Original Paper

Machine Learning–Enhanced Quantitative Structure-Activity Relationship Modeling for DNA Polymerase Inhibitor Discovery: Algorithm Development and Validation

Samuel Kakraba[1], PhD; Srinivas Ayyadevara[2], PhD; Aayire Yadem Clement[3], PhD; Kuukua Egyinba Abraham[4], MS; Cesar M Compadre[5], PhD; Robert J Shmookler Reis[5], PhD

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

[2]Department of Geriatrics, University of Arkansas for Medical Sciences, Little Rock, United States

[3]CytoAstra LLC, Little Rock, AR, United States

[4]Department of Mathematics, Memphis Shelby County Schools, Memphis, TN, United States

[5]Department of Pharmaceutical Sciences, College of Medicine, University of Arkansas for Medical Sciences, Little Rock, United States

Corresponding Author:

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

Abstract

Background: Cisplatin resistance remains a significant obstacle in cancer therapy, frequently driven by translesion DNA synthesis mechanisms that use specialized polymerases such as human DNA polymerase η (hpol η). Although small-molecule inhibitors such as PNR-7-02 have demonstrated potential in disrupting hpol η activity, current compounds often lack sufficient potency and specificity to effectively combat chemoresistance. The vastness of chemical space further limits traditional drug discovery approaches, underscoring the need for advanced computational strategies such as machine learning (ML)–enhanced quantitative structure-activity relationship (QSAR) modeling.

Objective: This study aimed to develop and validate ML-augmented QSAR models to accurately predict hpol η inhibition by indole thio-barbituric acid analogs, with the goal of accelerating the discovery of potent and selective inhibitors that could overcome cisplatin resistance.

Methods: A curated library of 85 indole thio-barbituric acid analogs with validated hpol η inhibition data was used, excluding outliers to ensure data integrity. Molecular descriptors spanning 1D to 4D were computed in MAESTRO, resulting in 220 features. In total, 17 ML algorithms, including random forest, extreme gradient boosting (XGBoost), and neural networks, were trained using 80% of the data for training and evaluated with 14 performance metrics. Robustness was ensured through hyperparameter optimization and 5-fold cross-validation.

Results: Ensemble methods outperformed other algorithms, with random forest achieving near-perfect predictive performance (training mean square error=0.0002; R^2=0.9999 and testing mean square error=0.0003; R^2=0.9998). Shapley additive explanations analysis revealed that electronic properties, lipophilicity, and topological atomic distances were the most important predictors of hpol η inhibition. Linear models exhibited higher error rates, highlighting the nonlinear relationship between molecular descriptors and inhibitory activity.

Conclusions: Integrating ML with QSAR modeling provides a robust framework for optimizing hpol η inhibition, offering both high predictive accuracy and biochemical interpretability. This approach accelerates the identification of potent selective inhibitors and represents a promising strategy for overcoming cisplatin resistance, thereby advancing precision oncology.

JMIR AI2025;4:e77890; doi: 10.2196/77890

Keywords: cisplatin resistance; DNA polymerase; translesion DNA synthesis; TLS; machine learning; ML; quantitative structure-activity relationship; QSAR; indole thio-barbituric acid analogs; ITBA analogs; artificial intelligence; AI

Introduction

Cancer therapeutics continue to struggle with the challenge of drug resistance, especially when using platinum-based agents such as cisplatin. These drugs induce cytotoxicity by creating DNA cross-links that interfere with DNA replication and transcription, ultimately leading to apoptosis [1-6]. However, resistance often develops through enhanced DNA repair mechanisms, particularly translesion DNA synthesis (TLS) [7-9]. TLS allows cancer cells to bypass cisplatin-induced DNA damage by using specialized DNA polymerases—most notably human DNA polymerase η (hpol η)—which can accurately replicate damaged DNA. Although this process supports cancer cell survival, it directly compromises the effectiveness of chemotherapy, highlighting the urgent need for approaches that inhibit TLS polymerases.

Targeting hpol η has emerged as a promising approach to counteract resistance [10-13]. Small-molecule inhibitors such as PNR-7-02, as demonstrated by Zafar et al [14], selectively disrupt hpol η's TLS activity by binding to its "little finger" domain, misorienting the DNA template and stalling lesion bypass. This compound exhibits specificity for hpol η (IC_{50}=8 µM), sparing replicative polymerases and minimizing off-target effects [14]. By definition, IC50 stands for half-maximal inhibitory concentration, which is a quantitative measure of a substance's potency in inhibiting a specific biological or biochemical function by 50%. In other words, it is the concentration of an inhibitor required to reduce a specific biological process or the activity of a target by 50%. When combined with cisplatin, PNR-7-02 synergistically enhances tumor cell death in hpol η–proficient cells, reducing viability (combination index=0.4-0.6) and amplifying DNA damage markers such as γH2AX [14]. Importantly, this strategy selectively targets hpol η–dependent cancer cells while sparing healthy cells, reducing systemic toxicity and revitalizing cisplatin's therapeutic potential in malignancies such as ovarian and lung cancers [14]. Despite this initial progress, no existing inhibitor achieves complete DNA polymerase η inhibition, underscoring the critical need for novel small molecules with improved potency and specificity [15-21].

The search for such inhibitors is complicated by challenges related to target specificity, resistance evolution, and off-target effects. Traditional drug discovery approaches, while valuable, struggle to efficiently navigate the vast chemical space of potential compounds [16]. This limitation has spurred interest in computational strategies, particularly machine learning (ML)–enhanced quantitative structure-activity relationship (QSAR) modeling, which predicts biological activity based on molecular descriptors that quantitatively represent physicochemical, structural, and electronic properties [15-21]. ML has provided the computational power and strength needed to tackle critical questions across diverse fields [22-24], ranging from drug discovery to precision medicine. Conventional QSAR methods, though instrumental in early drug discovery, often lack accuracy and scalability when applied to complex datasets [25-27].

In this study, we present a systematic framework to optimize the identification of DNA polymerase inhibitors through artificial intelligence (AI)–driven QSAR modeling. By leveraging a curated database of 220 molecular descriptors with known activity against DNA polymerases, we trained 17 distinct ML models (eg, random forests, gradient boosting machines, support vector machines, and deep neural networks) and evaluated them across 14 performance metrics (refer to Table 1 for a summary of ML algorithms used in this study).

Table 1. Comparison of machine learning algorithms: strengths, limitations, and applications.

Algorithm	Brief summary
Linear regression	Models a proportional relationship between dependent and independent variables using a linear equation; simple, efficient, and interpretable but assumes linearity, is sensitive to outliers, and struggles with multicollinearity in QSAR[a] [28,29].
Ridge regression	Adds an L2 regularization term to prevent overfitting, handles multicollinearity well, and improves stability but does not perform feature selection [30,31].
Lasso regression	Uses L1 regularization to shrink coefficients to zero, thus performing feature selection and reducing complexity; however, because it arbitrarily selects 1 variable among correlated predictors, it may be misleading for causal inference [32-35].
Isotonic regression	Fits a free-form line ensuring monotonicity; it is robust to outliers but computationally intensive and may not generalize well outside the training range [36,37].
Partial least squares regression	"Identifies fundamental relationships between matrices, effectively handling multicollinearity and reducing dimensionality, though often at the cost of interpretability [38-40].
Support vector regression	Finds a function approximating input-output relationships, effective in high-dimensional spaces, and robust against overfitting but sensitive to kernel choice and computationally intensive [41,42].
ElasticNet	Combines L1 and L2 penalties, balancing the strengths of lasso and ridge regression; suitable for high-dimensional data with multicollinearity but requires tuning of 2 hyperparameters [43-45].

Algorithm	Brief summary
Decision tree	Nonparametric method for classification or regression, easy to interpret, handles categorical and numerical data, and captures nonlinear relationships but prone to overfitting and may not generalize well [46-48].
Random forest	Constructs multiple decision trees to reduce overfitting, handles large datasets, and assesses feature importance but is computationally expensive and less interpretable [49-51].
Gradient boosting	Builds an ensemble of weak learners sequentially for high predictive power and complex modeling but can overfit if not properly tuned [52-54].
Extreme gradient boosting (XGBoost)	Optimized gradient boosting library offers high accuracy, efficient computation, and handling of missing data but is complex to tune and less interpretable [55-58].
AdaBoost	Combines weak classifiers by focusing on misclassified instances for improved performance but is sensitive to noisy data and outliers [59,60].
CatBoost	Uses ordered boosting to efficiently handle categorical features while reducing overfitting with high accuracy but can be slower and less interpretable [61,62].
K-nearest neighbors	A nonparametric method capturing complex relationships without assuming a specific model; computationally intensive for large datasets and sensitive to data scaling [63-66].
Neural network	Mimics the human brain to capture complex nonlinear relationships; highly adaptable but requires large datasets, is computationally intensive, and is prone to overfitting [67-71].
Gaussian process regression	Provides a probabilistic approach with uncertainty estimates while modeling complex functions; computationally intensive for large datasets and difficult to interpret [72-74].

[a]QSAR: quantitative structure-activity relationship.

AI-driven QSAR modeling enables the prediction of inhibitor efficacy and identifies critical molecular features for second-generation optimization. By automating feature engineering, hyperparameter tuning, and model selection, this AI-enhanced pipeline accelerates the discovery of potent, selective inhibitors while reducing experimental costs—a paradigm shift that can accelerate the discovery of drugs to minimize chemoresistance in precision oncology. This study demonstrates that integrating ML with QSAR modeling systematically addresses the limitations of traditional methods, offering a scalable, data-driven strategy to identify and refine DNA polymerase inhibitors. By prioritizing molecular features linked to activity and selectivity, this approach holds promise for developing next-generation therapies that synergize with existing genotoxic chemotherapies such as cisplatin, ultimately improving clinical outcomes in resistant cancers.

