Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 22 May 2026
Revised 25 June 2026
Accepted 22 July 2026

Natural Resources for Human Health 2026, Vol. 6 (10s): 379–393

KEYWORDS:
Natural products; antibacterial activity; machine learning; explainable artificial intelligence; XGBoost; molecular fingerprints; molecular descriptors; SHAP; bioactivity prediction; NPASS 3.0.

Explainable Machine Learning for Predicting Antibacterial Activity of Natural Products

Deepthi Goteti[1], A. Geetha[2], Lakshmi Naga Jayaprada Gavarraju[3], Avula Chitty[4], Guttikonda Prashanti[5]

[1]Department of Computer Science & Engineering, Malla Reddy University, Maisammaguda, Dulapally, Medchal–Malkajgiri District, Hyderabad, Telangana 500100, India.
Email: goteti.deepthi@mallareddyuniversity.ac.in
[2]Department of CSE(AI&ML), CVR College of Engineering, Mangalpalli, Ibrahimpatnam, Ranga Reddy District, Hyderabad, Telangana 501510, India. Email: songagitaabothu@gmail.com
[3]Department of CSE (CyS, DS) and AI & DS, Vallurupalli Nageswara Rao Vignana Jyothi Institute of Engineering & Technology, Bachupally, Hyderabad, Telangana 500090, India. Email: jayaprada_g@vnrvjiet.in
[4]Department of Computer Science & Engineering, Malla Reddy University, Maisammaguda, Dulapally, Medchal–Malkajgiri District, Hyderabad, Telangana 500100, India.
Email: avula.chitty@mallareddyuniversity.ac.in
[5]Department of Cyber Security & Data Science, Vignan's Foundation for Science, Technology and Research, Vadlamudi, Guntur, Andhra Pradesh 522213, India. Email: prashantiguttikonda77@gmail.com
*Corresponding Author: goteti.deepthi@mallareddyuniversity.ac.in

ABSTRACT: Natural products are a valuable source of bioactive compounds with considerable potential for antibacterial drug discovery. However, experimental screening of large natural-product collections is often time-consuming and resource-intensive. In this study, an explainable machine-learning framework was developed to predict antibacterial activity of natural products using data from the NPASS 3.0 database. Bacterial minimum inhibitory concentration (MIC) records were processed and classified using an operational threshold of ≤8 μg/mL, resulting in a final dataset of 3,130 compounds, including 624 active and 2,506 inactive compounds. Ten molecular descriptors were combined with 2,048-bit Morgan fingerprints to represent the molecular structures. Logistic Regression, Support Vector Machine, Random Forest and XGBoost models were evaluated using five-fold cross-validation and an independent test set. On the independent test set, Random Forest achieved the highest accuracy (0.8530) and ROC-AUC (0.8698), while XGBoost achieved the highest recall (0.6320), F1-score (0.6245) and MCC (0.5295). XGBoost was further interpreted using SHAP analysis to examine the contribution of molecular descriptors and fingerprint features to individual predictions. Morgan fingerprint features showed the highest overall contributions, while LogP was the most influential conventional molecular descriptor. The model was also used to prioritize natural products with high predicted probabilities of antibacterial activity. The proposed framework provides an interpretable computational approach for antibacterial activity prediction and can support the prioritization of natural products for further biological investigation.

Natural Resources for Human Health
eISSN: 2583-1194

1. INTRODUCTION

Natural products have played an important role in the discovery of biologically active compounds and therapeutic agents. Plants, microorganisms and other natural sources contain a wide range of structurally diverse molecules, many of which show antibacterial, anticancer, anti-inflammatory, antidiabetic and other pharmacological activities. The structural diversity of natural products is particularly valuable in drug discovery because many of their chemical structures differ from those commonly represented in synthetic compound libraries [1–4]. Although modern drug-discovery methods have expanded rapidly, natural products continue to provide important opportunities for identifying new bioactive molecules [5,6].

The identification of biologically active natural compounds generally involves several stages, including extraction, isolation, structural characterization and biological testing. These experimental approaches are essential for confirming activity, but screening a large number of compounds can require considerable time and resources. The increasing number of natural-product structures and biological activity measurements has therefore created a need for computational approaches that can help prioritize compounds before experimental testing [7,8]. Databases containing natural-product structures and experimentally measured biological activities have made it possible to investigate relationships between molecular characteristics and biological activity using computational methods [9,10].

Quantitative structure–activity relationship (QSAR) methods and machine learning have become useful approaches for studying these relationships. Molecular descriptors, fingerprints and other molecular representations can be used to describe chemical structures and provide inputs for predictive models. Machine-learning algorithms can then identify patterns between these molecular characteristics and experimentally observed activities. Random forest, support vector machine, logistic regression and gradient-boosting approaches have been widely used in molecular property and bioactivity prediction [11–15]. In recent years, these approaches have also been increasingly applied to natural-product research. For example, machine-learning methods have been used to classify the bioactivity of natural products and to screen phytochemicals for potential therapeutic activity [16,17].

However, predictive accuracy alone does not provide sufficient information about how a machine-learning model reaches its decision. This issue is particularly relevant in natural-product research because a computational prediction is generally used to guide further biological investigation. If a model predicts a compound as active, it is useful to know which molecular characteristics contributed to that prediction. Conventional machine-learning models may provide good predictive performance while offering limited insight into the relationship between the input molecular features and the final prediction. This lack of interpretability can reduce the practical usefulness of computational predictions in drug-discovery studies [18,19].

Explainable artificial intelligence (XAI) provides an opportunity to address this limitation. XAI methods can be used to examine the contribution of individual features to a model prediction and can therefore provide additional information beyond conventional performance measures. SHapley Additive exPlanations (SHAP) is one of the widely used approaches for interpreting machine-learning models. It provides an estimate of the contribution of individual features to a prediction and can be used to examine both overall model behaviour and individual compound predictions [20,21]. Recent studies have demonstrated the usefulness of SHAP and other explainability approaches in drug discovery and molecular property prediction [22–24]. However, the interpretation of machine-learning predictions specifically for natural-product bioactivity remains comparatively less explored than the prediction itself.

Another important challenge is the quality and consistency of natural-product activity data. Natural-product datasets may contain compounds from different biological sources and measurements obtained using different experimental conditions. Differences in activity measurements, duplicate structures, missing molecular information and class imbalance can affect the reliability of machine-learning models. The choice of molecular representation and validation strategy can also influence the ability of a model to generalize to previously unseen compounds [25–27]. Therefore, careful data preparation and model validation are necessary when developing predictive models for natural-product bioactivity.

