Journal of Saudi Chemical Society (2026) 30:50
https://doi.org/10.1007/s44442-026-00105-6

RESEARCH

Advancing QSPR with sum-connectivity descriptors: physicochemical and antibacterial modelling via molecular graph connectivity

Azzam Altairi[1,3] · **Mohammed Alsharafi**[2,3] · **Zaied Alhaj**[1,3]

Received: 27 January 2026 / Accepted: 29 June 2026
© The Author(s) 2026

Abstract

Quantitative structure-activity and structure-property relationship (QSAR/QSPR) modelling uses molecular descriptors to relate chemical structure to physicochemical behaviour and biological activity. In this study, we propose a new family of descriptors, termed *sum-connectivity descriptor*, obtained by multiplying a classical edge-degree kernel $\Phi(d_u, d_v)$ by the normalization factor $(d_u + d_v)^{-1/2}$. This construction generates a systematic family of descriptors induced from Zagreb-type, Sombor-type, Albertson, arithmetic-geometric, geometric-arithmetic, Forgotten, and inverse Nirmala indices. From a theoretical perspective, we establish comparison relations with the corresponding classical descriptors, degree-extremal bounds, and regular-graph proportionality results. The empirical study uses a curated antibacterial *E. coli* dataset from ChEMBL containing 6,657 unique compounds, with MIC values standardized to μM and transformed to $pMIC$. In the QSPR benchmark, the sum-connectivity indices consistently outperformed the classical indices, increasing the mean R^2 from 0.6951 to 0.7015. Moreover, the best QSPR performance was obtained by the **All Combined** descriptor set, which integrates indices, RDKit descriptors, and physicochemical properties, with properties-specific R^2 values ranging from 0.9349 to 0.9997. For antibacterial activity, the best model is **ExtraTrees+RDKit Descriptors**, with $R^2 = 0.5959 \pm 0.0246$ under repeated cross-validation and $R^2 = 0.4799 \pm 0.0252$ under Murcko scaffold validation. The proposed sum-connectivity descriptors are therefore most strongly supported as systematically improved graph-topological descriptors in matched QSPR comparisons, while their added value for the more heterogeneous $pMIC$ endpoint is complementary and modest.

Keywords Sum-connectivity descriptors · Molecular graph theory · Degree-based descriptors · Extremal bounds · QSPR/QSAR · Antibacterial activity · E. coli

1 Introduction

Quantitative structure-activity/property relationship (QSAR/QSPR) modeling aims to explain and predict biological activity and physicochemical properties of compounds using numerical descriptors derived from molecular structure.

Originally developed in medicinal chemistry to support drug discovery and lead optimization, QSAR/QSPR methodologies are now applied in pharmaceutical, environmental, materials, agrochemical, and regulatory settings [1–3]. Recent QSAR studies also combine descriptor-based modelling with ADMET, docking, and molecular-dynamics analyses in anticancer drug-discovery contexts [4].

A common and successful paradigm is to represent a molecule by a simple graph whose vertices correspond to atoms and whose edges correspond to chemical bonds,

✉ Azzam Altairi
azzo.tairi@gmail.com

Mohammed Alsharafi
alsharafi205010@gmail.com

Zaied Alhaj
zaiedalhaj@gmail.com

[1] Department of Biomedical Engineering, Institute of Graduate Studies, Istanbul University–Cerrahpasa, Istanbul, Türkiye

[2] Department of Mathematics, Faculty of Arts and Science, Yildiz Technical University, Istanbul 34220, Türkiye

[3] Department of Biomedical Engineering, Faculty of Engineering, University of Science and Technology, Aden, Yemen

Published online: 08 July 2026

and then to construct topological descriptors that capture branching, local environment, and global connectivity patterns. Connectivity-based descriptors play a prominent role in structure-activity analysis, dating back to the molecular connectivity framework of Kier and Hall [5] and the seminal branching descriptor introduced by Randić [6]. Comprehensive perspectives on chemical graph theory and its relevance to QSAR/QSPR can be found in the survey literature [7–9].

In chemical graph theory, *topological indices* are graph invariants designed to encode structural information in a form amenable to statistical learning and regression. Among the many available families, *degree-based* indices (defined via edge contributions that depend on endpoint degrees) have attracted sustained interest due to their interpretability and computational simplicity; see Gutman [10] and standard references on graph theory and chemical graph theory for background [8, 11]. Typical examples include Zagreb-type descriptors, arithmetic-geometric and geometric-arithmetic constructions, Sombor-type measures, Albertson irregularity, and Nirmala-type forms. In this paper, we adopt a unified viewpoint in which each such descriptor is defined by an edge-symmetric contribution function $\Phi(d_u, d_v)$, leading to the general degree-based index

$$I_\Phi(G) = \sum_{uv \in E(G)} \Phi(d_u, d_v),$$

with the specific instances considered here summarized in Table 1. Recent work has further developed degree-based descriptors under graph operations and complements, including complement-graph formulations [22, 23]. Applications to nanotube and nanotorus structures have also been reported for related degree-based indices [24, 25].

Connectivity indices constitute a central bridge between graph structure and chemical behavior. Beyond the Randić index and its extensive developments [6, 26, 27], Zhou and Trinajstić introduced and investigated the *sum-connectivity* paradigm and its general forms, initiating a large body of work on extremal graphs, bounds, and structural characterization [19, 28]. Subsequent contributions refined these ideas for trees and unicyclic graphs [29, 30]. Related extremal results involving pendant vertices, independence number, and comparisons with product-connectivity measures were also established in [31–33]. Extremal and inequality-based approaches for degree-based descriptors are also closely connected with the general theory of weighted graph invariants and optimization over graph classes [34].

Recent studies have broadened degree-based molecular descriptors across structurally diverse chemical graphs,

Table 1 Degree-based topological indices(classical indices)

No.	Name	Short name	$\Phi(d_u, d_v)$
1	Redefined first Zagreb index [12]	$ReZG_1$	$\frac{d_u + d_v}{d_u d_v}$
2	Geometric-Arithmetic F-index [13]	GAF	$2\frac{d_u d_v}{d_u^2 + d_v^2}$
3	Redefined second Zagreb index [12]	$ReZG_2$	$\frac{d_u d_v}{d_u + d_v}$
4	Elliptic Sombor index [14]	ESO	$(d_u + d_v)\sqrt{d_u^2 + d_v^2}$
5	Redefined third Zagreb index [12]	$ReZG_3$	$(d_u d_v)(d_u + d_v)$
6	Euler Sombor index [15]	EU	$\sqrt{d_u^2 + d_v^2 + d_u d_v}$
7	Albertson index [16]	A	$\lvert d_u - d_v \rvert$
8	Banhatti-Sombor index [17]	BSO	$\left(\frac{1}{d_u^2} + \frac{1}{d_v^2}\right)^{1/2}$
9	Arithmetic-geometric index [18]	AG	$\frac{d_u + d_v}{2\sqrt{d_u d_v}}$
10	Sum-connectivity index [19]	SC	$(d_u + d_v)^{-1/2}$
11	Forgotten index (F-index) [20]	F	$d_u^2 + d_v^2$
12	First inverse Nirmala index [21]	IN_1	$\left(\frac{1}{d_u} + \frac{1}{d_v}\right)^{1/2}$
13	Geometric-arithmetic index [18]	GA	$\frac{2\sqrt{d_u d_v}}{d_u + d_v}$
14	Second inverse Nirmala index [21]	IN_2	$\left(\frac{1}{d_u} + \frac{1}{d_v}\right)^{-1/2}$

including benzenoid systems, graphene-family structures, and dendrimer-related molecular architectures [35–37]. Related computations for coronoid and sodalite systems further illustrate the usefulness of degree- and neighbourhood-sum-based molecular descriptors [38, 39]. This context motivates the present unified sum-connectivity for classical degree-based indices rather than an isolated descriptor construction.

Motivated by both mathematical structure and descriptor design in QSAR/QSPR, we propose a systematic construction that *reweights* classical degree-based indices by the ordinary sum-connectivity edge weight $w_{SC}(u,v) = (d_u + d_v)^{-1/2}$. Concretely, for each degree-based index generated by $\Phi(d_u, d_v)$, we define its *sum-connectivity* analogue by

$$SC_\Phi(G) = \sum_{uv \in E(G)} \Phi(d_u, d_v)\,(d_u + d_v)^{-1/2}.$$

This mechanism yields a family of new descriptors that preserve the interpretability of the base index while introducing an explicit connectivity-based normalization through $(d_u + d_v)^{-1/2}$. The resulting descriptors-constructed from the well-known indices listed in Table 1-are collected in Table 2. The approach also aligns with earlier connectivity-modified and atom-bond variants of classical indices [40–42], as well as review-level developments on atom-bond connectivity bounds and extremal results [43].

From a methodological standpoint, it is important to understand how the newly defined sum-connectivity indices relate to their unweighted counterparts. Accordingly, we establish comparison inequalities linking $SC_\Phi(G)$ and $I_\Phi(G)$ in terms of the degree extremals $\delta(G)$ and $\Delta(G)$, and we identify graph classes where exact proportionality emerges (e.g., regular graphs). We further provide min-max bounds for representative indices in Table 2, yielding explicit estimates in terms of $m = |E(G)|$, $\delta(G)$, and $\Delta(G)$, as well as optional n-based forms via the handshaking lemma.

Moreover, this study leverages a carefully curated dataset comprising 6,657 unique small-molecule compounds with experimentally determined minimum inhibitory concentration (MIC) values against *Escherichia coli*, extracted from the ChEMBL bioactivity database, to develop predictive QSAR models of antibacterial potency. MIC measurements

Table 2 Sum-connectivity descriptors induced from the corresponding edge-degree kernels

No.	Name	Short name	$\Phi(d_u, d_v)\, w_{SC}(u,v)$
1	SC-redefined first Zagreb index	$SCReZG_1$	$\frac{d_u + d_v}{d_u d_v \sqrt{d_u + d_v}}$
2	SC-geometric-arithmetic F-index	$SCGAF$	$2\,\frac{d_u d_v}{(d_u^2 + d_v^2)\sqrt{d_u + d_v}}$
3	SC-redefined second Zagreb index	$SCReZG_2$	$\frac{d_u d_v}{(d_u + d_v)^{3/2}}$
4	SC-elliptic Sombor index	$SCESO$	$\sqrt{d_u + d_v}\,\sqrt{d_u^2 + d_v^2}$
5	SC-redefined third Zagreb index	$SCReZG_3$	$d_u d_v \sqrt{d_u + d_v}$
6	SC-Euler Sombor index	$SCEU$	$\frac{\sqrt{d_u^2 + d_v^2 + d_u d_v}}{\sqrt{d_u + d_v}}$
7	SC-Albertson index	SCA	$\frac{\lvert d_u - d_v\rvert}{\sqrt{d_u + d_v}}$
8	SC-Banhatti-Sombor index	$SCBSO$	$\left(\frac{1}{d_u^2} + \frac{1}{d_v^2}\right)^{1/2} \frac{1}{\sqrt{d_u + d_v}}$
9	SC-arithmetic-geometric index	$SCAG$	$\frac{d_u + d_v}{2\sqrt{d_u d_v}\sqrt{d_u + d_v}}$
10	Hyper sum-connectivity index	HSC	$(d_u + d_v)^{-1}$
11	SC-forgotten index	SCF	$\frac{d_u^2 + d_v^2}{\sqrt{d_u + d_v}}$
12	SC-first inverse Nirmala index	$SCIN_1$	$\left(\frac{1}{d_u} + \frac{1}{d_v}\right)^{1/2} \frac{1}{\sqrt{d_u + d_v}}$
13	SC-geometric-arithmetic index	$SCGA$	$\frac{2\sqrt{d_u d_v}}{(d_u + d_v)\sqrt{d_u + d_v}}$
14	SC-second inverse Nirmala index	$SCIN_2$	$\left(\frac{1}{d_u} + \frac{1}{d_v}\right)^{-1/2} \frac{1}{\sqrt{d_u + d_v}}$

represent a cornerstone of antibacterial susceptibility testing and antimicrobial drug development, providing a quantitative metric that bridges early-stage in vitro screening, preclinical optimization, and clinical breakpoint determination [44, 45]. Defined as the lowest concentration of an antimicrobial agent required to inhibit visible bacterial growth under standardized in vitro conditions, MIC remains the gold-standard measure for quantifying antibacterial activity [44]. Accordingly, MIC values underpin clinical susceptibility testing, inform breakpoint definitions established by regulatory and standards-setting bodies such as the Clinical and Laboratory Standards Institute (CLSI) and the European Committee on Antimicrobial Susceptibility Testing (EUCAST), and serve as critical inputs for pharmacokinetic/pharmacodynamic (PK/PD) target derivation and dosing optimization [46].

Despite its central role, substantial evidence indicates that MIC values for the same compound may vary considerably with experimental conditions, including assay methodology, bacterial strain, and molecular target [47]. To reduce experimental noise and enhance biological consistency, the present study exclusively considers compounds with reported MIC values against *E. coli*. This organism is of particular clinical and public-health relevance due to its dual role as a commensal member of the human microbiota and an opportunistic pathogen. *E. coli* is a leading cause of urinary tract infections, bloodstream infections, intra-abdominal sepsis, and neonatal meningitis, and it is increasingly associated with multidrug resistance worldwide [48]. Consequently, the development of robust and generalizable QSAR models for predicting antibacterial activity against *E. coli* is of significant importance for accelerating antimicrobial discovery and optimization.

To demonstrate practical relevance in cheminformatics, we include an application-oriented section in which the proposed sum-connectivity descriptors are employed as candidate features in QSPR, aimed at predicting selected physicochemical properties, and QSAR, focused on the base-10 negative logarithm of the experimentally measured MIC (pMIC) prediction (see Sect. 4). This application is consistent with the established role of connectivity descriptors in structure-activity analysis [5, 7] and assesses whether the additional connectivity moderation inherent in Table 2 can improve predictive performance relative to the classical degree-based descriptors in Table 1.

1.1 Application-oriented motivation and QSAR/QSPR case study

Beyond the preceding structural motivation, the proposed normalization is also motivated by descriptor design in molecular modelling: edge contributions associated with high-degree atoms (or highly connected substructures) can dominate classical sums, which may reduce robustness across chemically heterogeneous datasets. The factor $(d_u + d_v)^{-1/2}$ provides a principled connectivity-dependent moderation that preserves the underlying Φ-contribution while reducing sensitivity to degree extremes. In Sect. 4, we evaluate this motivation empirically on a curated antibacterial *E. coli* dataset using a more rigorous validation framework. First, we conduct a controlled indices-only QSPR benchmark that compares the 14 classical degree-based indices with their 14 sum-connectivity analogues across eight physicochemical endpoints under repeated cross-validation. Second, we extend the analysis to broader QSPR/QSAR experiments by combining these graph-based descriptors with RDKit features and assessing performance under both repeated random cross-validation and, for antibacterial activity prediction, Murcko scaffold-based validation. This design allows us to examine both the standalone utility of the proposed descriptors and their contribution within richer cheminformatics feature sets. The resulting analysis shows that sum-connectivity reweighting consistently improves upon the corresponding classical indices in the QSPR setting, while for the more challenging QSAR task the proposed descriptors remain useful complementary topological features alongside stronger RDKit-based representations.

1.2 Organization of the paper

Section 2 recalls the required definitions and compiles the degree-based indices together with their sum-connectivity analogues. Section 3 develops comparison relations, degree-extremal bounds, and special cases (including regular graphs and complement-graph transformations). Section 4 presents the application-oriented QSAR/QSPR case study on an antibacterial *E. coli* dataset and discusses the empirical implications of the proposed descriptors, including an indices-only benchmark (classical vs. sum-connectivity families) and complementary experiments using combined descriptor pools.

2 Properties of sum-connectivity descriptors

Throughout this section, let $G = (V, E)$ be a simple, finite, connected, undirected graph. For each vertex $u \in V(G)$, let $d(u) = d_G(u)$ denote its degree. For brevity, in tabular formulas and edge kernels we also write $d_u := d(u)$ and $d_v := d(v)$.

A broad class of edge-degree-based indices can be written as

$$I_\Phi(G) = \sum_{uv\in E(G)} \Phi(d(u), d(v)),$$

where Φ is symmetric in (d_u, d_v).

The general sum-connectivity index is

$$\chi_\alpha(G) = \sum_{uv\in E(G)} \big(d(u) + d(v)\big)^\alpha.$$

In particular, the ordinary sum-connectivity weight is

$$w_{SC}(u, v) = \frac{1}{\sqrt{d(u) + d(v)}} = \big(d(u) + d(v)\big)^{-1/2}.$$

This weighting leads naturally to the corresponding sum-connectivity descriptor defined below.