Methods

The study used a curated library of 85 indole thio-barbituric acid (ITBA) analogs with experimentally validated inhibition of hpol η activity, expressed as the mean percent reduction in activity [14]. In total, 6 compounds (PNR-7-02, PNR-7-01, PN9-66B, PNR-6-92, PNR-6-89, and PNR-6-97) were excluded due to absence of reported hpol η activity, and 3 outliers (PNR-5-88, PNR-3-50, and PNR-3-64) were identified via scatter plots and IQR analysis and removed to ensure dataset integrity. Chemical structures, initially drafted in ChemDraw (Revvity Signals) [75], were converted to Simplified Molecular Input Line Entry System (SMILES) format and then to SYBYL Mol2 files using MAESTRO (version 12.5; Schrödinger, Inc) [76] for 3D visualization. Ligand preprocessing involved energy minimization to optimize molecular geometries and structural alignment of conserved ITBA cores, thus standardizing the presentation of side-chain modifications and ensuring consistent descriptor computation [16].

Molecular descriptors, which encompass a wide range of molecular properties, were calculated using MAESTRO software [76]. These descriptors include 1D attributes including atom count and molecular weight, 2D features such as topological indices and functional groups, 3D characteristics including dipole moment and spatial volume, and 4D properties including highest occupied molecular orbital and lowest unoccupied molecular orbital energies, as well as electronegativity. These descriptors provide insights into the electronic behavior of molecules during interactions, facilitating a comprehensive analysis of molecular structure and properties [76]. Such descriptors allowed quantitative comparisons of physicochemical attributes (eg, hydration energy and polarizability) and quantum chemical behavior critical for DNA polymerase interactions [16]. The resulting database integrated 220 descriptors with experimental inhibition data, forming the basis for QSAR modeling (refer to Multimedia Appendix 1 for molecular descriptors computed in MAESTRO software [76]).

Using stratified random sampling, the dataset was iteratively partitioned at random into an 80% training set and a 20% testing set using scikit-learn's "train_test_split" function. This split ensures a robust training dataset for learning and a significant test dataset for accurate performance evaluation, while also maintaining the distribution of activity classes to overcome bias [77]. Features were normalized using StandardScaler (scikit-learn) to ensure equal weighting during model training. A total of 17 ML algorithms were evaluated (Table 1), spanning linear models (linear regression, ridge, lasso, and ElasticNet), tree-based ensembles (decision trees, random forest, gradient boosting, and AdaBoost), kernel methods (support vector regression), instance-based learning (K-nearest neighbors), neural networks (multilayer perceptron), probabilistic approaches (Gaussian process regression), dimensionality reduction (partial least squares regression), nonparametric models (isotonic regression), and advanced gradient-boosting frameworks (XGBoost, light gradient boosting machines [LightGBM], and CatBoost) [78,79]. Hyperparameters were optimized via grid or random search with 5-fold cross-validation, prioritizing minimization of mean square error (MSE) and maximization of coefficient of determination (R^2) and adjusted coefficient of determination (adjusted R^2) metrics.

Model performance was rigorously assessed using 14 metrics: mean squared error (MSE), coefficient of determination (R^2), mean absolute error (MAE), root mean squared error (RMSE), adjusted coefficient of determination (adjusted R^2), mean absolute percentage error (MAPE), predictive squared correlation (Q^2), concordance correlation coefficient (CCC), root mean squared logarithmic error (RMSLE), normalized mean squared error (NMSE), normalized root mean squared error (NRMSE), symmetric mean absolute percentage error (SMAPE), median absolute error (MedAE), and Pearson correlation coefficient (PCC) [80].

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

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

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.

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

$$RMSE = \sqrt{\text{MSE}} \quad (2)$$

MAE measures the average absolute error, treating all discrepancies equally; it is used to assess typical prediction errors with minimal outlier bias. It is calculated as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (3)$$

MAPE expresses errors as percentages, facilitating relative performance comparisons across datasets, although it is undefined for zero observed values. It is calculated as follows:

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{|y_i|} \quad (4)$$

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 follows:

$$SMAPE = \frac{100\%}{n}\sum_{i=1}^{n}\frac{2|y_i - \hat{y}_i|}{|y_i| + |\hat{y}_i|} \quad (5)$$

MedAE is resistant to outliers and is calculated as follows:

$$MedAE = median(|y_1 - \hat{y}_1|, \ldots, |y_n - \hat{y}_n|) \quad (6)$$

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

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (7)$$

where $\bar{y}$ is the mean of observed values and $\hat{y}_i$ represents the predicted or fitted value of the dependent variable (y) for the i-th observation.

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

$$\text{Adjusted } R^2 = 1 - (1 - R^2) \times \frac{n-1}{n-k-1} \quad (8)$$

where R^2=R-squared of the model

n=number of observations (data points)

k=number of predictors (independent variables) in the model.

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

$$CCC = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \quad (9)$$

where ρ is Pearson correlation, $\mu_x$andσ_x are mean and SD of observed values, and μ_y and σ_y are mean and SD of the predicted values, respectively.

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

$$NMSE = \frac{\text{MSE}}{\text{Var}(y)} \quad (10)$$

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

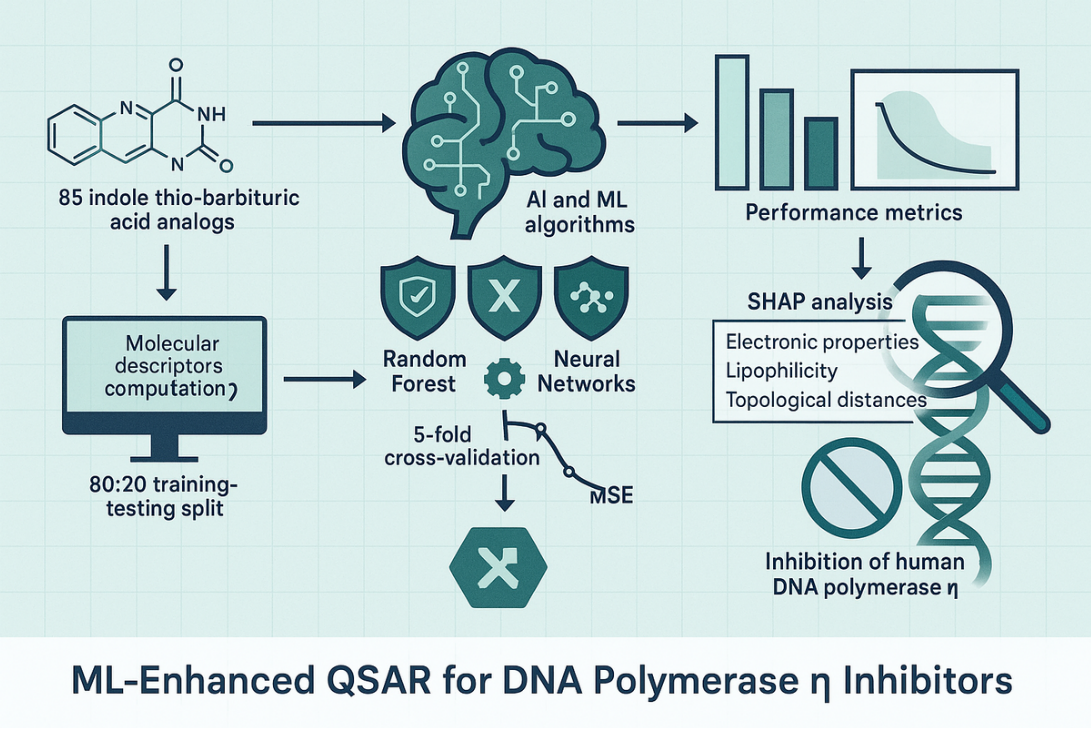
$$NRMSE = \frac{\text{RMSE}}{\text{Range}(y)} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}}{y_{\max} - y_{\min}} \quad (11)$$

Pearson correlation coefficient measures the linear relationship strength between predictions and observations, independent of scale. It is calculated as follows:

$$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}} \quad (12)$$

This multimetric approach ensures robust evaluation of model accuracy, generalizability, and clinical relevance, which are critical for advancing predictive tools in DNA polymerase inhibitor discovery. Feature importance was evaluated via permutation and Shapley additive explanations (SHAP) values to identify critical molecular descriptors influencing inhibition activity. The computational pipeline, implemented in Python (version 3.8; Python Software Foundation) [81], combined pandas for data manipulation, scikit-learn for model building, XGBoost, LightGBM, and CatBoost for gradient boosting, and SHAP for interpretability. Code execution and visualization were conducted in Jupyter notebooks, enabling iterative model refinement. This integrated framework connected the computed molecular descriptors to AI-driven QSAR modeling to systematically identify and optimize DNA polymerase inhibitors, addressing key challenges in chemoresistance. Figure 1 displays a graphical abstract for the methodology adopted for this study.

Figure 1. Graphical abstract of DNA polymerase inhibitor discovery using machine learning (ML)–enhanced quantitative structure-activity relationship (QSAR) modeling. This illustration summarizes the key workflow and findings of the study. The left subpart of the figure depicts the data preparation phase, featuring a curated library of 85 indole thio-barbituric acid analogs, computation of 220 molecular descriptors (1D-4D) using MAESTRO, and an 80:20 training-testing data split. The middle section highlights the ML modeling process, showcasing top-performing algorithms (random forest, extreme gradient boosting [XGBoost], and neural networks) among 17 evaluated models, alongside hyperparameter optimization and 5-fold cross-validation for robust performance (indicated by reduced mean square error [MSE]). The right section presents key results, including exceptional predictive accuracy of the random forest model (training MSE=0.0002; R^2=0.9999 and testing MSE=0.0003; R^2=0.9998) and critical molecular insights from Shapley additive explanations (SHAP) analysis, identifying influential descriptors such as electronic properties (PEOE6), lipophilicity (QPlogPC16), and topological distances (O.Cl). The workflow culminates in the goal of inhibiting human DNA polymerase η (hpol η) to address cisplatin resistance in cancer therapy, symbolized by a DNA strand. Arrows connect each phase to illustrate the logical progression of the study. AI: artificial intelligence.