Recent developments in natural-product databases provide an opportunity to address these challenges using experimentally supported activity information. The Natural Product Activity and Species Source database (NPASS) contains natural products together with information on their biological activities and source organisms, providing a useful resource for computational investigation of natural-product bioactivity [28]. The availability of activity-associated natural-product data makes it possible to

develop models that are directly connected to experimentally observed biological responses rather than relying only on structural similarity.

Antibacterial activity was selected as the biological problem in the present study because antimicrobial resistance and the continuing need for new antibacterial agents remain important challenges. Natural products have historically provided numerous biologically active antimicrobial molecules, and their chemical diversity makes them an attractive source for identifying potential antibacterial candidates [29–31]. Computational prioritization can help reduce the number of compounds requiring immediate experimental evaluation and may provide a practical first step for further biological investigation.

Therefore, the present study develops an explainable machine-learning framework for predicting the antibacterial activity of natural products using experimentally associated activity data from NPASS 3.0. Molecular descriptors and Morgan fingerprints are used to represent the chemical characteristics of the compounds. Four selected machine-learning algorithms, namely logistic regression, random forest, support vector machine and XGBoost, are evaluated to compare their ability to distinguish active and less-active compounds. Model development is performed using stratified five-fold cross-validation together with an independent test set, while scaffold-based validation is considered where the dataset structure permits. SHAP analysis is subsequently applied to the best-performing model to identify the molecular features contributing to antibacterial activity predictions. The study aims to provide both predictive and interpretable information that can be used to prioritize natural products for further antibacterial investigation.

2. LITERATURE REVIEW

Recent research has shown a growing interest in applying machine learning to natural-product research. The availability of natural-product databases containing chemical structures, source information and experimentally measured biological activities has made it possible to develop computational models for activity prediction and compound prioritization [5–9]. Molecular descriptors and structural fingerprints are commonly used to represent chemical compounds, while algorithms such as logistic regression, support vector machine, random forest and XGBoost have been applied to identify relationships between molecular characteristics and biological activity [10–17]. These approaches can reduce the number of compounds that need to be considered during the initial stages of biological screening.

Several recent studies have specifically explored machine learning for natural-product and phytochemical screening. Machine-learning models have been used for natural-product bioactivity classification and for identifying phytochemicals with potential therapeutic activities [20–22]. These studies demonstrate that computational models can provide useful predictions from chemical information. However, differences in datasets, activity measurements, molecular representations and validation procedures can substantially influence the reported performance. Therefore, appropriate preprocessing and validation are important when developing models for natural-product bioactivity prediction [18,19].

The use of explainable artificial intelligence has further expanded the scope of molecular machine learning. Methods such as LIME and SHAP allow researchers to investigate the contribution of individual features to model predictions [25,26]. SHAP-based approaches have recently been used in drug discovery to provide global feature importance as well as explanations for individual molecular predictions [27–29]. Such information can be particularly useful in natural-product research because identifying the molecular characteristics associated with predicted activity may help researchers prioritize compounds for subsequent investigation.

Despite these developments, there remains a need for computational studies that combine natural-product activity data, molecular descriptors, structural fingerprints, model comparison and explainability within a single framework. In particular, the interpretation of antibacterial activity predictions for natural products has received less attention than prediction performance alone. In addition, issues such as activity-class imbalance, structural similarity between compounds and model generalization need to be considered carefully [30–33]. The present study therefore combines selected machine-learning algorithms with molecular representations and SHAP analysis to develop an interpretable framework for predicting antibacterial activity from natural products using NPASS 3.0 data.

3. MATERIALS AND METHODS

3.1 Data Source

The data used in this study were obtained from the NPASS 3.0 (Natural Product Activity and Species Source) database, which provides information on natural products, their chemical structures, source organisms and experimentally reported biological activities [34,35]. The database was selected because it contains activity-associated natural products suitable for computational bioactivity modelling. The present study focused specifically on antibacterial activity. Relevant activity records were extracted from NPASS 3.0 based on antibacterial activity information and the availability of corresponding molecular structures. The overall computational workflow followed in this study is presented in Figure 1.

3.2 Data Collection and Preprocessing

The antibacterial activity records were first screened to identify compounds with a valid chemical structure and experimentally reported activity information. Records with missing molecular structures, missing activity values or insufficient information for classification were excluded.

Duplicate compounds were identified using their molecular structure, and duplicate records were removed to avoid repeated observations during model development. Where activity values were reported using different units, the values were converted to a common unit before classification. When multiple activity measurements were available for the same compound, the available measurements were examined and a consistent rule was applied for assigning the final activity label. After data cleaning, the resulting dataset was used for antibacterial activity classification and molecular feature generation. The final number of compounds and class distribution will be reported after completion of the dataset processing.

3.3 Antibacterial Activity Classification

The antibacterial activity prediction problem was formulated as a binary classification task, with compounds categorized as either active or inactive according to a predefined activity threshold. For compounds with minimum inhibitory concentration (MIC) measurements, the activity class can be expressed as:

$$Y_i = 1, \text{ if } MIC_i \leq C \quad ; \quad Y_i = 0, \text{ if } MIC_i > C \qquad (1)$$

where Y_i represents the activity class of compound i and C represents the predefined antibacterial activity cut-off. The final cut-off will be selected based on the distribution and measurement characteristics of the NPASS 3.0 antibacterial dataset rather than being chosen solely to balance the two classes. The final numbers of active and inactive compounds will be reported after dataset processing.

3.4 Molecular Representation

3.4.1 Molecular Descriptors

A selected group of molecular descriptors was calculated from the molecular structures. The descriptors were chosen to represent important physicochemical and structural characteristics of the compounds, including molecular weight, calculated LogP, topological polar surface area (TPSA), hydrogen-bond donors, hydrogen-bond acceptors, rotatable bonds, ring count, aromatic ring count, fraction of sp^3-hybridized atoms and heavy-atom count.

These descriptors provide information related to molecular size, lipophilicity, polarity, hydrogen-bonding capacity and structural complexity. Rather than including a very large number of descriptors, the analysis focused on informative features that could be used effectively with the selected machine-learning models.

3.4.2 Morgan Fingerprints

Morgan fingerprints were generated to capture structural patterns within the natural-product molecules. These fingerprints represent molecular environments as binary features indicating the presence or absence of particular structural substructures [26].

The fingerprint representation complements the conventional molecular descriptors by capturing structural information that may not be adequately represented by individual physicochemical descriptors. The final fingerprint radius and bit length will be reported according to the actual computational implementation used in the study.

