**Received:** xxx **Accepted:** xxx **Published:** xxx.
DOI. xxxxxxx
Volume xxx, Issue xxx, p.p. 1-30: xxx
**Research Article**

Open Access

**Control and Optimization in Applied Mathematics - COAM**

# Hyperparameter Optimization of SVR-Based Machine Learning Models via Graph-Theoretical Topological Indices for Predicting Physicochemical Properties of Anti-Anxiety Drugs

**Negar Kheirkhahan** **Masoud Ghods**

Department of Mathematics, Statistics, and Computer Science, Semnan University, Semnan 35131-19111, Iran

**Correspondence:**
Masoud Ghods
**E-mail:**
mghods@semnan.ac.ir

**How to Cite**
Kheirkhahan, N., Ghods, M. (2027). "Hyperparameter optimization of SVR-based machine learning models via graph-theoretical topological indices for predicting physicochemical properties of anti-anxiety drugs". *Control and Optimization in Applied Mathematics*, 12(x), 1-30. https://doi.org/10.30473/coam.xxxxxxxxx

**Abstract.** This research evaluates the integration of graph-theoretic topological indices (TIs) with machine learning (ML) frameworks to forecast the physicochemical attributes of anxiolytic drugs. By representing molecular structures as graphs, the study extracted TIs to serve as primary features for four distinct predictive algorithms: basic and optimized support vector regression (SVR-Basic and SVR-Tuned), random forest (RF), and linear regression (LR). To address the constraints of a small sample size, the authors utilized Leave-One-Out cross-validation (LOOCV) and bootstrap resampling to ensure robust performance metrics, including confidence intervals and coefficients of variation (CV%) for stability assessment. The findings indicate that the combination of hyperparameter refinement and rigorous validation significantly elevates the precision and reliability of ML models in chemical property prediction.

**Keywords.** Topological indices, Anti-anxiety drugs, Optimization, SVR-tuned, Machine learning

**MSC.** 05C90; 92C55; 68T07.

## 1 Introduction

Anxiety is a universal human experience; virtually all individuals encounter, at some point, concerns pertaining to health, finances, or interpersonal relationships. Anxiety disorders, however, extend well beyond transient worry or situational fear. They are characterized by persistent, often debilitating symptoms that compromise daily functioning, academic achievement, and social engagement [3]. Clinically, anxiety disorders constitute a heterogeneous group of psychopathological conditions encompassing generalized anxiety disorder, panic disorder, social anxiety disorder, and specific phobia-related syndromes. Despite sustained investigative effort, the etiopathogenesis of these conditions remains incompletely elucidated; current evidence implicates genetic predisposition and adverse life experiences as significant contributing factors [8]. In clinical practice, management typically involves a combination of psychological interventions and pharmacological treatments — including anxiolytic agents — directed at reducing symptom severity and restoring quality of life. The development and optimization of therapeutic strategies for anxiety disorders therefore represents a scientifically and clinically important research objective.

Recent advances in computational science, particularly quantitative structure–property relationship (QSPR) modeling, have rendered this framework an indispensable instrument in modern drug discovery and development. These developments underscore the growing integration of graph theory into chemistry — an inherently interdisciplinary discipline concerned with the relationships among molecular structure, physicochemical properties, and biological activity. A foundational construct in this domain is the *molecular graph*, in which atoms and chemical bonds are represented as vertices and edges, respectively. Chemical graph theory, through the deployment of topological indices (TIs), provides a rigorous and computationally tractable basis for characterizing molecular architecture [17]. Such indices encode structural information as graph-derived numerical descriptors [6, 15, 22]. Recent work has broadened the scope of classical graph-theoretic applications: irregular face coloring and graph optimization techniques have demonstrated utility in the structural analysis of molecules [25], while hybrid models coupling convolutional neural networks with support vector machines (CNN–SVM) have yielded reliable predictions of complex biological and chemical properties [16]. Temperature-based topological indices — including the inverse sum and harmonic temperature indices — have been applied to diverse nanostructures, among them $HC_5C_5[p, q]$ nanotubes, and have proven effective for molecular property prediction [13, 14, 19].

A substantial body of recent literature has examined the role of topological indices in QSPR and quantitative structure–activity relationship (QSAR) modeling. Investigations have addressed eigenvalues and Zagreb-based descriptors, derived bounds for Sombor-type indices, and introduced novel descriptors for nanostructures including $TiO_2$ nanotubes and glass-based molecular graphs [11, 12, 20, 21]. Within QSPR frameworks, these descriptors have been em-

ployed to predict the physicochemical properties of anticancer compounds using a range of machine learning models, including LR, SVR, and RF [27]. Taken together, this body of work attests to the broad applicability of topological indices across multiple scientific disciplines.

QSPR modeling establishes correlations between TIs and molecular physicochemical properties through regression techniques that relate structural features to physical or chemical behavior. The closely related QSAR framework exploits analogous descriptor sets to evaluate and predict drug activity and efficacy [7, 9]. Both paradigms have been applied extensively to predict physicochemical properties across a wide chemical space, encompassing potent anticancer agents, therapeutic candidates targeting the Omicron variant of COVID-19, breast cancer treatments, entropy characterization of benzene derivatives, nanotube structures, and thermal indicators. The present study emphasizes the central role of topological indices in QSPR-driven drug development and their utility in chemical structure analysis [23, 24]. Zhang et al. (2023) similarly examined graph-based descriptors within a QSPR framework, focusing on the physicochemical properties of antipsychotic drugs [30]. Temperature-based TIs have also been applied in QSPR analyses of anticancer compounds, illuminating the influence of structural features on molecular stability and reactivity [26, 29]. Their application to COVID-19 treatments — particularly those targeting the Omicron variant — further corroborates the predictive capacity of graph-theoretical approaches in drug design [10, 20].

Notwithstanding the widespread use of topological indices in QSPR/QSAR modeling, temperature-based descriptors have received comparatively limited attention in the context of anti-anxiety drugs [1, 2, 18, 28]. The present work addresses this gap by deploying such indices within a QSPR framework to investigate the physicochemical properties of anti-anxiety compounds, with model performance systematically improved through hyperparameter tuning and rigorous validation strategies. The complete methodological workflow is illustrated in Figure 1.

Each compound's molecular structure is cast within a graph-theoretical formalism in which atoms are modeled as *vertices* and chemical bonds or intermolecular interactions are encoded as *edges*. The resulting molecular networks are treated as undirected and fully connected graphs. Vertex degree — defined as the number of edges incident to a given vertex — serves as a local measure of connectivity within the molecular network.

Here, each compound's molecular structure is recast within a graph-theoretical formalism. In this abstraction, atoms are modeled as *vertices*, while chemical bonds or intermolecular interactions are encoded as *edges*. The constructed molecular networks are considered undirected and fully connected. Vertex degree, defined as the total number of edges adjacent to a vertex, provides an index of the node's connectivity within the overall molecular network.

The remainder of this paper is organized as follows. Section 2 presents the analytical evaluation of the molecular topological indices employed, including the justification for selecting temperature-based descriptors and the detailed computation of all eleven indices for the 15 anti-anxiety compounds. Section 3 describes the methodology, covering the four regression models

