A Composite Topological Descriptor for Modeling Stability and Reactivity in Polycyclic Conjugated Hydrocarbons

Hamad Mukhtar*, Tahira Perveen Nawaz, Muhammad Saleem

Department of Mathematics and Statistics,National College of Business Administration & Economics (NCBA&E),Multan Sub-Campus, Lahore, Pakistan

khamad245@gmail.com

Keywords: *Polycyclic conjugated hydrocarbons, Topological indices, QSPR, Molecular stability, Aromatic hydrocarbons.*

Polycyclic conjugated hydrocarbons (PCHs) are π-conjugated molecules whose stability and electronic reactivity are closely related to molecular topology. In this study, a composite topological descriptor based on the Wiener, Harary, Randić, Zagreb and metric degree polynomial indices was developed for QSPR modeling of condensed aromatic hydrocarbons. A set of structurally different PCH molecules was analyzed using reported thermodynamic and electronic properties. Thermodynamic stability was represented by the enthalpy of sublimation (ΔHsub), while molecular reactivity was approximated using a scaled reciprocal ionization energy descriptor. Correlation and regression analyses showed that the composite descriptor effectively reflects stability and reactivity trends in fused aromatic hydrocarbons.

Introduction

Accurate structure–property modelling remains a key objective in theoretical and computational chemistry. In π-conjugated systems such as polycyclic conjugated hydrocarbons (PCHs), molecular topology and connectivity strongly influence thermodynamic stability, electronic structure and chemical reactivity. Understanding how structural features relate to measurable physicochemical properties is therefore essential for the rational design of condensed aromatic materials and functional molecular systems [1–3].

Chemical graph theory provides a convenient framework for representing molecular structures as graphs in which atoms correspond to vertices and chemical bonds correspond to edges. Within this framework, numerical descriptors known as topological indices encode structural information and are widely used in quantitative structure–property relationship (QSPR) modelling [4–6]. Classical descriptors such as the Wiener index characterize molecular compactness through path-length distributions, whereas degree-based indices including the Randić and Zagreb indices describe local connectivity patterns within molecular graphs [7–10]. These descriptors have been successfully applied to correlate structural features with thermodynamic and electronic properties of organic molecules [11,12].

However, individual topological indices may suffer from descriptor degeneracy, where structurally different molecular graphs produce identical numerical values. This limitation becomes more pronounced in highly fused polycyclic hydrocarbons with similar connectivity patterns [13]. To address this issue, composite descriptors integrating several graph invariants have been proposed to improve structural discrimination and predictive capability in QSPR models [14,15].

Thermodynamic stability of aromatic hydrocarbons can be evaluated using experimental enthalpies of sublimation (ΔHsub), while electronic reactivity trends may be approximated using ionisation-energy-related descriptors [16,17]. In the present work, a log-transformed composite topological descriptor combining Wiener, Harary, Randić, Zagreb and metric degree polynomial indices is introduced for modelling stability and reactivity trends in condensed PCH systems. Statistical relationships are analysed using regression modelling with ridge regularisation and validated using leave-one-out cross-validation.

Recent studies further demonstrate the continuing relevance of graph-theoretical descriptors in modelling physicochemical properties of aromatic and heteroaromatic compounds, highlighting the need for improved composite descriptors capable of capturing structural variations in polycyclic chemical space [18–22].

Despite extensive use of individual topological indices in quantitative structure–property relationship (QSPR) studies of polycyclic aromatic systems, most existing approaches rely on single-descriptor models or independently applied graph invariants, which often fail to fully capture the combined effects of global molecular topology and local connectivity patterns in highly fused aromatic frameworks. In particular, no unified framework has been established that systematically integrates distance-based indices, degree-based connectivity measures, and higher-order polynomial graph descriptors within a single chemically interpretable representation for polycyclic conjugated hydrocarbons (PCHs).

In this study, we propose a novel log-transformed composite topological descriptor that integrates Wiener, Harary, Randić, Zagreb, and metric degree polynomial indices into a unified structural representation. The proposed formulation is designed to simultaneously capture molecular size, compactness, branching behavior, and fusion-dependent connectivity effects in condensed aromatic systems. Unlike conventional approaches that evaluate descriptors individually or rely on purely statistical feature selection, the present framework provides a structurally motivated aggregation of complementary graph invariants.

Furthermore, the proposed descriptor is systematically evaluated using ridge-regularized regression and leave-one-out cross-validation (LOOCV), enabling robust assessment of predictive stability and multicollinearity effects. The integration of statistical validation with chemical interpretability distinguishes this work

from previous QSPR studies and provides a more comprehensive descriptor framework for modeling stability and electronic reactivity in polycyclic conjugated hydrocarbons.

The novelty of this study lies in the construction of a unified log-transformed composite topological descriptor that integrates both distance-based and degree-based graph invariants into a single chemically interpretable framework. Unlike traditional QSPR approaches that rely on isolated descriptors or purely data-driven feature selection, the present work introduces a structured mathematical aggregation of complementary topological indices. Furthermore, the integration of ridge-regularized regression with LOOCV validation provides a statistically robust modeling strategy that addresses multicollinearity and improves predictive stability in polycyclic conjugated hydrocarbons.

Experimental Section

Molecular Dataset

A structurally diverse set of polycyclic conjugated hydrocarbons (PCHs) was considered, encompassing linear, angular, compact, and highly condensed fused-ring architectures. The molecular set includes naphthalene, azulene, acenaphthylene, fluorene, anthracene, phenanthrene, fluoranthene, pyrene, chrysene, triphenylene, tetracene, benzo[a]pyrene, benzo[e]pyrene, benzo[b]fluoranthene, perylene, dibenz[a,h]anthracene, pentacene, benzo[ghi]perylene, hexacene, coronene, circumcoronene, dicoronylene, and ovalene (**Table 1**).

Experimental property values were compiled primarily from the NIST Chemistry WebBook. Thermodynamic stability was characterized by the enthalpy of sublimation (ΔH_sub / kJ mol^{-1}). Electronic reactivity was quantified using ionization energy (IE, eV) as a proxy descriptor and expressed in the modeling framework as a scaled reciprocal ionization-energy measure (scaled 1/IE).

Table 1. Dataset of polycyclic conjugated hydrocarbons used in this study, including molecular formula, ring count, experimental enthalpy of sublimation, scaled reciprocal ionization-energy proxy, and data source.

