Research Article BEU Fen Bilimleri Dergisi 15(3): 1336-1351 (2026) DOI: 10.17798/bitlisfen.1897533

BİTLİS EREN UNIVERSITY
FEN BİLİMLERİ DERGİSİ
ISSN: 2147-3129 / e-ISSN: 2147-3188

A COMPARATIVE STUDY OF ATOM-CENTERED AND EDGE-NEIGHBORHOOD GRAPH REPRESENTATIONS IN QSPR MODELING OF GREEN PESTICIDES

Fatma ÖZEN ERDOĞAN [1*], Betül ÖZTÜRK[1]

[1] *Bursa Uludag University, Mathematics Department, Bursa, Türkiye*

**** Corresponding Author:*** *fatmaozen@uludag.edu.tr*

Article Info

Received: *February 26, 2026*
Revised: *June 4, 2026*
Accepted: *June 16, 2026*

Keywords
QSPR,
Edge-Neighborhood Graph,
Topological Indices,
Pesticides,
Ecotoxicity.

ABSTRACT

Pesticides are indispensable for agricultural productivity, yet their widespread use creates environmental risks such as persistence, bioaccumulation, and ecotoxicity. Early prediction of these risks is therefore important, especially because experimental evaluation of new chemical compounds is costly, time-consuming, and ethically constrained. In this context, Quantitative Structure-Property Relationship (QSPR) modeling provides a useful computational tool for the rational assessment of environmentally friendly pesticides.

This study compares two hydrogen-inclusive molecular graph representations for a structurally diverse set of sixteen pesticide molecules: the atom-centered G representation and the bond-centered edge-neighborhood $N_e(G)$ representation. Simple Linear Regression-based QSPR models were developed for six physicochemical and environmental fate properties, and the predictive behavior of topological indices derived from both representations was comparatively evaluated.

The results indicate a complementary, property-dependent pattern. The G-based models were slightly more favorable for extensive bulk-related properties such as molar refractivity and polarizability, whereas $N_e(G)$-based models yielded modestly higher coefficient of determination (R^2) values for soil organic carbon-water partition coefficient ($logK_{oc}$), bioconcentration factor ($logBCF$), and octanol-water partition coefficient ($logP$) within the descriptor families examined. These differences are interpreted as controlled, dataset-specific trends rather than evidence of universal representational superiority. Overall, the findings suggest that hydrogen-inclusive edge-neighborhood graph representations may provide an informative complementary framework for QSPR modeling and early environmental risk assessment of pesticides.

1. INTRODUCTION

1.1. The Necessity and Environmental Cost of Modern Agrochemicals

Agricultural production increasingly relies on the consistent application of pesticides to meet global food security demands and maintain high crop yields [1]. While chemically diverse and highly effective, the widespread and often uncontrolled release of these chemicals imposes severe and long-lasting stress on environmental systems. This stress manifests primarily through three critical environmental endpoints: ecotoxicity (harm to non-target organisms), persistence (long-term resistance to degradation in soil and water systems), and bioaccumulation (concentration in living tissues after entering the food chain) [2].

These persistent threats have necessitated a paradigm shift within the chemical industry, directing efforts toward the principles of "Green Chemistry." The goal is the rational design of novel, environmentally friendly pesticides, often termed "green pesticides," that exhibit a minimized environmental footprint [1].

 License: CC BY-NC 4.0

1.2. The Shift to Rational Design and Computational Toxicology

Traditional experimental toxicological and physicochemical evaluation of new chemical entities presents numerous challenges. It is inherently costly, time-intensive, and, crucially, ethically complicated due to regulatory frameworks, such as REACH, and ethical guidelines, like the 3R Rule, which mandate the reduction of animal testing [3, 4].

In response to these constraints, computational methods have emerged as powerful, ethical alternatives. Quantitative Structure-Property Relationship (QSPR) modeling is chief among these, offering the capability to predict a chemical's physicochemical and ecotoxicological properties based solely on its molecular structure, often before synthesis occurs [5].

1.3. Molecular Topology and the Standard Limitations

The predictive power of QSPR models fundamentally rests on topological indices (TIs) [6], which are numerical invariants that quantify the topological characteristics of a molecule, such as atom connectivity, branching, and molecular shape. Recent advancements in the descriptor literature, including the development of novel Zagreb-type indices (such as Zagreb rho, upsilon, eta and omicron indices), have further demonstrated the evolving capacity of TIs to model complex physicochemical characteristics [7-10]. Historically, molecular structures in computational chemistry have commonly been represented by an atom-centered G representation [11].

A common practice in traditional QSPR studies, aimed at reducing computational complexity, is the use of hydrogen-suppressed methodology, where hydrogen atoms are explicitly excluded from the graph representation [12]. This simplification is useful in many settings; however, it may remove structural details associated with local molecular interactions. Hydrogen bonding capacity, local steric effects, and subtle electronic environments can be influenced by the presence and arrangement of hydrogen atoms. These local features are relevant for phase partition coefficients, such as water-lipid and soil organic carbon-water partitioning, and for subsequent biological interactions [13].

1.4. Introducing the Methodological Innovations of This Study

This investigation introduces two key methodological advancements designed to overcome the limitations of the standard H-suppressed methodology.

The first innovation involved constructing the atom-centered G representation to represent the complete molecular structure, explicitly including all hydrogen atoms. This hydrogen-inclusive approach increases structural resolution and allows surface-related features, such as steric interactions and potential hydrogen bonding sites, to be reflected in the graph topology.

The second, and central, innovation utilized the bond-centered edge-neighborhood $N_e(G)$ representation [14]. This representation, which is derived from the complete hydrogen-inclusive G representation, adopts a bond-centered perspective by shifting the descriptor focus from individual atoms to relationships among adjacent chemical bonds and their local environments [15, 16].

1.5. Hypothesis and Objective of the Study

While the inclusion of hydrogen atoms in atom-centered representations enriches structural detail, the same increase in resolution may also introduce descriptor-level complexity. In contrast, the bond-centered $N_e(G)$ representation can be interpreted as a structural abstraction that emphasizes local bonding environments rather than individual atomic contributions. Accordingly, this study examines the working hypothesis that hydrogen-inclusive G and $N_e(G)$ may encode different aspects of molecular structure, with possible property-dependent differences in QSPR performance.

The objective of this study is to systematically compare the predictive behavior of QSPR models developed using indices derived from the hydrogen-inclusive atom-centered G representation and the bond-centered edge-neighborhood $N_e(G)$ representation for 16 environmentally significant pesticide molecules across six critical physicochemical properties. Rather than aiming to construct a universally generalizable predictive model, the study comparatively evaluates representation-dependent trends under controlled conditions involving dataset selection, descriptor family, and modeling protocol.

2. MATERIALS AND METHODS

The subsequent sections provide the necessary details concerning the data handling, molecular representation, topological index computation, and statistical methods employed, ensuring the reproducibility of the study.

