

Article

Not peer-reviewed version

Topological Characterization of Parkinson's Disease Drugs: A Graph-Theoretical Pilot Study

[Sakthivela A](#) and [Kavitha K](#)*

Posted Date: 12 March 2026

doi: 10.20944/preprints202603.0785.v1

Keywords: Parkinson's disease; QSPR; artificial neural network; topological indices; random forest; molecular descriptors



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a [Creative Commons CC BY 4.0 license](#), which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Topological Characterization of Parkinson's Disease Drugs: A Graph-Theoretical Pilot Study

Sakthivela A ¹ and Kavitha K ^{2,*}

¹ Research Scholar, Department of Mathematics, School of Advanced Sciences (SAS), Vellore Institute of Technology (VIT), Vellore, India

² Associate Professor, Department of Mathematics, School of Advanced Sciences (SAS), Vellore Institute of Technology (VIT), Vellore, India

* Correspondence: kavinphd@gmail.com

Abstract

Parkinson's disease (PD) is a neurodegenerative disorder with limited disease-modifying therapies. Computational models can provide predictive insights into drug properties, although critically limited datasets pose challenges. Fifteen FDA-approved Parkinson's disease drugs were represented as hydrogen-suppressed molecular graphs. Twelve degree-based topological indices were computed and used as descriptors for predicting seven physicochemical properties (MR, P, MV, MW, nHA, nRotB, Complexity). Multi-layer perceptron artificial neural network (ANN) and Random Forest (RF) models were trained. Model performance was evaluated using Leave-One-Out Cross-Validation (LOOCV). The statistical robustness of the models was verified using a Y-randomization test. Shapley Additive Explanations (SHAP) were applied for interpretability. The ANN demonstrated high predictive correlation on the small dataset for MR ($R^2 = 0.876$), P ($R^2 = 0.875$), MW ($R^2 = 0.837$), and nHA ($R^2 = 0.901$). Lower predictive performance was observed for MV ($R^2 = 0.729$), molecular Complexity ($R^2 = 0.706$), and nRotB ($R^2 = 0.308$). RF provided comparable results but was generally outperformed by ANN. The Y-randomization test yielded consistently negative average R^2_{rand} values (lowest $R^2_{\text{rand}} = -1.708$), confirming the absence of chance correlation. SHAP analysis identified the most influential topological indices for each property in ANN. ANN-based QSPR modeling with degree-based descriptors can accurately predict physicochemical properties of PD drugs for certain endpoints. These models were proven statistically robust through Y-randomization validation. Limitations include the small dataset size and high-dimensional descriptor space, highlighting the need for external validation, larger datasets, and inclusion of additional 3D/quantum descriptors for more complex pharmacokinetic endpoints.

Keywords: Parkinson's disease; QSPR; artificial neural network; topological indices; random forest; molecular descriptors

1. Introduction

The incorporation of computational technologies into the area of healthcare and pharmaceutical sciences revolutionized in the way researchers think about solving intricate biomedical issues. In particular, artificial intelligence (AI) and machine learning (ML) have revolutionised the classical approaches to drug discovery, replacing the method of labour-intensive trial-and-error experimentation with data-driven, predictive, and optimised methodologies.[1,2] These computational advances enable researchers to exploit large-scale biomedical and chemical data sets and detect relationships that would otherwise be undetectable using conventional analytical strategies.[3] Drug research has entered a new era of increased precision and efficiency by using the power of machine learning to detect the hidden patterns and non-linear correlations in multidimensional data, creating new possibilities for previously incurable diseases.

Among such conditions, neurodegenerative diseases like Parkinson's disease is a critical challenge. Characterised by progressive neuronal loss with diverse etiologies, Parkinson's disease poses great challenges in terms of early diagnosis, therapeutic efficacy, and correct prediction of disease progression.[4] Its multifactorial nature of genetic, molecular and environmental factors makes traditional research methodologies inadequate for elucidating the underlying mechanisms or for systematic prediction of the drug-related properties.[4,5] Recent advances in data-based research, however, are starting to provide alternative structures. Researchers have begun to refine early indicators for diagnosis, better understand therapy targets and evaluate medication response through a combination of biological data and computational modelling that are not possible with traditional laboratory techniques.[6,7]

In this context, computational chemistry has emerged as an essential area of research in the development of drugs, particularly for neurodegenerative diseases. Molecular modelling has, however, moved beyond the static representation and is now more a mathematical simulation of the molecular behaviour that enables the investigation of the behaviour at the atomic level with biological targets.[8] Central to these models are molecular descriptors and structural fingerprints, which are chemical structures that are transformed into a number vector that describes the chemical properties, such as the connectivity of the atoms, the distribution of the electrons, and the three-dimensional conformations.[9] These descriptors have predictive value in predicting the potential of molecules to interact with enzymes, receptors, or misfolded proteins involved in the pathogenesis of Parkinson's disease, and present an analytical interface between theoretical chemistry and drug development.[10]

The combination of such chemical descriptors with sophisticated machine learning algorithms, especially using Quantitative Structure Activity Relationship (QSAR) and quantitative structure property relationship (QSPR) models, has revolutionised computational pharmacology. These approaches allow biological activity or physicochemical properties to be predicted systematically on the basis of chemical structure, increasing the efficiency of the molecular datasets analysis.[11] While deep learning architectures like DNNs and GNNs are frequently used for large-scale feature extraction, in this study, the focus is on the classical ANN approach, enunciating more on the use of descriptor-based predictive modelling, than on full-scale drug discovery.[12,13] Predictive modelling of physicochemical characteristics of drug molecules can yield valuable information on the behaviour of drugs in relation to pharmacokinetics and drug formulation design and thus complement experimental data.[14,15]

The present study uses ANN and Random Forest (RF) models trained using predefined degree-based topological indices for predicting physicochemical properties of a small and well-defined set of approved drugs of Parkinson's Disease. Unlike Graph Neural Networks (GNNs) that learn features directly from the molecular graphs, in this study the models are based on precomputed molecular descriptors and are used to predict physicochemical properties for a small and well-defined set of approved Parkinson's disease drugs. Model performance is assessed with the help of standard regression metrics like Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and coefficient of determination (R^2). In order to interpret the contribution of molecular descriptors to the predictions of the model, the method of Shapley Additive Explanations (SHAP) is used, which gives information on the importance of the features and improves the interpretability of the models.

The challenge of making use of small, but high quality, datasets typical of specialized areas, such as Parkinson's Disease research, requires the use of sophisticated forms of machine learning, such as the Artificial Neural Networks (ANN) and the Random Forest (RF), which can be used to extract non-linear patterns. However, models based on small datasets have a higher risk of chance correlation/overfitting. Therefore, strict internal validation is of extreme importance. This study focuses on achieving robust QSPR models for seven vital physicochemical properties using topological indices as descriptors. Crucially, the predictive reliability and generalization ability of such models will be checked by implementation of the Y-randomization test which will ensure that the derived relationships are statistically significant and not artefacts of the selected data structure.

1.1. Motivation and Significance

Parkinson's disease (PD) is one of the most common neurodegenerative disorders, affecting millions of people around the world, and for which effective disease-modifying therapies are still lacking. The complexity of PD is due to its slow progression, heterogeneous symptoms and the difficulty of its early diagnosis. Therapeutic approaches are further limited by the multi-factoriality of its pathology that involves depletion of dopamine in the substantia nigra, mitochondrial dysfunction, and abnormal aggregation of alpha-synuclein proteins.[2,4] These challenges present the need to identify candidates for medications that are specific to disease pathology and can act early in the disease process. While traditional drug discovery methods are expensive and time-consuming, computational frameworks with the combination of molecular descriptors and ANN and RF models can offer predictive insights in an efficient way.

Advances in molecular modeling have enabled the simulation and analysis of drug-target interactions on a molecular level, providing insights into binding affinities, solubility and bioavailability. [6,8] Such methods are especially pertinent for PD, where candidate compounds would need to be optimized in order to cross the blood brain barrier and interact effectively with the dopaminergic receptors or to prevent protein misfolding and aggregation, respectively. In this context, computational chemistry facilitates the understanding of mechanisms and the prediction of properties, to diminish the dependence on trial-and-error experimentation.

A very promising approach to overcome these problems is the combination of Quantitative Structure--Property Relationship (QSPR) analysis with supervised learning machine learning techniques (ANNs). QSPR frameworks map chemical structures to numerical descriptors that encode material characteristics such as molecular shape, electronic properties, and hydrophobicity that in turn inform prediction of features of pharmacologic importance such as bioavailability and binding to receptors.[10] By combining these descriptors with ANN and RF models to build predictive models, they can be used to estimate physicochemical properties of drugs for Parkinson's disease, which will provide a basic computational tool to evaluate the molecular features. This integration leads to a better predictive accuracy and helps in the fast assessment of molecular datasets to aid in making informed decisions in preclinical research. [16]

While earlier studies have delved into the use of QSPR models for a diverse set of chemical data sets, their application to a narrow set of PD drugs comprising of a comprehensive set of degree-based topological indices with ANN and RF models has remained under-explored. Addressing such gap, the present work aims to validate and evaluate the predictive capability of such models to estimate key physicochemical properties of the Parkinson's disease drugs.

The following objectives have been designed:

1. To use degree based topological indices as part of QSPR models to quantify structural features of molecules in order to predict physicochemical properties of molecules of interest for Parkinson's disease drugs.
2. To train and validate ANN and RF models using the LOOCV technique to ensure realistic performance estimates on small data. To assess the predictive performance with standard metrics such as MSE, MAE, RMSE and R2 and compare the model efficiency and interpretability.
3. To validate the developed QSPR models using Y-randomization technique to exclude the possibility of chance correlation and confirm the robustness of the models.
4. To analyze the contribution of features using SHAP, giving interpretative information on the importance of molecular descriptors, which will aid useful computational analysis of drug candidates for PD.