Sr.	Molecule	Formula	Rings / Type	ΔHsub / kJ mol^{-1}	Scaled 1/IE / cm^3·molecule^{-1}·s^{-1}	Data Source
1	Naphthalene	$C_{10}H_8$	2	72.6	1.54×10^{-11}	NIST Chemistry WebBook
2	Azulene	$C_{10}H_8$	2	78.4	1.68×10^{-11}	NIST Chemistry WebBook
3	Acenaphthylene	$C_{12}H_8$	3	72	1.54×10^{-11}	NIST Chemistry WebBook
4	Fluorene	$C_{13}H_{10}$	3	84	1.58×10^{-11}	NIST Chemistry WebBook
5	Anthracene	$C_{14}H_{10}$	3	98.6	1.68×10^{-11}	NIST Chemistry WebBook
6	Phenanthrene	$C_{14}H_{10}$	3	92.9	1.58×10^{-11}	NIST Chemistry WebBook
7	Fluoranthene	$C_{16}H_{10}$	4	101.2 ± 2.8	1.72×10^{-11}	NIST Chemistry WebBook
8	Pyrene	$C_{16}H_{10}$	4	100.3	1.68×10^{-11}	NIST Chemistry WebBook
9	Chrysene	$C_{18}H_{12}$	4	123.4	1.65×10^{-11}	NIST Chemistry WebBook
10	Triphenylene	$C_{18}H_{12}$	4, compact	120.1 ± 3.3	1.59×10^{-11}	NIST Chemistry WebBook
11	Tetracene	$C_{18}H_{12}$	4, linear	118	1.79×10^{-11}	NIST Chemistry WebBook
12	Benzo[a]pyrene	$C_{20}H_{12}$	5	124	1.76×10^{-11}	NIST Chemistry WebBook
13	Benzo[e]pyrene	$C_{20}H_{12}$	5	121	1.68×10^{-11}	NIST Chemistry WebBook
14	Benzo[b]fluoranthene	$C_{20}H_{12}$	5	121	1.60×10^{-11}	NIST Chemistry WebBook
15	Perylene	$C_{20}H_{12}$	5	135.9	1.80×10^{-11}	NIST Chemistry WebBook
16	Dibenz[a,h]anthracene	$C_{22}H_{14}$	5	148.9	1.69×10^{-11}	NIST Chemistry WebBook
17	Pentacene	$C_{22}H_{14}$	5, linear	184 ± 10	1.89×10^{-11}	NIST Chemistry WebBook
18	Benzo[ghi]perylene	$C_{22}H_{12}$	6	127	1.74×10^{-11}	NIST Chemistry WebBook
19	Hexacene	$C_{26}H_{16}$	6	≈120, 300 K est.	1.84×10^{-11}	NIST Chemistry WebBook
20	Coronene	$C_{24}H_{12}$	7, compact	131.0 ± 1.7	1.72×10^{-11}	NIST Chemistry WebBook
21	Circumcoronene	$C_{54}H_{18}$	8	317.5	2.03×10^{-11}	NIST Chemistry WebBook
22	Dicoronylene	$C_{48}H_{20}$	9	305	2.05×10^{-11}	NIST Chemistry WebBook
23	Ovalene	$C_{32}H_{14}$	10	211.7 ± 7.9	1.82×10^{-11}	NIST Chemistry WebBook

Table 2. Structural Categories

Category	Examples
Linear	Naphthalene, Anthracene, Tetracene, Pentacene, Hexacene
Angular	Phenanthrene, Chrysene
Compact	Pyrene, Perylene, Coronene
Highly condensed	Circumcoronene, Dicoronylene, Ovalene

Although the present benchmark set comprises 23 polycyclic conjugated hydrocarbons, it spans a broad topological spectrum, including 2–10 fused rings and representative linear, angular, compact, and highly condensed architectures. This structural diversity ensures substantial variation in graph diameter, degree distribution, and fusion topology, thereby providing a meaningful testing ground for composite descriptor evaluation. In moderate-sized but chemically diverse datasets, leave-one-out cross-validation (LOOCV) offers an appropriate and statistically rigorous internal validation strategy, as each compound is systematically excluded and predicted under maximal data utilization. Consequently, while the dataset size is moderate, its topological heterogeneity supports robust structure–property assessment within condensed polycyclic chemical space.

Reactivity Proxy Based on Ionization Energy

Electronic reactivity was represented using a reciprocal ionization-energy-based proxy.

Ionization energy (IE) reflects the energy required to remove an electron from a molecule. In π-conjugated hydrocarbons, a lower IE generally indicates easier electron removal and a greater tendency toward electronic response. Therefore, while IE itself may be viewed as an electronic stability-related quantity, its reciprocal provides a convenient way to express the trend in the direction of increasing electronic reactivity.

The reactivity proxy was defined as: $R_{IE} = \frac{1}{IE}$

where IE is the ionization energy expressed in electron volts (eV). Since smaller IE values correspond to easier ionization, larger R_{IE}RIE values indicate greater relative electronic reactivity.

For tabular readability and numerical comparison across the molecular series, a uniformly scaled reciprocal form was used: $R_{scaled} = \lambda\left(\frac{1}{IE}\right)$

where $\lambda$is a constant scaling factor applied equally to all molecules. In the present dataset, the scaling factor was used only to place the reciprocal IE values on a convenient 10^{-11}10−11 numerical scale. This transformation does not alter the rank order, correlation coefficient, regression trend, or comparative interpretation of the reactivity proxy.

The scaled reciprocal ionization-energy descriptor should not be interpreted as an experimentally measured kinetic rate constant. Rather, it is used here as a relative electronic reactivity indicator suitable for comparing topological trends among the studied polycyclic conjugated hydrocarbons. For this reason, the values are reported as R_{scaled} in relative units.

Graph Construction

Each molecule was represented as a simple, undirected molecular graph G = (V, E), where vertices (V) correspond to non-hydrogen atoms and edges (E) represent covalent bonds. Hydrogen atoms were suppressed, and bond multiplicities were treated as single edges, consistent with standard chemical graph-theoretic practice for topological index calculations. Molecular graphs were generated from the structural representations of the 23 PCHs, and the corresponding adjacency and shortest-path distance information was used to compute distance-based and degree-based descriptors.

Composite Index Definition

To capture complementary aspects of both global distance organization and local connectivity patterns in fused polycyclic systems, a composite topological descriptor $C(G)$C(G) was constructed by integrating multiple classical graph invariants within a unified framework. Each selected descriptor encodes distinct but structurally meaningful information about polycyclic conjugated hydrocarbons (PCHs), including molecular compactness, branching behavior, degree distribution, and fusion-dependent connectivity.

The composite index is defined as:

$C(G) = A_1 \log_{10} W(G) + A_2 \log_{10} R(G) + A_3 \log_{10} H(G) + A_4 \log_{10} M1(G) + A_5 \log_{10} MDP(G, 2)$.

where $W(G)$ is the Wiener index capturing global path-length based molecular size, $R(G)$ is a Randic-type connectivity index describing branching and edge-degree interactions, $H(G)$ is the Harary index reflecting reciprocal distance-based compactness, $M_1(G)$ is the first Zagreb index representing degree distribution characteristics, and $MDP(G,2)$ denotes the metric degree polynomial term evaluated at parameter 2, capturing higher-order degree–distance fusion effects in condensed aromatic graphs.