3.5 Feature Selection and Scaling

Feature selection was performed to reduce redundant information and retain molecular features that were informative for antibacterial activity prediction. Features with very low variance and highly correlated descriptors were considered for removal.

For two features X_j and X_k, their Pearson correlation coefficient was calculated as:

$$r_{jk} = \Sigma_i[(X_{ij} - \bar{X_j})(X_{ik} - \bar{X_k})] / \sqrt{\{\Sigma_i(X_{ij} - \bar{X_j})^2 \times \Sigma_i(X_{ik} - \bar{X_k})^2\}} \quad (2)$$

Highly correlated descriptors were reduced to minimize redundancy. The final feature-selection criterion will be reported based on the actual analysis.

Continuous molecular descriptors were standardized when required by the machine-learning algorithm using:

$$z = (x - \mu) / \sigma \quad (3)$$

where x is the original feature value, μ is the mean calculated from the training data and σ is the corresponding standard deviation. Preprocessing parameters were derived from the training data only and then applied to validation and test data to avoid information leakage.

3.6 Data Splitting and Cross-Validation

The processed dataset was divided into training and independent test sets using a stratified splitting strategy. An approximately 80:20 ratio was used, with the training set used for model development and the independent test set reserved for final performance evaluation. Stratified 5-fold cross-validation was performed within the training set for model selection and hyperparameter optimization. In each fold, the model was trained using four subsets and evaluated on the remaining subset. This process was repeated until each subset had been used for validation.

The mean cross-validation performance was used for model comparison, while the independent test set was kept separate from model selection. Where feasible, scaffold-based evaluation will also be considered to determine whether the models can generalize to structurally different natural products.

3.7 Machine-Learning Models

Four machine-learning algorithms were selected for antibacterial activity prediction: Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF) and Extreme Gradient Boosting (XGBoost). The selected models provide a comparison between linear, kernel-based and tree-based learning approaches [23–25].

- Logistic Regression (LR) – used as a baseline linear classification model.
- Support Vector Machine (SVM) – used to model classification boundaries in the molecular feature space.
- Random Forest (RF) – used to capture nonlinear relationships through an ensemble of decision trees.
- XGBoost – used as a gradient-boosting approach capable of modelling nonlinear feature interactions.

Hyperparameter optimization was restricted to the parameters most relevant to each model. For LR, regularization strength and penalty type were considered. For SVM, kernel, regularization parameter and gamma were evaluated. For RF, the number of trees, maximum tree depth and minimum samples per leaf were considered. For XGBoost, the number of estimators, learning rate, maximum tree depth and subsampling parameters were considered. Hyperparameter selection was performed using the training data and stratified 5-fold cross-validation.

3.8 Model Performance Evaluation

The developed models were evaluated using the independent test set. The following classification measures were considered:

Accuracy:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (4)$$

Precision:

$$\text{Precision} = TP / (TP + FP) \quad (5)$$

Recall (Sensitivity):

$$\text{Recall} = TP / (TP + FN) \quad (6)$$

F1-score:

$$\text{F1-score} = 2 \times \text{Precision} \times \text{Recall} / (\text{Precision} + \text{Recall}) \quad (7)$$

Matthews correlation coefficient (MCC):

$$MCC = (TP \times TN - FP \times FN) / \sqrt{[(TP + FP)(TP + FN)(TN + FP)(TN + FN)]} \quad (8)$$

where TP, TN, FP and FN represent true positives, true negatives, false positives and false negatives, respectively. ROC-AUC and PR-AUC were also calculated to provide threshold-independent measures of model discrimination, particularly because natural-product activity datasets may contain unequal numbers of active and inactive compounds.

3.9 Explainable Artificial Intelligence Using SHAP

The best-performing machine-learning model was further analyzed using SHAP (SHapley Additive exPlanations) to understand the factors contributing to its predictions [27]. SHAP assigns a contribution value to each feature for an individual prediction. The model output can be represented as:

$$f(x) = \varphi_0 + \Sigma_j \varphi_j \quad (9)$$

where f(x) is the model prediction for a compound, φ_0 is the expected model output and φ_j represents the contribution of feature j.

Global SHAP analysis was used to identify the molecular features that had the greatest overall influence on antibacterial activity prediction. SHAP summary and dependence analyses were used to examine the direction and magnitude of feature contributions.

Local SHAP explanations were additionally generated for selected compounds to determine why individual natural products were predicted as active or inactive. This provided an interpretable layer to the predictive modelling rather than relying only on numerical performance measures.

3.10 Candidate Prioritization

The best-performing model was subsequently used to estimate the probability of antibacterial activity for compounds in the dataset. Natural products with high predicted activity probabilities were ranked for computational prioritization.

The ranking was based on the predicted probability generated by the selected model. High-confidence compounds were considered potential candidates for further biological investigation. These predictions represent computational prioritization and do not constitute experimental confirmation of antibacterial activity.

The complete workflow of the study is summarized in Figure 1. The analysis begins with the extraction of antibacterial activity data from NPASS 3.0, followed by data cleaning, bacterial taxonomic filtering, MIC unit conversion, and activity classification. An operational MIC threshold of ≤8 μg/mL was used to classify compounds as active, while compounds with higher MIC values were classified as inactive. Compounds with conflicting or ambiguous activity assignments were excluded to obtain a high-confidence dataset. Valid molecular structures were then used to calculate ten molecular descriptors and generate 2048-bit Morgan fingerprints. These features were combined and used for machine-learning analysis.

The final dataset was divided into training and independent test sets using an 80:20 stratified split, and the training data were evaluated using five-fold stratified cross-validation. Logistic Regression, Support Vector Machine, Random Forest, and XGBoost models were trained and compared using accuracy, precision, recall, F1-score, ROC-AUC, PR-AUC, and MCC. The independent test set was subsequently used for final performance assessment. Based on the comparative results, XGBoost was selected for SHAP-based interpretation to examine the molecular features contributing to its predictions. Finally, compounds

with high predicted probabilities of antibacterial activity were ranked for computational prioritization and further investigation. This workflow provides a systematic connection between natural-product activity data, molecular characteristics, predictive modelling, and interpretable candidate selection.

The workflow also ensures that the same data-processing and evaluation procedure is followed for all models, making the comparison more consistent. The final interpretation helps to understand not only which compounds are predicted to be active, but also which molecular features contribute to these predictions. This provides a simple and transparent computational approach for supporting the selection of natural products for further antibacterial studies.