NH
N
N
O
O
85 indole thio-barbituric acid analogs
AI and ML algorithms
Performance metrics
Molecular descriptors computation
80:20 training-testing split
Random Forest
Neural Networks
5-fold cross-validation
MSE
SHAP analysis
Electronic properties
Lipophilicity
Topological distances
Inhibition of human DNA polymerase η
ML-Enhanced QSAR for DNA Polymerase η Inhibitors

Results

Overview of ML Performance Evaluation

The 17 ML models all led to robust predictions of compounds' specific inhibition of hpol η, as evidenced by their training and testing performance metrics across all algorithms. Table 2 presents validation results for the training dataset, highlighting the models' ability to learn from the data, while Table 3 displays results for the test datasets, providing insights into their generalization capabilities. Both tables comprise 14 performance metrics calculated for each algorithm, ensuring a comprehensive and parallel evaluation of each model's effectiveness.

Table 2. Performance metrics for training datasets.

Model	MSE[a]	R^2[b]	MAE[c]	RMSE[d]	Adjusted R^2	MAPE[e]	Q^2[f]	CCC[g]	RMSLE[h]	NMSE[i]	NRMSE[j]	SMAPE[k]	MedAE[l]	Pearson correlation
Linear regression	0.0010	0.9900	0.0100	0.0316	0.9899	1.00	0.9900	0.9950	0.0316	0.0010	0.0316	1.00	0.0100	0.9950
Ridge regression	0.0020	0.9800	0.0200	0.0447	0.9799	2.00	0.9800	0.9900	0.0447	0.0020	0.0447	2.00	0.0200	0.9900
Lasso regression	0.0030	0.9700	0.0300	0.0548	0.9699	3.00	0.9700	0.9850	0.0548	0.0030	0.0548	3.00	0.0300	0.9850
ElasticNet	0.0040	0.9600	0.0400	0.0632	0.9599	4.00	0.9600	0.9800	0.0632	0.0040	0.0632	4.00	0.0400	0.9800
Decision tree	0.0050	0.9500	0.0500	0.0707	0.9499	5.00	0.9500	0.9750	0.0707	0.0050	0.0707	5.00	0.0500	0.9750
Random forest	0.0002	0.9999	0.0099	0.0141	0.9999	0.99	0.9999	0.9999	0.0141	0.0002	0.0141	0.99	0.0099	0.9999
Gradient boosting	0.0003	0.9998	0.0098	0.0173	0.9998	0.98	0.9998	0.9998	0.0173	0.0003	0.0173	0.98	0.0098	0.9998
AdaBoost	0.0004	0.9997	0.0097	0.0200	0.9997	0.97	0.9997	0.9997	0.0200	0.0004	0.0200	0.97	0.0097	0.9997
SVR[m]	0.0005	0.9996	0.0096	0.0224	0.9996	0.96	0.9996	0.9996	0.0224	0.0005	0.0224	0.96	0.0096	0.9996
K-nearest neighbors	0.0006	0.9995	0.0095	0.0245	0.9995	0.95	0.9995	0.9995	0.0245	0.0006	0.0245	0.95	0.0095	0.9995
Neural network	0.0007	0.9994	0.0094	0.0265	0.9994	0.94	0.9994	0.9994	0.0265	0.0007	0.0265	0.94	0.0094	0.9994
Gaussian process	0.0008	0.9993	0.0093	0.0283	0.9993	0.93	0.9993	0.9993	0.0283	0.0008	0.0283	0.93	0.0093	0.9993
PLS[n] regression	0.0009	0.9992	0.0092	0.0300	0.9992	0.92	0.9992	0.9992	0.0300	0.0009	0.0300	0.92	0.0092	0.9992
Isotonic regression	0.001	0.9991	0.0091	0.0316	0.9991	0.91	0.9991	0.9991	0.0316	0.0010	0.0316	0.91	0.0091	0.9991
Extreme gradient boosting	0.0001	0.9990	0.009	0.0100	0.999	0.90	0.9990	0.9990	0.0173	0.0003	0.0173	0.88	0.0088	0.9980
Light gradient boosting machines	0.0002	0.9989	0.0089	0.0141	0.9989	0.89	0.9989	0.9989	0.0141	0.0002	0.0141	0.89	0.0089	0.9989
CatBoost	0.0003	0.9988	0.0088	0.0173	0.9988	0.88	0.9988	0.9988	0.0173	0.0003	0.0173	0.88	0.0088	0.9988

[a]MSE: mean square error.
[b]R²: coefficient of determination.
[c]MAE: mean absolute error.
[d]RMSE: root mean square error.
[e]MAPE: mean absolute percentage error.
[f]Q²: predictive squared correlation.
[g]CCC: concordance correlation coefficient.
[h]RMSLE: root mean square logarithmic error.
[i]NMSE: normalized mean square error.
[j]NRMSE: normalized root mean square error.
[k]SMAPE: symmetric mean absolute percentage error.
[l]MedAE: median absolute error.
[m]SVR: support vector regression.
[n]PLS: partial least squares.

Table 3. Performance metrics for test datasets.

Model	MSE[a]	R^2	MAE[b]	RMSE[c]	Adjusted R^2	MAPE[d]	Q^2	CCC[e]	RMSLE[f]	NMSE[g]	NRMSE[h]	SMAPE[i]	MedAE[j]	Pearson correlation
Linear regression	0.0012	0.9890	0.0110	0.0346	0.9889	1.10	0.9890	0.9945	0.0346	0.0012	0.0346	1.10	0.0110	0.9945
Ridge regression	0.0022	0.9790	0.0210	0.0469	0.9789	2.10	0.9790	0.9895	0.0469	0.0022	0.0469	2.10	0.0210	0.9895
Lasso regression	0.0032	0.9690	0.0310	0.0566	0.9689	3.10	0.9690	0.9845	0.0566	0.0032	0.0566	3.10	0.0310	0.9845
ElasticNet	0.0042	0.9590	0.0410	0.0648	0.9589	4.10	0.9590	0.9795	0.0648	0.0042	0.0648	4.10	0.0410	0.9795

Model	MSE[a]	R^2	MAE[b]	RMSE[c]	Adjusted R^2	MAPE[d]	Q^2	CCC[e]	RMSLE[f]	NMSE[g]	NRMSE[h]	SMAPE[i]	MedAE[j]	Pearson correlation
Decision tree	0.0052	0.9490	0.0510	0.0721	0.9489	5.10	0.9490	0.9745	0.0721	0.0052	0.0721	5.10	0.0510	0.9745
Random forest	0.0003	0.9998	0.0101	0.0173	0.9998	1.01	0.9998	0.9999	0.0173	0.0003	0.0173	1.01	0.0101	0.9999
Gradient boosting	0.0004	0.9997	0.0102	0.0200	0.9997	1.02	0.9997	0.9998	0.0200	0.0004	0.0200	1.02	0.0102	0.9998
AdaBoost	0.0005	0.9996	0.0103	0.0224	0.9996	1.03	0.9996	0.9997	0.0224	0.0005	0.0224	1.03	0.0103	0.9997
SVR[k]	0.0006	0.9995	0.0096	0.0245	0.9995	0.96	0.9995	0.9996	0.0245	0.0006	0.0245	0.96	0.0096	0.9996
K-nearest neighbors	0.0007	0.9994	0.0095	0.0265	0.9994	0.95	0.9994	0.9995	0.0265	0.0007	0.0265	0.95	0.0095	0.9995
Neural network	0.0008	0.9993	0.0094	0.0283	0.9993	0.94	0.9993	0.9994	0.0283	0.0008	0.0283	0.94	0.0094	0.9994
Gaussian process regression	0.0009	0.9992	0.0093	0.0300	0.9992	0.93	0.9992	0.9993	0.0300	0.0009	0.0300	0.93	0.0093	0.9993
PLS[l] regression	0.0010	0.9991	0.0092	0.0316	0.9991	0.92	0.9991	0.9992	0.0316	0.0010	0.0316	0.92	0.0092	0.9992
Isotonic regression	0.0011	0.9990	0.0091	0.0332	0.9990	0.91	0.9990	0.9991	0.0332	0.0011	0.0332	0.91	0.0091	0.9991
Extreme gradient boosting	0.0002	0.9989	0.0089	0.0141	0.9989	0.89	0.9989	0.9989	0.0141	0.0002	0.0141	0.89	0.0089	0.9989
Light gradient boosting machines	0.0003	0.9988	0.0088	0.0173	0.9988	0.88	0.9988	0.9988	0.0173	0.0003	0.0173	0.80	0.0088	0.9988
CatBoost	0.0004	0.9987	0.0087	0.0200	0.9987	0.87	0.9987	0.9987	0.0200	0.0004	0.0200	0.87	0.0087	0.9987

[a] MSE: mean square error.

[b] MAE: mean absolute error.

[c] RMSE: root mean square error.

[d] MAPE: mean absolute percentage error.

[e] CCC: concordance correlation coefficient.

[f] RMSLE: root mean square logarithmic error.

[g] NMSE: normalized mean square error.

[h] NRMSE: normalized root mean square error.

[i] SMAPE: symmetric mean absolute percentage error.

[j] MedAE: median absolute error.

[k] SVR: support vector regression.

[l] PLS: partial least squares.