A logarithmic transformation was applied to each descriptor to reduce scale heterogeneity, stabilize variance, and prevent dominance of magnitude-sensitive indices such as Wiener and Zagreb terms. This transformation also improves numerical comparability across structurally diverse polycyclic systems and enhances regression stability in subsequent QSPR modeling.

In the present study, a baseline equal-weight formulation was adopted, where $A_1 = A_2 = A_3 = A_4 = A_5 = 1$. This choice provides an unbiased structural representation without introducing parameter fitting at the descriptor construction stage, thereby ensuring that the composite index reflects an intrinsic aggregation of classical topological information. Importantly, this baseline formulation serves as a reference model against which optimized regression-based weighting schemes (e.g., ridge regularization) are later evaluated.

No prior normalization was applied to individual log-transformed components in the baseline construction, allowing direct preservation of their intrinsic structural contributions. However, standardized representations can be introduced for weight optimization and cross-dataset transferability analysis, as demonstrated in the regression framework.

All molecular graphs were constructed as hydrogen-suppressed, undirected graphs $G = (V, E)$, and all descriptors were computed consistently on the same structural representation. The resulting composite descriptor $C(G)$ was subsequently used as a unified structural index for correlation analysis, regression modeling, and validation against experimental stability and electronic reactivity properties of PCH systems.

Descriptor Standardization and Weight Optimization

To examine the statistical robustness of the equal-weight composite formulation, log-components were standardized (z-score normalization) prior to weight optimization. Ridge-regularized regression ($\alpha = 0.1$) was applied under leave-one-out cross-validation (LOOCV) to derive optimized descriptor weights.

The resulting weight profile highlighted dominant contributions from the Randić and Harary components, while Zagreb and MDP contributions exhibited compensatory behavior.

LOOCV yielded RMSEcv = 27.7 kJ / mol^{-1}, confirming predictive robustness.

To further assess multivariate structure and descriptor redundancy, principal component analysis (PCA) was performed on the

standardized log-descriptor matrix. The first principal component accounted for approximately 90.8% of the total variance, indicating that the five classical graph invariants align predominantly along a single dominant structural axis in descriptor space. This result confirms that the composite formulation captures the principal variance direction governing polycyclic fusion and condensation patterns. The equal-weight baseline therefore approximates the dominant structural manifold, while ridge optimization refines predictive balance.

The optimized ridge coefficients (standardized space) were $\beta = [51.47, 48.77, -125.26, 109.85, -32.81]$ corresponding to [logW, logH, logM1, logR, logMDP], respectively.

$$CG_{ridge}^{(Stability)}(G) = 51.47Z_W + 48.77Z_H - 125.26Z_{M1} + 109.85Z_R - 32.81Z_{MDP}$$

$$CG_{ridge}^{(reactivity)}(G) = 5.677 \times 10^{-1}\ Z_W + 6.3777 \times 10^{-13}Z_H - 1.755 \times 10^{-12}Z_{M1} + 1.380 \times 10^{-12}Z_R + 2.802 \times 10^{-13}Z_{MDP}$$

$$Z_i = \frac{log(descripter_i) - \mu_i}{\sigma_i}$$

The magnitude and sign of the coefficients indicate dominant contributions from Randić-type connectivity and Harary reciprocal-distance effects, with compensatory influence from Zagreb and metric degree polynomial terms.

The explained variance of the first three principal components is shown in **Figure 1**

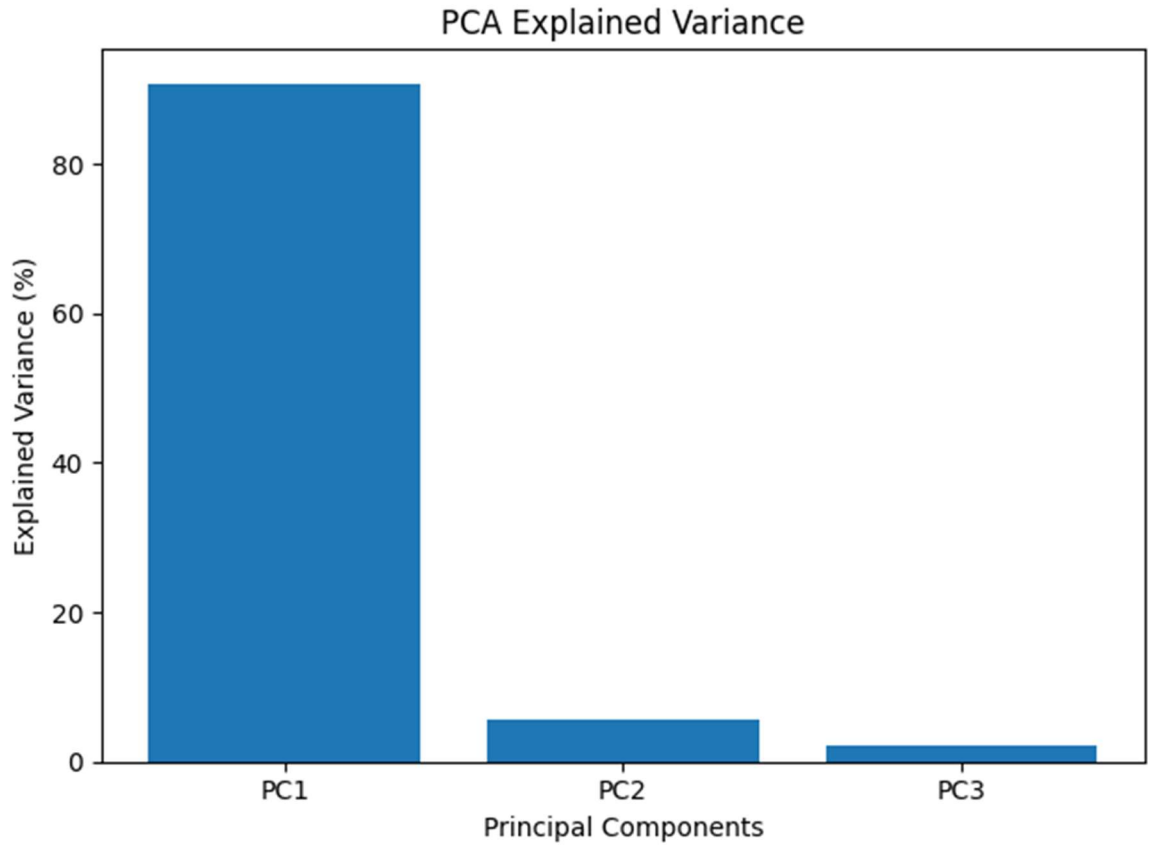

Figure 1. Percentage of total variance explained by the first three principal components of the standardized log-descriptor matrix. The first principal component accounts for approximately 90.8% of total variance, indicating strong alignment of classical topological invariants along a dominant structural axis.

