Punjab University Journal of Mathematics (2025), 57(06), 652-669
https://doi.org/10.52280/pujm.2025.57(06)03

A Concise *Bond-Distance* Summation Descriptor for Effective Melting Point Prediction of Boronic Acids

Received
25 August, 2025

Revised
06 November, 2025

Accepted
10 November, 2025

Published Online
28 November, 2025

*Muhammad Zia Afzal
Department of Mathematics,
University of Central Punjab, Pakistan

Shahid Saeed Siddiqi
Department of Mathematics,
University of Central Punjab, Pakistan

Abstract. Predicting the melting points of boronic acids is crucial for guiding synthetic strategies and understanding their physicochemical behaviors. In this study, we introduce a novel *bond-distance summation descriptor*, a concise 20-component vector that numerically encodes the molecular structure by summing atomic numbers over the shortest paths from the boron atom. We benchmarked this descriptor against four established feature extraction methods Coulomb Matrix, Mordred, Morgan Fingerprints, and Molecular ACCess System (MACCS) and evaluated the predictive accuracy of five machine learning models: Decision Tree, Random Forest, XGBoost, LightGBM, and Support Vector Machine. Despite having far fewer features than the high-dimensional Mordred and Morgan representations, our 20-length descriptor achieves competitive results, particularly when paired with XGBoost, which consistently exhibits superior performance in terms of Mean Absolute Error (MAE) and R^2 score. These findings underscore the potential of a concise, interpretable descriptor for effective melting point prediction, paving the way for the future integration of this scheme into broader cheminformatics applications.

AMS (MOS) Subject Classification Codes: 35S29; 40S70; 25U09
Key Words: molecular descriptors, boronic acids, melting point, machine learning algorithms.

1. Introduction

The accurate prediction of physicochemical properties, such as melting points, is a cornerstone challenge in computational chemistry and materials science. Melting points provide essential insight into a compound's stability, phase behavior, and potential applications, and they play a pivotal role in areas ranging from drug discovery to material design. Traditional experimental methods for determining melting points are often laborious and

*Corresponding Author:l1f22phma0001@ucp.edu.pk resource-intensive, which has spurred the development of computational approaches capable of predicting these properties with greater efficiency.

Among the various classes of organic compounds, boronic acids have attracted significant attention due to their unique chemical properties and broad applications in medicinal chemistry, organic synthesis, and materials science. Characterized by the presence of a boron-containing functional group, $-B(OH)_2$, boronic acids are integral to Suzuki coupling reactions and other key chemical transformations [17, 22, 23]. Their versatile reactivity and potential therapeutic applications, including enzyme inhibition and anticancer activity, have been widely explored [14, 17, 18].

However, the accurate prediction of melting points remains challenging because of the complex interplay of molecular structure, intermolecular interactions, steric effects, and crystal packing. Traditional quantitative structure-property relationship (QSPR) models, such as group contribution methods and geometry based approaches, often struggle with these complexities [10, 18, 23]. Recent advances in machine learning (ML) offer a promising alternative, enabling the capture of non-linear relationships between molecular features and melting points that are difficult to model using conventional methods [1, 2, 7, 8, 11, 21].

Beyond general organic compounds, boronic acids have also been evaluated in context-specific QSPR studies involving melting and boiling points [6, 28]. Other approaches, such as hybrid ML-QSPR pipelines, offer enhanced transferability [9, 12]. Boronic acid-containing compounds have also been considered in molecular design for anticancer agents and deep eutectic solvents, demonstrating the broad predictive utility of ML in such areas [5, 18, 24].

In this study, we introduce a novel bond-distance summation descriptor that aggregates atomic numbers along all shortest paths from a reference boron atom to atoms at predefined bond-distance thresholds (e.g., 2, 3) by applying multipliers for double, triple, and aromatic bonds (2, 3, and 1.5, respectively). This descriptor is concise, consisting of a fixed-length vector of only 20 components; however, it captures the subtle electronic and steric effects that govern melting behavior. We benchmark this descriptor against established molecular representations including Coulomb Matrix, Mordred 3D descriptors, Morgan fingerprints, and MACCS fingerprints (which contain 166 descriptors) while the other methods typically yield feature sets exceeding one thousand descriptors [7, 16, 20, 26, 30].

Our work employs five different ML algorithms Decision Tree (DT), Random Forest (RF), XGBoost, LightGBM, and Support Vector Regression (SVR) to model the melting points of boronic acids. We compared the performance of our bond-distance summation descriptor with established representations using these algorithms, demonstrating that our concise 20-component descriptor achieves competitive accuracy. Recent studies have demonstrated the potential of ML in catalysis, crystal structure prediction, solvent screening, reaction yield optimization, and melting point estimation [3, 4, 9, 11, 18, 27, 29, 31, 32].

The ability of descriptors to reflect electronic, steric, and intermolecular interactions including hydrogen bonding and π-stacking is crucial in modeling melting points of boronic acids [11, 17]. Traditional descriptors often include topological indices, volume-based parameters, and dipole moments [18, 19]. Our summation descriptor provides an interpretable yet concise representation, avoiding the curse of dimensionality while maintaining physical relevance.

Furthermore, methodological advances such as active learning [15], deep neural networks [9], and literature-bias-aware ML models [7] have revealed important design rules that improve generalization. Ensemble learning and hybrid representations continue to outperform single-model baselines in tasks like thermal property prediction and molecular screening [13, 25, 28].

This study aims to enhance the precision in predicting boronic acids' melting points and to showcase the effectiveness of the newly developed bond-distance summation descriptor.

2. METHODOLOGY

2.1. **Data Collection.** The dataset was sourced from `https://organoborons.com/`, containing information on 605 boronic acids with recorded melting points.