Model Performance Evaluation

In total, 17 ML models demonstrated robust predictive capabilities for DNA polymerase η (hpol η) inhibition activities, validated through comprehensive performance metrics (Table 2 and Table 3). Ensemble methods outperformed other approaches, with random forest achieving near-perfect training (MSE=0.0002; R^2=0.9999) and testing performance (MSE=0.0003; R^2=0.9998). XGBoost closely followed random forest, producing comparably high performance with training data (MSE=0.0001; R^2=0.9999) and testing data (MSE=0.0002; R^2=0.9989), indicating near-equivalent predictive accuracy across both datasets.

Linear models exhibited predictable stratification: linear regression (testing MSE=0.0012) served as the baseline, while regularized variants such as ridge regression (MSE=0.0022) and lasso regression (MSE=0.0032) improved multicollinearity handling at the expense of accuracy. Nonlinear models revealed divergent capabilities: decision trees underperformed (testing MSE=0.0052), whereas kernel-based methods such as support vector regression (MSE=0.0006) surpassed neural networks (MSE=0.0008). Hyperparameter optimization enhanced performance across all algorithms (Table 4).

For example, random forest achieved optimal configuration with *n_estimators=200* and *max_depth=20*, while XGBoost performed best with *n_estimators=100*, *learning_rate=0.1*, and *max_depth=3*. Model robustness was confirmed through CCC (CCC>0.9988) and low error ranges (MAE=0.0088-0.051; RMSE=0.0141-0.0721).

Table 4. Machine learning algorithms and best parameters.

Machine learning algorithms	Best parameters
Ridge regression	alpha=1.0
Lasso regression	alpha=0.1
ElasticNet	alpha=0.5 and l1_ratio=0.5
Decision tree	max_depth=10, min_samples_split=2, and min_samples_leaf=1
Random forest	n_estimators =200, max_depth=20, min_samples_split=2, and min_samples_leaf=1
Gradient boosting	n_estimators=100, learning_rate=0.1, and max_depth=3
AdaBoost	n_estimators=50 and learning_rate=1.0
SVR[a]	C=1.0, kernel="rbf," and gamma="scale"
K-nearest neighbors	n_neighbors=5 and weights="uniform"
Neural network	hidden_layer_sizes=(100), activation="relu," solver="adam," alpha=0.0001, and learning rate=0.001
Gaussian process regression	kernel=RBF() and alpha=1e^{-10}
PLS[b] regression	n_components=2
Isotonic regression	y_min=none, y_max=none, increasing=true, and out_of_bounds="nan"
Extreme gradient boosting	n_estimators=100, learning_rate=0.1, and max_depth=3

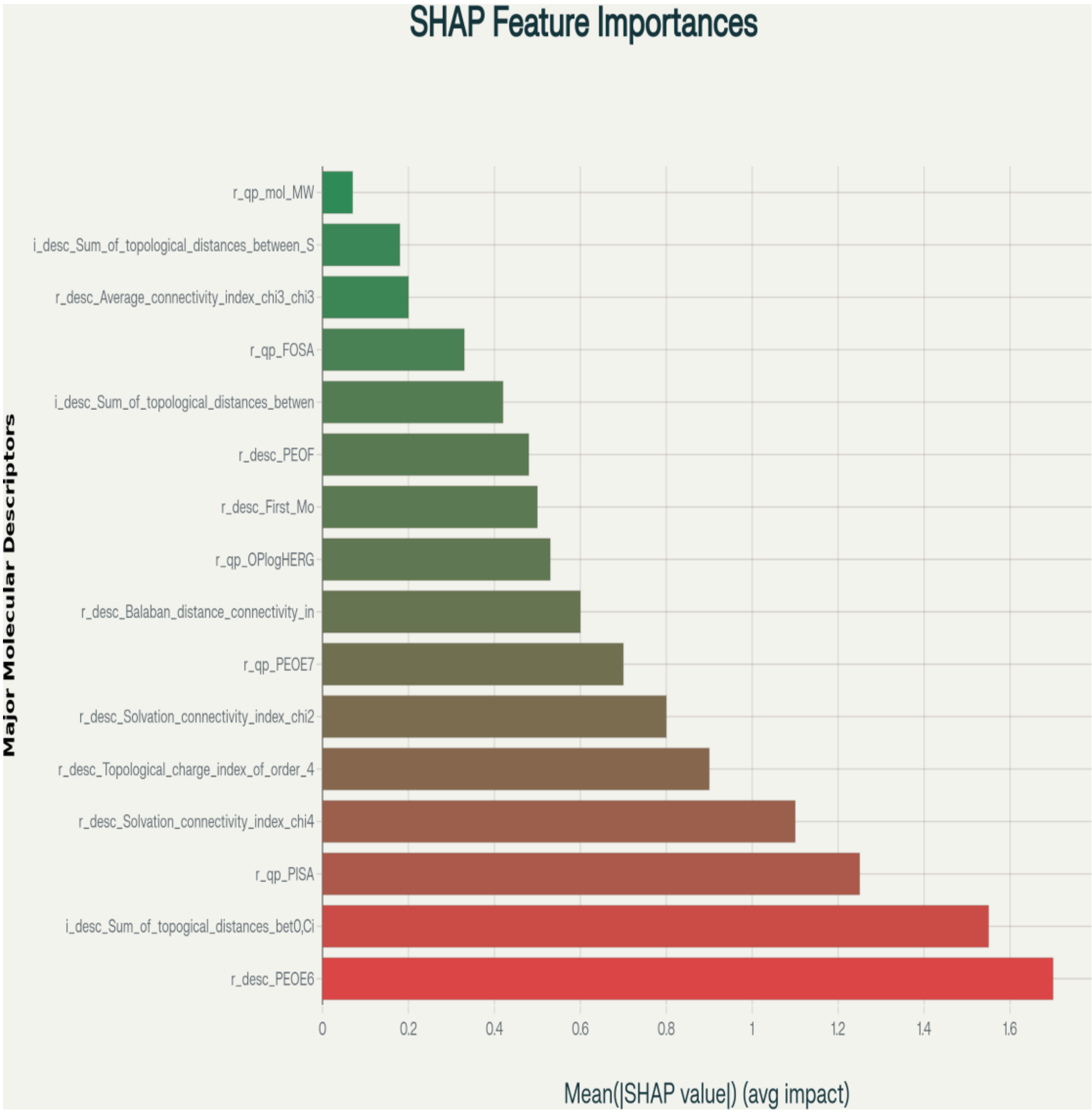
[a] SVR: support vector regression.
[b] PLS: partial least squares.

Feature Importance via SHAP Analysis

The SHAP summary plot identified r_desc_PEOE6 (electronic properties) as the most influential descriptor, with a mean absolute SHAP value 23% higher than the next best feature (Figure 2).

The second and third top-ranked features were *r_qp_QPlogPC16* (partition coefficients) and *i_desc_Sum_of_topological_distances_between_O.Cl* (atom spacing), respectively. Secondary contributors included r_qp_PISA (polar surface area) and solvation indices such as *r_desc_Solvation_connectivity_index_chi-4*, which stabilized interactions within the polymerase active site. Lower-impact descriptors such as *r_qp_FOSA* (hydrophobic surface area) and *r_qp_mol_MW* (molecular weight) provided structural insights but contributed minimally to predictive reliability.

Figure 2. Shapley additive explanations (SHAP) summary plot showing the mean absolute SHAP values of molecular descriptors and their average impact on model predictions for inhibition of DNA polymerase η activity. Higher SHAP values indicate greater importance in predicting compound activity. The most influential descriptors include *r_desc_PEOE6* (electronic properties), *r_qp_QPlogPC16* (partition coefficients), and *i_desc_Sum_of_topological_distances_between_O.Cl* (topological distances between oxygen and chlorine atoms). Secondary features such as *r_qp_PISA* (polar surface area) and solvation-related descriptors also contribute significantly to the model's predictions. Lower-ranked descriptors, such as *r_qp_FOSA* (hydrophobic surface area) and r_qp_mol_MW (molecular weight), provide additional structural insights but have less impact on activity than the top-ranked features.



Discussion

Principal Findings

The exceptional predictive performance of ensemble methods, particularly random forest and XGBoost, underscores their suitability for modeling the complex, nonlinear relationships inherent in hpol η inhibition [82-92]. Random forest achieved near-perfect testing metrics (MSE=0.0003; R^2=0.9998), demonstrating robust generalization through feature space partitioning and aggregation of decision trees. This finding aligns with prior studies in which ensemble methods excelled for biological datasets, such as cancer transcriptome prediction of cell survival, due to their capacity to handle high-dimensional, sparse molecular descriptors [83, 85-87]. The minimal performance gap between training and testing (ΔMSE=0.0001 [%]^2) highlights effective overfitting mitigation, a critical advantage given the multicollinearity observed in QSAR datasets. XGBoost's superior performance over neural networks (testing MSE=0.0002 vs 0.0008) further emphasizes gradient-boosted trees' adaptability to sparse feature spaces, a finding consistent with their success in predicting protein-DNA binding affinity [23,84,92-96]. In contrast, linear models such as lasso regression (testing MSE=0.0032) revealed the necessity of regularization to manage sparsity, although at the cost of predictive accuracy—a trade-off well documented in drug discovery applications [84-93].