2.1. Data Set and Physicochemical Properties

The analysis was conducted using a dataset of 16 structurally diverse, commercially relevant, and environmentally significant pesticide molecules: Chlorpyrifos, Malathion, Diazinon, Cypermethrin, Deltamethrin, Permethrin, Imidacloprid, Clothianidin, Thiamethoxam, Glyphosate, Atrazine, 2,4-D, Tebuconazole, Fipronil, Fenvalerate, and Amitraz. The two-dimensional (2D) molecular structures of the selected pesticide molecules were obtained from the ChemSpider database and subsequently standardized using the RDKit cheminformatics toolkit [17].

Table 1. Investigated Pesticide Molecules and Their Physicochemical Property Values

Pesticide	MV	MR	P	$logK_{oc}$	$logBCF$	$logP$
Chlorpyrifos	236.8	77.2	30.6	3.9794	3.4057	4.77
Malathion	259.7	77.5	30.7	2.8783	1.8674	2.92
Diazinon	260.5	79.6	31.6	3.4433	2.6568	3.81
Cypermethrin	313.1	110.2	43.7	4.6702	4.3707	6.27
Deltamethrin	316.7	116.0	46.0	4.7047	4.4189	6.20
Permethrin	302.5	105.7	41.9	4.8965	4.6869	7.15
Imidacloprid	160.1	62.3	24.7	9.956	0	-0.43
Clothianidin	147.9	57.7	22.9	1.2613	0	0.40
Thiamethoxam	170.3	67.6	26.8	1.1992	0	-1.16
Glyphosate	100.7	31.1	12.3	0	0	-2.36
Atrazine	169.9	58.5	23.2	2.8264	1.7949	2.63
2,4-D	148.5	48.9	19.4	0	0	2.58
Tebuconazole	268.1	87.2	34.6	3.4115	2.6123	3.58
Fipronil	233.7	81.9	32.5	3.3939	2.5878	4.75
Fenvalerate	346.8	116.4	46.2	4.7914	4.5401	6.67
Amitraz	297.3	94.2	37.3	4.2346	3.7628	5.64

Six fundamental physicochemical and ecotoxicological properties were selected for modeling (Table 1), as they are critical indicators of a chemical's environmental fate and biological interaction potential: Molar Volume (MV), Molar Refractivity (MR), Polarizability (P), Octanol-Water Partition Coefficient ($logP$), Bioconcentration Factor ($logBCF$), and Soil Organic Carbon-Water Partition Coefficient ($logK_{oc}$). MV, MR, and P provide information about internal characteristics such as molecular volume and electron density, while $logP$, $logBCF$, and $logK_{oc}$ reflect the interactions the chemical enters into with different environmental phases (lipid, water, soil) and its resulting bioaccumulation and persistence potentials [13]. The 2D structures of the investigated pesticide molecules are shown in Figure 1.

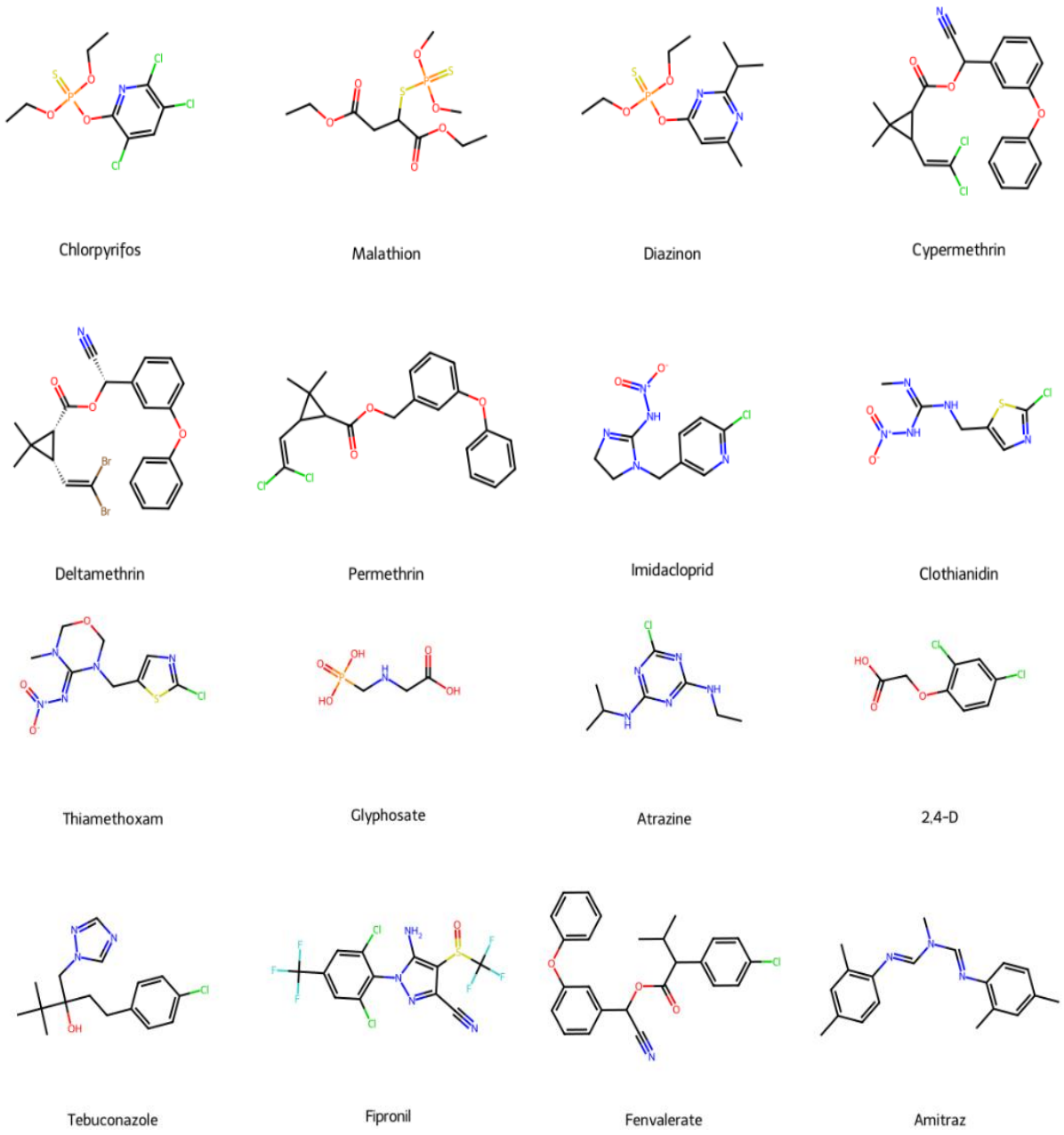

Figure 1. 2D Structures of the Investigated Pesticide Molecules.

2.2. Molecular Graph Representations

Two distinct molecular representations were generated for each pesticide molecule to enable the comparative calculation of topological indices. Both representations were explicitly constructed to include hydrogen atoms, based on preliminary evidence reported in related studies [14, 15] (Figure 2).