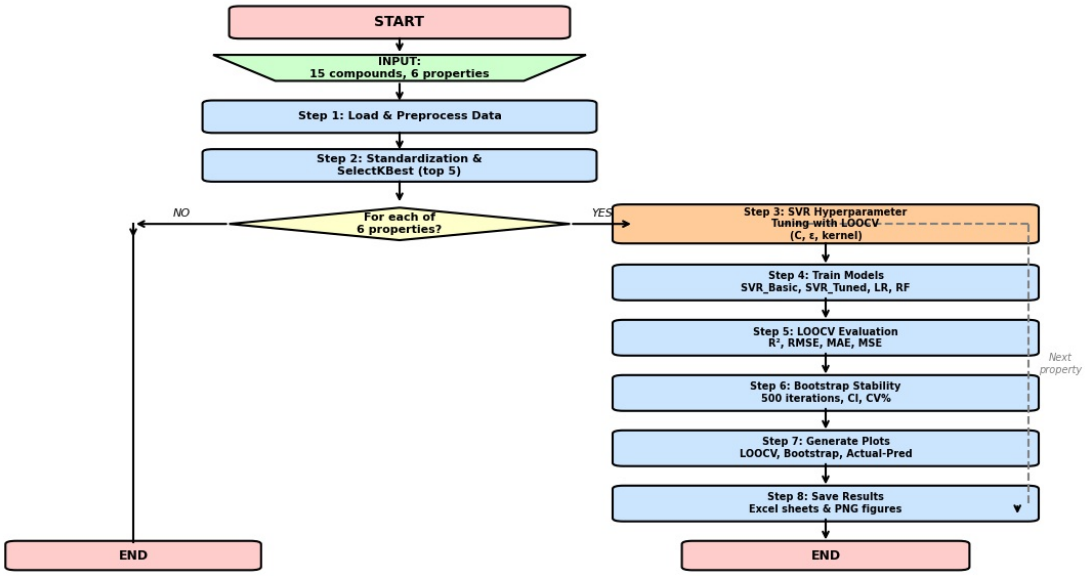



**Figure 1:** Workflow diagram of the methodology employed in this study.

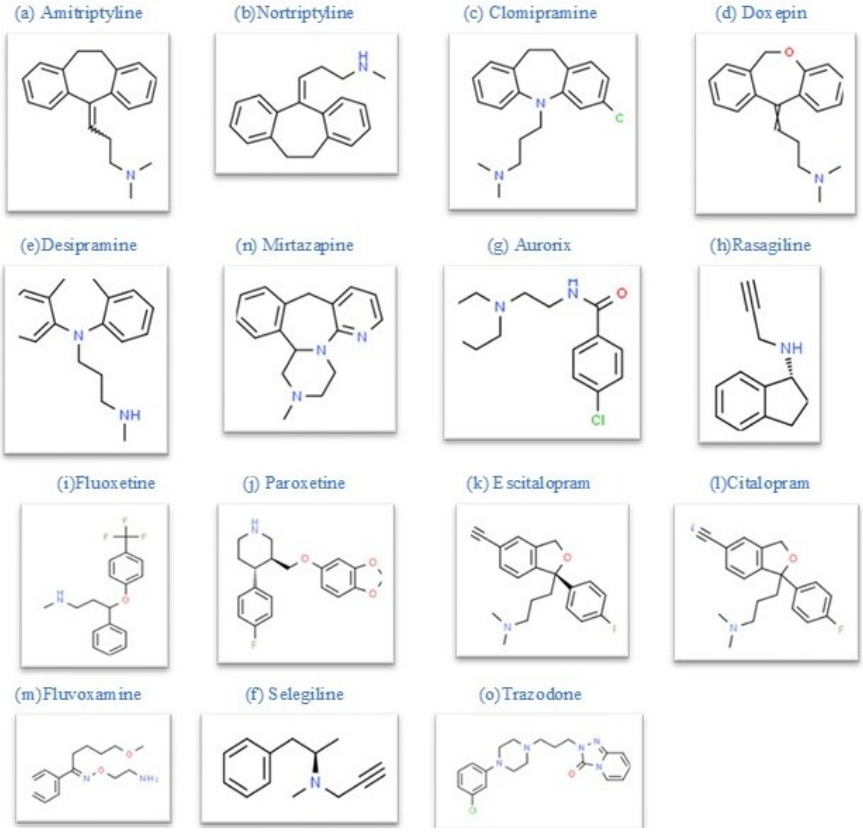



**Figure 2:** Chemical structures of the anti-anxiety drugs retrieved from ChemSpider.

(SVR-Basic, SVR-Tuned, RF, and LR), the linear QSPR models developed for each topological index, and the performance metrics used for evaluation. Section 4 reports the assessment and comparison of model performance, including predicted versus actual property values and graphical summaries of LOOCV and bootstrap results. Section 5 details the algorithmic workflow, model validation strategies, external testing on unseen compounds, residual analysis, and feature selection results. Section 6 concludes the paper with a synthesis of the principal findings, a discussion of the study's limitations, and directions for future research.



## 2 Analytical Evaluation of Molecular Topological Indices

In this study, the chemical architectures of compounds employed in anxiety therapy are modeled as undirected, unweighted graphs. Topological indices (TIs) were computed through vertex and edge decomposition strategies in combination with algorithmic procedures. The analysis is restricted to simple, undirected, connected molecular graphs.

### Justification for Using Temperature-Based Topological Indices

Temperature-based topological indices were selected for this study on the grounds that they are particularly sensitive to molecular stability and thermal reactivity — properties directly relevant to the physicochemical characteristics of anti-anxiety drugs, including boiling point, flash point, and enthalpy. Although other structural descriptors, such as the Zagreb and Wiener indices, capture valuable information about molecular branching and connectivity, they exhibit weaker correlations with temperature-dependent properties. The strong predictive performance achieved using temperature-based indices exclusively, as reported in Section 5.1, further substantiates this choice and confirms their suitability for QSPR modeling of anti-anxiety compounds.

Let $G = (V, E)$ denote an undirected graph, where $V$ is the vertex set and $E$ the edge set. The degree of a vertex $u \in V$, denoted $d_u$, is the number of vertices adjacent to $u$. Table 1 summarizes the topological and temperature-based molecular descriptors employed in this work, together with their defining formulae, as reported in prior studies [5, 27].

### 2.1 Topological Index Computation

This section details the computation of topological indices for anti-anxiety drugs and the QSPR modeling applied to their molecular architectures. Consider the graph $M$ representing the

**Table 1:** Topological descriptors and temperature indices.

| No. | Descriptor Definition | Mathematical Formula |
|---|---|---|
| (1) | Temperature of a vertex $u$ in a connected graph $G$ | $T_u = \dfrac{d_u}{n - d_u}$ |
| (2) | Secondary $T$-index | $ST = \sum_{uv \in E(G)} (T_u + T_v)$ |
| (3) | First-order hyper $T$-index | $H_1(T) = \sum_{uv \in E(G)} (T_u\, T_v)$ |
| (4) | Second-order hyper $T$-index | $H_2(T) = \sum_{uv \in E(G)} (T_u + T_v)^2$ |
| (5) | $F$-$T$ index | $F(T) = \sum_{u \in V(G)} (T_u)^2$ |
| (6) | Harmonic $T$-index | $H(T) = \sum_{uv \in E(G)} \dfrac{2}{T_u + T_v}$ |
| (7) | Product-connectivity $T$-index | $P(T) = \sum_{uv \in E(G)} \dfrac{1}{\sqrt{T_u\, T_v}}$ |
| (8) | Modified third $T$-index | $MT_3 = \sum_{uv \in E(G)} \dfrac{\lvert T_u - T_v \rvert}{T_u + T_v}$ |
| (9) | Modified second $T$-index | $MT_2 = \sum_{uv \in E(G)} \dfrac{(T_u - T_v)^2}{T_u\, T_v}$ |
| (10) | First $T$-index | $M_1(T) = \sum_{u \in V(G)} T_u$ |
| (11) | Reciprocal product-connectivity index | $RPC(T) = \sum_{uv \in E(G)} \dfrac{1}{\sqrt{T_u + T_v}}$ |
| (12) | Symmetric division $T$-index | $SDD(T) = \sum_{uv \in E(G)} \dfrac{T_u^2 + T_v^2}{T_u\, T_v}$ |

molecular structure of Mirtazapine (Figure 2), whose edges are partitioned into distinct categories according to the temperature values of their endpoint vertices:

$$E_1 = \left\{ uv \in E(M) \;\middle|\; T_u = \frac{1}{19},\ T_v = \frac{3}{17} \right\},$$
$$E_2 = \left\{ uv \in E(M) \;\middle|\; T_u = \frac{2}{18},\ T_v = \frac{2}{18} \right\},$$
$$E_3 = \left\{ uv \in E(M) \;\middle|\; T_u = \frac{2}{18},\ T_v = \frac{3}{17} \right\},$$
$$E_4 = \left\{ uv \in E(M) \;\middle|\; T_u = \frac{3}{17},\ T_v = \frac{3}{17} \right\}.$$

Table 2 presents the complete edge partition of $M$.

Applying the descriptors defined in Table 1 to the edge partition of $M$ yields the following index values:

**Table 2:** Edge partition of graph $M$.

| **Edge type** $(T_u, T_v)$ | **Number of edges** |
|---|---|
| $(3/17,\ 3/17)$ | 3 |
| $(2/18,\ 3/17)$ | 10 |
| $(2/18,\ 2/18)$ | 7 |
| $(1/19,\ 3/17)$ | 1 |
| **Total** | 21 |

1. $$PT(M) = \sum_{uv\in E(M)} \frac{1}{\sqrt{T_u \cdot T_v}}$$
$$= \frac{1}{\sqrt{\frac{1}{19}\cdot\frac{3}{17}}} + 7\cdot\frac{1}{\sqrt{\frac{2}{18}\cdot\frac{2}{18}}} + 10\cdot\frac{1}{\sqrt{\frac{2}{18}\cdot\frac{3}{17}}} + 3\cdot\frac{1}{\sqrt{\frac{3}{17}\cdot\frac{3}{17}}} = 161.7905.$$

2. $$HT(M) = \sum_{uv\in E(M)} \frac{2}{T_u + T_v}$$
$$= \frac{2}{\frac{1}{19}+\frac{3}{17}} + 7\cdot\frac{2}{\frac{2}{18}+\frac{2}{18}} + 10\cdot\frac{2}{\frac{2}{18}+\frac{3}{17}} + 3\cdot\frac{2}{\frac{3}{17}+\frac{3}{17}} = 158.2751.$$

3. $$SDT(M) = \sum_{uv\in E(M)} \left(\frac{T_u}{T_v} + \frac{T_v}{T_u}\right)$$
$$= \left(\frac{\frac{1}{19}}{\frac{3}{17}} + \frac{\frac{3}{17}}{\frac{1}{19}}\right) + 7\left(\frac{\frac{2}{18}}{\frac{2}{18}} + \frac{\frac{2}{18}}{\frac{2}{18}}\right) + 10\left(\frac{\frac{2}{18}}{\frac{3}{17}} + \frac{\frac{3}{17}}{\frac{2}{18}}\right) + 3\left(\frac{\frac{3}{17}}{\frac{3}{17}} + \frac{\frac{3}{17}}{\frac{3}{17}}\right) = 45.8298.$$

4. $$FT(M) = \sum_{uv\in E(M)} (T_u^2 + T_v^2)$$
$$= \left(\left(\frac{1}{19}\right)^2 + \left(\frac{3}{17}\right)^2\right) + 7\left(\left(\frac{2}{18}\right)^2 + \left(\frac{2}{18}\right)^2\right)$$
$$+ 10\left(\left(\frac{2}{18}\right)^2 + \left(\frac{3}{17}\right)^2\right) + 3\left(\left(\frac{3}{17}\right)^2 + \left(\frac{3}{17}\right)^2\right) = 0.8285.$$

5. $$T_1(M) = \sum_{uv\in E(M)} (T_u + T_v)$$
$$= \left(\frac{1}{19}+\frac{3}{17}\right) + 7\left(\frac{2}{18}+\frac{2}{18}\right) + 10\left(\frac{2}{18}+\frac{3}{17}\right) + 3\left(\frac{3}{17}+\frac{3}{17}\right) = 5.719298.$$

$$6.\ {}^{m}T_3(M) = \sum_{uv\in E(M)} \frac{1}{T_u + T_v}$$
$$= \frac{1}{\frac{1}{19}+\frac{3}{17}} + 7\cdot\frac{1}{\frac{2}{18}+\frac{2}{18}} + 10\cdot\frac{1}{\frac{2}{18}+\frac{3}{17}} + 3\cdot\frac{1}{\frac{3}{17}+\frac{3}{17}} = 79.1376.$$

$$7.\ {}^{m}T_2(M) = \sum_{uv\in E(M)} \frac{1}{T_u \cdot T_v}$$
$$= \frac{1}{\frac{1}{19}\cdot\frac{3}{17}} + 7\cdot\frac{1}{\frac{2}{18}\cdot\frac{2}{18}} + 10\cdot\frac{1}{\frac{2}{18}\cdot\frac{3}{17}} + 3\cdot\frac{1}{\frac{3}{17}\cdot\frac{3}{17}} = 1281.$$