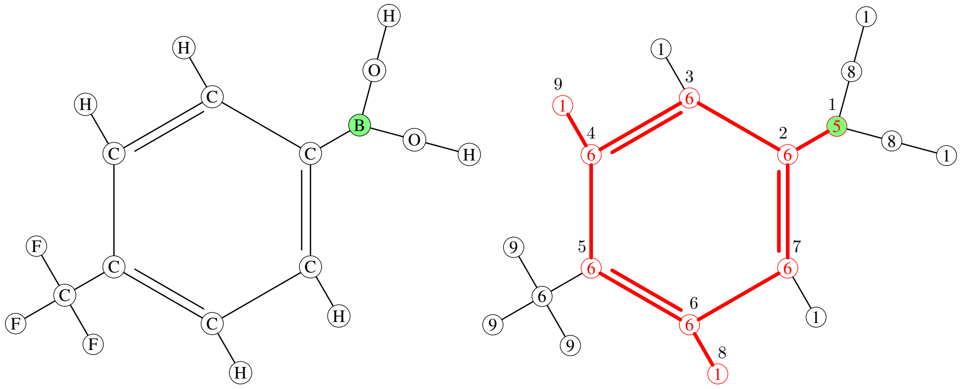
2.2. **Molecular Descriptor Calculation.** In our descriptor scheme, each component of the descriptor vector is calculated by summing the atomic numbers along all shortest paths from the reference boron atom to all atoms exactly d bonds away. For each bond along a path, the atomic number is multiplied by a factor depending on the bond order:

- Non-aromatic single bond: $\times 1$
- Non-aromatic double bond: $\times 2$
- Non-aromatic triple bond: $\times 3$
- Aromatic bond: $\times 1.5$

All atoms, including hydrogens, are explicitly included in the bond-distance summation to ensure that both heavy-atom and peripheral hydrogen contributions to the local electronic and steric environment are represented. The bond-distance summation and descriptor generation were implemented using in-house Python scripts built upon RDKit cheminformatics routines, enabling reproducibility through a deterministic graph traversal procedure.

Physically, this descriptor encodes how the atomic number and bond order propagate outward from the boron center, effectively capturing both the electronic effects (through heavier atoms and bond multiplicities) and steric effects (through molecular topology and branching). This design allows the descriptor to reflect the cumulative structural influence of substituents and bonding environment on the melting point of boronic acids.

We will detail the method for acquiring the descriptor of 4-Trifluoromethylphenylboronic acid as depicted in Figure 1a:



(A) 4-Trifluoromethylphenylboronic acid

(B) Highlighted 4-bond paths

FIGURE 1. Schematic representation of the molecular structure and bond-path analysis for 4-trifluoromethylphenylboronic acid. (a) shows the complete molecular structure with the boronic acid functional group (-B(OH)$_2$) and the trifluoromethyl substituent (-CF$_3$) on the aromatic ring. (b) highlights the four-bond connectivity paths extending from the central boron atom through the aromatic framework, which form the basis for computing atom-centered topological descriptors in the study. The red-highlighted atoms and bonds indicate the atomic environment captured within a four-bond distance, used to encode the boron-centered local structure during descriptor generation.

Figure 1b illustrates the atoms situated four bonds distant from Boron, considering these four shortest pathways:

$$\text{Path 1: } (12768)\ 5 + 6 + 6 \times 1.5 + 6 \times 1.5 + 1 = 30,$$

$$\text{Path 2: } (12349)\ 5 + 6 + 6 \times 1.5 + 6 \times 1.5 + 1 = 30,$$

$$\text{Path 3: } (12765)\ 5 + 6 + 6 \times 1.5 + 6 \times 1.5 + 6 \times 1.5 = 38,$$

$$\text{Path 4: } (12345)\ 5 + 6 + 6 \times 1.5 + 6 \times 1.5 + 6 \times 1.5 = 38.$$

Thus, the descriptor value for the 4-bond component is:

$$30 + 30 + 38 + 38 = 136.$$

Since the longest descriptor length over our entire dataset is 20, every descriptor vector is defined to have 20 components. For molecules with a maximum bond distance less than 20, the remaining components are padded with 0.

Figure 2 illustrates this fixed-length descriptor, with the first five components corresponding to bond distances 2, 3, 4, 5, and 6 (here, 68, 100, 136, 88, and 318, respectively), and the remaining positions filled with 0.

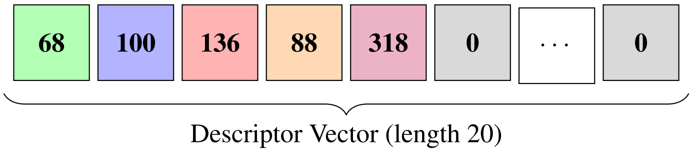


FIGURE 2. Compound description for 4-(Trifluoromethyl)phenylboronic acid

Notice that the descriptor value for a bond at a distance of one remains consistently identical; therefore, it has been omitted. The descriptor's representation is unaffected by the labeling of vertices, thus ensuring it is invariant under permutations.

3. MACHINE LEARNING METHODOLOGY

To predict the **melting points of Boronic Acids**, we perform a comparative analysis using five different machine learning models:

- **Decision Tree (DT)**
- **Random Forest (RF)**
- **XGBoost**
- **LightGBM**
- **Support Vector Machine (SVM)**

These models were selected for their robust performance in handling **non-linear relationships** and **high-dimensional molecular descriptors**. Our objective is to compare different descriptor extraction methods including our custom bond-distance summation descriptor, Mordred 3D descriptors, MACCS fingerprints, Coulomb matrices, and various configurations of Morgan fingerprints while evaluating the performance of different machine learning models (DT, RF, XGBoost, LightGBM, and SVR) in predicting melting points.

3.1. Model Selection and Justification.

- **Decision Tree (DT):** A rule-based model that serves as a **baseline**, providing insights into how simple partitioning can capture descriptor-melting point relationships.
- **Random Forest (RF):** An ensemble of multiple decision trees that reduces variance and enhances robustness by aggregating predictions over bootstrapped samples.
- **XGBoost:** A boosting algorithm that **iteratively refines weak learners** using regularization and parallelization, thereby improving performance.
- **LightGBM:** A gradient boosting framework optimized for speed and large feature spaces. Its **leaf-wise** tree growth strategy improves efficiency.
- **Support Vector Machine (SVM):** Although not tree-based, **Support Vector Regression (SVR)** is included for its ability to model complex, non-linear relationships via the RBF kernel.

By systematically comparing these models in conjunction with different descriptor extraction methods, we aim to determine the most effective approach for predicting melting points.