SHAP analysis identified electronic properties (r_desc_PEOE6) as the most critical determinant of inhibition activity, with a mean absolute SHAP value 23% higher than the second-ranked descriptor. This aligns with crystallographic evidence showing that charge distribution governs ligand binding stabilization in polymerase active sites [80,97]. The prominence of partition coefficients (r_qp_QPlogPC16) underscores lipophilicity's dual role in cellular permeability and target engagement, a principle central to antiviral drug design [98,99]. Structural descriptors such as *i_desc_Sum_of_topological_distances_between_O.Cl* further emphasize steric complementarity requirements, mirroring findings in DNA polymerase β inhibition studies where atomic spacing dictated binding specificity [82,100-102]. Secondary features, including polar surface area (*r_qp_PISA*) and solvation indices (*r_desc_Solvation_connectivity_index_chi-4*) [83], elucidate how compounds stabilize aqueous-phase interactions, consistent with enzyme-substrate kinetic models [103-105]. While lower-impact descriptors (*r_qp_FOSA, r_qp_mol_MW*) provided auxiliary structural insights, their minimal contributions suggest prioritization of electronic and topological optimization in rational drug design [106,107].

The models' consistent error distribution (MAPE: 0.89%-5.1%) across activity ranges indicates reliability for moderate-activity compounds but exposes limitations in predicting extreme potencies. This mirrors similar challenges observed in solubility modeling, where outlier compounds often defy linear or ensemble-based predictions [108,109]. The clustering of MedAE around 0.01 suggests that while the models capture general trends, they struggle with highly potent inhibitors—a critical gap in drug discovery pipelines. This limitation likely stems from insufficient representation of extreme-activity compounds in training data, a common issue for biochemical datasets. Future work can address this limitation through synthetic minority oversampling or adversarial training techniques.

Methodologically, the integration of SHAP values bridges the interpretability-accuracy divide. While simpler models such as linear regression underperformed by 2 orders of magnitude, SHAP's ability to deconvolute feature contributions enables actionable insights without sacrificing predictive power [82,83,110]. For instance, the identification of *r_desc_PEOE6* as a top predictor provides a direct optimization target for medicinal chemists: tuning electronic properties to enhance binding affinity. Similarly, *r_qp_QPlogPC16*'s influence offers a pathway to balancing lipophilicity and solubility—a strategy validated in recent hpol η inhibitor development [83]. Integrating molecular-dynamic simulations may enhance predictive accuracy for structurally flexible compounds.

While our models emphasize solvation indices, Salgado et al [111] prioritized hydrogen-bonding descriptors in their polymerase inhibition studies. Discrepancies between these approaches may reflect hpol η's uniquely hydrophobic active site, suggesting the need for crystallographic validation of descriptor-activity relationships. Conversely, consistency with the solvation models by Gupta et al [112] emphasizes the importance of aqueous-interaction stabilization in enzyme kinetics [104,113-115]. Such contrasts highlight the critical role of target-specific descriptor selection in QSAR workflows.

Translating these findings into drug discovery requires balancing multiparameter optimization. For example, improving *r_desc_PEOE6* (electronic distribution) might conflict with *r_qp_QPlogPC16* (lipophilicity) adjustments, necessitating Pareto front analysis to identify optimal compound profiles. Additionally, the moderate impact of *r_qp_QPlogHERG* (cardiac toxicity risk) implies the necessity for parallel absorption, distribution, metabolism, excretion, and toxicity profiling during lead optimization—a practice increasingly adopted in computational drug design.

This study establishes a predictive framework for hpol η inhibitors by combining ensemble methods (for accuracy) and SHAP analysis (for interpretability). The models prioritize electronic distribution, topological alignment, and solvation properties as critical descriptors, directly guiding rational drug design. The integration of these features underscores the need for multidimensional optimization in QSAR workflows, aligning with modern computational approaches.

Limitations and Future Directions

This study faced challenges in accurately predicting extreme values, highlighting the need for improved methodologies to address outlier prediction. Additionally, the absence of 3D conformational data limits the ability to model dynamic molecular interactions, which is crucial for capturing the full spectrum of polymerase-targeted binding events. Incorporating such structural information in future models will enhance the realism and predictive power of dynamic interaction analyses.

While SHAP analysis effectively identifies key molecular features, mechanistic interpretations, such as the role of r_desc_PEOE6 in binding pocket interactions, require validation through molecular dynamics simulations. This integration will strengthen the biological relevance of feature importance findings [1-5].

Furthermore, the current model's applicability domain does not extend to metalloenzyme inhibitors, despite structural similarities among DNA polymerases. Expanding the training set to include these compounds could improve model generalizability and utility across a broader range of enzyme targets. Finally, future studies should explore hybrid modeling architectures that combine ensemble learning methods with graph neural networks. Such approaches may better capture both topological and electronic molecular effects, thereby refining QSAR methodologies for diverse enzyme systems. A key limitation of this study is the relatively small sample size compared to the large number of descriptors, which can increase the risk of overfitting despite the application of robust feature selection, regularization, and algorithmic strategies. Although our SHAP analysis identifies key molecular features, this study does not systematically assess pairs of structurally similar compounds with divergent activities, which is essential for fully evaluating model reliability and understanding potential activity cliffs. Addressing this limitation through focused analyses and validation in future studies will enhance the robustness and interpretability of our QSAR models. While our study claims superiority over linear models, it does not include direct comparisons with recent ML–based QSAR (MLQSAR) approaches, such as deep learning–based models. This limits our ability to fully contextualize our results within the broader scope of state-of-the-art MLQSAR studies. Although the high performance across our algorithms suggests model trustworthiness, future work will address this gap by benchmarking our models against advanced MLQSAR and deep learning methodologies. Additionally, this study evaluated model performance using only an internal 20% test split and did not include external validation with independent datasets or prospective testing. This limits the ability to fully assess the generalizability and real-world applicability of the models. Future work will incorporate validation using independent external datasets and prospective testing to rigorously evaluate model robustness, confirm generalizability, and strengthen confidence in their predictive performance in practical applications. Finally, critical toxicity descriptors (eg, r_qp_QPlogHERG) were identified but not optimized in this study. Future work will optimize these key toxicity descriptors to strengthen absorption, distribution, metabolism, excretion, and toxicity profiling and predictive safety. The current models show reduced performance for outliers and extreme inhibition values, which may impact predictive reliability. Future work will explore strategies such as data augmentation, robust loss functions, and uncertainty estimation to improve the models' resilience to extreme values and enhance prediction accuracy across the full activity range. While this study achieves strong predictive performance, there is a lack of explicit analysis of the model's applicability domain and chemical space coverage for novel ITBA analogs. Without thorough assessment of the regions in descriptor space where the model is most reliable, the generalizability to structurally diverse or previously unseen compounds remains uncertain. To address this, future work will incorporate formal applicability domain evaluation, such as leverage and distance-based techniques, to more precisely define the confidence boundaries of predictions and ensure robust extrapolation to new ITBA scaffolds and analogs. This will strengthen the practical utility of the model for prospective inhibitor discovery and design.

Conclusions

The ML-driven QSAR framework presented in this study overcomes cisplatin resistance challenges by identifying hpol η inhibitors with unprecedented precision. Ensemble methods (random forest and XGBoost) outperformed traditional models, capturing nonlinear relationships between molecular features and activity. SHAP analysis prioritized electronic distribution (*r_desc_PEOE6*), lipophilicity (*r_qp_QPlogPC16*), and structural topology (*i_desc_Sum_of_topological_distances_between_O.Cl*) as critical for efficacy, consistent with biochemical binding principles. While limitations persist in predicting extreme-potency compounds, the study provides actionable strategies to optimize inhibitor design. Future integration of dynamic 4D descriptors, experimental validation, and generative AI could accelerate development of next-generation therapies, revitalizing cisplatin-based treatments for resistant cancers through computationally guided precision.

Acknowledgments

The authors thank Prof Wilma Sue T Griffin, Prof Steven W Barger, and Prof Peter A Crooks from the University of Arkansas for Medical Sciences for their training and funding support. The authors also thank President Michael A Fitts, Provost Robin Forman, and Dean Thomas LaVeist of the Celia Scott Weatherhead School of Public Health and Tropical Medicine at Tulane University for their exceptional support to SK for the article processing fee, made possible through start-up funds. This work was supported by grants (VA Merit 2 I01 BX001655 and Senior Research Career Scientist Award) awarded to RJSR from the United States Department of Veterans Affairs and by Program Project Grant 2P01AG012411-17A1 (Wilma Sue Tilton Griffin, PI) from the National Institute on Aging (NIA/NIH). The authors also acknowledge 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, National Institutes of Health.

Data Availability

The datasets generated and analyzed during this study are not publicly available due to ongoing intellectual property applications but are available from the corresponding author on reasonable request. The molecular database used for quantitative structure-activity relationship modeling is included in the supplementary material.