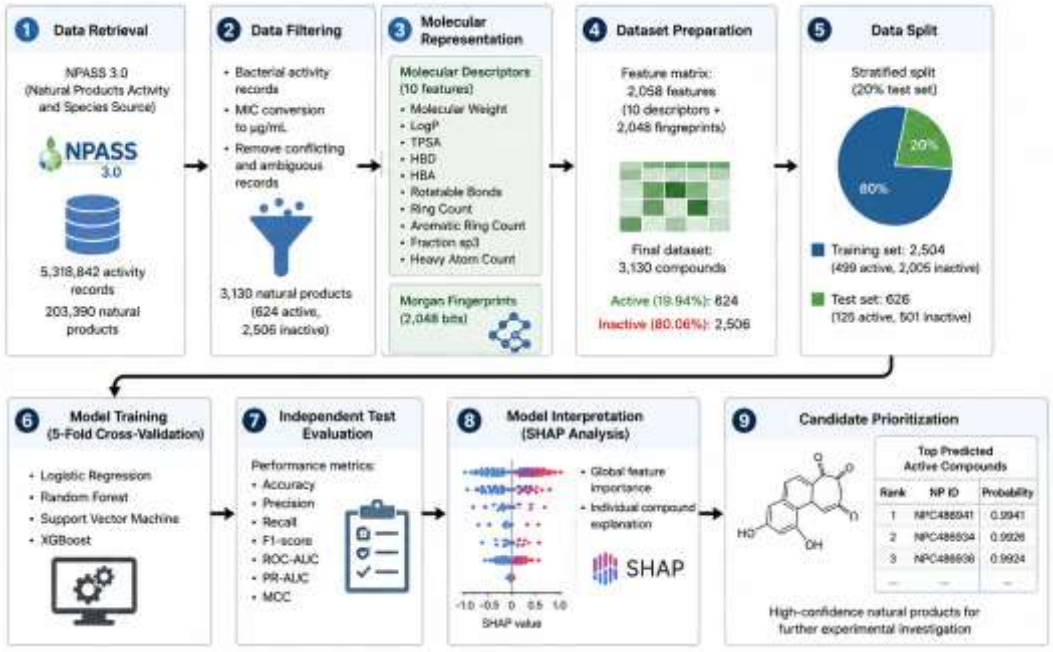

Fig 1. Overall workflow of the proposed machine-learning framework for antibacterial activity prediction and explainability analysis of natural products using NPASS 3.0 data.

4. RESULTS AND DISCUSSION

4.1 Dataset Characteristics and Molecular Feature Generation

The processed NPASS 3.0 activity data yielded 83,278 bacterial MIC records corresponding to 4,295 natural products. After conversion of compatible MIC measurements to μg/mL, 82,643 records were available for activity classification. An operational MIC threshold of ≤8 μg/mL was used to define an active observation, whereas values above 8 μg/mL were classified as inactive. Ambiguous and conflicting compound-level observations were excluded to obtain a high-confidence classification set. This resulted in 3,131 compounds, comprising 624 active and 2,507 inactive compounds. One compound with an invalid molecular structure was removed during molecular feature generation, leaving 3,130 compounds for machine-learning analysis. A summary of the dataset filtering, activity classification and final compound composition is presented in Table 1 below. The final modeling dataset contained 624 active compounds (19.94%) and 2,506 inactive compounds (80.06%). Thus, the dataset was imbalanced toward the inactive class. This imbalance was considered during model development by stratifying the train-test split and using class weighting for the relevant classifiers. Ten molecular descriptors were combined with 2,048-bit Morgan fingerprints, giving 2,058 input features per compound. The molecular descriptors included molecular weight, LogP, TPSA, hydrogen-bond donors, hydrogen-bond acceptors, rotatable bonds, ring count, aromatic ring count, fraction sp3 and heavy atom count.

Table 1. Summary of the final dataset and molecular representation used for machine-learning analysis.

Dataset characteristic	Value
Bacterial MIC records after taxonomic filtering	83,278
Usable MIC records after unit conversion	82,643
Natural products before high-confidence labeling	4,295
High-confidence compounds after labeling	3,131
Final compounds after structure validation	3,130
Active compounds	624 (19.94%)
Inactive compounds	2,506 (80.06%)
Molecular descriptors	10
Morgan fingerprint length	2,048 bits
Total model features	2,058

3.2 Comparative Performance of Machine-Learning Models

Four classification algorithms—Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM) and XGBoost—were evaluated using stratified five-fold cross-validation. The training set contained 2,504 compounds (499 active and 2,005 inactive), while 626 compounds (125 active and 501 inactive) were retained as an independent test set. No compounds from the independent test set were used during cross-validation or model fitting. The cross-validation results showed a clear advantage for the tree-based approaches. Random Forest achieved the highest mean accuracy (0.8674), precision (0.7825), ROC-AUC (0.8713) and PR-AUC (0.7018). XGBoost produced the highest mean recall (0.6071), F1-score (0.6271) and MCC (0.5401). Logistic Regression had the lowest ROC-AUC (0.7694) and PR-AUC (0.5461), while SVM showed intermediate performance. The difference between RF and XGBoost is important because it reflects two different performance priorities: RF was stronger in overall discrimination and precision, whereas XGBoost was better at recovering active compounds.

Table 2: Five-fold cross-validation performance of the four machine-learning models.

Model	Accuracy	Precision	Recall	F1	ROC-AUC	PR-AUC	MCC
Logistic Regression	0.7899	0.4813	0.5992	0.5320	0.7694	0.5461	0.4044
Random Forest	0.8674	0.7825	0.4689	0.5845	0.8713	0.7018	0.5364
SVM	0.8586	0.7137	0.4909	0.5804	0.8470	0.6557	0.5123
XGBoost	0.8558	0.6534	0.6071	0.6271	0.8492	0.6829	0.5401

3.3 Independent Test-Set Evaluation

The independent test set was used to assess how well the models generalized to compounds that were not used during training. The results broadly followed the cross-validation pattern, although the absolute values changed slightly, as expected for an independent sample. The complete independent test-set performance of the four machine-learning models is presented in Table 3 below. Random Forest achieved the highest test accuracy (0.8530), precision (0.6701), ROC-AUC (0.8698) and PR-AUC (0.6967), whereas XGBoost achieved the highest recall (0.6320), F1-score (0.6245) and MCC (0.5295).

The confusion matrices shown in Figure 2 provide a clearer view of the differences in classification performance among the four models. Random Forest correctly identified 469 of 501 inactive compounds and 65 of 125 active compounds, producing 32 false positives and 60 false negatives. XGBoost correctly identified 452 inactive compounds and 79 active compounds, with 49 false positives and 46 false negatives.

