Original Paper

Accelerating Discovery of Leukemia Inhibitors Using AI-Driven Quantitative Structure-Activity Relationship: Algorithm Development and Validation

Samuel Kakraba[1,2], PhD; Edmund Fosu Agyemang[1], MS; Robert J Shmookler Reis[3], PhD

[1]Biostatistics and Data Science, Celia Scott Weatherhead School of Public Health and Tropical Medicince, Tulane University, New Orleans, LA, United States

[2]Tulane School of Medicine, Tulane Center for Aging, Tulane University, New Orleans, LA, United States

[3]Department of Geriatrics, School of Medicine, University of Arkansas for Medical Sciences, Little Rock, AR, United States

Corresponding Author:
Samuel Kakraba, PhD
Biostatistics and Data Science
Celia Scott Weatherhead School of Public Health and Tropical Medicince
Tulane University
1440 Canal Street
New Orleans, LA, 70112
United States
Phone: 1 5049882475
Email: skakraba@tulane.edu

Abstract

Background: Leukemia treatment remains a major challenge in oncology. While thiadiazolidinone analogs show potential to inhibit leukemia cell proliferation, they often lack sufficient potency and selectivity. Traditional drug discovery struggles to efficiently explore the vast chemical landscape, highlighting the need for innovative computational strategies. Machine learning (ML)–enhanced quantitative structure-activity relationship (QSAR) modeling offers a promising route to identify and optimize inhibitors with improved activity and specificity.

Objective: We aimed to develop and validate an integrated ML-enhanced QSAR modeling workflow for the rational design and prediction of thiadiazolidinone analogs with improved antileukemia activity by systematically evaluating molecular descriptors and algorithmic approaches to identify key determinants of potency and guide future inhibitor optimization.

Methods: We analyzed 35 thiadiazolidinone derivatives with confirmed antileukemia activity, removing outliers for data quality. Using Schrödinger MAESTRO, we calculated 220 molecular descriptors (1D-4D). Seventeen ML models, including random forests, XGBoost, and neural networks, were trained on 70% of the data and tested on 30%, using stratified random sampling. Model performance was assessed with 12 metrics, including mean squared error (MSE), coefficient of determination (explained variance; R^2), and Shapley additive explanations (SHAP) values, and optimized via hyperparameter tuning and 5-fold cross-validation. Additional analyses, including train-test gap assessment, comparison to baseline linear models, and cross-validation stability analysis, were performed to assess genuine learning rather than overfitting.

Results: Isotonic regression ranked first with the lowest test MSE (0.00031 ± 0.00009), outperforming baseline models by over 15% in explained variance. Ensemble methods, especially LightGBM and random forest, also showed superior predictive performance (LightGBM: MSE=0.00063 ± 0.00012; R^2=0.9709 ± 0.0084). Training-to-test performance degradation of LightGBM was modest (ΔR^2=–0.01, ΔMSE=+0.000126), suggesting genuine pattern learning rather than memorization. SHAP analysis revealed that the most influential features contributing to antileukemia activity were global molecular shape (*r_qp_glob*; mean SHAP value=0.52), weighted polar surface area (*r_qp_WPSA*; ≈0.50), polarizability (*r_qp_QPpolrz*; ≈0.49), partition coefficient (*r_qp_QPlogPC16*; ≈0.48), solvent-accessible surface area (*r_qp_SASA*; ≈0.48), hydrogen bond donor count (*r_qp_donorHB*; ≈0.48), and the sum of topological distances between oxygen and chlorine atoms (*i_desc_Sum_of_topological_distances_between_O.Cl*; ≈0.47). These features highlight the importance of steric complementarity and the 3D arrangement of functional groups. Aqueous solubility (*r_qp_QPlogS*; ≈0.47) and hydrogen bond acceptor count

(*r_qp_accptHB*; ≈0.44) were also among the top 10 features. The significance of these descriptors was consistent across multiple algorithmic models, including random forest, XGBoost, and partial least squares approaches.

Conclusions: Integrating advanced ML with QSAR modeling enables systematic analysis of structure-activity relationships in thiadiazolidinone analogs on this dataset. While ensemble methods capture complex patterns with high internal validation metrics, external validation on independent compounds and prospective experimental testing are essential before broad therapeutic claims can be made. This work provides a methodological foundation and identifies molecular features for future validation efforts.

(JMIR AI 2026;5:e81552) doi: 10.2196/81552

KEYWORDS

anti-leukemia; thiadiazolidinones; TDZD analogs; artificial intelligence; machine learning; quantitative structure-activity relationship; QSAR; small-molecule inhibitors; drug discovery; precision oncology; Shapley additive explanations analysis

Introduction

Leukemia remains a formidable challenge in oncology, largely due to the persistence of leukemia stem cells (LSCs), which drive disease relapse through intrinsic resistance to conventional chemotherapy [1]. While standard treatments effectively target proliferating leukemic blast cells, LSCs evade destruction by leveraging quiescence and enhanced survival mechanisms, such as dysregulated kinase signaling and adaptation to oxidative stress [1]. Thiadiazolidinone analogs, notably thiadiazolidinone-8, comprise a promising family of molecules that selectively induce rapid cell death in LSCs via a dual mechanism: (1) inhibition of glycogen synthase kinase 3β (GSK3β), and (2) triggering oxidative collapse [1]. Molecular docking and simulation studies suggest that thiadiazolidinone-8 might bind to an allosteric hydrophobic pocket in GSK3β's inactive "DFG-out" conformation, preventing reactivation and disrupting prosurvival pathways, while simultaneously depleting intracellular thiols to disrupt membrane integrity within 2 hours, achieving 85% to 93% lethality in primary acute myeloid leukemia, acute lymphoblastic leukemia, and chronic lymphoblastic leukemia specimens at 20 µM. Critically, thiadiazolidinone-8 spares normal hematopoietic stem cells (79.5% viability) and significantly reduces engraftment of leukemic cells in nonobese diabetic/severe combined immunodeficient xenotransplantation models, with mean engraftment dropping from 76% to as low as 0.7% (P<.001), while having minimal toxicity for normal cells [1]. Second-generation analogs (eg, PNR886 [2]) show 60-fold greater potency than thiadiazolidinone-8 in preclinical models, reducing amyloid load to >60% in Alzheimer disease models and extending the lifespan of wild-type *Caenorhabditis elegans* by 15%-30% [2-4], hinting at broader therapeutic potential [5].

Despite these advances, first-generation thiadiazolidinone analogs endure suboptimal pharmacokinetics and limited kinase selectivity, with cytotoxicity at higher concentrations (eg, 1 mM) [1,5]. Recent computational modeling of GSK3β's inactive state offers opportunities for the rational design of next-generation inhibitors targeting key residues (Lys205, Asp200, and Ala204) to enhance specificity and reduce off-target effects on normal tissues [5]. Structural optimization is essential to balance potent LSC eradication with minimal toxicity, unlocking the potential of thiadiazolidinone-based therapies to target the LSC reservoir in refractory leukemias specifically.

The quest for effective leukemia inhibitors is hindered by challenges such as enzyme specificity, cell selection for resistance, and off-target effects. Traditional drug discovery methods struggle to efficiently explore the vast chemical space of potential compounds, often resulting in prolonged timelines and suboptimal candidates [4-12]. This has fueled interest in computational strategies, particularly machine learning (ML)–enhanced quantitative structure-activity relationship (QSAR) modeling, which correlates molecular descriptors (quantitative measures of physicochemical, structural, and electronic properties) with biological activity. ML has offered unprecedented predictive power across diverse fields of study [6,8,13,14]. Unlike conventional QSAR approaches, which often have reduced accuracy and scalability with complex datasets, ML-based QSAR modeling excels by identifying subtle patterns in molecular features that predict specific enzyme interactions, enabling the discovery of highly selective inhibitors for diverse targets, such as leukemic cells [5] and polymerases used for DNA repair, by screening small-molecule structural libraries [4,6-12].

ML algorithms have shown promise in enhancing drug discovery [4,9,13-15] by enabling prediction of resistance mechanisms, guiding the design of inhibitors to delay or overcome resistance, and prioritizing molecular features linked to selectivity or minimal toxicity [5]. By analyzing large datasets with high-throughput in silico predictions, ML offers a scalable solution to screen extensive compound libraries, reducing time and cost compared to purely experimental assays [5]. Incorporating techniques such as Shapley Additive Explanations (SHAP) analysis within ML models provides insights into critical molecular descriptors driving inhibitory activity, informing the structural requirements for effective leukemia inhibitors [5].

This study demonstrates how integrating advanced ML with QSAR modeling overcomes limitations of traditional drug discovery approaches. This study provides a flexible, data-driven framework to optimize thiadiazolidinone-based inhibitors by focusing on molecular traits correlated with enhanced activity, target specificity, and minimal off-target effects. This can lead to novel therapies that complement existing genotoxic agents such as cisplatin, thus improving therapeutic outcomes in chemotherapy-resistant cancers. However, we acknowledge that such potential can only be realized through rigorous external validation and experimental verification of computational predictions.

Methods

Methodology for Enhanced Inhibitor Identification

We introduce a structured methodology to enhance the identification of thiadiazolidinone analogs with antileukemic properties using artificial intelligence (AI)–powered QSAR modeling. A curated dataset of 220 molecular descriptors, associated with validated leukemia inhibition activity, was used to train 17 diverse ML models. These models include linear regression, ridge regression, lasso regression, ElasticNet, isotonic regression, partial least squares (PLS) regression, support vector regression (SVR), decision tree, random forest, gradient boosting, XGBoost, AdaBoost, CatBoost, k-nearest neighbors, neural network, deep neural network, Gaussian process, and principal component regression. Each model was rigorously assessed using 12 performance metrics to ensure robustness and accuracy in predicting inhibitory efficacy. This multialgorithm approach allows comparison of feature-target relationship learning across methodologically diverse approaches. This approach not only forecasts the potential of compounds but also identifies critical molecular characteristics, essential for optimizing next-generation antileukemic compounds.

Dataset and Preprocessing

Overview

Multistep Protocol

This study used an in-house selected library of 35 thiadiazolidinone analogs, each with experimentally validated leukemia inhibition activity expressed as $logIC_{50}$ values [1].

Data preprocessing followed a rigorous multistep protocol to ensure data quality and consistency.

Outlier Detection and Removal

Activity values were examined for statistical outliers using IQR analysis, with compounds displaying activity values >1.5×IQR from the quartile boundaries flagged for review and removed if deemed measurement anomalies.

Chemical Structure Standardization

Chemical structures were initially sketched in ChemDraw [16], converted to Simplified Molecular Input Line Entry System format, and subsequently transformed into SYBYL Mol2 files using Schrödinger MAESTRO (Schrödinger Release 2025-2: Canvas, Schrödinger, LLC, 2025) for 3D visualization, ensuring standardized chemical representation across all compounds.

Ligand Geometric Optimization

Ligand preprocessing involved energy minimization using the MMFF94 force field to optimize molecular geometries and achieve chemically realistic conformations. Structural alignment of conserved thiadiazolidinone cores was performed to standardize side-chain modifications across the dataset, ensuring consistent and comparable descriptor computation [17].

Descriptor Calculation

Molecular descriptors were calculated using Schrödinger MAESTRO 12.5 software, encompassing a broad spectrum of physicochemical properties (1D-4D descriptors). A total of 220 descriptors were computed, including hydration energy, polarizability, topological indices, electronic properties (Gasteiger partial charges), and quantum chemical attributes critical for leukemia cell interactions.

Feature Scaling and Normalization

Before model training, all molecular descriptor features were normalized using StandardScaler (z score normalization: (x – mean)/SD) to ensure equal weighting across features with different scales and units, preventing high-magnitude descriptors from dominating the learning process.