2. Materials and Methods

A dataset that included 15 drugs generally used in the treatment of Parkinson's disease was chosen. These include Levodopa, Apomorphine, Pramipexole, Ropinirole, Rotigotine, Entacapone, Selegiline, Rasagiline, Amantadine, Safinamide, Bzotropine, Istradefylline, Pimavanserin,

Rivastigmine and Trihexyphenidyl. The drugs belong to multiple pharmacological classes (dopamine precursors, enzyme inhibitors, dopamine agonists), ensuring diversity in molecular structure, physicochemical properties and modes of action pertinent for QSPR modeling.

Molecular structures were illustrated as hydrogen-suppressed graphs by applying RDKit and NetworkX. Non-hydrogen atoms were nodes and chemical bonds were edges. The degrees of atoms and edge partitions (degree pairs for connected atoms) were calculated for all of the compounds. These edge partitions are used as the foundation for calculating topological descriptors (as shown in Figures 1 to 15). These descriptors are calculated with Python 3.12.0, and constitute the structural input for further modelling. The computation was done using some of the core scientific libraries such as NumPy, NetworkX for processing graph data and scikit-learn for implementing machine learning. The experimental physicochemical properties of each drug Molar Refractivity (MR), Polarizability (P), Molar Volume (MV), Molecular Weight (MW), Heavy Atom Count (nHA), Rotatable Bond Count (nRotB) and Complexity (C) were retrieved from automated Python scripts of public chemical databases including PubChem (<https://pubchem.ncbi.nlm.nih.gov>) and ChemSpider (<http://www.chemspider.com>). These experimentally verified properties were used as target variables for the QSPR models as seen in Table 3.

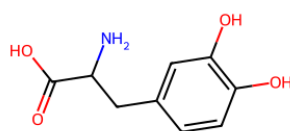


Figure 1. Levodopa.

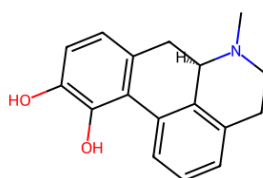


Figure 2. Apomorphine.

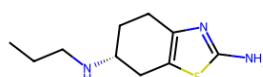


Figure 3. Pramipexole.

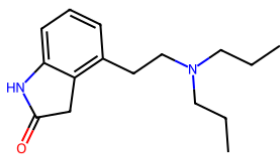


Figure 4. Ropinirole.

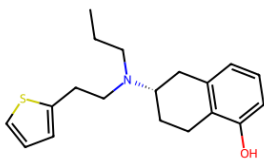


Figure 5. Rotigotine.

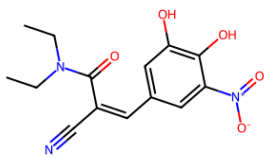


Figure 6. Entacapone.

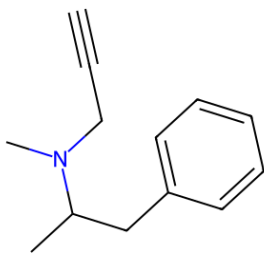


Figure 7. Selegiline.

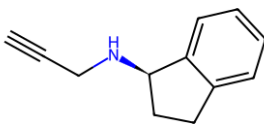


Figure 8. Rasagiline.



Figure 9. Amantadine.

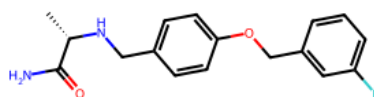


Figure 10. Safinamide.

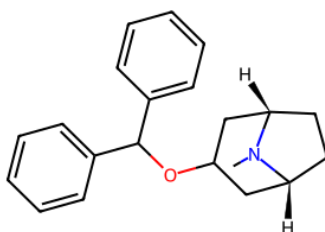


Figure 11. Benztropine.

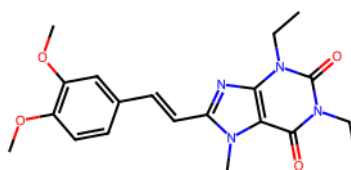


Figure 12. Istradefylline.

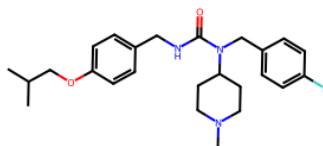


Figure 13. Pimavanserin.

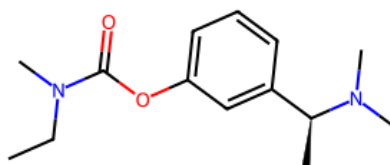


Figure 14. Rivastigmine.

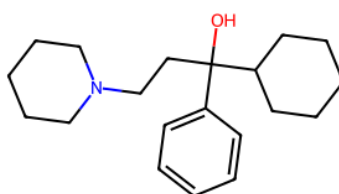


Figure 15. Trihexyphenidyl.

2.1. Calculation of Topological Indices

From the hydrogen-suppressed molecular graphs, a full set of 12 degree-based topological indices was calculated the first and second Zagreb indices (M1, M2), the harmonic index (H), the forgotten index (F), the Shilpa-Shanmukha index (SS), the atom-bond connectivity index (ABC), the Randic index (RI), the sum connectivity index (SC), the geometric-arithmetic index (GA), the hyper-Zagreb index (HZ) and the redefined Zagreb indices (ReZ1 and ReZ2). All indices were calculated with custom Python algorithms that were cross validated with manual calculations to ensure accuracy. The resulting dataset of indices was then used as the descriptor matrix in QSPR modelling. The calculated indices are presented in Table 1. With 15 compounds and 12 topological indices, the research works on a high dimension and low sample size (HDLSS) regime. This poses a great danger

of over-fitting which is taken care of by the robust Leave-One-Out Cross-Validation (LOOCV)-strategy explained in the later sections.

Table 1. Topological Indices.

Authors	Formula	Indices
Gutman and Polansky [17]	$M_1(G) = \sum_{uv \in E(G)} (d_u + d_v)$	First Zagreb index
Gutman and Polansky [17]	$M_2(G) = \sum_{uv \in E(G)} (d_u \cdot d_v)$	Second Zagreb index
Fajtlowicz [18]	$H(G) = \sum_{uv \in E(G)} \frac{2}{d_u + d_v}$	Harmonic index
Furtula and Gutman [19]	$F(G) = \sum_{uv \in E(G)} (d_u)^2 + (d_v)^2$	Forgotten index
Zhao [20]	$SS(G) = \sum_{uv \in E(G)} \sqrt{\frac{d_u \cdot d_v}{d_u + d_v}}$	Shilpa-Shanmukha index
Estrada et al. [21]	$ABC(G) = \sum_{uv \in E(G)} \sqrt{\frac{d_u + d_v - 2}{d_u \cdot d_v}}$	Atom bond connectivity index
M. Randic et al.[22]	$RI(G) = \sum_{uv \in E(G)} \frac{1}{\sqrt{d_u \cdot d_v}}$	Randic index
Farahani [23]	$SC(G) = \sum_{uv \in E(G)} \frac{1}{\sqrt{d_u + d_v}}$	Sum connectivity index
Vukicevic et al. [24]	$GA(G) = \sum_{uv \in E(G)} \frac{2\sqrt{d_u \cdot d_v}}{d_u + d_v}$	Geometric arithmetic index
Rajasekharaiah et al. [25]	$HZ(G) = \sum_{uv \in E(G)} (d_u + d_v)^2$	Hyper-Zagreb index
Ranjini [26]	$ReZ1(G) = \sum_{uv \in E(G)} \frac{d_u \cdot d_v}{d_u + d_v}$	Redefined first Zagreb index
Ranjini [26]	$ReZ2(G) = \sum_{uv \in E(G)} (d_u \cdot d_v)(d_u + d_v)$	Redefined second Zagreb index

2.2. Extraction of Physicochemical Properties

This study was focused on seven experimentally verified physicochemical properties, such as Molar Refractivity (MR), Polarizability (P), Molar Volume (MV), Molecular Weight (MW), Heavy Atom Count (nHA), Rotatable Bond Count (nRotB), and Molecular Complexity (C), in view of their

direct relevance to Central Nervous System (CNS) pharmacology properties in Parkinson’s disease (PD) therapeutics. PD is characterized by the gradual deterioration of dopaminergic neurons in the substantia nigra, and to be effective, drugs must cross the blood-brain barrier (BBB) to execute their beneficial effects within the CNS. The selected properties represent important aspects of molecular behavior that can affect BBB permeability, binding affinity to receptors, solubility, and flexibility of structures that are important for CNS-targeted activity. Data was collected using API based queries from PubChem and chemspider databases and presented in Table 2. Topological indices were calculated according to reference indices and the obtained values are shown in Table 3.

Table 2. Physiochemical properties of drugs.