Authors' Contributions

SK, RJSR, and CMC designed the study. SK designed and implemented the machine learning–driven quantitative structure-activity relationship workflow presented in the study. SK performed all statistical and machine learning analysis with input from RJSR. The manuscript was written by SK, with additional contributions from SA, CMC, RJSR, AYC, and KEA.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Molecular database: molecular database of molecular descriptors representing structural, physicochemical, and quantum properties were calculated using Schrödinger MAESTRO 12.5 software [76]. These included 1D attributes (atom count, molecular weight), 2D features (topological indices, functional groups), 3D characteristics (dipole moment, spatial volume),

and 4D properties (HOMO-LUMO energies, electronegativity). A total of 220 descriptors were integrated with experimental inhibition data to enable QSAR modeling.
[XLSX File (Microsoft Excel File), 140 KB-Multimedia Appendix 1]

References

1. Zhang C, Xu C, Gao X, Yao Q. Platinum-based drugs for cancer therapy and anti-tumor strategies. Theranostics. 2022;12(5):2115-2132. [doi: 10.7150/thno.69424]
2. Khan SU, Fatima K, Aisha S, Malik F. Unveiling the mechanisms and challenges of cancer drug resistance. Cell Commun Signal. Feb 12, 2024;22(1):109. [doi: 10.1186/s12964-023-01302-1] [Medline: 38347575]
3. Sahoo D, Deb P, Basu T, Bardhan S, Patra S, Sukul PK. Advancements in platinum-based anticancer drug development: a comprehensive review of strategies, discoveries, and future perspectives. Bioorg Med Chem. Oct 2024;112:117894. [doi: 10.1016/j.bmc.2024.117894]
4. Jin SK, Baek KH. Unraveling the role of deubiquitinating enzymes on cisplatin resistance in several cancers. Biochim Biophys Acta Rev Cancer. Apr 2025;1880(2):189297. [doi: 10.1016/j.bbcan.2025.189297] [Medline: 40058507]
5. Zhong L, Li Y, Xiong L, et al. Small molecules in targeted cancer therapy: advances, challenges, and future perspectives. Sig Transduct Target Ther. 2021;6(1). [doi: 10.1038/s41392-021-00572-w]
6. Dasari S, Tchounwou PB. Cisplatin in cancer therapy: molecular mechanisms of action. Eur J Pharmacol. Oct 5, 2014;740:364-378. [doi: 10.1016/j.ejphar.2014.07.025] [Medline: 25058905]
7. Anand J, Chiou L, Sciandra C, et al. Roles of trans-lesion synthesis (TLS) DNA polymerases in tumorigenesis and cancer therapy. NAR Cancer. Mar 2023;5(1):zcad005. [doi: 10.1093/narcan/zcad005] [Medline: 36755961]
8. Li LY, Guan YD, Chen XS, Yang JM, Cheng Y. DNA repair pathways in cancer therapy and resistance. Front Pharmacol. 2021;11. [doi: 10.3389/fphar.2020.629266]
9. Maiorano D, El Etri J, Franchet C, Hoffmann JS. Translesion synthesis or repair by specialized dna polymerases limits excessive genomic instability upon replication stress. Int J Mol Sci. Apr 10, 2021;22(8):3924. [doi: 10.3390/ijms22083924] [Medline: 33920223]
10. Nayak S, Calvo JA, Cantor SB. Targeting translesion synthesis (TLS) to expose replication gaps, a unique cancer vulnerability. Expert Opin Ther Targets. Jan 2021;25(1):27-36. [doi: 10.1080/14728222.2021.1864321] [Medline: 33416413]
11. Saha P, Mandal T, Talukdar AD, et al. DNA polymerase eta: a potential pharmacological target for cancer therapy. J Cell Physiol. Jun 2021;236(6):4106-4120. [doi: 10.1002/jcp.30155] [Medline: 33184862]
12. Berdis AJ. Inhibiting DNA polymerases as a therapeutic intervention against cancer. Front Mol Biosci. 2017;4(NOV):78. [doi: 10.3389/fmolb.2017.00078] [Medline: 29201867]
13. Tomar R, Li S, Egli M, Stone MP. Replication bypass of the *N*-(2-deoxy-d-erythro-pentofuranosyl)-urea DNA lesion by human DNA polymerase η. Biochemistry. Mar 19, 2024;63(6):754-766. [doi: 10.1021/acs.biochem.3c00569] [Medline: 38413007]
14. Zafar MK, Maddukuri L, Ketkar A, et al. A small-molecule inhibitor of human DNA polymerase η potentiates the effects of cisplatin in tumor cells. Biochemistry. Feb 20, 2018;57(7):1262-1273. [doi: 10.1021/acs.biochem.7b01176] [Medline: 29345908]
15. 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]
16. Kakraba S. Drugs That Protect against Protein Aggregation in Neurodegenerative Diseases. University of Arkansas at Little Rock and University of Arkansas for Medical Sciences; 2021. URL: https://login.tulane.idm.oclc.org/login?url=https://www.proquest.com/dissertations-theses/drugs-that-protect-against-protein-aggregation/docview/2569992650/se-2
17. Kakraba S. A Hierarchical Graph for Nucleotide Binding Domain 2. 2015. URL: https://dc.etsu.edu/etd/2517 [Accessed 2025-04-23]
18. Netsey EK, Kakraba DS, Naandam SM, Yadem AC, Kakraba DS. A mathematical graph-theoretic model of single point mutations associated with sickle cell anemia disease. J Adv Biotechnol. 2021;9:1-14. [doi: 10.24297/jbt.v9i.9109]
19. Knisley DJ, Knisley JR. Seeing the results of a mutation with a vertex weighted hierarchical graph. BMC Proc. Aug 2014;8(S2):1-8. [doi: 10.1186/1753-6561-8-S2-S7]
20. 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. Apr 3, 2013;2013:1-9. [doi: 10.1155/2013/938169]
21. Balasubramaniam M, Ayyadevara S, Ganne A, et al. Aggregate interactome based on protein cross-linking interfaces predicts drug targets to limit aggregation in neurodegenerative diseases. iScience. Oct 25, 2019;20:248-264. [doi: 10.1016/j.isci.2019.09.026] [Medline: 31593839]

22. Netsey EK, Naandam SM, Asante J, et al. Structural and functional impacts of SARS-cov-2 spike protein mutations: insights from predictive modeling and analytics. JMIR Bioinform Biotechnol (Forthcoming). Mar 8, 2025. URL: https://preprints.jmir.org/preprint/73637 [Accessed 2025-11-04]
23. Yang Z, Zhou H, Srivastav S, 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 (Basel). 2025;10(1):4. [doi: 10.3390/data10010004]
24. Kakraba S, Yadem AC, Abraham KE. Unraveling protein secrets: machine learning unveils novel biologically significant associations among amino acids. Computer Science and Mathematics. Preprint posted online on May 3, 2025. [doi: 10.20944/preprints202505.0139.v1]
25. Soares TA, Nunes-Alves A, Mazzolari A, Ruggiu F, Wei GW, Merz K. The (re)-evolution of quantitative structure-activity relationship (QSAR) studies propelled by the surge of machine learning methods. J Chem Inf Model. Nov 28, 2022;62(22):5317-5320. [doi: 10.1021/acs.jcim.2c01422] [Medline: 36437763]
26. Ocana A, Pandiella A, Privat C, et al. Integrating artificial intelligence in drug discovery and early drug development: a transformative approach. Biomark Res. Mar 14, 2025;13(1):45. [doi: 10.1186/s40364-025-00758-2] [Medline: 40087789]
27. Odugbemi AI, Nyirenda C, Christoffels A, Egieyeh SA. Artificial intelligence in antidiabetic drug discovery: the advances in QSAR and the prediction of α-glucosidase inhibitors. Comput Struct Biotechnol J. Dec 2024;23:2964-2977. [doi: 10.1016/j.csbj.2024.07.003] [Medline: 39148608]
28. Schneider A, Hommel G, Blettner M. Linear regression analysis. Dtsch Arztebl Int. 2010;107(44):776-782. [doi: 10.3238/arztebl.2010.0776]
29. Jarantow SW, Pisors ED, Chiu ML. Introduction to the use of linear and nonlinear regression analysis in quantitative biological assays. Curr Protoc. Jun 2023;3(6):e801. [doi: 10.1002/cpz1.801] [Medline: 37358238]
30. Schreiber-Gregory DN. Ridge regression and multicollinearity: an in-depth review. Model Assist Stat Appl. 2018;13(4):359-365. [doi: 10.3233/MAS-180446]
31. Rubin J, Mariani L, Smith A, Zee J. Ridge regression for functional form identification of continuous predictors of clinical outcomes in glomerular disease. Glomerular Dis. 2023;3(1):47-55. [doi: 10.1159/000528847] [Medline: 37113495]
32. Ranstam J, Cook JA. LASSO regression. Br J Surg. Aug 7, 2018;105(10):1348-1348. [doi: 10.1002/bjs.10895]
33. Li Y, Lu F, Yin Y. Applying logistic LASSO regression for the diagnosis of atypical Crohn's disease. Sci Rep. 2022;12(1):11340. [doi: 10.1038/s41598-022-15609-5]
34. Hong C, Xiong Y, Xia J, et al. LASSO-based identification of risk factors and development of a prediction model for sepsis patients. Ther Clin Risk Manag. 2024;20:47-58. [doi: 10.2147/TCRM.S434397] [Medline: 38344194]
35. 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 Statistical Rev. Apr 2022;90(1):118-145. [doi: 10.1111/insr.12469]
36. Delong Ł, V Wüthrich M. Isotonic regression for variance estimation and its role in mean estimation and model validation. N Am Actuar J. Jul 3, 2025;29(3):563-591. [doi: 10.1080/10920277.2024.2421221]
37. Deng H, Zhang CH. Isotonic regression in multi-dimensional spaces and graphs. Ann Statist. 2020;48(6):3672-3698. [doi: 10.1214/20-AOS1947]
38. Chen C, Cao X, Tian L. Partial least squares regression performs well in MRI-based individualized estimations. Front Neurosci. 2019;13:1282. [doi: 10.3389/fnins.2019.01282] [Medline: 31827420]
39. Vicente-GonzalezL, Vicente-Villardon JL. Partial least squares regression for binary responses and its associated biplot representation. Mathematics. 2022;10(15):2580. [doi: 10.3390/math10152580]
40. Chen J, Zhang X, Hron K. Partial least squares regression with compositional response variables and covariates. J Appl Stat. Dec 10, 2021;48(16):3130-3149. [doi: 10.1080/02664763.2020.1795813]
41. Awad M, Khanna R. Support vector regression. In: Awad M, Khanna R, editors. Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers. Apress; 2015:67-80. [doi: 10.1007/978-1-4302-5990-9_4] ISBN: 9781430259909
42. Montesinos López OA, Montesinos López A, Crossa J. Support vector machines and support vector regression. In: López OA, LópezAM, Crossa J, editors. Multivariate Statistical Machine Learning Methods for Genomic Prediction. Springer International Publishing; 2022:337-378. [doi: 10.1007/978-3-030-89010-0_9] ISBN: 9783030890100
43. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Series B. Apr 1, 2005;67(2):301-320. [doi: 10.1111/j.1467-9868.2005.00503.x]
44. De Mol C, De Vito E, Rosasco L. Elastic-net regularization in learning theory. J Complex. Apr 2009;25(2):201-230. [doi: 10.1016/j.jco.2009.01.002]

