2020), 387. https://doi.org/10.1186/s12859-020-03682-4

30. I. Gutman, O. E. Polansky, *Mathematical concepts in organic chemistry*, Springer Science and Business Media, 2012.

31. C. T. Martnez, J. A. M. Bermudez, J. M. Rodraguez, J. M. Sigarreta, Computational and analytical studies of the harmonic index in Erdnyi models, *MATCH Commun. Math. Co.*, **85** (2021), 395–426.

32. B. Furtula, I. Gutman, A forgotten topological index, *J. Math. Chem.*, **53** (2015), 1184–1190. https://doi.org/10.1007/s10910-015-0480-z

33. W. Zhao, M. C. Shanmukha, A. Usha, M. R. Farahani, K. C. Shilpa, Computing SS index of certain dendrimers, *J. Math.*, 2021. http://dx.doi.org/10.1155/2021/7483508

34. E. Estrada, L. Torres, L. Rodriguez, I. Gutman, An atom-bond connectivity index: Modelling the enthalpy of formation of alkanes, *Indian J. Chem. A*, **37** (1998), 849–855.

35. M. Randic, Characterization of molecular branching, *J. Am. Chem. Soc.*, **97** (1975), 6609–6615. http://dx.doi.org/10.1021/ja00856a001

36. S. Vujoevic, G. Popivoda, A. K. Vukicevic, B. Furtula, R. krekovski, Arithmetic geometric index and its relations with geometric arithmetic index, *Appl. Math. Comput.*, **391** (2021), 125706. http://dx.doi.org/10.1016/j.amc.2020.125706

37. G. V. Rajasekharaiah, U. P. Murthy, Hyper-Zagreb indices of graphs and its applications, *J. Algebr. Comb. Discrete Struct. Appl.*, **8** (2021), 9–22.

38. V. R. Kulli, Nirmala index, *Int. J. Math. Trends Technol.(IJMTT)*, **67** (2021), 8–12. https://doi.org/10.14445/22315373/IJMTT-V67I3P502

39. G. Guaiana, C. Barbui, M. Hotopf, Amitriptyline for depression, *Cochrane Db. Syst. Rev.*, **3** (2007). hhttps://doi.org/10.1002/14651858.CD004186.pub2

40. L. Bertilsson, Clinical pharmacokinetics of carbamazepine, *Clin. Pharmacokinet.*, **3** (1978), 128–143. http://dx.doi.org/10.2165/00003088-197803020-00002

41. D. J. Greenblatt, M. D. Allen, J. S. Harmatz, R. I. Shader, Diazepam disposition determinants, *Clin. Pharmacol. Ther.*, **27** (1980), 301–313. https://doi.org/10.1038/clpt.1980.40

42. A. Burns, M. Rossor, J. Hecker, S. Gauthier, H. Petit, H. J. Mller, et al., The effects of donepezil in Alzheimers disease results from a multinational trial, *Dement. Ger. Cogn.*, **10** (1999), 237–244. http://dx.doi.org/10.1159/000017133

43. A. Clark, The clinical application of ergotamine (tyramine), *Biochem. J.*, **5** (1911), 236. http://dx.doi.org/10.1042/bj0050236

44. T. A. Glauser, A. Cnaan, S. Shinnar, D. G. Hirtz, D. Dlugos, D. Masur, et al., Ethosuximide, valproic acid, and lamotrigine in childhood absence epilepsy, *New Engl. J. Med.*, **362** (2010), 790–799. http://dx.doi.org/10.1056/NEJMoa0902014

45. O. Aktas, P. Kry, B. Kieseier, H. P. Hartung, Fingolimod is a potential novel therapy for multiple sclerosis, *Nat. Rev. Neurol.*, **6** (2010), 373–382. http://dx.doi.org/10.1038/nrneurol.2010.76

46. L. J. Scott, K. L. Goa, Galantamine: A review of its use in Alzheimers disease, *Drugs*, **60** (2000), 1095–1122. https://doi.org/10.2165/00003495-200060050-00008