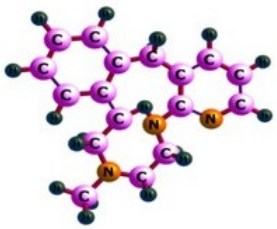
$$8.\ T_2(M) = \sum_{uv\in E(M)} (T_u \times T_v)$$
$$= \left(\frac{1}{19}\times\frac{3}{17}\right) + 7\left(\frac{2}{18}\times\frac{2}{18}\right) + 10\left(\frac{2}{18}\times\frac{3}{17}\right) + 3\left(\frac{3}{17}\times\frac{3}{17}\right) = 0.3852.$$

$$9.\ RPT(M) = \sum_{uv\in E(M)} \sqrt{T_u \times T_v}$$
$$= \sqrt{\frac{1}{19}\times\frac{3}{17}} + 7\sqrt{\frac{2}{18}\times\frac{2}{18}} + 10\sqrt{\frac{2}{18}\times\frac{3}{17}} + 3\sqrt{\frac{3}{17}\times\frac{3}{17}} = 2.803844.$$

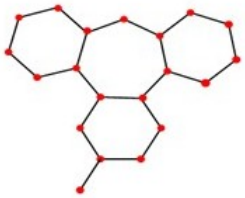
$$10.\ HT_1(M) = \sum_{uv\in E(M)} (T_u + T_v)^2$$
$$= \left(\frac{1}{19}+\frac{3}{17}\right)^2 + 7\left(\frac{2}{18}+\frac{2}{18}\right)^2 + 10\left(\frac{2}{18}+\frac{3}{17}\right)^2 + 3\left(\frac{3}{17}+\frac{3}{17}\right)^2$$
$$= 1.5989.$$

$$11.\ HT_2(M) = \sum_{uv\in E(M)} (T_u \times T_v)^2$$
$$= \left(\frac{1}{19}\times\frac{3}{17}\right)^2 + 7\left(\frac{2}{18}\times\frac{2}{18}\right)^2 + 10\left(\frac{2}{18}\times\frac{3}{17}\right)^2 + 3\left(\frac{3}{17}\times\frac{3}{17}\right)^2$$
$$= 0.0079.$$

Figure 3 depicts the molecular structure of Mirtazapine alongside its corresponding graph representation.



(a) chemical composition



(b) Graph of the molecule

**Figure 3:** Molecular representation and graph-based model of mirtazapine. (a) Two-dimensional (2D) structural representation of the mirtazapine molecule. (b) The corresponding molecular graph, in which vertices represent atoms and edges represent chemical bonds.

Temperature-based topological indices for the 15 anti-anxiety drugs (Figure 2) were computed using the descriptors listed in Table 1 (Section 2) across all 11 temperature indices. The resulting values are reported in Table 3.

Table 4 provides the physicochemical properties of the same 15 compounds, which served as target variables in the QSPR models. The properties considered are boiling point (BP), enthalpy (EN), flash point (FP), molar refractivity (MR), polarizability (PO), and molar volume (MV), all retrieved from the ChemSpider database [4]. Prior to model training, missing values were removed, units were standardized, and all features were normalized by the Z-score method. The units are as follows: BP and FP in °C; EN in kJ/mol; MR in $cm^3$/mol; PO in $RA^3$; and MV in $cm^3$/mol.

**Table 3:** Temperature index values for the 15 anti-anxiety drugs.

| Drug | T2 | HT1 | HT2 | FT | HT | PT | mT3 | mT2 | T1 | RPT | SDT |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Selegiline | 0.5601 | 2.4139 | 0.0288 | 1.2936 | 60.7 | 64.18 | 30.37 | 358.7 | 5.495 | 2.634 | 29.68 |
| Rasagiline | 0.8848 | 3.6402 | 0.0676 | 1.8705 | 59.6 | 60.78 | 29.81 | 282.0 | 6.958 | 3.426 | 29.99 |
| Fluoxetine | 0.2860 | 1.2653 | 0.0069 | 0.6416 | 185.6 | 195.3 | 92.79 | 1864.0 | 4.878 | 2.446 | 53.39 |
| Fluvoxamine | 0.2860 | 1.2653 | 0.0046 | 0.6416 | 185.6 | 207.6 | 92.79 | 1864.0 | 4.878 | 2.446 | 53.13 |
| Amitriptyline | 0.3743 | 1.5584 | 0.0070 | 0.8098 | 184.7 | 189.9 | 92.33 | 1619.0 | 5.898 | 2.883 | 51.20 |
| Nortriptyline | 0.3911 | 1.5978 | 0.0081 | 0.8156 | 173.6 | 175.7 | 86.78 | 1466.0 | 5.811 | 2.874 | 46.11 |
| Desipramine | 0.3911 | 1.5978 | 0.0081 | 0.8156 | 173.6 | 175.7 | 86.78 | 1466.0 | 5.811 | 2.874 | 46.11 |
| Doxepin | 0.3679 | 1.5288 | 0.0068 | 0.7931 | 186.8 | 191.8 | 93.40 | 1652.0 | 5.837 | 2.856 | 50.99 |
| Escitalopram | 0.3503 | 1.5089 | 0.0060 | 0.8082 | 216.3 | 227.2 | 108.2 | 2156.0 | 6.010 | 2.886 | 60.07 |
| Citalopram | 0.3503 | 1.5089 | 0.0060 | 0.8082 | 216.3 | 227.2 | 108.2 | 2156.0 | 6.010 | 2.886 | 60.07 |
| Mirtazapine | 0.3852 | 1.5989 | 0.0079 | 0.8285 | 158.3 | 161.8 | 79.14 | 1281.0 | 5.719 | 2.804 | 45.83 |
| Clomipramine | 0.3852 | 1.6282 | 0.0067 | 0.7703 | 202.1 | 209.5 | 101.0 | 1893.0 | 5.859 | 3.006 | 54.97 |
| Aurorix | 0.3635 | 1.5334 | 0.0082 | 0.8063 | 127.5 | 131.7 | 63.77 | 987.0 | 5.193 | 2.524 | 40.96 |
| Paroxetine | 0.3236 | 1.3422 | 0.0043 | 0.6949 | 250.5 | 255.5 | 125.3 | 2475.0 | 5.953 | 2.921 | 58.50 |
| Trazodone | 0.3026 | 1.2799 | 0.0032 | 0.6337 | 293.1 | 300.6 | 146.6 | 3202.0 | 5.877 | 2.955 | 63.99 |

**Table 4:** Physicochemical properties of anti-anxiety drugs used as targets and features.

| **Drug** | **BP** (°C) | **EN** (kJ/mol) | **FP** (°C) | **MR** ($cm^3$/mol) | **PO** ($RA^3$) | **MV** ($cm^3$/mol) |
|---|---|---|---|---|---|---|
| Selegiline | 272.5 | 51.1 | 108.4 | 53.9 | 21.4 | 162.7 |
| Rasagiline | 305.5 | 54.6 | 121.3 | 60.5 | 24.0 | 196.2 |
| Fluoxetine | 395.1 | 64.5 | 146.8 | 71.0 | 28.2 | 216.6 |
| Fluvoxamine | 395.1 | 64.5 | 160.5 | 79.9 | 31.7 | 222.8 |
| Amitriptyline | 398.2 | 64.9 | 174.0 | 79.9 | 31.7 | 242.9 |
| Nortriptyline | 403.4 | 65.5 | 192.8 | 80.7 | 32.0 | 248.8 |
| Desipramine | 407.4 | 65.9 | 192.8 | 84.2 | 33.4 | 254.3 |
| Doxepin | 413.3 | 66.6 | 194.9 | 86.8 | 34.4 | 257.8 |
| Citalopram | 428.3 | 68.3 | 212.8 | 87.9 | 34.9 | 266.7 |
| Escitalopram | 428.3 | 68.3 | 212.8 | 88.5 | 35.1 | 266.7 |
| Mirtazapine | 432.4 | 68.8 | 215.3 | 91.5 | 36.3 | 271.5 |
| Clomipramine | 434.2 | 69.0 | 216.4 | 92.1 | 36.5 | 272.6 |
| Aurorix | 447.7 | 70.6 | 224.6 | 92.1 | 36.5 | 272.6 |
| Paroxetine | 451.7 | 71.1 | 227.0 | 93.8 | 37.2 | 278.8 |
| Trazodone | 528.5 | 80.3 | 273.4 | 104.2 | 41.3 | 281.2 |

## 3 Methodology

This study employs four predictive modeling approaches: *SVR-Basic*, *SVR-Tuned*, *RF*, and *LR*.

*SVR-Basic* projects input data into a higher-dimensional feature space to capture nonlinear relationships, minimizing prediction errors within a specified margin while controlling model complexity through the regularization parameter $C$ and the insensitivity threshold $\varepsilon$.

*SVR-Tuned* extends the standard SVR formulation by systematically optimizing these hyperparameters over a predefined candidate grid: $C \in \{1, 5, 10\}$, $\varepsilon \in \{0.05, 0.1\}$, and kernel $\in \{\text{linear}, \text{rbf}\}$. Model selection is conducted via Leave-One-Out Cross-Validation (LOOCV), with performance assessed by the mean fold-wise $R^2$ computed from predictions on each held-out sample. The hyperparameter configuration yielding the highest mean LOOCV $R^2$ is retained as the optimal model for each target property.

*Random Forest (RF)* constructs an ensemble of decision trees trained on random subsets of the data, aggregating their outputs to improve predictive accuracy, reduce overfitting, and capture complex nonlinear patterns. Key governing parameters include *n_estimators*, *max_depth*, and *min_samples_split*.

For *Linear Regression (LR)*, separate univariate models were developed for each topological index and each target property. The model takes the form:

$$P = B + A \cdot \text{TI}, \tag{1}$$

where $P$ denotes the predicted drug property, $B$ is the intercept, $A$ is the regression coefficient, and TI represents a single topological index.

Model performance is quantified using four metrics: *Mean Squared Error (MSE)*, *Root Mean Squared Error (RMSE)*, *Mean Absolute Error (MAE)*, and the coefficient of determination $R^2$. These measures, corresponding to models 13–16, jointly characterize both predictive accuracy and overall model effectiveness:

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

$$\text{RMSE} = \sqrt{\text{MSE}}, \tag{3}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \tag{4}$$

where $y_i$ denotes the observed value, $\hat{y}_i$ the predicted value, and $n$ the total number of samples.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{5}$$

where $\bar{y}$ is the mean of the observed values.

An optimal model is characterized by $R^2 \to 1$, indicating that nearly all variance in the response is explained, together with MSE, RMSE, and MAE values approaching zero. A model satisfying both criteria simultaneously is regarded as the most reliable for predictive purposes.

