Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 30 May 2026
Revised 05 July 2026
Accepted 20 July 2026

Natural Resources for Human Health 2026, Vol.6(12s):1505–1517

KEYWORDS:
Natural products, medicinal plants, phytochemicals, NPASS, machine learning, scaffold-aware validation, molecular fingerprints, anticancer activity, QSAR, drug discovery.

Machine Learning Driven Discovery of Anticancer Phytochemicals from Medicinal Plants: A Computational Approach

Humaira Shaheen[1], Rabia Shaheen Niazi[2], Sobia Niaz[3], Amber Sharif[4], Riffat Farooqui[5], Muhammad Salman[6], Ahmad Jawad Khan[7], Aafaq Ali[8], Dr. Abid Ejaz[9], Dr. Mian Jahan Zaib Rasheed[10]

[1] Department of Botany, University of Sargodha, Sargodha, Pakistan
humairakhaan86@gmail.com
[2] Department of Biological Sciences, Thal University of Bhakkar, Punjab, Pakistan
rabia.shaheen@tu.edu.pk
[3] Department of Botany, Bahauddin Zakariya University, Multan, Punjab, Pakistan
sobasco786@gmail.com
[4] School of Pharmacy, University of Management and Technology Lahore, Punjab, Pakistan. amber.sharif@umt.edu.pk
[5] Professor, Department of Pharmacology, Dr. Ishratul Ibad Institute of Oral Health Sciences, Dow University of Health Sciences, Karachi.
riffat.farooqui@duhs.edu.pk
[6] Department of Biochemistry, NUR International University, Lahore, Pakistan
7006-mbc-f24c@niu.edu.pk
[7] Assistant Professor, Department of Biological Sciences, Cadet College Swat, KPK, Pakistan. ahjawadarf@gmail.com
[8] Department of Botany, University of Sargodha, Sargodha, Punjab, Pakistan
Aafaqalibio@gmail.com
[9] Department of Botany, University of Sargodha, Sargodha, Punjab, Pakistan
abid155yahoo.com
[10] Department of Botany, University of Sargodha, Sargodha, Punjab, Pakistan
jahanzaibrasheedgc@gmail.com
Corresponding author: Sobia Niaz, Department of Botany, Bahauddin Zakariya University, Multan, Punjab, Pakistan.
sobasco786@gmail.com

ABSTRACT: Natural products remain an important source of chemically diverse molecules for anticancer drug discovery, but the large number of compounds reported in natural product databases makes systematic experimental screening difficult. Computational approaches can assist this process by identifying structural patterns associated with biological activity and prioritizing compounds for subsequent investigation. In this study, a machine-learning framework was developed for the computational prioritization of plant-derived compounds associated with cancer related bioactivity using data from NPASS v3.0. Plant-derived records were filtered against cancer associated biological activity, followed by molecular structure standardization, duplicate consolidation, activity label assignment, molecular descriptor calculation, fingerprint generation, and scaffold analysis. The final curated dataset contained 6,120 unique compounds representing 2,145 Bemis Murcko scaffolds, including 2,693 active and 3,427 inactive compounds. Molecular information was represented using 208 physicochemical and topological descriptors together with circular and structural

Natural Resources for Human Health
eISSN: 2583-1194

fingerprints. Random forest, support vector machine, XGBoost, and deep neural network models were evaluated using an 80/10/10 scaffold aware train validation test design. Combining ECFP4 fingerprints with RDKit descriptors consistently improved predictive performance compared with individual representations. The deep neural network achieved the highest reported classification performance, with ROC-AUC of 0.883, PR-AUC of 0.857, accuracy of 0.817, F1-score of 0.782, and MCC of 0.627 on the held-out test set. Quantitative modelling also showed useful predictive performance for pIC_{50}, with the best DNN model obtaining a test-set R^2 of 0.712, RMSE of 0.701, and MAE of 0.524. The scaffold-aware evaluation provides a more conservative estimate of molecular generalization than a conventional random split, although independent external validation remains necessary. Overall, the proposed framework provides a reproducible computational strategy for prioritizing plant-derived phytochemicals for subsequent anticancer investigation.