Drugs	Molar Refractivit y (cm³)	Polarizabi lity (cm³)	Molar Volume (cm³)	Molecular Weight (g/mol)	Heavy Atom Count	Rotatab le Bond Count	Compl exity
Levodopa	49.3	19.5	134.3	197.19	14	3	209
Apomorphine	77.9	30.9	205.6	267.32	20	0	374
Pramipexole	60.3	23.9	180.5	211.33	14	3	188
Ropinirole	78.4	31.1	250.2	260.37	19	7	286
Rotigotine	94.6	37.5	272.4	315.5	22	6	337
Entacapone	79.1	31.4	219.2	305.29	22	4	500
Selegiline	60.5	24	196.2	187.28	14	4	195
Rasagiline	53.9	21.4	162.7	171.24	13	2	212
Amantadine	45.7	18.1	141.8	151.25	11	0	144
Safinamide	83.3	33	254.2	302.34	22	7	346
Benzotropine	94.6	37.5	274.8	403.5	28	4	433
Istradefylline	105.6	41.8	309.2	384.4	28	6	613
Pimavanserin	123	48.7	371.8	427.6	31	8	523
Rivastigmine	73.1	29	241.2	250.34	18	5	269
Trihexyphenid yl	91.8	36.4	289.7	301.5	22	5	314

Table 3. Calculated values of topological indices.

	M1	M2	H	F	SS	ABC	RI	SC	GA	HZ	ReZ1	ReZ2
Levodopa	66	73	6.066667	172	14.3522 9	10.3655 6	6.50290 8	6.49977 8	13.2089	318	14.95	364
Apomorph ine	116	145	9.366667	310	25.1596	16.2681	9.66470 4	10.3437	22.4364 4	600	17.3476 2	474
Pramipexo le	70	80	6.6	174	15.7662 7	10.6755 5	6.79202 5	7.01052 1	14.627	334	27.85	772
Ropinirole	92	105	9	226	20.9029 6	14.1706 5	9.24071 3	9.44333 3	19.5496	436	16.7166 7	390
Rotigotine	112	130	10.56667	276	25.4066 5	16.9586 3	10.7751 9	11.2214 4	23.5865 9	536	22.0833 3	512
Entacapon e	104	120	9.766667	272	22.7623 5	15.9011	10.3489 7	10.2982 1	20.9767 2	512	27.1166 7	638
Selegiline	62	67	6.5	148	14.2505 2	10.0778 3	6.73638 2	6.72166 7	13.5738 4	282	24.05	618
Rasagiline	64	74	6.333333	154	14.7432 1	9.81861 5	6.41501 5	6.62991 5	13.8417 9	302	14.6666 7	314
Amantadi ne	68	82	5	194	14.2175 3	9.35130 7	5.23489 5	5.69688 1	12.4465 9	358	15.6666 7	360

Safinamide	106	117	10.13333	262	23.83419	16.66058	10.54171	10.77481	22.22165	496	15.6	434
Benzotropine	124	147	11.13333	310	27.91035	18.33241	11.30966	11.99956	25.62358	604	24.9	558
Istradefylline	146	178	12.96667	384	32.03211	21.17741	13.45946	13.85026	29.12686	740	30.15	732
Pimavanserin	154	173	14.36667	384	34.49763	23.80059	14.90189	15.36634	31.96645	730	34.81667	938
Rivastigmine	84	94	7.966667	216	18.48897	13.15355	8.451596	8.43259	17.13151	404	36.35	836
Trihexyphenidyl	112	130	10.60476	282	25.33586	17.00626	10.78864	11.24189	23.58087	542	19.31667	470

2.5. Predictive Modeling

Before ANN and RF models were used, all the computed topological descriptors were normalized using z-score standardization to make mean of zero and standard deviation of one. The target physicochemical properties were normalized to a [0,1] range using Min-Max normalization because faster convergence is achieved during training on the normalized data and scale-related bias is reduced. Targets with strong skew (ratio of max/min > 100) were log-transformed prior to modelling and inverse-transformed after prediction.

The relationship between the computed topological indices and physicochemical properties of the drugs used in Parkinson’s disease was modeled by two supervised learning algorithms, the artificial neural networks (ANNs) and a random forest (RF) regressor. The ANN has been selected because of its ability to capture nonlinear relationships and complex interactions in relatively small datasets and is therefore suitable for this application. The ANN model was applied as a multi-layer perceptron with two hidden layers of eight neurons each using ReLU activation functions. Learning rate was increased to 0.005 with 1000 iterations max while hyperparameters were tuned for reaching a balance between under and over fitting to reach the goal of stable convergence. The RF model consisted of 100 decision trees with a maximum depth of six, resulting in robustness against overfitting in our high-dimensional low-sample-size (HDLSS) dataset. With the small size of the dataset (15 compounds), the classical train-test split was unsuitable. To get a reliable estimate of the predictive performance, the Leave-One-Out Cross-Validation (LOOCV) was used. In this method, each of the drugs was sequentially used as a test sample and the rest of the compounds as a training set. This process was repeated for all drugs by ensuring that a single evaluation each for unseen data was performed. Model predictions of LOOCV folds were summed to get the overall performance metrics. Model performance was assessed using several regression metrics: mean squared error (MSE), mean absolute error (MAE), root mean squared error (RMSE) and the coefficient of determination (R²). These metrics offered an overall view of the prediction accuracy and generalization of the model. Where required, target properties with a high level of skew were log-transformed before the model was trained and inverse-transformed after predictions were made to ensure consistency with the experimental scale.

2.6. Model Interpretability

To understand the role of single topological indices in predicting the contributions to the properties, the Shapley Additive Explanations (SHAP) were used. For the ANN models, KernelExplainer was used and for the RF models, TreeExplainer was employed. SHAP values for each LOOCV fold were calculated and aggregated to get mean absolute contributions and point out the most influential descriptors for all compounds. For each physicochemical property, SHAP-based bar plots have been generated, which provide a visual depiction of the relative importance of each topological index, making the mechanistic interpretation of the predictive models easier. High-resolution SHAP plots for each property and model were created which can clearly visualize the

influence of the descriptor. This workflow guarantees that, in spite of the high-dimensional and low-sample-size of the data, the predictive models are rigorously validated, interpretable and able to generate meaningful information on the structure-property relationships of Parkinson's disease drugs.

2.7. Model Validation: Y-Randomization Test

To ensure the robustness of the models of QSPR developed and to avoid results due to a chance correlation between the descriptors and the dependent variables (overfitting in the regime high-dimensional and low-sample-size), a Y-randomization test was carried out. This procedure involved the random shuffling (permutation) of the values of the experimental target property(s) (Y). Both ANN and RF models were then re-trained based on these randomly assigned Y-values. This process was repeated five dozen times (independent permutations of each of the seven properties).

The resulting predictive performance was determined by the average coefficient of determination (R^2_{rand}) of the 50 models. A valid and robust QSPR model must show that its original R^2 is significantly larger than the R^2_{rand} , and the latter should be equal to zero or highly negative. A negative R^2_{rand} is a signal that models based on randomized data are worse than a trivial model based on mean prediction, based on which the original model is proved to capture meaningful relations between structural properties.

3. Results

3.1. Artificial Neural Network (ANN) Model Performance

The ANN model was able to show high predictive power of the selected physicochemical properties of the 15 Parkinson's disease medicines. Degree-based topological indices were taken as input features, which allowed ANN to detect the nonlinear relationship between molecular structure and physicochemical endpoints. The predicted values for each property are shown in Table 4. Notably, Istradefylline and Pimavanserin had higher values for molar refractivity (MR) and polarizability (P) which were consistent with their larger and more complex molecular structures, whereas smaller molecules such as Amantadine and Levodopa had lower MR and P values.

The overall predictive performance was measured in terms of standard metrics of regression (MSE, MAE, RMSE, R^2). For most properties, the ANN showed high R^2 values, showing good agreement between predicted and experimental values (Table 5). Molar Refractivity (MR) was predicted with a R^2 value of 0.876, RMSE value of 7.40, MAE value of 5.24 showing high correlation with experimental value. Polarizability (P) had R^2 of 0.875 and RMSE of 2.94, Molecular Weight (MW) and Molar Volume (MV) prediction were R^2 of 0.837 and 0.729, respectively. Heavy Atom Count (nHA) was well predicted (R^2 = 0.901), as compared to the number of Rotatable Bonds (nRotB) showing lower predictive power (R^2 = 0.308), indicating a higher structural variation among this property. Molecular Complexity (C) was moderately (R^2 = 0.706) predicted. Across all properties ANN maintained relatively low errors (MSE and RMSE) which indicates effective learning of nonlinear relationships between topological indices and physicochemical features. The predictive performance is visually verified in Figure 16 (a-g) which shows the Actual vs. Predicted plots of ANN model.

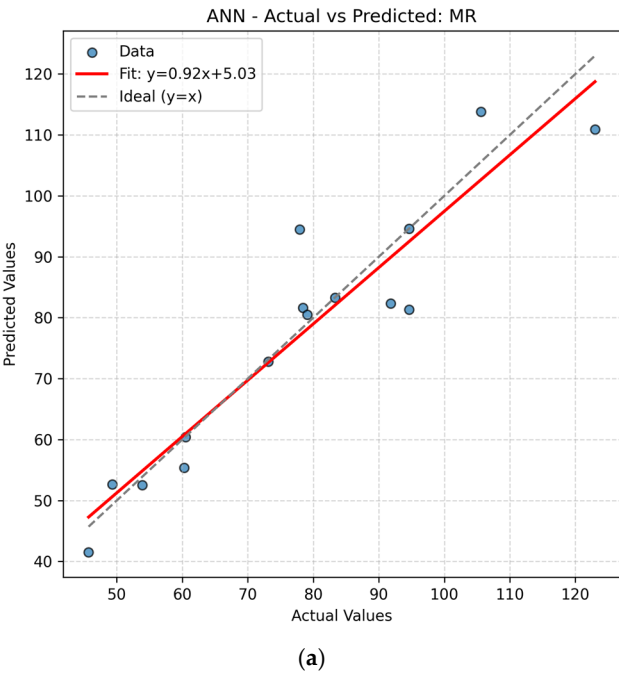
Table 4. Predicted properties of Molecular Structure Using Artificial Neural Network (ANN).