Statistical Analysis

The relationship between the composite index C(G) and experimental property proxies was evaluated using both rank-based association and linear regression. Spearman's rank correlation coefficient (rho) was computed to quantify monotone dependence between C(G) and each endpoint across the dataset. In addition, ordinary least squares regression was used to assess linear structure–property modeling performance, and the coefficient of determination (R-squared) was reported as a goodness-of-fit measure.

Stability was represented by the enthalpy (heat) of sublimation, ΔH_sub / kJmol^{-1}. Reactivity was represented by a scaled reciprocal ionization-energy proxy (scaled 1/IE), reported in cm^3·molecule^{-1}·s^{-1} as listed in **Table 1**. Experimental values were compiled primarily from the NIST Chemistry WebBook, with literature supplementation for compounds where NIST entries were unavailable or incomplete (as indicated in the reference column of **Table 1**).

Computational Tools

All computations were performed using widely used computational chemistry and scientific computing tools.

Table 3. Computational tools used in this study and their purposes.

Tool	Purpose
RDKit	Molecular structure parsing and molecular graph construction
NumPy	Numerical computation and array operations
SciPy	Statistical analysis (Spearman correlation, linear regression utilities)
scikit-learn	Regression modeling (OLS, Ridge), standardization, and LOOCV validation
pandas	Tabular data handling and model comparison tables
NetworkX	Graph algorithms and topological descriptor computations
Matplotlib	Visualization (scatter plots, regression lines, model diagnostics)

These tools support accuracy, reproducibility, and scalability of the computational workflow.

Results and discussion

Composite Index Construction

In this section, the composite index is constructed from five classical graph invariants selected to reflect complementary aspects of polycyclic topology. Specifically, Wiener and Harary indices capture global distance organization, while Randić and the first Zagreb index encode local degree-driven connectivity patterns. The metric degree polynomial term MDP(G,2) further introduces a degree–distance layering component that is sensitive to fusion and condensation effects in PCH graphs. **Table 4** summarizes the component descriptors, their mathematical definitions, and their qualitative structural interpretations.

Table 4. Descriptors comprising the composite index and their structural meanings

Descriptor	Formula	Structural meaning
Wiener index, W(G)	$W(G) = \sum_{u,v \subset V} d_{uv}$	Global molecular size + overall distance spread
Randić index, R(G)	$R(G) = \sum_{(u,v)\in E} 1/\sqrt{deg(u) \cdot deg(v)}$	Branching/degree interaction sensitivity
Harary index, H(G)	$H(G) = \sum_{u,v \subset V} 1/d_{uv}$	Compactness via reciprocal distances
First Zagreb index, M1(G)	$M_1(G) = \sum_{v\in V}\big(deg(v)\big)^2$	Local connectivity + degree distribution
Metric degree polynomial term, MDP(G,2)	$MDP(G,k) = \sum_{v\in V}\big(S(v)\big)^k$, $S(v) = \sum_{u\in V, u\neq v} deg\,(u)/d\,(u,v)$	Degree–distance layering; fusion/condensation sensitivity

The composite index defined in the Methods section was calculated for all selected PCH molecules and used for subsequent correlation, regression, validation, and degeneracy analyses.

Since all molecular graphs considered in this study are connected hydrogen-suppressed graphs, the selected topological indices are positive and finite. Therefore, their logarithmic transformations are well-defined. The mathematical details are straightforward and are not emphasized here; instead, the discussion focuses on the chemical role of each descriptor component in distinguishing polycyclic fusion patterns.

Statistical Evaluation

In this section, the composite index is evaluated against experimentally compiled Stability and reactivity proxies for the considered PCH systems. To assess monotonic structure–property relationships without imposing linearity assumptions, Spearman's rank correlation coefficient (ρ) is computed for each endpoint. Stability is characterized by the enthalpy of sublimation (ΔH_sub), while reactivity is quantified using the scaled reciprocal ionization-energy proxy (scaled 1/IE), as summarized in **Table 5**

Tables 5 and 6 report the correlation statistics, providing a quantitative measure of association strength and direction between C(G) and the experimental properties. In addition, linear regression analysis is performed to evaluate explained variance and overall goodness-of-fit.

Table 5. Composite graph index (C(G)) with experimental Stability (ΔH_sub) and reactivity proxy (scaled 1/IE) for 23 PCHs.

Sr	Molecule	C(G)	Stability / kJ mol^{-1}	Reactivity proxy (cm^3·molecule^{-1}·s^{-1})
1	Naphthalene	8.9673	72.6	1.54×10^{-11}
2	Azulene	8.9317	78.4	1.68×10^{-11}
3	Acenaphthylene	9.7095	72	1.54×10^{-11}
4	Fluorene	10.1638	84	1.58×10^{-11}
5	Anthracene	10.5813	98.6	1.68×10^{-11}
6	Phenanthrene	10.5044	92.9	1.58×10^{-11}
7	Fluoranthene	11.0046	101.2 ± 2.8	1.72×10^{-11}
8	Pyrene	11.0078	100.3	1.68×10^{-11}
9	Chrysene	11.7941	123.4	1.65×10^{-11}
10	Triphenylene	11.5481	120.1 ± 3.3	1.59×10^{-11}
11	Tetracene	11.9092	118	1.79×10^{-11}
12	Benzo[a]pyrene	12.1655	124	1.76×10^{-11}
13	Benzo[e]pyrene	11.9845	121	1.68×10^{-11}
14	Benzo[b]fluoranthene	12.1044	121	1.60×10^{-11}
15	Perylene	11.9979	135.9	1.80×10^{-11}
16	Dibenz[a,h]anthracene	12.9566	148.9	1.69×10^{-11}
17	Pentacene	13.0798	184 ± 10	1.89×10^{-11}
18	Benzo[ghi]perylene	12.3867	127	1.74×10^{-11}
19	Hexacene	14.1504	≈120 (300 K est.)	1.84×10^{-11}
20	Coronene	12.7539	131.0 ± 1.7	1.72×10^{-11}
21	Circumcoronene	14.6901	317.5	2.03×10^{-11}
22	Dicoronylene	15.1822	305	2.05×10^{-11}
23	Ovalene	14.1233	211.7 ± 7.9	1.82×10^{-11}

Correlation with Experimental Properties

In this subsection, the relationship between the selected topological descriptors and the experimental properties of the studied PCHs is examined. Spearman rank correlation analysis is used to evaluate the strength and direction of association with the Stability and reactivity proxies. The results are summarized in **Table 5** and provide the basis for subsequent regression modeling.