1. INTRODUCTION

Natural products have historically provided a substantial proportion of bioactive chemical matter used in pharmaceutical research. Their structural complexity, stereochemical diversity, and broad biological activity make plants an especially important source of candidate molecules for therapeutic discovery. At the same time, natural-product collections contain a large number of structurally related compounds and heterogeneous biological measurements, making systematic experimental evaluation expensive and time-consuming. Recent developments in computational chemistry and machine learning have therefore created opportunities to organize natural-product information and prioritize compounds before experimental screening [1–3].

Cancer is a particularly suitable area for computational natural-product research because the biological activity of plant-derived molecules spans numerous mechanisms, including effects on cell proliferation, apoptosis, oxidative stress, signal transduction, and protein-mediated pathways. Computational approaches have increasingly been used to connect natural-product structures with anticancer activity, identify candidate targets, predict molecular interactions, and prioritize compounds for further investigation [2,3]. However, the reliability of such models depends strongly on the quality of the underlying data and on how chemical similarity is handled during model evaluation. Large-scale natural-product databases provide an opportunity to address this limitation. The Natural Product Activity and Species Source database (NPASS) integrates information concerning natural products, their biological activities, and their source organisms. The continued expansion of NPASS has increased its usefulness for data-driven natural-product research and computational drug discovery [4,5]. The current version contains substantially richer quantitative bioactivity and related information than earlier releases, providing an appropriate foundation for systematic machine-learning analysis.

Machine-learning methods have been increasingly applied to natural-product discovery because they can learn nonlinear relationships between molecular structure and biological response. Random forests can capture nonlinear interactions while remaining relatively robust to high-dimensional descriptors, support vector machines are effective in nonlinear classification spaces, and gradient-boosting approaches can identify complex relationships among molecular features. Neural networks provide another route for learning distributed representations from large molecular feature spaces. The comparative performance of these methods, however, depends strongly on the molecular representation and on the independence of the evaluation data [1]. A particular concern in ligand-based machine learning is structural leakage between training and test data. If closely related molecules are randomly divided between the two sets, a model may obtain apparently strong predictive performance because structurally similar analogues occur in both subsets. Such an evaluation does not necessarily demonstrate the ability to generalize to previously unseen chemical frameworks. Scaffold-based partitioning provides a more stringent alternative by assigning compounds according to their molecular frameworks and thereby reducing the presence of closely related analogues across different partitions.

In this study, a data-driven framework was developed for the discovery and prioritization of plant-derived phytochemicals associated with cancer-related bioactivity. The work combines NPASS v3.0 data curation, molecular structure standardization, molecular fingerprints, physicochemical descriptors, Bemis–Murcko scaffold analysis, scaffold-aware data partitioning,

supervised classification, and quantitative structure–activity relationship modelling. Four machine-learning approaches were compared: random forest (RF), support vector machine (SVM), XGBoost, and deep neural network (DNN).

The study was designed around three main objectives: (i) to construct a curated plant-derived cancer-associated compound library from NPASS v3.0; (ii) to determine whether combined molecular fingerprints and conventional molecular descriptors improve predictive performance; and (iii) to evaluate model performance under a scaffold-aware hold-out design that provides a more demanding assessment of chemical generalization. The resulting framework is intended as a prioritization tool for computational screening rather than a replacement for experimental validation.

2. MATERIALS AND METHODS

2.1. Data Source and Dataset Construction

Bioactivity and species information were obtained from NPASS v3.0 [7]. The database was selected because it provides natural-product structures together with biological activity and source-organism information and has been developed specifically to support natural-product discovery and computational analysis [4,5]. The initial NPASS collection contained 135,240 bioactivity records. Records associated with organisms belonging to Kingdom Plantae were retained, while microbial, fungal, animal, and other non-plant sources were excluded. The resulting plant-derived records were subsequently intersected with cancer-associated biological activity. Cancer-associated records were identified using cancer-related targets, human cancer cell lines, and neoplasm-associated biological annotations. Quantitative IC_{50} and GI_{50} measurements were retained where available. The resulting records were subjected to molecular structure standardization and duplicate consolidation before construction of the final modelling dataset. Table 1 summarizes the progressive construction of the dataset.

Table 1. Construction of the NPASS-derived plant–cancer dataset

Dataset Stage / Sub-dataset	Bioactivity Records	Unique Compounds	Target Count	Plant Species
NPASS v3.0 Raw	135,240	35,680	3,240	6,850
Plant-Only	42,180	18,450	1,820	5,120
Cancer-Only	28,340	12,120	480	N/A
Plant × Cancer	14,280	6,840	312	2,140
IC_{50} subset	9,250	4,820	245	1,680
GI_{50} subset	5,030	2,410	115	890
Consolidated SMILES-standardized	11,150	6,120	298	1,980

The table shows the progressive reduction from the complete NPASS activity collection to the final plant-derived cancer associated molecular library. The final modelling library contained 6,120 standardized unique compounds.