**Linear Models**

Univariate linear regression models were constructed for six physicochemical properties (BP, EN, FP, MR, PO, and MV) across fifteen anti-anxiety drugs using eleven computed topological indices, yielding the results presented in Table 5. Models with a $p$-value exceeding 0.05 were deemed statistically non-significant; for such cases, no reliable regression coefficients could be established and the corresponding regression line was not reported. Of all candidate models, fifty-three attained statistical significance and are retained in the table. Complete details — including all coefficients and statistical measures for every index–property combination, including non-significant cases — are provided in the Appendix. The Appendix further documents the physicochemical features associated with each topological index and the full set of statistical measures used within the linear QSPR modeling framework. The physicochemical properties dataset is available online.

https://doi.org/10.6084/m9.figshare.28903265

**Table 5:** Linear regression models for each topological index.

| Index | Regression Model | Index | Regression Model |
|---|---|---|---|
| $[PT(G)]$ | BP $= 261.490 + 0.800$ TI | $[HT(G)]$ | BP $= 259.527 + 0.841$ TI |
| | EN $= 49.511 + 0.091$ TI | | EN $= 49.278 + 0.095$ TI |
| | FP $= 102.330 + 0.483$ TI | | FP $= 102.320 + 0.501$ TI |
| | MR $= 47.989 + 0.190$ TI | | MR $= 47.738 + 0.199$ TI |
| | PO $= 19.060 + 0.075$ TI | | PO $= 18.960 + 0.079$ TI |
| | MV $= 157.259 + 0.488$ TI | | MV $= 158.846 + 0.497$ TI |
| $[SDT(G)]$ | BP $= 175.188 + 4.718$ TI | $[FT(G)]$ | BP $= 531.650 - 140.670$ TI |
| | EN $= 39.851 + 0.532$ TI | | EN $= 79.969 - 15.772$ TI |
| | FP $= 48.236 + 2.887$ TI | | FP $= 257.105 - 75.415$ TI |
| | MR $= 24.912 + 1.173$ TI | | MR $= 111.255 - 32.369$ TI |
| | PO $= 9.928 + 0.464$ TI | | PO $= 44.113 - 12.823$ TI |
| | MV $= 93.216 + 3.107$ TI | | MV $= 328.430 - 93.178$ TI |
| $[mT3(G)]$ | BP $= 259.527 + 1.682$ TI | $[mT2(G)]$ | BP $= 300.259 + 0.066$ TI |
| | EN $= 49.278 + 0.191$ TI | | EN $= 53.862 + 0.008$ TI |
| | FP $= 102.320 + 1.001$ TI | | FP $= 123.513 + 0.041$ TI |
| | MR $= 47.738 + 0.397$ TI | | MR $= 57.388 + 0.016$ TI |
| | PO $= 18.960 + 0.157$ TI | | PO $= 22.782 + 0.006$ TI |
| | MV $= 158.846 + 0.994$ TI | | MV $= 181.129 + 0.040$ TI |
| $[T2(G)]$ | BP $= 522.324 - 282.105$ TI | $[HT1(G)]$ | BP $= 530.068 - 71.610$ TI |
| | EN $= 78.921 - 31.624$ TI | | EN $= 79.788 - 8.027$ TI |
| | FP $= 252.213 - 151.510$ TI | | FP $= 255.439 - 37.900$ TI |
| | MR $= 109.516 - 65.930$ TI | | MR $= 111.038 - 16.565$ TI |
| | PO $= 43.423 - 26.116$ TI | | PO $= 44.027 - 6.562$ TI |
| | MV $= 325.494 - 194.962$ TI | | MV $= 328.200 - 47.918$ TI |
| $[HT2(G)]$ | BP $= 440.424 - 2577.510$ TI | | |
| | EN $= 69.738 - 288.780$ TI | | |
| | MR $= 90.665 - 626.568$ TI | | |
| | PO $= 35.956 - 248.130$ TI | | |
| | MV $= 268.345 - 1735.770$ TI | | |

The FP regression model for $HT2(G)$ is not reported, as the corresponding $p$-value exceeded 0.05, indicating statistical non-significance. Complete results are provided in the supplementary material.

## 4 Assessment and Comparison of Machine Learning Model Performance in QSPR Studies

All computations were performed within the Python 3.12.7 environment. Measured physicochemical parameters and their corresponding model predictions were obtained using four distinct regression algorithms. For illustrative purposes, predicted values for BP and EN are presented in Table 6; results for the remaining physicochemical properties are available online.

Model performance was evaluated using MSE, RMSE, MAE, and $R^2$. Complete LOOCV and bootstrap results for all four models across the six physicochemical properties are reported in Table 7. Graphical comparisons appear in Figures 4–7: Figure 4 contrasts LOOCV $R^2$ scores; Figure 5 presents LOOCV RMSE values; Figure 6 depicts model stability via the bootstrap coefficient of variation (CV%); and Figure 7 provides a detailed cross-model comparison of LOOCV RMSE. The algorithmic procedures underlying these predictions are described in Algorithm 1.

**Table 6:** Predicted and actual values of BP and EN using the four regression models.

| Drug | Act BP | SVR-B | SVR-T | RF | LR | Act EN | SVR-B | SVR-T | RF | LR |
|---|---|---|---|---|---|---|---|---|---|---|
| Selegiline | 272.5 | 409.28 | 279.60 | 319.25 | 275.76 | 51.1 | 66.78 | 49.99 | 64.12 | 50.06 |
| Rasagiline | 305.5 | 409.33 | 305.30 | 314.25 | 304.52 | 54.6 | 65.91 | 54.59 | 59.50 | 54.60 |
| Fluoxetine | 395.1 | 410.44 | 394.91 | 396.75 | 395.22 | 64.5 | 65.70 | 64.49 | 65.19 | 64.47 |
| Fluvoxamine | 395.1 | 410.44 | 395.30 | 401.10 | 394.98 | 64.5 | 65.70 | 64.57 | 64.86 | 64.53 |
| Amitriptyline | 398.2 | 410.38 | 398.89 | 401.18 | 397.33 | 64.9 | 66.13 | 64.80 | 65.10 | 64.82 |
| Nortriptyline | 403.4 | 410.72 | 403.94 | 403.40 | 402.62 | 65.5 | 66.63 | 65.44 | 65.40 | 65.46 |
| Desipramine | 407.4 | 410.35 | 407.20 | 403.69 | 406.53 | 65.9 | 66.13 | 65.91 | 65.83 | 65.93 |
| Doxepin | 413.3 | 410.35 | 413.10 | 403.71 | 413.03 | 66.6 | 66.37 | 66.59 | 66.17 | 66.61 |
| Citalopram | 428.3 | 411.45 | 428.10 | 428.97 | 426.30 | 68.3 | 67.76 | 68.31 | 68.23 | 68.34 |
| Escitalopram | 428.3 | 411.45 | 428.10 | 428.97 | 426.30 | 68.3 | 67.76 | 68.31 | 68.23 | 68.34 |
| Mirtazapine | 432.4 | 411.46 | 432.23 | 433.05 | 431.58 | 68.8 | 67.82 | 68.81 | 68.45 | 68.83 |
| Clomipramine | 434.2 | 411.55 | 433.83 | 431.23 | 433.10 | 69.0 | 68.02 | 69.11 | 69.35 | 69.12 |
| Aurorix | 447.7 | 411.63 | 447.50 | 441.50 | 448.07 | 70.6 | 67.73 | 70.59 | 69.97 | 70.59 |
| Paroxetine | 451.7 | 411.84 | 451.77 | 440.85 | 451.25 | 71.1 | 68.62 | 71.18 | 70.49 | 71.17 |
| Trazodone | 528.5 | 411.97 | 530.15 | 494.50 | 531.69 | 80.3 | 68.80 | 80.29 | 76.58 | 80.25 |

https://doi.org/10.6084/m9.figshare.30571391

## 5 Algorithms

This section presents, in structured algorithmic form, the complete workflow for *Optimization, Validation, and Prediction of Physicochemical Drug Properties Using SVR, RF, and LR Models*. Given the limited dataset size of 15 compounds, the validation framework incorporates both *Leave-One-Out Cross-Validation (LOOCV)* and *Bootstrap Resampling* to ensure statistical reliability of the reported estimates.

### 5.1 Model Validation and Statistical Reliability

Robust validation is indispensable when machine learning models are trained on small chemical datasets, where conventional train–test splits yield estimates that are both high-variance and potentially misleading. This subsection describes the validation strategies adopted to address this challenge, presents their results across all four models and six target properties, and offers a statistical interpretation of the comparative findings.

#### 5.1.1 Limitation of Small Dataset

The experimental dataset comprises 15 compounds, a sample size that, while typical of early-stage QSPR studies, constrains the statistical reliability and generalizability of learned models. To demonstrate that the reported conclusions remain scientifically valid despite this constraint, two complementary validation strategies were employed: Leave-One-Out Cross-Validation (LOOCV) and Bootstrap Resampling.

#### 5.1.2 Validation Methodology

**Leave-One-Out Cross-Validation (LOOCV).** In LOOCV, a single compound is withheld for evaluation while the model is trained on the remaining 14; this procedure is repeated until every compound has served once as the validation sample. The approach is particularly appropriate for small datasets because every observation contributes to both training and evaluation, thereby maximizing the use of available data and yielding nearly unbiased estimates of predictive performance.

**Bootstrap Resampling.** Model stability was assessed via bootstrap resampling with replacement over 500 iterations. In each iteration, a sample of 15 compounds was drawn from the original dataset, the model was retrained on the replicate, and predictive performance was recorded. From the resulting $R^2$ distribution, the mean $R^2$, 95 % confidence interval (CI), and coeffi-

**Algorithm 1** Model Optimization and Validation for Physicochemical Property Prediction

**Input:** Dataset of 15 compounds; six target properties: BP, EN, FP, MR, PO, MV.
**Output:** Model performance metrics, optimal SVR parameters, and visualizations.

**Step 1. Data Loading and Preprocessing.**
- a. Load the dataset from the Excel source file.
- b. Clean the data and convert all features to numeric format.
- c. Define the six target properties and extract topological indices as the feature matrix.

**Step 2. Feature Processing** (per target property).
- a. Standardize all features using $z$-score normalization.
- b. Apply `SelectKBest` with `f_regression` to retain the top 5 most informative features.