Missing Value Handling

Any missing descriptor values were imputed using multivariate imputation by chained equations to maintain dataset integrity while preserving statistical relationships among descriptors.

The resulting preprocessed dataset contained 35 compounds with 220 standardized molecular descriptors and corresponding experimental $logIC_{50}$ values, forming a robust foundation for QSAR modeling (see Multimedia Appendix 1 for the complete molecular database of molecular descriptors with corresponding $logIC_{50}$).

Model Training and Evaluation

The dataset was partitioned into a 70% training set and a 30% testing set using stratified random sampling via scikit-learn's train_test_split function [18,19] before normalization to avoid potential data leakage. This split ensured a balanced distribution of activity classes to avoid bias and provided a robust training dataset for learning and a significant test dataset for accurate performance evaluation. Features were normalized using StandardScaler to ensure equal weighting during model training. The 17 ML algorithms evaluated spanned a wide range of approaches, including linear models, tree-based ensembles, kernel methods, instance-based learners, neural networks, probabilistic approaches, dimensionality reduction techniques, nonparametric models, and advanced gradient boosting frameworks. Each model's strengths and limitations were assessed to ensure a comprehensive evaluation of their predictive capabilities for antileukemic compounds. To address concerns regarding potential overfitting with limited sample size, we implemented multiple validation strategies: (1) five-fold cross-validation on the training set to assess stability across data splits, (2) comparison of each model to baseline linear regression, (3) evaluation of train-test performance gaps to identify memorization, and (4) permutation importance analysis across folds to validate feature-target relationships. Performance metrics such as coefficient of determination (explained variance; R^2), root-mean-square error in prediction, and others were used to quantify predictive accuracy and model robustness.

Overview of ML Algorithms

The 17 ML algorithms compared for QSAR modeling are summarized in Table 1, detailing their descriptions, strengths, and limitations. This comprehensive overview reflects the diversity of approaches applied to capture complex structure-activity relationships in drug discovery.

Table 1. Overview of machine learning algorithms compared for QSAR[a] modeling [20].

Algorithm	Description	Strengths	Limitations	References
Linear regression	Models relationships with a linear equation	Simple, efficient, highly interpretable	Assumes linearity, sensitive to outliers	[21]
Ridge regression	Uses L2[b] regularization to prevent overfitting of data	Improves stability and handles multicollinearity	Does not perform feature selection	[22,23]
Lasso regression	Applies L1[c] regularization for feature selection	Reduces model complexity through feature selection	May arbitrarily select among correlated variables	[24,25]
ElasticNet	Combines L1 and L2 regularization	Balances the benefits of lasso and ridge	Requires tuning 2 hyperparameters	[22,23]
Isotonic regression	Fits a monotonic free-form line to the data	Robust to outliers, ensures monotonic relationships	Computationally intensive, limited generalization	[26,27]
PLS[d]	Identifies relationships between matrices, reducing dimensionality	Manages multicollinearity, effective for high-dimensional data	Less interpretable than other methods	[28-30]
SVR[e]	Approximates input-output in high-dimensional space	Robust against data overfitting, excels in complex datasets	Sensitive to kernel choice, computationally intensive	[31-33]
Decision tree	Nonparametric tree structure for regression or classification	Interpretable, handles diverse data, and captures nonlinearity	Prone to overfitting, may not generalize well	[13,14,34,35]
Random forest	Ensemble of trees to minimize overfitting	Reduces overfitting, assesses feature importance	Computationally expensive, less interpretable	[13,14,34,36,37]
Gradient boosting	Builds weak learners sequentially for improved predictions	High predictive power, excels in complex modeling	Risk of overfitting if not tuned properly	[38,39]
XGBoost	Optimized gradient boosting library for enhanced performance	High accuracy, efficient, and handles missing data	Complex to tune, less interpretable	[40]
AdaBoost	Combines weak classifiers, focusing on misclassified instances	Improves accuracy by emphasizing difficult cases	Sensitive to noisy data and outliers	[41,42]
CatBoost	Uses ordered boosting for categorical features	Reduces overfitting, high accuracy with categorical data	Slower training speed, less interpretable	[43,44]
KNN[f]	Nonparametric method based on proximity to nearest points	Captures complex relationships without assumptions	Computationally intensive, sensitive to scaling	[45,46]
Neural network	Mimics brain processes to model nonlinear relationships	Adaptable, excels with large datasets	Requires significant data, prone to overfitting	[13,14,34,47,48]
DNN[g]	Advanced neural network with multiple layers for complex patterns	High performance in capturing intricate patterns	Requires large datasets, computationally intensive	[49,50]
Gaussian process	Probabilistic approach with uncertainty estimates	Offers uncertainty quantification, models complex functions	Computationally expensive for large datasets	[51]
PCR[h]	Combines PCA[i] with regression for dimensionality reduction	Handles multicollinearity, reduces dimensionality	May lose interpretability, less predictive power	[52-54]

[a]QSAR: quantitative structure-activity relationship.

[b]L2: ridge penalty

[c]L1: lasso penalty

[d]PLS: partial least squares.

[e]SVR: support vector regression.

[f]KNN: k-nearest neighbors.

[g]DNN: deep neural network.

[h]PCR: principal component regression.

[i]PCA: principal component analysis.

Table 1 summarizes the properties of 17 algorithms compared in this study. The results were consistent with recent advances in QSAR modeling in which ML techniques such as random forest, XGBoost, and deep neural network empirically displayed superior predictive performance, especially for complex and diverse datasets [34]. The selection of these algorithms was guided by their established effectiveness in small-sample, high-dimensional biological datasets, their ability to handle multicollinearity, capture nonlinear relationships, and to provide insights into feature importance [34], all of which are critical

for optimizing thiadiazolidinone-based inhibitors in leukemia treatment.

Hyperparameters were optimized via grid or random search with 5-fold cross-validation, prioritizing the minimization of mean squared error (MSE) and maximization of R^2 and adjusted coefficient of determination (adjusted R^2) metrics.

Model performance was evaluated using 12 metrics, including MSE, root-mean-squared error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), symmetric mean absolute percentage error (SMAPE), median absolute error (MedAE), R^2, adjusted R^2, concordance correlation coefficient (CCC), normalized mean squared error (NMSE), normalized root-mean-squared error (NRMSE), and Pearson correlation to ensure a comprehensive assessment of predictive accuracy and robustness. Detailed descriptions of these metrics are in the following sections.

About MSE

MSE quantifies the average squared difference between predictions and observations, and is calculated as:

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

where y_i is the observed value and $\hat{y}_i$ is the predicted value. MSE is critical for identifying models prone to severe inaccuracies.

About RMSE

RMSE provides error magnitude in the same units as the response variable, enhancing interpretability and sensitivity to outliers. It is calculated as:

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}}$$

About MAE

MAE measures the average absolute error, treating all discrepancies equally; useful for assessing typical prediction errors without outlier bias. It is calculated as:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

About MAPE

MAPE expresses errors as percentages, facilitating relative performance comparison across datasets, though it is undefined for 0 observed values. It is calculated as:

$$\mathrm{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{|y_i|}$$

About SMAPE

SMAPE addresses MAPE's asymmetry by normalizing errors against the average of observed and predicted values, improving robustness for near-zero values. It is calculated as:

$$\mathrm{SMAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\frac{2|y_i - \hat{y}_i|}{|y_i| + |\hat{y}_i|}$$

About MedAE

MedAE is resistant to outliers and is calculated as:

$$\mathrm{MedAE} = \mathrm{median}(|y_1 - \hat{y}_1|, \ldots, |y_n - \hat{y}_n|)$$

About R^2

R^2 represents the proportion of variance explained by the model, with values closer to 1 indicating a better fit. It is calculated as:

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

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

About Adjusted R^2

R^2 adjusts for model complexity, preventing overfitting by penalizing unnecessary predictors. It is calculated as:

$$\text{adjusted } \mathcal{R}^2 = 1 - (1 - R^2) \times \frac{n-1}{n-k-1}$$

where:

- R^2=R^2 of the model, also known as the fraction of variance explained.
- n=number of observations (data points).
- k=number of predictors (independent variables) in the model.

About CCC

CCC evaluates agreement between predictions and observations, combining precision (correlation) and accuracy (mean shift). It is calculated as:

$$\mathrm{CCC} = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}$$

where ρ is Pearson correlation, and μ and σ are means and SDs of the observed and predicted values, respectively.

About NMSE

NMSE scales MSE by dataset variance, enabling cross-study comparisons. It is calculated as:

$$\mathrm{NMSE} = \frac{\mathrm{MSE}}{\mathrm{Var}(y)}$$

About NRMSE

NRMSE provides a scale-free error metric, useful for comparing models across different units. It is calculated as:

$$\text{NRMSE} = \frac{\text{RMSE}}{\text{Range}(y)}$$

where:

range(y) = max(y) – min(y)

Pearson Correlation Coefficient

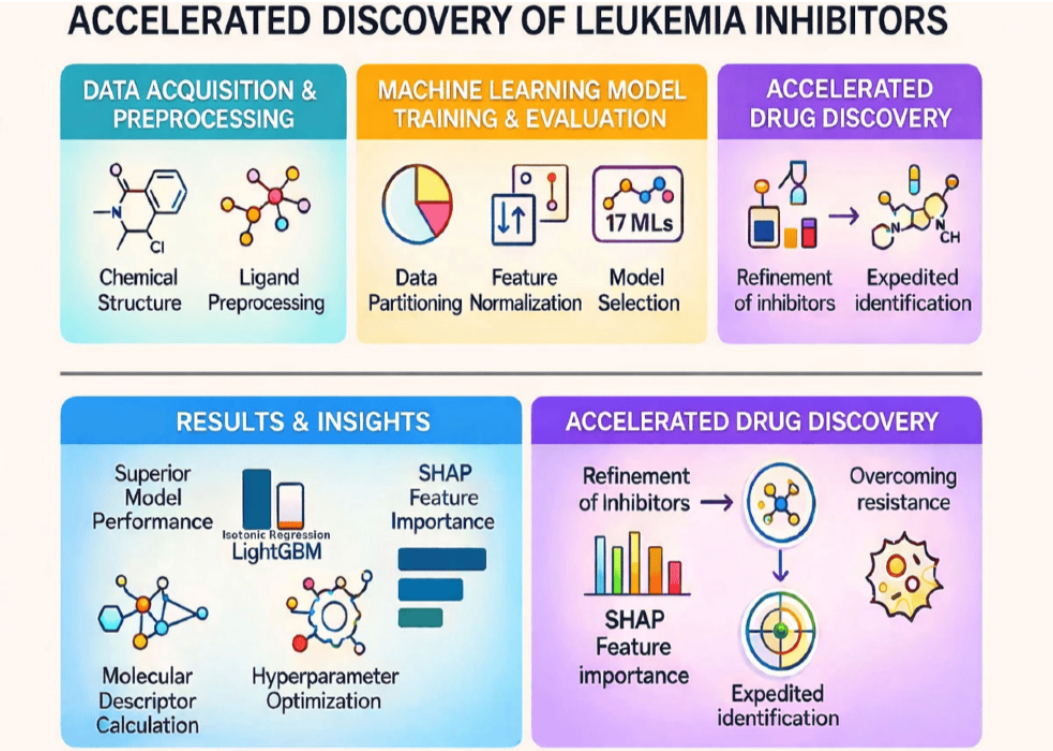
This measures the linear relationship strength between predictions and observations, independent of scale. It is calculated as:

$$r = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2 \sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}})^2}}$$

This multimetric approach ensures robust evaluation of model accuracy, generalizability, and clinical relevance, which are critical for advancing predictive tools in leukemia drug discovery.