Table 6. Correlation with Experimental Properties

Property	Spearman Correlation (ρ)	Interpretation
Stability	0.9073	Strong positive monotonic association
Reactivity	0.8283	Strong positive monotonic association

The Spearman correlation results indicate a strong positive monotonic association between the composite index and thermodynamic Stability (**Table 6**). The relationship between the composite index and thermodynamic Stability is illustrated in **Figure 2**.

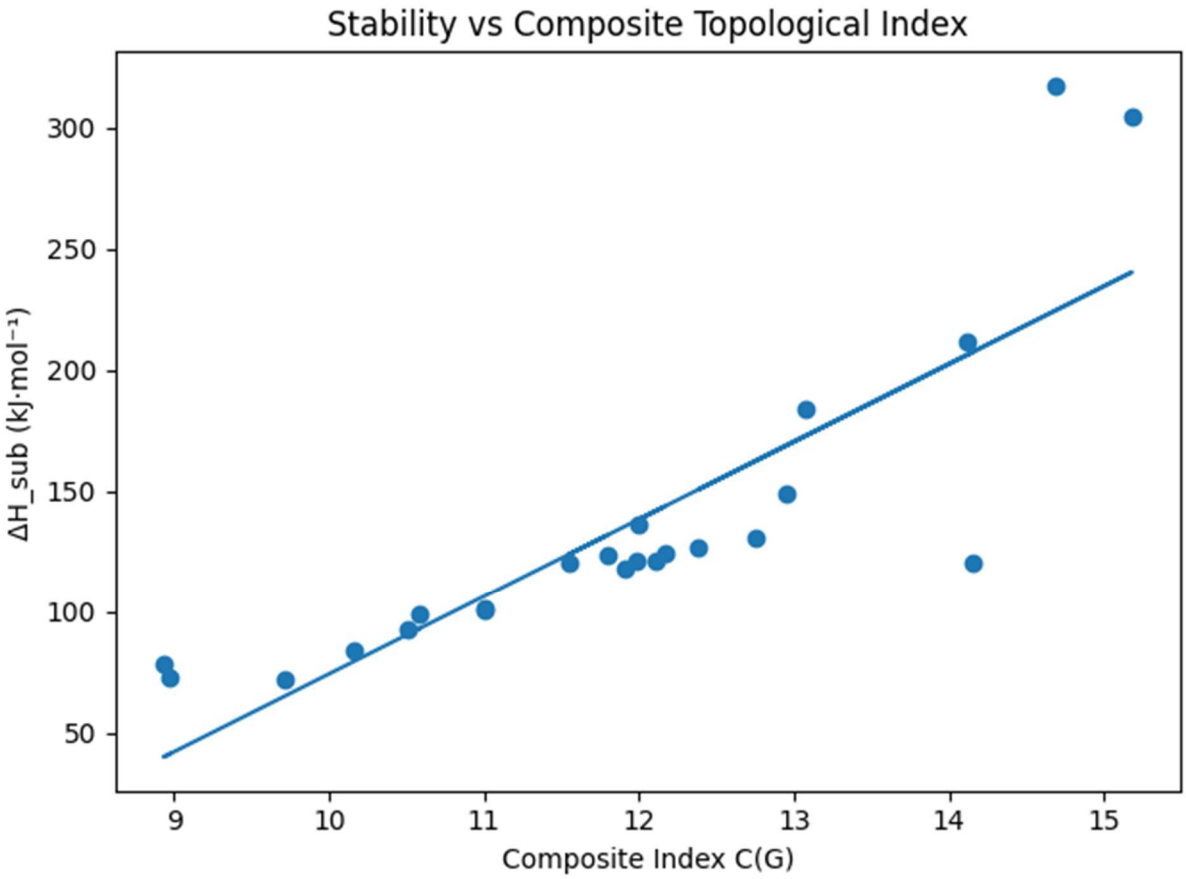

Figure 2. Relationship between the composite topological index C(G) and experimental enthalpy of sublimation (ΔH_sub / kJ mol-1) for the 23 polycyclic conjugated hydrocarbons. The solid line represents the ordinary least squares (OLS) regression fit ($R^2 = 0.708$), illustrating a strong positive monotonic association between structural compactness/connectivity and thermodynamic Stability.

Regression Performance

In this subsection, the predictive performance of the proposed models is evaluated for the selected experimental properties of the studied PCHs. The results of OLS and ridge regression are compared using standard goodness-of-fit and validation measures, including leave-one-out cross-validation (LOOCV). The numerical outcomes are presented in **Table 7** and are used to identify the most reliable modeling approach.

Table 7. Regression Performance

Model	Equation	R	R^2	p-value
Stability vs C(G)	Stability = 32.0655·C(G) − 246.423	0.8412	0.7077	4.95×10^{-7}
Reactivity vs C(G)	Reactivity = 6.7363×10^{-13}·C(G) + 9.2230×10^{-12}	0.8250	0.6807	1.27×10^{-6}
$\log_{10}$(Reactivity) vs C(G)	$\log_{10}$(Reactivity) = 0.01656·C(G) − 10.9618	0.8262	0.6826	1.19×10^{-6}

The regression statistics confirm that the composite index provides a reliable linear description of both Stability and reactivity trends (7). The corresponding structure–reactivity relationship is illustrated in **Figure 3**.