3.2. **Visual Representation of the Models.** Figure 3 provides a general overview of tree-based algorithms. The diagram starts with a basic Decision Tree and illustrates how ensemble methods (Bagging and Boosting) extend this basic model into more robust algorithms like Random Forest, XGBoost, and LightGBM.

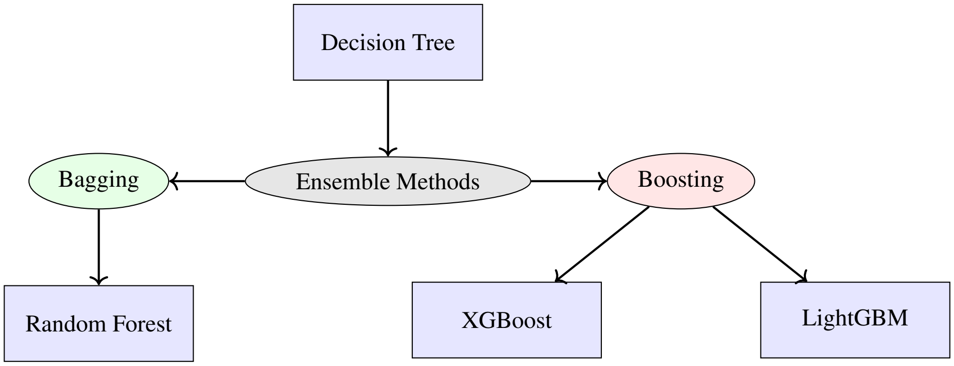


FIGURE 3. General Representation of Tree-Based Algorithms

Figure 4 illustrates a basic decision tree structure, showing how the root node is split into internal nodes and ultimately leads to leaf nodes that provide the final predictions. This simple decision tree forms the foundation for the more complex ensemble techniques described above.

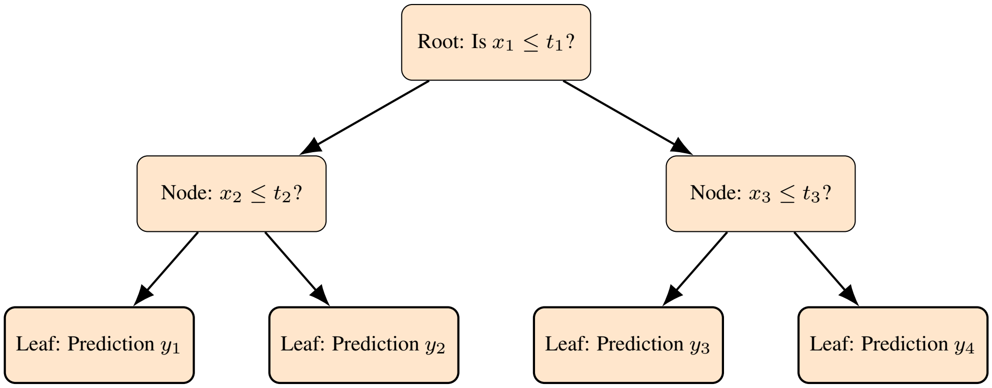


FIGURE 4. Decision Tree Diagram

Short Explanation of the Decision Tree Diagram. The decision tree diagram shows:

- A **Root Node** initiating the decision process.
- **Internal Nodes** splitting the data based on specific thresholds.
- **Leaf Nodes** that yield the final predictions.

This structure underpins more advanced ensemble methods such as Random Forest, XGBoost, and LightGBM.

Our study utilizes a novel bond-distance summation descriptor, which numerically encodes molecular structure, and compares it against established representations (Mordred 3D descriptors, MACCS fingerprints, Coulomb matrices, and various Morgan fingerprint configurations). All models are trained on the same dataset (80% training and 20% validation) and evaluated using standard metrics (MAE and R^2 Score).

Figure 5 shows the various descriptor extraction methods, with our bond-distance summation descriptor highlighted. Figure 6 illustrates the complete machine learning pipeline?from descriptor extraction and dataset splitting to the training of multiple models (DT, RF, XGB, LGBM, and SVR) and their subsequent evaluation.

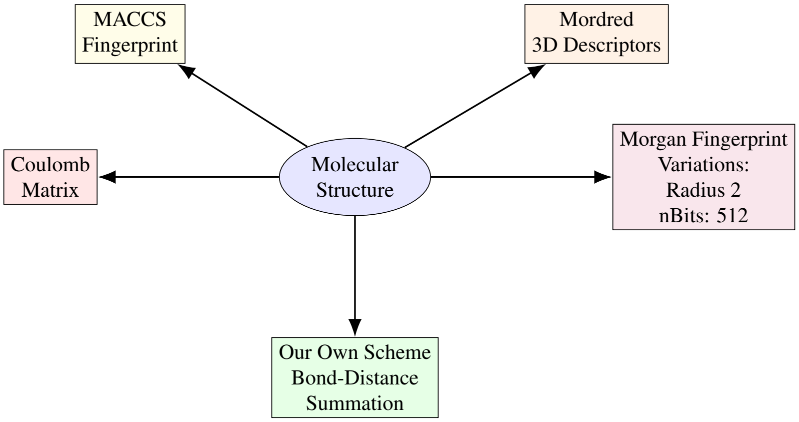

MACCS Fingerprint
Mordred 3D Descriptors
Coulomb Matrix
Molecular Structure
Morgan Fingerprint Variations: Radius 2 nBits: 512
Our Own Scheme Bond-Distance Summation

FIGURE 5. Descriptor Extraction Methods: Our bond-distance summation descriptor is compared against established representations.

4. RESULTS AND DISCUSSION

This section presents a comparative analysis of the five descriptors (Coulomb Matrix, Mordred, Morgan Fingerprint, MACCS, and our Bond-Distance Summation Descriptor) across multiple machine learning models. Table 1 summarizes the **Mean Absolute Errors (MAE)**, while Table 2 reports the corresponding **R^2** values. Lower MAE values indicate more accurate predictions, whereas higher R^2 values signify better explanatory power of the model with respect to the observed melting points.