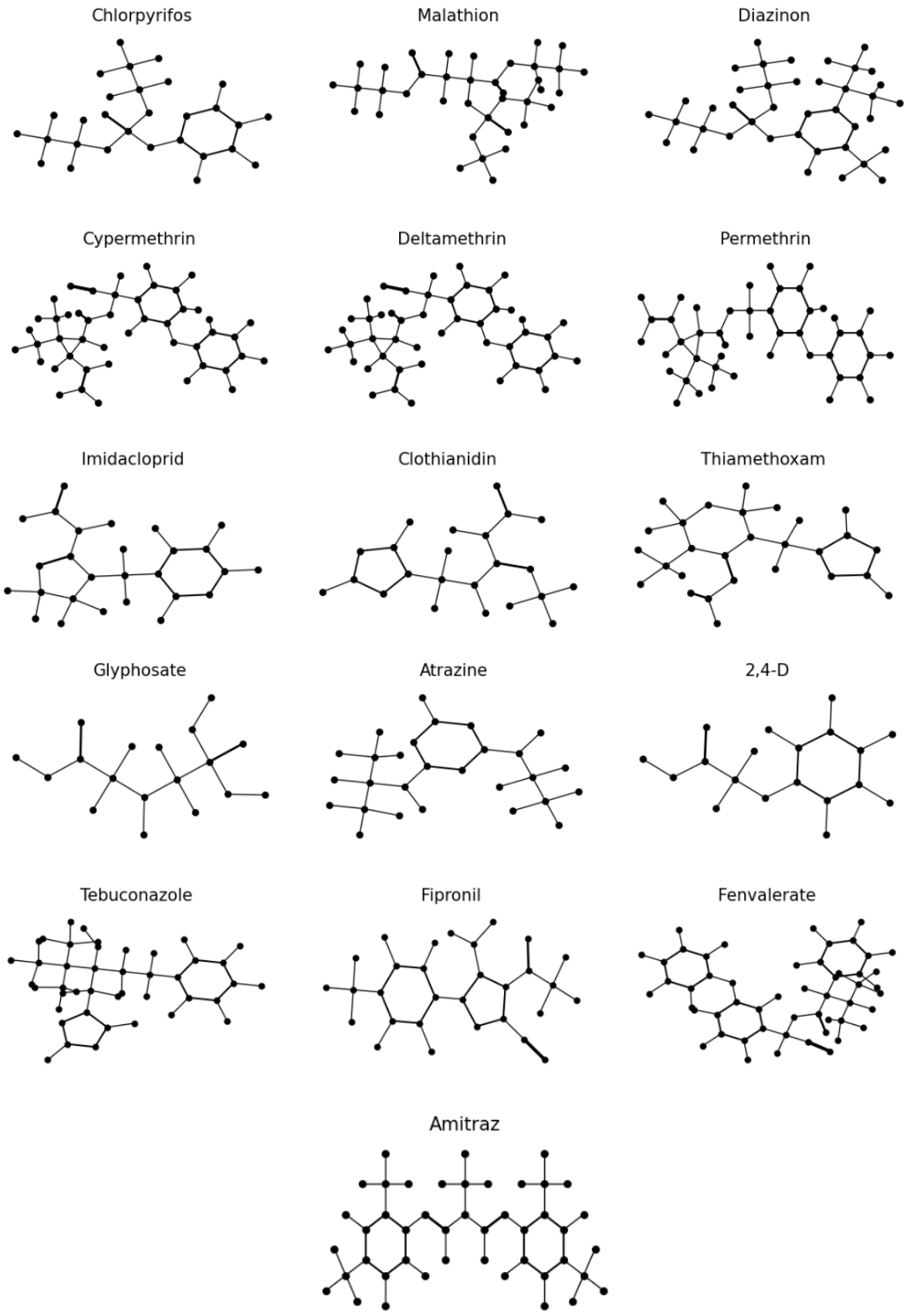

Figure 2. Hydrogen-Inclusive Atom-Centered G Representations of the 16 Investigated Pesticide Molecules.

2.2.1. Atom-Centered G Representation (Hydrogen-Inclusive Approach)

The atom-centered G representation is defined as $G = (V, E)$. In this representation, the vertex set $V(G)$ represents all atoms in the molecule, including hydrogens, while the edge set $E(G)$ represents all chemical bonds. The deliberate inclusion of hydrogen atoms substantially increases the dimension and complexity of the molecular representation, aiming to incorporate surface properties such as steric interactions and potential hydrogen bonding capacity into the graph topology.

2.2.2. Bond-Centered Edge-Neighborhood $N_e(G)$ Representation

The $N_e(G)$ representation provides an alternative, bond-centered perspective of the molecular structure and is derived directly from the atom-centered G representation. Formally, if the atom-centered representation is defined as $G = (V, E)$, the vertex set of $N_e(G)$, denoted by $V(N_e(G))$, is equivalent to the edge set of G, that is, $V(N_e(G)) = E(G)$.

Adjacency in $N_e(G)$ is defined by shared atoms: two vertices in $N_e(G)$ (which correspond to edges e_i and e_j in G) are connected by an edge if and only if the original edges e_i and e_j share a common vertex (atom) in G. For example, if $e_i = (u, v)$ and $e_j = (v, w)$, the corresponding vertices in $N_e(G)$ are adjacent at atom v. This bond-centered representation aims to provide higher resolution structural information by focusing on the local chemical environment and the bonding surroundings, rather than individual atoms. The representational difference between G and $N_e(G)$ representations for the Chlorpyrifos molecule is conceptually represented in Figure 3 [14].