Therefore, XGBoost recovered 14 additional active compounds compared with Random Forest, while accepting 17 additional false-positive predictions. This trade-off explains why Random Forest achieved higher precision, whereas XGBoost showed higher recall and F1-score. The difference in these predictions was also important for selecting XGBoost for the subsequent SHAP-based interpretation and candidate prioritization analysis.

Table 3: Independent test-set performance of the four machine-learning models.

Model	Accuracy	Precision	Recall	F1	ROC-AUC	PR-AUC	MCC
Logistic Regression	0.7939	0.4877	0.6320	0.5505	0.7653	0.5443	0.4256
Random Forest	0.8530	0.6701	0.5200	0.5856	0.8698	0.6967	0.5039
SVM	0.8419	0.6204	0.5360	0.5751	0.8297	0.6252	0.4805
XGBoost	0.8482	0.6172	0.6320	0.6245	0.8569	0.6921	0.5295

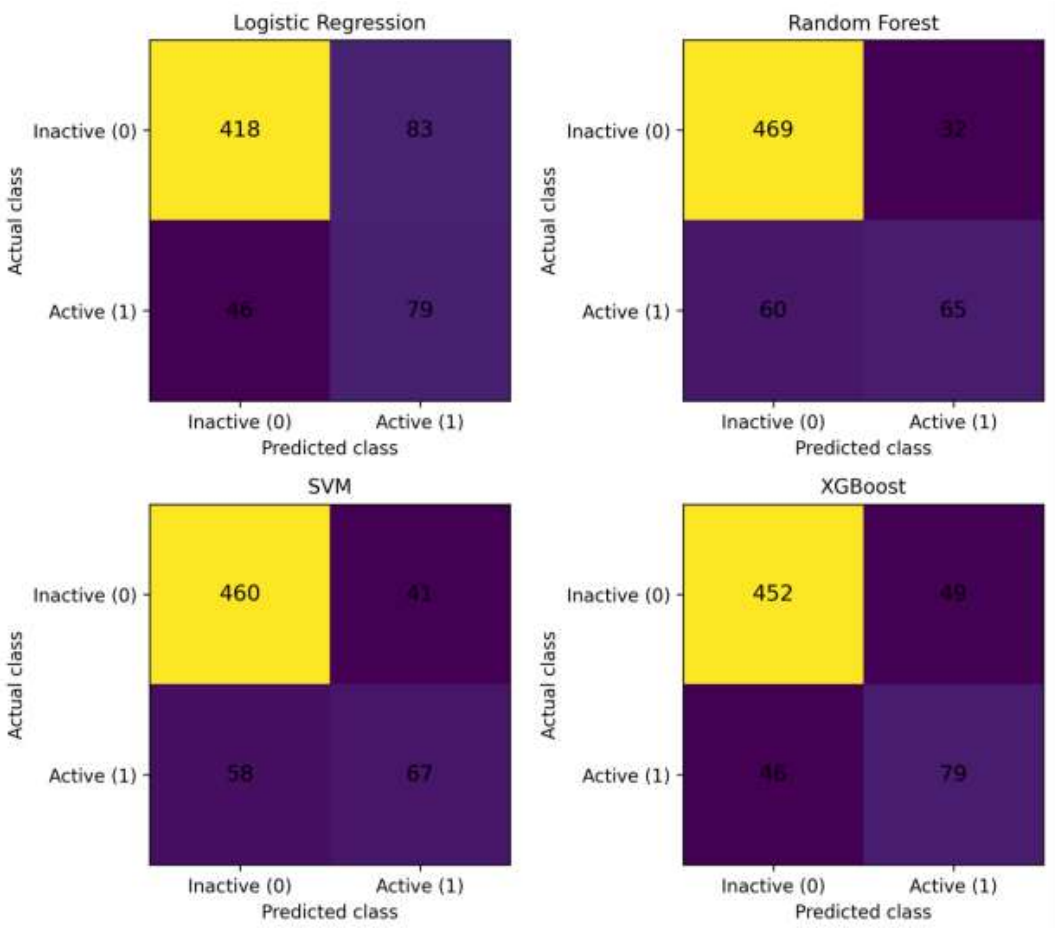

Fig 2: Confusion matrices for the four machine-learning models on the independent test set.

For antibacterial candidate prioritization, this distinction is relevant. A model that is too conservative may miss active natural products, while a model with higher sensitivity may return additional candidates for subsequent experimental screening. On

this basis, XGBoost was selected for the explainability analysis, while Random Forest was retained as the strongest comparator in terms of overall discrimination.

3.4 ROC and Precision–Recall Analysis

The ROC analysis confirmed good discrimination across the evaluated models. **The receiver operating characteristic curves for the four machine-learning models on the independent test set are shown in Figure 4.** Random Forest produced the highest ROC-AUC (0.8698), followed by XGBoost (0.8569), SVM (0.8297), and Logistic Regression (0.7653). The relatively close ROC-AUC values of Random Forest and XGBoost indicate that both models provided strong discrimination between active and inactive natural products. Overall, the ROC results support the suitability of the tree-based models for capturing the complex relationships between molecular representations and antibacterial activity.

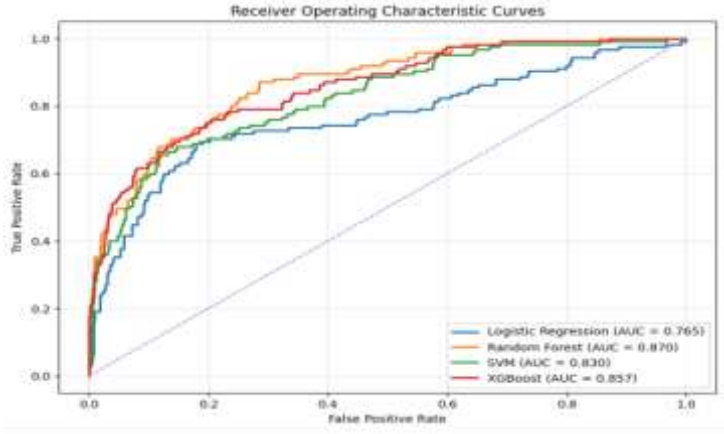

Fig 3: Receiver operating characteristic curves of the four machine-learning models evaluated on the independent test set.