Definition 1 Let $\Phi : \mathbb{N}_{\geq 1} \times \mathbb{N}_{\geq 1} \to \mathbb{R}$ be a symmetric function. The sum-connectivity descriptor induced by Φ is defined by

$$SC_\Phi(G) := I^{SC}_\Phi(G) = \sum_{uv\in E(G)} \Phi(d_u, d_v)\, w_{SC}(u, v).$$

Remark 2.1 The ordinary sum-connectivity index satisfies $SC(G) = \chi_{-1/2}(G)$ and is recovered in this framework by choosing $\Phi(d_u, d_v) = 1$. Moreover, choosing $\Phi(d_u, d_v) = (d_u + d_v)^{-1/2}$ yields

$$SC_\Phi(G) = \chi_{-1}(G),$$

that is, the hyper sum-connectivity index $HSC(G)$.

2.1 Mathematical motivation for sum-connectivity descriptors

Many widely used degree-based descriptors admit an edge-decomposable representation

$$I_\Phi(G) = \sum_{uv\in E(G)} \Phi(d(u), d(v)),$$

where Φ is symmetric. Although this framework is flexible, for several choices of Φ the resulting index may be strongly influenced by edges incident to high-degree vertices. From a graph-theoretic viewpoint, it is therefore natural to introduce a degree-dependent normalization that moderates such dominance while preserving the local information encoded by Φ. In this sense, the present contribution is primarily methodological: rather than proposing an isolated descriptor, we formulate a unified transformation $\Phi \mapsto SC_\Phi$ that systematically generates a family of sum-connectivity descriptors from existing edge-degree kernels.

A classical normalization mechanism is provided by the general sum-connectivity family

$$\chi_\alpha(G) = \sum_{uv\in E(G)} \big(d(u) + d(v)\big)^\alpha,$$

introduced and studied extensively in the literature; see, e.g. [19, 28, 29]. In particular, the choice $\alpha = -\frac{1}{2}$ yields the standard sum-connectivity weight

$$w_{SC}(u, v) = \big(d(u) + d(v)\big)^{-1/2},$$

which is positive and decreases with $d(u) + d(v)$. Motivated by this, we associate with each symmetric kernel Φ the descriptor

$$SC_\Phi(G) = \sum_{uv\in E(G)} \Phi(d(u), d(v))\, w_{SC}(u, v) = \sum_{uv\in E(G)} \frac{\Phi(d(u), d(v))}{\sqrt{d(u) + d(v)}}.$$

This construction has several useful structural consequences. First, since

$$2\delta(G) \leq d(u) + d(v) \leq 2\Delta(G) \quad \text{for all } uv \in E(G),$$

the weight satisfies

$$\frac{1}{\sqrt{2\Delta(G)}} \leq w_{SC}(u, v) \leq \frac{1}{\sqrt{2\delta(G)}},$$

and therefore, whenever $\Phi(d(u), d(v)) \geq 0$ on $E(G)$,

$$\frac{1}{\sqrt{2\Delta(G)}}\, I_\Phi(G) \leq SC_\Phi(G) \leq \frac{1}{\sqrt{2\delta(G)}}\, I_\Phi(G).$$

Thus, $SC_\Phi(G)$ remains comparable in scale to $I_\Phi(G)$ with explicit degree-extremal control. Second, if G is r-regular, then $d(u) + d(v) = 2r$ for every edge, and hence

$$SC_\Phi(G) = \frac{1}{\sqrt{2r}}\, I_\Phi(G),$$

so on regular graphs the new descriptors reduce to a constant multiple of the corresponding classical ones.

Finally, the mapping $\Phi \mapsto SC_\Phi$ is linear and preserves edge-locality and symmetry. It therefore provides a systematic way to generate new descriptors by combining established degree-based contributions (Table 1) with the sum-connectivity normalization, thereby yielding the descriptors listed in Table 2. The present subsection establishes the graph-theoretic rationale and structural

consequences of this transformation, while the empirical chemical assessment is treated separately in Sect. 4.

3 Relations between degree-based indices and their sum-connectivity counterparts

Using the notation introduced in Sect. 2, we now compare a degree-based index $I_\Phi(G)$ with its associated sum-connectivity counterpart $SC_\Phi(G)$. In particular, $\delta(G)$ and $\Delta(G)$ denote the minimum and maximum degrees of G, respectively, and

$$w_{SC}(u,v) = \frac{1}{\sqrt{d(u)+d(v)}}.$$

Thus, $SC_\Phi(G)$ is obtained from $I_\Phi(G)$ by reweighting each edge contribution through the standard sum-connectivity factor.

The comparison result shows that the proposed reweighting preserves the scale of the original descriptor up to explicit degree-extremal factors, which provides a basic structural justification for the framework.

3.1 A basic comparison inequality

Lemma 3.1 Assume that $\Phi(d(u), d(v)) \geq 0$ for every edge $uv \in E(G)$. Then

$$\frac{1}{\sqrt{2\Delta(G)}} I_\Phi(G) \leq SC_\Phi(G) \leq \frac{1}{\sqrt{2\delta(G)}} I_\Phi(G).$$

Moreover, if G is r-regular, then

$$SC_\Phi(G) = \frac{1}{\sqrt{2r}} I_\Phi(G).$$

Proof For every edge $uv \in E(G)$, one has

$$2\delta(G) \leq d(u) + d(v) \leq 2\Delta(G),$$

and hence

$$\frac{1}{\sqrt{2\Delta(G)}} \leq w_{SC}(u,v) \leq \frac{1}{\sqrt{2\delta(G)}}.$$

Since $\Phi(d(u), d(v)) \geq 0$ on $E(G)$, multiplying the above inequality by $\Phi(d(u), d(v))$ and summing over all edges yields

$$\frac{1}{\sqrt{2\Delta(G)}} I_\Phi(G) \leq SC_\Phi(G) \leq \frac{1}{\sqrt{2\delta(G)}} I_\Phi(G).$$

If G is r-regular, then $d(u) + d(v) = 2r$ for every edge $uv \in E(G)$, so $w_{SC}(u,v) = 1/\sqrt{2r}$. Substituting this into the definition of $SC_\Phi(G)$ gives

$$SC_\Phi(G) = \frac{1}{\sqrt{2r}} I_\Phi(G).$$

□

3.2 Correspondence between classical degree-based indices and their sum-connectivity descriptors

Let $uv \in E(G)$ and abbreviate $d_u = d_G(u)$ and $d_v = d_G(v)$. This correspondence makes explicit that the proposed family is generated by a single reweighting principle applied uniformly to the underlying edge-degree kernels. Each degree-based index in Table 1 is of the form

$$I_\Phi(G) = \sum_{uv \in E(G)} \Phi(d_u, d_v),$$

for the corresponding edge-contribution function $\Phi(d_u, d_v)$ listed there. Its sum-connectivity descriptor in Table 2 is defined by

$$SC_\Phi(G) = \sum_{uv \in E(G)} \Phi(d_u, d_v)\, w_{SC}(u,v), \quad w_{SC}(u,v) = (d_u + d_v)^{-1/2}.$$

Thus, each pair consisting of a classical degree-based index and its sum-connectivity descriptor in Tables 1 and 2 is generated by the same kernel Φ. This observation avoids repeating separate definitions for each descriptor and highlights the uniform structure of the construction.

Special note (classical $SC(G)$ versus hyper $HSC(G)$). The classical sum-connectivity index is obtained by choosing

$$\Phi_{SC}(d_u, d_v) = (d_u + d_v)^{-1/2},$$

so that

$$SC(G) = I_{\Phi_{SC}}(G) = \sum_{uv \in E(G)} (d_u + d_v)^{-1/2}.$$

Applying the present reweighting once more to the same kernel Φ_{SC} gives

$$SC_{\Phi_{SC}}(G) = \sum_{uv\in E(G)} (d_u + d_v)^{-1/2}\, w_{SC}(u,v) = \sum_{uv\in E(G)} \frac{1}{d_u + d_v}.$$

We denote this induced descriptor by

$$HSC(G) := \sum_{uv\in E(G)} \frac{1}{d_u + d_v},$$

and refer to it as the hyper sum-connectivity index.

3.3 Comparison bounds via degree extremals

As an immediate consequence of Lemma 3.1, whenever $\Phi(d_u, d_v) \geq 0$ for all $uv \in E(G)$, one has

$$\frac{1}{\sqrt{2\Delta(G)}}\, I_\Phi(G) \leq SC_\Phi(G) \leq \frac{1}{\sqrt{2\delta(G)}}\, I_\Phi(G).$$

In particular, these bounds apply to every paired descriptor in Tables 1 and 2, including (SC, HSC).

3.4 Regular graphs (Exact proportionality)

The r-regular case is already contained in Lemma 3.1, which yields the exact proportionality

$$SC_\Phi(G) = \frac{1}{\sqrt{2r}}\, I_\Phi(G).$$

Thus, on regular graphs, the proposed reweighting preserves the underlying descriptor up to a uniform constant factor.

Complement relations are included because many degree-based descriptors are naturally examined under graph complementation, and in the present setting they show how the sum-connectivity reweighting behaves under the degree transformation $d_{\overline{G}}(u) = n - 1 - d_G(u)$.

3.5 Complement graph relations

Let G have order $n = |V(G)|$ and size $m = |E(G)|$, and let $\overline{G}$ be its complement. Then

$$d_{\overline{G}}(u) = n - 1 - d_G(u), \quad |E(\overline{G})| = \binom{n}{2} - m.$$

Lemma 3.2 For any distinct vertices $u, v \in V(G)$,

$$d_{\overline{G}}(u) + d_{\overline{G}}(v) = 2(n-1) - \big(d_G(u) + d_G(v)\big).$$

Consequently, for every edge $uv \in E(\overline{G})$,

$$w^{\overline{G}}_{SC}(u,v) = \frac{1}{\sqrt{d_{\overline{G}}(u) + d_{\overline{G}}(v)}} = \frac{1}{\sqrt{2(n-1) - \big(d_G(u) + d_G(v)\big)}}.$$

Lemma 3.3 For any symmetric bivariate function Φ,

$$\begin{aligned} I_\Phi(\bar{G}) &= \sum_{uv\in E(\bar{G})} \Phi(d_{\bar{G}}(u), d_{\bar{G}}(v)) \\ &= \sum_{\substack{u<v \\ uv\notin E(G)}} \Phi(n-1-d_G(u), n-1-d_G(v)), \end{aligned}$$

and

$$\begin{aligned} SC_\Phi(\bar{G}) &= \sum_{uv\in E(\bar{G})} \Phi(d_{\bar{G}}(u), d_{\bar{G}}(v)) w^{\bar{G}}_{SC}(u,v) \\ &= \sum_{\substack{u<v \\ uv\notin E(G)}} \frac{\Phi(n-1-d_G(u), n-1-d_G(v))}{\sqrt{2(n-1) - (d_G(u) + d_G(v))}}. \end{aligned}$$

Remark 3.4 The minimum and maximum degrees satisfy

$$\delta(\overline{G}) = n - 1 - \Delta(G), \quad \Delta(\overline{G}) = n - 1 - \delta(G).$$

If $\overline{G}$ has no isolated vertices, equivalently if $\delta(\overline{G}) \geq 1$ (or $\Delta(G) \leq n-2$), and if $\Phi\big(d_{\overline{G}}(u), d_{\overline{G}}(v)\big) \geq 0$ for all $uv \in E(\overline{G})$, then

$$\frac{1}{\sqrt{2\Delta(\overline{G})}}\, I_\Phi(\overline{G}) \leq SC_\Phi(\overline{G}) \leq \frac{1}{\sqrt{2\delta(\overline{G})}}\, I_\Phi(\overline{G}).$$

The lower bound remains meaningful without this extra assumption, whereas the stated upper bound requires $\delta(\overline{G}) \geq 1$.

Corollary 3.5 If G is r-regular, then $\overline{G}$ is $(n-1-r)$-regular. Moreover,

$$SC_\Phi(G) = \frac{1}{\sqrt{2r}}\, I_\Phi(G).$$

If $r \leq n-2$, so that $\overline{G}$ has positive degree, then

$$SC_\Phi(\overline{G}) = \frac{1}{\sqrt{2(n-1-r)}}\, I_\Phi(\overline{G}).$$

In the extreme case $r = n-1$, one has $\overline{G}$ edgeless and hence $I_\Phi(\overline{G}) = SC_\Phi(\overline{G}) = 0$.

Remark 3.6 Connectivity and isolated vertices in $\overline{G}$. Although G is assumed connected, its complement $\overline{G}$ need not be connected. If G contains a vertex of degree $n-1$, then $\overline{G}$ has an isolated vertex since

$$d_{\overline{G}}(u) = n - 1 - d_G(u) = 0.$$

Nevertheless, the identities in Lemmas 3.2 and 3.3 remain valid without any connectivity assumption, since they depend only on the degree transformation

$$d_{\overline{G}}(u) = n - 1 - d_G(u)$$

and the relation

$$E(\overline{G}) = \binom{V(G)}{2} \setminus E(G).$$

In particular, isolated vertices do not create difficulties for the edge-sum descriptors themselves, because they are incident to no edges and therefore do not contribute any terms. However, upper bounds involving $\delta(\overline{G})$ require the separate assumption $\delta(\overline{G}) \geq 1$, as noted in Remark 3.4.

3.6 Degree-extremal bounds for selected sum-connectivity descriptors

In this subsection, we record degree-extremal bounds for selected sum-connectivity descriptors listed in Table 2. Let

$$n := |V(G)|, \quad m := |E(G)|, \quad \delta := \delta(G), \quad \Delta := \Delta(G).$$

The results collected here should be viewed as representative extremal estimates for the proposed framework, rather than as a complete extremal characterization for all graph classes. For the sharper closed-form bounds derived below, we impose the additional assumption $\delta \geq 2$, which excludes pendant vertices and simplifies the extremal analysis. This assumption is used only in the present graph-theoretic bounds subsection and is not required for the later QSPR/QSAR computations. All descriptors in and 2 remain well defined whenever $\delta \geq 1$, since every endpoint of an edge has positive degree.

For an edge $uv \in E(G)$, write $d_u := d_G(u)$ and $d_v := d_G(v)$. We repeatedly use Lemma 3.1 together with elementary degree-extremal estimates for the corresponding edge-contribution functions.

Lemma 3.7 Let $\delta \leq x, y \leq \Delta$ with $\delta \geq 2$.

1. $f(x,y) = \dfrac{2xy}{x^2+y^2} \cdot \dfrac{1}{\sqrt{x+y}}$ satisfies

$$f(x,y) \geq \min\left\{ \frac{2\delta\Delta}{\delta^2+\Delta^2} \cdot \frac{1}{\sqrt{\delta+\Delta}}, \frac{1}{\sqrt{2\Delta}} \right\}.$$

2. $g(x,y) = \dfrac{xy}{(x+y)^{3/2}}$ satisfies

$$g(x,y) \geq \min\{g(\delta,\Delta),\, g(\delta,\delta)\} = \min\left\{ \frac{\delta\Delta}{(\delta+\Delta)^{3/2}}, \frac{\sqrt{\delta}}{2\sqrt{2}} \right\}.$$

3. $h(x,y) = \dfrac{2\sqrt{xy}}{(x+y)^{3/2}}$ satisfies

$$h(x,y) \geq \min\{h(\delta,\Delta),\, h(\Delta,\Delta)\} = \min\left\{ \frac{2\sqrt{\delta\Delta}}{(\delta+\Delta)^{3/2}}, \frac{1}{\sqrt{2\Delta}} \right\}.$$

In particular, the relevant lower bounds are attained among corner configurations of the rectangle $[\delta,\Delta]^2$.

Proof All three functions are symmetric in (x,y), hence we may assume without loss of generality that $\delta \leq x \leq y \leq \Delta$.