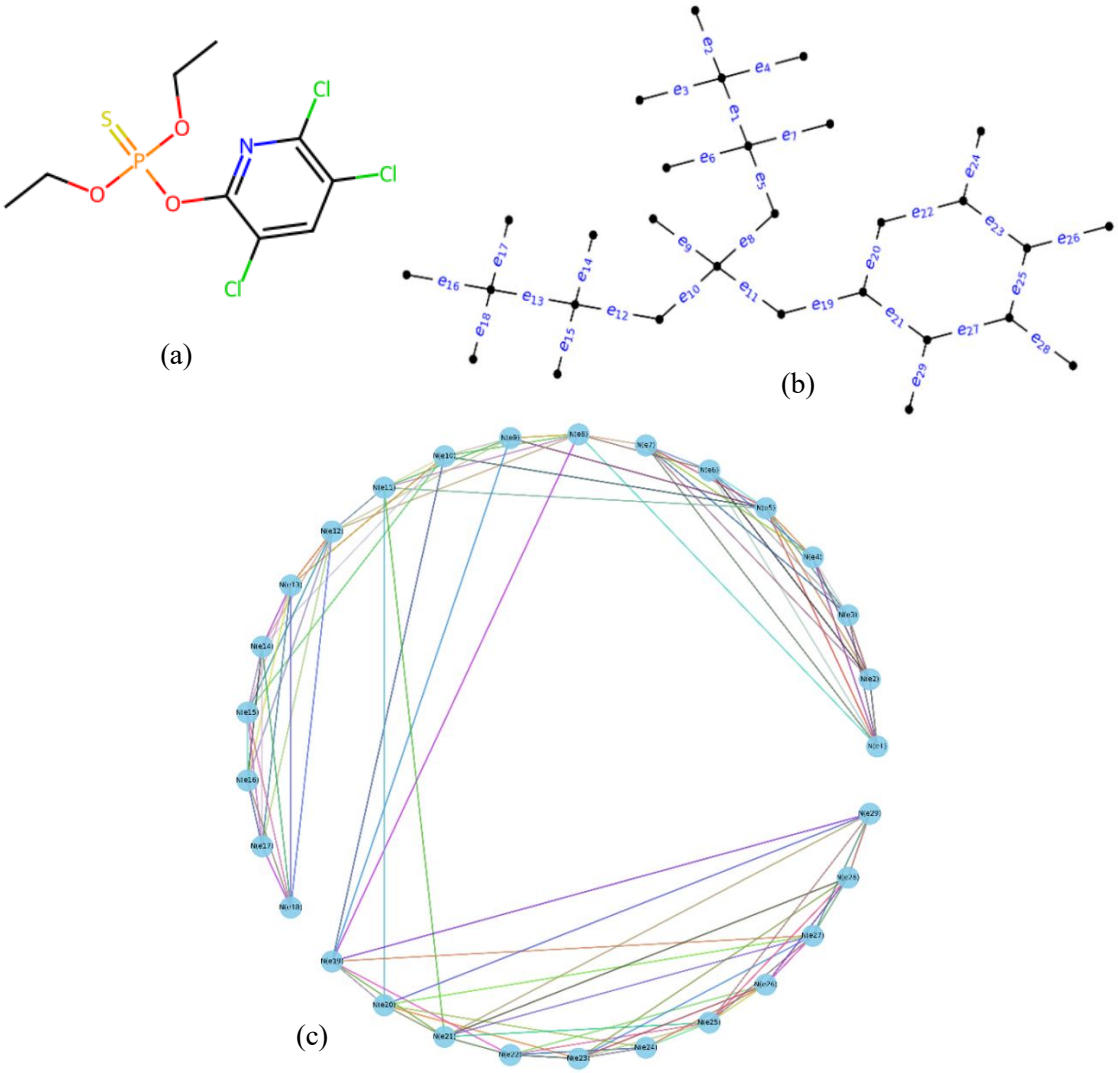

Figure 3. (a) 2D Structure of Chlorpyrifos, (b) Hydrogen-Inclusive G Representation, and (c) $N_e(G)$ Representation.

2.3. Topological Index Calculation

Two major families of topological indices were calculated from both representations, G and $N_e(G)$, for subsequent QSPR modeling: the Classic Degree-Based Indices (Randić, Zagreb M_1, M_2, GA, ABC) and the Neighborhood Degree-Sum Based Indices (NTI).

2.3.1. Neighborhood Degree-Sum Based Indices (NTI)

NTIs rely on a structural measure known as the neighborhood degree sum, which calculates the sum of the degrees of a given vertex's neighbors. For a vertex v in graph G, the neighborhood degree sum $S_G(v)$ is formally defined in Equation (1):

$$S_G(v) = \sum_{u \epsilon N_G(v)} d_G(u) \tag{1}$$

where $d_G(u)$ is the degree of vertex u in G, and $N_G(v)$ is the set of neighbors of v in G.

The mathematical formulations for the Neighborhood-Based Topological Indices (NTI) used in this study are detailed in Table 2.

Table 2. Formulas for Neighborhood-Based Topological Indices (NTI)

Index Name	Abbreviation	Mathematical Formula
Neighborhood First Zagreb Index	$NM_1(G)$	$\sum_{uv \in E(G)} S_G(u) + S_G(v)$
Neighborhood Second Zagreb Index	$NM_2(G)$	$\sum_{uv \in E(G)} S_G(u) S_G(v)$
Neighborhood Geometric-Arithmetic Index	$NGA(G)$	$\sum_{uv \in E(G)} \frac{2\sqrt{S_G(u) S_G(v)}}{S_G(u) + S_G(v)}$
Neighborhood Randic Index	$NR(G)$	$\sum_{uv \in E(G)} \frac{1}{\sqrt{S_G(u) S_G(v)}}$
Neighborhood Reciprocal Randic Index	$NRR(G)$	$\sum_{uv \in E(G)} \sqrt{S_G(u) S_G(v)}$
Neighborhood Atom-Bond Connectivity Index	$NABC(G)$	$\sum_{uv \in E(G)} \sqrt{\frac{S_G(u) + S_G(v) - 2}{S_G(u) S_G(v)}}$
Neighborhood Sum-Degree Division Index	$NSDD(G)$	$\sum_{uv \in E(G)} \frac{S_G(u)}{S_G(v)} + \frac{S_G(v)}{S_G(u)}$
Neighborhood Sum-Connectivity Index	$NSCI(G)$	$\sum_{uv \in E(G)} \sqrt{\frac{1}{S_G(u) + S_G(v)}}$
Neighborhood Forgotten Index	$NF(G)$	$\sum_{uv \in E(G)} (S_G(u))^2 + (S_G(v))^2$

2.4. Statistical Analysis

Simple Linear Regression (SLR) was utilized to develop the QSPR models and systematically compare the structural information capacity of the different graph representations. Each physicochemical property (y) was regressed against a single calculated topological index (TI) using Equation (2):

$$y = a(TI) + b \tag{2}$$

The primary aim of this study is to compare the single-descriptor behavior of G and $N_e(G)$ under an intentionally transparent modeling protocol, without introducing the additional flexibility of complex multivariate models.

The use of Simple Linear Regression was intentionally preferred to ensure that observed differences in predictive performance could be traced to individual descriptor-graph combinations rather than to the increased flexibility of multivariate or machine learning models. Nevertheless, this design also limits the scope of inference. Multiple Linear Regression, partial least squares, or graph-based machine learning approaches could combine descriptors in ways that change the apparent ranking of graph representations. Therefore, the present results should be interpreted as descriptor-by-descriptor comparative trends, not as definitive validation of representation-level superiority.

Model quality and goodness-of-fit were assessed using the coefficient of determination (R^2), the Root Mean Square Error (RMSE), and statistical significance parameters (F-statistic and p-value). A statistical significance level of $p < 0.05$ was adopted for all models. Given the limited dataset size, these statistics are interpreted as in-sample indicators of comparative behavior; broader predictivity would require validation strategies such as cross-validation and external test sets [18].

Because a topological index is not independent of the graph on which it is calculated, the observed performance reflects the combined effect of graph representation, descriptor definition, endpoint type, and model type. The selected classical and NTI descriptors therefore serve as a controlled testing panel; they should not be understood as exhaustive evidence for or against either representation.

This point is also important in light of recent descriptor-design studies. Zagreb-type families have expanded beyond the classical M_1 and M_2 indices to include ev/ve-degree Zagreb variants, Zagreb tau indices, K-Banhatti-related descriptors, and numerous other modern extensions that have demonstrated promising QSPR performance for organic compounds, pharmaceuticals, and benzenoid systems [19-22]. The present study is positioned within this evolving literature by systematically investigating how selected degree- and neighborhood-degree-based descriptors behave when calculated on hydrogen-inclusive G and $N_e(G)$ graph constructions.