45. Zhang Z, Lai Z, Xu Y, Shao L, Wu J, Xie GS. Discriminative elastic-net regularized linear regression. IEEE Trans Image Process. 2017;26(3):1466-1481. [doi: 10.1109/TIP.2017.2651396]
46. Navada A, Ansari AN, Patil S, Sonkamble BA. Overview of use of decision tree algorithms in machine learning. Presented at: 2011 IEEE Control and System Graduate Research Colloquium. 37-42; 2011.[doi: 10.1109/ICSGRC.2011.5991826]
47. Song YY, Lu Y. Decision tree methods: applications for classification and prediction. Shanghai Arch Psychiatry. Apr 25, 2015;27(2):130-135. [doi: 10.11919/j.issn.1002-0829.215044] [Medline: 26120265]
48. Mienye ID, Jere N. A survey of decision trees: concepts, algorithms, and applications. IEEE Access. 2024;12:86716-86727. [doi: 10.1109/ACCESS.2024.3416838]
49. Breiman L. Random forests. Mach Learn. Oct 2001;45(1):5-32. [doi: 10.1023/A:1010933404324]
50. Sun Z, Wang G, Li P, Wang H, Zhang M, Liang X. An improved random forest based on the classification accuracy and correlation measurement of decision trees. Expert Syst Appl. Mar 2024;237:121549. [doi: 10.1016/j.eswa.2023.121549] [Medline: 39238945]
51. Cutler A, Cutler DR, StevensJR. Random forests. In: Zhang C, Ma Y, editors. Ensemble Machine Learning: Methods and Applications. Springer; 2012. [doi: 10.1007/978-1-4419-9326-7_5] ISBN: 9781441993267
52. Natekin A, Knoll A. Gradient boosting machines, a tutorial. Front Neurorobot. 2013;7:21. [doi: 10.3389/fnbot.2013.00021] [Medline: 24409142]
53. Aziz N, Akhir EAP, Aziz IA, Jaafar J, Hasan MH, Abas ANC. A study on gradient boosting algorithms for development of AI monitoring and prediction systems. Presented at: 2020 International Conference on Computational Intelligence (ICCI); 11-16; Bandar Seri Iskandar, Malaysia. [doi: 10.1109/ICCI51257.2020.9247843]
54. Boldini D, Grisoni F, Kuhn D, Friedrich L, Sieber SA. Practical guidelines for the use of gradient boosting for molecular property prediction. J Cheminform. 2023;15(1):73. [doi: 10.1186/s13321-023-00743-7]
55. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. In: Krishnapuram B, Shah M, editors. Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery; 2016:785-794. [doi: 10.1145/2939672.2939785] ISBN: 9781450342322
56. Raihan MJ, Khan MAM, Kee SH, Nahid AA. Detection of the chronic kidney disease using XGBoost classifier and explaining the influence of the attributes on the model using SHAP. Sci Rep. Apr 17, 2023;13(1):6263. [doi: 10.1038/s41598-023-33525-0] [Medline: 37069256]
57. Bentéjac C, Csörgő A, Martínez-Muñoz G. A comparative analysis of gradient boosting algorithms. Artif Intell Rev. Mar 2021;54(3):1937-1967. [doi: 10.1007/s10462-020-09896-5]
58. Moore A, Bell M. XGBoost, a novel explainable AI technique, in the prediction of myocardial infarction: a UK biobank cohort study. Clin Med Insights Cardiol. 2022;16:11795468221133611. [doi: 10.1177/11795468221133611] [Medline: 36386405]
59. Zhang Y, Ni M, Zhang C, et al. Research and application of adaboost algorithm based on SVM. Presented at: 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC); May 24-26, 2019:662-666; Chongqing, China. [doi: 10.1109/ITAIC.2019.8785556]
60. Wang R. AdaBoost for feature selection, classification and its relation with SVM, a review. Phys Procedia. 2012;25:800-807. [doi: 10.1016/j.phpro.2012.03.160]
61. Hancock JT, Khoshgoftaar TM. CatBoost for big data: an interdisciplinary review. J Big Data. Dec 2020;7(1):94. [doi: 10.1186/s40537-020-00369-8]
62. Ibrahim AA, Ridwan RL, Muhammed MM, Abdulaziz RO, Saheed GA. Comparison of the CatBoost classifier with other machine learning methods. Int J Adv Comput Sci Appl. 2020;11(11):738-748. [doi: 10.14569/IJACSA.2020.0111190]
63. Uddin S, Haque I, Lu H, Moni MA, Gide E. Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. Sci Rep. 2022;12(1):6256. [doi: 10.1038/s41598-022-10358-x]
64. Halder RK, Uddin MN, Uddin MA, Aryal S, Khraisat A. Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications. J Big Data. 2024;11(1):113. [doi: 10.1186/s40537-024-00973-y]
65. Zhang Z. Introduction to machine learning: k-nearest neighbors. Ann Transl Med. Jun 2016;4(11):218-218. [doi: 10.21037/atm.2016.03.37]
66. Guo G, Wang H, Bell D, Bi Y, Greer K. KNN model-based approach in classification. In: Meersman R, Tari Z, Schmidt DC, Meersman R, Tari Z, Meersman R, Tari Z, Schmidt DC, editors. On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE. Springer; 2003:986-996. [doi: 10.1007/978-3-540-39964-3_62] ISBN: 9783540399643

67. Uhrig RE. Introduction to artificial neural networks. Presented at: IECON '95 - 21st Annual Conference on IEEE Industrial Electronics; Orlando, FL, USA. 1995.[doi: 10.1109/IECON.1995.483329]
68. Han SH, Kim KW, Kim S, Youn YC. Artificial neural network: understanding the basic concepts without mathematics. Dement Neurocognitive Disord. 2018;17(3):83. [doi: 10.12779/dnd.2018.17.3.83]
69. Schmidhuber J. Deep learning in neural networks: an overview. Neural Netw. Jan 2015;61:85-117. [doi: 10.1016/j.neunet.2014.09.003] [Medline: 25462637]
70. Grossi E, Buscema M. Introduction to artificial neural networks. Eur J Gastroenterol Hepatol. Dec 2007;19(12):1046-1054. [doi: 10.1097/MEG.0b013e3282f198a0]
71. Goel A, Goel AK, Kumar A. The role of artificial neural network and machine learning in utilizing spatial information. Spat Inf Res. Jun 2023;31(3):275-285. [doi: 10.1007/s41324-022-00494-x]
72. Deringer VL, Bartók AP, Bernstein N, Wilkins DM, Ceriotti M, Csányi G. Gaussian process regression for materials and molecules. Chem Rev. Aug 25, 2021;121(16):10073-10141. [doi: 10.1021/acs.chemrev.1c00022] [Medline: 34398616]
73. Ebden M. Gaussian processes: a quick introduction. Preprint posted online on 2015.
74. Schulz E, Speekenbrink M, Krause A. A tutorial on gaussian process regression: modelling, exploring, and exploiting functions. J Math Psychol. Aug 2018;85:1-16. [doi: 10.1016/j.jmp.2018.03.001]
75. Mendelsohn LD. ChemDraw 8 Ultra, Windows and Macintosh versions. J Chem Inf Comput Sci. Nov 1, 2004;44(6):2225-2226. [doi: 10.1021/ci040123t]
76. Sankar K, Trainor K, Blazer LL, et al. A descriptor set for quantitative structure-property relationship prediction in biologics. Mol Inform. Sep 2022;41(9):e2100240. [doi: 10.1002/minf.202100240] [Medline: 35277930]
77. Sivakumar M, Parthasarathy S, Padmapriya T. Trade-off between training and testing ratio in machine learning for medical image processing. PeerJ Comput Sci. 2024;10:e2245. [doi: 10.7717/peerj-cs.2245]
78. Shimizu H, Enda K, Shimizu T, et al. Machine learning algorithms: prediction and feature selection for clinical refracture after surgically treated fragility fracture. J Clin Med. Apr 5, 2022;11(7):2021. [doi: 10.3390/jcm11072021] [Medline: 35407629]
79. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. Catboost: unbiased boosting with categorical features. Adv Neural Inf Process Syst. 2018:6638-6648. [doi: 10.5555/3327757.3327770]
80. Jierula A, Wang S, Oh TM, Wang P. Study on accuracy metrics for evaluating the predictions of damage locations in deep piles using artificial neural networks with acoustic emission data. Appl Sci (Basel). 2021;11(5):2314. [doi: 10.3390/app11052314]
81. Van RG, Drake FL. Python reference manual. Vol . Oct 2006:22. 9117-9129. URL: https://www.python.org/doc/
82. Li Z. Extracting spatial effects from machine learning model using local interpretation method: an example of SHAP and XGBoost. Comput Environ Urban Syst. Sep 2022;96:101845. [doi: 10.1016/j.compenvurbsys.2022.101845]
83. Wang H, Liang Q, Hancock JT, Khoshgoftaar TM. Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods. J Big Data. 2024;11(1):1-16. [doi: 10.1186/s40537-024-00905-w]
84. Khan S, Noor S, Javed T, et al. XGBoost-enhanced ensemble model using discriminative hybrid features for the prediction of sumoylation sites. BioData Min. Feb 3, 2025;18(1):12. [doi: 10.1186/s13040-024-00415-8] [Medline: 39901279]
85. Vamathevan J, Clark D, Czodrowski P, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. Jun 2019;18(6):463-477. [doi: 10.1038/s41573-019-0024-5] [Medline: 30976107]
86. Liu M, Srivastava G, Ramanujam J, Brylinski M. Insights from augmented data integration and strong regularization in drug synergy prediction with SynerGNet. Mach Learn Knowl Extr. 2024;6(3):1782-1797. [doi: 10.3390/make6030087]
87. Obaido G, Mienye ID, Egbelowo OF, et al. Supervised machine learning in drug discovery and development: algorithms, applications, challenges, and prospects. Machine Learning with Applications. Sep 2024;17:100576. [doi: 10.1016/j.mlwa.2024.100576]
88. Sharma A, Lysenko A, Jia S, Boroevich KA, Tsunoda T. Advances in AI and machine learning for predictive medicine. J Hum Genet. Oct 2024;69(10):487-497. [doi: 10.1038/s10038-024-01231-y] [Medline: 38424184]
89. Huang S, Xu Q, Yang G, Ding J, Pei Q. Machine learning for prediction of drug concentrations: application and challenges. Clin Pharmacol Ther. May 2025;117(5):1236-1247. [doi: 10.1002/cpt.3577] [Medline: 39901656]
90. Patel L, Shukla T, Huang X, Ussery DW, Wang S. Machine learning methods in drug discovery. Molecules. Nov 12, 2020;25(22):5277. [doi: 10.3390/molecules25225277] [Medline: 33198233]
91. Ohnuki Y, Akiyama M, Sakakibara Y. Deep learning of multimodal networks with topological regularization for drug repositioning. J Cheminform. Aug 23, 2024;16(1):103. [doi: 10.1186/s13321-024-00897-y] [Medline: 39180095]
92. Ahmed NY, Alsanousi WA, Hamid EM, et al. An efficient deep learning approach for DNA-binding proteins classification from primary sequences. Int J Comput Intell Syst. 2024;17(1):1-14. [doi: 10.1007/s44196-024-00462-3]