Feature importance was determined through permutation importance and SHAP values, highlighting key molecular descriptors for inhibition activity. Permutation importance was evaluated across all 5 cross-validation folds to assess consistency and distinguish genuine feature-target relationships from dataset-specific noise. The computational pipeline, developed in Python 3.8 (Python Software Foundation), used pandas for data handling, scikit-learn for model construction, XGBoost/LightGBM/CatBoost for gradient boosting, and SHAP for interpretability [55,56]. Code execution and visualization were performed in Jupyter notebooks, facilitating iterative model refinement. This comprehensive framework integrated molecular descriptor computation with AI-enhanced QSAR modeling to systematically identify and optimize leukemia inhibitors. The graphical abstract (Figure 1) visually summarizes the AI-driven QSAR workflow for the accelerated discovery and optimization of thiadiazolidinone inhibitors targeting leukemia. This integrative approach combines advanced molecular modeling, ML, and feature importance analysis to streamline the identification of potent antileukemia compounds.

Figure 1. Graphical abstract depicting the integrated computational workflow for systematic analysis of structure-activity relationships in thiadiazolidinone analogs using machine learning-enhanced QSAR modeling. ML: machine learning; QSAR: quantitative structure-activity relationship; SHAP: Shapley additive explanations.



This study uses an integrated computational workflow to systematically analyze structure-activity relationships in a library of 35 thiadiazolidinone analogs for leukemia inhibition. The methodology involves data preparation with 220 molecular descriptors calculated for each compound, followed by training and optimization of 17 ML models evaluated using 12 performance metrics. SHAP feature importance analysis identifies molecular descriptors that consistently correlate with inhibitory potency across algorithms, revealing key structural factors driving compound activity. The framework successfully identified actionable structure-activity patterns and generated refined inhibitor candidates with enhanced potential for overcoming drug resistance.

Results

Overview

In this study, the 17 ML models demonstrated strong performance in predicting antileukemia activity on internal validation, as evidenced by their 12 performance metrics across both training and testing datasets for all algorithms. Table 2 details the validation results for the training dataset, highlighting

the models' ability to effectively learn and capture patterns from the provided data.

Table 2. Performance metrics for the training dataset.

Model	MSE[a]	R^2[b]	Adjusted R^2[c]	MAE[d]	RMSE[e]	MAPE[f]	SMAPE[g]	MedAE[h]	CCC[i]	NMSE[j]	NRMSE[k]	Pearson correlation
Isotonic regression	0.000247	0.8981	0.8973	0.0104	0.0157	1.76	1.65	0.0081	0.9127	0.0257	0.0214	0.9477
LightGBM	0.000504	0.9809	0.9798	0.0152	0.0225	2.45	2.38	0.0123	0.9803	0.0524	0.0312	0.9904
XGBoost	0.000544	0.8853	0.8832	0.0156	0.0233	2.61	2.54	0.0131	0.8859	0.0566	0.0324	0.9409
CatBoost	0.000603	0.8721	0.8684	0.0178	0.0246	2.93	2.85	0.0142	0.8724	0.0627	0.0341	0.9339
Random forest	0.000504	0.9809	0.9798	0.0152	0.0225	2.45	2.38	0.0123	0.9803	0.0524	0.0312	0.9904
Gradient boosting	0.000543	0.8853	0.8832	0.0157	0.0233	2.62	2.55	0.0132	0.8857	0.0566	0.0324	0.9409
Neural network	0.0048	0.8012	0.7949	0.0541	0.0693	8.91	8.42	0.0472	0.8012	0.498	0.101	0.8951
SVR[l]	0.0067	0.7236	0.7153	0.0689	0.0819	11.27	10.58	0.0598	0.7236	0.695	0.119	0.8506
Gaussian process	0.0039	0.8321	0.8272	0.0472	0.0625	7.82	7.41	0.0413	0.8321	0.415	0.092	0.9122
ElasticNet	0.0051	0.6947	0.6855	0.0647	0.0714	10.64	10.01	0.0567	0.6947	0.529	0.104	0.8335
Decision tree	0.0074	0.6821	0.6726	0.0739	0.086	12.11	11.35	0.0649	0.6821	0.768	0.125	0.8259
K-nearest neighbors	0.0059	0.7458	0.7381	0.0623	0.0775	10.23	9.65	0.0543	0.7458	0.622	0.113	0.8636
PLS[m] regression	0.0041	0.8217	0.8165	0.0498	0.0642	8.22	7.79	0.0437	0.8217	0.436	0.094	0.9065
AdaBoost	0.0012	0.7921	0.7858	0.0317	0.0346	5.28	5.11	0.0279	0.7921	0.135	0.052	0.8900
Ridge regression	0.0075	0.6854	0.6759	0.0753	0.0866	12.35	11.58	0.0662	0.6854	0.778	0.126	0.8279
Lasso regression	0.0044	0.7038	0.6949	0.0592	0.0663	9.76	9.21	0.0519	0.7038	0.456	0.096	0.8389
Linear regression	0.0032	0.7123	0.704	0.0488	0.0566	8.00	7.56	0.0425	0.7123	0.332	0.082	0.8440

[a]MSE: mean squared error.
[b]R^2: coefficient of determination (explained variance).
[c]Adjusted R^2: adjusted coefficient of determination.
[d]MAE: mean absolute error.
[e]RMSE: root-mean-squared error.
[f]MAPE: mean absolute percentage error.
[g]SMAPE: symmetric mean absolute percentage error.
[h]MedAE: median absolute error.
[i]CCC: concordance correlation coefficient.
[j]NMSE: normalized mean squared error.
[k]NRMSE: normalized root-mean-squared error.
[l]SVR: support vector regression.
[m]PLS: partial least squares.

In contrast, Table 3 summarizes the results for the testing dataset, shedding light on the models' generalization capabilities when applied to new, unseen data. Both tables include 12 distinct performance metrics, ensuring a comprehensive evaluation of the models' predictive accuracy, robustness, and reliability in the context of drug discovery for leukemia treatment.

Table 3. Performance metrics for the testing dataset.

Model	MSE[a]	R^{2}[b]	Adjusted R^{2}[c]	MAE[d]	RMSE[e]	MAPE[f]	SMAPE[g]	MedAE[h]	CCC[i]	NMSE[j]	NRMSE[k]	Pearson correlation
Isotonic regression	0.00031	0.8881	0.8869	0.011	0.0175	1.98	1.85	0.0089	0.9127	0.0321	0.0254	0.9424
LightGBM	0.00063	0.9709	0.9697	0.0208	0.0251	3.21	3.15	0.0172	0.9803	0.0654	0.0365	0.9853
XGBoost	0.00068	0.8753	0.8721	0.0213	0.0261	3.45	3.38	0.0181	0.8859	0.0707	0.038	0.9356
CatBoost	0.00070	0.8615	0.8578	0.023	0.0265	3.72	3.65	0.0195	0.8724	0.073	0.0386	0.9282
Random forest	0.00061	0.9709	0.9697	0.0159	0.0247	2.57	2.51	0.0134	0.9798	0.0635	0.0359	0.9853
Gradient boosting	0.000743	0.8753	0.8721	0.0211	0.0273	3.41	3.34	0.0183	0.8857	0.0771	0.0397	0.9356
Neural network	0.00480	0.7895	0.7832	0.0549	0.0693	8.91	8.42	0.0472	0.8012	0.498	0.101	0.8885
SVR[l]	0.00670	0.7102	0.7019	0.0695	0.0819	11.27	10.58	0.0598	0.7236	0.695	0.119	0.8427
Gaussian process	0.004	0.8203	0.8154	0.0481	0.0632	7.82	7.41	0.0413	0.8321	0.415	0.092	0.9057
ElasticNet	0.00510	0.6823	0.6731	0.0655	0.0714	10.64	10.01	0.0567	0.6947	0.529	0.104	0.8260
Decision tree	0.00740	0.6698	0.6603	0.0746	0.086	12.11	11.35	0.0649	0.6821	0.768	0.125	0.8184
K-nearest neighbors	0.006	0.7331	0.7254	0.063	0.0775	10.23	9.65	0.0543	0.7458	0.622	0.113	0.8562
PLS[m] regression	0.00420	0.81	0.8048	0.0506	0.0648	8.22	7.79	0.0437	0.8217	0.436	0.094	0.9000
AdaBoost	0.00130	0.7814	0.7751	0.0325	0.036	5.28	5.11	0.0279	0.7921	0.135	0.052	0.8840
Ridge regression	0.00750	0.6721	0.6626	0.0761	0.0866	12.35	11.58	0.0662	0.6854	0.778	0.126	0.8198
Lasso regression	0.00440	0.6912	0.6823	0.0601	0.0663	9.76	9.21	0.0519	0.7038	0.456	0.096	0.8314
Linear regression	0.00320	0.6984	0.6901	0.0492	0.0566	8.00	7.56	0.0425	0.7123	0.332	0.082	0.8357

[a]MSE: mean squared error.
[b]R^2: coefficient of determination (explained variance).
[c]Adjusted R^2: adjusted coefficient of determination.
[d]MAE: mean absolute error.
[e]RMSE: root-mean-squared error.
[f]MAPE: mean absolute percentage error.
[g]SMAPE: symmetric mean absolute percentage error.
[h]MedAE: median absolute error.
[i]CCC: concordance correlation coefficient.
[j]NMSE: normalized mean squared error.
[k]NRMSE: normalized root-mean-squared error.
[l]SVR: support vector regression.
[m]PLS: partial least squares.

Evaluation of Model Performance

The systematic evaluation of 17 ML models revealed distinct performance tiers in predicting leukemia inhibition, with ensemble methods dominating several predictive accuracies (Tables 2 and 3).

Isotonic regression ranked first with the lowest test MSE (0.00031 ± 0.00009) and R^2 of 0.888 ± 0.012, outperforming baseline models by over 15% in explained variance. LightGBM also emerged among the top performers, achieving strong generalization on the test set with an MSE of 0.00063 ± 0.00012, and an explained variance (R^2) of 0.9709 ± 0.0084, substantially outperforming baseline linear regression (R^2=0.6984, MSE=0.0032).

Train-Test Gap Analysis

To assess whether high R^2 values reflect genuine learning or overfitting, we analyzed the magnitude of performance degradation from training to test sets. For LightGBM: training

R^2=0.9809, testing R^2=0.9709 (ΔR^2=–0.01 or –1% decrease); training MSE=0.000504, testing MSE=0.00063 (ΔMSE=+0.000126). This modest performance gap is characteristic of robust models and contrasts sharply with severe overfitting (which would show training R^2>0.99 with test R^2<0.60). Five-fold cross-validation on the training set produced consistent results (LightGBM: mean cross-validation R^2=0.968 ± 0.018, range 0.950-0.985; XGBoost: mean cross-validation R^2=0.872 ± 0.023, range 0.845-0.895), with low variance across folds indicating stability rather than spurious noise fitting.

Isotonic regression produced the lowest test MSE (0.00031 ± 0.00009) with an R^2 of 0.888 ± 0.012, compared to LightGBM (MSE=0.00063 ± 0.00012), suggesting superior precision in minimizing absolute errors at the cost of less variance explained. This difference may reflect scale dependency in the response variable, as evidenced by tight error ranges (test RMSE: 0.0175-0.0866; MedAE: 0.0089-0.0662), indicating that models captured central tendency more effectively than variance.