Because active compounds represented only 19.94% of the final dataset, the precision–recall analysis provides an important complementary measure. The precision–recall curves of the four machine-learning models evaluated on the independent test set are shown in Figure 4. Random Forest achieved a PR-AUC of 0.6967, while XGBoost achieved 0.6921, with both values substantially above the test-set prevalence baseline of 0.200. SVM and Logistic Regression produced lower PR-AUC values of 0.6252 and 0.5443, respectively. The PR results therefore support the stronger performance of the two tree-based models under class imbalance. The higher PR-AUC values also indicate that Random Forest and XGBoost maintained a better balance between identifying active compounds and limiting false-positive predictions.

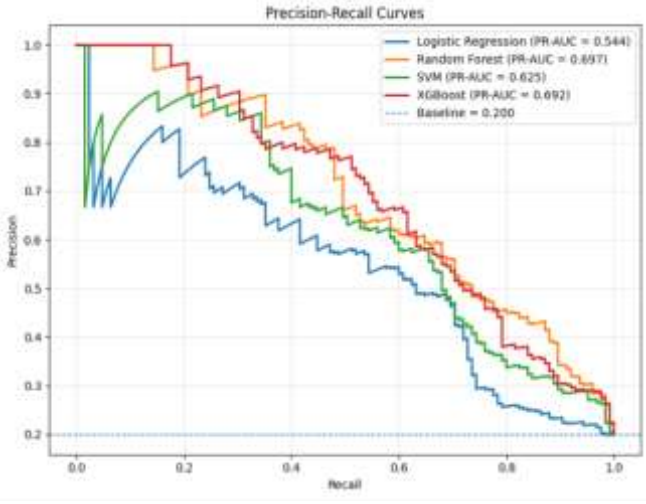

Fig 4: Precision–recall curves of the four machine-learning models on the independent test set. The dashed line represents the active-class prevalence baseline.

3.5 SHAP-Based Interpretation of the XGBoost Model

SHAP analysis was applied to the XGBoost model to examine how individual features contributed to its predictions. The analysis was performed on the independent test set containing 626 compounds. The global SHAP analysis showed that several Morgan fingerprint features had the highest overall contributions, including Morgan_314, Morgan_1602, Morgan_1162, Morgan_745 and Morgan_1292. Among the conventional molecular descriptors, LogP showed the highest overall contribution, followed by Fraction_sp3 and Molecular_Weight. TPSA, hydrogen-bond acceptors, hydrogen-bond donors, ring count and rotatable bonds also contributed to the model predictions.

The presence of several Morgan fingerprint features among the leading variables indicates that the XGBoost model relied on local structural patterns in addition to global physicochemical properties. This is useful for natural products because their structures can be chemically diverse, and fingerprint features can capture structural patterns that may not be represented by individual molecular descriptors. The combination of molecular descriptors and Morgan fingerprints therefore provided the model with both general physicochemical information and structural information.

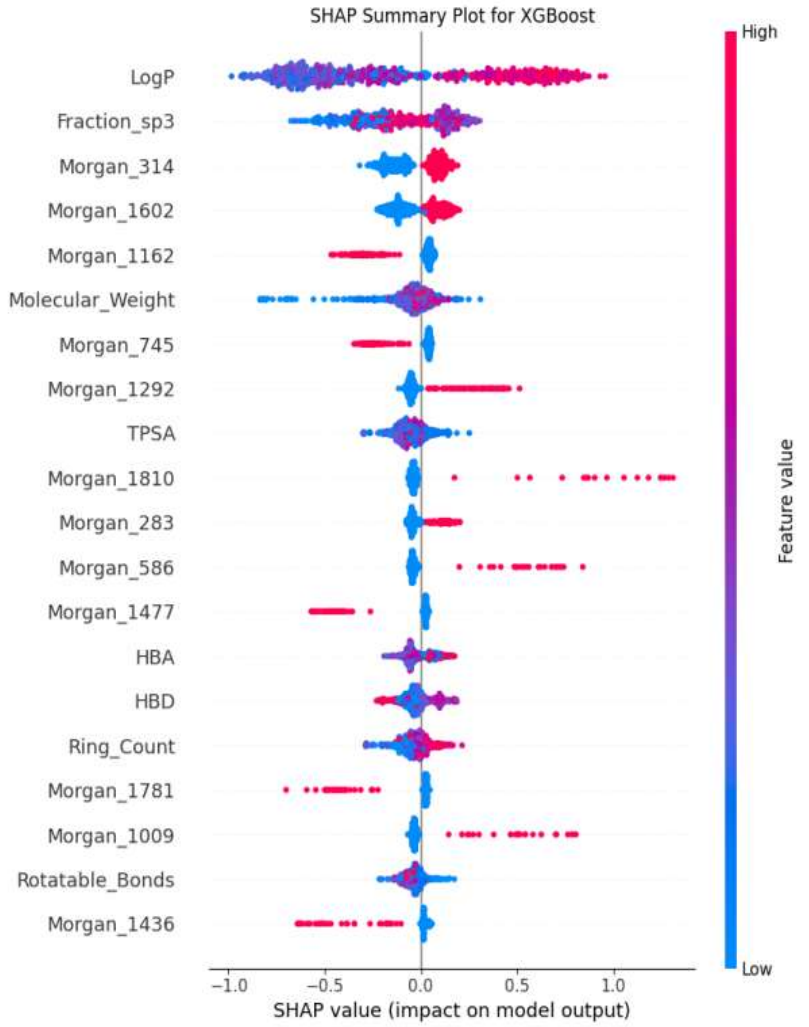

Fig 5: SHAP summary plot showing the most influential features contributing to XGBoost predictions on the independent test set.

The SHAP distribution also provides information about the direction of each feature's contribution. Positive SHAP values push a prediction toward the active class, whereas negative SHAP values push it toward the inactive class. For example, LogP showed contributions in both positive and negative directions, indicating that its effect varied across compounds rather than consistently increasing or decreasing the predicted antibacterial activity. Similarly, the contribution of individual Morgan fingerprint features

depended on the molecular context in which they occurred. Therefore, the SHAP results should be interpreted as model-level associations rather than evidence that a particular descriptor or fingerprint feature directly causes antibacterial activity.

3.6 Individual Compound Interpretation and Candidate Prioritization

To demonstrate the practical application of the model, predicted probabilities were examined for compounds in the independent test set. The highest predicted probability was obtained for NPC486941 (0.9941), followed by NPC486934 (0.9926), NPC486936 (0.9924), NPC486938 (0.9923), NPC486948 (0.9922) and NPC486944 (0.9906). All ten compounds listed in Table 4 had an observed active class in the NPASS-derived dataset. These compounds were therefore considered high-confidence computationally prioritized natural products based on their predicted probabilities and observed activity labels.