3. RESULTS AND DISCUSSION

While hydrogen-inclusive representations increase graph complexity, the central issue is not the presence of additional information alone, but how the chosen graph representation and descriptor family organize that information. The analysis below therefore separates observed statistical outcomes from chemical interpretation. Numerical differences, particularly small R^2 differences, are treated as dataset-specific tendencies that require cautious interpretation.

3.1. Correlation Matrix Comparison

As an initial step, the Pearson correlation coefficients (R values) between all calculated indices and the six target properties were examined (Tables 3, 4, 5, and 6).

Table 3. Correlation Coefficients for Classic Indices on the Atom-Centered G Representation

	$M_1(G)$	$M_2(G)$	$GA(G)$	$R(G)$	$RR(G)$	$ABC(G)$	$SDD(G)$	$F(G)$	$SCI(G)$
MV	0.962	0.947	0.962	0.967	0.960	0.967	0.951	0.949	0.966
MR	0.949	0.951	0.964	0.957	0.957	0.948	0.906	0.921	0.961
P	0.949	0.951	0.964	0.956	0.957	0.948	0.905	0.921	0.961
$logK_{oc}$	0.891	0.879	0.888	0.891	0.889	0.895	0.885	0.883	0.890
$logBCF$	0.879	0.876	0.885	0.887	0.882	0.880	0.856	0.863	0.886
$logP$	0.800	0.795	0.814	0.813	0.805	0.805	0.774	0.774	0.814

Table 4. Correlation Coefficients for Neighborhood-Based Indices on the Atom-Centered G Representation

	$NM_1(G)$	$NM_2(G)$	$NGA(G)$	$NR(G)$	$NRR(G)$	$NABC(G)$	$NSDD(G)$	$NF(G)$	$NSCI(G)$
MV	0.947	0.914	0.967	0.968	0.950	0.968	0.955	0.969	0.900
MR	0.951	0.936	0.959	0.959	0.954	0.958	0.951	0.959	0.921
P	0.951	0.936	0.958	0.959	0.953	0.958	0.951	0.959	0.921
$logK_{oc}$	0.879	0.850	0.894	0.887	0.881	0.892	0.882	0.893	0.837
$logBCF$	0.876	0.854	0.883	0.881	0.877	0.885	0.885	0.883	0.845
$logP$	0.795	0.767	0.811	0.811	0.797	0.813	0.806	0.812	0.755

Table 5. Correlation Coefficients for Classic Indices on the $N_e(G)$ Representation

	M_1 $N_e(G)$	M_2 $N_e(G)$	GA $N_e(G)$	R $N_e(G)$	RR $N_e(G)$	ABC $N_e(G)$	SDD $N_e(G)$	F $N_e(G)$	SCI $N_e(G)$
MV	0.875	0.814	0.940	0.969	0.881	0.951	0.912	0.787	0.957
MR	0.900	0.849	0.947	0.959	0.906	0.952	0.925	0.822	0.955
P	0.900	0.850	0.947	0.958	0.906	0.952	0.925	0.822	0.955
$logK_{oc}$	0.803	0.742	0.869	0.896	0.810	0.880	0.840	0.716	0.886
$logBCF$	0.810	0.752	0.870	0.887	0.817	0.878	0.842	0.724	0.882
$logP$	0.724	0.664	0.790	0.815	0.731	0.799	0.758	0.637	0.806

Table 6. Correlation Coefficients for Neighborhood-Based Indices on the $N_e(G)$ Representation

	NM_1 $N_e(G)$	NM_2 $N_e(G)$	NGA $N_e(G)$	NR $N_e(G)$	NRR $N_e(G)$	$NABC$ $N_e(G)$	$NSDD$ $N_e(G)$	NF $N_e(G)$	$NSCI$ $N_e(G)$
MV	0.814	0.703	0.940	0.962	0.819	0.971	0.917	0.973	0.687
MR	0.849	0.745	0.946	0.919	0.854	0.958	0.934	0.957	0.729
P	0.850	0.745	0.946	0.918	0.854	0.958	0.934	0.957	0.729
$logK_{oc}$	0.742	0.633	0.869	0.889	0.747	0.899	0.843	0.902	0.617
$logBCF$	0.752	0.644	0.871	0.863	0.757	0.891	0.843	0.894	0.628
$logP$	0.664	0.561	0.791	0.814	0.669	0.820	0.758	0.825	0.545

For bulk properties (MV, MR, and P), correlation values were generally high across all index families, reflecting the fundamental dependence of these properties on overall molecular size and complexity. In contrast, for interaction-related and environmental fate properties ($logK_{oc}$, $logBCF$ and $logP$), the correlation coefficients were generally lower, indicating that these endpoints are more difficult to represent using single topological descriptors.

A noteworthy observation emerged from comparing the Classic and NTI descriptors calculated on the G representation. Although NTIs are designed to capture localized structural features, they did not uniformly improve correlations on the hydrogen-inclusive G graph. This result suggests that descriptor choice and graph representation interact with one another; hence, differences cannot be attributed to the graph representation alone without caution.

3.2. Detailed Simple Linear Regression (SLR) Model Performance

The SLR models exhibiting the strongest predictive performance for each property are summarized in Tables 7 through 10. Across these models, a high level of in-sample statistical significance was observed, with p-values consistently indicating significance ($p \leq 0.05$). However, significance should not be interpreted as external predictive validity, especially given the sample size of 16 molecules.

Table 7. Optimal SLR Performance of Classic Indices Calculated on the Atom-Centered G Representation

Target Feature	a	b	R	R^2	$RMSE$	F	p-value
MV	14.7151	5.2967	0.9672	0.9354	19.5925	202.6487	0.0000
MR	2.2157	7.5758	0.9642	0.9297	6.8124	185.2251	0.0000
P	0.8790	2.9928	0.9641	0.9296	2.7058	184.7633	0.0000
$logK_{oc}$	0.1758	-1.8185	0.8954	0.8018	0.7869	56.6202	0.0000
$logBCF$	0.3311	-2.8358	0.8870	0.7868	0.8731	51.6621	0.0000
$logP$	0.4725	-4.0442	0.8142	0.6629	1.7612	27.5337	0.0001

Table 8. Optimal SLR Performance of Neighborhood-Based Indices Calculated on the Atom-Centered G Representation

Target Feature	a	b	R	R^2	$RMSE$	F	p-value
MV	22.9406	2.2135	0.9690	0.9390	19.0419	215.3607	0.0000
MR	13.1579	0.2451	0.9591	0.9198	7.2757	160.6621	0.0000
P	5.2193	0.0872	0.9589	0.9195	2.8920	159.9999	0.0000
$logK_{oc}$	0.1362	-1.7727	0.8943	0.7997	0.7909	55.9021	0.0000
$logBCF$	0.0577	-2.4650	0.8854	0.7840	0.8788	50.8074	0.0000
$logP$	0.3924	-4.1525	0.8130	0.6409	1.7665	27.2847	0.0001

Table 9. Optimal SLR Performance of Classic Indices Calculated on the $N_e(G)$ Representation