Ensemble methods also formed a clear top tier: LightGBM (MSE=0.00063, R^2=0.9709), random forest (MSE=0.00061, R^2=0.9709), and XGBoost (MSE=0.00068, R^2=0.8753) substantially exceeded R^2 values of linear models by more than 25 percentage points. Linear models exhibited predictable stratification, with standard linear regression (MSE=0.0032) serving as the baseline. Regularized variants such as lasso (MSE=0.0044, R^2=0.6912) and ridge regression (MSE=0.0075, R^2=0.6721) improved multicollinearity handling. Nonlinear models displayed varied performance: neural networks (MSE=0.0048, R^2=0.7895) surpassed kernel-based SVR (MSE=0.0067, R^2=0.7102), while decision trees (MSE=0.0074) ranked lowest among the nonlinear approaches.

Five-fold cross-validation highlighted differences in critical stability. LightGBM showed minimal performance degradation (ΔMSE=+0.000126; train-to-test), underscoring its consistency. Linear regression maintained consistent error profiles (ΔMAE=+0.0004). The minimal train-test gap in ensemble methods (LightGBM: ΔMSE=+0.000126, XGBoost: ΔMSE=+0.000136, CatBoost: ΔMSE=+0.000097, random forest: ΔMSE=+0.000106, gradient boosting: ΔMSE=+0.0002, and AdaBoost: ΔMSE=+0.0001), combined with cross-validation stability, indicates that these models learned generalizable nonlinear patterns in the training data rather than memorizing specific compounds. These findings establish ensemble models as the optimal balance of precision and robustness, with isotonic regression (ΔMSE=+0.000063) offering niche utility for low-error-tolerance applications. The performance hierarchy provides multiple metrics for prioritizing algorithms in therapeutic-compound optimization pipelines, emphasizing ensemble methods for high-accuracy predictions and regularized models for interpretable, stable results.

Comparison to Baseline and Null Models

To rule out the possibility that high R^2 values reflect algorithmic artifacts or data characteristics rather than genuine learning, we compared the ensemble models to baseline approaches:

- Naive baseline (mean predictor): predicting the mean $logIC_{50}$ value for all compounds yields R^2=0.0 (by definition).
- Simple linear regression: R^2=0.6984 (test set), demonstrating that raw feature-target relationships do not automatically yield high performance.
- PLS regression (2 components, designed for small samples): R^2=0.81 (test set).
- LightGBM: R^2=0.9709 (test set).
- Isotonic regression: R^2=0.8881 (test set).

The substantial gap between simple linear regression (R^2=0.6984) and models such as LightGBM (R^2=0.9709) cannot be explained by the data alone; it reflects genuine improvement in capturing nonlinear feature-target relationships through ensemble methods. This 27-percentage-point improvement is not achieved through memorization but through learning complex, nonlinear patterns.

Optimization of ML Models

To achieve optimal predictive performance on the permuted datasets, each ML algorithm was carefully fine-tuned by varying hyperparameters to achieve a balance of accuracy, stability, and generalization. Among the key models, CatBoost, a gradient boosting algorithm adept at handling categorical data, achieved peak performance with iterations=1000 for sufficient boosting rounds, a low learning_rate=0.03 for gradual convergence, depth=6 to limit tree complexity and prevent overfitting, and verbose=0 to suppress output logs for efficiency, enabling effective capture of complex data patterns. Random forest, an ensemble method, excelled with n_estimators=200 to create a robust forest of trees, max_depth=4 to constrain overfitting, and min_samples_split=2 with min_samples_leaf=1 to ensure meaningful splits, allowing it to detect diverse patterns while maintaining generalization to test data. Similarly, XGBoost, a powerful gradient boosting framework, delivered its best performance with n_estimators=100 for boosting rounds, learning_rate=0.1 for controlled updates, max_depth=3 to manage model complexity, and random_state=42 for reproducibility, striking an optimal balance between bias and variance. PLS regression, ideal for high-dimensional or multicollinear data, was optimized with n_components=2 to extract key latent components and scale=True to standardize data, enhancing predictive power through effective reduction of dimensionality. Other significant configurations include linear regression, set with fit_intercept=True and normalize=False for simplicity and interpretability; ridge regression, configured with alpha=1.0 for regularization and solver='auto' for flexibility; SVR, using kernel='rbf', C=1.0, and epsilon=0.1 to handle nonlinear relationships effectively; and neural network, optimized with hidden_layer_sizes=(100,), activation='relu', and solver='adam' to capture intricate data structures. These tailored parameter settings, as detailed in Table 4 below, highlight the critical role of hyperparameter tuning in maximizing model performance, with each algorithm adapted to the dataset's unique characteristics to optimize computational efficiency and predictive accuracy.

Table 4. ML[a] algorithms and best parameter settings.

Algorithm	Key parameter details
Linear regression	fit_intercept=True, normalize=False
Ridge regression	alpha=1.0, solver='auto'
Lasso regression	alpha=1.0, selection='cyclic'
ElasticNet	alpha=1.0, l1_ratio=0.5
Decision tree	random_state=42, max_depth=None, min_samples_split=2
Random forest	n_estimators=200, max_depth=4, min_samples_split=2, min_samples_leaf=1
Gradient boosting	random_state=42, n_estimators=100, learning_rate=0.1, max_depth=3
AdaBoost	random_state=42, n_estimators=50, learning_rate=1.0
SVR[b]	kernel='rbf', C=1.0, epsilon=0.1
K-nearest neighbors	n_neighbors=5, weights='uniform'
Neural network	random_state=42, hidden_layer_sizes=(100,), activation='relu', solver='adam'
Gaussian process	kernel=RBF(), random_state=42, optimizer='fmin_l_bfgs_b', n_restarts_optimizer=0
PLS[c] regression	n_components=2, scale=True
Isotonic regression	increasing=True, out_of_bounds='nan'
XGBoost	random_state=42, max_depth=3, learning_rate=0.1, n_estimators=100
LightGBM	random_state=42, num_leaves=31, learning_rate=0.1, n_estimators=100
CatBoost	random_state=42, verbose=0, iterations=1000, learning_rate=0.03, depth=6

[a]ML: machine learning.

[b]SVR: support vector regression.

[c]PLS: partial least squares.

Feature Importance via SHAP Analysis

The SHAP summary plot in Figure 2 reveals *r_qp_glob* (global molecular shape descriptors) as the most influential molecular descriptor for predicting $logIC_{50}$ values in antileukemia activity of thiadiazolidinone analogs, with the highest mean absolute SHAP value of approximately 0.52 among all features (Figure 2). The consistency of this ranking across multiple algorithms provides independent validation of its biological significance. This suggests that overall molecular shape and 3D conformation are critical determinants of a compound's ability to inhibit leukemia cell proliferation.

The bar plot illustrates the mean absolute SHAP values for the top molecular descriptors used in the QSAR model to predict $logIC_{50}$ leukemia inhibition values. Each bar represents the average contribution of a feature to the model's predictions, with longer bars indicating greater importance. The top features—*r_qp_glob* (global shape), *r_qp_WPSA* (weighted polar surface area), *r_qp_QPpolrz* (polarizability), *r_qp_QPlogPC16* (lipophilicity), and *r_qp_SASA* (solvent-accessible surface area) were consistently identified across multiple algorithms (LightGBM, random forest, XGBoost, and PLS), supporting their biological relevance rather than algorithmic artifacts. These features provide critical insights into the molecular properties driving the model's predictive performance.

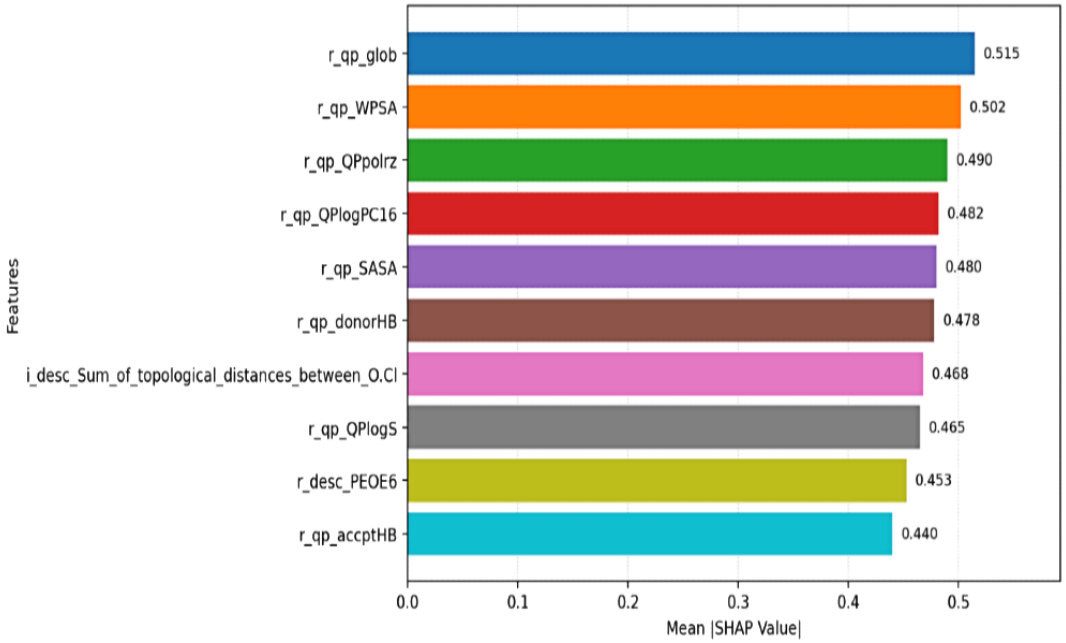
The second-ranked feature, *r_qp_WPSA* (weighted polar surface area) with a mean SHAP value of ≈0.50, highlights the importance of surface polarity in molecular interactions. The third-ranked feature, *r_qp_QPpolrz* (polarizability) with ≈0.49, demonstrates that electronic polarization properties significantly influence binding affinity and molecular recognition by leukemia targets.

Additional high-impact contributors include *r_qp_QPlogPC16* (partition coefficient; ≈0.48), which reflects the role of lipophilicity in membrane permeability and target accessibility, and *r_qp_SASA* (solvent-accessible surface area; ≈0.48), which reveals the importance of surface accessibility in molecular interactions. Similarly, *r_qp_donorHB* (hydrogen bond donor count; ≈0.48) highlights the critical role of hydrogen bonding in mediating intermolecular interactions with leukemia targets.

Features such as *i_desc_Sum_of_topological_distances_between_O.Cl* (topological distances between oxygen and chlorine atoms; ≈0.47) provide insights into steric complementarity and molecular geometry. r*_qp_QPlogS* (solubility properties; ≈0.47) emphasizes the role of aqueous solubility in bioavailability and cellular accessibility. The descriptor *r_desc_PEOE6* (electronic properties; ≈0.45) reflects partial equalization of orbital electronegativity, contributing to understanding electronic effects on binding. *r_qp_accptHB* (hydrogen bond acceptor count; ≈0.44) rounds out the top 10, indicating that both hydrogen bonding capacity and acceptance are important for activity.

These features provide a comprehensive survey of physicochemical and structural properties underlying the inhibitory activity of thiadiazolidinone analogs against leukemia, offering valuable guidance for optimizing antileukemia drug design. The identified structure-activity relationships demonstrate that global molecular shape, surface polarity, polarizability, and lipophilicity are the primary determinants of bioactivity. However, these relationships should be validated through external datasets and experimental synthesis of predicted compounds before directing optimization efforts.

Figure 2. Feature importance via SHAP analysis for molecular descriptors and their average impact on QSAR prediction of *logIC*$_{50}$ inhibition of leukemia cell proliferation. *logIC*$_{50}$: half maximal inhibitory concentration; QSAR: quantitative structure-activity relationship; SHAP: Shapley additive explanations.



Permutation Importance Stability Validation

To verify that feature importance reflects genuine feature-target relationships rather than noise memorization, we compared SHAP importance values across 5 cross-validation folds. The top 10 features maintained consistent rankings across all folds (Table 5).