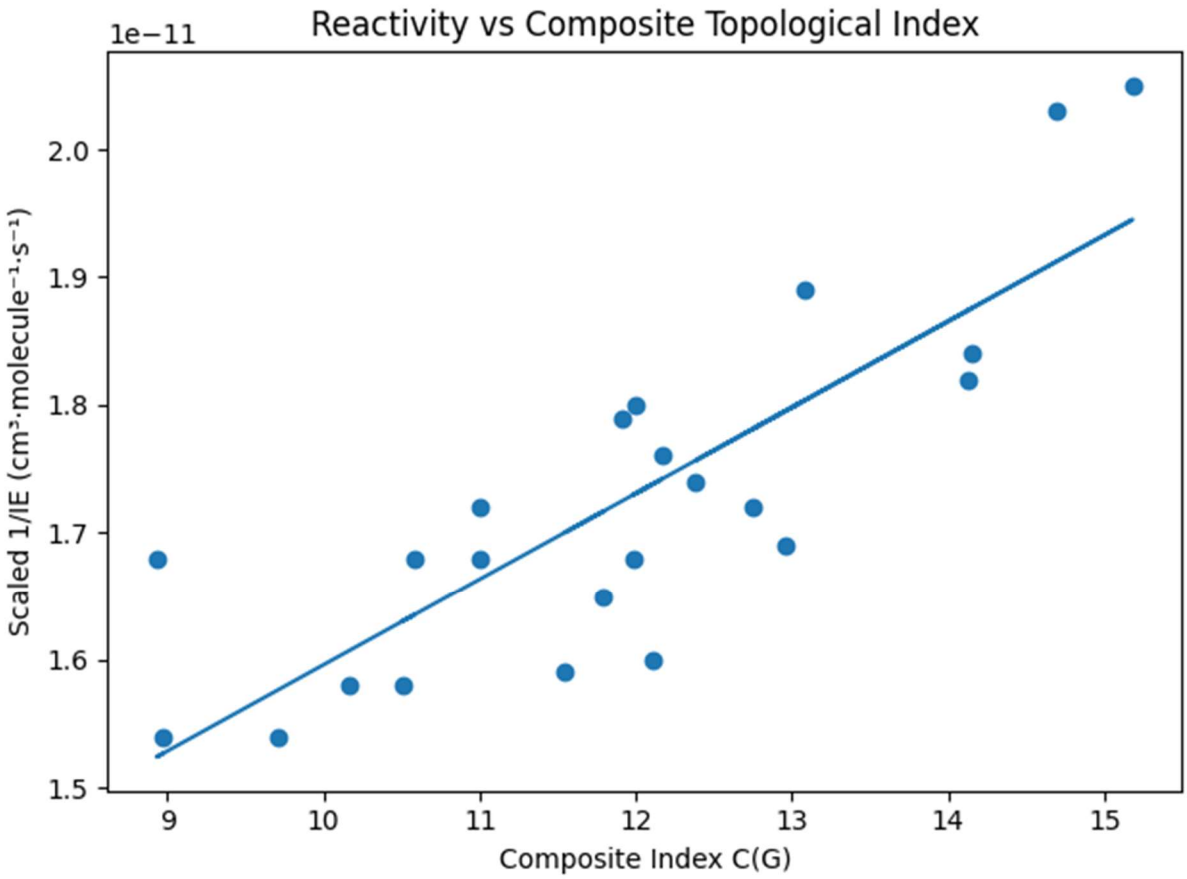

Figure 3. Relationship between the composite topological index C(G) and the scaled reciprocal ionization-energy proxy (scaled 1/IE) for the studied PCHs. The linear regression line ($R^2 = 0.681$) indicates a strong positive structure–reactivity trend captured by the composite descriptor.

Internal Validation by Leave-One-Out Cross-Validation

To further assess predictive robustness, multiple modeling strategies were compared under leave-one-out cross-validation (LOOCV), including the equal-weight composite baseline, ordinary least-squares (OLS) multi-descriptor regression, ridge-regularized modeling, and projection onto a ridge-derived optimized single-index axis (C(G)ridge=β·Z).

Stability (ΔH_{sub})

For Stability, the equal-weight composite baseline yielded moderate predictive performance (R2=0.708, Q2=0.594; RMSECV = 40.14 kJ·mol−1), indicating that simple aggregation of log-transformed descriptors captures meaningful structural trends but does not fully exploit descriptor complementarity. In contrast, the OLS multi-descriptor model achieved the strongest overall predictive performance ($R^2 = 0.907$, $Q^2 = 0.723$; $RMSE_{CV}$ = 33.16 kJ·mol−1), substantially improving upon the equal-weight baseline in both explained variance and cross-validated error.

Ridge regression (α=0.1; five standardized descriptors) also performed well (R2=0.838, Q2=0.690; RMSECV = 35.07 kJ·mol−1), demonstrating that regularization preserves strong predictive ability while offering a more stable coefficient framework under descriptor collinearity. Projection onto the ridge-derived one-axis composite, C(G)ridge=β·Z, produced closely comparable performance (R2=0.839, Q2=0.688; RMSECV = 35.19 kJ·mol−1),

showing that a reduced single-axis representation can retain much of the predictive information contained in the full multi-descriptor space.

Overall, the LOOCV results indicate that Stability trends in the present PCH dataset are captured most effectively by the full OLS multi-descriptor formulation, while ridge-based models provide competitive performance with improved interpretability and regularization control. Thus, although a ridge-derived optimized axis remains chemically meaningful, the strongest cross-validated Stability predictions in this dataset arise from the full descriptor combination rather than the single-index projection.

Table 8. LOOCV validation statistics for Stability models.

Model	R^2	Q^2 (LOOCV)	RMSECV ($kJ \cdot mol^{-1}$)	Interpretation
Equal-weight composite	0.708	0.594	40.14	Baseline composite performance
OLS multi-descriptor	0.907	0.723	33.16	Strong fit, but collinearity risk remains
Ridge (α = 0.1; five standardized descriptors)	0.838	0.690	35.07	Regularized multi-descriptor model
Single-index C(G)ridge (β·Z)	0.839	0.688	35.19	Ridge-derived one-axis composite

Reactivity (scaled 1/IE)

For reactivity, the equal-weight composite model yielded moderate predictive performance (R2=0.681, Q2=0.595; RMSECV = 8.586×10−13), indicating that simple descriptor aggregation captures a substantial portion of the reactivity trend (**Table IX**). In contrast to the Stability endpoint, the OLS multi-descriptor model showed improved apparent fit (R2=0.753) but weaker predictive generalization (Q2=0.552; RMSECV = 9.027×10−13), suggesting that multicollinearity and model variance reduce LOOCV robustness for the unregularized formulation.

Ridge regression (α=0.1) provided the best overall predictive performance for reactivity (R2=0.717, Q2=0.601; RMSECV = 8.526×10−13), indicating a modest but consistent benefit from regularization. The ridge-derived single-index projection (C(G)ridge) produced nearly identical performance (R^2=0.717, Q^2=0.600; RMSECV = 8.529×10−13), showing that most of the predictive information can be retained in a reduced one-axis representation.

Compared with the equal-weight baseline, the ridge model increased Q2 by 0.006 and slightly reduced RMSECV, confirming only a modest regularization advantage for this endpoint. PCA-based modeling (PC1) yielded intermediate performance (R2=0.671, Q2=0.583; RMSECV = 8.709×10−13), again indicating that variance-maximizing directions are not necessarily prediction-optimal.

Collectively, these results (**Table 9**) show that reactivity trends are predictively tractable

using composite graph descriptors, with ridge regularization offering the most robust LOOCV performance, while the ridge-derived single-index model provides a compact and nearly equivalent alternative.

