KFUPM
Journal of Undergraduate Research International

Diamond Open Access 41

Article Chemical Engineering

XGBoost-Based Prediction of pKa Values from Molecular Descriptors

Fahad Ali Atwi and Mohammed Al-Khater*

Cite https://doi.org/10.64589/juri/221642

Submitted: May 19, 2025 Revised: March 14, 2026 Accepted: April 04, 2026

ABSTRACT

The acid dissociation constant (pKa) is a fundamental physicochemical property that governs molecular ionization, solubility, and bioavailability, rendering its accurate prediction essential in drug design and environmental toxicology. This study entailed a systematic comparison of molecular descriptor representations for pKa prediction using eXtreme Gradient Boosting (XGBoost) regression. Four descriptor categories were evaluated: Morgan circular fingerprints, known as extended-connectivity fingerprints (ECFP), with varying radii (one to four) and bit lengths (12–16,384), Molecular ACCess System (MACCS) structural keys, physicochemical descriptors (molecular weight, logP, rotatable bond count, and fragment counts), and combined representations. A comprehensive grid search optimized the XGBoost hyperparameters across all configurations using five-fold cross-validation on a curated dataset of 7,912 compounds with experimental pKa values obtained from DataWarrior. Morgan fingerprints with radius two and 16,384 bits achieved the highest standalone performance (test $R^2 = 0.795$, root mean squared error, RMSE = 1.539), whereas combining Morgan fingerprints (radius one, 2,048 bits) with MACCS keys and physicochemical descriptors yielded the optimal overall result (test $R^2 = 0.817$; RMSE = 1.455). An extended analysis employing 1,135 Mordred and RDKit cheminformatic descriptors with XGBoost feature selection further improved the prediction accuracy (test $R_2 = 0.829$; RMSE = 1.404). Moreover, feature importance analysis identified the number of acidic groups, spectral descriptors, and aromatic amine fragments as the most influential predictors. These results demonstrate that multisource descriptor fusion consistently outperforms individual representations and that XGBoost provides a computationally efficient and interpretable framework for pKa modeling.

Keywords: pKa prediction, XGBoost, molecular fingerprints, Morgan fingerprints, MACCS keys, QSAR, molecular descriptors

1. INTRODUCTION

The acid dissociation constant (pKa) quantifies the tendency of a molecule to donate a proton in an aqueous solution and is one of the most critical physicochemical properties in medicinal chemistry, environmental science, and chemical engineering[1]. Because pKa governs the molecular charge state at physiological pH, it directly influences drug absorption, distribution, metabolism, excretion, and toxicity (ADMET) profiles. Therefore, accurate pKa values are indispensable for lead optimization in pharmaceutical development and for predicting the environmental fate of organic pollutants.

The experimental determination of pKa via potentiometric titration or spectrophotometric methods, although reliable, is time-consuming and resource-intensive, particularly for the screening of large compound libraries. Consequently, computational approaches to pKa prediction have attracted sustained research interest over the past few decades. Empirical methods based on linear free-energy relationships represent some of the earliest pKa-prediction strategies; however, their applicability is often restricted to structurally similar compound series[2]. Alternatively, quantum-mechanical approaches, including density functional theory (DFT) combined with continuum solvation models, offer higher accuracy; however, they remain computationally prohibitive for large-scale virtual screening[3,4].

More recently, machine learning (ML) methods have emerged as attractive compromises between accuracy and computational cost. For instance, quantitative structure–activity relationship (QSAR) models trained on molecular descriptors or fingerprints can rapidly predict pKa values of thousands of compounds following the establishment of a robust training set[1,5]. Among available ML algorithms, gradient boosting methods, particularly eXtreme Gradient Boosting (XGBoost), have demonstrated competitive performance across diverse molecular property prediction tasks[6,7]. XGBoost employs regularized gradient boosting of decision trees, incorporating L1 and L2 regularization terms, which mitigate overfitting while maintaining predictive power, thereby making the method particularly well-suited for high-dimensional descriptor spaces common in cheminformatics[6].

Previous studies have reported that the choice of molecular representation profoundly affects QSAR model performance. For instance, binary fingerprints, such as Morgan (extended-connectivity) fingerprints[8] and Molecular ACCess System (MACCS) structural keys[9], encode substructural information as fixed-length bit vectors and have been widely employed in

similarity searching and property prediction. Additionally, continuous physicochemical descriptors computed from molecular graphs (e.g., molecular weight, polar surface area, and partition coefficients) capture complementary global properties. More recently, comprehensive descriptor calculators such as Mordred[10] and PaDEL have enabled the generation of thousands of descriptors per molecule, raising questions regarding optimal feature-selection strategies. Nevertheless, despite the wide range of descriptor options, systematic comparisons of representation types for pKa prediction using modern gradient-boosting frameworks remain limited.