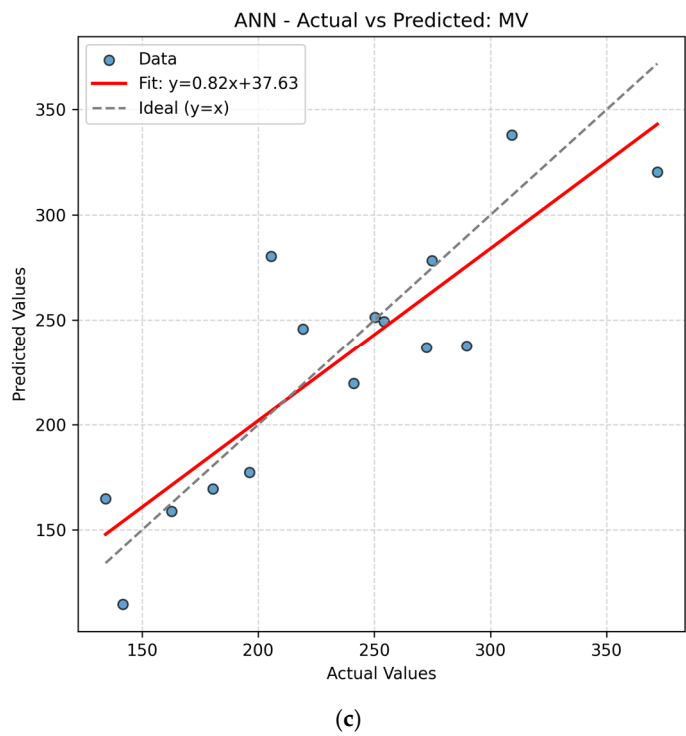
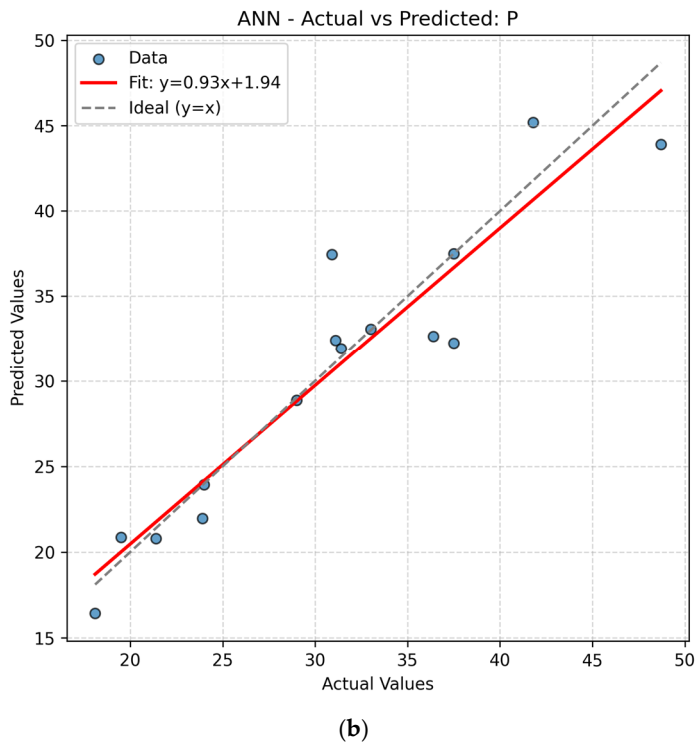
Drug	MR	P	MV	MW	nHA	nRotB	C
Levodopa	52.63117	20.87008	164.9146	170.4163	12.59502	1.64032	176.5882
Apomorphine	94.49452	37.44374	280.268	345.7778	24.61308	4.250875	447.1501
Pramipexole	55.37984	21.95077	169.4526	178.8109	13.11403	2.627078	189.2228
Ropinirole	81.64808	32.4055	251.3758	289.9559	20.7694	5.632905	338.4935
Rotigotine	81.33713	32.24891	236.763	294.2737	21.28952	4.828927	360.1011
Entacapone	80.47317	31.91037	245.7266	282.5815	20.33943	5.082306	292.2081

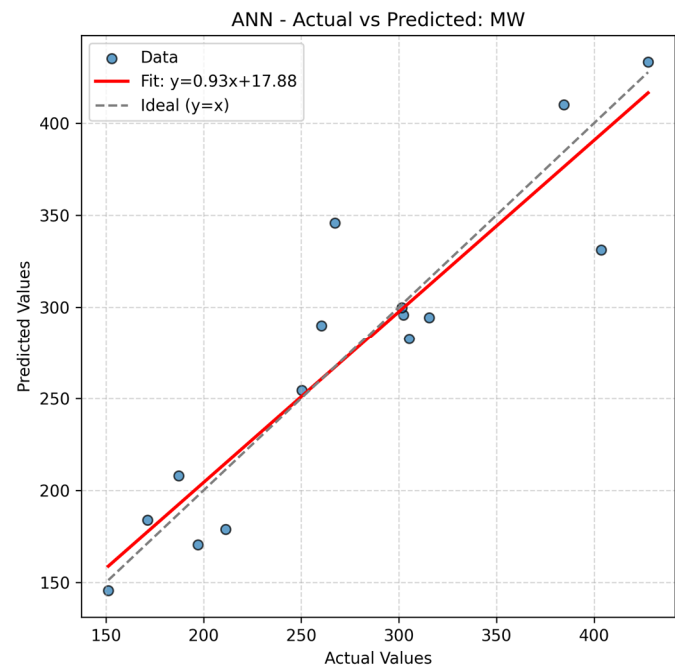
Selegiline	60.41703	23.94212	177.4657	208.0032	14.49071	4.750501	232.5675
Rasagiline	52.49665	20.78585	158.7776	183.9739	13.14485	3.143748	178.8192
Amantadine	41.47466	16.41405	114.6885	145.5297	10.53274	-2.09549	125.8131
Safinamide	83.30815	33.05338	249.2978	295.8306	21.15048	6.021882	324.9536
Benzotropine	94.61402	37.48776	278.3036	331.0505	24.03365	3.124366	427.4362
Istradefylline	113.7861	45.2003	337.968	409.9358	29.5562	1.400639	508.1925
Pimavanserin	110.8927	43.91051	320.4126	433.21	31.5906	6.458686	613.7453
Rivastigmine	72.76938	28.85653	219.5816	254.3704	17.81942	4.9572	268.3656
Trihexyphenidyl	82.32708	32.63992	237.6065	299.7029	21.53465	5.169024	367.5329

Table 5. Artificial Neural Network (ANN) Error metrics.

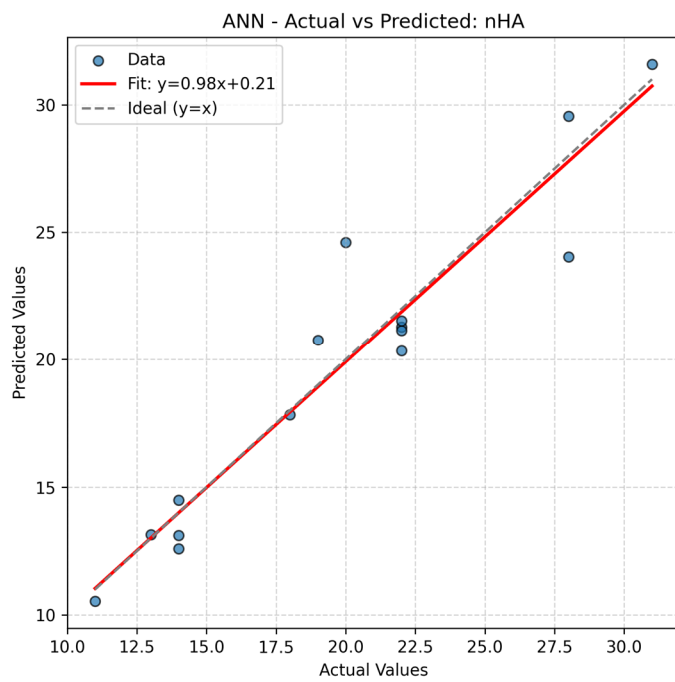
Properties	MSE	MAE	RMSE	R2
MR	54.8198	5.237384	7.404039	0.876329
P	8.661455	2.096463	2.943035	0.87531
MV	1093.113	26.11405	33.06227	0.728741
MW	1092.677	24.42538	33.05566	0.836963
nHA	3.359065	1.317061	1.832775	0.900763
nRotB	3.734373	1.453329	1.932453	0.30788
C	5218.14	50.36244	72.2367	0.706349