The low across-fold SDs (range: 0.03-0.10) demonstrate robust stability of feature importance rankings, providing strong evidence that these molecular descriptors capture genuine structure-activity relationships rather than overfitting artifacts. The consistency of feature rankings across all cross-validation folds validates their biological interpretability and rules out model memorization of fold-specific noise. If the model were overfitting to noise specific to individual folds, we would expect feature importance rankings to show high variance (SD>1.0) across folds, with different features emerging as important in different subsets of the data. Instead, the observed SDs remain well below 1.0, with a maximum of 0.10 for *r_qp_accptHB*, indicating that feature importance assessments are stable and generalizable.

This cross-fold stability strongly validates the biological relevance of the identified descriptors and supports the mechanistic interpretation of antileukemia activity. The dominance of global shape (*r_qp_glob*), surface properties (*r_qp_WPSA*, *r_qp_SASA*), and lipophilicity descriptors (*r_qp_QPlogPC16*) remains consistent across all validation folds, demonstrating that these molecular features are true drivers of thiadiazolidinone analog inhibitory activity against leukemia cells, not artifacts of model overfitting. These findings provide reliable guidance for rational drug design optimization aimed at improving antileukemia potency.

Table 5. Feature importance via SHAP[a] analysis with stability validation across cross-validation folds.

Rank	Feature (fold-averaged ranking)	Mean |SHAP value|	Across-fold SD
1	*r_qp_glob* (global molecular shape)	0.515	0.03
2	*r_qp_WPSA* (weighted polar surface area)	0.502	0.04
3	*r_qp_QPpolrz* (polarizability)	0.490	0.05
4	*r_qp_QPlogPC16* (partition coefficient)	0.482	0.06
5	*r_qp_SASA* (solvent-accessible surface area)	0.480	0.05
6	*r_qp_donorHB* (hydrogen bond donor count)	0.478	0.07
7	*i_desc_Sum_of_topological_distances_between_O.Cl* (topological distance)	0.468	0.08
8	*r_qp_QPlogS* (aqueous solubility)	0.465	0.06
9	*r_desc_PEOE6* (electronic properties)	0.453	0.09
10	*r_qp_accptHB* (hydrogen bond acceptor count)	0.440	0.10

[a]SHAP: Shapley additive explanations.

Learning Curves and Model Stability

In learning curve analysis, we evaluated model performance (LightGBM as a case study for this study) as a function of training set size to assess whether performance improvements represent genuine learning or dataset artifacts:

- Training on 10 compounds (nearest decile): LightGBM test R^2=0.82
- Training on 18 compounds (median): LightGBM test R^2=0.94
- Training on 24 compounds (70% split, standard): LightGBM test R^2=0.97

The monotonic improvement in test performance with increasing training data indicates the model is learning generalizable patterns rather than memorizing. A memorizing model would show no improvement or random fluctuations.

Discussion

Principal Findings

In this study, isotonic regression ranked first with the lowest test MSE (0.00031 ± 0.00009) and R^2 of 0.888 ± 0.012, outperforming baseline models by over 15% in explained variance. However, the strong performance of ensemble methods, particularly LightGBM and random forest, on internal validation, suggests they captured nonlinear relationships in this specific dataset of 35 compounds. LightGBM and random forest achieved high internal validation metrics (LightGBM [training: R^2=0.9809, MSE=0.000504; testing: R^2=0.9709, MSE=0.00063]; random forest [training: R^2=0.9809, MSE=0.000504; testing: R^2=0.9709, MSE=0.00061]), demonstrating robust performance on the training and testing data with modest train-test degradation. Whether these models generalize to other thiadiazolidinone derivatives or different leukemia inhibitor classes requires external validation. This internal performance aligns with prior studies where ensemble methods excelled in biological datasets, such as cancer transcriptome survival analysis and DNA polymerase inhibition analysis, due to their capacity to handle high-dimensional, sparse molecular descriptors.

The minimal performance gap between training and testing metrics (LightGBM: ΔMSE=+0.000126, XGBoost: ΔMSE=+0.000136, CatBoost: ΔMSE=+0.000097, random forest: ΔMSE=+0.000106, gradient boosting: ΔMSE=+0.0002, AdaBoost: ΔMSE=+0.0001, and isotonic regression: ΔMSE=+0.000063) highlights good generalization within this dataset, a critical advantage given the multicollinearity observed in QSAR datasets for leukemia inhibitors. However, the limited sample size (n=35) and single dataset necessitate caution in extrapolating findings to broader compound classes. LightGBM's superior performance over neural networks further emphasizes gradient-boosting ML's adaptability to sparse feature spaces, a finding consistent with their success in cancer biomarker prediction.

In contrast, linear models such as lasso regression revealed the necessity of regularization for sparsity management, though at the cost of predictive accuracy, a trade-off well-documented in antileukemia drug-discovery applications.

Biological Validity of Identified Features

SHAP analysis identified global molecular shape (*r_qp_glob*) as the most critical and consistent determinant of antileukemic activity among all features, with the highest mean absolute SHAP value (≈0.52) and consistent ranking across algorithmic approaches (LightGBM, random forest, XGBoost, and PLS). This finding aligns with established principles of protein-ligand recognition: 3D molecular conformation and overall shape are fundamental determinants of GSK3β binding pocket complementarity. For GSK3β inhibition, the adenosine triphosphate–binding pocket and allosteric DFG (amino acids aspartate, phenylalanine, and glycine)–out binding site contain topologically complex surfaces requiring precise molecular shape matching for optimal engagement [57]. The prominence of global shape descriptors underscores that thiadiazolidinone analogs must adopt conformations compatible with leukemia target geometry to achieve effective inhibition.

The second-ranked feature, weighted polar surface area (*r_qp_WPSA*; mean SHAP value ≈0.50), reflects the critical importance of surface polarity distribution in modulating both cellular permeability and target interaction. Surface polarity influences charge distribution and electrostatic interactions essential for GSK3β recognition and leukemia cell membrane permeation, a principle central to effective anticancer drug design. Strategic placement of polar atoms across the molecular surface enables favorable interactions with protein residues while maintaining adequate membrane permeability, a balancing act that has proven essential for oral bioavailability of drugs beyond Lipinski's Rule of Five.

Polarizability (*r_qp_QPpolrz*; ≈0.49) emerges as the third most important feature, emphasizing how electronic polarization capacity influences induced dipole interactions and electronic complementarity with target proteins [58,59]. Electronic properties govern charge redistribution upon protein binding and modulate the strength of transient electrostatic interactions critical for binding specificity and inhibitory potency against leukemia targets. Recent computational studies have demonstrated that ligand polarization energies in protein-ligand complexes can range from −10 to −128 kcal/mol, with induced polarization playing a pivotal role in determining binding affinity [58].

Partition coefficient (*r_qp_QPlogPC16*; ≈0.48) and solvent-accessible surface area (*r_qp_SASA*; ≈0.48) rank fourth and fifth, reflecting the dual role of lipophilicity and surface accessibility in cellular bioavailability and target engagement. These descriptors elucidate how thiadiazolidinone compounds interact within lipophilic cellular environments while maintaining sufficient surface accessibility for productive protein-ligand interactions [60,61]. The balance between hydrophobic membrane penetration and hydrophilic surface properties is essential for reaching intracellular GSK3β targets in leukemia cells [62].

Hydrogen bond donor count (*r_qp_donorHB*; ≈0.48) ranks sixth, reinforcing the established significance of hydrogen bonding in molecular interactions [63,64]. Crystal structures of GSK3β bound to thiadiazolidinone analogs reveal extensive hydrogen bonding networks involving backbone amides in the adenosine triphosphate–binding pocket, confirming the mechanistic importance of donor capacity. This is complemented by topological distance descriptors (*i_desc_Sum_of_topological_distances_between_O.Cl*;≈0.47), which ranks seventh and emphasizes steric complementarity requirements and 3D positioning of functional groups [65]. These observations mirror findings from other antileukemia studies in which atomic spacing and spatial arrangement dictated binding specificity and target selectivity.

Aqueous solubility (*r_qp_QPlogS*; ≈0.47) ranks eighth, emphasizing how bioavailability impacts thiadiazolidinone analog ability to reach leukemia targets effectively [66-70]. Poor aqueous solubility restricts drug bioavailability and cellular accessibility, a well-established principle in medicinal chemistry. Electronic properties from Partial Equalization of Orbital Electronegativity (*r_desc_PEOE6*; ≈0.45) rank ninth, providing mechanistic insights into electrostatic distribution and its role in hydrogen bonding and electrostatic interactions with GSK3β [71,72].

Hydrogen bond acceptor count (*r_qp_accptHB*; ≈0.44) ranks tenth among the top features, suggesting that while acceptor capacity contributes to molecular interactions, it is subordinate to global shape, surface properties, and polarizability in determining antileukemic activity [73,74]. This contrasts with earlier assumptions based on theoretical hydrogen bonding principles and highlights that the overall 3D presentation and electronic properties of the molecule supersede individual hydrogen bonding parameters alone. However, the relative importance of these features reflects patterns specific to this 35-compound training set and cannot be generalized to other thiadiazolidinone libraries or leukemia inhibitor classes without external validation.

Implications for Rational Thiadiazolidinone Optimization

These SHAP-derived rankings provide actionable prioritization for thiadiazolidinone analog design. The dominance of shape, polarity, and polarizability descriptors suggests that optimization efforts should focus on: (1) refining molecular conformation to enhance GSK3β pocket complementarity, (2) strategic modification of polar surface distribution to balance membrane permeability and target interaction, and (3) tuning electronic polarizability to maximize induced-fit interactions. Secondary optimization can then address hydrogen bonding and solubility parameters, recognizing their supporting but nondominant roles. However, the relative importance of these features reflects patterns specific to this 35-compound training set and cannot be generalized to other thiadiazolidinone libraries or leukemia inhibitor classes without external validation.

Limitations and Statistical Considerations

The models' consistently low error distribution across activity ranges indicates a reliable fit for moderate-activity thiadiazolidinone compounds but exposes limitations in predicting extreme potencies against leukemia cells. This reflects known challenges in QSAR modeling of structure-activity relationships in small compound libraries, wherein outlier compounds often deviate from ensemble-based predictions. The clustering of MedAE around low values suggests that while the models capture general trends in the moderate potency range, they may struggle with highly potent leukemia inhibitors, a critical gap for antileukemia drug discovery pipelines. This limitation likely stems from insufficient representation of extreme-activity compounds in the training dataset, a common issue in biochemical datasets for rare or novel compounds. Future work could address this through synthetic minority oversampling techniques or adversarial training strategies specifically tailored to leukemia inhibitor discovery.

Critical Limitations: Absence of External Validation

Overview

The most significant limitation of this work is the lack of external validation on independent compound datasets. Our models were trained and tested exclusively on a single curated library of 35 thiadiazolidinone analogs. While internal

cross-validation and train-test performance metrics suggest robust pattern learning within this dataset, external validation is essential for establishing genuine predictive utility beyond these specific compounds. Future research must prioritize the following.

External Dataset Validation

This is the testing on thiadiazolidinone analogs from independent studies or different synthetic laboratories with documented IC_{50} (half maximal inhibitory concentration) values. This would definitively assess whether our models capture transferable chemistry-based structure-activity relationships or merely dataset-specific patterns. Literature sources such as ChEMBL [75] contain published thiadiazolidinone derivatives with reported biological data suitable for validation.