As one example, Mansouri et al.[1] established an important benchmark by developing open-source pKa prediction models using support vector machines, XGBoost, and deep neural networks on a curated dataset of 7,912 compounds from DataWarrior[11]. These models achieved root mean squared error (RMSE) values of approximately 1.5 pKa units on external test sets, comparable to those of commercial software packages. Additionally, Yang et al.[12] applied XGBoost with structural and physical-organic parameter-based (SPOC) descriptors for pKa prediction across 39 solvents, achieving a mean absolute error (MAE) of 0.87 pKa units. Furthermore, graph-based neural network approaches, such as MolGpKa[13] have also shown promise for pKa prediction. However, the relative contributions of different descriptor types (fingerprints versus continuous descriptors, as well as their combinations) to the XGBoost-based pKa models have not yet been comprehensively evaluated.

The present study addresses this gap by systematically comparing four molecular descriptor categories for pKa prediction using XGBoost regression: (i) Morgan circular fingerprints with varying radii and bit lengths, (ii) MACCS structural keys, (iii) a set of physicochemical descriptors and fragment counts, and (iv) combined multisource descriptor vectors. A grid search of the XGBoost hyperparameters with five-fold cross-validation was performed for each configuration. Furthermore, an extended analysis employing 1,135 Mordred and RDKit descriptors was conducted to assess the potential of a high-dimensional descriptor space. The key objectives of this work are to determine the descriptor representation that yields the most accurate pKa prediction, quantify the benefit of descriptor fusion, and identify the most informative molecular features using XGBoost feature-importance analysis.

2. METHODOLOGY

2.1. Dataset. The experimental pKa values for 7,912 organic compounds were obtained from the DataWarrior software package[11], which aggregates data from multiple public repositories. This dataset, originally curated by Mansouri et al.[1], includes the strongest acidic pKa for each compound measured in aqueous solution at 25 °C. The chemical structures were represented as Simplified Molecular Input Line Entry System (SMILES) strings. Prior to descriptor computation, salt forms were removed using a custom salt-removal protocol based on the RDKit SaltRemover module[14] with a curated salt definition file. Molecules that could not be parsed by RDKit or that yielded invalid molecular objects after salt removal were excluded, yielding the final working dataset. For the fingerprint comparison experiments (Sections 2.2–2.5), this dataset was split into 75% training and 25% test sets using a random partition (random_state = 11) to ensure reproducibility.

2.2. Morgan Circular Fingerprints. Morgan fingerprints, also known as ECFPs[8], were generated using the RDKit library[14]. Each fingerprint encodes a substructural environment around every atom up to a specified radius and is hashed into a fixed-length bit vector. To evaluate the effect of the fingerprint parameters on predictive performance, a comprehensive grid was constructed over four radii (one, two, three, and four) and ten bit lengths (12, 64, 128, 256, 512, 1,024, 2,048, 4,096, 8,192, and 16,384), yielding 40 distinct fingerprint configurations.

2.3. MACCS Structural Keys. MACCS keys[9] are a set of 166 predefined binary structural keys that indicate the presence or absence of specific substructural patterns (e.g., aromatic rings, functional groups, and ring sizes). Unlike Morgan fingerprints, MACCS keys do not require parameter tuning as the key definitions are fixed. MACCS fingerprints were generated for all valid molecules using the RDKit MACCSKey module[14].

2.4. Physicochemical Descriptors. A set of five continuous molecular descriptors and fragment counts was computed for each molecule using RDKit[14]: molecular weight (MolWt), calculated octanol–water partition coefficient (MolLogP), number of rotatable bonds (NumRotatableBonds), number of aliphatic hydroxyl groups (fr_Al_OH), and number of aromatic nitrogen atoms (fr_Ar_N). These descriptors were selected to capture fundamental size, lipophilicity, flexibility, and functional group properties relevant to acid–base equilibria.

2.5. Combined Descriptor Vectors. For each of the 40 Morgan fingerprint configurations, a combined feature vector was constructed by concatenating the Morgan fingerprint with a 166-bit MACCS key vector and five continuous descriptors, resulting in a total dimensionality of (Morgan size + 171). This experimental design allowed the assessment of whether multisource descriptor fusion improved pKa prediction relative to individual descriptor types.