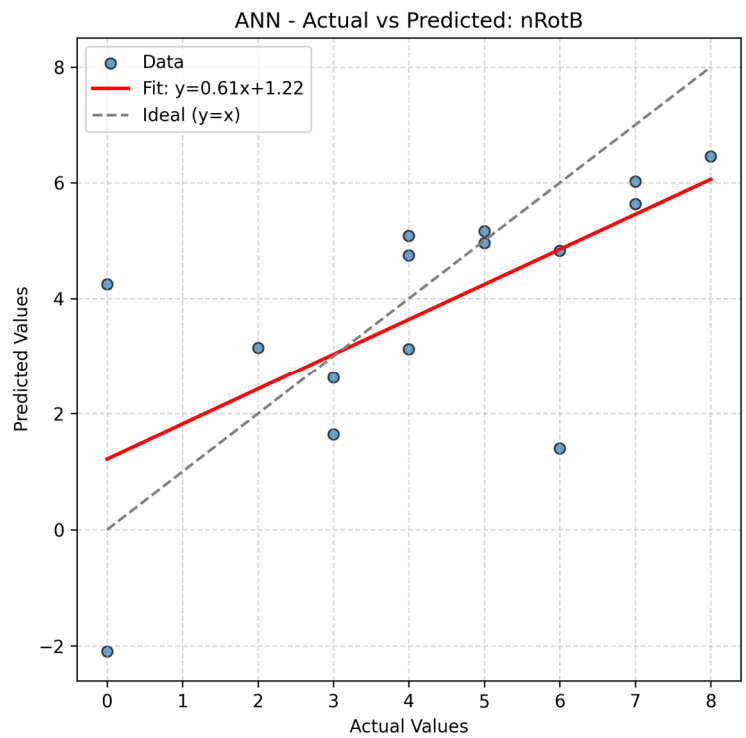




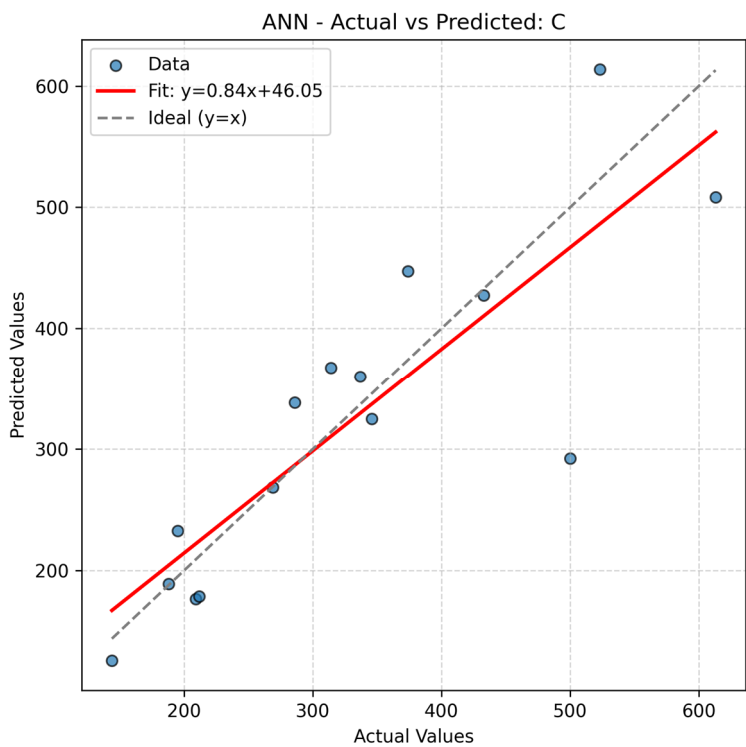
(d)



(e)



(f)



(g)

Figure 16. a–g illustrating Actual Vs Predicted (ANN).

3.2. Random Forest (RF) Model Performance

The Random Forest model also offered reliable predictions, which captured the non-linear dependencies between topological descriptors and physicochemical properties. The predicted values for 15 drugs for each property are summarized in Table 6. Trends that were observed using the ANN; increased MR and P values for Istradefylline and Pimavanserin and decreased values for Amantadine

and Levodopa, were consistently reproduced by the RF model. The RF model also showed good performance, but overall a little lower than ANN for most endpoints

Performance measures for RF (Table 7) revealed generally similar predictive accuracy to the ANN. The R² values of MR and P predictions were 0.856 and 0.857 respectively with RMSE values of 7.98 and 3.15 respectively. Molecular volume (MV) and weight (MW) were predicted with a moderate accuracy (R² = 0.742 and 0.793), whilst hydrogen atoms (nHA) predictions were reliable (R² = 0.839). However, predictions for nRotB and complexity (C) were less accurate (R² = 0.112 and 0.742), reflecting the inherent difficulty to model these properties from topological descriptors alone.

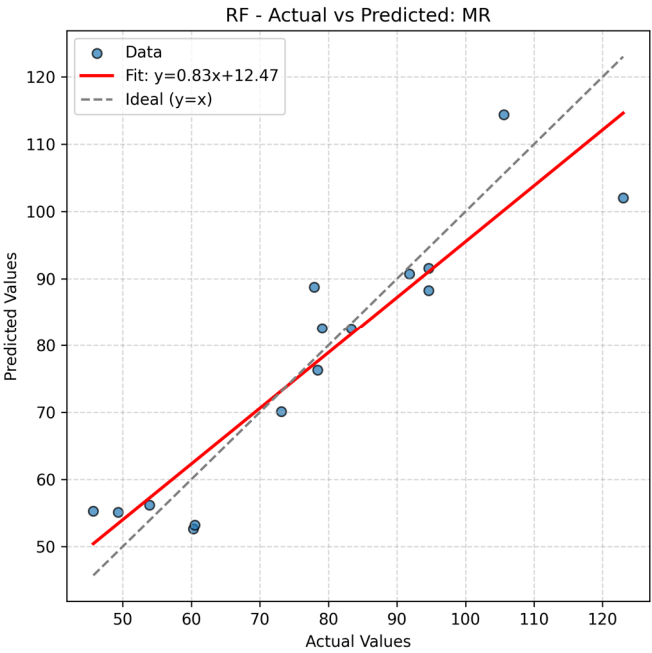
The comparison between ANN and RF models shows that both algorithms reflect the major structural-property relationships well, with ANN being slightly better at the majority of endpoints than RF. Both the models (LOOCV validation) can provide meaningful predictive insight into the physicochemical characteristics of Parkinson’s disease drugs and avoid overfitting despite the small size of the data set. Overall, ANN outperformed RF in five of the seven physicochemical properties indicating the better generalization and consistency. Particularly for MR, P, MW and nHA, ANN gave higher R² and lower error measurements which again demonstrate the capacity to represent complex nonlinear relationships in small data sets. RF had similar performance for MV and C but was less reliable for other properties. Both models had low accuracy for nRotB, indicating that this property may need more descriptors or larger data sets for reliable prediction. These outcomes suggest that ANN is the model of choice in the prediction of the physicochemical properties useful in the design of drugs for the treatment of Parkinson’s disease, whereas RF will be used as a complementary method in the structure-dependent endpoints. The predictive performance is graphically validated in Figure 17 (a-g) which shows the Actual vs. Predicted plots for the RFF model.

Table 6. Predicted properties of Molecular Structure Using Random Forest (RF).

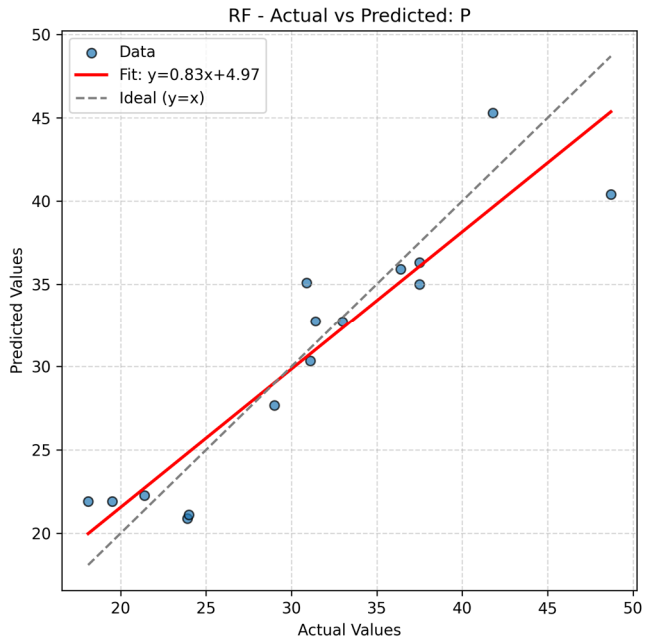
Drugs	MR	P	MV	MW	nHA	nRotB	C
Levodopa	55.122	21.91	168.58	175.9011	13.12	2.26	197.15
Apomorphine	88.787	35.082	259.412	326.7082	23.32	5.37	382.44
Pramipexole	52.628	20.892	168.338	174.7597	12.99	2.87	190.37
Ropinirole	76.284	30.342	228.67	269.2426	19.86	4.015	361.39
Rotigotine	88.258	34.991	271.731	302.3791	22.02	5.173333	357.46
Entacapone	82.593	32.739	250.144	288.4176	21.3	5.58	339.44
Selegiline	53.21	21.119	157.63	184.6312	13.06	2.14	200.15
Rasagiline	56.168	22.261	175.08	190.2648	13.8	3.24	195.92
Amantadine	55.279	21.912	161.086	196.9744	13.95	2.7	211.11
Safinamide	82.541	32.731	240.932	302.7113	21.73	4.385	416.94
Benzotropine	91.55	36.305	274.271	308.0587	21.68	5.18	348.42
Istradefylline	114.37	45.296	342.773	411.2973	29.53	6.43	486.43
Pimavanserin	102.016	40.397	298.327	378.7655	27.32	5.01	540.37
Rivastigmine	70.105	27.669	220.965	236.8066	16.69	4.965	247.25
Trihexyphenidyl	90.735	35.901	260.818	312.5013	22.01	5.84	367.73

Table 7. Random Forest (RF) Error metrics.