(1) Proof for $f(x,y) = \dfrac{2xy}{x^2+y^2} \cdot \dfrac{1}{\sqrt{x+y}}$. Write $y = tx$ with $t \in [1, \Delta/x]$. Then

$$f(x,y) = f(x,tx) = \frac{1}{\sqrt{x}}\, F(t), \quad F(t) := \frac{2t}{1+t^2} \cdot \frac{1}{\sqrt{1+t}}.$$

A direct logarithmic differentiation gives, for $t \geq 1$,

$$\frac{F'(t)}{F(t)} = \frac{1}{t} - \frac{2t}{1+t^2} - \frac{1}{2(1+t)} = \frac{1-t^2}{t(1+t^2)} - \frac{1}{2(1+t)} < 0.$$

Thus F is strictly decreasing on $[1,\infty)$, and therefore for fixed x and $y \in [x,\Delta]$,

$$f(x,y) = \frac{1}{\sqrt{x}} F\left(\frac{y}{x}\right) \geq \frac{1}{\sqrt{x}} F\left(\frac{\Delta}{x}\right) = f(x,\Delta).$$

Hence it suffices to bound $f(x,\Delta)$ from below for $x \in [\delta,\Delta]$. Set $x = r\Delta$ with $r \in [\delta/\Delta, 1]$. Then

$$f(x,\Delta) = \frac{1}{\sqrt{\Delta}}\, P(r), \quad P(r) := \frac{2r}{1+r^2} \cdot \frac{1}{\sqrt{1+r}}.$$

Again, compute

$$\frac{P'(r)}{P(r)} = \frac{1}{r} - \frac{2r}{1+r^2} - \frac{1}{2(1+r)} = \frac{2+r-2r^2-3r^3}{2r(1+r)(1+r^2)}.$$

Let $N(r) := 2 + r - 2r^2 - 3r^3$. Note that $N''(r) = -4 - 18r < 0$, so N is strictly concave on $(0,\infty)$. Moreover, $N(0) = 2 > 0$ and

$N(1) = 2 + 1 - 2 - 3 = -2 < 0$, hence N has a unique zero in $(0, 1)$. Therefore $P'(r)$ changes sign exactly once on $(0, 1)$, so P increases and then decreases; consequently, the minimum of P on the closed interval $[\delta/\Delta, 1]$ is attained at an endpoint:

$$P(r) \geq \min\{P(\delta/\Delta), P(1)\}.$$

Returning to $x = r\Delta$ gives

$$f(x, \Delta) \geq \min\{f(\delta, \Delta), f(\Delta, \Delta)\} = \min\left\{\frac{2\delta\Delta}{\delta^2 + \Delta^2} \cdot \frac{1}{\sqrt{\delta + \Delta}}, \frac{1}{\sqrt{2\Delta}}\right\}.$$

Since $f(x, y) \geq f(x, \Delta)$, the stated bound for f follows.

(2) Proof for $g(x, y) = \dfrac{xy}{(x + y)^{3/2}}$. Write again $y = tx$ with $t \in [1, \Delta/x]$. Then

$$g(x, y) = g(x, tx) = \sqrt{x}\ \psi(t), \quad \psi(t) := \frac{t}{(1 + t)^{3/2}}.$$

Compute

$$\frac{\psi'(t)}{\psi(t)} = \frac{1}{t} - \frac{3}{2(1 + t)} = \frac{2 - t}{2t(1 + t)}.$$

Hence ψ is increasing on $(0, 2]$ and decreasing on $[2, \infty)$, so on any interval $[1, T]$ its minimum is attained at an endpoint. Therefore, for $t \in [1, \Delta/x]$,

$$g(x, y) = \sqrt{x}\,\psi(t) \geq \sqrt{x}\min\{\psi(1), \psi(\Delta/x)\} = \min\{g(x, x), g(x, \Delta)\}.$$

Now $g(x, x) = \dfrac{\sqrt{x}}{2\sqrt{2}}$ is increasing in x, so

$$g(x, x) \geq g(\delta, \delta) = \frac{\sqrt{\delta}}{2\sqrt{2}}.$$

Also,

$$g(x, \Delta) = \frac{x\Delta}{(x + \Delta)^{3/2}}$$

is increasing in $x \in (0, \Delta]$ because

$$\frac{d}{dx}\Big(\ln g(x, \Delta)\Big) = \frac{1}{x} - \frac{3}{2(x + \Delta)} = \frac{2\Delta - x}{2x(x + \Delta)} > 0 \quad (x \leq \Delta),$$

hence $g(x, \Delta) \geq g(\delta, \Delta) = \dfrac{\delta\Delta}{(\delta + \Delta)^{3/2}}$. Combining these yields

$$g(x, y) \geq \min\left\{\frac{\delta\Delta}{(\delta + \Delta)^{3/2}}, \frac{\sqrt{\delta}}{2\sqrt{2}}\right\}.$$

(3) Proof for $h(x, y) = \dfrac{2\sqrt{xy}}{(x + y)^{3/2}}$. Again set $y = tx$ with $t \in [1, \Delta/x]$. Then

$$h(x, y) = h(x, tx) = \frac{1}{\sqrt{x}}\ \eta(t), \quad \eta(t) := \frac{2\sqrt{t}}{(1 + t)^{3/2}}.$$

Compute

$$\frac{\eta'(t)}{\eta(t)} = \frac{1}{2t} - \frac{3}{2(1 + t)} = \frac{1 - 2t}{2t(1 + t)} < 0 \quad (t \geq 1),$$

so η is strictly decreasing on $[1, \infty)$. Hence for fixed x and $y \in [x, \Delta]$,

$$h(x, y) = \frac{1}{\sqrt{x}}\eta\left(\frac{y}{x}\right) \geq \frac{1}{\sqrt{x}}\eta\left(\frac{\Delta}{x}\right) = h(x, \Delta).$$

Now write $x = r\Delta$ with $r \in [\delta/\Delta, 1]$. Then

$$h(x, \Delta) = \frac{1}{\sqrt{\Delta}}\ Q(r), \quad Q(r) := \frac{2\sqrt{r}}{(1 + r)^{3/2}}.$$

Since

$$\frac{Q'(r)}{Q(r)} = \frac{1}{2r} - \frac{3}{2(1 + r)} = \frac{1 - 2r}{2r(1 + r)},$$

we have Q increasing on $(0, \frac{1}{2}]$ and decreasing on $[\frac{1}{2}, \infty)$, hence on $[\delta/\Delta, 1]$ the minimum is attained at an endpoint:

$$Q(r) \geq \min\{Q(\delta/\Delta), Q(1)\}.$$

Therefore,

$$h(x, \Delta) \geq \min\{h(\delta, \Delta), h(\Delta, \Delta)\} = \min\left\{\frac{2\sqrt{\delta\Delta}}{(\delta + \Delta)^{3/2}}, \frac{1}{\sqrt{2\Delta}}\right\}.$$

Since $h(x, y) \geq h(x, \Delta)$, the stated bound for h follows.

This completes the proof of the lemma. □

Proposition 3.8 All bounds displayed in Table 3 hold for the corresponding indices.

Proof Each descriptor in Table 3 is an edge-sum

$$X(G) = \sum_{uv \in E(G)} k(d_u, d_v),$$

where $k(d_u, d_v) \geq 0$ and $\delta \leq d_u, d_v \leq \Delta$.

For rows (2), (3), and (13), the stated lower bounds follow from Lemma 3.7, while the corresponding upper bounds are obtained by evaluating the edge kernels at the relevant degree-extremal configurations.

For the remaining rows, the edge kernels are either monotone on $[\delta, \Delta]^2$ or admit direct elementary bounds. For example:

– row (1) uses

$$k(d_u, d_v) = \frac{d_u + d_v}{d_u d_v \sqrt{d_u + d_v}} = \frac{\sqrt{d_u + d_v}}{d_u d_v},$$

which is decreasing in each variable, so its minimum and maximum occur at (Δ, Δ) and (δ, δ), respectively;

– row (7) follows from

$$0 \leq |d_u - d_v| \leq \Delta - \delta, \quad d_u + d_v \geq \delta + \Delta,$$

hence

$$0 \leq \frac{|d_u - d_v|}{\sqrt{d_u + d_v}} \leq \frac{\Delta - \delta}{\sqrt{\Delta + \delta}};$$

– row (14) uses

$$k(d_u, d_v) = \left(\frac{1}{d_u} + \frac{1}{d_v}\right)^{-1/2} \frac{1}{\sqrt{d_u + d_v}} = \frac{\sqrt{d_u d_v}}{d_u + d_v},$$

Table 3 Degree-extremal bounds for selected sum-connectivity descriptors of G under the standing assumption $\delta = \delta(G) \geq 2$ (with $m = |E(G)|$ and $\Delta = \Delta(G)$). The sharper lower bounds in rows (2), (3), and (13) follow from Lemma 3.7

No.	Index	Lower bound	Upper bound
(1)	$SCReZG_1(G)$	$m\frac{\sqrt{2}}{\Delta^{3/2}}$	$m\frac{\sqrt{2}}{\delta^{3/2}}$
(2)	$SCGAF(G)$	$m \cdot \min\left\{\frac{2\delta\Delta}{\delta^2+\Delta^2} \cdot \frac{1}{\sqrt{\delta+\Delta}}, \frac{1}{\sqrt{2\Delta}}\right\}$	$m\frac{1}{\sqrt{2\delta}}$
(3)	$SCReZG_2(G)$	$m \cdot \min\left\{\frac{\delta\Delta}{(\delta+\Delta)^{3/2}}, \frac{\sqrt{\delta}}{2\sqrt{2}}\right\}$	$m\frac{\sqrt{\Delta}}{2\sqrt{2}}$
(4)	$SCESO(G)$	$2m\,\delta\sqrt{\delta}$	$2m\,\Delta\sqrt{\Delta}$
(5)	$SCReZG_3(G)$	$m\,\delta^2\sqrt{2\delta}$	$m\,\Delta^2\sqrt{2\Delta}$
(6)	$SCEU(G)$	$m\sqrt{\frac{3\delta}{2}}$	$m\sqrt{\frac{3\Delta}{2}}$
(7)	$SCA(G)$	0	$m\,\frac{\Delta-\delta}{\sqrt{\Delta+\delta}}$
(8)	$SCBSO(G)$	$m\,\frac{1}{\Delta\sqrt{\Delta}}$	$m\,\frac{1}{\delta\sqrt{\delta}}$
(9)	$SCAG(G)$	$m\,\frac{\sqrt{2}}{2\sqrt{\Delta}}$	$m\,\frac{\sqrt{2}}{2\sqrt{\delta}}$
(10)	$HSC(G)$	$m\,\frac{1}{2\Delta}$	$m\,\frac{1}{2\delta}$
(11)	$SCF(G)$	$m\,(\sqrt{2}\,\delta^{3/2})$	$m\,(\sqrt{2}\,\Delta^{3/2})$
(12)	$SCIN_1(G)$	$m\,\frac{1}{\Delta}$	$m\,\frac{1}{\delta}$
(13)	$SCGA(G)$	$m \cdot \min\left\{\frac{2\sqrt{\delta\Delta}}{(\delta+\Delta)^{3/2}}, \frac{1}{\sqrt{2\Delta}}\right\}$	$m\,\frac{1}{\sqrt{2\delta}}$
(14)	$SCIN_2(G)$	$m\,\frac{\sqrt{\delta\Delta}}{\delta+\Delta}$	$\frac{m}{2}$

for which

$$\frac{\sqrt{\delta\Delta}}{\delta+\Delta} \le \frac{\sqrt{d_u d_v}}{d_u+d_v} \le \frac{1}{2},$$

the upper bound being a direct consequence of the AM–GM inequality.
The remaining rows are handled similarly by evaluating the corresponding kernel at the degree-extremal pairs. Summing the resulting edgewise bounds over all m edges gives the entries of Table 3. □

Remark 3.9 If G is r-regular, then $\delta = \Delta = r$ and all bounds in Proposition 3.8 collapse to equalities.

Although Table 3 is naturally expressed in terms of the size m, it is sometimes convenient to rewrite these estimates in terms of n and the degree extremes. The following corollary provides such coarser n-based reformulations (Table 4).

Corollary 3.10 Since $2m = \sum_{u\in V(G)} d_G(u)$, one has

$$\frac{n\delta}{2} \le m \le \binom{n}{2}.$$

Hence, any estimate of the form

Table 4 n-based reformulations of the bounds in Table 3. Using $\frac{n\delta}{2} \le m \le \binom{n}{2}$, each estimate $mL \le X(G) \le mU$ yields $\frac{n\delta}{2}L \le X(G) \le \binom{n}{2}U$

No.	Index	n-based lower bound	n-based upper bound
(1)	$SCReZG_1(G)$	$\frac{n\delta}{2}\cdot\frac{\sqrt{2}}{\Delta^{3/2}}$	$\binom{n}{2}\cdot\frac{\sqrt{2}}{\delta^{3/2}}$
(2)	$SCGAF(G)$	$\frac{n\delta}{2}\cdot\min\left\{\frac{2\delta\Delta}{\delta^2+\Delta^2}\cdot\frac{1}{\sqrt{\delta+\Delta}},\ \frac{1}{\sqrt{2\Delta}}\right\}$	$\binom{n}{2}\cdot\frac{1}{\sqrt{2\delta}}$
(3)	$SCReZG_2(G)$	$\frac{n\delta}{2}\cdot\min\left\{\frac{\delta\Delta}{(\delta+\Delta)^{3/2}},\ \frac{\sqrt{\delta}}{2\sqrt{2}}\right\}$	$\binom{n}{2}\cdot\frac{\sqrt{\Delta}}{2\sqrt{2}}$
(4)	$SCESO(G)$	$\frac{n\delta}{2}\cdot(2\delta\sqrt{\delta})$	$\binom{n}{2}\cdot(2\Delta\sqrt{\Delta})$
(5)	$SCReZG_3(G)$	$\frac{n\delta}{2}\cdot\delta^2\sqrt{2\delta}$	$\binom{n}{2}\cdot\Delta^2\sqrt{2\Delta}$
(6)	$SCEU(G)$	$\frac{n\delta}{2}\cdot\sqrt{\frac{3\delta}{2}}$	$\binom{n}{2}\cdot\sqrt{\frac{3\Delta}{2}}$
(7)	$SCA(G)$	0	$\binom{n}{2}\cdot\frac{\Delta-\delta}{\sqrt{\Delta+\delta}}$
(8)	$SCBSO(G)$	$\frac{n\delta}{2}\cdot\frac{1}{\Delta\sqrt{\Delta}}$	$\binom{n}{2}\cdot\frac{1}{\delta\sqrt{\delta}}$
(9)	$SCAG(G)$	$\frac{n\delta}{2}\cdot\frac{\sqrt{2}}{2\sqrt{\Delta}}$	$\binom{n}{2}\cdot\frac{\sqrt{2}}{2\sqrt{\delta}}$
(10)	$HSC(G)$	$\frac{n\delta}{2}\cdot\frac{1}{2\Delta}$	$\binom{n}{2}\cdot\frac{1}{2\delta}$
(11)	$SCF(G)$	$\frac{n\delta}{2}\cdot(\sqrt{2}\,\delta^{3/2})$	$\binom{n}{2}\cdot(\sqrt{2}\,\Delta^{3/2})$
(12)	$SCIN_1(G)$	$\frac{n\delta}{2}\cdot\frac{1}{\Delta}$	$\binom{n}{2}\cdot\frac{1}{\delta}$
(13)	$SCGA(G)$	$\frac{n\delta}{2}\cdot\min\left\{\frac{2\sqrt{\delta\Delta}}{(\delta+\Delta)^{3/2}},\ \frac{1}{\sqrt{2\Delta}}\right\}$	$\binom{n}{2}\cdot\frac{1}{\sqrt{2\delta}}$
(14)	$SCIN_2(G)$	$\frac{n\delta}{2}\cdot\frac{\sqrt{\delta\Delta}}{\delta+\Delta}$	$\binom{n}{2}\cdot\frac{1}{2}$

$$mL \le X(G) \le mU$$

yields the corresponding n-based bounds

$$\frac{n\delta}{2} L \le X(G) \le \binom{n}{2} U.$$

For instance, applying this to the bounds for $SCGA(G)$ gives

$$\frac{n\delta}{2} \min\left\{ \frac{2\sqrt{\delta\Delta}}{(\delta+\Delta)^{3/2}}, \frac{1}{\sqrt{2\Delta}} \right\} \le SCGA(G) \le \binom{n}{2} \frac{1}{\sqrt{2\delta}}.$$

4 QSAR/QSPR models

This section provides an application-oriented validation of the proposed sum-connectivity descriptors as molecular features for both quantitative structure-property relationship (QSPR) and quantitative structure-activity relationship (QSAR) modelling. The empirical aim is to compare the proposed descriptors directly with their matched classical counterparts, assess whether they add complementary signal beyond RDKit descriptors, and determine how stable these conclusions remain under scaffold-based validation.