2.6. Extended Descriptor Analysis. In a separate analysis, a comprehensive set of molecular descriptors was computed for an independent subset of the data (Opt2 acidic partition:

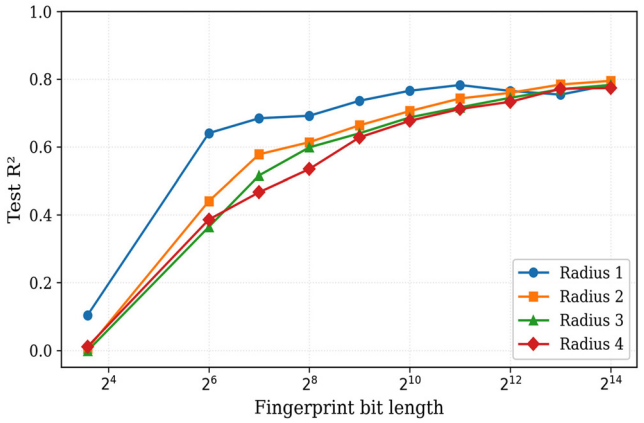

Figure 1. Effect of Morgan fingerprint bit length on the test R^2 *value for radii 1–4*

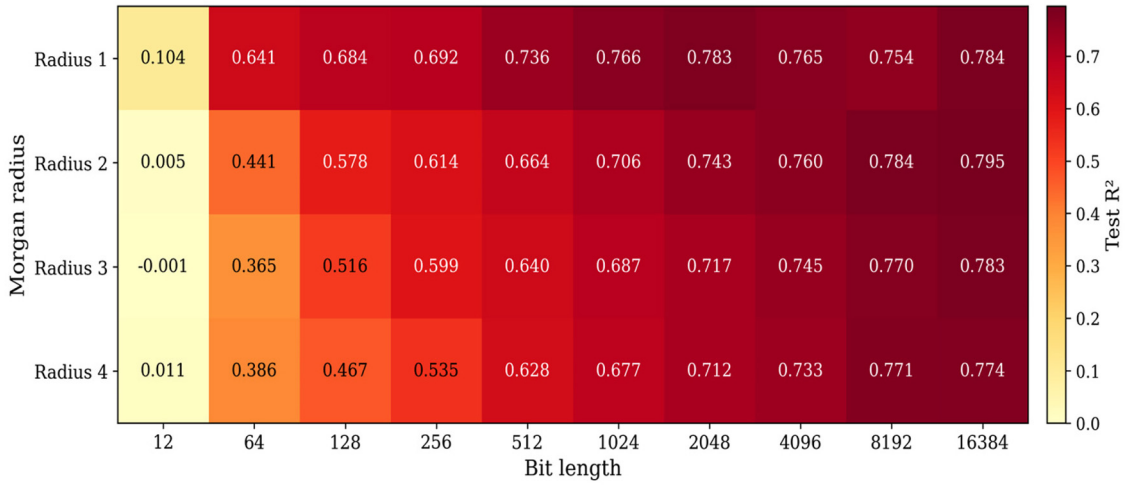

Figure 2. Heatmap of the Morgan fingerprint test R^2 values across all radius and bit-length combinations. Darker shading reflects superior performance

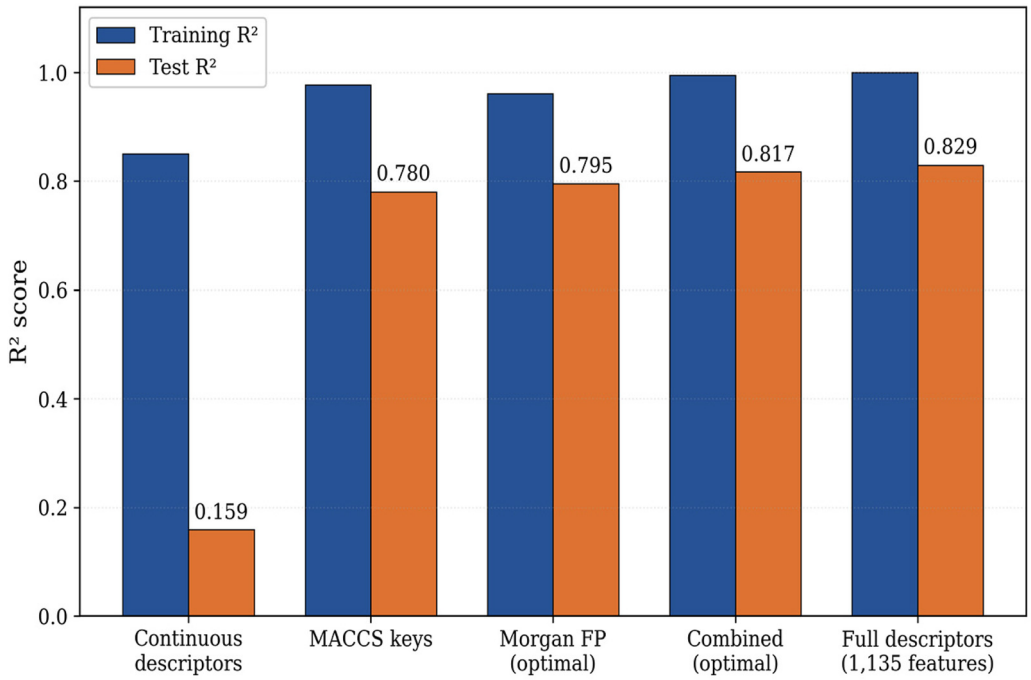

Figure 3. Comparison of the training and test R^2 values across descriptor categories