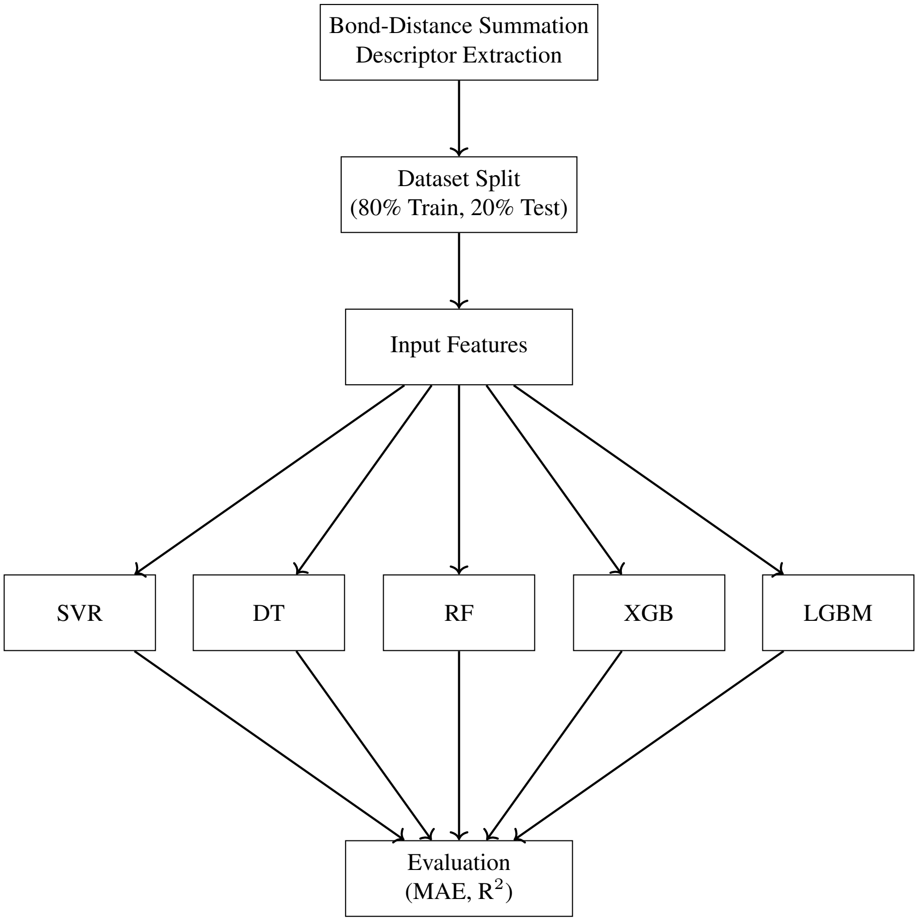


FIGURE 6. Pictorial Representation of the Machine Learning Pipeline for Melting Point Prediction.

- **MAE (Table 1):** The mean absolute error (MAE) values clearly indicate that the *XGBoost* regressor consistently provides the most accurate melting point predictions across nearly all descriptor types. Among all tested molecular representations, the *Mordred* descriptor produces the lowest MAE values overall, achieving 7.26 °C when coupled with XGBoost. This can be attributed to Mordred?s extensive feature set, which captures a diverse range of 2D and 3D structural, topological, and physicochemical attributes relevant to thermodynamic properties.

The proposed *Bond-Distance Summation Descriptor* (referred to as “Descriptor” in Table 1) performs competitively, yielding MAE values comparable to the

Coulomb Matrix and better than MACCS fingerprints in several models. This suggests that our descriptor successfully encodes local electronic and steric information centered around the boron atom, which plays a pivotal role in determining the intermolecular interactions and packing efficiency that influence melting points.

Classical models such as *Decision Trees (DT)* and *Random Forests (RF)* achieve moderate accuracy but show higher variability across descriptor types. LightGBM performs well for Mordred and Coulomb Matrix but less consistently for Morgan and MACCS, potentially due to its sensitivity to sparse or binary feature distributions. In contrast, *Support Vector Machines (SVM)* exhibit significantly higher MAE values across all representations, likely reflecting limitations of kernel-based methods in capturing nonlinear structure-property relationships in high-dimensional molecular spaces.

- **R^2 (Table 2):** The coefficient of determination (R^2) results reinforce the MAE trends, demonstrating that the XGBoost model provides the best fit to experimental melting point data. XGBoost achieves $R^2 = 0.89$ for both Mordred and the proposed descriptor, indicating that these feature spaces effectively explain a large portion of the observed variance in melting points.

  The Coulomb Matrix and the new Bond-Distance Summation Descriptor perform comparably, with R^2 values around 0.82, confirming that both capture essential molecular-level interactions influencing phase-transition behavior. Notably, the relatively strong performance of the new descriptor compared to conventional fingerprints (Morgan and MACCS) highlights its chemical interpretability and its ability to retain key structure-property relationships centered on the boron atom.

  Models based on ensemble methods (RF, LGBM, and XGBoost) consistently outperform simpler or kernel-based algorithms, emphasizing the advantage of gradient-boosting frameworks in optimizing nonlinear regressions with complex descriptor sets. Overall, these findings demonstrate that coupling our physically inspired descriptor with modern boosting algorithms yields performance on par with widely used, high-dimensional descriptors like Mordred, while maintaining a more interpretable and chemically meaningful feature representation.

Overall, XGBoost demonstrates the best performance among the tested models, consistently yielding lower MAE and higher R^2 scores across most descriptors. To further illustrate the predictive capability of XGBoost, Figure 7 compares **observed vs. predicted melting points** for each descriptor under XGBoost. The blue data points represent predictions closely matching the diagonal reference line, whereas red points indicate higher deviations. As shown, Mordred and our Bond-Distance Summation Descriptor exhibit closer clustering around the diagonal, reflecting strong predictive performance.
In summary, these results confirm that:

(1) **XGBoost** stands out for its balance of accuracy (MAE) and explanatory power (R^2).
(2) The **Mordred** and **Bond-Distance Summation Descriptor** consistently rank among the top performers, highlighting their suitability for capturing relevant molecular features.

TABLE 1. Comparison of mean absolute errors (MAE, in °C) for different regression models trained on various molecular descriptors. Lower MAE indicates higher predictive accuracy.