Table 5 Feature families considered in the QSPR/QSAR modelling study

Feature family	Description
RDKit	Standard RDKit molecular descriptors only. These descriptors capture molecular size, connectivity, branching, ring content, and related structural and physicochemical characteristics (Table 9).
Classical	The 14 classical degree-based topological indices listed in Table 7.
Sum-Conn.	The 14 proposed sum-connectivity descriptors obtained by reweighting the corresponding classical degree-based indices, as listed in Table 8.
RDKit+Classical	The combined feature block formed by concatenating the RDKit descriptor set with the 14 classical degree-based indices Tables 7 and 9.
RDKit+Sum-Conn.	The combined feature block formed by concatenating the RDKit descriptor set with the 14 proposed sum-connectivity descriptors Tables 8 and 9.
All Combined	The full integrated descriptor set containing RDKit descriptors together with both index families, i.e., classical degree-based indices and their sum-connectivity analogues Tables 7, 8, and 9.

4.1 Data acquisition and curated modelling dataset

The underlying antibacterial dataset was assembled from the ChEMBL bioactivity database, a manually curated resource of bioactive drug-like molecules [49]. To reduce biological and endpoint heterogeneity and to obtain a dataset suitable for regression modelling, only assays targeting *Escherichia coli* were retained. In addition, the activity endpoint was restricted to minimum inhibitory concentration (MIC) measurements so that the response variable reflects antibacterial potency under a common experimental definition [44, 45]. Reported MIC values were standardized to a common micromolar (μM) scale, duplicate records for the same compound were aggregated using the median, and the final activity endpoint was expressed as

$$\mathrm{pMIC} = -\log_{10}(\mathrm{MIC}),$$

where MIC is given in μM units.

The modelling dataset used in the study contains 6,657 compounds and 134 columns after integrating antibacterial activity, physicochemical properties, and descriptor blocks. For QSAR modelling of antibacterial activity, rows with missing $pMIC$ were excluded, yielding an effective sample size of 6,656 compounds. In addition to random repeated cross-validation, scaffold identifiers were computed from the molecular structures using Murcko scaffolds, yielding 3,053 unique scaffolds for scaffold-based validation.

4.2 Descriptor blocks, physicochemical properties, and feature-set design

Each compound in the curated dataset is associated with the antibacterial activity endpoint ($pMIC$) and eight physicochemical properties (Table 6): QED, SPS, $MolWt$ (MW), $TPSA$, Phi, $MolLogP$ (MLP), $MolMR$ (MR), and NP-likeness (NpL). In addition, each molecule is represented by three descriptor blocks:

1. 14 classical degree-based indices (Table 7);
2. 14 sum-connectivity analogues obtained from the proposed reweighting framework (Table 8);
3. 90 RDKit descriptors capturing molecular size, connectivity, branching, ring content, and related structural characteristics (Table 9).

For readability, Tables 7, 8, and 9 report representative columns from each descriptor family rather than the full blocks.

Based on these blocks, six feature families were considered throughout the modelling study Table 5:

This design enables both a controlled comparison between classical and sum-connectivity descriptors and

an assessment of their contribution when combined with a richer RDKit representation.

4.2.1 Illustrative calculation for one representative compound

To make the descriptor construction explicit, we consider compound C_1 with molecular formula $C_{24}H_{27}FN_6O_4$, represented by the standard hydrogen-suppressed molecular graph associated with the given SMILES string. This graph has $n = 35$ vertices and $m = 39$ edges. Its edge-degree partition is

$$m_{1,2} = 1, \quad m_{1,3} = 3, \quad m_{2,2} = 7, \quad m_{2,3} = 23, \quad m_{3,3} = 5,$$

where $m_{i,j}$ denotes the number of edges joining a vertex of degree i to a vertex of degree j. Using Table 1, the redefined first Zagreb index is

$$\begin{aligned} ReZG_1(C_1) &= \sum_{uv\in E(G)} \frac{d_u + d_v}{d_u d_v} \\ &= 1\left(\frac{1+2}{1\cdot 2}\right) + 3\left(\frac{1+3}{1\cdot 3}\right) + 7\left(\frac{2+2}{2\cdot 2}\right) \\ &+ 23\left(\frac{2+3}{2\cdot 3}\right) + 5\left(\frac{3+3}{3\cdot 3}\right) \\ &= \frac{3}{2} + 3\left(\frac{4}{3}\right) + 7(1) + 23\left(\frac{5}{6}\right) + 5\left(\frac{2}{3}\right) = 35.0000. \end{aligned}$$

Similarly, the redefined second Zagreb index is

$$\begin{aligned} ReZG_2(C_1) &= \sum_{uv\in E(G)} \frac{d_u d_v}{d_u + d_v} \\ &= 1\left(\frac{1\cdot 2}{1+2}\right) + 3\left(\frac{1\cdot 3}{1+3}\right) + 7\left(\frac{2\cdot 2}{2+2}\right) \\ &+ 23\left(\frac{2\cdot 3}{2+3}\right) + 5\left(\frac{3\cdot 3}{3+3}\right) \\ &= \frac{2}{3} + 3\left(\frac{3}{4}\right) + 7(1) + 23\left(\frac{6}{5}\right) + 5\left(\frac{3}{2}\right) = 45.0167. \end{aligned}$$

Now, using Table 2, the corresponding sum-connectivity descriptors are obtained by multiplying each edge contribution by

$$w_{SC}(u,v) = \frac{1}{\sqrt{d_u + d_v}}.$$

Hence,

$$\begin{aligned} SCReZG_1(C_1) &= \sum_{uv\in E(G)} \frac{d_u + d_v}{d_u d_v \sqrt{d_u + d_v}} \\ &= 1\left(\frac{3}{2\sqrt{3}}\right) + 3\left(\frac{4}{3\sqrt{4}}\right) + 7\left(\frac{4}{4\sqrt{4}}\right) \\ &+ 23\left(\frac{5}{6\sqrt{5}}\right) + 5\left(\frac{6}{9\sqrt{6}}\right) \\ &= 16.2984, \end{aligned}$$

and

$$\begin{aligned} SCReZG_2(C_1) &= \sum_{uv\in E(G)} \frac{d_u d_v}{(d_u + d_v)^{3/2}} \\ &= 1\left(\frac{2}{3\sqrt{3}}\right) + 3\left(\frac{3}{4^{3/2}}\right) + 7\left(\frac{4}{4^{3/2}}\right) \\ &+ 23\left(\frac{6}{5^{3/2}}\right) + 5\left(\frac{9}{6^{3/2}}\right) \\ &= 20.4149. \end{aligned}$$

These values agree with the corresponding entries reported for C_1 in the descriptor tables. This example shows explicitly how one representative molecular graph generates both the classical degree-based descriptors and their sum-connectivity counterparts before the full QSAR/QSPR modelling stage.

4.2.2 Illustrative compound-level descriptor partition

To make the feature construction concrete for one representative molecule, consider compound C_1. From Table 6, this compound has $QED = 0.5192$, $SPS = 43.43$, $MolWt = 482.516$, $TPSA = 110.61$, $Phi = 2.1259$, $MolLogP = 1.3332$, $MolMR = 127.4974$, $NpL = -1.09$, and $pMIC = -0.0155$. The same molecule is represented in the classical block of Table 7 by values such as $RZG_1 = 35.00$, $GAF = 35.83$, $RZG_2 = 45.02$, and $ESO = 665.77$, in the sum-connectivity block of Table 8 by $SCRZG_1 = 16.30, SCGAF = 16.40, SCRZG_2 = 20.41$, and $SCESO = 299.84$, and in the RDKit block of Table 9 by descriptors including $BJ = 1.7697$, $BC = 2449.7264$, $\chi_0 = 48.2207$, and $\chi_1 = 27.4895$. In the modelling stage, the feature vector for this compound is taken either from one selected block or from the concatenated blocks used in the RDKit+Classical, RDKit+Sum-Conn., and All Combined settings. This makes explicit how one representative compound is partitioned across response, property, and descriptor spaces before the cross-validation workflow is applied to the full dataset.

Table 6 Physicochemical properties and antibacterial activity endpoint ($pMIC$) for the compound set ($C_1, C_2, \ldots, C_{6657}$)

Compound	QED	SPS	MW	$TPSA$	Phi	MLP	MR	NpL	$pMIC$
C_1	0.5192	43.43	482.516	110.61	2.1259	1.3332	127.4974	−1.09	−0.0155
C_2	0.7038	48.27	220.703	41.81	0.6115	3.1574	63.1751	0.02	−2.2244
C_3	0.5994	44.50	463.513	92.41	2.0912	2.2416	124.7307	−1.19	−0.9360
C_4	0.6224	40.06	445.523	92.41	1.9184	2.2936	124.4517	−1.26	−0.0501
C_5	0.8101	35.52	402.829	87.46	1.8375	1.7698	100.5480	−0.72	0.4912
C_6	0.8932	43.63	331.347	74.57	1.1758	1.5833	88.4790	−0.72	−1.9849
C_7	0.5847	23.33	368.239	50.09	2.0119	4.8655	95.2424	−1.70	−1.9390
C_8	0.4453	26.58	450.829	134.81	2.1087	1.8423	111.9539	−0.27	1.7509
C_9	0.2245	52.80	299.352	113.72	1.2362	−0.1760	74.3280	0.89	0.0782
C_{10}	0.0598	70.63	995.029	402.41	4.6268	−6.2394	238.1466	−0.05	−0.1783
…	…	…	…	…	…	…	…	…	…
C_{6652}	0.3957	43.03	510.428	44.76	1.6453	1.6263	122.8409	0.75	−1.0864
C_{6653}	0.5971	18.59	226.235	69.14	0.7802	2.6411	64.8593	−0.36	−2.0000
C_{6654}	0.0573	34.28	717.746	164.42	3.8596	5.4989	205.2296	−0.12	−0.9031
C_{6655}	0.3192	44.00	537.438	44.76	1.7563	1.5730	116.4689	0.88	−1.3674
C_{6656}	0.3656	30.69	490.585	78.18	2.1624	3.9243	137.2300	−1.43	−2.4065
C_{6657}	0.7566	60.43	376.456	51.54	0.9932	2.8250	105.2755	1.54	−1.8195

QED=Quantitative Estimate of Drug-likeness, SPS=Synthetic Accessibility Score, MW=Molecular Weight (g/mol), $TPSA$=Topological Polar Surface Area (Å^2), Phi=Flexibility Index, MLP=Calculated LogP, MR=Molar Refractivity (cm^3/mol), NpL=NP-Likeness Score, $\mathrm{pMIC} = -\log_{10}(\mathrm{MIC})$ with MIC expressed in μM

4.3 Model development and validation framework

Two complementary supervised-learning problems were investigated. The first was a multi-target QSPR setting, in which the objective was to predict eight physicochemical properties (QED, SPS, $MolWt$, $TPSA$, Phi, $MolLogP$, $MolMR$, and NP-likeness) from graph-derived and cheminformatics descriptors. The second was a QSAR setting, in which antibacterial activity against *E. coli* was modelled using the continuous endpoint $pMIC$.

To ensure a rigorous and internally consistent comparison across descriptor families, all experiments were performed within a unified machine-learning workflow following standard statistical-learning formulations [50]. The candidate regressors were selected to represent complementary modelling assumptions: Linear Regression, Partial Least Squares Regression (PLS) [51], Ridge Regression [52], support vector regression with a radial-basis-function kernel (SVR-RBF) [53, 54], Random Forest [55], Extra Trees [56], and XGBoost [57]. These models range from transparent linear baselines to kernel-based and ensemble learners. Predictors were standardized within the modelling pipeline, and the same validation schemes were applied across competing feature blocks so that performance differences could be attributed to descriptor content rather than to inconsistent evaluation.

For explicit notation, the linear baseline has the form $\hat{y} = \beta_0 + \mathbf{x}^\top \boldsymbol{\beta}$, whereas Ridge regression estimates $\boldsymbol{\beta}$ by minimizing $\sum_i (y_i - \beta_0 - \mathbf{x}_i^\top \boldsymbol{\beta})^2 + \lambda \|\boldsymbol{\beta}\|_2^2$. PLS projects $\mathbf{x}$ onto latent components before regression. SVR-RBF uses $\hat{y}(\mathbf{x}) = \sum_i (\alpha_i - \alpha_i^*) \exp[-\gamma \|\mathbf{x}_i - \mathbf{x}\|^2] + b$. Random Forest and Extra Trees estimate $\hat{y}(\mathbf{x}) = (1/T) \sum_{t=1}^{T} h_t(\mathbf{x})$ by averaging randomized decision trees, while XGBoost uses an additive tree ensemble $\hat{y}(\mathbf{x}) = \sum_{m=1}^{M} f_m(\mathbf{x})$ fitted by gradient boosting. These equations clarify the mathematical form of the learners while recognizing that the tree-based methods are algorithmic ensembles rather than single closed-form regression equations.

The principal validation protocol was *5×repeated 5-fold cross-validation*, which provides a more stable estimate of generalization than a single random partition and reduces sensitivity to a particular split realization. In addition, for the $pMIC$ QSAR task, a more stringent *Murcko scaffold-based 5-fold GroupKFold* protocol was used to probe generalization across distinct core chemotypes. This scaffold-based setting is especially important for antibacterial modelling because random splits may overestimate performance when closely related analogues are distributed across training and test folds.

Model performance was quantified using the coefficient of determination (R^2), mean absolute error (MAE), and root-mean-square error (RMSE). The mean R^2 was used for model ranking because it measures the fraction of response variance explained on held-out folds, while MAE and RMSE were retained to report error magnitudes on the original endpoint scale. For the QSAR endpoint, Pearson,

Table 7 Summary of selected classical degree-based indices for antibacterial *E. coli* compounds $C_1, C_2, \ldots, C_{6657}$. The table reports representative descriptor columns; the full descriptor matrix is supplied in the supplementary data

Compound	RZG_1	GAF	RZG_2	ESO	RZG_3	EU	A	BSO	AG	SC	$\cdots$	GA	IN_2
C_1	35.00	35.83	45.02	665.77	1114.00	163.95	30.00	25.41	40.00	17.90	...	38.08	41.73
C_2	15.00	15.58	20.10	305.28	528.00	73.26	12.00	10.92	17.47	7.71	...	16.57	18.39
C_3	34.00	35.38	43.87	639.69	1074.00	158.55	26.00	24.57	38.80	17.51	...	37.25	40.68
C_4	33.00	34.78	42.62	612.90	1024.00	153.21	24.00	23.75	37.65	17.10	...	36.38	39.59
C_5	26.00	24.29	32.41	528.04	908.00	122.71	26.00	19.31	29.42	12.69	...	26.78	29.87
...	...	...	...	...	...	...	...	...	...	...	...	...	...
C_{6653}	17.00	17.28	21.90	326.48	540.00	80.30	16.00	12.38	19.56	8.68	...	18.49	20.33
C_{6654}	48.00	48.06	58.50	843.74	1392.00	212.73	32.00	34.84	53.34	24.21	...	50.79	54.90
C_{6655}	30.00	29.22	38.36	609.41	1038.00	143.52	30.00	22.09	34.38	14.97	...	31.79	35.35
C_{6656}	35.00	36.38	45.45	668.74	1150.00	163.70	26.00	25.28	39.77	17.98	...	38.27	41.89
C_{6657}	28.00	30.63	41.51	668.67	1270.00	149.81	22.00	20.31	33.75	14.70	...	32.31	36.75

RZG_1, RZG_2, and RZG_3 denote redefined Zagreb indices; GAF denotes the geometric-arithmetic F-index; ESO and EU denote the elliptic and Euler Sombor indices; A denotes the Albertson index; BSO denotes the Banhatti-Sombor index; AG and GA denote arithmetic-geometric and geometric-arithmetic indices; SC denotes the classical sum-connectivity index; and IN_2 denotes the second inverse Nirmala index

Table 8 Summary of selected sum-connectivity indices for antibacterial *E. coli* compounds $C_1, C_2, \ldots, C_{6657}$. These descriptors are obtained by applying the common factor $(d_u + d_v)^{-1/2}$ to the corresponding classical edge-degree kernels