2,293 training compounds and 765 test compounds with predefined splits from DataWarrior). Three descriptor families were calculated: (i) all 210 RDKit two-dimensional (2D) descriptors[14], (ii) 85 RDKit fragment counts (fr_* functions), and (iii) approximately 1,800 Mordred of 2D descriptors[10]. After removal of zero-variance columns, conversion of all values to numeric format, and removal of rows with missing values, the final descriptor matrix contained 1,135 features. This high-dimensional representation was used to train an XGBoost model with the same hyperparameters described in Section 2.7, and to extract feature importance for chemical interpretation.

Table 1. *Performance comparison of XGBoost models across descriptor categories*

Descriptor type	Features	Training R^2	Test R^2	MAE	RMSE
Continuous descriptors	5	0.850	0.159	2.385	3.117
MACCS keys	166	0.977	0.780	1.056	1.596
Morgan FP (optimal)	16384	0.961	0.795	1.040	1.539
Combined (optimal)	2219	0.995	0.817	0.913	1.454
Full descriptors	1,135	0.9998	0.829	0.887	1.404

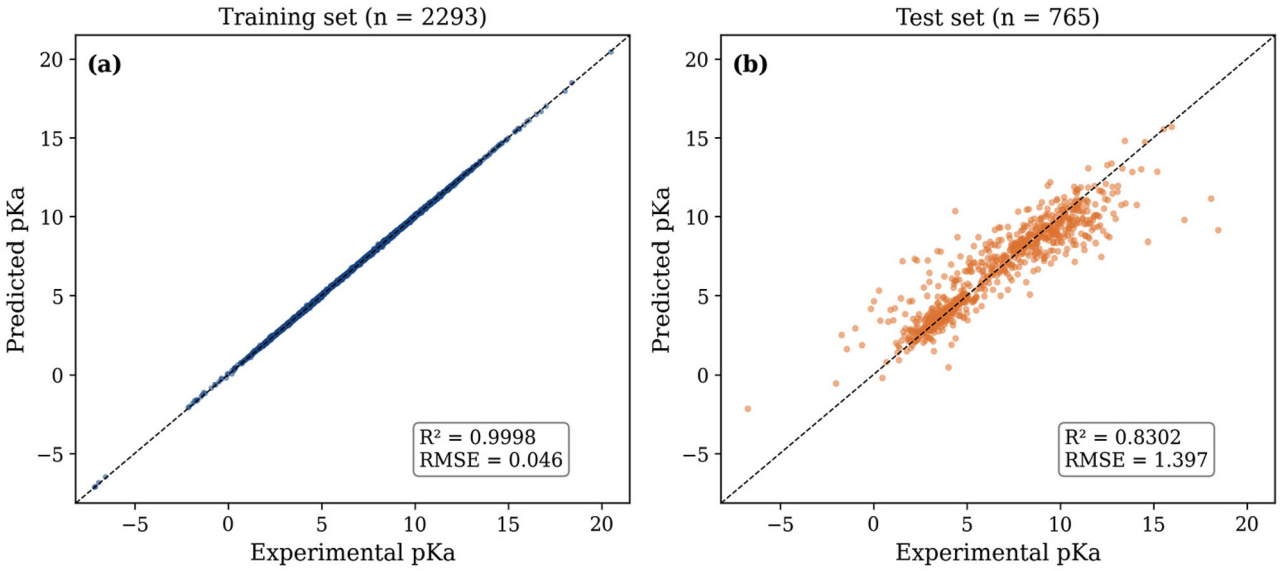

Figure 4. Predicted versus experimental pKa values for the full-descriptor XGBoost model. (a) Training set (n = 2293). (b) Test set (n = 765). The dashed line represents perfect prediction