Target Feature	a	b	R	R^2	$RMSE$	F	p-value
MV	12.7028	12.1756	0.9383	0.9354	19.1494	212.7918	0.0000
MR	4.1914	6.5420	0.9586	0.9188	7.3211	158.5035	0.0000
P	1.6627	2.5836	0.9584	0.9186	2.9084	158.0371	0.0000
$logK_{oc}$	0.2694	-1.7717	0.8959	0.8027	0.7850	56.9645	0.0000
$logBCF$	0.2853	-2.6714	0.8867	0.7862	0.8744	51.4733	0.0000
$logP$	0.4205	-3.9805	0.8147	0.6637	1.7591	27.6299	0.0001

Table 10. Optimal SLR Performance of Neighborhood-Based Indices Calculated on the $N_e(G)$ Representation

Target Feature	a	b	R	R^2	$RMSE$	F	p-value
MV	18.9249	10.4904	0.9732	0.9472	17.7107	251.1339	0.0000
MR	3.0682	6.4447	0.9584	0.9185	7.3366	157.7724	0.0000
P	1.2171	2.5450	0.9583	0.9183	2.9146	157.3142	0.0000
$logK_{oc}$	0.4024	-1.8194	0.9025	0.8145	0.7612	61.4600	0.0000
$logBCF$	0.4265	-2.7265	0.8939	0.7991	0.8475	55.6922	0.0000
$logP$	0.6317	-4.0986	0.8255	0.6814	1.7123	29.9396	0.0001

3.3. Optimal Regression Equations

The selected simple linear regression equations are presented in Equations (3a)-(8d). Each group corresponds to one physicochemical property and relates the response variable (y) to the corresponding topological index (TI) through the coefficients a (slope) and b (intercept).

Molar Volume (MV)

$$MV = 14.7151(R) + 5.2967 \quad (3a)$$

$$MV = 22.9406(NSCI) + 2.2135 \quad (3b)$$

$$MV = 12.7028(R) + 12.1756 \quad (3c)$$

$$MV = 18.9249(NSCI) + 10.4904 \quad (3d)$$

Molar Refractivity (MR)

$$MR = 2.2157(GA) + 7.5758 \quad (4a)$$

$$MR = 13.1579(NR) + 0.2451 \quad (4b)$$

$$MR = 4.1914(R) + 6.5420 \quad (4c)$$

$$MR = 3.0682(NABC) + 6.4447 \quad (4d)$$

Polarizability (P)

$$P = 0.8790(GA) + 2.9928 \quad (5a)$$

$$P = 5.2193(NR) + 0.0872 \quad (5b)$$

$$P = 1.6627(R) + 2.5836 \quad (5c)$$

$$P = 1.2171(NABC) + 2.5450 \quad (5d)$$

Soil Organic Carbon-Water Partition Coefficient ($logK_{oc}$)

$$logK_{oc} = 0.1758(ABC) - 1.8185 \quad (6a)$$

$$logK_{oc} = 0.1362(NGA) - 1.7727 \quad (6b)$$

$$logK_{oc} = 0.2694(R) - 1.7717 \quad (6c)$$

$$logK_{oc} = 0.4024(NSCI) - 1.81944 \quad (6d)$$

Bioconcentration Factor ($logBCF$)

$$logBCF = 0.3311(R) - 2.8358 \quad (7a)$$

$$logBCF = 0.0577(NSDD) - 2.4650 \quad (7b)$$

$$logBCF = 0.2853(R) - 2.6714 \quad (7c)$$

$$logBCF = 0.4265(NSCI) - 2.7265 \quad (7d)$$

Octanol-Water Partition Coefficient ($logP$)

$$logP = 0.4725(SCI) - 4.0442 \quad (8a)$$

$$logP = 0.3924(NABC) - 4.1525 \quad (8b)$$

$$logP = 0.4205(R) - 3.9805 \quad (8c)$$

$$logP = 0.6317(NSCI) - 4.0986 \quad (8d)$$

3.4. Critical Comparison and Scientific Explanation

Table 11. Comparative QSPR Performance Summary: Best In-Sample R^2 Values for the G and $N_e(G)$ Representations

Property	$G(R^2)$	$N_e(G)(R^2)$	ΔR^2	Higher R^2
MV	0.9390	0.9472	+0.0082	$N_e(G)$
MR	0.9297	0.9188	-0.0109	G
P	0.9296	0.9186	-0.0110	G
$logK_{oc}$	0.8018	0.8145	+0.0127	$N_e(G)$
$logBCF$	0.7868	0.7991	+0.0123	$N_e(G)$
$logP$	0.6629	0.6814	+0.0185	$N_e(G)$

Note: Values represent the best in-sample R² across all tested descriptors within each graph-descriptor combination. $\Delta R^2 = R^2[N_e(G)] - R^2[G]$; positive ΔR^2 values indicate higher R^2 for the $N_e(G)$ representation, whereas negative values indicate higher R^2 values for the G representation. All differences should be interpreted as descriptor- and dataset-dependent tendencies.

Although the differences in R^2 values between the G and $N_e(G)$ representations summarized in Table 11 are mathematically marginal, with ΔR^2 values ranging approximately from 0.008 to 0.018, they provide useful structural insights for this specific dataset. Therefore, rather than claiming absolute superiority of one representation over the other, these results should be evaluated as descriptor- and dataset-dependent tendencies that require validation on larger and structurally broader datasets before any generalization is warranted.

3.4.1. Performance of the Atom-Centered G Representation for Bulk Properties (MR, P)

Molar Refractivity (MR) and Polarizability (P), which are intrinsically linked to the molecule's total electron cloud, volume, and size, were marginally better predicted by models based on the atom-centered G representation than by those based on the $N_e(G)$ representation (Table 11). These properties are classically considered additive and extensive, meaning that they depend on the cumulative contribution of all structural components. The observed differences are small, but their direction is chemically reasonable within the present descriptor set.

The atom-centered G representation, which explicitly represents all atoms (including hydrogens) as vertices, is well suited for encoding global, extensive information. In this context, the inclusion of hydrogen atoms contributes to a comprehensive representation of the molecular structure. The Geometric-Arithmetic index calculated on G, $GA(G)$, exhibited the strongest predictive performance for both *MR* ($R^2 = 0.9297$) and *P* ($R^2 = 0.9296$) within the present work. These findings suggest that the G representation remains useful for modeling properties primarily governed by overall molecular dimensions and electron density.

3.4.2. Property-Dependent Performance of Bond-Centered $N_e(G)$ Representation for Environmental Fate Properties

The analysis indicates a dataset-specific tendency consistent with the working hypothesis for interaction-driven properties. The $N_e(G)$-based models yielded modestly higher R^2 values for $logK_{oc}$,

$logBCF$ and $logP$ compared with the corresponding best-performing G-based models (Table 11). These outcomes should be understood as comparative observations within the present dataset and descriptor panel, rather than as evidence supporting a universally superior bond-centered representation.

These environmental fate parameters are not governed solely by bulk molecular size; they may also be influenced by localized interaction mechanisms, including hydrogen bonding potential, hydrophobic interactions, and electronic complementarity with environmental media such as soil organic carbon or lipid bilayers. This chemical interpretation provides a plausible explanation for the observed trend, but it remains inferential because the study uses single topological descriptors rather than direct mechanistic interaction variables.