**Step 3. SVR Hyperparameter Optimization** (LOOCV-based grid search).
- a. Candidate grid: $C \in \{1, 5, 10\}$, $\varepsilon \in \{0.05, 0.1\}$, kernel $\in \{\texttt{linear}, \texttt{rbf}\}$.
- b. Perform LOOCV for each parameter combination.
- c. Select the configuration maximising mean LOOCV $R^2$.

**Step 4. Model Training.**
- a. Train `SVR_Basic` (`kernel=rbf`, $C = 10$, $\varepsilon = 0.1$).
- b. Train `SVR_Tuned` with optimal parameters from Step 3.
- c. Train a Linear Regression baseline model.
- d. Train a Random Forest (`n_estimators`$= 50$, `max_depth`$= 3$).

**Step 5. Model Evaluation** (LOOCV).
- a. For each model, train on 14 samples and evaluate on the held-out sample.
- b. Compute $R^2$, RMSE, MAE, and MSE.

**Step 6. Bootstrap Stability Analysis** (500 iterations).
- a. Draw a bootstrap sample of 15 compounds with replacement.
- b. Train the model on each bootstrap replicate.
- c. Record the resulting $R^2$ distribution.
- d. Compute mean $R^2$, 95 % CI, and CV %.

**Step 7. Visualization.**
- a. Comparative bar charts of LOOCV $R^2$ and RMSE across all models.
- b. Bootstrap stability plots displaying CV % per property.
- c. Actual vs. predicted values with 95 % confidence bands.

**Step 8. Output Storage.**
- a. Export all numerical results to a multi-sheet Excel workbook.
- b. Save all figures as high-resolution PNG files.

cient of variation (CV %) were computed. A CV % below 15 % was considered indicative of acceptable stability, while CV % below 10 % was classified as excellent.

### 5.1.3 Model Comparison Results

Four regression models — SVR_Basic, SVR_Tuned, Linear Regression, and Random Forest — were evaluated for their ability to predict six physicochemical properties (BP, EN, FP, MR, PO, and MV) using the computed topological indices as input features. The complete LOOCV and bootstrap results are reported in Table 7.

**Table 7:** Complete LOOCV and bootstrap results for all four models across six physicochemical properties.

| Property | Model | LOOCV $R^2$ | RMSE | MAE | CV% |
|---|---|---|---|---|---|
| **BP** | SVR_Basic | -0.0364 | 58.44 | 36.85 | 47.0 |
| | SVR_Tuned | 0.6637 | 33.29 | 23.98 | 32.9 |
| | Linear Regression | -0.5675 | 71.87 | 45.15 | 13.9 |
| | Random Forest | 0.6261 | 35.10 | 24.45 | 5.7 |
| **EN** | SVR_Basic | 0.5277 | 4.48 | 2.88 | 4.0 |
| | SVR_Tuned | 0.6615 | 3.79 | 2.72 | 29.2 |
| | Linear Regression | -0.6071 | 8.26 | 5.21 | 13.4 |
| | Random Forest | 0.6157 | 4.04 | 2.79 | 5.6 |
| **FP** | SVR_Basic | -0.0308 | 42.49 | 31.98 | 39.5 |
| | SVR_Tuned | 0.2807 | 35.49 | 27.60 | 50.9 |
| | Linear Regression | -0.4670 | 50.69 | 38.66 | 26.4 |
| | Random Forest | 0.3676 | 33.28 | 24.39 | 5.5 |
| **MR** | SVR_Basic | 0.4780 | 9.13 | 6.10 | 13.6 |
| | SVR_Tuned | 0.8255 | 5.28 | 4.31 | 11.7 |
| | Linear Regression | 0.3946 | 9.83 | 6.70 | 4.9 |
| | Random Forest | 0.6701 | 7.26 | 5.86 | 3.5 |
| **PO** | SVR_Basic | 0.5518 | 3.35 | 2.14 | 4.6 |
| | SVR_Tuned | 0.8276 | 2.08 | 1.70 | 11.5 |
| | Linear Regression | 0.3991 | 3.88 | 2.64 | 4.9 |
| | Random Forest | 0.6721 | 2.86 | 2.31 | 3.5 |
| **MV** | SVR_Basic | 0.3386 | 26.77 | 16.62 | 23.6 |
| | SVR_Tuned | 0.7710 | 15.75 | 11.71 | 8.6 |
| | Linear Regression | -1.3069 | 49.99 | 27.12 | 5.6 |
| | Random Forest | 0.7716 | 15.73 | 10.50 | 2.1 |

Stability classification based on bootstrap CV%: CV% $<$ 30% = Stable; 30–50% = Moderate; $>$ 50% = Unstable.

### 5.1.4 Visual Comparison of Model Performance

Figures 4–7 provide graphical summaries of model performance across all properties. Figure 4 contrasts LOOCV $R^2$ scores, offering a direct measure of each model's explanatory power under cross-validation. Figure 5 presents the corresponding LOOCV RMSE values, which quantify absolute prediction error on the same scale as the target properties. Figure 6 depicts bootstrap CV% profiles, conveying the degree to which each model's performance fluctuates across resampled datasets — a measure of stability rather than accuracy alone. Finally, Figure 7 provides a property-by-property comparison of LOOCV RMSE across all four models, facilitating identification of properties that are inherently more difficult to predict.

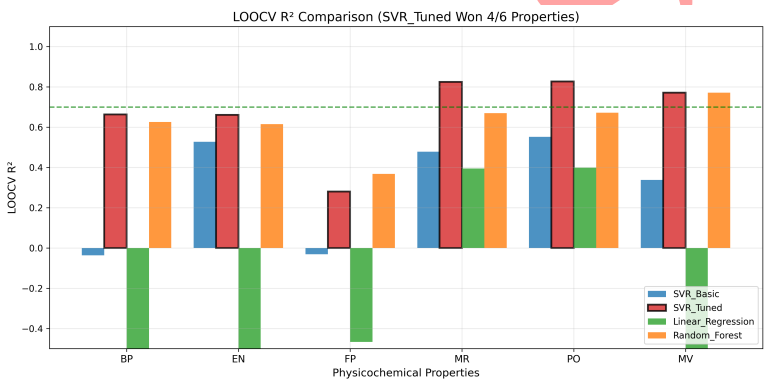



**Figure 4:** Comparison of the four regression models using LOOCV $R^2$ scores across six physicochemical properties.

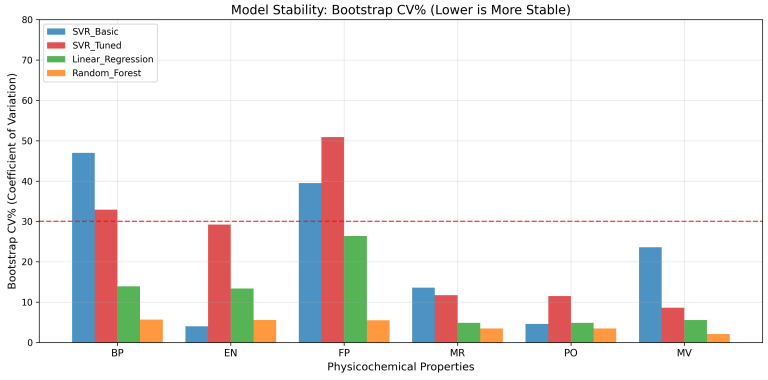



**Figure 5:** Comparison of LOOCV RMSE values for the four regression models across six physicochemical properties.

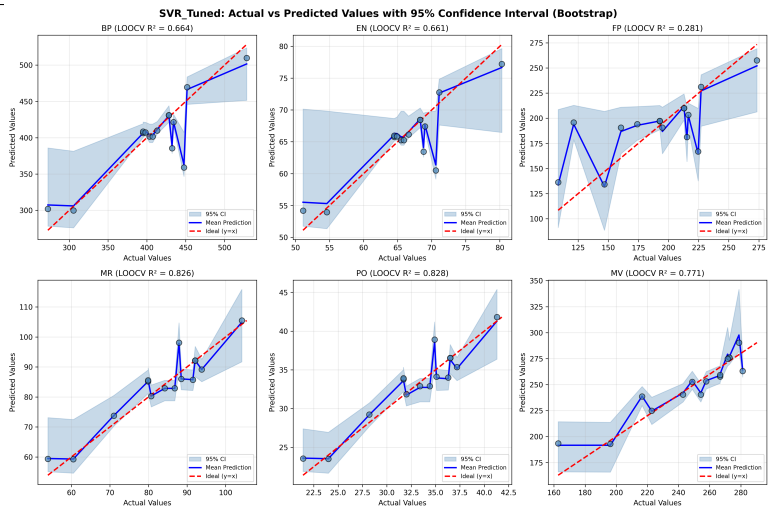



**Figure 6:** Bootstrap coefficient of variation (CV%) for the four models, illustrating relative prediction stability across physicochemical properties.

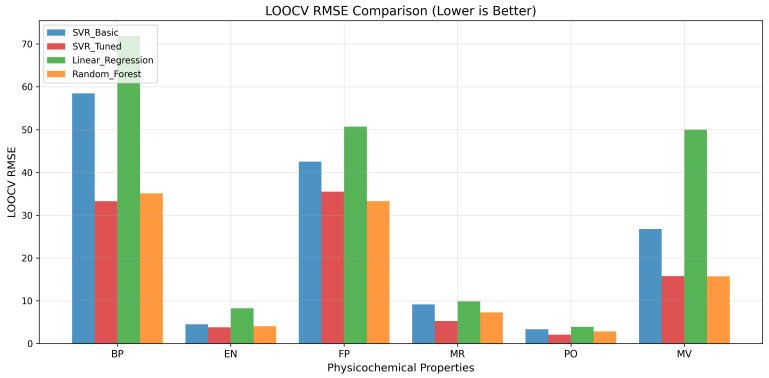



**Figure 7:** Property-level comparison of LOOCV RMSE across the four regression models.

### 5.1.5 Statistical Interpretation

The results in Table 7 and Figures 4–7 yield several clear conclusions. SVR_Tuned achieved the highest LOOCV $R^2$ for MR (0.8255) and PO (0.8276), and ranked first or second across all remaining properties, demonstrating that systematic hyperparameter optimization provides consistent gains in predictive accuracy. Random Forest proved the most stable model overall, exhibiting the lowest bootstrap CV% values across properties — a consequence of its ensemble averaging mechanism, which suppresses variance at the cost of a modest accuracy penalty relative to SVR_Tuned. Linear Regression, by contrast, yielded negative LOOCV $R^2$ for BP ($-0.5675$), EN ($-0.6071$), FP ($-0.4670$), and MV ($-1.3069$), confirming that the underlying structure–property relationships are substantially nonlinear and cannot be adequately captured by a linear model. The untuned SVR_Basic performed poorly despite its nonlinear kernel, un-

derscoring how sensitive kernel machines are to hyperparameter selection. Flash point emerged as the most challenging property across all models, likely reflecting its greater dependence on molecular interactions that are not fully encoded by temperature-based topological indices alone. While external validation on a larger, independent dataset would further strengthen these conclusions, the LOOCV and bootstrap results collectively provide statistically informative evidence for the predictive utility of the proposed framework across the six physicochemical properties studied.