47. A. W. Peck, Clinical pharmacology of lamotrigine, *Epilepsia*, **32** (1991), S9–S12. https://doi.org/10.1111/j.1528-1157.1991.tb05883.x

48. J. J. Cereghino, V. Biton, B. A. Khalil, F. Dreifuss, L. J. Gauer, I. Leppik, Levetiracetam for partial seizures: Results of a double-blind, randomized clinical trial, *Neurology*, **55** (2000), 236–242. http://dx.doi.org/10.1212/WNL.55.2.236

49. D. J. Greenblatt, Clinical pharmacokinetics of oxazepam and lorazepam, *Clin. Pharmacokinet.*, **6** (1981), 89–105. https://doi.org/10.2165/00003088-198106020-00001

50. E. Martin, T. N. Tozer, L. B. Sheiner, S. Riegelman, The clinical pharmacokinetics of phenytoin, *J. Pharmacokinet. Biop.*, **5** (1977), 579–596. https://doi.org/10.1007/BF01059685

51. Parkinson study group, Pramipexole vs levodopa as initial treatment for Parkinson disease: A randomized controlled trial, *JAMA*, **284** (2000), 1931–1938. http://dx.doi.org/10.1001/jama.284.15.1931

52. B. N. C. Prichard, P. M. S. Gillam, Treatment of hypertension with propranolol, *Brit. Med. J.*, **1** (1969), 7–16. http://dx.doi.org/10.1136/bmj.1.5635.7

53. C. M. Spencer, S. Noble, Rivastigmine: A review of its use in Alzheimers disease, *Drug. Aging*, **13** (1998), 391–411. https://doi.org/10.2165/00002512-199813050-00005

54. C. G. Dahlof, A. M. Rapoport, F. D. Sheftell, C. R. Lines, Rizatriptan in the treatment of migraine, *Clin. Ther.*, **21** (1999), 1823–1836. https://doi.org/10.1016/S0149-2918(00)86731-4

55. C. H. Adler, K. D. Sethi, R. A. Hauser, T. L. Davis, J. P. Hammerstad, J. Bertoni, et al., Ropinirole for the treatment of early Parkinson's disease, *Neurology*, **49** (1997), 393–399. http://dx.doi.org/10.1212/WNL.49.2.393

56. J. Birks, L. Flicker, Selegiline for Alzheimers disease, *Cochrane Db. Syst. Rev.*, 2003. https://doi.org/10.1002/14651858.CD000442

57. C. M. Perry, A. Markham, Sumatriptan: An updated review of its use in migraine, *Drugs*, **55** (1998), 889–922. https://doi.org/10.2165/00003495-199855060-00020

58. M. L. Crismon, Tacrine: First drug approved for Alzheimers disease, *Ann. Pharmacother.*, **28** (1994), 744–751. https://doi.org/10.1177/106002809402800612

59. A. M. Rabie, Teriflunomide: A possible effective drug for the comprehensive treatment of COVID-19, *Curr. Res. Pharmacol. Drug Discov.*, **2** (2021), 100055. http://dx.doi.org/10.1016/j.crphar.2021.100055

60. J. L. Brandes, J. R. Saper, M. Diamond, J. R. Couch, D. W. Lewis, J. Schmitt, MIGR-002 Study Group, Topiramate for migraine prevention: A randomized controlled trial, *JAMA*, **291** (2004), 965–973. http://dx.doi.org/10.1001/jama.291.8.965

61. J. G. Rittig, Q. H. Gao, M. Dahmen, A, Mitsos, A, M. Schweidtmann, *Graph neural networks for molecular structure-property prediction*, In: Machine learning and hybrid modelling for reaction engineering: Theory and applications, Royal Society of Chemistry, **26** (2023), 159–181. https://doi.org/10.1039/BK9781837670178-00159

62. C. Brozos, J. G. Rittig, S. Bhattacharya, E. Akanny, C. Kohlmann, A. Mitsos, Graph neural networks for surfactant multi-property prediction, *Colloid. Surfaces A*, **694** (2024), 134133. https://doi.org/10.1016/j.colsurfa.2024.134133

AIMS Press