	CM	Mordred	MFP	MACCS	Descriptor
DT	11.16	9.88	10.17	15.35	11.54
RF	23.37	17.56	19.75	21.61	22.24
LGBM	11.13	9.97	27.79	28.26	28.18
XGBoost	9.43	7.26	13.84	16.71	11.42
SVM	36.11	36.15	35.57	34.46	35.86

TABLE 2. Coefficient of determination (R^2) values for the same models and descriptors, showing the proportion of variance in melting points explained by each model.

	CM	Mordred	MFP	MACCS	Descriptor
DT	0.75	0.80	0.81	0.74	0.77
RF	0.77	0.86	0.81	0.78	0.77
LGBM	0.86	0.89	0.69	0.68	0.68
XGBoost	0.82	0.89	0.85	0.76	0.82
SVM	0.61	0.61	0.63	0.64	0.61

(3) The **Coulomb Matrix** and **Morgan Fingerprints** also provide competitive performance in some cases, whereas **MACCS** tends to lag behind, particularly in terms of MAE.

4.1. **Hyperparameter Tuning and Model Optimization.** To ensure robust and unbiased estimation of the predictive performance of the XGBoost regressor across all molecular descriptor types, a nested cross-validation (CV) framework was implemented. The outer CV loop was used for model evaluation, while the inner CV loop performed random hyperparameter search with early stopping for hyperparameter optimization.

4.1.1. *Experimental Setup.* All experiments were executed using the XGBoost regressor (`XGBRegressor`) configured for regression with a squared error objective. To avoid overfitting and to ensure computational reproducibility, a fixed random seed (`RANDOM_STATE = 42`) was used throughout.

The nested CV procedure was defined as follows:

- **Outer loop:** 5-fold cross-validation for independent model evaluation.
- **Inner loop:** 3-fold cross-validation for hyperparameter optimization using random search.

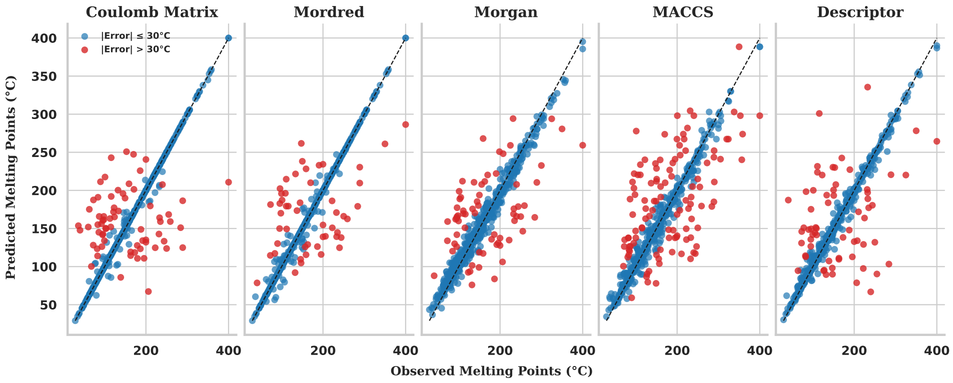


FIGURE 7. Comparison of predicted and experimental melting points for boronic acids using five descriptor sets (Coulomb Matrix, Mordred, Morgan, MACCS, and composite). Blue points denote accurate predictions ($\leq 30°C$ error), red points indicate outliers, and the dashed line represents the ideal parity line.

- **Evaluation metric:** Mean Absolute Error (MAE) was used for optimization and comparison.
- **Early stopping:** 30 rounds of no improvement on the validation loss triggered early termination during training.

Prior to training, all datasets were standardized to remove missing values and low-variance features using a variance threshold of 1×10^{-5}. For high-dimensional descriptors such as Mordred, the feature set was further reduced to the number of samples using correlation-based feature selection to accelerate model convergence.

4.1.2. *Hyperparameter Search Space.* Random search was conducted over the following XGBoost hyperparameters:

- `n_estimators` $\in \{200, 400, 800\}$
- `max_depth` $\in \{3, 5, 7, 9\}$
- `learning_rate` $\in \{0.01, 0.03, 0.05, 0.1\}$
- `subsample` $\in \{0.6, 0.8, 1.0\}$
- `colsample_bytree` $\in \{0.6, 0.8, 1.0\}$
- `reg_alpha` $\in \{0.0, 1.0\}$ (L1 regularization)
- `reg_lambda` $\in \{1.0, 3.0\}$ (L2 regularization)

Each outer fold evaluated the mean and standard deviation of validation MAE across 50 randomly sampled configurations from this parameter space. The best-performing configuration was retrained on the full outer training set and evaluated on the outer test split.

4.1.3. *Model Performance Summary.* The aggregated results of nested cross-validation for all descriptor types are summarized in Table 3. Mordred descriptors achieved the best overall performance, with the lowest mean MAE (38.14) and highest R^2 score (0.46), indicating their strong correlation with melting point prediction. The Coulomb matrix descriptor yielded the weakest performance, suggesting limited expressivity for boronic acid systems.

TABLE 3. Summary of Nested Cross-Validation Results for XGBoost Across Different Descriptor Types.

Descriptor	**MAE$_{\mathbf{mean}}$**	**MAE$_{\mathbf{std}}$**	**RMSE$_{\mathbf{mean}}$**	$R^2_{\mathbf{mean}}$
MACCS	39.65	2.40	52.85	0.37
Morgan	38.44	1.44	51.41	0.41
Coulomb Matrix	47.33	4.13	60.86	0.17
Mordred	**38.14**	3.86	**49.21**	**0.46**
Descriptor	44.40	3.50	58.32	0.24

4.1.4. *Per-Fold Optimization Details.* Table 4 lists representative hyperparameter configurations selected during nested CV for each descriptor type. These results demonstrate that the model typically favored shallow trees (`max_depth = 3−7`) and moderate learning rates (0.03-0.05) across most descriptor sets, balancing bias-variance trade-offs efficiently.

TABLE 4. Representative Fold-wise Hyperparameter Settings and MAE Values (Example Shown for Mordred Descriptors).