### 5.2 Testing and Assessment of ML Algorithms on Unseen Data

Internal cross-validation, while rigorous, evaluates model performance within the training domain. To assess how well the learned models generalize to genuinely new compounds, a separate set of 10 anti-anxiety drugs not included in the training phase was used as an external test set. All four models were trained on the original 15-compound dataset and then applied to predict six physicochemical properties for these held-out compounds. Table 8 presents the predicted BP and EN values alongside the experimental uncertainty ranges retrieved from the ChemSpider database. These uncertainty ranges reflect the inherent variation in experimental measurements and are independent of model prediction error.

**Table 8:** Predicted and actual values of BP and EN for 10 unseen test compounds.

| Drug | ACT-BP | SVR-B | SVR-T | RF | LR | ACT-EN | SVR-B | SVR-T | RF | LR |
|---|---|---|---|---|---|---|---|---|---|---|
| Aplenzin | $334.8 \pm 27.0$ | 409.7 | 339.0 | 346.8 | 332.4 | $57.8 \pm 3.0$ | 66.3 | 56.9 | 64.2 | 57.0 |
| Effexor | $397.6 \pm 27.0$ | 411.1 | 428.1 | 418.4 | 426.5 | $68.3 \pm 3.0$ | 66.2 | 65.5 | 65.4 | 65.5 |
| Melipramin | $403.1 \pm 44.0$ | 410.5 | 402.9 | 404.3 | 402.4 | $65.4 \pm 3.0$ | 66.4 | 64.6 | 65.3 | 64.6 |
| Cymbalta | $466.2 \pm 40.0$ | 412.1 | 466.3 | 435.3 | 466.8 | $72.8 \pm 3.0$ | 68.9 | 72.9 | 70.0 | 72.8 |
| Nozinan | $468.0 \pm 45.0$ | 412.1 | 468.0 | 435.6 | 468.3 | $73.0 \pm 3.0$ | 68.9 | 73.1 | 70.1 | 73.0 |
| Lodopin | $478.4 \pm 45.0$ | 412.2 | 479.1 | 438.9 | 479.7 | $74.3 \pm 3.0$ | 69.1 | 74.3 | 70.0 | 74.3 |
| Seroquel | $556.5 \pm 60.0$ | 411.3 | 597.2 | 494.5 | 601.2 | $88.2 \pm 3.0$ | 68.2 | 83.6 | 76.0 | 83.6 |
| Invega | $612.3 \pm 65.0$ | 410.7 | 660.2 | 494.5 | 665.1 | $95.6 \pm 3.0$ | 67.3 | 90.4 | 76.0 | 90.3 |
| Buspar | $613.9 \pm 65.0$ | 411.0 | 622.0 | 494.5 | 626.2 | $91.1 \pm 3.0$ | 67.2 | 91.6 | 75.9 | 91.5 |
| Latuda | $623.4 \pm 55.0$ | 410.5 | 631.8 | 494.5 | 634.9 | $92.3 \pm 3.0$ | 67.4 | 90.2 | 75.9 | 90.2 |

Model performance on the 10 external compounds is summarized in Figures 8 and 9 and Table 9. Figure 8 plots actual values against SVR_Tuned predictions — the best-performing model — for all six properties; each point represents one test compound ($n = 10$), with red diamond markers identifying the new samples. The proximity of points to the $y = x$ diagonal indicates that the model generalizes effectively beyond its training domain. Figure 9 presents a heatmap of prediction errors for all four models across the six properties, with cell colors ranging from yellow to dark red to indicate increasing error magnitude. SVR_Tuned consistently

displays lighter colors, particularly for BP, MR, and MV, corroborating its superior generalization performance.

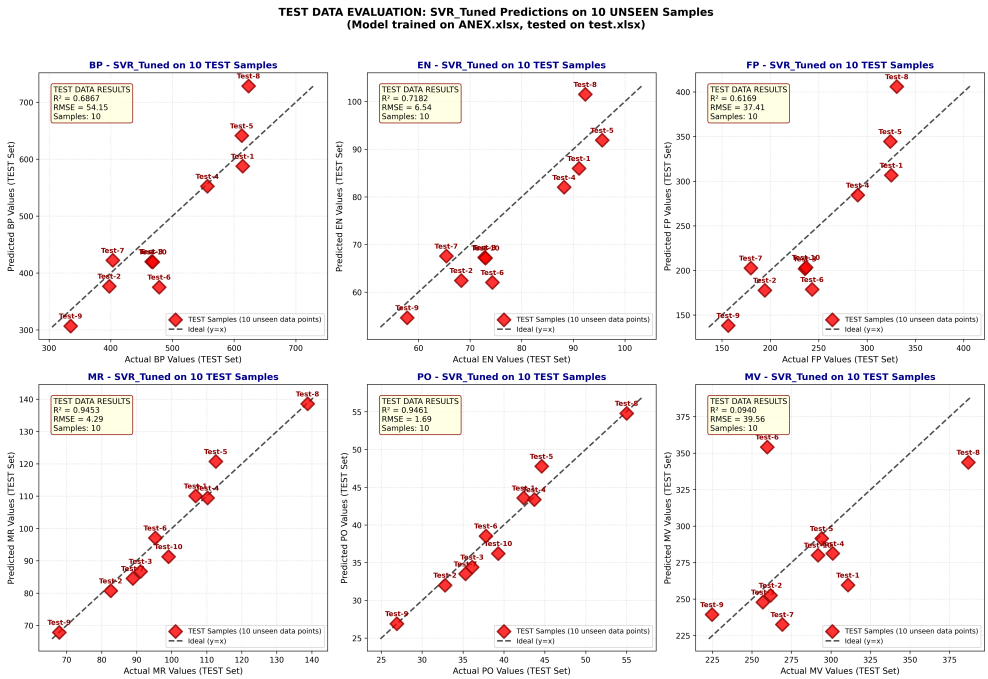



**Figure 8:** Actual versus predicted values for SVR_Tuned applied to 10 unseen test compounds across all six physicochemical properties. Red diamond markers denote the external test samples; proximity to the diagonal $y = x$ line indicates predictive accuracy.

As shown in Table 9, SVR_Tuned achieves its highest $R^2$ values for PO and MR (approximately 0.95) and its lowest for MV (0.09). Linear regression performs acceptably for BP ($R^2 = 0.7472$), slightly exceeding SVR_Tuned for that property ($R^2 = 0.6867$). Random Forest and SVR_Basic exhibit poor performance across most properties on the external set. These findings confirm that LOOCV-guided hyperparameter optimization substantially improves the generalization capability of the SVR model.

**Discussion of Test Results**

The external validation results presented in Figure 8, Figure 9, and Table 9 support three main observations. First, regarding generalization: SVR_Tuned produced predictions closely aligned with experimental values across most properties (Figure 8), demonstrating that the model captures genuine structure–property relationships rather than artifacts of the training set. Second, regarding model ranking: SVR_Tuned achieved the best $R^2$ for EN, FP, MR, and PO; linear regression was marginally superior for BP (0.7472 vs. 0.6867); and all models failed for

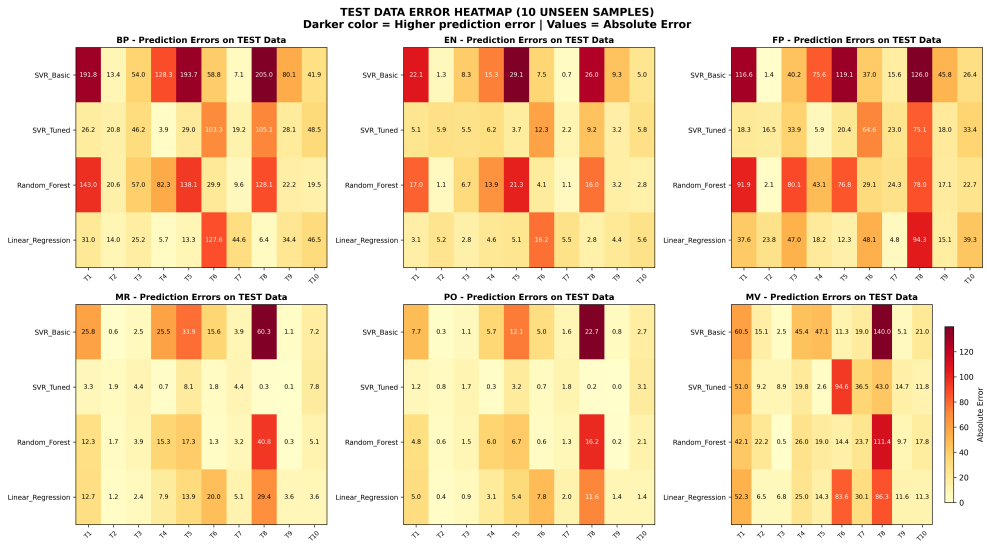



**Figure 9:** Heatmap of prediction errors for the four regression models on 10 external test compounds across six physicochemical properties. Color intensity from yellow to dark red reflects increasing error magnitude.

**Table 9:** $R^2$ values for all four models evaluated on 10 external test compounds across six physicochemical properties.

| Property | SVR-Basic | SVR-Tuned | Lin. Reg. | Rand. Forest |
|---|---|---|---|---|
| BP | -0.5767 | 0.6867 | **0.7472** | 0.2709 |
| EN | -0.6373 | **0.7182** | 0.7082 | 0.1585 |
| FP | -0.5172 | **0.6169** | 0.5161 | 0.1525 |
| MR | -0.9056 | **0.9453** | 0.4824 | 0.2851 |
| PO | -0.4957 | **0.9461** | 0.4899 | 0.2909 |
| MV | -0.6620 | **0.0940** | -0.1152 | 0.0219 |

Bold values indicate the highest $R^2$ per property. Negative $R^2$ values indicate that the model performs worse than a constant mean predictor.