For NPC486941, the strongest positive SHAP contributions were Morgan_1810 (1.2669), Morgan_1009 (0.7852), Morgan_448 (0.7622), Morgan_1292 (0.4280) and Morgan_627 (0.3763). The strongest negative contributions were Morgan_1399 (−0.1913), TPSA (−0.0727), Morgan_1017 (−0.0711) and Morgan_1602 (−0.0705). Thus, the final prediction reflects the combined contribution of multiple structural features rather than a single descriptor. The SHAP values also help explain why a compound can receive a high predicted probability even when some individual features contribute in the negative direction.

The ranking of the compounds was based on the predicted probability generated by the final XGBoost model. NPC486941 showed the highest probability among the evaluated test compounds, indicating that the molecular representation of this compound was strongly associated with the active class according to the trained model. The remaining compounds in Table 4 also showed high predicted probabilities, suggesting that they may be useful for further computational or experimental investigation.

Importantly, these compounds should be considered **model-prioritized natural products rather than newly discovered antibacterial compounds**. Their prioritization is based on the learned patterns in the NPASS-derived dataset and does not by itself provide experimental confirmation of antibacterial activity. The combination of predicted probability and individual SHAP contributions provides an interpretable way to identify compounds that can be considered for subsequent biological evaluation.

Table 4: High-confidence natural products prioritized by the XGBoost model from the independent test set.

Rank	NP ID	Observed class	Predicted probability
1	NPC486941	Active	0.9941
2	NPC486934	Active	0.9926
3	NPC486936	Active	0.9924
4	NPC486938	Active	0.9923
5	NPC486948	Active	0.9922
6	NPC486944	Active	0.9906
7	NPC20175	Active	0.9714
8	NPC244994	Active	0.9663
9	NPC119225	Active	0.9478
10	NPC66699	Active	0.9419

3.7 Discussion

The present results show that antibacterial activity classification of natural products can be approached effectively using a combination of molecular descriptors and Morgan fingerprints. The performance differences among the models suggest that nonlinear relationships within the chemical feature space are important. Logistic Regression provided a useful linear baseline but showed lower discrimination than the tree-based approaches. SVM improved upon the linear model, whereas Random Forest and XGBoost provided the strongest overall performance.

Random Forest consistently produced the highest ROC-AUC and PR-AUC, indicating strong discrimination and ranking ability. Its higher precision also resulted from fewer false-positive predictions on the independent test set. XGBoost, however, identified more active compounds and achieved higher recall, F1-score and MCC. This behavior is relevant to natural-product screening, where the objective is often to prioritize promising compounds for additional investigation rather than to minimize the number of computational candidates at all costs. The agreement between cross-validation and independent test results provides additional confidence in the model evaluation. For example, the RF ROC-AUC changed only slightly from 0.8713 during cross-validation to 0.8698 on the independent test set, while XGBoost changed from 0.8492 to 0.8569. The preservation of performance across these two evaluation settings suggests that the models were not dependent on a single favorable training split. Nevertheless, the dataset is structurally diverse and class imbalanced, and the present evaluation should be interpreted within the scope of the NPASS-derived dataset.

The SHAP analysis adds an interpretability layer that is not available from predictive performance metrics alone. LogP, Fraction_sp3 and Molecular_Weight were the leading conventional descriptors, while several Morgan fingerprint features were also highly influential. This combination suggests that both broad physicochemical properties and local structural patterns contributed to the model decisions. Importantly, SHAP identifies features used by the predictive model and does not establish a direct mechanistic or causal relationship with antibacterial activity.

The prioritized compounds should therefore be regarded as computationally prioritized natural products for further study, rather than as newly discovered antibacterial agents. Since the candidates were selected from the NPASS-derived test set, their ranking demonstrates the ability of the model to assign high activity probabilities to compounds with known active labels in the analyzed dataset. Future experimental studies would be required to confirm activity under standardized assay conditions.

3.8 Summary of the Main Findings

Overall, the final 3,130-compound dataset supported reproducible comparison of four machine-learning algorithms using an independent test set. Random Forest showed the strongest overall discrimination, whereas XGBoost provided the strongest active-class detection at the selected decision threshold and was therefore used for SHAP interpretation. The SHAP results identified LogP and Fraction_sp3 as important conventional descriptors and showed substantial contributions from Morgan fingerprint features. Together, the predictive and explainability results support the use of the proposed framework for computational prioritization of natural products for antibacterial research.

ACKNOWLEDGEMENTS

The authors acknowledge the support and computational resources that facilitated this study. The authors also acknowledge the developers and curators of the NPASS 3.0 database for making natural-product activity and structural information available for research purposes.

AUTHOR CONTRIBUTIONS

DG: Conceptualization, Methodology, Data Curation, Formal Analysis, Writing—Original Draft; AG: Methodology, Validation, Investigation; LNGJG: Supervision, Project Administration, Writing—Review & Editing; AC: Data Curation, Formal Analysis, Investigation; GP: Validation, Investigation, Writing—Review & Editing.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

ETHICS STATEMENT

Ethical approval was not required for this study because it involved computational analysis of publicly available natural-product and bioactivity data and did not involve human participants, animals, or identifiable personal data.

DATA AVAILABILITY STATEMENT

The study was based on data obtained from the NPASS 3.0 database. The processed datasets, molecular descriptors, Morgan fingerprints, model results, and analysis outputs generated during this study are available from the corresponding author upon reasonable request. The original NPASS 3.0 data are available through the NPASS database.

REFERENCES

[1] Mullowney, M.W., Duncan, K.R., Elsayed, S.S., et al., 2023. Artificial intelligence for natural product drug discovery. *Nature Reviews Drug Discovery* 22, 895–916. https://doi.org/10.1038/s41573-023-00774-7.

[2] Arora, S., Chettri, S., Percha, V., Kumar, D., Latwal, M., 2024. Artificial intelligence: a virtual chemist for natural product drug discovery. *Journal of Biomolecular Structure and Dynamics* 42, 3826–3835. https://doi.org/10.1080/07391102.2023.2216295.

[3] Ancajas, C.M.F., Oyedele, A.S., Butt, C.M., Walker, A.S., 2024. Advances, opportunities, and challenges in methods for interrogating the structure activity relationships of natural products. *Natural Product Reports* 41, 1543–1578. https://doi.org/10.1039/D4NP00009A.

[4] AI-driven drug discovery from natural products, 2024. *Advanced Agrochem* 3, 185–187. https://doi.org/10.1016/j.aac.2024.06.003.

[5] Ponce-Bobadilla, A.V., Schmitt, V., Maier, C.S., Mensing, S., Stodtmann, S., 2024. Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development. *Clinical and Translational Science* 17, e70056. https://doi.org/10.1111/cts.70056.