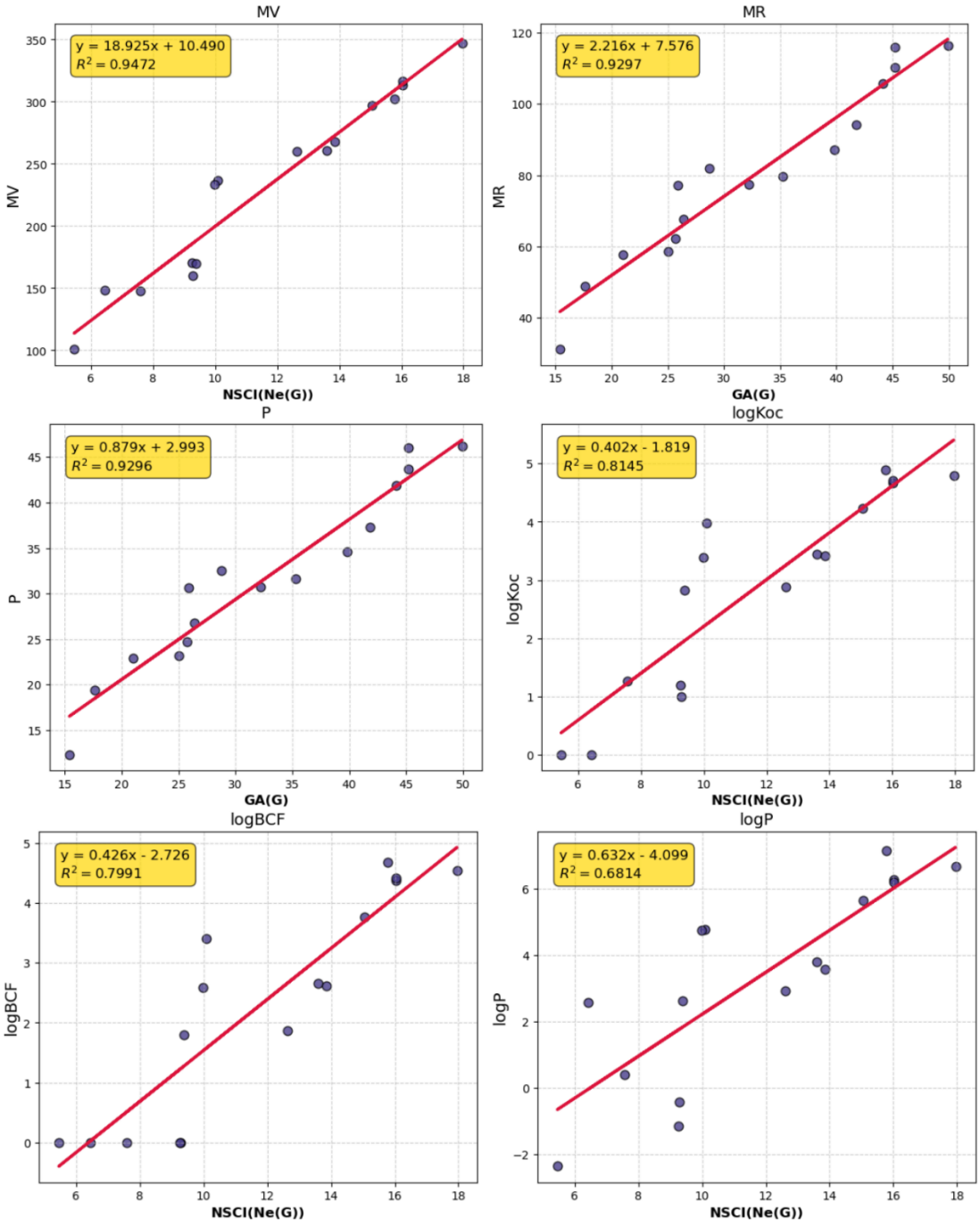

Figure 4. Regression Plots for the Highest R^2 Values Achieved for the Six Physicochemical Properties (MV: $N_e(G)NSCI$; MR/P: G GA; $logK_{oc}/logBCF/logP$: $N_e(G)NSCI$).

The $N_e(G)$ representation shifts the representational focus from individual atoms to the bonding environment. By treating bonds, corresponding to localized chemical functionalities such as C-H or C-O linkages, as vertices, $N_e(G)$ emphasizes connectivity patterns that may be more directly related to local functional group behavior. This perspective may help descriptors respond to localized steric and electronic environments relevant to phase partitioning processes.

The Neighborhood Sum Connectivity Index calculated on the $N_e(G)$ representation, $NSCI$ ($N_e(G)$), was identified as the most effective descriptor for the interaction-based properties considered, yielding R^2 values of 0.8145 for $logK_{oc}$, 0.7991 for $logBCF$, and 0.6814 for $logP$. The improvement over G-based models is modest but consistent across these three endpoints. Therefore, the result is best interpreted as a useful comparative tendency, and as a basis for future validation with larger datasets and additional descriptor families. This comparative behavior is illustrated in Figure 4.

4. CONCLUSION

This investigation shows that the choice of molecular graph representation can influence QSPR model behavior, particularly when the target property differs in chemical character, such as bulk-related versus interaction-driven properties. The primary contribution of this study is the controlled comparison of bond-centered edge-neighborhood $N_e(G)$ representation and the atom-centered G representation for hydrogen-inclusive molecular structures.

Within the selected descriptor panel, the atom-centered G representation exhibited effective performance for extensive, volume-dependent properties such as Molar Refractivity and Polarizability, whereas the $N_e(G)$ representation produced slightly higher R^2 values for environmental fate parameters, including soil sorption potential ($logK_{oc}$), bioconcentration factor ($logBCF$), and octanol-water partitioning ($logP$). These differences are interpreted as property-dependent tendencies rather than universal representation-level conclusions.

Overall, the findings indicate a complementary rather than substitutive relationship between the atom-centered G representation and the bond-centered $N_e(G)$ representation. The suitability of each approach appears to depend on the physicochemical nature of the target property, the descriptor family used, and the modeling framework. In this context, the $N_e(G)$ representation may serve as a useful early-stage screening framework for cheminformaticians involved in the rational design and environmental risk assessment of minimum-risk pesticides, provided that future validation confirms the observed trends.

4.1. Limitations and Future Work

This study adopts a comparative and exploratory framework aimed at assessing the relative representational behavior of different molecular graph representations within QSPR modeling. The results provide preliminary support for the usefulness of $N_e(G)$-based descriptors for selected interaction-driven endpoints, but they should not be generalized beyond the present dataset, descriptor panel, and SLR framework.

The dataset contains 16 pesticide molecules. This limited size is acceptable for a controlled comparison of graph representations, but it restricts the strength of general predictive claims. In particular, small differences between R^2 values should be interpreted as directional tendencies rather than decisive evidence.

In light of the scope of the present study, several directions may be considered for future research to further explore the applicability of the $N_e(G)$ approach.