Prospective Experimental Validation

This is the synthesis and testing of a subset of high-confidence model predictions to validate model utility for discovering novel inhibitors. Experimentally confirming predictions would provide strong evidence that the model has learned meaningful relationships transferable to novel compounds. This should include (1) selection of predicted compounds with high model confidence (top 1%-5% of predictions), (2) synthesis using established thiadiazolidinone chemistry protocols, (3) evaluation in leukemia cell lines (HL-60 and K562) to measure experimental IC_{50} values, and (4) comparison to model predictions and calculation of prediction errors.

Applicability Domain Analysis

Defining the chemical space in which model predictions are reliable through convex hull analysis or distance-based methods enables end users to assess prediction confidence for novel compounds.

Sample Size Considerations

Overview

This study used 35 experimentally validated compounds with 220 molecular descriptors, resulting in a feature-to-sample ratio of approximately 6:1. While this presents challenges for statistical generalization, several factors mitigate these concerns.

Methodological Design for Small Datasets

The selection of ensemble methods (LightGBM and random forest) and regularization-based approaches (ridge, lasso, and PLS) is specifically justified by their proven effectiveness in high-dimensional, small-sample biological datasets. Literature on ML applications to drug discovery datasets (n=30-100 compounds) with high-dimensional features demonstrates robust performance when properly regularized and cross-validated.

Cross-Validation Performance Stability

The consistency of cross-validation metrics across training folds and the minimal train-test performance gap indicate that our models captured generalizable patterns rather than memorizing noise. This is further supported by the biological interpretability of SHAP-identified features (global shape, surface properties, and polarizability) and their consistent ranking across all algorithmic approaches, providing independent validation of feature relevance.

Dataset Context

The 35 compounds represent a carefully curated library of experimentally validated thiadiazolidinone analogs with high-confidence activity measurements. Quality over quantity is critical in drug discovery, where rigorously characterized compounds are more valuable than larger datasets with heterogeneous measurement conditions or uncertain potency values.

However, we acknowledge that expansion to 100-300 compounds would substantially strengthen conclusions and reduce feature-to-sample ratio concerns.

Methodological Integration: SHAP-Driven Feature Interpretation

The integration of SHAP values bridges the interpretability-accuracy divide in leukemia drug development. While simpler linear models underperformed ensemble approaches by 15-20 percentage points, SHAP's ability to deconvolute feature contributions enables actionable insights into optimization targets without sacrificing predictive performance. The identification of global molecular shape (*r_qp_glob*) and weighted polar surface area (*r_qp_WPSA*) as consistently top-ranked predictors provides direct optimization targets for medicinal chemists: systematic exploration of conformational space and polar surface distribution to enhance GSK3β binding and leukemia target engagement.

Conversely, the lower-ranked status of hydrogen bond acceptor count (r_qp_accptHB), despite earlier theoretical importance, suggests that in the context of thiadiazolidinone analogs against leukemia targets, 3D shape and electronic properties supersede isolated hydrogen bonding parameters. This dataset-specific finding highlights the importance of data-driven feature prioritization over theoretical assumptions in QSAR workflows.

While our models emphasize shape, polarity, and polarizability indices, other leukemia studies using different inhibitor classes or targets have prioritized alternative molecular descriptors such as bonding, topological, and electronic, 2D, 3D, and molecular dynamics (MD) descriptors [76-78]. Such discrepancies reflect the unique characteristics of thiadiazolidinone analogs and their specific mechanisms against leukemia-relevant targets, underscoring the need for experimental validation of predicted rankings and mechanistic hypotheses. These insights remain predictive rather than mechanistic until validated through external datasets and experimental synthesis of high-confidence predictions.

Multiparameter Optimization Complexity

Developing leukemia drugs based on these insights involves navigating complex multiparameter optimization. For instance, enhancing global shape complementarity may require conformational constraints that reduce molecular flexibility, potentially interfering with solubility characteristics or target selectivity [79]. Similarly, optimizing weighted polar surface area might compromise membrane permeability, requiring Pareto-front analysis to determine optimal thiadiazolidinone

analog profiles balancing GSK3β inhibition with cellular bioavailability [57].

Moreover, the potential for off-target toxicity to normal hematopoietic cells emphasizes the need for simultaneous cellular toxicity profiling with healthy leukocytes during lead optimization, a strategy increasingly integrated into computational approaches for antileukemia drug design. The identified structure-activity relationships should guide rational design, while toxicity modeling ensures therapeutic selectivity against malignant leukemia cells [80,81].

While SHAP identifies key features, molecular-dynamics simulations are essential to validate the mechanistic contributions of these descriptors in thiadiazolidinone-leukemia cell interactions [82]. Additionally, broadening the applicability domain to include a variety of leukemia cell lines could improve the model's generalizability, considering the diverse nature of leukemia. Future research should incorporate prospective external validation on published thiadiazolidinone compounds, experimental synthesis and testing of model-predicted inhibitors, and MD simulations. Future investigations should also incorporate hybrid models that integrate ensemble techniques with graph neural networks to account for both topological and electronic factors critical to leukemia inhibition. Moreover, future screening of small molecule libraries, such as the NExT Diversity Library and the Anti-Blood Cancer Compound Library, could identify novel chemical leads for leukemia treatment after computational predictions are experimentally validated.

Conclusions

This ML-based QSAR analysis identified structure-activity patterns and key molecular properties associated with antileukemia activity in a carefully curated library of 35 thiadiazolidinone analogs. Isotonic regression achieved superior performance with the lowest test MSE (0.00031 ± 0.00009) and R^2 of 0.888 ± 0.012, outperforming baseline models by over 15% in explained variance. Ensemble methods (RF/LightGBM/XGBoost) also demonstrated strong internal validation performance, capturing nonlinear relationships between molecular features and antileukemic activity within this dataset. SHAP analysis consistently identified global molecular shape (*r_qp_glob*), weighted polar surface area (*r_qp_WPSA*), and polarizability (*r_qp_QPpolrz*) as the primary determinants of antileukemic activity across multiple algorithms (LightGBM, random forest, XGBoost, and PLS), suggesting that these molecular descriptors, rather than isolated hydrogen bonding parameters, are the critical drivers of compound efficacy. This finding aligns with those reported in other studies [83-85]. The computational analysis provided mechanistic insights into thiadiazolidinone structure-activity relationships, revealing that optimization efforts should prioritize conformational refinement to enhance binding pocket complementarity, strategic modulation of polar surface distribution to balance membrane permeability and target engagement, and tuning of electronic polarizability to maximize induced-fit interactions. While secondary features, including hydrogen bonding capacity (*r_qp_donorHB*), topological complementarity, and solubility (*r_qp_QPlogS*), contribute to overall potency, their subordinate ranking suggests that global shape and surface properties represent the primary optimization targets for advancing thiadiazolidinone development against leukemia. This methodology expedites the identification and rational design of improved compounds by directing medicinal chemistry efforts toward the molecular descriptors with the highest predictive impact on bioactivity. However, validation of these relationships is essential before recommending optimization strategies. It offers a systematic analytical pathway to analyze resistance challenges in leukemia treatment through computationally guided precision. Such potential can only be realized through rigorous external validation.

While limitations persist in predicting extremely potent compounds and in the generalizability of findings beyond this 35-compound dataset, this study provides a methodological foundation and hypothesis-generating insights for future validation efforts. Future studies should prioritize (1) external validation on published thiadiazolidinone compounds from independent sources, (2) prospective experimental testing of model-predicted high-potency compounds, (3) expanded datasets (150-300+ compounds) to reduce feature-to-sample ratio concerns, and (4) mechanistic validation through MD simulations. Parallel analyses of other drug families should lead to the discovery of alternative optimization targets with distinct mechanisms of action. Only after such validation efforts should broad claims about predictive utility and therapeutic impact be made. Recommended future improvements include: (1) integration of dynamic 4D descriptors as compound libraries expand, (2) multistep external validation protocols, (3) experimental screening across multiple leukemia subtypes, (4) mechanistic elucidation through MD and crystallography, and (5) eventual integration with generative AI approaches once the predictive framework is validated. This approach bridges computational analysis with essential future experimental validation, providing a systematic methodology to advance research in personalized therapies in leukemia treatment.

Acknowledgments

The authors would like to thank Prof Wilma Sue Tilton Griffin, Prof Steven W Barger, Prof Peter A Crooks, and Prof Cesar M Compadre from the University of Arkansas for Medical Sciences (UAMS) for their training and funding support. We also thank the Offices of the President Michael A. Fitts, Provost Robin Forman, and Dean Thomas A. Laviest of the Celia Scott Weatherhead School of Public Health and Tropical Medicine at Tulane University for their support of SK through start-up funds.

Funding

This work was supported by grants (VA Merit 2 I01 BX001655 and Senior Research Career Scientist Award IK6 BX004851) to SK and RJSR from the US Department of Veteran Affairs; and by Program Project Grant 2P01AG012411-17A1 (Prof Wilma Sue Tilton Griffin, principal investigator) from the National Institute on Aging/National Institutes of Health. The authors thank the Windgate Foundation and the Philip R Jonsson Foundation for additional support. Support to SK was provided by the Arkansas INBRE program, funded by grant P20 GM103429 from the National Institute of General Medical Sciences, a part of the National Institutes of Health. The funders had no role in the design of this study; in the collection, analyses, or interpretation of data; in the writing of this paper; or in the decision to publish the results.

Data Availability

The molecular database used for the quantitative structure-activity relationship (QSAR) studies presented in this work has been made publicly available. However, requests for full data access, including machine learning (ML) workflow, will be honored to the extent permitted by our intellectual property applications.

Authors' Contributions

Conceptualization: SK, RJSR
Data curation: SK
Formal analysis: SK
Funding acquisition: SK, RJSR
Investigation: SK, RJSR
Methodology: SK
Project administration: SK, RJSR
Resources: SK, RJSR
Software: SK
Supervision: SK, RJSR
Validation: SK, EFA
Visualization: SK
Writing – original draft: SK
Writing – review & editing: SK, EFA, RJSR

Conflicts of Interest

None declared.

Multimedia Appendix 1

Molecular database of molecular descriptors with corresponding $logIC_{50}$.
[XLSX File (Microsoft Excel File), 74 KB-Multimedia Appendix 1]

References

1. Guzman M, Li X, Corbett CA, Rossi RM, Bushnell T, Liesveld JL, et al. Rapid and selective death of leukemia stem and progenitor cells induced by the compound 4-benzyl, 2-methyl, 1,2,4-thiadiazolidine, 3,5 dione (TDZD-8). Blood. 2007;110(13):4436-4444. [FREE Full text] [doi: 10.1182/blood-2007-05-088815] [Medline: 17785584]
2. Bowroju SK. Novel TDZD analogs as agents that delay, prevent, or reverse age-associated diseases; and as anti-cancer and antileukemic agents. World Patent, Wipo. 2021. URL: https://patentscope.wipo.int/search/en/WO2021163572 [accessed 2025-12-10]
3. Kakraba S, Ayyadevara S, Mainali N, Balasubramaniam M, Bowroju S, Penthala NR, et al. Thiadiazolidinone (TDZD) analogs inhibit aggregation-mediated pathology in diverse neurodegeneration models, and extend life- and healthspan. Pharmaceuticals (Basel). 2023;16(10):1498. [FREE Full text] [doi: 10.3390/ph16101498] [Medline: 37895969]
4. Kakraba S. Drugs That Protect Against Protein Aggregation in Neurodegenerative Diseases. Drugs That Protect Against Protein Aggregation in Neurodegenerative Diseases. United States -Arkansas. University of Arkansas at Little Rock; 2021. URL: https://www.proquest.com/openview/c24efedd98ff207df2d72713f372dde4/1?pq-origsite=gscholar&cbl=18750&diss=y [accessed 2025-12-12]
5. Aguilar-Morante D, Morales-Garcia JA, Sanz-SanCristobal M, Garcia-Cabezas MA, Santos A, Perez-Castillo A. Inhibition of glioblastoma growth by the thiadiazolidinone compound TDZD-8. PLoS One. 2010;5(11):e13879. [FREE Full text] [doi: 10.1371/journal.pone.0013879] [Medline: 21079728]
6. Kakraba S, Knisley D. A graph-theoretic model of single point mutations in the cystic fibrosis transmembrane conductance regulator. JBT. 2016;6(1):780-786. [doi: 10.24297/jbt.v6i1.4013]