MV, where the best $R^2$ reached only 0.09, pointing to a systematic limitation of the current descriptor set for this property. Third, regarding error structure: the heatmap in Figure 9 confirms that SVR_Tuned and, to a lesser extent, linear regression yield the smallest prediction errors, with no evidence of systematic directional bias in the residuals.

Taken together, these results establish that LOOCV-optimized SVR generalizes reliably to new anti-anxiety compounds for most target properties. Molar volume remains an open challenge that warrants further investigation.

**Table 10:** RFE rankings of the eleven temperature-based topological indices across six target properties.

| Feature | BP | EN | FP | MR | PO | MV |
|---|---|---|---|---|---|---|
| T2 | 1 | 1 | 1 | 1 | 1 | 1 |
| HT1 | 1 | 1 | 1 | 1 | 1 | 1 |
| HT2 | 1 | 1 | 1 | 1 | 1 | 1 |
| FT | 1 | 1 | 1 | 1 | 1 | 1 |
| HT | 1 | 1 | 1 | 1 | 1 | 1 |
| PT | 3 | 3 | 3 | 3 | 3 | 3 |
| mT3 | 1 | 1 | 1 | 1 | 1 | 1 |
| mT2 | 4 | 4 | 4 | 4 | 4 | 4 |
| T1 | 1 | 1 | 1 | 1 | 1 | 1 |
| RPT | 1 | 1 | 1 | 1 | 1 | 1 |
| SDT | 2 | 2 | 2 | 2 | 2 | 2 |

Rank 1 denotes features retained by RFE as equally top-ranked; features eliminated at the same recursive step receive identical ranks. `SelectKBest` independently identified the top-5 features via univariate $F$-regression scores; these consistently coincided with the Rank 1 group across all target properties.



### 5.3 Analyzing Machine Learning Model Performance through Error and Residual Plots

Residual and error distribution analysis provides a complementary perspective on model behavior that aggregate metrics such as $R^2$ and RMSE do not fully convey. Systematic patterns in residuals can reveal model misspecification, whereas near-zero, symmetrically distributed residuals indicate that the model captures the underlying relationship without bias. Figure 10 presents a comparative residual analysis for BP across all four models. The results show that hyperparameter tuning brought SVR residuals close to zero and markedly improved prediction accuracy relative to SVR_Basic. Figure 11 further illustrates that error dispersion was substantially reduced following tuning, with predictions shifting consistently toward the actual values. Among all models and properties examined, SVR_Tuned applied to MR achieved the highest

accuracy and the smallest error deviation. These findings reinforce the conclusion that systematic hyperparameter optimization is a decisive factor in SVR model performance within this QSPR framework.

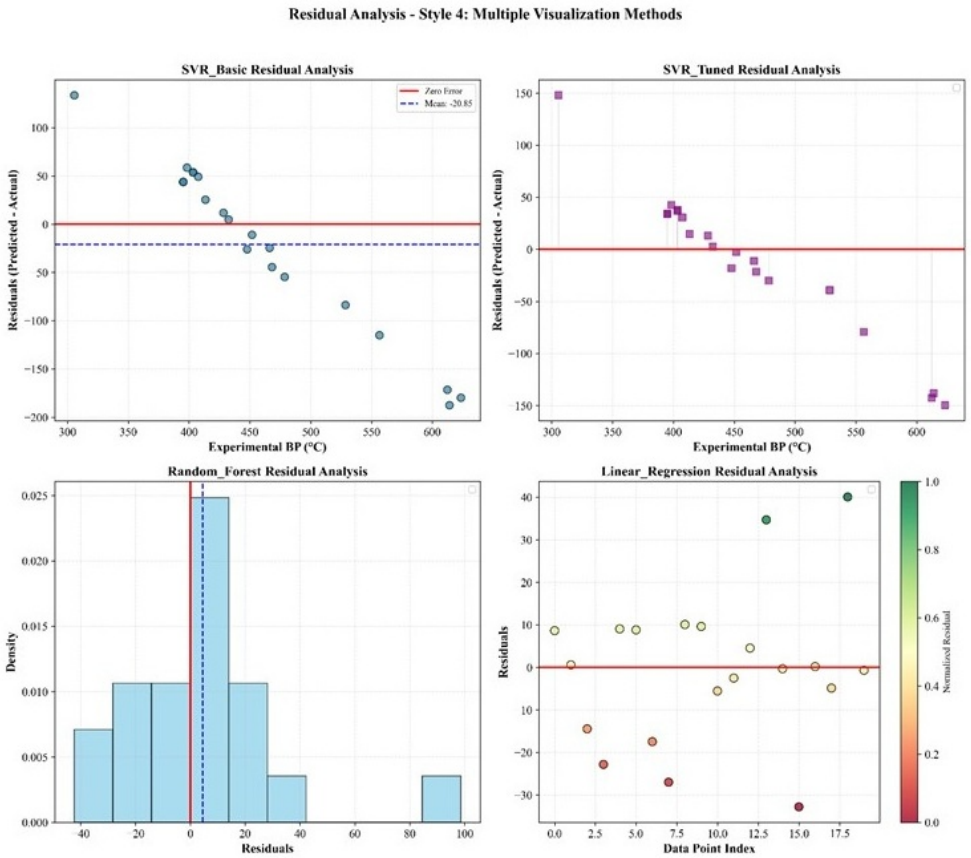



**Figure 10:** Comparative residual analysis for BP predictions across the four regression models.

## 5.4 Analysis of Feature Selection and Correlation

Understanding which descriptors drive model performance is essential both for scientific interpretability and for avoiding redundancy in the feature set. Among the eleven temperature-based descriptors, Recursive Feature Elimination (RFE) was applied to rank each descriptor's contribution to each target property; the results are reported in Table 10. A rank of 1 indicates that the corresponding features were retained at the same elimination step and are therefore considered equally informative — not that they are identical in value or meaning. The `SelectKBest` method, applied independently using univariate $F$-regression scores, confirmed that the top-5 features it selected consistently fell within the Rank 1 group across all six properties. Figure 12

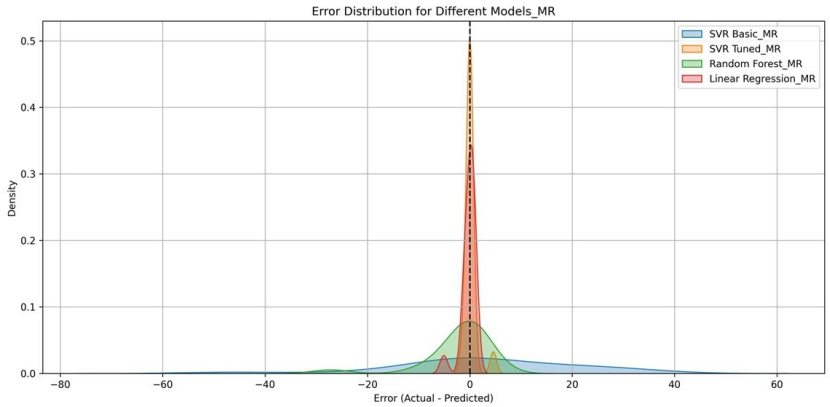



**Figure 11:** Error distribution analysis illustrating the effect of hyperparameter optimization on SVR predictive accuracy.

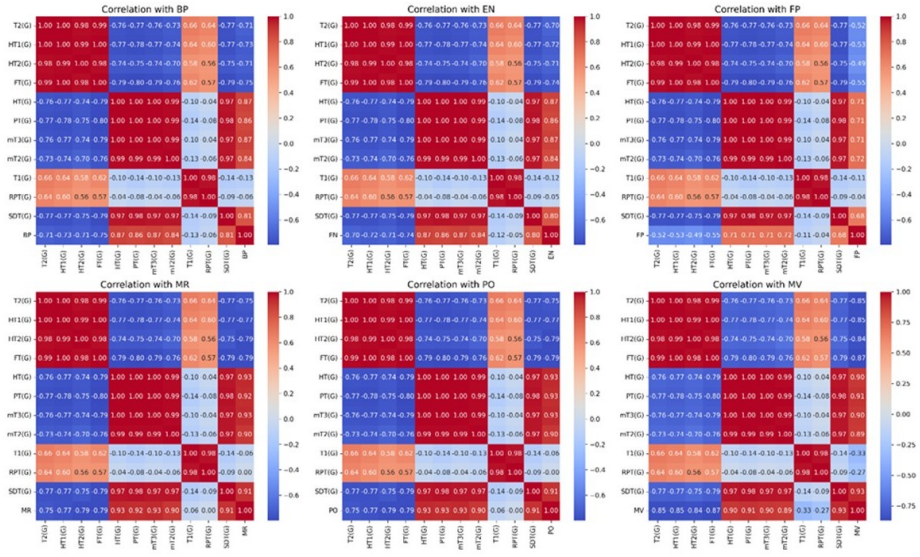



**Figure 12:** Pearson correlation matrices between the eleven temperature-based topological descriptors and the six target properties (BP, EN, FP, MR, PO, MV). Each subplot displays the correlation coefficients for one target property, highlighting the linear relationships exploited during feature selection.

presents the Pearson correlation matrices between each descriptor and each target property, providing a transparent account of the linear associations that underpin the feature selection process. Collectively, these analyses ensure that only the most informative descriptors are incorporated into the final models.

## 6 Conclusion

This investigation confirms that temperature-based topological indices are effective molecular descriptors within a QSPR (Quantitative Structure-Property Relationship) framework for analyzing anti-anxiety medications. A primary finding of this work is that LOOCV-driven hyperparameter tuning is essential for enhancing the performance of SVR models when working with sparse chemical datasets. Specifically, the SVR-Tuned model outperformed its counterparts, achieving high LOOCV $R^2$ scores for Molar Refractivity (MR, $0.8255$) and Parachor (PO, $0.8276$).

While SVR-Tuned provided the highest accuracy, Random Forest demonstrated superior stability during bootstrap analysis due to its ensemble nature. Conversely, Linear Regression proved inadequate for this task, failing to model the complex, nonlinear dependencies between molecular structure and properties like Boiling Point (BP) or Molar Volume (MV). External validation further supported these results, with SVR-Tuned maintaining high accuracy ($R^2 \approx 0.95$) for most properties, though Molar Volume remained a challenge for all models. Feature selection techniques identified a core set of eight descriptors (including T2, HT1, and FT) that consistently provided the most predictive value across different properties.