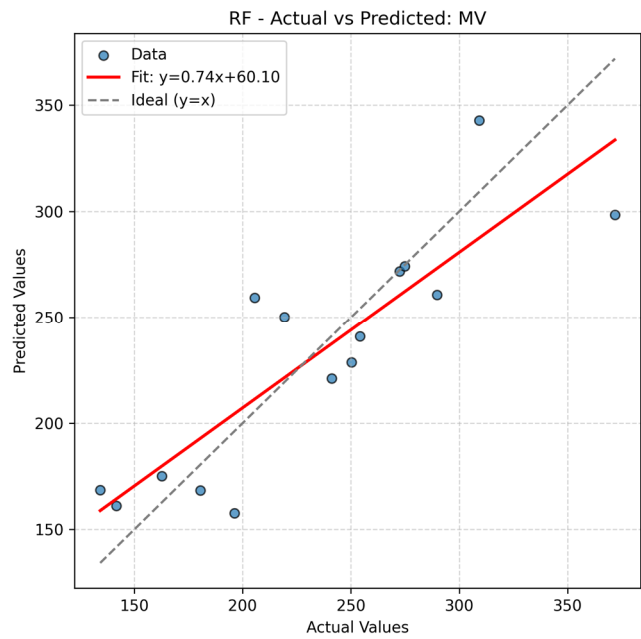
Properties	MSE	MAE	RMSE	R2
MR	63.69671	6.206133	7.981022	0.856303
P	9.950897	2.456867	3.154504	0.856747
MV	1039.901	26.23953	32.2475	0.741946
MW	1388.962	27.97269	37.26879	0.792755
nHA	5.459187	1.64	2.33649	0.838719
nRotB	4.789404	1.701444	2.188471	0.112343
C	4583.721	49.49	67.70318	0.742051



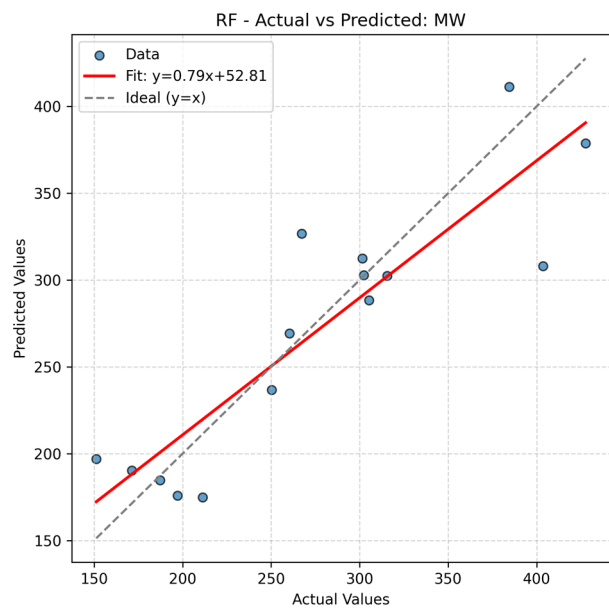
(a)



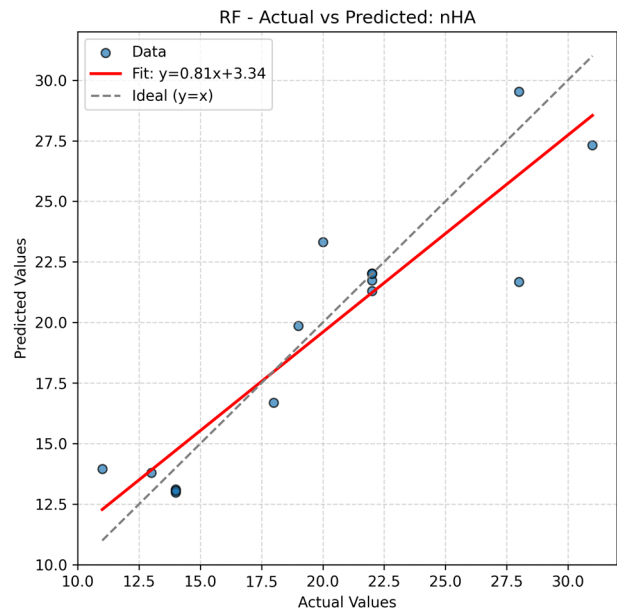
(b)



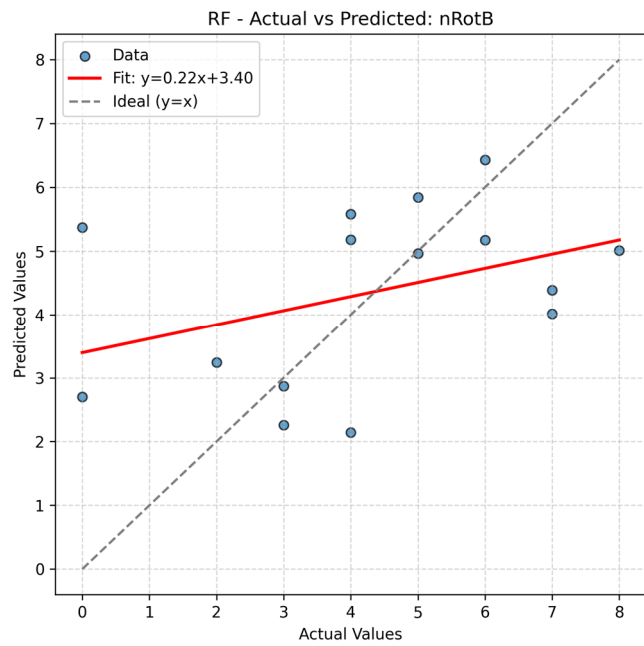
(c)



(d)



(e)



(f)

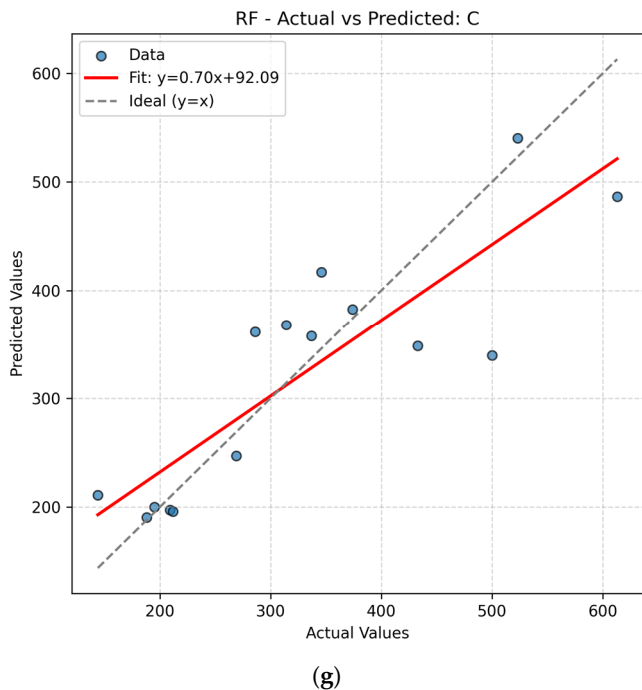


Figure 17. a–g illustrating Actual Vs Predicted (RF).

3.3. Model Robustness: Y-Randomization Test

The statistical reliability and the generalization ability of the QSPR models were verified by conducting a Y-randomization test (response permutation). This stringent method of internal validation overcomes the risk of chance correlation in small and high-dimensional data.

The average coefficient of determination (R^2_{rand}) was computed using 50 permutation of the target property values for ANN and RF models which was summarized in Table 8.

Table 8. Average R^2_{rand} values from the Y-Randomization Test.

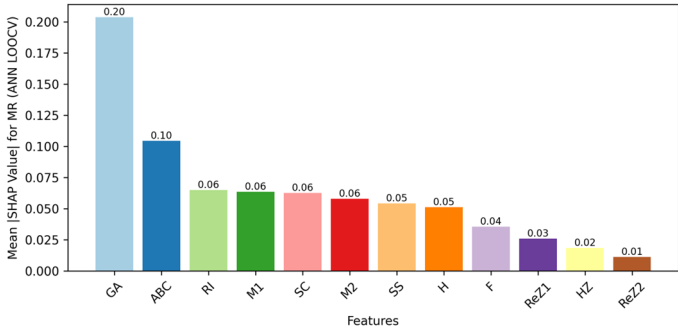
Property	Average R^2_{rand} (ANN Random)	Average R^2_{rand} (RF Random)
C	-1.70758	-0.490221
MR	-1.3982	-0.533738
MV	-1.5136	-0.486532
MW	-1.66355	-0.476982
P	-1.34117	-0.569124
nHA	-1.62529	-0.457509
nRotB	-1.12054	-0.467537

As shown in Table 8, the average R^2_{rand} values obtained as a result of the models trained from the randomized data were always and strongly negative, varying from -1.70758 to -0.457509. This outcome is crucial to the validation as a negative R^2_{rand} means that the predictive models based on scrambled meaningless data fit much worse than a simple prediction of the average value of the property.

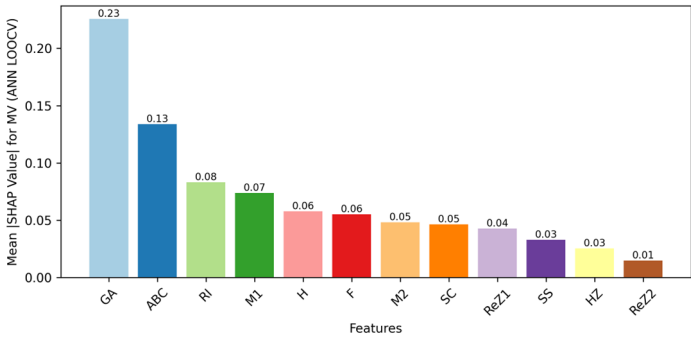
The enormous difference between the original cross-validated R^2 values (Tables 5 and 7) and the negative R^2_{rand} confirms that the high predictive power of both ANN and RF models is based on true non-chance correlations between the topological indices and the physicochemical properties. This establishes the statistical validity and robustness of developed QSPR models.