Table 9. LOOCV validation statistics for reactivity models (scaled 1/IE)

Model	R^2	Q^2 (LOOCV)	RMSEcv
Equal-weight composite	0.681	0.595	8.586×10^{-13}
OLS multi-descriptor	0.753	0.552	9.027×10^{-13}
Ridge ($\alpha = 0.1$)	0.717	0.601	8.526×10^{-13}
Single-index C(G)_ridge	0.717	0.600	8.529×10^{-13}
PCA PC1	0.671	0.583	8.709×10^{-13}

Comparative Predictive Performance Across Modeling Strategies

To synthesize model performance across both endpoints, the best-performing approaches are compared against the equal-weight composite baseline. Improvements in cross-validated predictive ability are quantified using ΔQ^2 under LOOCV.

Table 10. Comparative Predictive Performance for Stability and Reactivity

Endpoint	Best Model	Best Q^2 (LOOCV)	Baseline Q^2	ΔQ^2 Improvement	Interpretation
Stability (ΔH_sub)	OLS multi-descriptor	0.723	0.594	+0.129	Best predictive performance on this dataset; collinearity risk remains despite strong LOOCV
Reactivity (proxy)	Ridge ($\alpha = 0.1$)	0.601	0.595	+0.006	Small but real regularization benefit; weak-to-moderate improvement over baseline composite

Degeneracy Analysis of Individual Descriptors and Composite Index

To evaluate structural discrimination capacity, degeneracy statistics were computed for each descriptor using a numerical tolerance of 10^{-3}. The number of distinct values and corresponding degeneracy ratios provide a quantitative measure of descriptor resolution across the examined polycyclic systems.

Table 11. Combined Degeneracy Comparison Table (Exact vs 3 d.p. vs 2 d.p.)

Descriptor	Distinct (Exact)	Degeneracy (Exact) (%)	Distinct (3 d.p.)	Degeneracy (3 d.p.) %	Distinct (2 d.p.)	Degeneracy (2 d.p.) (%)
$\log_{10}W$	23	0.000	23	0.000	20	0.130
$\log_{10}H$	23	0.000	21	0.087	14	0.391
$\log_{10}Z_1$	14	0.391	14	0.391	13	0.435
$\log_{10}R$	20	0.130	20	0.130	13	0.435
$\log_{10}MDP$	23	0.000	23	0.000	23	0.000
C(G)	23	0.000	23	0.000	23	0.000

Interpretation

Considering rounding to two decimal places, classical descriptors exhibit varying degrees of degeneracy across structurally distinct polycyclic systems. Specifically, degree-based descriptors such as $\log_{10}H$ (D = 0.391%), $\log_{10}Z_1$ (D = 0.435%), and $\log_{10}R$ (D = 0.435%) show noticeable value collapse, whereas $\log_{10}W$ exhibits only minimal degeneracy (D = 0.130%). In contrast, $\log_{10}MDP$ and the composite index C(G) retain complete discrimination (D = 0.000%), confirming that the composite formulation preserves structural uniqueness throughout the examined PCH dataset. These results highlight that, while individual classical descriptors may suffer from precision-dependent degeneracy, combining complementary invariants into a composite index effectively maintains robust structural discrimination across polycyclic chemical space.

Chemical Interpretation

The results show that the proposed composite topological descriptor has clear chemical meaning for polycyclic conjugated hydrocarbons. The strong positive relationships with stability and the scaled reciprocal ionization-energy reactivity proxy indicate that molecular topology plays an important role in controlling these properties.

Distance-based indices such as Wiener and Harary describe molecular size, compactness, and graph-distance distribution, while degree-based indices such as Randić and Zagreb represent local connectivity and branching. Together with the metric degree polynomial term, these descriptors capture important structural features such as ring fusion, molecular extension, and condensation pattern.

For stability, larger and more condensed polycyclic systems generally show higher enthalpy of sublimation due to stronger intermolecular interactions. For reactivity, extended π-conjugation can lower ionization energy, making electron removal easier and increasing the scaled reciprocal IE value.

Overall, the composite descriptor links molecular graph structure with stability and electronic reactivity trends. This confirms that the model is not only statistically useful but also chemically interpretable within the studied PCH series.

Comparison with Molecular Size Baseline

To examine whether the proposed composite descriptor provides information beyond simple molecular size, the predictive performance of C(G) was compared with basic size-based variables such as carbon count and ring count. Although stability and reactivity in polycyclic conjugated hydrocarbons naturally increase with molecular size to some extent, molecular size alone cannot fully distinguish isomeric or topologically different structures. For example, anthracene and phenanthrene have the same molecular formula but different fusion patterns, while tetracene, chrysene, and triphenylene contain the same number of carbon atoms but differ in linearity, angularity, and compactness.

The composite descriptor is therefore not only a size measure; it also incorporates graph-distance distribution, local degree connectivity, compactness, and degree–distance layering. These features allow it to distinguish molecules with similar molecular size but different topology. Thus, the observed correlation with stability and reactivity reflects both molecular extension and fusion topology rather than carbon count alone.

Limitations and future work

The present study is based on a moderate-sized but structurally diverse dataset of polycyclic conjugated hydrocarbons. Although the selected molecules cover linear, angular, compact, and highly condensed fused-ring systems, extension to a larger and more chemically varied dataset would further strengthen the general applicability of the proposed descriptor.

Leave-one-out cross-validation was used as an internal validation strategy because of the moderate dataset size. However, future studies should include independent external test sets to provide stronger predictive validation and to examine the transferability of the model beyond the present PCH series.

Another limitation is that the reactivity endpoint was represented using a scaled reciprocal ionization-energy proxy rather than a direct experimental kinetic reactivity measure. Therefore, the reactivity results should be interpreted as relative electronic reactivity trends rather than absolute reaction-rate predictions.

Future work may include testing alternative weighting schemes, nonlinear regression models, and broader graph classes. The descriptor may also be extended to additional experimental or quantum-chemical properties such as HOMO–LUMO gaps, electron affinity, aromaticity indices, and NICS values.

The observed strong correlation between the composite descriptor C(G) and both thermodynamic stability and electronic reactivity can be rationalized from a molecular topology

perspective. Each component of C(G) encodes a distinct physically meaningful aspect of π-conjugated polycyclic systems. The Wiener and Harary indices capture global π-electron delocalization through shortest-path and reciprocal-distance effects, which are directly related to molecular compactness and stabilization via extended conjugation. The Randić and Zagreb indices reflect local bond connectivity and branching patterns, which influence electron distribution uniformity and aromatic stabilization across fused rings.

The inclusion of the metric degree polynomial term further enhances sensitivity toward fusion density and higher-order connectivity interactions, particularly in highly condensed aromatic systems such as coronene and circumcoronene. These structural features are known to enhance π–π stacking interactions and electron delocalization, which are directly linked to increased thermodynamic stability (higher ΔH_{sub}) and modified ionization behavior.