2.2. Identification of Cancer-Associated Bioactivity

Cancer-associated activity was defined using biological annotations corresponding to cancer-related targets, cancer cell-line experiments, and neoplasm-associated pathways. The filtering procedure was applied before structure-level consolidation so that compounds represented in multiple cancer-related records could be retained and subsequently aggregated.

IC_{50} and GI_{50} measurements were considered the principal quantitative endpoints. Where several measurements were associated with the same standardized compound, the values were examined for consistency before consolidation.

2.3. Molecular Structure Standardization

Molecular structures were standardized before descriptor and fingerprint calculation. Inorganic, organometallic, and unsuitable molecular records were removed. Salt and counterion components were separated where appropriate, and the largest chemically meaningful fragment was retained. Structures were subjected to valence and sanitization checks. Explicit hydrogen handling was standardized before generation of canonical SMILES. Tautomeric and valence inconsistencies were

examined during quality control. InChIKeys were used to identify duplicate molecular structures. Where multiple quantitative activity measurements remained for the same standardized compound, conflicting observations were examined. Measurements showing substantial disagreement were excluded when the standard deviation in log-transformed activity exceeded 0.5 log units; otherwise, compatible measurements were aggregated. Table 2 summarizes the structure-curation process.

Table 2. Molecular structure curation and consolidation

Curation Stage	Removed	Retained
Raw ingestion	0	6,840
Inorganic/organometallic compounds	142	6,698
Salt/counterion removal	210	6,488
Valence/sanitization failures	85	6,403
Charge/tautomer quality control	0	6,403
InChIKey duplicate aggregation	215	6,188
Activity conflict resolution	68	6,120

The cumulative reduction from 6,840 raw plant–cancer compounds to 6,120 final compounds was 720 compounds. The 68 compounds shown in the final row represent only the compounds removed during the activity-conflict step and should therefore not be interpreted as the total number removed during curation.

2.4. Activity Definition and Label Assignment

For classification, compounds were categorized according to their quantitative activity. A compound was considered active when an IC_{50} or GI_{50} value was $\leq 10\ \mu$M. Measurements above this threshold were assigned to the inactive class. For quantitative modelling, IC_{50} and GI_{50} values expressed in micromolar units were converted to molar concentration before logarithmic transformation:

$$pIC_{50} = -10(IC_{50})^{-6}$$

and

$$GI_{50} = -10(GI_{50})^{-6}$$

The activity distribution is shown in Table 3.

Table 3. Activity distribution in the curated dataset

Dataset	**Active**	**Inactive**	**Total**	**Active Ratio**
IC_{50}	2,072	2,748	4,820	42.99%
GI_{50}	1,012	1,398	2,410	41.99%
Combined standardized	2,693	3,427	6,120	44.00%
High potency ($\leq 1\ \mu$M)	1,105	5,015	6,120	18.06%

The final classification dataset was moderately imbalanced, with approximately 44% active and 56% inactive compounds. This imbalance was incorporated into model training through appropriate class-weighting or imbalance-aware learning procedures where applicable.

2.5. Molecular Representation

Two complementary classes of molecular information were used: continuous molecular descriptors and binary structural fingerprints. A total of 208 two-dimensional descriptors were calculated using RDKit. These descriptors represented physicochemical, constitutional, topological, electronic-state, and surface-area properties. Zero-variance features were removed before modelling, and continuous descriptors were standardized using parameters derived from the training data only.

Table 4. Molecular descriptor groups

Descriptor Group	Number of Features	Representative Features	Range	Mean ± SD
Physicochemical	18	MW, LogP, TPSA, LabuteASA	118.15–1248.35	412.45 ± 142.18
H-bonding	8	HBD, HBA, NumRotatableBonds	0–18	4.25 ± 2.85
Ring systems	14	Aromatic/aliphatic ring descriptors	0–9	3.12 ± 1.45
Topological/constitutional	42	Chi0n–Chi4path, HallKierAlpha	−2.85–18.42	4.18 ± 2.11
E-State	86	MaxAbsEStateIndex, MinEStateIndex	−3.45–15.82	6.24 ± 3.12
VSA/surface	40	PEOE_VSA1–14, SlogP_VSA1–12	0–285.40	45.20 ± 28.60
Total	**208**	—	—	—

Circular fingerprints were generated using ECFP4/Morgan fingerprints with radius 2 at both 1,024- and 2,048-bit lengths. ECFP6 fingerprints with radius 3 were also evaluated. MACCS keys and RDKit topological fingerprints were included for comparison.