Compound	$SCRZG_1$	$SCGAF$	$SCRZG_2$	$SCESO$	$SCRZG_3$	$SCEU$	SCA	$SCBSO$	$SCAG$	HSC	$\cdots$	$SCGA$	$SCIN_2$
C_1	16.30	16.40	20.41	299.84	496.27	74.50	13.86	11.84	18.38	8.27	...	17.46	19.04
C_2	6.90	7.04	8.99	135.69	231.51	32.88	5.58	5.03	7.94	3.52	...	7.50	8.29
C_3	15.89	16.27	19.97	288.70	479.44	72.26	11.97	11.49	17.89	8.12	...	17.15	18.63
C_4	15.45	16.06	19.48	277.64	459.39	70.07	10.97	11.12	17.41	7.95	...	16.81	18.19
C_5	12.06	10.93	14.37	231.67	390.35	54.67	12.03	8.97	13.36	5.81	...	12.11	13.39
...	...	...	...	...	...	...	...	...	...	...	...	...	...
C_{6653}	7.85	7.87	9.91	147.15	241.09	36.44	7.37	5.72	8.95	3.98	...	8.44	9.24
C_{6654}	22.63	22.32	26.88	384.15	624.69	97.95	14.94	16.44	24.86	11.33	...	23.63	25.40
C_{6655}	13.86	13.20	17.11	268.98	450.80	64.17	13.72	10.22	15.62	6.84	...	14.40	15.90
C_{6656}	16.50	16.71	20.59	299.60	508.45	74.30	12.12	11.93	18.36	8.37	...	17.63	19.14
C_{6657}	12.85	13.56	18.03	286.27	533.81	65.30	10.02	9.34	15.06	6.63	...	14.36	16.15

The prefix SC indicates the proposed sum-connectivity analogue of the corresponding classical index; HSC denotes the hyper sum-connectivity index; and $SCIN_2$ denotes the sum-connectivity version of the second inverse Nirmala index

Table 9 Summary of selected RDKit descriptors for antibacterial *E. coli* compounds $C_1, C_2, \ldots, C_{6657}$. The listed columns are representative examples from the RDKit block; the abbreviation list below the table defines the symbols used

Compound	BJ	BC	χ_0	χ_{0n}	χ_{0v}	χ_1	χ_{1n}	χ_{1v}	χ_{2n}	χ_{2v}	⋯	κ_3	$AIPC$
C_1	1.7697	2449.7264	48.2207	43.6942	16.6942	27.4895	22.4386	9.0442	6.1456	6.1456	⋯	5.1022	3.1107
C_2	2.7365	1013.3382	21.7735	20.2724	8.0283	12.4908	10.4514	4.4878	2.9716	3.3496	⋯	1.3634	2.8952
C_3	1.7736	2440.4832	46.6434	42.3388	16.3388	26.6723	21.8665	8.9193	6.0335	6.0335	⋯	4.7379	3.0940
C_4	1.7736	2434.3127	46.6434	42.9608	15.9608	26.6723	22.1775	8.7303	5.8445	5.8445	⋯	4.8453	3.0940
C_5	0.0000	1610.7961	38.3424	33.6475	13.4034	21.2560	16.7098	6.4764	4.7111	4.7111	⋯	3.7445	2.9004
⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯
C_{6653}	2.4840	1097.6046	20.2042	18.2109	8.2109	12.6660	9.3164	4.5527	3.1527	3.1527	⋯	1.6934	2.9675
C_{6654}	0.0000	3458.1061	76.0441	70.1501	25.6620	43.2952	34.8724	12.5387	8.2492	8.2492	⋯	10.4687	0.6564
C_{6655}	0.0000	2160.2765	49.1269	46.9582	17.1157	26.9875	23.8233	7.8233	5.7122	5.7122	⋯	4.6966	3.3581
C_{6656}	1.8221	2503.4204	46.8670	42.8301	17.6466	27.3238	22.0760	9.8925	6.2413	7.2404	⋯	4.5200	3.4726
C_{6657}	2.0967	2103.8820	40.2900	37.6192	13.6192	22.9520	19.9081	7.9081	6.2599	6.2599	⋯	1.6842	3.3361

BJ = Balaban's J index, BC = Bertz Complexity index, χ_0 = zero-order Kier connectivity index, χ_{0n} = zero-order connectivity index (non-valence form), χ_{0v} = zero-order valence Kier connectivity index, χ_1 = first-order Kier connectivity index, χ_{1n} = first-order connectivity index (non-valence form), χ_{1v} = first-order valence Kier connectivity index, χ_{2n} = second-order connectivity index (non-valence form), χ_{2v} = second-order valence connectivity index, $\kappa_{1,2,3}$ = first-, second-, and third-order Kier shape indices, $AIPC$ = average information content on atomic composition

Spearman, and concordance correlations were treated as complementary diagnostics rather than replacements for R^2, because correlation alone can overstate practical predictive accuracy when bias or scale errors are present. Additional model-validation analyses were carried out for the strongest QSAR configuration, including Y-randomization, permutation importance, and learning-curve analysis, in order to assess chance correlation, descriptor relevance, and data-efficiency behavior, respectively.

Figure 1 summarizes the complete modelling workflow used to generate the final tables and figures.

4.4 QSPR modelling of physicochemical properties

We first examined the proposed descriptors in a multi-target QSPR framework. Three feature configurations were considered: **Classical Indices**, **Sum-Connectivity Indices**, and **All Combined**, where the last block integrates RDKit descriptors with both index families. This design enables both a direct assessment of the proposed reweighting in isolation and an evaluation of its role within a richer descriptor pool.

Table 10 summarizes the best-performing model for each physicochemical endpoint under repeated 5-fold cross-validation. The integrated **All Combined** representation yielded the highest cross-validated R^2 for every target. However, these values should be interpreted as within-space predictive accuracy rather than automatic evidence of orthogonal descriptor complementarity, because several physicochemical endpoints are tightly related to constitutional descriptors already present in the combined block. The strongest empirical support for the proposed reweighting therefore comes from the matched standalone comparison against the classical family, whereas the combined-descriptor results serve as an upper-bound benchmark within the present descriptor space.

To specifically evaluate the effect of the proposed reweighting, we next compared the **Sum-Connectivity Indices** against the corresponding **Classical Indices** under identical dimensionality and validation conditions. Table 11 shows that the sum-connectivity family produced higher mean R^2 values for all eight physicochemical targets. The gains were especially pronounced for $TPSA$, $MolLogP$,and SPS, whereas $MolWt$ remained essentially unchanged. Averaged over the eight targets, the mean R^2 increased from 0.6951 for the classical family to 0.7015 for the sum-connectivity family, supporting the conclusion that the proposed normalization yields a modest but systematic improvement in standalone QSPR performance.

Figure 2 provides a compact visual summary of QSPR performance across targets and descriptor families. The

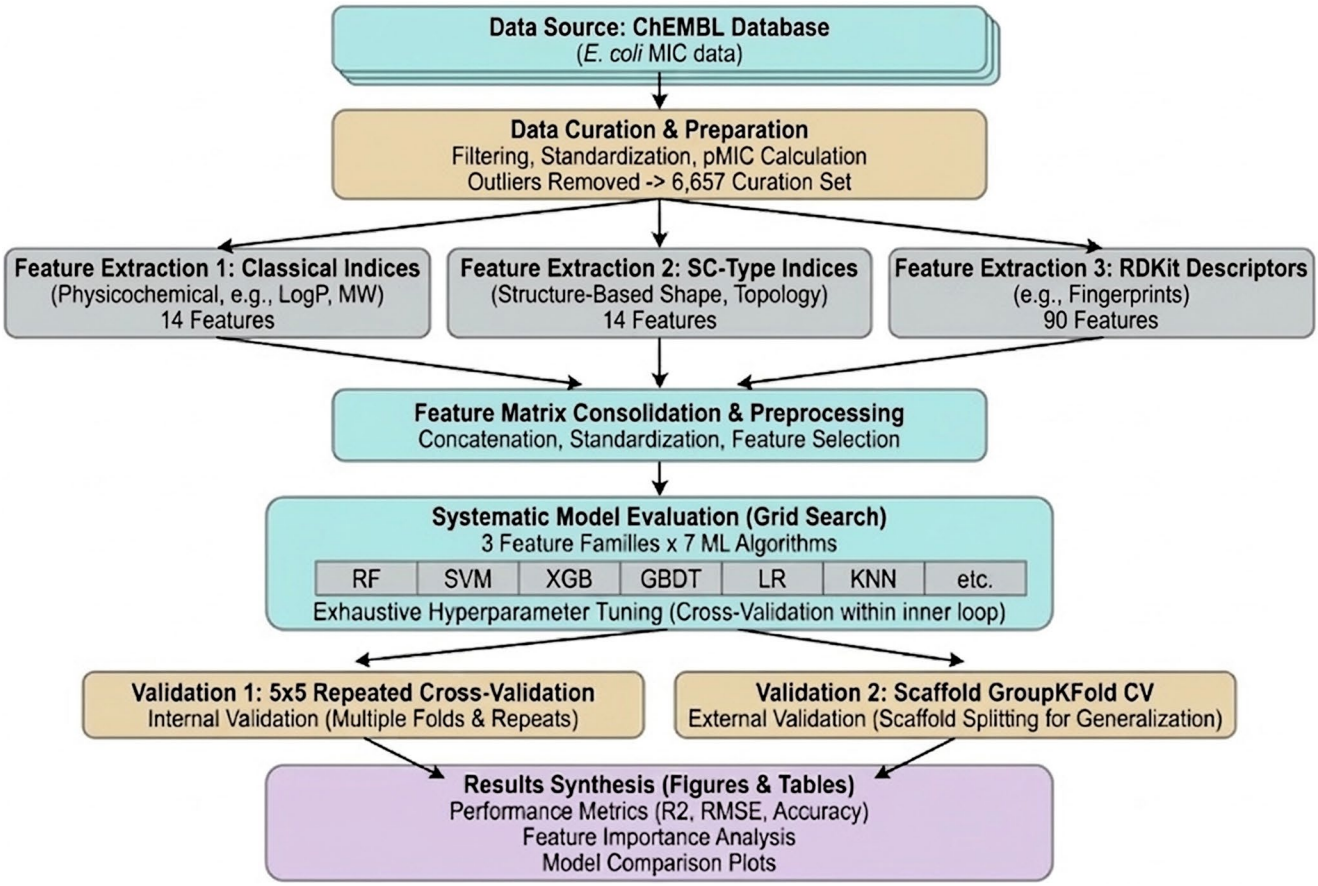

Fig. 1 Unified modelling workflow for the QSPR/QSAR study, from ChEMBL curation and descriptor construction to repeated cross-validation, scaffold validation, and diagnostic analyses

dominance of the integrated representation is evident, but the direct comparison between the two index families also shows that the proposed sum-connectivity descriptors consistently match or exceed their classical counterparts.

4.5 Correlation analysis and descriptor overlap

Because several QSPR endpoints are constitution-driven, we examined feature-target Spearman correlations to assess whether the strongest QSPR scores could reflect descriptor overlap with the targets. Figure 3 shows that $MolWt$ is almost directly encoded by $ExactMolWt$ ($\rho = 0.99999$) and $HeavyAtomMolWt$ ($\rho = 0.99787$), while $TPSA$ is also strongly associated with $NOCount$ ($\rho = 0.9304$). Accordingly, the very high QSPR R^2 values for several combined-descriptor models should be interpreted cautiously as upper-bound performance within the present descriptor space rather than as evidence of broad extrapolative generalization.

The proposed descriptors are most informative in matched standalone comparisons against the classical family, whereas combined-model gains should not be interpreted as proof of large independent information gain.

4.5.1 PCA sensitivity analysis

To test whether the strongest combined-descriptor results were driven mainly by raw-feature redundancy, we performed a separate rerun in which the descriptor blocks were first compressed by principal component analysis (PCA) and the same modelling workflow was then reapplied. Table 12 compares the strongest non-PCA and PCA results. For the **All Combined** QSPR benchmark, PCA reduced the best-model R^2 for every endpoint, with a mean change of approximately -0.0519 across the eight targets and the largest drop observed for NP-likeness (from 0.9349 to 0.8078). The same pattern was observed for the main QSAR benchmarks: under repeated cross-validation, the best **RDKit** result decreased from 0.5959 to 0.4744, and the best **All Combined** result decreased from 0.5938 to 0.4734; under scaffold validation, the corresponding drops were from 0.4799 to 0.4003 and from 0.4767 to 0.3970, respectively.

Table 10 Best-performing QSPR models under $5\times$ repeated 5-fold cross-validation for the three descriptor families. For each target, the listed model is the one with the highest mean R^2

Feature family	Target	Best model	R^2 (mean±std)	MAE (mean±std)	RMSE (mean±std)
Sum-Conn.	QED	RF	0.7225 ± 0.0151	0.0945 ± 0.0022	0.1254 ± 0.0026
	SPS	XGBoost	0.5032 ± 0.0288	8.5560 ± 0.2127	12.3293 ± 0.4072
	$MolWt$	ExtraTrees	0.9528 ± 0.0042	22.9396 ± 0.4599	34.1301 ± 1.0305
	$TPSA$	XGBoost	0.7166 ± 0.0219	24.4883 ± 0.4639	34.6357 ± 1.3056
	Phi	XGBoost	0.8300 ± 0.0287	0.2934 ± 0.0063	0.4448 ± 0.0173
	$MolLogP$	XGBoost	0.4393 ± 0.0314	1.4670 ± 0.0297	2.1191 ± 0.0483
	$MolMR$	XGBoost	0.9687 ± 0.0027	5.0975 ± 0.1201	7.0048 ± 0.2642
	NP-likeness	XGBoost	0.4790 ± 0.0249	0.5304 ± 0.0130	0.7261 ± 0.0175
Classical	QED	RF	0.7195 ± 0.0153	0.0950 ± 0.0023	0.1261 ± 0.0027
	SPS	XGBoost	0.4919 ± 0.0296	8.6857 ± 0.2277	12.4704 ± 0.4370
	$MolWt$	RF	0.9528 ± 0.0040	22.8881 ± 0.4422	34.1397 ± 1.1140
	$TPSA$	XGBoost	0.7019 ± 0.0200	25.1112 ± 0.4284	35.5259 ± 1.0285
	Phi	RF	0.8282 ± 0.0256	0.2955 ± 0.0067	0.4475 ± 0.0141
	$MolLogP$	XGBoost	0.4264 ± 0.0295	1.4864 ± 0.0366	2.1440 ± 0.0602
	$MolMR$	XGBoost	0.9671 ± 0.0027	5.1661 ± 0.1095	7.1888 ± 0.2586
	NP-likeness	XGBoost	0.4729 ± 0.0249	0.5341 ± 0.0128	0.7303 ± 0.0173
All Combined	QED	XGBoost	0.9487 ± 0.0039	0.0365 ± 0.0011	0.0539 ± 0.0021
	SPS	XGBoost	0.9851 ± 0.0027	1.2053 ± 0.0461	2.1244 ± 0.1708
	$MolWt$	RF	0.9997 ± 0.0002	0.4806 ± 0.0901	2.3109 ± 1.0532
	$TPSA$	XGBoost	0.9891 ± 0.0138	2.5812 ± 0.1656	5.9225 ± 3.6913
	Phi	XGBoost	0.9882 ± 0.0030	0.0531 ± 0.0023	0.1167 ± 0.0142
	$MolLogP$	SVR-RBF	0.9741 ± 0.0198	0.1869 ± 0.0114	0.4309 ± 0.1527
	$MolMR$	XGBoost	0.9975 ± 0.0003	1.2722 ± 0.0342	1.9952 ± 0.1042
	NP-likeness	SVR-RBF	0.9349 ± 0.0058	0.1710 ± 0.0056	0.2566 ± 0.0126

Table 11 Indices-only QSPR comparison under $5\times$ repeated 5-fold cross-validation. The final column reports $\Delta R^2 = R^2(\text{Sum-Conn.}) - R^2(\text{Classical})$

Target	R^2 **(Classical)**	R^2 **(Sum-Conn.)**	ΔR^2
QED	0.7195	0.7225	+0.0030
SPS	0.4919	0.5032	+0.0113
$MolWt$	0.9528	0.9528	+0.0000
$TPSA$	0.7019	0.7166	+0.0147
Phi	0.8282	0.8300	+0.0018
$MolLogP$	0.4264	0.4393	+0.0129
$MolMR$	0.9671	0.9687	+0.0016
NP-likeness	0.4729	0.4790	+0.0061
Mean across targets	0.6951	0.7015	+0.0064

These PCA results are useful for assessing descriptor redundancy, but they should be interpreted as a *sensitivity analysis* rather than as a replacement main result set. Unsupervised PCA reduces dimensionality and weakens some inflated within-space accuracies, yet it does not by itself remove constitution-driven target overlap when those relationships dominate the latent components; for example, even after PCA the **All Combined** $MolWt$ benchmark remains very high at $R^2 = 0.9819$. At the same time, PCA materially weakens the main QSAR benchmarks. We therefore retain the non-PCA descriptor-space results as the primary benchmark, use PCA as evidence that blind compression is not an adequate standalone remedy, and regard targeted proxy-descriptor removal together with correlation-pruned or regularized feature selection as the more informative next step.