3.4. Feature Importance and SHAP Analysis

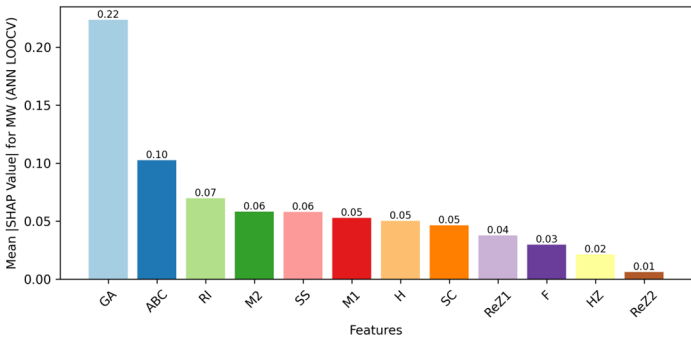
In order to understand the contribution of each topological descriptor on the predicted physicochemical properties, SHapley Additive exPlanations (SHAP) values were calculated for ANN using LOOCV. The analysis provided relative importance of the descriptors for all the targets, generating mechanistic insight into the role of molecular topology in driving physicochemical behavior. SHAP bar plots (Figures 18a to 18g) depicts the average absolute contribution of each descriptor for the 15 drugs.



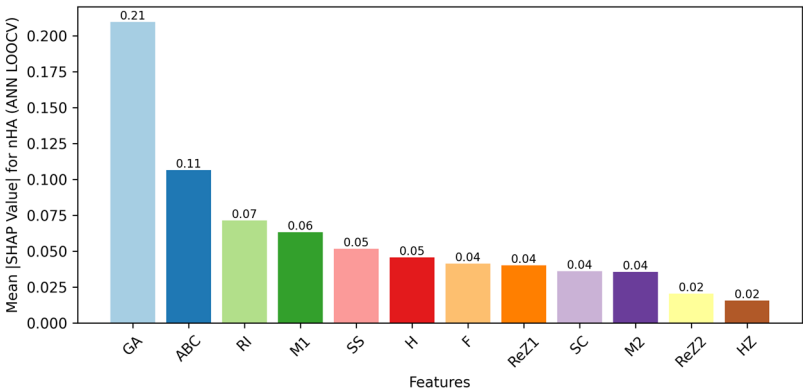
(a)



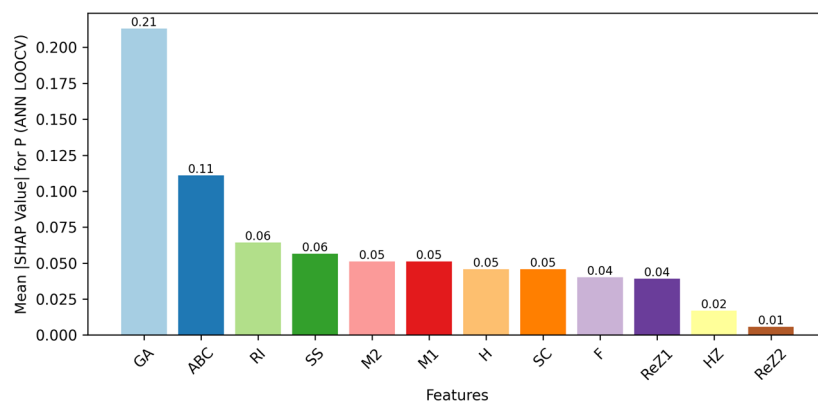
(b)



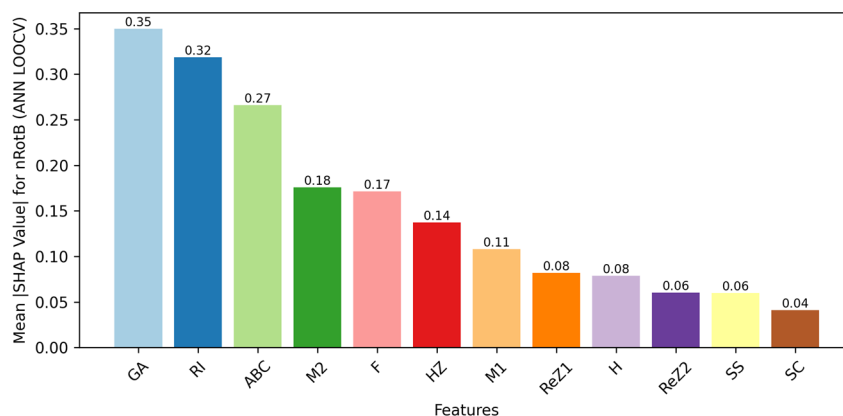
(c)



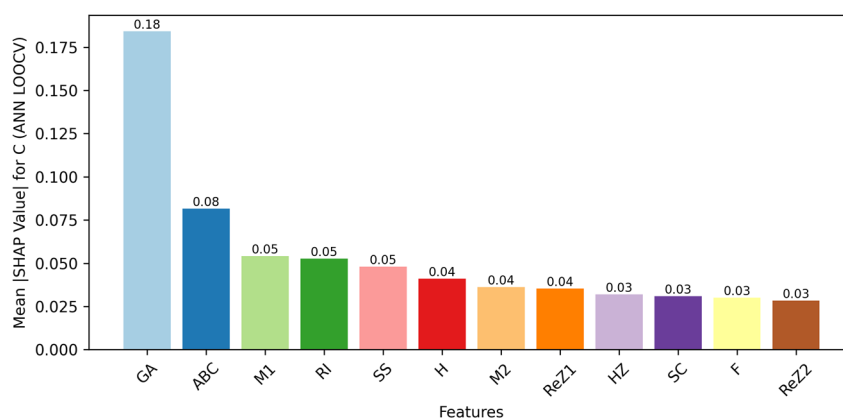
(d)



(e)



(f)



(g)

Figure 18. a–g illustrating SHAP Analysis—ANN.

4. Discussion

The present work had two linked objectives to assess the supervised machine learning models by predictive accuracy, computational efficiency, and interpretability and show the integration between QSPR descriptors and supervised learning can enhance the drug discovery for Parkinson's disease (PD) by decreasing the experimental screening. Based on the LOOCV results, ANN and Random Forest (RF) model are able to predict several important properties of PD drugs successfully. MR, Polarizability (P), Molecular Weight (MW) and Heavy Atom Count (nHA) were well predicted with high accuracy, with R^2 values always higher than 0.84 for both models, while discrete or structurally complex features [i.e., Rotatable Bond Count (nRotB) and Molecular Complexity (C)] were less well predicted, especially in ANN models (R^2 around 0.31 for nRotB). These results represent different abilities of models to represent continuous and discrete or highly variable

molecular features. The reported near-perfect R^2 values for a large number of endpoints by the ANN reflect outstanding predictive accuracy on the available data set. Both models had a difficult time predicting the correct amount of rotatable bonds. This would imply that the chosen descriptor set does not strongly correlate with the nRotB property and further descriptor engineering or different model architectures might be required for this particular target.

Due to the small dataset size, model performance was assessed using Leave-One-Out Cross-Validation (LOOCV) so as to make a realistic prediction of model performance on unseen compounds. LOOCV enables individual testing of each compound while using the remaining data for training and is ideal for testing small datasets. This strategy guarantees that the reported impact of performance measures such as MSE, RMSE, MAE, and R^2 , are reported on unseen data, even though the number of data points available constrains the use of more powerful methods such as nested cross-validation, or even external validation.

However, such results need to be put into context of dataset size, dimensionality of the descriptor and the validation strategy. High R^2 plus small sample of compounds plus high dimensional topological inputs creates a realistic risk of overfitting, for robust model selection, nested CV plus external holdout sets, y-randomization tests, and reporting out of sample RMSE/MAE rather than in-sample fit alone.^{8,10} To address this critical concern, Y-randomization test was undertaken with 50 permutations for all models and properties. The resulting average R^2_{rand} values for all the randomized models were consistently and strongly negative, representing a value range of -1.708-0.457. This is conclusive evidence that the predictive relationships that were created by the original models are statistically sound and not the result of chance correlation. The test thus establishes the good stability of both these models (ANN and RF) against spurious fitting, which statistically validates the utility of the models for physicochemical properties for which $R^2 > 0.8$.

Comparative evaluation between classes of algorithms reveals some obvious trade-offs. Deep ANNs usually perform well on modeling complex, nonlinear structure-property relationships and, thus, for many tasks in cheminformatics, are often better than linear regressions and simpler tree models, provided the richness of the data.^[11,14] In this study, ANN slightly outperformed RF for MR, P, and nHA which reflects the ability of ANN to capture nonlinear relationship and RF for MW, MV, and C provided more stable predictions with less extreme errors. Ensemble tree approaches are often on par or exceed ANNs on tabular QSPR problems with significantly reduced tuning cost and improved out-of-the-box stability.^[27,28] The results of the LOOCV analysis showed that there were compounds with larger deviations, such as Pimavanserin and Istradefylline, which could indicate that when small datasets containing compounds with structural diversity are used, the generalization of models can be challenged.