2.7. XGBoost Model Training and Evaluation. All models were trained using XGBoost[6] with a squared-error regression objective (squarederror). For fingerprint comparison experiments, hyperparameter optimization was performed via an exhaustive grid search with five-fold cross-validation using GridSearchCV[15]. The search space included n_estimators = {300}, max_depth = {3, 6}, learning_rate = {0.1, 0.2}, subsample = {0.8}, colsample_bytree = {0.8, 1.0}, and reg_lambda = {1, 5}. The optimal model determined by cross-validation was evaluated using the held-out test set. For the extended descriptor analysis, fixed hyperparameters were employed (n_estimators = 300, max_depth = 6, learning_rate = 0.1, subsample = 0.8, colsample_bytree = 0.8, and reg_lambda = 1), following published guidelines for gradient boosting in QSAR[7]. Model performance was assessed using the coefficient of determination (R^2), RMSE, and MAE for both the training and test sets. Feature importance was extracted from the trained XGBoost model using a gain-based importance metric.

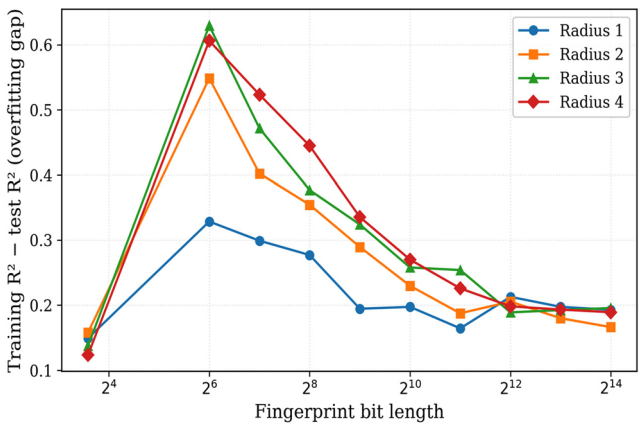

Figure 5. Overfitting gap (training R^2 − test R^2) for Morgan fingerprint models as a function of bit length and radius

3. RESULTS AND DISCUSSION

3.1. Effect of Morgan Fingerprint Parameters. The predictive performance of the Morgan fingerprints varied substantially with both radius and bit length. Figure 1 illustrates the relationship between the fingerprint bit length and R^2 for each radius. At the smallest bit length tested (12 bits), all radii yielded poor model performance (test R^2 <0.20), reflecting severe information loss due to hash collisions. Performance increased monotonically with bit length up to approximately 2,048–4,096 bits, beyond which the gain plateaued. Radius 2 consistently outperformed the other radii at higher bit lengths, achieving superior individual Morgan fingerprint results: test R^2 = 0.795 and RMSE = 1.539 at 16,384 bits. This finding is consistent with the observation that ECFP4 (radius 2) captures a balance between local atomic environments and broader substructural context[8].

Figure 2 presents a heatmap of the test R^2 values across the full-radius-bit-length grid. A clear pattern emerged: radii 1 and 2 generally outperformed radii 3 and 4 at equivalent bit lengths. This can be attributed to the fact that larger radii encode increasingly specific substructural environments, which may introduce noise and reduce generalizability when the training set does not contain sufficient structural diversity. Additionally, extremely small bit lengths (12–64 bits) were found to be consistently inadequate, regardless of radius, confirming that fingerprint dimensionality represents a critical design parameter.

3.2. Comparison Across Descriptor Types. Table 1 summarizes the best-performing models for each descriptor category. Use of the five continuous physicochemical descriptors alone yielded a low R^2 value of 0.159 (RMSE = 3.117), indicating that this minimal feature set was insufficient for pKa prediction across a structurally diverse compound library. This result underscores the importance of substructural information for capturing the electronic and steric factors that govern the acid–base equilibria. MACCS keys (166 bits) produced a significantly superior R^2 of 0.780 (RMSE = 1.596), demonstrating the utility of predefined structural keys despite their modest dimensionality. The