4.6 QSAR modelling of antibacterial activity (pMIC)

The antibacterial-activity task is intrinsically more challenging than the physicochemical QSPR tasks because $pMIC$ is influenced by multiple partially overlapping determinants, including scaffold class, lipophilicity, polarity, hydrogen-bonding profile, and ring composition. We therefore benchmarked six feature families: **RDKit Descriptors**, **Classical Indices**, **Sum-Connectivity Indices**, **RDKit+Classical**, **RDKit+Sum-Connectivity**, and **All Combined**.

Table 13 reports the repeated cross-validation results for all feature families and learners. The strongest performance was obtained with **RDKit Descriptors** using the **Extra-Trees** regressor, which achieved $R^2 = 0.5959 \pm 0.0246$, MAE $= 0.5548 \pm 0.0151$, and RMSE $= 0.7389 \pm 0.0228$. Closely related performance was obtained by

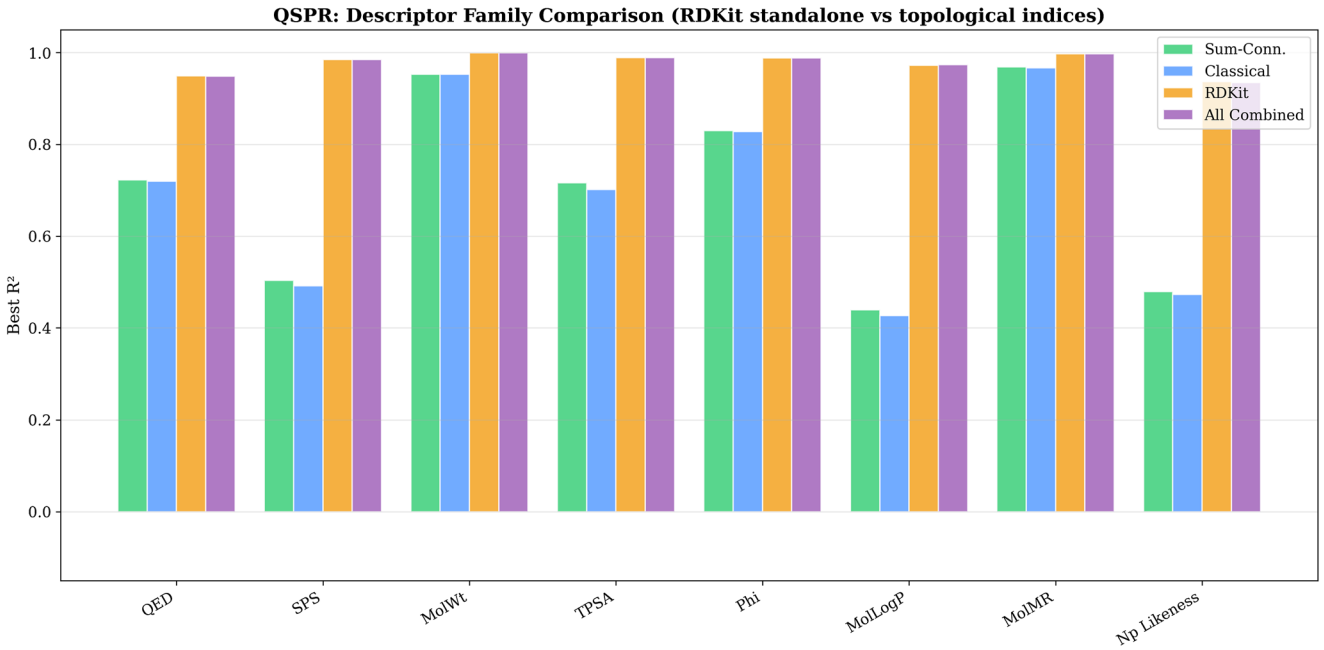

Fig. 2 Comparison of QSPR predictive performance across physicochemical targets for the descriptor families Classical Indices, Sum-Connectivity Indices, and All combined. In this figure, the All combined bar denotes the concatenated descriptor pool formed by RDKit, Classical, and Sum-Conn. descriptors and is included as a benchmark for the largest available feature set under a common workflow

RDKit+Sum-Conn., **RDKit+Classical**, and **All Combined**, but none of these exceeded the RDKit-only baseline. By contrast, the standalone topological descriptor families produced substantially lower predictive power, with the best **Classical Indices** model reaching $R^2 = 0.2767$ and the best **Sum-Connectivity Indices** model reaching $R^2 = 0.2887$. Thus, for antibacterial activity prediction on this chemically heterogeneous dataset, the dominant signal is carried by the RDKit representation, while the proposed index family provides only a modest improvement over classical topological descriptors in isolation.

Among the three tree-ensemble learners, ExtraTrees was the strongest overall model: it was the best learner for five of the six feature families under repeated cross-validation, while XGBoost produced the best standalone **Sum-Conn.** QSAR result. Random Forest was typically close behind. This pattern explains why the main QSAR discussion refers explicitly to tree ensembles rather than only to a single learner.

Because random cross-validation may still benefit from close analogue overlap, we next examined the same feature families under Murcko scaffold-based GroupKFold. As shown in Table 14, performance decreased across all feature families, as expected for a chemically stricter evaluation. Nevertheless, the relative ranking remained stable. Again, the strongest model was **ExtraTrees+RDKit Descriptors**, with $R^2 = 0.4799 \pm 0.0252$, followed closely by **RDKit+Classical**, **RDKit+Sum-Conn.**, and **All Combined**. The standalone topological families remained clearly weaker and were essentially tied, with only a marginal advantage for the proposed sum-connectivity family over the classical one (0.2025 vs. 0.2019 in best-model R^2). This scaffold-separated analysis therefore supports a moderate but chemically more realistic QSAR conclusion: the descriptor ranking is stable, RDKit remains strongest, and the proposed descriptors offer only modest standalone gains for the $pMIC$ task.

4.7 Diagnostic validation and model interpretation

The strongest QSAR configuration, **ExtraTrees+RDKit Descriptors**, was subjected to additional diagnostic validation. First, a Y-randomization test with 50 permutations was performed to assess whether the observed performance could arise from chance correlation. As summarized in Table 15, the real model achieved $R^2 = 0.5959$, whereas the permuted-label baseline yielded a mean $R^2 = -0.0356 \pm 0.0069$. The corresponding Z-score of 91.8 indicates clear separation from the null distribution, supporting a non-random statistical signal while not changing the conclusion that the absolute QSAR performance is moderate.

Figure 4 visualizes the same result, contrasting the real-label performance against the null distribution obtained from permuted responses.

For the same repeated-CV model, the mean Pearson correlation, Spearman correlation, and concordance correlation coefficient were 0.7761 ± 0.0174, 0.7448 ± 0.0188

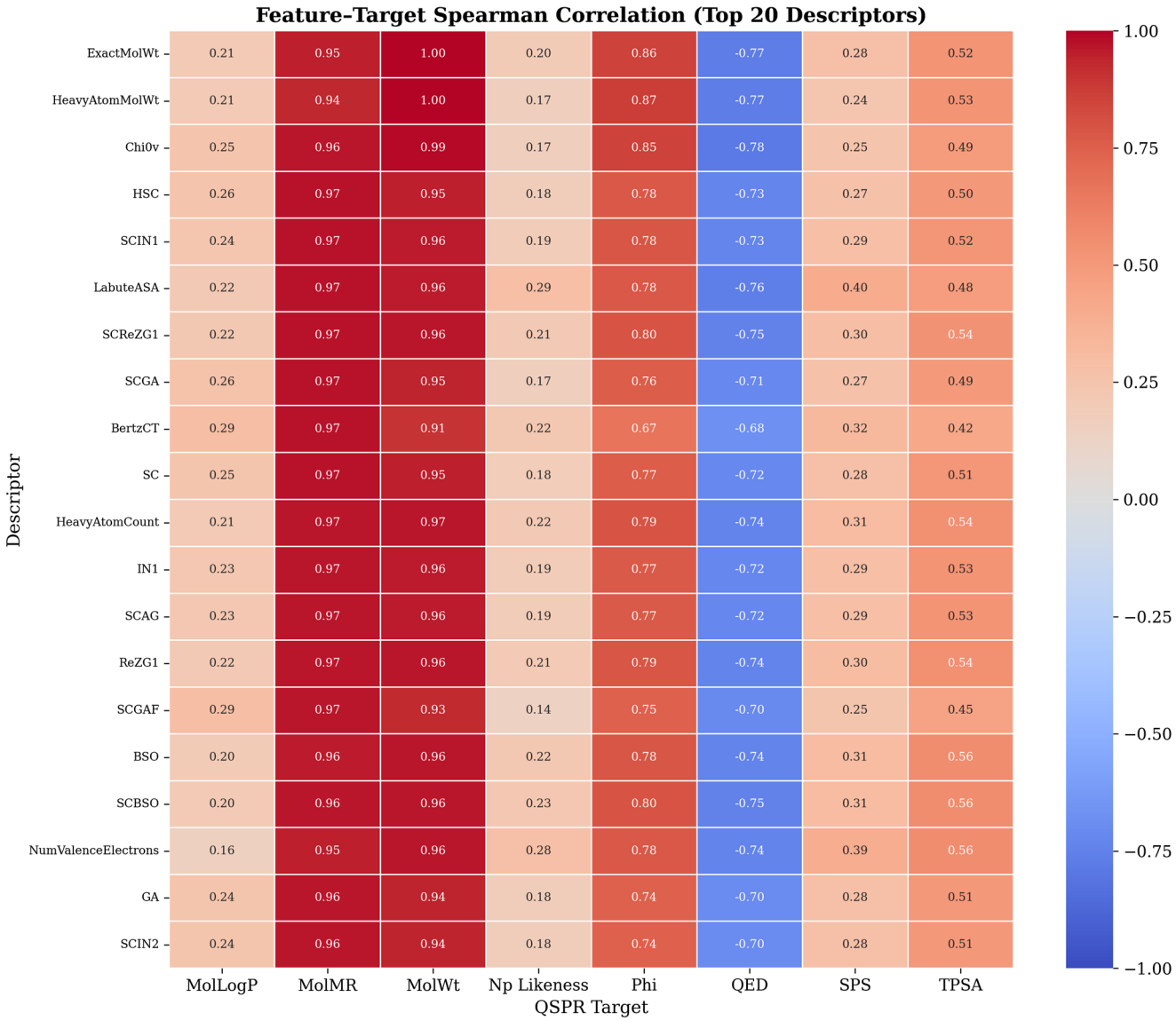

Fig. 3 Feature-target Spearman correlations for the most strongly associated descriptors across the eight QSPR endpoints. The figure highlights the strong overlap between some constitution-driven targets and closely related descriptors in the current feature pool

, and 0.7283 ± 0.0174, respectively. These values indicate a stable association between predicted and observed activity, but they are interpreted only as supportive diagnostics because the scaffold-based R^2 remains the stricter indicator of prospective chemical generalization.

Second, permutation importance was computed for the same model to identify the descriptors contributing most strongly to $pMIC$ prediction. Table 16 lists the top-ranked variables. The most influential features are dominated by RDKit descriptors related to heterocycle count, aliphatic ring content, charge-sensitive surface-area terms, and electronic-state measures. This pattern is chemically plausible, since antibacterial potency in a diverse small-molecule dataset is expected to depend on a combination of ring topology, heteroatom content, charge distribution, and lipophilicity-related surface descriptors rather than on degree-based topology alone.

The same information is presented visually in Fig. 5, which facilitates rapid comparison of effect magnitudes across the most relevant descriptors.

Figure 6 shows the predicted-versus-observed scatter plot for the best QSAR model. The predictions follow the identity trend well in the central activity range, with increased dispersion at the extremes, which is expected for a chemically heterogeneous antibacterial dataset with multiple mechanisms of action and substantial assay-level variability.

Table 12 Sensitivity analysis comparing the strongest non-PCA results with a PCA-compressed rerun of the same workflow. For the QSPR rows, the comparison uses the All Combined descriptor block; for the QSAR rows, the comparison reports the main RDKit and All Combined benchmarks under repeated and scaffold-based validation. The final column reports $\Delta R^2 = R^2_{\text{PCA}} - R^2_{\text{non-PCA}}$

Setting	Non-PCA best model	Non-PCA R^2	PCA best model	PCA R^2	ΔR^2
QSPR: QED	XGBoost	0.9487	XGBoost	0.8816	−0.0671
QSPR: SPS	XGBoost	0.9851	XGBoost	0.9465	−0.0386
QSPR: $MolWt$	RF	0.9997	XGBoost	0.9819	−0.0178
QSPR: $TPSA$	XGBoost	0.9891	XGBoost	0.9304	−0.0587
QSPR: Phi	XGBoost	0.9882	SVR-RBF	0.9482	−0.0400
QSPR: $MolLogP$	SVR-RBF	0.9741	SVR-RBF	0.9210	−0.0531
QSPR: $MolMR$	XGBoost	0.9975	XGBoost	0.9849	−0.0125
QSPR: NP-likeness	SVR-RBF	0.9349	XGBoost	0.8078	−0.1270
QSAR: RDKit (5×5 CV)	ExtraTrees	0.5959	ExtraTrees	0.4744	−0.1214
QSAR: All Combined (5×5 CV)	ExtraTrees	0.5938	ExtraTrees	0.4734	−0.1205
QSAR: RDKit (scaffold)	ExtraTrees	0.4799	ExtraTrees	0.4003	−0.0795
QSAR: All Combined (scaffold)	ExtraTrees	0.4767	ExtraTrees	0.3970	−0.0797

Table 13 Best QSAR performance for $pMIC$ under $5\times$ repeated 5-fold cross-validation. For each feature family, the table reports the model with the highest mean R^2

Feature family	Best model	R^2 (mean±std)	MAE (mean±std)	RMSE (mean±std)
RDKit Descriptors	ExtraTrees	0.5959 ± 0.0246	0.5548 ± 0.0151	0.7389 ± 0.0228
Classical Indices	ExtraTrees	0.2767 ± 0.0240	0.7736 ± 0.0137	0.9889 ± 0.0178
Sum-Conn. Indices	XGBoost	0.2887 ± 0.0222	0.7596 ± 0.0141	0.9806 ± 0.0173
RDKit+Classical	ExtraTrees	0.5945 ± 0.0243	0.5559 ± 0.0144	0.7402 ± 0.0225
RDKit+Sum-Conn.	ExtraTrees	0.5945 ± 0.0244	0.5555 ± 0.0148	0.7402 ± 0.0226
All Combined	ExtraTrees	0.5938 ± 0.0243	0.5562 ± 0.0142	0.7408 ± 0.0221

Table 14 Best QSAR performance for $pMIC$ under Murcko scaffold-based 5-fold GroupKFold cross-validation

Feature family	Best model	R^2 (mean±std)	MAE (mean±std)	RMSE (mean±std)
RDKit Descriptors	ExtraTrees	0.4799 ± 0.0252	0.6383 ± 0.0063	0.8366 ± 0.0065
Classical Indices	SVR-RBF	0.2019 ± 0.0365	0.8053 ± 0.0328	1.0367 ± 0.0289
Sum-Conn. Indices	SVR-RBF	0.2025 ± 0.0366	0.8053 ± 0.0309	1.0363 ± 0.0272
RDKit+Classical	ExtraTrees	0.4791 ± 0.0268	0.6389 ± 0.0049	0.8371 ± 0.0051
RDKit+Sum-Conn.	ExtraTrees	0.4779 ± 0.0261	0.6394 ± 0.0065	0.8381 ± 0.0054
All Combined	ExtraTrees	0.4767 ± 0.0278	0.6407 ± 0.0047	0.8390 ± 0.0058

Table 15 Y-randomization results for the best QSAR model (ExtraTrees+RDKit Descriptors)

Metric	Value
Real model R^2	0.5959
Permuted mean R^2	-0.0356 ± 0.0069
Z-score	91.8
Number of permutations	50
Conclusion	Model captures non-random signal

Finally, Fig. 7 presents the learning curve for the best QSAR configuration. The monotonic improvement with increasing training size indicates that the model continues to benefit from additional data and has not yet fully saturated, suggesting that further expansion of the curated antibacterial dataset may still improve predictive accuracy.

5 Discussion

This study provides a rigorous empirical evaluation of the proposed sum-connectivity descriptors in molecular modelling. Their clearest empirical advantage appears in the matched standalone QSPR benchmarks, where the sum-connectivity family consistently outperforms the corresponding classical degree-based indices across all eight physicochemical endpoints under repeated cross-validation. The effect size, however, is modest rather than transformative, so the empirical claim is systematic improvement within a controlled descriptor benchmark rather than broad superiority over richer cheminformatics representations.