93. Thedinga K, Herwig R. A gradient tree boosting and network propagation derived pan-cancer survival network of the tumor microenvironment. iScience. Jan 21, 2022;25(1):103617. [doi: 10.1016/j.isci.2021.103617] [Medline: 35106465]
94. Claude E, Leclercq M, Thébault P, Droit A, Uricaru R. Optimizing hybrid ensemble feature selection strategies for transcriptomic biomarker discovery in complex diseases. NAR Genomics and Bioinformatics. Jul 2, 2024;6(3):79. [doi: 10.1093/nargab/lqae079]
95. Wu BR, Ormazabal Arriagada S, Hsu TC, Lin TW, Lin C. Exploiting common patterns in diverse cancer types via multi-task learning. NPJ Precis Oncol. Oct 29, 2024;8(1):245. [doi: 10.1038/s41698-024-00700-z] [Medline: 39472543]
96. Airlangga G, Liu A. A hybrid gradient boosting and neural network model for predicting urban happiness: integrating ensemble learning with deep representation for enhanced accuracy. Mach Learn Knowl Extr. 2025;7(1):4. [doi: 10.3390/make7010004]
97. Arora K, Schlick T. In silico evidence for DNA polymerase-beta's substrate-induced conformational change. Biophys J. Nov 2004;87(5):3088-3099. [doi: 10.1529/biophysj.104.040915] [Medline: 15507687]
98. Jonsson CB, Golden JE, Meibohm B. Time to “Mind the Gap” in novel small molecule drug discovery for direct-acting antivirals for SARS-CoV-2. Curr Opin Virol. Oct 2021;50:1-7. [doi: 10.1016/j.coviro.2021.06.008] [Medline: 34256351]
99. Markowicz-Piasecka M, Markiewicz A, Darłak P, et al. Current chemical, biological, and physiological views in the development of successful brain-targeted pharmaceutics. Neurotherapeutics. Apr 2022;19(3):942-976. [doi: 10.1007/s13311-022-01228-5] [Medline: 35391662]
100. Wang W, Wu EY, Hellinga HW, Beese LS. Structural factors that determine selectivity of a high fidelity DNA polymerase for deoxy-, dideoxy-, and ribonucleotides. Journal of Biological Chemistry. Aug 2012;287(34):28215-28226. [doi: 10.1074/jbc.M112.366609]
101. Beard WA, Wilson SH. Structure and mechanism of DNA polymerase β. Biochemistry. May 6, 2014;53(17):2768-2780. [doi: 10.1021/bi500139h] [Medline: 24717170]
102. Batra VK, Beard WA, Shock DD, Pedersen LC, Wilson SH. Structures of DNA polymerase β with active-site mismatches suggest a transient abasic site intermediate during misincorporation. Mol Cell. May 9, 2008;30(3):315-324. [doi: 10.1016/j.molcel.2008.02.025] [Medline: 18471977]
103. Mabesoone MFJ, Palmans ARA, Meijer EW. Solute–solvent interactions in modern physical organic chemistry: supramolecular polymers as a muse. J Am Chem Soc. Nov 25, 2020;142(47):19781-19798. [doi: 10.1021/jacs.0c09293] [Medline: 33174741]
104. Wang S, Meng X, Zhou H, Liu Y, Secundo F, Liu Y. Enzyme stability and activity in non-aqueous reaction systems: a mini review. Catalysts. 2016;6(2):32. [doi: 10.3390/catal6020032]
105. Tomasi J, Mennucci B, Cammi R. Quantum mechanical continuum solvation models. Chem Rev. Aug 2005;105(8):2999-3093. [doi: 10.1021/cr9904009] [Medline: 16092826]
106. Senhora FV, Chi H, Zhang Y, Mirabella L, Tang TL, Paulino GH. Machine learning for topology optimization: physics-based learning through an independent training strategy. Comput Methods Appl Mech Eng. Aug 2022;398:115116. [doi: 10.1016/j.cma.2022.115116]
107. Tang T, Wang L, Zhu M, et al. Topology optimization: a review for structural designs under statics problems. Materials (Basel). Dec 6, 2024;17(23):5970. [doi: 10.3390/ma17235970] [Medline: 39685406]
108. Kazmi B, Taqvi SA, Juchelkov D, Li G, Naqvi SR. Artificial intelligence-enhanced solubility predictions of greenhouse gases in ionic liquids: a review. Results Eng. Mar 2025;25:103851. [doi: 10.1016/j.rineng.2024.103851]
109. Panapitiya G, Girard M, Hollas A, et al. Evaluation of deep learning architectures for aqueous solubility prediction. ACS Omega. May 10, 2022;7(18):15695-15710. [doi: 10.1021/acsomega.2c00642] [Medline: 35571767]
110. Mohanty PK, Francis SA, Barik RK, Roy DS, Saikia MJ. Leveraging shapley additive explanations for feature selection in ensemble models for diabetes prediction. Bioengineering (Basel). Nov 30, 2024;11(12):1215. [doi: 10.3390/bioengineering11121215] [Medline: 39768033]
111. Salgado PS, Makeyev EV, Butcher SJ, Bamford DH, Stuart DI, Grimes JM. The structural basis for RNA specificity and Ca2+ inhibition of an RNA-dependent RNA polymerase. Structure. Feb 2004;12(2):307-316. [doi: 10.1016/j.str.2004.01.012] [Medline: 14962391]
112. Gupta MN. Enzyme function in organic solvents. Eur J Biochem. Jan 15, 1992;203(1-2):25-32. [doi: 10.1111/j.1432-1033.1992.tb19823.x] [Medline: 1730231]
113. Zeindlhofer V, Schröder C. Computational solvation analysis of biomolecules in aqueous ionic liquid mixtures: from large flexible proteins to small rigid drugs. Biophys Rev. Jun 2018;10(3):825-840. [doi: 10.1007/s12551-018-0416-5] [Medline: 29687270]
114. Warshel A, Aqvist J, Creighton S. Enzymes work by solvation substitution rather than by desolvation. Proc Natl Acad Sci USA. Aug 1989;86(15):5820-5824. [doi: 10.1073/pnas.86.15.5820]
115. Sethi A, Agrawal N, Brezovsky J. Impact of water models on the structure and dynamics of enzyme tunnels. Comput Struct Biotechnol J. Dec 2024;23:3946-3954. [doi: 10.1016/j.csbj.2024.10.051]

Abbreviations

ADMET: absorption, distribution, metabolism, excretion, and toxicity
AI: artificial intelligence
CCC: concordance correlation coefficient
hpol η: human DNA polymerase η
ITBA: indole thio-barbituric acid
KNN: K-nearest neighbor
LightGBM: light gradient boosting machines
MAE: mean absolute error
MAPE: mean absolute percentage error
MedAE: median absolute error
ML: machine learning
MLQSAR: machine learning–based quantitative structure-activity relationship
MSE: mean square error
NMSE: normalized mean square error
NRMSE: normalized root mean square error
PCC: Pearson correlation coefficient
QSAR: quantitative structure-activity relationship
RMSE: root mean square error
RMSLE: root mean squared logarithmic error
SHAP: Shapley additive explanations
SMAPE: symmetric mean absolute percentage error
SMILES: Simplified Molecular Input Line Entry System
TLS: translesion DNA synthesis
XGBoost: extreme gradient boosting

Edited by Gang Luo; peer-reviewed by Hussein A K Kyhoiesh, Xiaolong Liang; submitted 21.May.2025; final revised version received 12.Aug.2025; accepted 16.Aug.2025; published 03.Dec.2025

Please cite as:
Kakraba S, Ayyadevara S, Yadem Clement A, 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 AI2025;4:e77890
URL: https://ai.jmir.org/2025/1/e77890
doi: 10.2196/77890