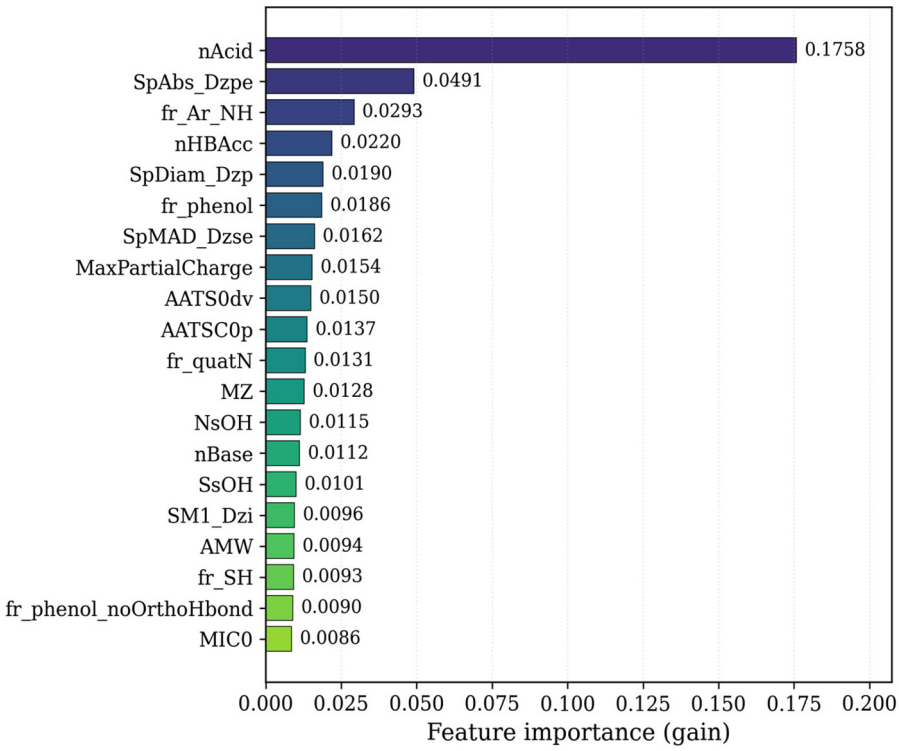

Figure 6. Top 20 features ranked by importance (gain metric) identified by the full-descriptor XGBoost model

optimal Morgan fingerprint configuration (radius two, 16,384 bits) slightly outperformed MACCS, with a test R^2 of 0.795, which can be attributed to its data-driven encoding of circular substructures rather than the fixed key set[8,9].

As detailed in the table, the combined descriptor vectors consistently outperformed the individual components. The optimal combined configuration (Morgan radius 1, 2,048 bits + MACCS + continuous descriptors) achieved a test R^2 of 0.817 (RMSE = 1.454), representing a 2.8% improvement over the optimal individual Morgan fingerprints. This result corroborates previous findings that heterogeneous descriptor fusion can capture complementary structural information[5,7]. Figure 3 presents a visual comparison of the training and test R^2 values across all five model categories.

3.3. Predicted Versus Experimental pKa. Figure 4 presents scatter plots of the predicted versus experimental pKa values obtained using the full-descriptor XGBoost model for both the training and test sets. The training set (n = 2293) exhibited a near-perfect fit (R^2 = 0.9998, RMSE = 0.047), whereas the test set (n = 765) showed good generalization (R^2 = 0.829, RMSE = 1.404). The scatter around the diagonal identity line remained relatively uniform across the pKa range of −3 to 14, suggesting that the model does not exhibit systematic bias toward any particular pKa region.

3.4. Overfitting Analysis. The gap between the training and test R^2 values was used as an indicator of overfitting. Figure 5 plots this gap for the Morgan fingerprint models across all radii and bit lengths. Models with max_depth = 6 exhibited larger train–test gaps (up to 0.20 R^2units) than those with max_depth = 3, consistent with the known tendency of deeper trees to memorize training data[6]. Nevertheless, the regularization provided by XGBoost (L2 penalty via reg_lambda) prevented catastrophic overfitting, even at high dimensionalities. The full-descriptor model showed the largest absolute training–test gap (training R^2 = 0.9998 vs. test R^2 = 0.829), suggesting that some degree of overfitting was inevitable with 1,135 features and 2,293 training samples. Future studies should therefore explore additional regularization strategies (e.g., feature selection or ensemble averaging) to narrow this gap.

3.5. Feature Importance Analysis. The feature importance extracted from the full-descriptor XGBoost model provided chemical insights into the structural drivers of pKa variation (Figure 6). The most important feature was nAcid (importance = 0.176), which is a Mordred descriptor that counts the number of acidic functional groups in the molecule. This is chemically intuitive because the presence and number of acidic groups directly determine the pKa landscape. The second-ranked feature, SpAbs_Dzpe (importance = 0.049), is a spectral descriptor related to Barysz-weighted electronegativity that captures electronic distribution patterns. The third most important feature, fr_Ar_NH (importance = 0.029), reflects the influence of aromatic amine groups, which are well-known weak acids.

Other highly ranked features included nHBAcc (hydrogen-bond acceptor count, importance = 0.022), fr_phenol (phenolic hydroxyl groups, importance = 0.019), and MaxPartialCharge (importance = 0.015). This feature-importance profile is consistent with chemical theory: molecular acidity is primarily governed by the presence of ionizable functional groups, the electronic environment around those groups (i.e., partial charges and electronegativity), and intramolecular hydrogen bonding[2]. The top ten features identified in this work are listed in Table 2.