The other decisive axis is interpretability. Interpretability was addressed using Shapley Additive Explanations (SHAP), which identified the most influential topological descriptors for each property, which provides mechanistic insight into model predictions.^[29] This way, it was possible to assess which molecular features were mostly responsible for predicted MR, P, MW, and nHA, increasing transparency even when the ANN is a nonlinear “black box” model. RF models naturally support feature importance measures, which were complementary in a way that gave access to interpretability. The ability to associate certain descriptors with predicted properties is of special interest in medicinal chemistry decision-making and in the early-stage PD drug prioritization.

Integrating QSPR descriptors with supervised learning gives three practical speed advantages to discovery. First, it allows high-throughput in silico prioritization of compounds to balance the BBB permeation, lipophilicity, and pharmacokinetic constraints, thus reducing the pool of candidates subject to experimental assays.^[30,31] After which, iterative cycles of model-guided selection and targeted synthesis which are augmented by active learning can maximize information per experiment and reduce total wet-lab burden.^[1,2] This is useful for the combination of QSPR outputs with target-specific predictions (e.g., docking scores or binding affinities for dopamine receptors or alpha-synuclein aggregations) supports multi- optimization and prioritization for PD-relevant mechanisms.^[13] While iterative model-guided selection or active learning was not applied in this

study, the current LOOCV results provide a framework to use computational prioritization in PD drug discovery. The findings highlight that descriptor-based supervised learning can be used to speed up the early-stage evaluation, however, caution needs to be taken as a result of small datasets size, high-dimensionality inputs, and endpoints with low-predictability, which might limit the generalization of the model.[8,10]

Operational recommendations follow directly from these observations where benchmark models with nested cross-validation, quantification of uncertainty (prediction intervals, ensembles), and definition of an applicability domain are useful before deploying predictions for experimental selection.[10] To enlarge descriptor space to 3D and quantum chemical features; consider graph neural networks or message passing architectures in case of large datasets, as they learn representations directly from molecular graphs.[13] This will help in explicitly including PD-specific endpoints and multiobjective criteria so that computational prioritization shares the same therapeutic constraints as central nervous system drug development.

4. Conclusions

This study shows that supervised machine learning models based on topological and physicochemical descriptors can be used to precisely predict the important properties of Parkinson's disease drugs. ANN models performed better in terms of predictive power to continuous descriptors such as molar refractivity, polarizability, molecular weight, and heavy atom count, while RF models gave comparable results with an advantage in terms of interpretability and computational efficiency. Both approaches were restricted in the prediction of discrete features such as rotatable bond count, which implies a scope for further model improvement, pointing to a specific limitation of the 2D degree-based topological indices in adequately describing complex molecular flexibility.

The incorporation of ANN- and RF-based QSPR modelling into PD drug discovery pipelines can help lessen the dependence on extensive experimental screening, prioritize compounds with favourable CNS profiles, and rational design strategies. Crucially, the rigorous Y-randomization test confirmed the robustness of all derived models. The strongly negative average R^2 rand values unequivocally show that the high predictive power found in the original models found is statistically significant and that the predictive power is not the result of chance correlation or spurious fitting based on the limited dataset.

These highly predictive and validated QSPR models, coupled with the interpretability provided by SHAP, are a valuable and rapid computational tool for the prioritization of drug candidates in the early stages of development, minimizing physicochemical constraints that are essential to succeed in the drug discovery process. However, the use of these models for experimental selection should continue with caution because of the small data set size ($N=15$), which does not allow complete validation by others. Future studies should aim at increasing the number of compounds and including 3D and quantum chemical descriptors to improve the prediction of endpoints such as molecular flexibility which are difficult to predict. These steps will facilitate mechanistic understanding, and facilitate accelerated therapeutic development for neurodegenerative diseases.

Funding: No funding was received for conducting this study.

Data: Availability Statement All data generated or analysed during this study are included in this published article.

Disclosure: of Interest The authors report there are no competing interests to declare.

References

1. Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al. Applications of machine learning in drug discovery and development. *Nat Rev Drug Discov.* 2019 Jun 11;18(6):463–77.

2. Zhavoronkov A, Ivanenkov YA, Aliper A, Veselov MS, Aladinskiy VA, Aladinskaya A V., et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat Biotechnol.* 2019 Sep 2;37(9):1038–40.
3. Walters WP, Barzilay R. Applications of Deep Learning in Molecule Generation and Molecular Property Prediction. *Acc Chem Res.* 2021 Jan 19;54(2):263–70.
4. Bloem BR, Okun MS, Klein C. Parkinson's disease. *The Lancet.* 2021 Jun;397(10291):2284–303.
5. Kalia L V, Lang AE. Parkinson's disease. *The Lancet.* 2015 Aug;386(9996):896–912.
6. Kouli A, Torsney KM, Kuan WL. Parkinson's Disease: Etiology, Neuropathology, and Pathogenesis. In: *Parkinson's Disease: Pathogenesis and Clinical Aspects.* Codon Publications; 2018. p. 3–26.
7. Dorsey ER, Bloem BR. The Parkinson Pandemic—A Call to Action. *JAMA Neurol.* 2018 Jan 1;75(1):9.
8. Sliwoski G, Kothiwale S, Meiler J, Lowe EW. Computational Methods in Drug Discovery. *Pharmacol Rev.* 2014 Jan;66(1):334–95.
9. Todeschini Roberto, Consonni Viviana. Molecular descriptors for chemoinformatics. Wiley-VCH ; John Wiley [distributor]; 2009.
10. Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M, et al. QSAR Modeling: Where Have You Been? Where Are You Going To? *J Med Chem.* 2014 Jun 26;57(12):4977–5010.
11. Goh GB, Hodas NO, Vishnu A. Deep learning for computational chemistry. *J Comput Chem.* 2017 Jun 15;38(16):1291–307.
12. Segler MHS, Preuss M, Waller MP. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature.* 2018 Mar 29;555(7698):604–10.
13. Yang K, Swanson K, Jin W, Coley C, Eiden P, Gao H, et al. Analyzing Learned Molecular Representations for Property Prediction. *J Chem Inf Model.* 2019 Aug 26;59(8):3370–88.
14. Gawehn E, Hiss JA, Schneider G. Deep Learning in Drug Discovery. *Mol Inform.* 2016 Jan 30;35(1):3–14.
15. Mamoshina P, Vieira A, Putin E, Zhavoronkov A. Applications of Deep Learning in Biomedicine. *Mol Pharm.* 2016 May 2;13(5):1445–54.
16. Ekins S, Puhl AC, Zorn KM, Lane TR, Russo DP, Klein JJ, et al. Exploiting machine learning for end-to-end drug discovery and development. *Nat Mater.* 2019 May 18;18(5):435–41.
17. Gutman Iva, Polansky OE. *Mathematical Concepts in Organic Chemistry.* 2012.
18. Fajtlowicz S. On Conjectures of Graffiti. In 1988. p. 113–8.
19. Furtula B, Gutman I. A forgotten topological index. *J Math Chem.* 2015 Apr 6;53(4):1184–90.
20. Zhao W, Shanmukha MC, Usha A, Farahani MR, Shilpa KC. Computing SS Index of Certain Dendrimers. *Journal of Mathematics.* 2021 Sep 24;2021:1–14.
21. Estrada E, Torres L, Rodriguez L, Gutman I. An Atom-Bond Connectivity Index: Modelling the Enthalpy of Formation of Alkanes. *Indian J Chem.* 1998;37(A):849–55.
22. Randic M. Characterization of molecular branching. *J Am Chem Soc.* 1975 Nov 1;97(23):6609–15.
23. Farahani MR. On the Randic and Sum-Connectivity Index of Nanotubes. *Annals of West University of Timisoara—Mathematics.* 2013 Jan 1;51(2).
24. Vujošević S, Popivoda G, Kovijanić Vukićević Ž, Furtula B, Škrekovski R. Arithmetic–geometric index and its relations with geometric–arithmetic index. *Appl Math Comput.* 2021 Feb;391:125706.
25. RAJASEKHARAIAH G V., MURTHY UP. Hyper-Zagreb indices of graphs and its applications. *Journal of Algebra Combinatorics Discrete Structures and Applications.* 2021 Jan 15;8(1):9–22.
26. P.S. Ranjini, V. Loksha, A. Usha. Relation between phenylene and hexagonal squeeze using harmonic index. *International Journal of Graph Theory.* 2013;1(4):116–21.
27. Chen CH, Tanaka K, Kotera M, Funatsu K. Comparison and improvement of the predictability and interpretability with ensemble learning models in QSPR applications. *J Cheminform.* 2020 Dec 30;12(1):19.
28. Sarkar C, Das B, Rawat VS, Wahlang JB, Nongpiur A, Tiewsoh I, et al. Artificial Intelligence and Machine Learning Technology Driven Modern Drug Discovery and Development. *Int J Mol Sci.* 2023 Jan 19;24(3):2026.
29. Scott M. Lundberg, Su-In Lee. Consistent feature attribution for tree ensembles. In: *ICML Workshop on Human Interpretability in Machine Learning.* Sydney, NSW, Australia; 2017.

30. Bagchi S, Chhibber T, Lahooti B, Verma A, Borse V, Jayant RD. In-vitro blood-brain barrier models for drug screening and permeation studies: an overview. *Drug Des Devel Ther.* 2019 Oct;Volume 13:3591–605.
31. Dichiara M, Cosentino G, Giordano G, Pasquinucci L, Marrazzo A, Costanzo G, et al. Designing drugs optimized for both blood–brain barrier permeation and intra-cerebral partition. *Expert Opin Drug Discov.* 2024 Mar 3;19(3):317–29.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.