### Limitations and Future Work

The study identifies several constraints, primarily the small dataset size (15 training compounds) and the reliance on 2D descriptors, which may overlook critical 3D conformational data. Furthermore, the consistent difficulty in predicting Flash Point suggests that this property may be influenced by structural factors not captured by the current TIs or by inherent experimental variability.

Future research should focus on:

- *Dataset Expansion*: Including a wider variety of psychotropic substances to improve the model's generalizability.
- *Descriptor Diversity*: Incorporating 3D structural features, electronic data, and molecular fingerprints to improve predictions for Molar Volume and Flash Point.

- *Advanced Architectures*: Implementing Graph Neural Networks (GNNs) to learn directly from molecular graphs, potentially capturing high-order relationships that manual TIs miss.
- *Biological Application*: Extending the current QSPR framework to QSAR (Quantitative Structure-Activity Relationship) modeling to predict pharmacological activity.

## Declarations

### Data Availability and Computational Details

The datasets used in this study were obtained from ChemSpider and PubChem, and all relevant data are included within the manuscript. The scripts and algorithms for data processing and model implementation are provided in the *Limitations and Future Work* section to ensure reproducibility of the results. All computations were performed using Python 3.12.7 with standard scientific libraries. The models were executed on a high-performance computing (HPC) server equipped with an Intel Xeon processor and 64 GB of RAM. The implementation code for this study is hosted on Figshare at: https://doi.org/10.6084/m9.figshare.28903214 and https://doi.org/10.6084/m9.figshare.32236647.

### Funding

The authors conducted this research without any funding, grants, or support.

### Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

### Author Contributions

**Negar Kheirkhahan** contributed to conceptualization, methodology, formal analysis, software implementation, data curation, writing of the original draft, and preparation of the numerical results. **Masoud Ghods** contributed to supervision, validation, review and editing of the manuscript, and the overall scientific guidance of the study. All authors read and approved the final manuscript.

### Artificial Intelligence Statement

Artificial intelligence (AI) tools, including large language models, were used solely for language editing and improving readability. AI tools were not used for generating ideas,

performing analyses, interpreting results, or writing the scientific content. All scientific conclusions and intellectual contributions were made exclusively by the authors.

**Publisher's Note**

The publisher remains neutral regarding jurisdictional claims in published maps and institutional affiliations.

## References

[1] Abubakar, M.S., Aremu, K.O., Aphane, M., Amusa, L.B. (2024). "A QSPR analysis of physical properties of antituberculosis drugs using neighbourhood degree-based topological indices and support vector regression". *Heliyon*, 10(7), e28260. https://doi.org/10.1016/j.heliyon.2024.e28260

[2] Breiman, L. (2001). "Random forests". *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

[3] Craske, M.G., Stein, M.B., Eley, T.C., Milad, M.R., Holmes, A., Rapee, M. R., Wittchen, H.-U. (2017). "Anxiety disorders". *Nature Reviews Disease Primers*, 3, 17024. https://doi.org/10.1038/nrdp.2017.24

[4] ChemSpider. (2021). "Search and share chemistry". *Royal Society of Chemistry*. https://www.chemspider.com/

[5] Fajtlowicz, S. (1988). "On conjectures of Graffiti". *Discrete Mathematics*, 72(1–3), 113–118. https://doi.org/10.1016/0012-365X(88)90199-9

[6] Ghorbani, M., Hosseinzadeh, M.A. (2012). "A new version of Zagreb indices". *Filomat*, 26, 93–100. https://doi.org/10.2298/FIL1201093G

[7] Havare, Ö.Ç. (2022). "Quantitative structure analysis of some molecules in drugs used in the treatment of COVID-19 with topological indices". *Polycyclic Aromatic Compounds*, 42, 5249–5260. https://doi.org/10.1080/10406638.2021.1934045

[8] Hettema, J.M., Neale, M.C., Kendler, K.S. (2001). "A review and meta-analysis of the genetic epidemiology of anxiety disorders". *American Journal of Psychiatry*, 158(10), 1568–1578. https://doi.org/10.1176/appi.ajp.158.10.1568

[9] Huang, L., Wang, Y., Pattabiraman, K., Danesh, P., Siddiqui, M.K., Cancan, M. (2023). "Topological indices and QSPR modeling of new antiviral drugs for cancer treat-

ment". *Polycyclic Aromatic Compounds*, 43, 8147–8170. https://doi.org/10.1080/10406638.2022.2145320

[10] Kansal, N., Garg, P., Singh, O. (2023). "Temperature-based topological indices and QSPR analysis of COVID-19 drugs". *Polycyclic Aromatic Compounds*, 43, 4148–4169. https://doi.org/10.1080/10406638.2022.2086271

[11] Kosari, S. (2023). "On spectral radius and Zagreb Estrada index of graphs". *Asian-European Journal of Mathematics*, 16(10), 2350176. https://doi.org/10.1142/S1793557123501760

[12] Kosari, S., Dehgardi, N., Khan, A. (2023). "Lower bound on the KG-Sombor index". *Communications in Combinatorics and Optimization*, 8(4), 751–757. https://doi.org/10.22049/CCO.2023.28666.1662

[13] Kulli, V.R. (2019). "Computation of some temperature indices of $HC_5C_5$ [p,q] nanotubes". *Annals of Pure and Applied Mathematics*, 20(2), 69–74. https://doi.org/10.22457/apam.639v20n2a4

[14] Kulli, V. (2025). "Computation of inverse sum indeg uphill index and its polynomial of certain graphs". *International Journal of Mathematics and Computer Research*, 13(06), 5346–5350. https://doi.org/10.47191/ijmcr/v13i6.10

[15] Kulli, V.R., Pal, M., Samanta, S., Pal, A. (2020). *Handbook of Research on Advanced Applications of Graph Theory in Modern Society*. IGI Global. https://doi.org/10.4018/978-1-5225-9380-5.ch015

[16] Mikaeyl Nejad, S. (2025). "Hybrid of CNN and SVM for cancer type prediction". *Control and Optimization in Applied Mathematics*, 10(1), 73–89. https://doi.org/10.30473/coam.2025.72710.1269

[17] Mondal, S., Dey, A., De, N., Pal, A. (2021). "QSPR analysis of some novel neighbourhood degree-based topological descriptors". *Complex & Intelligent Systems*, 7, 977–996. https://doi.org/10.1007/s40747-020-00262-0

[18] Murphy, K.P. (2022). *Probabilistic Machine Learning: An Introduction*. MIT Press. https://probml.github.io/pml-book/book1.html

[19] Narayankar, K.P., Kahsay, A.T., Selvan, D. (2018). "Harmonic temperature index of certain nanostructures". *International Journal of Mathematics Trends and Technology*, 56(8), 575–582. https://doi.org/10.14445/22315373/IJMTT-V56P523

[20] Ramezani Tousi, J., Ghods, M. (2023). "Computing $K$ Banhatti and $K$ hyper Banhatti indices of titania nanotubes $TiO_2[m,n]$". *Journal of Information and Optimization Sciences*, 44(1), 207–216. https://doi.org/10.47974/JIOS-1130

[21] Ramezani Tousi, J., Ghods, M. (2024). "Investigating Banhatti indices on the molecular graph and the line graph of glass with M-polynomial approach". *Proyecciones (Antofagasta)*, 43(1), 199–219. https://doi.org/10.22199/issn.0717-6279-5594

[22] Rauf, A., Naeem, M., Ramzan, R., Cham, A. (2024). "Exploring physicochemical characteristics of cyclodextrin through M-polynomial indices". *Scientific Reports*, 14, 200. https://doi.org/10.1038/s41598-024-68775-z

[23] Ravi, V., Desikan, K. (2023). "Curvilinear regression analysis of benzenoid hydrocarbons and computation of some reduced reverse degree-based topological indices". *Scientific Reports*, 13, 3239. https://doi.org/10.1038/s41598-023-28416-3

[24] Sadati, S., Talebi, A.A. (2023). "A description of connectivity indices in a cubic fuzzy graph with an application". *Journal of Multiple-Valued Logic and Soft Computing*, 40(5–6), 541-568. https://www.oldcitypublishing.com/journals/mvlsc-home/mvlsc-issue-contents/mvlsc-volume-40-number-5-6-2023/mvlsc-40-5-6-p-541-568/

[25] Shahabi, M., Rahbarnia, F. (2026). "A metaheuristic and LP-based approach to irregular face coloring in planar graphs". *Control and Optimization in Applied Mathematics* و 11(1), 141-151. https://doi.org/10.30473/coam.2025.75246.1325

[26] Shi, X., Cai, R., Ramezani Tousi, J., Talebi, A.A. (2024). "Quantitative structure–property relationship analysis in molecular graphs of anticancer drugs". *Mathematics*, 12(13), 1953. https://doi.org/10.3390/math12131953

[27] Shi, X., Kosari, S., Ghods, M., Kheirkhahan, N. (2025). "Innovative approaches in QSPR modelling using topological indices for cancer treatments". *PLOS ONE*, 20(2), e0317507. https://doi.org/10.1371/journal.pone.0317507

[28] Vapnik, V.N. (1995). *The Nature of Statistical Learning Theory*. Springer. https://doi.org/10.1007/978-1-4757-2440-0

[29] Zhang, Y., Khalid, A., Siddiqui, M. K., Rehman, H., Ishtiaq, M., Cancan, M. (2023). "On analysis of temperature-based topological indices of COVID-19 drugs". *Polycyclic Aromatic Compounds*, 43, 3810–3826. https://doi.org/10.1080/10406638.2022.2080238

[30] Zhang, X., Saif, M. J., Idrees, N., Kanwal, S., Parveen, S., Saeed, F. (2023). "QSPR analysis of drugs for schizophrenia using topological indices". *ACS Omega*, 8(44), 41417–41426. https://doi.org/10.1021/acsomega.3c05000

**Authors Bio-sketches**

**Negar Kheirkhahan** is a Ph.D. candidate in Applied Mathematics at Semnan University. Her research interests include applied mathematics and computational methods.

**Masoud Ghods** is a faculty member at Semnan University. His research interests include applied mathematics, computational methods, and graph theory. Corresponding author: Email: mghods@semnan.ac.ir