Table 2. *Top ten molecular descriptors ranked by XGBoost feature importance (gain metric)*

Rank	Descriptor	Importance	Description
1	nAcid	0.176	Number of acidic groups
2	SpAbs_Dzpe	0.049	Spectral Barysz electronegativity
3	fr_Ar_NH	0.029	Aromatic amine count
4	nHBAcc	0.022	H-bond acceptor count
5	SpDiam_Dzp	0.019	Spectral Barysz polarizability
6	fr_phenol	0.019	Phenol group count
7	SpMAD_Dzse	0.016	Spectral distance descriptor
8	MaxPartialCharge	0.015	Maximum partial charge
9	AATS0dv	0.015	Moreau–Broto autocorrelation
10	AATSC0p	0.014	Centered Moreau–Broto polarizability

3.6. Residual Analysis. Figure 7 shows the residual distribution and the residual-versus-predicted plot for the full-descriptor model on the test set. The residual histogram is approximately symmetric about zero, with a slight positive skew. The majority of predictions (>80%) fall within ±2 pKa units of the experimental value. Additionally, the residual-versus-predicted plot shows no pronounced systematic pattern, although larger absolute residuals were observed at extreme pKa values (i.e., below −1 and above 12), where the training-data density is lower. These observations indicate that the prediction errors of the model are predominantly random rather than systematic and suggest that the performance could be improved by augmenting the training set in these extreme pKa ranges.

3.7. Comparison with the Literature. The results of this study are consistent with the findings of Mansouri et al.[1], who reported RMSE values of approximately 1.5 pKa units for XGBoost models trained on the same DataWarrior dataset. The present combined fingerprint model (RMSE = 1.454) and full-descriptor model (RMSE = 1.404) both fall within this range. Although Yang et al.[12] achieved a lower MAE of 0.87 pKa units using solvent-specific SPOC descriptors, their study was restricted to a smaller, more homogeneous dataset. Additionally, Sanchez et al.[4] achieved high prediction accuracy by combining DFT-computed features with molecular fragmentation descriptors, albeit at a substantially higher computational cost than the purely descriptor-based approach presented herein. These results demonstrate that purely 2D descriptor-based XGBoost models provide a favorable balance between accuracy, interpretability, and computational efficiency for large-scale pKa predictions.

4. CONCLUSIONS

This study systematically evaluated the performance of four molecular descriptor categories for pKa prediction using XGBoost regression on a curated dataset of 7,912 compounds.

It was revealed that Morgan circular fingerprints at radius 2 with 16,384 bits achieved the optimal standalone fingerprint performance (test R^2 = 0.795, RMSE = 1.539), whereas bit lengths below 128 were insufficient for accurate prediction, regardless of radius. The MACCS structural keys (166 bits) provided competitive performance (test R^2 = 0.780) relative to their low dimensionality. However, the use of continuous physicochemical descriptors alone resulted in poor predictability (R^2 = 0.159), confirming the requirement for substructural

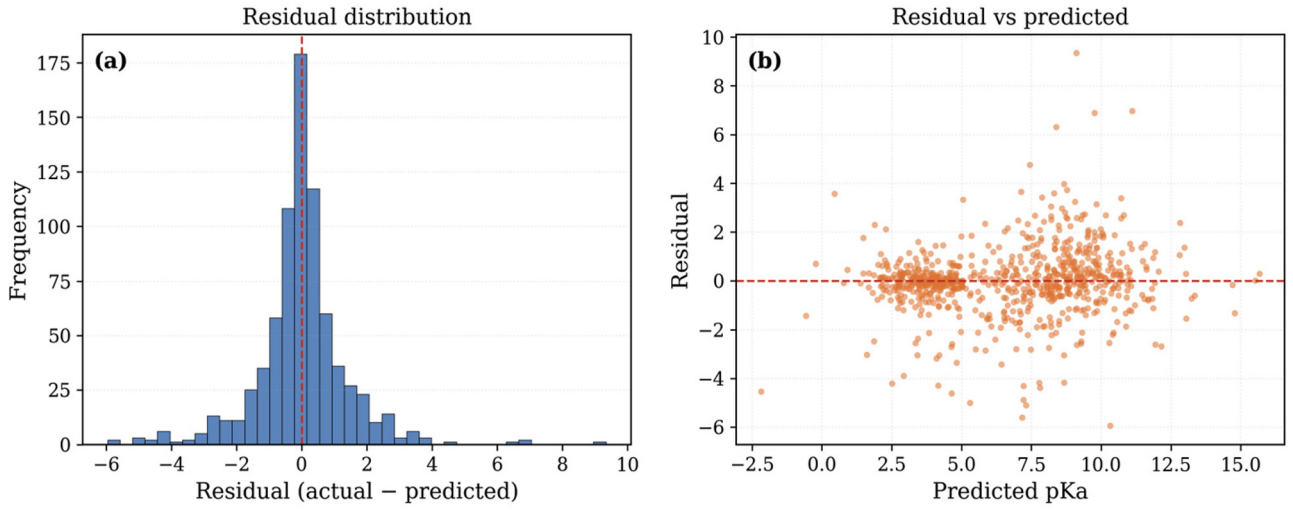

Figure 7. Residual analysis of the full-descriptor XGBoost model on the test set. (a) Residual histogram. (b) Residuals versus predicted pKa values