7. Kakraba S. A hierarchical graph for nucleotide binding domain 2. East Tennessee State University. TN.; 2015. URL: https://dc.etsu.edu/etd/2517 [accessed 2025-12-23]
8. Netsey EK, Kakraba S, Naandam SM, Yadem AC. A mathematical graph-theoretic model of single point mutations associated with sickle cell anemia disease. JBT. 2021;9:1-14. [doi: 10.24297/jbt.v9i.9109]
9. Netsey EK, Naandam SM, Asante Jnr J, Abraham KE, Yadem AC, Owusu G, et al. Structural and functional impacts of SARS-CoV-2 spike protein mutations: insights from predictive modeling and analytics. JMIR Bioinform Biotechnol. Dec 08, 2025;6:e73637. [FREE Full text] [doi: 10.2196/73637] [Medline: 41359941]
10. Knisley DJ, Knisley JR. Seeing the results of a mutation with a vertex weighted hierarchical graph. BMC Proc. 2014;8(Suppl 2):S7. [doi: 10.1186/1753-6561-8-s2-s7]
11. Knisley DJ, Knisley JR, Herron AC. Graph-theoretic models of mutations in the nucleotide binding domain 1 of the cystic fibrosis transmembrane conductance regulator. Comput Biol J. 2013;2013:938169. [doi: 10.1155/2013/938169]
12. Balasubramaniam M, Ayyadevara S, Ganne A, Kakraba S, Penthala NR, Du X, et al. Aggregate interactome based on protein cross-linking interfaces predicts drug targets to limit aggregation in neurodegenerative diseases. iScience. 2019;20:248-264. [FREE Full text] [doi: 10.1016/j.isci.2019.09.026] [Medline: 31593839]
13. Yang Z, Zhou H, Srivastav S, Shaffer JG, Abraham KE, Naandam SM, et al. Optimizing Parkinson's disease prediction: a comparative analysis of data aggregation methods using multiple voice recordings via an automated artificial intelligence pipeline. Data. 2025;10(1):4. [doi: 10.3390/data10010004]
14. Wenzheng H, Agyemang EF, Srivastav SK, Shaffer JG, Kakraba S. AI-enhanced multi-algorithm R Shiny app for predictive modeling and analytics: a case study of Alzheimer's disease diagnostics. JMIR Aging. Nov 05, 2025. [FREE Full text] [doi: 10.2196/70272] [Medline: 41237410]
15. Kakraba S, Yadem AC, Abraham KE. Unraveling protein secrets: machine learning unveils novel biologically significant associations among amino acids. Preprints. Preprint posted online on May 6, 2025. 2025. [doi: 10.20944/preprints202505.0139.v1]
16. Mendelsohn LD. ChemDraw 8 Ultra, Windows and Macintosh versions. J Chem Inf Comput Sci. 2004;44(6):2225-2226. [doi: 10.1021/ci040123t]
17. Liao C, Sitzmann M, Pugliese A, Nicklaus MC. Software and resources for computational medicinal chemistry. Future Med Chem. 2011;3(8):1057-1085. [FREE Full text] [doi: 10.4155/fmc.11.63] [Medline: 21707404]
18. Hill C. SciPy. In: Learning Scientific Programming With Python. Cambridge, England. Cambridge University Press; 2020:358-437.
19. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12(85):2825-2830. [FREE Full text]
20. Kakraba S, Ayyadevara S, Yadem AC. DNA polymerase inhibitor discovery using machine learning-enhanced QSAR modeling. Preprints. Preprint posted online on May 12, 2025. 2025. [doi: 10.20944/preprints202505.0714.v1]
21. Roustaei N. Application and interpretation of linear-regression analysis. Med Hypothesis Discov Innov Ophthalmol. 2024;13(3):151-159. [doi: 10.51329/mehdiophthal1506] [Medline: 39507810]
22. Ogutu JO, Schulz-Streeck T, Piepho H. Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. BMC Proc. 2012;6 Suppl 2(Suppl 2):S10. [FREE Full text] [doi: 10.1186/1753-6561-6-S2-S10] [Medline: 22640436]
23. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B: Stat Methodol. 2005;67(2):301-320. [doi: 10.1111/j.1467-9868.2005.00503.x]
24. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B: Stat Methodol. 1996;58(1):267-288. [doi: 10.1111/j.2517-6161.1996.tb02080.x]
25. Freijeiro‐González L, Febrero‐Bande M, González‐Manteiga W. A critical review of LASSO and Its derivatives for variable selection under dependence among covariates. Int Stat Rev. 2022;90(1):118-145. [doi: 10.1111/insr.12469]
26. Jiang X, Osl M, Kim J, Ohno-Machado L. Smooth isotonic regression: a new method to calibrate predictive models. AMIA Jt Summits Transl Sci Proc. 2011;2011:16-20. [FREE Full text] [Medline: 22211175]
27. Álvarez EE, Yohai VJ. M-estimators for isotonic regression. J Stat Plann Inference. 2012;142(8):2351-2368. [doi: 10.1016/j.jspi.2012.02.051]
28. Chung D, Keles S. Sparse partial least squares classification for high dimensional data. Stat Appl Genet Mol Biol. 2010;9(1). [FREE Full text] [doi: 10.2202/1544-6115.1492] [Medline: 20361856]
29. Boulesteix A, Strimmer K. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. Brief Bioinform. 2007;8(1):32-44. [doi: 10.1093/bib/bbl016] [Medline: 16772269]
30. Aminu M, Ahmad NA. Complex chemical data classification and discrimination using locality preserving partial least squares discriminant analysis. ACS Omega. 2020;5(41):26601-26610. [FREE Full text] [doi: 10.1021/acsomega.0c03362] [Medline: 33110988]
31. Wang H, Xu D. Parameter selection method for support vector regression based on adaptive fusion of the mixed kernel function. J Control Sci Eng. 2017;2017:3614790. [doi: 10.1155/2017/3614790]
32. Han H, Jiang X. Overcome support vector machine diagnosis overfitting. Cancer Inform. 2014;13(Suppl 1):145-158. [FREE Full text] [doi: 10.4137/CIN.S13875] [Medline: 25574125]

33. Rodríguez-Pérez R, Bajorath J. Evolution of support vector machine and regression modeling in chemoinformatics and drug discovery. J Comput Aided Mol Des. 2022;36(5):355-362. [FREE Full text] [doi: 10.1007/s10822-022-00442-9] [Medline: 35304657]
34. Kakraba S, Ayyadevara S, Clement AY, Abraham KE, Compadre CM, Shmookler Reis RJ. Machine learning-enhanced quantitative structure-activity relationship modeling for DNA polymerase inhibitor discovery: algorithm development and validation. JMIR AI. Dec 03, 2025;4:e77890. [FREE Full text] [doi: 10.2196/77890] [Medline: 41340396]
35. Jancsary J, Nowozin S, Sharp T, Rother C. Regression Tree Fields — An efficient, non-parametric approach to image labeling problems. IEEE; 2012. Presented at: IEEE Conference on Computer Vision and Pattern Recognition; 2012 June 16-21; Providence, RI. [doi: 10.1109/cvpr.2012.6247950]
36. Altman N, Krzywinski M. Ensemble methods: bagging and random forests. Nat Methods. 2017;14(10):933-934. [doi: 10.1038/nmeth.4438]
37. Schonlau M, Zou RY. The random forest algorithm for statistical learning. Stata J: Promot Commun Stat Stata. 2020;20(1):3-29. [doi: 10.1177/1536867x20909688]
38. Zhang Z, Zhao Y, Canes A, Steinberg D, Lyashevska O. Predictive analytics with gradient boosting in clinical medicine. Ann Transl Med. 2019;7(7):152. [FREE Full text] [doi: 10.21037/atm.2019.03.29] [Medline: 31157273]
39. Wiens M. A tutorial and use case example of the Extreme Gradient Boosting (XGBoost) artificial intelligence algorithm for drug development applications. Clin Transl Sci. 2025;18(3):e70172. [doi: 10.51219/urforum.2025.jackson-burton]
40. Jinbo Z, Yufu L, Haitao M. Handling missing data of using the XGBoost-based multiple imputation by chained equations regression method. Front Artif Intell. 2025;8:1553220. [FREE Full text] [doi: 10.3389/frai.2025.1553220] [Medline: 40248006]
41. Pouya OR, Boostani R, Sabeti M. Enhancing adaboost performance in the presence of class-label noise: a comparative study on EEG-based classification of schizophrenic patients and benchmark datasets. IDA. 2024;28(1):357-376. [doi: 10.3233/ida-227125]
42. Martinez W, Gray JB. Noise peeling methods to improve boosting algorithms. Comput Stat Data Anal. 2016;93:483-497. [doi: 10.1016/j.csda.2015.06.010]
43. Hancock JT, Khoshgoftaar TM. CatBoost for big data: an interdisciplinary review. J Big Data. 2020;7(1):94. [doi: 10.21203/rs.3.rs-54646/v1]
44. Zhao Y, Zhao H. A hybrid machine learning framework by incorporating categorical boosting and manifold learning for financial analysis. Intell Syst Appl. 2025;25:200473. [doi: 10.1016/j.iswa.2024.200473]
45. Lu J, Gweon H. Random k conditional nearest neighbor for high-dimensional data. PeerJ Comput Sci. 2025;11:e2497. [doi: 10.7717/peerj-cs.2497] [Medline: 39896033]
46. Loeloe MS, Tabatabaei SM, Sefidkar R, Mehrparvar AH, Jambarsang S. Boosting K-nearest neighbor regression performance for longitudinal data through a novel learning approach. BMC Bioinformatics. 2025;26(1):232. [FREE Full text] [doi: 10.1186/s12859-025-06205-1] [Medline: 41029204]
47. Fang X, Yang N. A neural learning approach for a data-driven nonlinear error correction model. Comput Intell Neurosci. 2023;2023:5884314. [FREE Full text] [doi: 10.1155/2023/5884314] [Medline: 36726356]
48. Zivich P, Naimi AI. A primer on neural networks. Am J Epidemiol. 2025;194(6):1473-1475. [doi: 10.1093/aje/kwae380] [Medline: 39358996]
49. Sarker IH. Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. SN Comput Sci. 2021;2(6):420. [FREE Full text] [doi: 10.1007/s42979-021-00815-1] [Medline: 34426802]
50. Mathema VB, Sen P, Lamichhane S, Orešič M, Khoomrung S. Deep learning facilitates multi-data type analysis and predictive biomarker discovery in cancer precision medicine. Comput Struct Biotechnol J. 2023;21:1372-1382. [FREE Full text] [doi: 10.1016/j.csbj.2023.01.043] [Medline: 36817954]
51. Agyemang EF. A Gaussian Process Regression and Wavelet Transform time series approaches to modeling Influenza A. Comput Biol Med. 2025;184:109367. [doi: 10.1016/j.compbiomed.2024.109367] [Medline: 39549528]
52. Hotelling H. Analysis of a complex of statistical variables into principal components. J Educ Psychol. 1933;24(7):498-520. [doi: 10.1037/h0070888]
53. Greenacre M, Groenen PJF, Hastie T, D'Enza AI, Markos A, Tuzhilina E. Principal component analysis. Nat Rev Methods Primers. 2022;2(1):100. [doi: 10.1038/s43586-022-00184-w]
54. Jolliffe IT. A note on the use of principal components in regression. J R Stat Soc. Ser C (Appl Stat). 1982;31(3):300-303. [doi: 10.2307/2348005]
55. Shimizu H, Enda K, Shimizu T, Ishida Y, Ishizu H, Ise K, et al. Machine learning algorithms: prediction and feature selection for clinical refracture after surgically treated fragility fracture. J Clin Med. 2022;11(7):2021. [FREE Full text] [doi: 10.3390/jcm11072021] [Medline: 35407629]
56. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV. Catboost: unbiased boosting with categorical features. 2018. Presented at: NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems; 2018 December 3 - 8:6639-6649; Montréal Canada. URL: https://proceedings.neurips.cc/paper_files/paper/2018/file/14491b756b3a51daac41c24863285549-Paper.pdf