Fold	**MAE**	**RMSE**	R^2	**Best Inner MAE**	**max_depth**	**learning_rate**	**subsample**
1	34.07	45.85	0.48	36.15	5	0.05	1.0
2	39.27	48.85	0.38	35.87	5	0.03	0.8
3	35.09	43.93	0.53	35.58	5	0.03	0.6
4	43.84	57.26	0.45	34.71	3	0.05	0.8
5	38.41	50.17	0.46	35.77	9	0.05	0.6

4.1.5. *Discussion.* The nested cross-validation approach effectively reduced bias and variance in model performance estimates, particularly by avoiding information leakage between hyperparameter tuning and evaluation. Among the tested molecular representations, the 3D Mordred descriptors provided the highest predictive accuracy for boronic acid melting points, followed closely by Morgan fingerprints. The custom graph-based descriptor (`Boron_En`) also demonstrated competitive performance, highlighting the role of bond path information centered around the boron atom. These results validate the use of hybrid molecular descriptors combined with tree-based ensemble learning for accurate thermochemical property prediction.

4.2. **Statistical Analysis of Descriptor Performance.** To determine whether the predictive performance of XGBoost models trained on different descriptor sets differed significantly, statistical significance testing was carried out. Since the absolute prediction errors did not follow a normal distribution (as confirmed by the Shapiro-Wilk test, $p < 0.05$ for all descriptors), non-parametric statistical methods were employed.

The Friedman test, a non-parametric alternative to repeated-measures ANOVA, was used to compare the mean ranks of absolute errors across all descriptor-based models. Upon observing a statistically significant Friedman statistic ($p < 0.05$), post-hoc pairwise comparisons were conducted using the Nemenyi test to identify which descriptor pairs differed significantly.

All statistical analyses were implemented in Python (version 3.12) using the `SciPy` and `scikit-posthocs` libraries. Visualization of error distributions and mean performance differences was performed using `Matplotlib` and `Seaborn`. The resulting figures provide a graphical summary of model performance dispersion and descriptor-wise ranking in melting point prediction of boronic acids.

4.2.1. *Error Distribution and Normality Testing.* Figure 8 illustrates the distribution of absolute prediction errors for all descriptors. The Mordred and CoulombMatrix descriptors exhibit notably lower median errors, suggesting superior predictive power compared to the MACCS and Morgan fingerprints.

To assess whether the errors followed a normal distribution, the Shapiro-Wilk test was applied to each descriptor's error distribution. The results ($p = 0.0000$ for all cases) indicated significant deviation from normality, thus justifying the use of non-parametric statistical methods for further analysis.

4.2.2. *Friedman and Post-hoc Nemenyi Tests.* A Friedman test, which is a non-parametric alternative to repeated-measures ANOVA, was conducted to evaluate whether statistically significant differences exist among descriptor performances. The test produced a statistic of 1246.7412 with a *p*-value of 0.000000, confirming significant performance differences ($p < 0.05$) across descriptors.

Subsequently, a Nemenyi post-hoc test was carried out to identify pairwise differences. Table 5 presents the pairwise *p*-values, showing that Mordred and CoulombMatrix descriptors significantly outperformed MACCS and Morgan fingerprints, while their differences with the proposed Descriptor set were also statistically significant.

4.2.3. *Average Error Comparison.* Figure 9 displays the average absolute errors for each descriptor type. The Mordred descriptor achieved the lowest mean error (7.26 °C), followed by CoulombMatrix (9.43 °C), indicating their robustness in capturing structure-property relationships. The summary of mean and standard deviation of absolute errors is reported in Table 6.

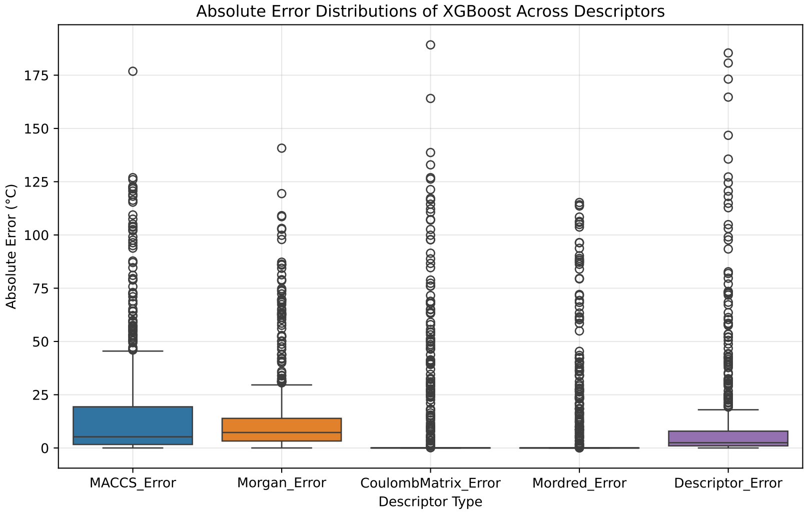


FIGURE 8. Absolute error distributions of XGBoost models trained on different descriptors.

TABLE 5. Pairwise Nemenyi post-hoc test *p*-values among descriptor errors.

	MACCS	Morgan	Coulomb Matrix	Mordred	Descriptor
MACCS	1.0000	0.1254	0.0000	0.0000	0.00003
Morgan	0.1254	1.0000	0.0000	0.0000	0.00000
Coulomb Matrix	0.0000	0.0000	1.0000	0.00000	0.0000
Mordred	0.0000	0.0000	0.00000	1.0000	0.0000
Descriptor	0.00003	0.00000	0.0000	0.0000	1.0000

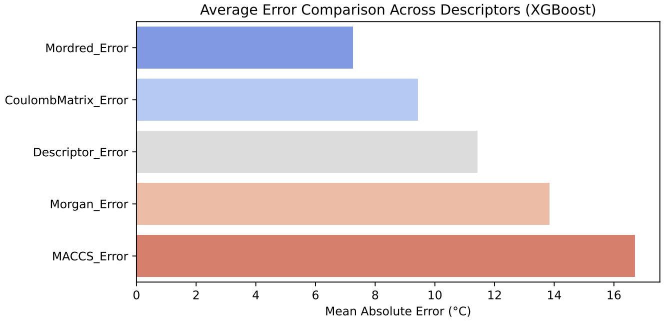


FIGURE 9. Average error comparison across descriptors using XGBoost models.

TABLE 6. Summary of mean and standard deviation of absolute errors across descriptors.