Expanded Datasets and Validation: Future studies may examine the performance of the $N_e(G)$ framework using larger and more structurally diverse datasets. Such extensions would facilitate additional evaluation of model robustness, applicability domain, and external predictivity. Where appropriate, established validation strategies commonly adopted in QSPR studies, including cross-validation and external test sets, should be incorporated to complement the comparative findings reported here [18].

Diverse Ecotoxicological Endpoints: Beyond the physicochemical properties investigated in this work, $N_e(G)$-based descriptors may also be explored in relation to direct ecotoxicological endpoints, such as

LC_{50} values for aquatic organisms or LD_{50} values for pollinators. Examining such endpoints could provide further insight into the relevance of bond-centered graph representations for biologically meaningful environmental risk parameters.

Advanced Modeling Techniques: While the present study focuses on Simple Linear Regression to ensure methodological transparency and interpretability, future investigations should consider more advanced modeling approaches. Techniques such as Multiple Linear Regression, partial least squares, random forest regression, support vector regression, or graph-based machine learning models could be employed to investigate potential synergies between descriptors derived from the G and $N_e(G)$ representations. Such analyses would clarify whether the observed tendencies persist when multiple descriptors are allowed to interact.

Descriptor-Set Dependence: Future work should also test whether the tendencies observed here persist when additional descriptor families are included. Newer Zagreb-type descriptors, including eta, rho, tau, upsilon and omicron variants, as well as Sombor-type descriptors, entropy-based indices, and electronic or geometric descriptors, may alter the relative performance of the G and $N_e(G)$ representations. This is important because the present results reflect descriptor-representation interactions rather than graph representation alone.

Acknowledgements

The authors declare that no financial or institutional support was received for the conduct of this study.

Conflict of Interest Statement

The authors declare no conflict of interest.

Statement of Research and Publication Ethics

The authors declare that this study does not require ethical approval because it is a theoretical study based on mathematical modeling and existing chemical data and does not involve human or animal subjects.

Artificial Intelligence (AI) Contribution Statement

During the revision process, AI-based language assistance was used only to improve grammar, clarity, organization, and readability. No AI tool was used for data generation, molecular graph construction, topological index calculation, statistical modeling, figure preparation, or scientific interpretation. The authors reviewed, verified, and approved all scientific content, interpretations, calculations, and conclusions.

Contributions of the Authors

Conceptualization, methodology, supervision, and writing – review and editing were carried out by Fatma Özen Erdoğan; data curation, formal analysis, investigation, software, and writing – original draft were carried out by Betül Öztürk.

REFERENCES

[1] S. Parveen, F. Saeed, F. B. Farooq, N. Parveen, N. Idrees, S. Nasir, and R. Fanja, "QSPR modeling of fungicides using topological descriptors," Int. J. Anal. Chem., vol. 2023, no. 1, p. 9625588, 2023.

[2] P. Gramatica, M. Corradi, and V. Consonni, "Modelling and prediction of soil sorption coefficients of non-ionic organic pesticides by molecular descriptors," Chemosphere, vol. 41, no. 5, pp. 763-777, 2000.

[3] European Chemicals Agency (ECHA), "ANNEX XV Restriction Report – Proposal for a Restriction: Per- and polyfluoroalkyl substances (PFASs)," Helsinki, Finland, 2023.

[4] W. M. S. Russell and R. L. Burch, The Principles of Humane Experimental Technique. London, U.K.: Methuen, 1959.

[5] A. R. Katritzky, V. S. Lobanov, and M. Karelson, "QSPR: The correlation and quantitative prediction of chemical and physical properties from structure," Chem. Soc. Rev., vol. 24, no. 4, pp. 279-287, 1995.

[6] L. B. Kier and L. H. Hall, Molecular Connectivity in Structure-Activity Analysis. Letchworth, U.K.: Research Studies Press, 1986.

[7] İ. Çiftçi, "Modeling the physicochemical characteristics of benzenes through the application of Zagreb rho indices," Comb. Chem. High Throughput Screen., vol. 28, no. 19, pp. 3430-3439, 2025.

[8] İ. Çiftçi, K. Yamaç, and İ. H. Denizler, "Modeling benzene physicochemical properties using Zagreb upsilon indices," Open Chem., vol. 23, no. 1, Art. no. 20250160, 2025.

[9] İ. Çiftçi, "Modeling the physicochemical characteristics of benzene compounds through the application of Zagreb omicron indices," Cumhuriyet Sci. J., vol. 46, no. 2, pp. 360-368, 2025.

[10] İ. Çiftçi, "Modeling the physicochemical characteristics of benzene compounds utilizing Zagreb eta indices," Journal of Chemistry, vol. 2025, Article ID 6814222, pp. 1-10, 2025.

[11] J. V. Knop, D. Plavšić, M. Randić, and N. Trinajstić, "Chemical graph theory. V. On the classification of topological biradicals," Croat. Chem. Acta, vol. 56, no. 3, pp. 411-441, 1983.

[12] I. Gutman, "Degree-based topological indices," Croat. Chem. Acta, vol. 86, no. 4, pp. 351-361, 2013.

[13] R. S. Boethling and D. Mackay, Property Estimation Methods for Chemicals. Boca Raton, FL, USA: Environmental and Health Science, 2000.

[14] Sarika, S. G., and R. Manjusha, "Graph theory and molecular behavior: Utilizing Platt numbers and polynomial approximations for boiling point estimations," Engineering Reports, vol. 7, no. 2, Art. no. e13124, 2025.

[15] S. Mondal, A. Dey, N. De, and A. Pal, "QSPR analysis of some novel neighbourhood degree-based topological descriptors," Complex Intell. Syst., vol. 7, pp. 977-996, 2021.

[16] M. W. Rasheed, A. Mahboob, and I. Hanif, "Uses of degree-based topological indices in QSPR analysis of alkaloids with poisonous and healthful nature," Front. Phys., vol. 12, 1381887, 2024.

[17] G. Landrum, RDKit: Open-source cheminformatics software, 2006. [Online]. Available: https://www.rdkit.org

[18] R. Kiralj and M. M. Ferreira, "Basic validation procedures for regression models in QSAR and QSPR studies: Theory and application," J. Braz. Chem. Soc., vol. 20, no. 4, pp. 770-787, 2009.

[19] S. Ediz, "Predicting some physicochemical properties of octane isomers: A topological approach using ev-degree and ve-degree Zagreb indices," arXiv:1701.02859, 2017.

[20] D. Wardecki, M. Dołowy, and K. Bober-Majnusz, "Evaluation of the usefulness of topological indices for predicting selected physicochemical properties of bioactive substances with anti-androgenic and hypouricemic activity," Molecules, vol. 28, no. 15, 5822, 2023.

[21] A. Ullah, S. Jabeen, S. Zaman, A. Hamraz, and S. Meherban, "Predictive potential of K-Banhatti and Zagreb type molecular descriptors in structure-property relationship analysis of some novel drug molecules," J. Chin. Chem. Soc., vol. 71, no. 3, pp. 250-276, 2024, doi: 10.1002/jccs.202300450.

[22] K. Yamaç, "Modelling physico-chemical properties of benzenes via Zagreb tau indices," CBUJOS, vol. 21, no. 4, pp. 51-61, 2025.