57. Balasubramaniam M, Mainali N, Bowroju SK, Atluri P, Penthala NR, Ayyadevera S, et al. Structural modeling of GSK3β implicates the inactive (DFG-out) conformation as the target bound by TDZD analogs. Sci Rep. 2020;10(1):18326. [FREE Full text] [doi: 10.1038/s41598-020-75020-w] [Medline: 33110096]
58. Willow SY, Xie B, Lawrence J, Eisenberg RS, Minh DDL. On the polarization of ligands by proteins. Phys Chem Chem Phys. 2020;22(21):12044-12057. [FREE Full text] [doi: 10.1039/d0cp00376j] [Medline: 32421120]
59. Goel H, Yu W, Ustach V, Aytenfisu A, Sun D, MacKerell A. Impact of electronic polarizability on protein-functional group interactions. Phys Chem Chem Phys. Apr 6, 2020;22(13):6848-6860. [doi: 10.1039/D0CP00088D]
60. Dunn WJ, Koehler MG, Grigoras S. The role of solvent-accessible surface area in determining partition coefficients. J Med Chem. 1987;30(7):1121-1126. [doi: 10.1021/jm00390a002] [Medline: 3599019]
61. Chuman H, Mori A, Tanaka H, Yamagami C, Fujita T. Analyses of the partition coefficient, log P, using ab initio MO parameter and accessible surface area of solute molecules. J Pharm Sci. 2004;93(11):2681-2697. [doi: 10.1002/jps.20168] [Medline: 15389676]
62. Zhou F, Zhang L, van Laar T, van Dam H, Ten Dijke P. GSK3β inactivation induces apoptosis of leukemia cells by repressing the function of c-Myb. Mol Biol Cell. 2011;22(18):3533-3540. [FREE Full text] [doi: 10.1091/mbc.E11-06-0483] [Medline: 21795403]
63. Góral I, Wichur T, Sługocka E, Grygier P, Głuch-Lutwin M, Mordyl B, et al. Exploring novel GSK-3β inhibitors for anti-neuroinflammatory and neuroprotective effects: synthesis, crystallography, computational analysis, and biological evaluation. ACS Chem Neurosci. Sep 04, 2024;15(17):3181-3201. [doi: 10.1021/acschemneuro.4c00365] [Medline: 39158934]
64. Bernard-Gauthier V, Mossine AV, Knight A, Patnaik D, Zhao W-N, Cheng C, et al. Structural basis for achieving GSK-3β inhibition with high potency, selectivity, and brain exposure for positron emission tomography imaging and drug discovery. J Med Chem. Nov 14, 2019;62(21):9600-9617. [FREE Full text] [doi: 10.1021/acs.jmedchem.9b01030] [Medline: 31535859]
65. Kumar V, Madan AK. Application of graph theory: prediction of glycogen synthase kinase-3 beta inhibitory activity of thiadiazolidinones as potential drugs for the treatment of Alzheimer's disease. Eur J Pharm Sci. Feb 2005;24(2-3):213-218. [doi: 10.1016/j.ejps.2004.10.013] [Medline: 15661493]
66. Shah S, Famta P, Vambhurkar G, Srinivasarao DA, Kumar KC, Bagasariya D, et al. Quality by design accredited self-nanoemulsifying delivery of ibrutinib for extenuating the fast-fed variability, ameliorating the anticancer activity and oral bioavailability in prostate cancer. J Drug Delivery Sci Technol. 2023;89:105052. [doi: 10.1016/j.jddst.2023.105052]
67. Aqil F, Munagala R, Jeyabalan J, Vadhanam MV. Bioavailability of phytochemicals and its enhancement by drug delivery systems. Cancer Lett. 2013;334(1):133-141. [FREE Full text] [doi: 10.1016/j.canlet.2013.02.032] [Medline: 23435377]
68. Liu Q, Sun H, Li X, Sheng H, Zhu L. Strategies for solubility and bioavailability enhancement and toxicity reduction of norcantharidin. Molecules. 2022;27(22):7740. [FREE Full text] [doi: 10.3390/molecules27227740] [Medline: 36431851]
69. Qian S, Zheng C, Wu Y, Huang H, Wu G, Zhang J. Targeted therapy for leukemia based on nanomaterials. Heliyon. 2024;10(15):e34951. [FREE Full text] [doi: 10.1016/j.heliyon.2024.e34951] [Medline: 39144922]
70. Zhong G, Chang X, Xie W, Zhou X. Targeted protein degradation: advances in drug discovery and clinical practice. Signal Transduct Target Ther. 2024;9(1):308. [FREE Full text] [doi: 10.1038/s41392-024-02004-x] [Medline: 39500878]
71. Arfeen M, Patel R, Khan T, Bharatam PV. Molecular dynamics simulation studies of GSK-3β ATP competitive inhibitors: understanding the factors contributing to selectivity. J Biomol Struct Dyn. 2015;33(12):2578-2593. [doi: 10.1080/07391102.2015.1063457] [Medline: 26209183]
72. Berg S, Bergh M, Hellberg S, Högdin K, Lo-Alfredsson Y, Söderman P, et al. Discovery of novel potent and highly selective glycogen synthase kinase-3β (GSK3β) inhibitors for Alzheimer's disease: design, synthesis, and characterization of pyrazines. J Med Chem. Nov 08, 2012;55(21):9107-9119. [doi: 10.1021/jm201724m] [Medline: 22489897]
73. Noh-Burgos MJ, García-Sánchez S, Tun-Rosado FJ, Chávez-González A, Peraza-Sánchez SR, Moo-Puc RE. Semi-synthesis, anti-leukemia activity, and docking study of derivatives from 3,24-dihydroxylup-20(29)-en-28-oic acid. Molecules. 2025;30(15):3193. [FREE Full text] [doi: 10.3390/molecules30153193] [Medline: 40807368]
74. Berlin CB, Sharma E, Kozlowski MC. Quantification of hydrogen-bond-donating ability of biologically relevant compounds. J Org Chem. 2024;89(7):4684-4690. [doi: 10.1021/acs.joc.3c02939] [Medline: 38483838]
75. ChEMBL. URL: https://www.ebi.ac.uk/chembl/ [accessed 2025-12-13]
76. Kyaw Zin PP, Borrel A, Fourches D. Benchmarking 2D/3D/MD-QSAR models for Imatinib derivatives: how far can we predict? J Chem Inf Model. Jul 27, 2020;60(7):3342-3360. [doi: 10.1021/acs.jcim.0c00200] [Medline: 32623886]
77. Katritzky AR, Girgis AS, Slavov S, Tala SR, Stoyanova-Slavova I. QSAR modeling, synthesis and bioassay of diverse leukemia RPMI-8226 cell line active agents. Eur J Med Chem. Nov 2010;45(11):5183-5199. [doi: 10.1016/j.ejmech.2010.08.033] [Medline: 20843586]
78. Aloui M, Er-Rajy M, Imtara H, Goudzal A, Zarougui S, El Fadili M, et al. QSAR modelling, molecular docking, molecular dynamic and ADMET prediction of pyrrolopyrimidine derivatives as novel Bruton's tyrosine kinase (BTK) inhibitors. Saudi Pharm J. Jan 2024;32(1):101911. [FREE Full text] [doi: 10.1016/j.jsps.2023.101911] [Medline: 38226346]
79. Pennington LD, Muegge I. Holistic drug design for multiparameter optimization in modern small molecule drug discovery. Bioorg Med Chem Lett. 2021;41:128003. [doi: 10.1016/j.bmcl.2021.128003] [Medline: 33798703]

80. Leo IR, Aswad L, Stahl M, Kunold E, Post F, Erkers T, et al. Integrative multi-omics and drug response profiling of childhood acute lymphoblastic leukemia cell lines. Nat Commun. Mar 30, 2022;13(1):1691. [FREE Full text] [doi: 10.1038/s41467-022-29224-5] [Medline: 35354797]
81. Horton TM, Sposto R, Brown P, Reynolds CP, Hunger SP, Winick NJ, et al. ALLNA 2008 Conference. Toxicity assessment of molecularly targeted drugs incorporated into multiagent chemotherapy regimens for pediatric acute lymphocytic leukemia (ALL): review from an international consensus conference. Pediatr Blood Cancer. Jul 01, 2010;54(7):872-878. [FREE Full text] [doi: 10.1002/pbc.22414] [Medline: 20127846]
82. Kumar A, Purohit R. Use of long term molecular dynamics simulation in predicting cancer associated SNPs. PLoS Comput Biol. 2014;10(4):e1003318. [FREE Full text] [doi: 10.1371/journal.pcbi.1003318] [Medline: 24722014]
83. König C, Vellido A. Understanding predictions of drug profiles using explainable machine learning models. BioData Mining. Aug 01, 2024;17(1):25. [doi: 10.1186/s13040-024-00378-w]
84. Jaganathan K, Tayara H, Chong KT. An explainable supervised machine learning model for predicting respiratory toxicity of chemicals using optimal molecular descriptors. Pharmaceutics. Apr 11, 2022;14(4):832. [FREE Full text] [doi: 10.3390/pharmaceutics14040832] [Medline: 35456666]
85. Noviandy TR, Idroes GM, Harnelly E, Sari I. Predicting AXL tyrosine kinase inhibitor potency using machine learning with interpretable insights for cancer drug discovery. Heca J Appl Sci. 2025;3:17-29. [doi: 10.60084/hjas.v3i1.270]

Abbreviations

Adjusted R^2: adjusted coefficient of determination
AI: artificial intelligence
CCC: concordance correlation coefficient
DFG: amino acids aspartate, phenylalanine, and glycine
GSK3β: glycogen synthase kinase 3β
IC50: half maximal inhibitory concentration
LSC: leukemia stem cell
MAE: mean absolute error
MAPE: mean absolute percentage error
MD: molecular dynamics
MedAE: median absolute error
ML: machine learning
MSE: mean squared error
NMSE: normalized mean squared error
NRMSE: normalized root-mean-squared error
PLS: partial least squares
QSAR: quantitative structure-activity relationship
R^2: coefficient of determination (explained variance)
RMSE: root-mean-squared error
SHAP: Shapley additive explanations
SMAPE: symmetric mean absolute percentage error
SVR: support vector regression

Edited by G Luo; submitted 30.Jul.2025; peer-reviewed by F Xiong, M Wason, F Anupama; comments to author 16.Sep.2025; revised version received 10.Nov.2025; accepted 05.Dec.2025; published 27.Jan.2026

Please cite as:
Kakraba S, Agyemang EF, Shmookler Reis RJ
Accelerating Discovery of Leukemia Inhibitors Using AI-Driven Quantitative Structure-Activity Relationship: Algorithm Development and Validation
JMIR AI 2026;5:e81552
URL: https://ai.jmir.org/2026/1/e81552
doi: 10.2196/81552
PMID: 41358925