information. Notably, a combination of Morgan fingerprints, MACCS keys, and physicochemical descriptors consistently outperformed the individual representations, with the optimal combined model achieving $R^2 = 0.817$ (RMSE = 1.454). An extended analysis with 1,135 Mordred and RDKit descriptors further improved the prediction accuracy to $R^2 = 0.829$ (RMSE = 1.404). Furthermore, feature importance analysis revealed that the acidic group count, spectral electronegativity descriptors, and aromatic amine fragments were the most influential predictors, consistent with established chemical theory of acid–base equilibria.

These findings demonstrate that descriptor fusion is a simple yet effective strategy for improving pKa prediction accuracy and that XGBoost offers a computationally efficient and interpretable framework suitable for large-scale virtual screening. Future studies should explore the application of these models to external validation sets, the integration of three-dimensional molecular descriptors, and comparisons with graph neural network approaches to further advance the field.

AFFILIATIONS AND AUTHOR DETAILS

Undergraduate Author

Fahad Ali Atwi – *Department of Chemical Engineering, King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia*; 0009-0003-7015-9092
Email: s202283300@kfupm.edu.sa

Corresponding Author

Mohammed Al-Khater – *Research Mentor, Professor, Department of Chemical Engineering, King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia;* 0000-0003-0926-4056
Email: mfkhater@kfupm.edu.sa;
– *Researcher, Interdisciplinary Research Center for Refining & Advanced Chemicals, King Fahd University of Petroleum & Minerals, 31261 Dhahran, Saudi Arabia.*

ACKNOWLEDGEMENTS

The authors acknowledge the support provided by the King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia. The authors also acknowledge the developers of the open-source software packages used in this study, including RDKit, Mordred, XGBoost, Scikit-learn, and DataWarrior.

REFERENCES

(1) Mansouri, K. et al. Open-source QSAR models for pKa prediction using multiple machine learning approaches. J. Cheminform. 11, 60 (2019).
(2) Göller, A. H. The art of atom descriptor design. Drug Discov. Today Technol. 32–33, 37–43 (2020).
(3) Lawler, R. et al. DFT–Machine learning approach for accurate prediction of pKa. J. Phys. Chem. A 125, 8712–8722 (2021).
(4) Sanchez, A. J., Maier, S. & Raghavachari, K. Leveraging DFT and molecular fragmentation for chemically accurate pKa prediction using machine learning. J. Chem. Inf. Model. 64, 712–723 (2024).
(5) Tropsha, A. Best practices for QSAR model development, validation, and exploitation. Mol. Inform. 29, 476–488 (2010).
(6) Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. 785–794 (2016).
(7) Boldini, D. et al. Practical guidelines for the use of gradient boosting for molecular property prediction. J. Cheminform. 15, 73 (2023).
(8) Rogers, D. & Hahn, M. Extended-connectivity fingerprints. J. Chem. Inf. Model. 50, 742–754 (2010).
(9) Durant, J. L., Leland, B. A., Henry, D. R. & Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. J. Chem. Inf. Comput. Sci. 42, 1273–1280 (2002).
(10) Moriwaki, H., Tian, Y.-S., Kawashita, N. & Takagi, T. Mordred: a molecular descriptor calculator. J. Cheminform. 10, 4 (2018).
(11) Sander, T., Freyss, J., von Korff, M. & Rufener, C. DataWarrior: an open-source program for chemistry aware data visualization and analysis. J. Chem. Inf. Model. 55, 460–473 (2015).
(12) Yang, Q. et al. Holistic prediction of the pKa in diverse solvents based on a machine-learning approach. Angew. Chem. Int. Ed. 59, 19282–19291 (2020).
(13) Pan, S. et al. MolGpKa: a web server for molecular pKa prediction using a graph-convolutional neural network. J. Chem. Inf. Model. 61, 3159–3165 (2021).
(14) Landrum, G. RDKit: Open-source cheminformatics. https://www.rdkit.org (2024).
(15) Pedregosa, F. et al. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830 (2011).