[6] Alizadehsani, R., Oyelere, S.S., Hussain, S., et al., 2024. Explainable Artificial Intelligence for Drug Discovery and Development: A Comprehensive Survey. *IEEE Access* 12, 35796–35812. https://doi.org/10.1109/ACCESS.2024.3373195.

[7] Proietti, M., Ragno, A., La Rosa, B., Ragno, R., Capobianco, R., 2024. Explainable AI in drug discovery: self-interpretable graph neural network for molecular property prediction using concept whitening. *Machine Learning* 113, 2013–2044. https://doi.org/10.1007/s10994-023-06369-y.

[8] Brittin, N.J., Anderson, J.M., Braun, D.R., Rajski, S.R., Currie, C.R., Bugni, T.S., 2025. Machine Learning-Based Bioactivity Classification of Natural Products Using LC-MS/MS Metabolomics. *Journal of Natural Products* 88, 361–372. https://doi.org/10.1021/acs.jnatprod.4c01123.

[9] PhyIndBC: Development of a machine learning tool for screening of potential breast cancer inhibitors from phytochemicals, 2025. *Computational and Structural Biotechnology Journal* 42, 100435. https://doi.org/10.1016/j.csbj.2025.100435.

[10] Application of artificial intelligence in bioprospecting for natural products for biopharmaceutical purposes, 2025. *BMC Artificial Intelligence* 1, 4. https://doi.org/10.1186/s44398-025-00004-7.

[11] Artificial intelligence for natural product drug discovery and development: current landscape, applications, and future directions, 2025. *Intelligence-Based Medicine* 12, 100316. https://doi.org/10.1016/j.ibmed.2025.100316.

[12] Muthuraj, R., Chandrasekaran, J., 2026. Nature meets machine: the AI renaissance in natural product drug discovery. *Natural Products and Bioprospecting* 16, 37. https://doi.org/10.1007/s13659-025-00589-6.

[13] Data-quality-guided AI strategies for natural product drug discovery, 2026. *Trends in Pharmacological Sciences*. https://doi.org/10.1016/j.tips.2026.04.008.

[14] Atanasov, A.G., Zotchev, S.B., Dirsch, V.M., et al., 2021. Natural products in drug discovery: advances and opportunities. *Nature Reviews Drug Discovery* 20, 200–216. https://doi.org/10.1038/s41573-020-00114-z.

[15] Newman, D.J., Cragg, G.M., 2020. Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. *Journal of Natural Products* 83, 770–803. https://doi.org/10.1021/acs.jnatprod.9b01285.

[16] Harvey, A.L., Edrada-Ebel, R., Quinn, R.J., 2015. The re-emergence of natural products for drug discovery in the genomics era. *Nature Reviews Drug Discovery* 14, 111–129. https://doi.org/10.1038/nrd4510.

[17] Rodrigues, T., Reker, D., Schneider, P., Schneider, G., 2016. Counting on natural products for drug design. *Nature Chemistry* 8, 531–541. https://doi.org/10.1038/nchem.2479.

[18] Pye, C.R., Bertin, M.J., Lokey, R.S., Gerwick, W.H., Linington, R.G., 2017. Retrospective analysis of natural products provides insights for future discovery trends. *Proceedings of the National Academy of Sciences* 114, 5601–5606. https://doi.org/10.1073/pnas.1614680114.

[19] Vamathevan, J., Clark, D., Czodrowski, P., et al., 2019. Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery* 18, 463–477. https://doi.org/10.1038/s41573-019-0024-5.

[20] Shen, J., Nicolaou, C.A., 2020. Molecular property prediction: recent trends in the era of artificial intelligence. *Drug Discovery Today: Technologies* 32–33, 29–36. https://doi.org/10.1016/j.ddtec.2020.05.001.

[21] Muratov, E.N., Bajorath, J., Sheridan, R.P., et al., 2020. QSAR without borders. *Chemical Society Reviews* 49, 3525–3564. https://doi.org/10.1039/D0CS00098A.

[22] Tropsha, A., 2010. Best Practices for QSAR Model Development, Validation, and Exploitation. *Molecular Informatics* 29, 476–488. https://doi.org/10.1002/minf.201000061.

[23] Cortes, C., Vapnik, V., 1995. Support-vector networks. *Machine Learning* 20, 273–297. https://doi.org/10.1007/BF00994018.

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

[25] Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785.

[26] Rogers, D., Hahn, M., 2010. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling* 50, 742–754. https://doi.org/10.1021/ci100050t.

[27] Wu, Z., Ramsundar, B., Feinberg, E.N., et al., 2018. MoleculeNet: a benchmark for molecular machine learning. *Chemical Science* 9, 513–530. https://doi.org/10.1039/C7SC02664A.

[28] Stokes, J.M., Yang, K., Swanson, K., et al., 2020. A deep learning approach to antibiotic discovery. *Cell* 180, 688–702.e13. https://doi.org/10.1016/j.cell.2020.01.021.

[29] Lipinski, C.F., Maltarollo, V.G., Oliveira, P.R., da Silva, A.B.F., Honorio, K.M., 2019. Advances and perspectives in applying deep learning for drug design and discovery. *Frontiers in Robotics and AI* 6, 108. https://doi.org/10.3389/frobt.2019.00108.

[30] Ribeiro, M.T., Singh, S., Guestrin, C., 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778.

[31] Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems* 30. https://doi.org/10.48550/arXiv.1705.07874.

[32] Ponzoni, I., Páez Prosper, J.A., Campillo, N.E., 2023. Explainable artificial intelligence: A taxonomy and guidelines for its application to drug discovery. *WIREs Computational Molecular Science* 13, e1681. https://doi.org/10.1002/wcms.1681.

[33] Proietti, M., Ragno, A., La Rosa, B., et al., 2024. Explainable AI in drug discovery: self-interpretable graph neural network for molecular property prediction using concept whitening. *Machine Learning* 113, 2013–2044. https://doi.org/10.1007/s10994-023-06369-y.

[34] Ponce-Bobadilla, A.V., Schmitt, V., Maier, C.S., et al., 2024. Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development. *Clinical and Translational Science* 17, e70056. https://doi.org/10.1111/cts.70056.

[35] Alizadehsani, R., Oyelere, S.S., Hussain, S., et al., 2024. Explainable Artificial Intelligence for Drug Discovery and Development: A Comprehensive Survey. *IEEE Access* 12, 35796–35812. https://doi.org/10.1109/ACCESS.2024.3373195.