Although the **All Combined** block attains the highest R^2 values for every property, the new correlation analyses

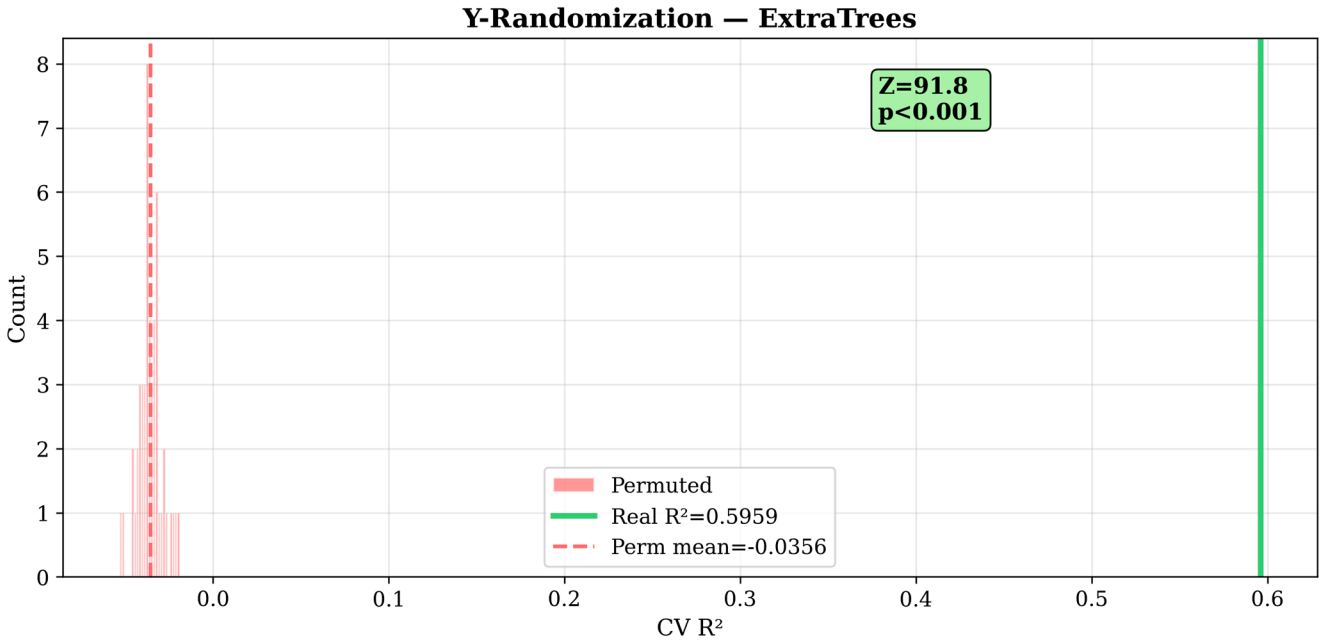

Fig. 4 Y-randomization test for the best QSAR model. The real-model R^2 is clearly separated from the null distribution obtained from permuted response values

Table 16 Top 15 permutation-importance features for the best QSAR model (ExtraTrees+RDKit Descriptors). Importance is reported as mean decrease in R^2 under permutation

Rank	Feature	Importance (ΔR^2)	± Std
1	NumSaturatedHeterocycles	0.069173	0.006137
2	NumAliphaticRings	0.030292	0.003626
3	NumHeterocycles	0.029609	0.003489
4	PEOE_VSA3	0.022402	0.003178
5	SlogP_VSA3	0.021686	0.002317
6	SlogP_VSA6	0.021251	0.002278
7	VSA_EState2	0.020933	0.002398
8	MaxAbsEStateIndex	0.018277	0.001832
9	PEOE_VSA13	0.018130	0.002166
10	MaxEStateIndex	0.017831	0.002118
11	NumHDonors	0.015014	0.002791
12	NHOHCount	0.013413	0.001749
13	MinAbsPartialCharge	0.012694	0.001778
14	PEOE_VSA1	0.012200	0.001485
15	MinPartialCharge	0.011231	0.001719

show that some endpoints are strongly aligned with closely related constitutional descriptors already present in the feature set. For example, $MolWt$ is almost directly encoded by $ExactMolWt$ ($\rho = 0.99999$) and $HeavyAtomMolWt$ ($\rho = 0.99787$), while $TPSA$ is strongly associated with $NOCount$ ($\rho = 0.9304$). In addition, each classical descriptor is almost collinear with its sum-connectivity counterpart ($\rho = 0.9967$–0.9986). Consequently, the near-unity QSPR scores for several combined models should be interpreted cautiously as upper-bound performance within the present descriptor space, not as direct evidence that all descriptor blocks contribute independent mechanistic information.

For antibacterial activity, the final QSAR results are moderate rather than strong. The best repeated-cross-validation model reached $R^2 = 0.5959$, and the stricter Murcko scaffold evaluation reached $R^2 = 0.4799$. These values do not support claims of excellent predictive performance, but they do show reproducible signal in a large and chemically heterogeneous small-molecule dataset under a more realistic validation regime than a single random split.

Among the non-linear regressors, ExtraTrees was the strongest overall learner, Random Forest was usually close behind, and XGBoost remained competitive, especially for the standalone Sum-Conn. QSAR benchmark. Together with the Y-randomization, permutation-importance, and learning-curve analyses, this indicates that the activity model is capturing chemically meaningful structure-activity trends rather than chance correlations. The dominant $pMIC$ signal, however, is carried by RDKit descriptors, while the proposed graph descriptors provide only modest standalone gains over the classical family.

The modelling analysis therefore supports a nuanced conclusion. The proposed sum-connectivity descriptors are most convincing as systematically improved graph-topological descriptors in matched standalone comparisons. Their added value becomes much less distinct once large RDKit descriptor blocks are already available. In the present study, a separate PCA-compressed rerun was used as an unsupervised decorrelation test, and it consistently reduced the strongest **All Combined** QSPR scores as well as the main

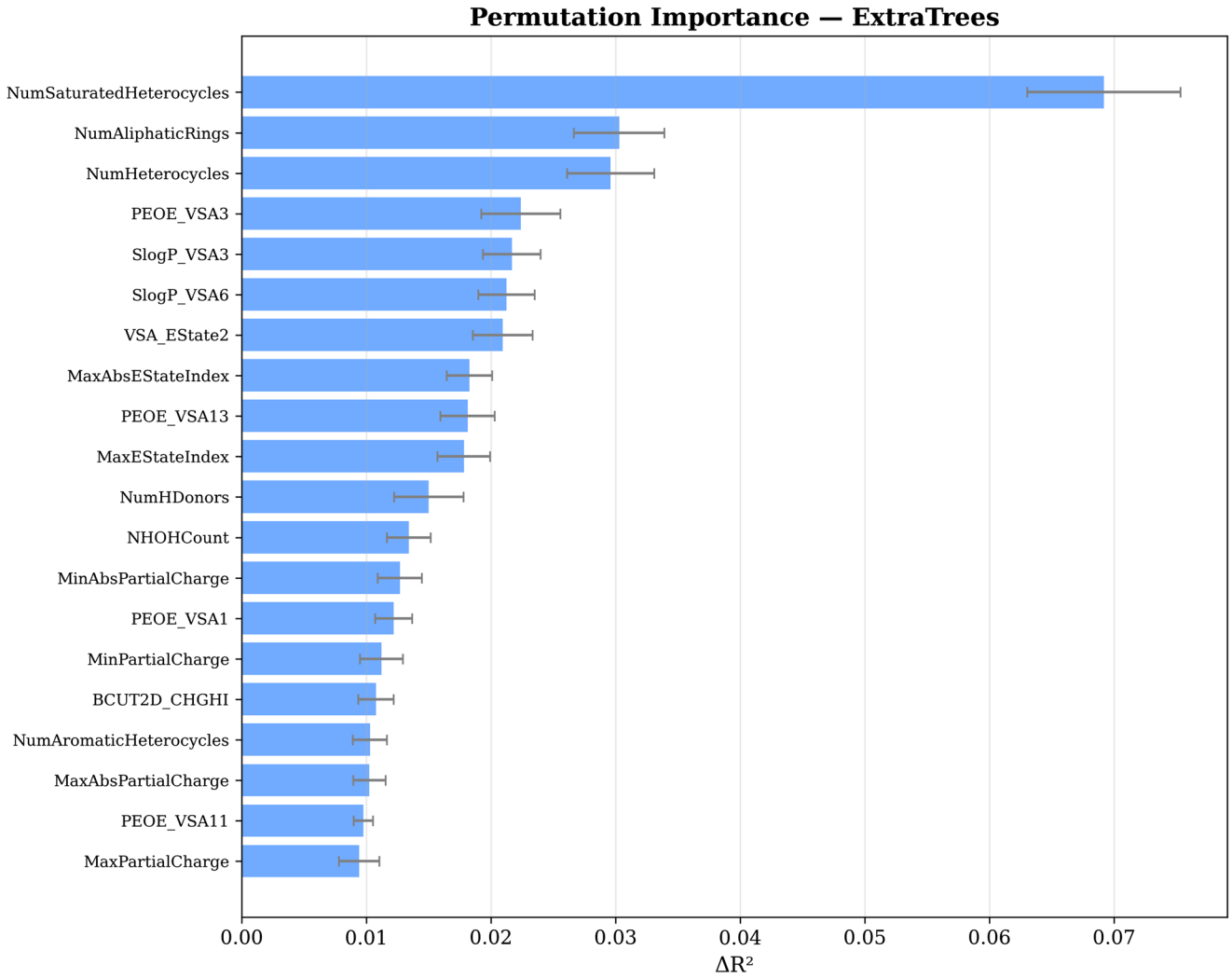

Fig. 5 Permutation importance for the best QSAR model (ExtraTrees+RDKit Descriptors), showing the top-ranked descriptors contributing to $pMIC$ prediction

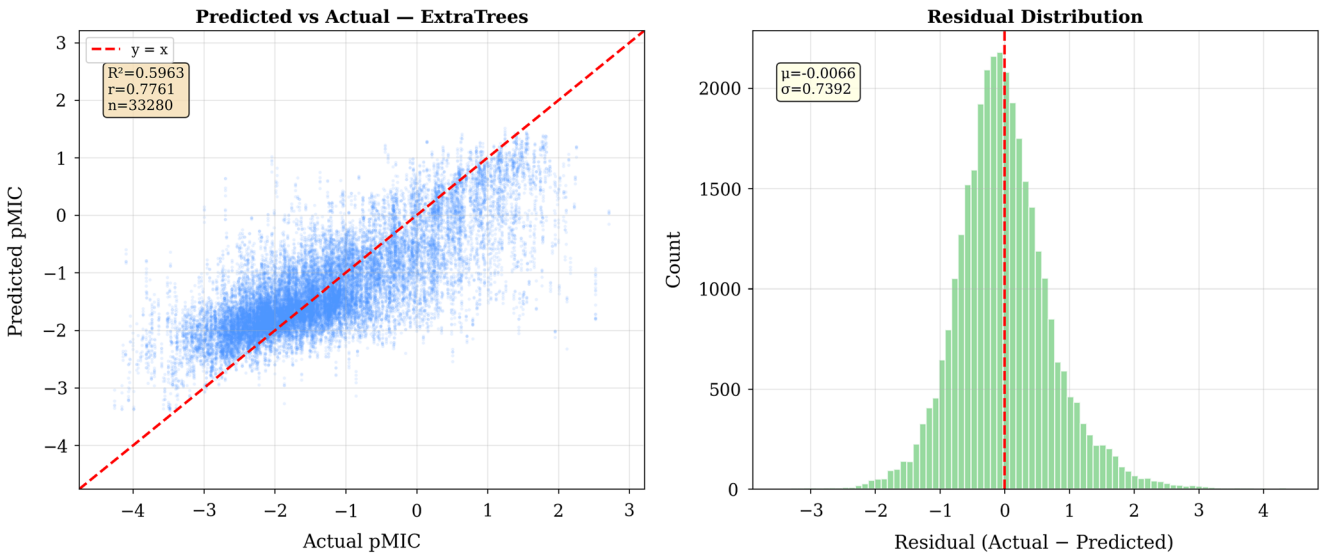

Fig. 6 Predicted versus observed $pMIC$ values for the best QSAR model (ExtraTrees+RDKit Descriptors)

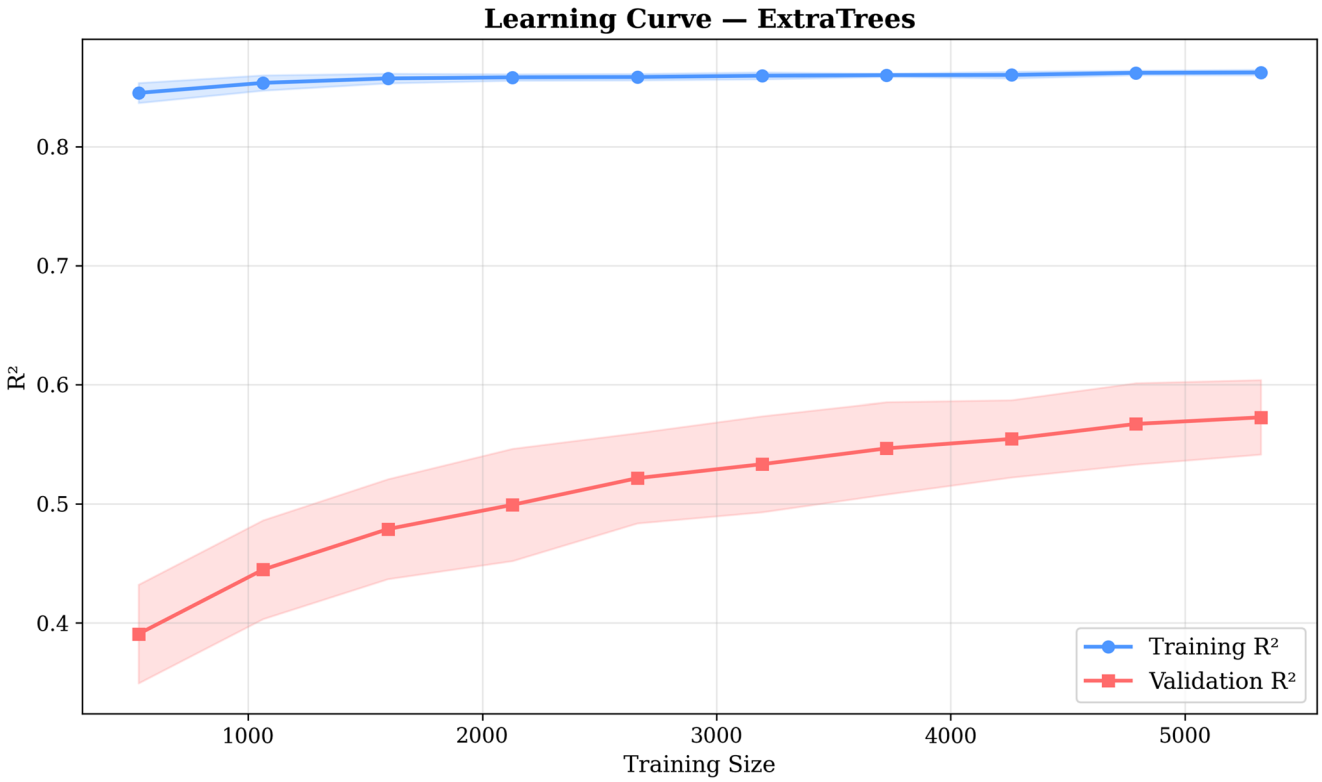

Fig. 7 Learning curve for the best QSAR model (ExtraTrees + RDKit Descriptors), showing the dependence of predictive performance on training-set size

QSAR benchmarks. This directly shows how the reported results behave under dimensionality reduction, but it also shows that PCA alone is not a satisfactory fix: it weakens predictive performance while constitution-driven targets remain comparatively easy to encode. The more informative follow-up is therefore targeted removal of target-proxy descriptors together with correlation-pruned or regularized feature selection, rather than blind compression of the full descriptor space.

Table 17 places the QSAR results in the context of recent studies on **E. coli** MIC-related prediction. Many published models were developed on smaller or chemically narrower datasets, particularly peptide-focused collections, and often under validation settings that differ substantially from those used here. In the present study, the best repeated-cross-validation result reached $R^2 \approx 0.596$, while the more stringent Murcko scaffold-based validation yielded $R^2 \approx 0.480$. Although these values are not strictly directly comparable to the literature because of differences in endpoint definition, dataset composition, assay standardization, and evaluation protocol, they indicate competitive, but not exceptional, predictive performance for a large and chemically heterogeneous small-molecule dataset.