Descriptor	Mean Error (°C)	Std. Dev. (°C)
MACCS	16.71	26.19
Morgan	13.84	19.82
CoulombMatrix	9.43	25.84
Mordred	7.26	20.60
Descriptor	11.42	25.13

4.2.4. *Interpretation.* The results demonstrate that the differences in model performance across descriptor types are statistically significant. Specifically, Mordred and CoulombMatrix descriptors yield consistently lower prediction errors, indicating their effectiveness in encoding chemically relevant information for boronic acid melting point prediction. Conversely, MACCS and Morgan fingerprints, while simpler, were less capable of capturing fine structural variations.

Overall, the statistical evaluation confirms that descriptor selection has a significant impact on predictive accuracy, with complex 3D and bond-energy-based descriptors outperforming conventional fingerprint methods.

5. Conclusion

In this study, we compared five different molecular descriptors Coulomb Matrix, Mordred, Morgan Fingerprints, MACCS, and our novel Bond-Distance Summation Descriptor across a range of machine learning models for predicting the melting points of boronic acids. Despite MACCS having only 166 descriptors, the other three established methods (Coulomb Matrix, Mordred, and Morgan) each generate well over a thousand descriptors. In contrast, our bond-distance summation approach uses a fixed-length vector of just 20 components.

The results demonstrated XGBoost consistently achieved superior predictive accuracy, as evidenced by lower Mean Absolute Errors (MAE) and higher R^2 scores compared to other algorithms. Mordred and Bond-Distance Summation Descriptor performed particularly well in capturing key molecular features relevant to melting point prediction. Despite having a comparatively small descriptor length, our 20-component vector produced results competitive with high-dimensional descriptors such as Coulomb Matrix, Mordred, and Morgan. MACCS, which has 166 predefined keys, generally exhibited less robust performance in terms of MAE, underscoring the importance of descriptor richness or relevance.

Overall, these findings highlight the potential of a concise, interpretable descriptor to compete effectively against more complex, higher-dimensional representations. Our results show that a compact 20-component descriptor can yield reasonable predictive performance compared to high-dimensional representations. These findings contribute to the growing body of work at the intersection of machine learning and cheminformatics, paving the way for the development of efficient computational tools for the design and synthesis of boronic acid derivatives in pharmaceutical and industrial applications. Future research can explore hybrid approaches that combine the simplicity of our bond-distance summation descriptor with complementary features, potentially enhancing both interpretability and predictive power in melting point prediction and related cheminformatics tasks.

6. Declaration of competing interest

Authors declare that they do not have any competing interests.

7. Declaration of AI in scientific writing

Authors declare that they used generative artificial intelligence (AI) to improve the language and quality of this article, aligned with international standards.

8. Declaration of Funding

This study has not received financial support from any funding sources.

9. Data and Software Availability statement

The datasets and software used in this study are publicly available at GitHub `https://github.com/MZiaAfzal71/Melting-Point-Prediction-of-Boronic-Acids/`. The repository contains the curated Boronic Acids dataset with experimentally reported melting points and result files produced from the Colab notebooks. These Google Colab-runnable notebooks perform key steps such as dataset extraction, fingerprint generation, model comparison, hyperparameter tuning, and statistical analysis. Detailed execution instructions are provided within the repository to ensure transparency and reproducibility.

10. Author Contributions

Each author made an equal contribution to the preparation of this manuscript.

References

[1] M. Z. Afzal, S. S. Siddiqi, and A. R. Nizami, *Predicting molecular properties using adjacency matrix powers and atomic number sequences*, Chemical Engineering Science **320**, Part C (2026) 122650.

[2] M. Z. Afzal and S. S. Siddiqi, *A comparison of optimizers in a pytorch based artificial neural network to predict normal boiling points of alkanes*, Journal of the National Science Foundation of Sri Lanka **53**, No. 2 (2025) 165–172.

[3] R. S. A. E. Ali, J. Meng, M. E. I. Khan, and X. Jiang, *Machine learning advancements in organic synthesis: A focused exploration of artificial intelligence applications in chemistry*, Artificial Intelligence Chemistry **2**, No. 1 (2024) 100049.

[4] J. P. António, I. L. Roque, F. M. F. Santos, and P. M. P. Gois, *Designing functional and responsive molecules with boronic acids*, Accounts of Chemical Research **58**, No. 5 (2025) 673–687.

[5] B. Basha, M. H. Tahir, T. Kadyrov, N. S. Alsaiari, and M. S. Al-Buriahi, *Machine learning assisted prediction of melting points of non-fullerene electron acceptors, chemical space generation and visualization*, Chemical Engineering Science **319** (2026) 122258.

[6] Y. Beghour and Y. Lahiouel, *Using machine learning in QSPR to estimate boiling and critical temperatures*, Chemical Engineering Science **309** (2025) 121228.

[7] W. Beker, R. Roszak, A. Wolos, N. H. Angello, V. Rathore, M. D. Burke, and B. A. Grzybowski, *Machine learning may sometimes simply capture literature popularity trends: a case study of heterocyclic suzuki–miyaura coupling*, Journal of the American Chemical Society **144**, No. 11 (2022) 4819–4827.

[8] N. S. S. Capman, X. V. Zhen, J. T. Nelson, V. R. S. K. Chaganti, R. C. Finc, M. J. Lyden, T. L. Williams, M. Freking, G. J. Sherwood, P. Buhlmann, C. J. Hogan, and S. J. Koester, *Machine learning-based rapid detection of volatile organic compounds in a graphene electronic nose*, ACS nano **16**, No. 11 (2022) 19567–19583.

[9] H. Chang, Z. Zhang, J. Tian, T. Bai, Z. Xiao, D. Wang, R. Qiao, and C. Li, *Machine learning-based virtual screening and identification of fourth-generation EGFR inhibitors*, ACS Omega **9**, No. 2 (2024) 2314–2324.

[10] Y. Chen, D. Zhang, Z. Wang, M. Tang, and H. Zhang, Pb^{2+} *transfer-enabled recoverable hydrogel-based* H_2S *colorimetric sensing with assistance of multimodal deep learning for multifunctional applications*, Advanced Functional Materials **34**, No. 49 (2024) 2409017.