The strong monotonic trends observed in Spearman correlation analysis ($\rho = 0.9073$ for stability and $\rho = 0.8283$ for reactivity) therefore reflect the ability of C(G) to simultaneously encode both global conjugation length and local structural heterogeneity. In particular, molecules with larger fused frameworks and higher ring condensation exhibit increased values of C(G), corresponding to enhanced electronic delocalization and higher cohesive energy, consistent with experimental thermodynamic trends.

These results confirm that the composite descriptor is not merely a statistical construct, but a chemically interpretable representation of polycyclic π-electron topology.

Conclusion

In this study, a novel composite topological descriptor C(G) was developed by integrating Wiener, Harary, Randić, Zagreb, and metric degree polynomial indices into a unified log-transformed framework for polycyclic conjugated hydrocarbons (PCHs). The proposed descriptor successfully captures both global molecular topology and local connectivity effects in fused aromatic systems.

Statistical evaluation demonstrated strong monotonic relationships between C(G) and experimental properties, with Spearman correlation coefficients of 0.9073 for thermodynamic stability and 0.8283 for electronic reactivity. Regression analysis further confirmed the predictive capability of the descriptor, with significant linear relationships observed for both endpoints.

Internal validation using leave-one-out cross-validation (LOOCV) showed that ridge-regularized models achieved the most stable predictive performance, while OLS models provided the highest explained variance for stability prediction. The results indicate that descriptor multicollinearity is effectively controlled through regularization, improving robustness without sacrificing interpretability.

Degeneracy analysis further demonstrated that the composite descriptor significantly reduces structural ambiguity compared to

individual topological indices, making it more suitable for distinguishing highly fused polycyclic systems.

Overall, the proposed framework provides a chemically interpretable, statistically robust, and computationally efficient approach for modeling structure–property relationships in polycyclic conjugated hydrocarbons. The methodology may be extended to other π-conjugated systems, including heteroaromatic compounds and nanographene derivatives.

Limitations of the study

Although the proposed composite descriptor demonstrates strong predictive performance, several limitations should be noted. First, the dataset size is relatively moderate (23 molecules), which may limit generalization to larger and more chemically diverse polycyclic systems. Second, the study relies on experimentally reported ionization energies and sublimation enthalpies, which may contain measurement inconsistencies across different sources. Third, the current framework assumes equal structural weighting in the baseline descriptor construction, which may not fully reflect all chemical environments in extreme cases of molecular distortion or heteroatom substitution.

Future work should extend the descriptor framework to larger datasets, include density functional theory (DFT)-derived properties, and explore machine learning integration for nonlinear structure–property mapping.

References

[1] Gutman I. Degree-Based Topological Indices. Croatica Chemica Acta 2013;86(4):351-361.

[2] Randić M. Aromaticity of Polycyclic Conjugated Hydrocarbons. Chemical Reviews 2003;103(9):3449-3606.

[3] Gao W, Younas M, Farooq A, Mahboob A, Nazeer W. M-Polynomials and Degree-Based Topological Indices of the Crystallographic Structure of Molecules. Biomolecules 2018;8(4):107.

[4] Wiener H. Structural Determination of Paraffin Boiling Points. Journal of the American Chemical Society 1947;69(1):17-20.

[5] Randic M. Characterization of molecular branching. Journal of the American Chemical Society 1975;97(23):6609-6615.

[6] Gutman I, Trinajstić N. Graph theory and molecular orbitals. Total φ-electron energy of alternant hydrocarbons. Chemical Physics Letters 1972;17(4):535-538.

[7] Furtula B, Gutman I. A forgotten topological index. Journal of Mathematical Chemistry 2015;53(4):1184-1190.

[8] Bouarra N, Kherouf S, Messadi D. QSPR-based prediction model for the melting point of polycyclic aromatic hydrocarbons using MLR and ANN methods. International Journal of Chemistry and Technology 2024;8(2):128-136.

[9] Redzepovic I. Chemical applicability of Sombor indices. Journal of the Serbian Chemical Society 2021;86(5):445-457.

[10] Malloci G, Joblin C, Mulas G. On-line database of the spectral properties of polycyclic aromatic hydrocarbons. Chemical Physics 2007;332(2-3):353-359.

[11] Li J, Zhao T, Yang Q, Du S, Xu L. A review of quantitative structure-activity relationship: The development and current status of data sets, molecular descriptors and mathematical models. Chemometrics and Intelligent Laboratory Systems 2025;256:105278.

[12] Zhang Y, Fan L, Su C, Shu Z, Zhang H. Application of machine learning in developing a quantitative structure–property relationship model for

predicting the thermal decomposition temperature of nitrogen-rich energetic ionic salts. RSC Advances 2024;14(51):37737-37751.

[13] Chickos J, Acree W. Enthalpies of Sublimation of Organic and Organometallic Compounds. 1910–2001. Journal of Physical and Chemical Reference Data 2002;31(2):537-698.

[14] Al-Saadi HM, Al-Janabi AA. QSPR modelling using topological descriptors. Natl. J. Chem. 2019;73:45–56.

[15] Abbas SH, Al-Mousawi MA. Molecular descriptors in aromatic systems. Natl. J. Chem. 2020;77:112–123.

[16] Ibrahim RK, Al-Ani AH. Topological indices and chemical reactivity prediction. Natl. J. Chem. 2021;81:89–98.

[17] Estrada E, Hatano N. Resistance Distance, Information Centrality, Node Vulnerability and Vibrations in Complex Networks. Network Science 2010:13-29.

[18] Pavlic B, Teslic N, Kojic P, Pezo L. Prediction of the GC-MS retention time for terpenoids detected in sage (Salvia officinalis L.) essential oil using QSRR approach. Journal of the Serbian Chemical Society 2020;85(1):9-23.

[19] Albert R, Barabási A. Statistical mechanics of complex networks. Reviews of Modern Physics 2002;74(1):47-97.

[20] Bakaxonov A, Gaziev L. Electrochemical Determination of Carboplatin. French-Ukrainian Journal of Chemistry 2025;13(2):15-23.

[21] Pavlova V, Zadorozhnii P, Kiselev V, Okhtina O, Kharchenko A. Synthesis and spectral characteristics of N-(2,2,2-trichloro-1-((5-(methylthio)-1,3,4-thiadiazol-2-yl)amino)ethyl)carboxamides. French-Ukrainian Journal of Chemistry 2025;13(2):51-57.

[22] İslamoğlu F. Comprehensive Computational Profiling of Some New Candidate Molecules for Huntington's Disease: Drug-Likeness, Quantum Descriptors, Molecular Docking, Molecular Dynamics Simulation and Synthetic Mechanism. French-Ukrainian Journal of Chemistry 2025;13(2):31-50.