5.1 Limitations of the current modelling study

The present modelling conclusions should be interpreted together with several limitations. First, the highest **All Combined** QSPR scores still involve constitution-driven targets for which descriptor-target overlap is difficult to exclude completely without explicitly removing proxy descriptors. Second, the PCA rerun reported here is informative as a dimensionality-reduction sensitivity analysis, but it is not a definitive remedy for leakage because unsupervised compression can preserve dominant target-related variance while simultaneously weakening useful predictive structure. Third, the antibacterial *E. coli* dataset is large and carefully standardized at the MIC level, yet it remains chemically heterogeneous and mechanism-of-action heterogeneous, which naturally limits the attainable QSAR accuracy. These points are why the non-PCA results are presented as the main benchmark, but the claims are framed conservatively and the next recommended step is targeted proxy removal under the same rigorous validation workflow.

Table 17 Comparison of published QSAR models for predicting MIC-related activity against Escherichia coli. Reported R^2 values are taken as stated in the corresponding references; evaluation protocols, dataset composition, and endpoint definitions differ across studies, so the values are not strictly directly comparable. For the present study, the reported value corresponds to $5\times$ repeated 5-fold cross-validation

Target	Type of agent	Dataset size	$\mathbf{R^2}$	Ref.
E. coli K-12	Conjugated oligoelectrolytes	136	0.65	[58]
E. coli	Peptides	3929	0.59	[59]
E. coli	Peptides	3536	0.588	[47]
E. coli	Peptides	3818	0.548	[60]
E. coli ATCC 25,922	Peptides	3818	0.472	[60]
E. coli	Heterogeneous small molecules	6657	0.596	This study

6 Conclusion

This work developed a unified sum-connectivity reweighting framework for classical degree-based molecular descriptors. The theoretical contribution is the systematic transformation of established edge-degree kernels into normalized descriptors with explicit comparison relations, degree-extremal bounds, regular-graph proportionality, and complement-graph formulations. These results provide a coherent mathematical basis for interpreting the proposed descriptors as connectivity-normalized counterparts of classical indices.

The empirical findings show that, in the matched indices-only QSPR benchmarks, the proposed sum-connectivity descriptors consistently outperformed their classical degree-based counterparts across the eight physicochemical targets, yielding systematic but moderate improvements. In the larger combined descriptor space, the highest QSPR scores were obtained when RDKit descriptors were included; however, some targets are closely related to constitutional RDKit descriptors, so these values should be interpreted as within-study predictive accuracy rather than as independent mechanistic evidence. Quantitatively, the indices-only comparison increased the average QSPR R^2 from 0.6951 to 0.7015, whereas the strongest **All Combined** QSPR models achieved R^2 values from 0.9349 for NP-likeness to 0.9997 for $MolWt$; these larger values are therefore reported as endpoint-specific within-study benchmarks rather than as evidence of independent descriptor causality.

For antibacterial activity prediction, RDKit descriptors carried the dominant predictive signal. The best QSAR configuration, **ExtraTrees+RDKit Descriptors**, achieved $R^2 = 0.5959$ under $5\times$ repeated 5-fold cross-validation and $R^2 = 0.4799$ under Murcko scaffold validation. These values indicate moderate, chemically non-random predictability, not strong standalone activity prediction. The sum-connectivity descriptors showed a small standalone advantage over classical topological indices for $pMIC$, but their contribution became less distinct when richer RDKit descriptor blocks were already present.

The main limitations are the strong redundancy between paired classical and sum-connectivity descriptors, possible target-proxy effects in some physicochemical endpoints, and the chemical and mechanism-of-action heterogeneity of the antibacterial dataset. Future work should therefore combine the present descriptor family with targeted removal of proxy descriptors, correlation-pruned or regularized feature selection, and external validation on additional biological targets and chemical domains. On the mathematical side, extensions to χ_α-weighted families remain an open direction for studying how the exponent in $d_u + d_v$ affects extremal behaviour and descriptor discrimination.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1007/s44442-026-00105-6.

Author contributions Conceptualization: A.A., M.A., Z.A.; Methodology: A.A., Z.A.; Validation: A.A., M.A.; Investigation: M.A., Z.A.; Resources: A.A., M.A.; Writing-original draft: M.A., A.A.; Writing-review and editing: Z.A. All authors have read and agreed to the published version of the manuscript.

Funding No funding is available for this study.

Data availability The original bioactivity records analysed in this work are publicly available from the ChEMBL database (https://www.ebi.ac.uk/chembl/). The curated dataset, descriptor matrices, and Python scripts used for RDKit descriptor calculation, classical and sum-connectivity index calculation, model fitting, cross-validation, scaffold validation, and diagnostic analyses are supplied as Supplementary Materials to support reproducibility. A permanent archived version of the full reproducibility package has been deposited in Zenodo and is available at https://doi.org/10.5281/zenodo.21210043.

Declarations

Ethical approval Not applicable.

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

References

1. Chang CM, Banerjee A, Kumar V, Roy K, Benfenati E (2025) The q-RASPR approach for predicting the property and fate of persistent organic pollutants. Sci Rep 15(1):1344. https://doi.org/10.1038/s41598-024-84778-2
2. Cronin MTD, Basiri H, Chrysochoou G, Enoch, Enoch SJ, Firman JW, Spînu N, Madden JC (2025) The predictivity of QSARs for toxicity: recommendations for improving model performance. Comput Toxicol 33:100338https://doi.org/10.1016/j.comtox.2024.100338
3. Xin AW, Rivera-Delgado E, von Recum HA (2024) Using QSAR to predict polymer-drug interactions for drug delivery. Front Soft Matter 4. https://doi.org/10.3389/frsfm.2024.1402702
4. Moussaoui M, Baammi S, Soufi H, Baassi, Baassi M, El Allali A, Belghiti ME, Daoud R, Belaaouad S (2024) QSAR, ADMET, molecular docking, and dynamics studies of 1,2,4-triazine-3(2H)-one derivatives as tubulin inhibitors for breast cancer therapy. Sci Rep 14(1):16418. https://doi.org/10.1038/s41598-024-66877-2
5. Kier LB, Hall LH (1986) Molecular connectivity in structure-activity analysis. Wiley, New York
6. Randić M (1975) On characterization of molecular branching. J Am Chem Soc 97(23):6609–6615. https://doi.org/10.1021/ja00856a001
7. Garcia-Domenech R, Gálvez J, de Julián-Ortiz JV, Pogliani L (2008) Some new trends in chemical graph theory. Chem Rev 108(3):1127–1169. https://doi.org/10.1021/cr0780006
8. Wagner S, Wang H (2018) Introduction to chemical graph theory. CRC Press, Boca Raton
9. Alameri A, Alsharafi M (2021) Topological indices types in graphs and their applications. Generis Publishing, Chisinau, Moldova
10. Gutman I (2013) Degree-based topological indices. Croat Chem Acta 86(4):351–361. https://doi.org/10.5562/cca2294
11. Bondy JA, Murty USR (2008) Graph theory. Springer, London
12. Ranjini PS, Lokesha V, Usha A (2013) Relation between phenylene and hexagonal squeeze using harmonic index. Int J Graph Theory 1(4):53–67
13. Kulli VR (2018) Degree based connectivity F-Indices of nanotubes 18(2):201–206. https://doi.org/10.22457/apam.v18n2a10
14. Gutman I, Furtula B, Oz MS (2024) Geometric approach to vertex-degree-based topological indices–Elliptic Sombor index, theory and application. Int J Quantum Chem 124(2):e27346. https://doi.org/10.1002/qua.27346
15. Gutman I (2024) Relations between Sombor and Euler indices. Vojnoteh Glas 72(1):1–12. https://doi.org/10.5937/vojtehg72-48818
16. Albertson MO (1997) The irregularity of a graph. Ars Combin 46:219–225
17. V R K (2021) On Banhatti-Sombor indices. SSRG Int J Appl Chem 8(1):21–25. https://doi.org/10.14445/23939133/IJAC-V8I1P105
18. Vukičević D, Furtula B (2009) Topological index based on the ratios of geometrical and arithmetical means of end-vertex degrees of edges. J Math Chem 46(4):1369–1376. https://doi.org/10.1007/s10910-009-9520-x
19. Zhou B, Trinajstić N (2009) On a novel connectivity index. J Math Chem 46(4):1252–1270. https://doi.org/10.1007/s10910-008-9515-z
20. Furtula B, Gutman I (2015) A forgotten topological index. J Math Chem 53(4):1184–1190. https://doi.org/10.1007/s10910-015-0480-z
21. Kulli VR, Lokesha V, Nirupadi K (2021) Computation of inverse Nirmala indices of certain nanostructures. Int J Math Combin 2:32–39
22. Alameri A, Alsharafi M, Ali E, Gumaan MS (2021) A note on topological indices and coindices of disjunction and symmetric difference of graphs. Discrete Appl Math 304:230–235. https://doi.org/10.1016/j.dam.2021.07.037
23. Alsharafi M, Zeren Y, Alameri A (2021) The second hyper-Zagreb index of complement graphs and its applications of some nano structures. Asian J Probab Stat 15(4). https://doi.org/10.9734/ajpas/2021/v15i430364
24. Alsharafi M, Alameri A (2021) The F-index and coindex of V-phenylenic nanotubes and nanotorus and their molecular complement graphs. Nanosyst Phys Chem Math 12(3):263–270. https://doi.org/10.17586/2220-8054-2021-12-3-263-270
25. Alameri AQS, Al-Sharafi MSY (2021) Topological indices of some new graph operations and their possible applications. Asian J Probab Stat 13(2):16–30. https://doi.org/10.9734/ajpas/2021/v13i230302
26. Li X, Gutman I (2006) Mathematical aspects of Randić-type molecular structure descriptors. Univ. Kragujevac, Kragujevac
27. Gutman I, Furtula B (Eds.) (2008) Recent results in the theory of Randić index. Univ. Kragujevac, Kragujevac
28. Zhou B, Trinajstić N (2010) On general sum-connectivity index. J Math Chem 47(1):210–218. https://doi.org/10.1007/s10910-009-9542-4
29. Du Z, Zhou B, Trinajstić N (2010) Minimum general sum-connectivity index of unicyclic graphs. J Math Chem 48(3):697–703. https://doi.org/10.1007/s10910-010-9702-6
30. Du Z, Zhou B, Trinajstić N (2011) On the general sum-connectivity index of trees. Appl Math Lett 24(3):402–405. https://doi.org/10.1016/j.aml.2010.10.038
31. Tomescu I, Arshad M (2015) On the general sum-connectivity index of connected unicyclic graphs with k pendant vertices. Discrete Appl Math 181:306–309
32. Tomescu I, Jamil MK (2014) Maximum general sum-connectivity index for trees with given independence number. MATCH Commun Math Comput Chem 72:715–722
33. Lučić B, Trinajstić N, Zhou B (2009) Comparison between the sum-connectivity index and product-connectivity index for benzenoid hydrocarbons. Chem Phys Lett 475(1–3):146–148. https://doi.org/10.1016/j.cplett.2009.05.022
34. Bollobás B, Erdős P (1998) Graphs of extremal weights, *Ars Combin* 50:225–233
35. Chu YM, Julietraja K, Venugopal P, Siddiqui MK, Prabhu S (2021) Degree- and irregularity-based molecular descriptors for benzenoid systems. Eur Phys J Plus 136(1):78. https://doi.org/10.1140/epjp/s13360-020-01033-z
36. Yang J, Konsalraj J, Raja S AA (2022) Neighbourhood sum degree-based indices and entropy measures for certain family of graphene molecules. Molecules 28(1):168. https://doi.org/10.3390/molecules28010168
37. Anuradha DS, Julietraja K, Jaganathan B, Alsinai A (2024) Curcumin-conjugated PAMAM dendrimers of two generations: comparative analysis of physiochemical properties using Adriatic topological indices. ACS Omega 9(12):14558–14579. https://doi.org/10.1021/acsomega.4c00686
38. Julietraja K, Venugopal P (2022) Computation of degree-based topological descriptors using M-polynomial for coronoid systems. Polycyclic Aromatic Compd 42(4):1770–1793. https://doi.org/10.1080/10406638.2020.1804415
39. Julietraja K, Susai AAR, Alsinai A (2025) Neighbourhood sum-based structural analysis for sodalite system. Curr Org Synth 22(4):548–555. https://doi.org/10.2174/0115701794320235240926054912
40. Estrada E, Torres L, Rodríguez L, Gutman I (1998) An atom-bond connectivity index: modelling the enthalpy of formation of alkanes. Indian J Chem Sec A 37:849–855

41. Ali A, Furtula B, Redžepović I, Gutman I (2022) Atom-bond sum-connectivity index. J Math Chem 60(10):2081–2093. https://doi.org/10.1007/s10910-022-01403-1
42. Ali A, Gutman I, Redžepović I (2023) Atom-bond sum-connectivity index of unicyclic graphs and some applications. Electron J Math 5:1–7
43. Ali A, Das KC, Dimitrov D, Furtula B (2021) Atom-bond connectivity index of graphs: a review over extremal results and bounds. Discrete Math. Lett 5:68–93
44. Kadeřábková N, Mahmood AJS, Mavridou DAI (2024) Antibiotic susceptibility testing using minimum inhibitory concentration (MIC) assays. Npj Antimicrob Resist 2(1):37. https://doi.org/10.1038/s44259-024-00051-6
45. Hossain TJ (2024) Methods for screening and evaluation of antimicrobial activity: a review of protocols, advantages, and limitations. Eur J Microbiol Immunol 14(2):97–115. https://doi.org/10.1556/1886.2024.00035
46. CLSI & EUCAST (2025) Release New guidance on modifications to antimicrobial susceptibility testing. (n.d.). https://clsi.org/about/news/clsi-and-eucast-release-new-guidance-on-modifications-to-antimicrobial-susceptibility-testing-methods/
47. Yan J, Zhang B, Zhou M, Campbell-Valois F-X, Siu SWI (2023) A deep learning method for predicting the minimum inhibitory concentration of antimicrobial peptides against Escherichia coli using multi-branch-CNN and attention. Msystems 8(4):e00345–23. https://doi.org/10.1128/msystems.00345-23
48. Nasrollahian S, Graham JP, Halaji M (2024) A review of the mechanisms that confer antibiotic resistance in pathotypes of E. coli. Front Cell Infect Microbiol 14: 1387497.https://doi.org/10.3389/fcimb.2024.1387497
49. Zdrazil B, Felix E, Hunter F, Manners EJ, Blackshaw J, Corbett S, de Veij M, Ioannidis H, Lopez DM, Mosquera J, Magarinos M, Bosc N, Arcila R, Kizilören T, Gaulton A, Bento A, Adasme M, Monecke P, Landrum G, Leach A (2024) The ChEMBL database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. Nucleic Acids Res 52(D1):D1180–D1192. https://doi.org/10.1093/nar/gkad1004
50. Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction, 2nd edn. Springer, New York
51. Geladi P, Kowalski BR (1986) Partial least-squares regression: a tutorial. Anal Chim Acta 185:1–17. https://doi.org/10.1016/0003-2670(86)80028-9
52. Hoerl AE, Kennard RW (1970) Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12(1):55–67. https://doi.org/10.1080/00401706.1970.10488634
53. Drucker H, Burges CJC, Kaufman L, Smola A, Vapnik V (1997) Support vector regression machines. In: Advances in neural information processing systems 9. MIT Press, pp 155–161
54. Smola AJ, Schölkopf B (2004) A tutorial on support vector regression. Stat Comput 14(3):199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88
55. Breiman L (2001) Random forests. Mach Learn 45(1):5–32. https://doi.org/10.1023/A:1010933404324
56. Geurts P, Ernst D, Wehenkel L (2006) Extremely randomized trees. Mach Learn 63(1):3–42. https://doi.org/10.1007/s10994-006-6226-1
57. Chen T, Guestrin C (2016) Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 785–794
58. Tiihonen A, Santos HA, Manzanera JA (2021) Predicting antimicrobial activity of conjugated oligoelectrolytes with machine learning. Cell Rep Phys Sci 2(6):100475
59. Kumar V, Singh A, Gupta A, Sharma P (2025) Prediction of inhibitory peptides against E. coli with desired minimum inhibitory concentrations. Front Microbiol 16:1180598
60. Chung CR, Chien CY, Tang Y, Wu L-C, Hsu JB-K, Lu J-J, Lee T-Y, Bai C, Horng J-T (2024) An ensemble deep learning model for predicting minimum inhibitory concentrations of antimicrobial peptides against pathogenic bacteria. Iscience 27(9):110718. https://doi.org/10.1016/j.isci.2024.110718

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.