[11] S. Escayola, N. Bahri-Laleh, and A. Poater, *% V Bur index and steric maps: from predictive catalysis to machine learning*, Chemical Society Reviews **53** (2024) 853–882.

[12] M. U. Ghani, M. K. Maqbool, R. George, and A. E. Ofem, and M. Cancan, *Entropies via various molecular descriptors of layer structure of* H_3BO_3, Mathematics **10**, No. 24 (2022) 4831.

[13] B. B. Hansen, S. Spittle, B. Chen, D. Poe, Y. Zhang, J. M. Klein, A. Horton, L. Adhikari, T. Zelovich, B. W. Doherty, B. Gurkan, E. J. Maginn, A. Ragauskas, M. Dadmun, T. A. Zawodzinski, G. A. Baker, M. E. Tuckerman, R. F. Savinell, and J. R. Sangoro, *Deep eutectic solvents: A review of fundamentals and applications*, Chemical Reviews **121**, No. 3 (2020) 1232–1285.

[14] A. C. King, M. Woods, W. Liu, Z. Lu, D. Gill, and M. R. H. Krebs, *High-throughput measurement, correlation analysis, and machine-learning predictions for pH and thermal stabilities of Pfizer-generated antibodies*, Protein Science **20** 9 (2011) 1546–1557.

[15] B. Liang, W. Song, S. Liu, K. Li, H. Yu, P, Li, and R. Xing, *Improving the performance of chitin bioprocessing: Pretreating chitin and optimizing active enzyme*, Polymer Reviews **65**, No. 3 (2025) 777–813.

[16] A. Lopalco, V. J. Stella, and W. H. Thompson, *Origins, and formulation implications, of the* pK_a *difference between boronic acids and their esters: a density functional theory study*, European Journal of Pharmaceutical Sciences **124** (2018) 10–16.

[17] D. A. M. Muyassiroh, F. A. Permatasari, and F. Iskandar, *Machine learning-driven advanced development of carbon-based luminescent nanomaterials*, Journal of Materials Chemistry C **10**, No. 46 (2022) 17431–17450.

[18] H. Nada, A. R. Gul, A. Elkamhawy, S. Kim, M. Kim, Y. Choi, T. J. Park, and K. Lee, *Machine learning-based approach to developing potent EGFR inhibitors for breast cancer-design, synthesis, and in vitro evaluation*, ACS omega **8**, No. 35 (2023) 31784–31800.

[19] B. J. Neves, J. P. Agnes, M. d. N. Gomes, M. R. H. Donza, R. M. Gonçalves, M. Delgobo, Lauro R. d. S. Neto, M. R. Senger, F. P. Silva-Junior, S. B. Ferreira, A. Zanotto-Filho, and C. H. Andrade, *Efficient identification of novel anti-glioma lead compounds by machine learning models*, European Journal of Medicinal Chemistry **189** (2020) 111981.

[20] S. Nikolić and N. Trinajstić, *The wiener index: Development and applications*, Croatica Chemica Acta **68**, No. 1 (1995) 105–129.

[21] A. R. Nizami, S. F. Ali, M. Z. Afzal, and M. Inc, *Novel descriptors for the prediction of molecular properties*, Open Chemistry **23**, No. 1 (2025) 20250194.

[22] J. C. A. Oliveira, J. Frey, S. Q. Zhang, L. C. Xu, X. Li, S. W. Li, X, Hong, and L. Ackermann, *When machine learning meets molecular synthesis*, Trends in Chemistry **4**, No. 10 (2022) 863–885.

[23] J. G. Pereira, J. M. J. M. Ravasco, L. Bustillo, I. S. Marques, P. Y. Kao, P. Y. Li, Y. C. Lin, T. Rodrigues, V. D. B. Bonifácio, A. F. Peixoto, C. A. M. Afonso, and R. F. A. Gomes, *Active learning assists chemical intuition identify a scalable conversion of chitin to 3-acetamido-5-acetylfuran*, Green Chemistry **27** (2025) 1740–1746.

[24] A. Prabhune and R. Dey, *Green and sustainable solvents of the future: Deep eutectic solvents*, Journal of Molecular Liquids **379** (2023) 121676.

[25] A. D. A. Ramahi, V. V. Shinde, T. C. Pearce, and I. C. Sinka, *Virtual screening of drug materials for pharmaceutical tablet manufacturability with reference to sticking*, International Journal of Pharmaceutics **667**, Part A (2024) 124722.

[26] M. Randic, *Characterization of molecular branching*, Journal of the American Chemical Society **97**, No. 23 (1975) 6609–6615.

[27] H. S. Samuel, E. E. Etim, U. N. Maraizu, and S. Yakubu, *Machine learning in chemical kinetics: Predictions, mechanistic analysis, and reaction optimization*, Applied Journal of Environmental Engineering Science **10**, No. 1 (2024) 36–61.

[28] E. Shim, A. Tewari, T. Cernak, and P. M. Zimmerman, *Machine learning strategies for reaction development: toward the low-data limit*, Journal of chemical information and modeling **63**, No. 12 (2023) 3659–3668.

[29] Y. Song, J. Lindsay, Y. Zhao, A. Nasiri, S. Y. Louis, J. Ling, M. Hu, and J. Hu, *Machine learning based prediction of noncentrosymmetric crystal materials*, Computational Materials Science **183** (2020) 109792.

[30] T. Toyao, Z. Maeno, S. Takakusagi, T. Kamachi, I. Takigawa, and K. Shimizu, *Machine learning for catalysis informatics: Recent applications and prospects*, ACS Catalysis **10**, No. 3 (2019) 2260–2297.

[31] Q. Zeng, Y. Zhang, Y. Peng, Q. Zeng, G. Sun, M. Guo, and T. Cai, *Interpretable machine learning for solvent prediction and mechanistic insights in multi-component crystal screening*, Chemical Engineering Journal **524** (2025) 169397.

[32] Y. Zhang, M. Bertani, A. Pedone, R. E. Youngman, G. Tricot, A. Kumar, and A. Goel, *Decoding crystallization behavior of aluminoborosilicate glasses: From structural descriptors to Quantitative Structure Property Relationship (QSPR) based predictive models*, Acta Materialia **268** (2024) 119784.