Table 5. Molecular fingerprint representations

Representation	Size	Radius / Path	Mean Nonzero Bits	Density
ECFP4	1,024	Radius 2	48.2	4.71%
ECFP4	2,048	Radius 2	52.6	2.57%
ECFP6	2,048	Radius 3	78.4	3.83%
MACCS	166	—	38.5	23.19%
RDKit path fingerprint	2,048	Path 1–7	112.3	5.48%

The combined representation used in the principal model comparisons consisted of the 2,048-bit ECFP4 fingerprint together with the 208 standardized RDKit descriptors.

2.6. Scaffold Analysis

Chemical diversity was assessed using Bemis–Murcko molecular scaffolds [6]. The scaffold representation removes peripheral substituents while retaining the principal molecular framework, allowing compounds to be grouped according to shared structural architecture. The standardized dataset contained 2,145 unique scaffolds, of which 1,280 were represented by single compounds. Scaffold counts were also calculated for the major dataset subsets.

Table 6. Scaffold diversity across dataset subsets

Dataset	Compounds	Scaffolds	Singleton Scaffolds	Diversity Index	Top Scaffold Coverage
Plant-only	18,450	5,420	3,110	0.294	4.85%
Cancer-only	12,120	3,890	2,150	0.321	5.12%
Plant × Cancer	6,840	2,340	1,380	0.342	3.95%
Standardized	6,120	2,145	1,280	0.350	3.68%
Active	2,693	1,085	625	0.403	2.84%
Inactive	3,427	1,312	755	0.383	4.12%

The relatively large number of singleton scaffolds indicates substantial structural diversity in the curated collection. The active subset also exhibited a comparatively high scaffold diversity value, suggesting that active compounds were not concentrated within only a small number of molecular frameworks.

2.7. Scaffold Aware Dataset Partitioning

The dataset was divided into training, validation, and test subsets using an 80/10/10 scaffold-aware design. The purpose was to reduce the possibility that highly similar molecular frameworks would be distributed across the development and evaluation sets.

Table 7. Scaffold aware train validation test partition

Set	Proportion	Compounds	Active	Inactive	Scaffolds
Training	80%	4,896	2,154 (44.0%)	2,742 (56.0%)	1,716
Validation	10%	612	269 (43.9%)	343 (56.1%)	214
Test	10%	612	270 (44.1%)	342 (55.9%)	215
Total	**100%**	**6,120**	**2,693 (44.0%)**	**3,427 (56.0%)**	**2,145**

The class proportions remained highly similar across the three subsets. More importantly, the partitioning was performed with scaffold information rather than treating individual molecules as independent random observations. This distinction is important because a conventional random molecular split can place close analogues in both training and test sets, allowing models to benefit from local structural similarity. A scaffold-aware evaluation is more demanding because prediction must extend to molecular frameworks that are less directly represented in the training data. The present results should therefore be interpreted as evidence of performance under a scaffold-aware internal hold-out design rather than as evidence of fully independent external validation. The test compounds originate from the same NPASS-derived collection, and independent datasets containing previously unseen compounds and scaffolds would provide a stronger test of prospective generalization.

2.8. Machine-Learning Models

Four machine learning approaches were evaluated.

Random Forest

The random forest model consisted of 500 decision trees. Maximum tree depth was optimized over 10, 20, 30, and unrestricted depth. Gini impurity was used for node splitting, and class-balanced learning was applied to reduce the influence of the moderate class imbalance.

Support Vector Machine

An RBF-kernel support vector machine was trained using a logarithmic search over C values ranging from 10^{-2} to 10^{3} and gamma values ranging from 10^{-4} to 1.

XGBoost

The XGBoost classifier used 300 boosting trees with a learning rate of 0.05, maximum depth values of 3, 6, and 9, subsampling of 0.8, and column subsampling of 0.8.

Deep Neural Network

The DNN architecture consisted of an input layer followed by fully connected layers of 512, 256, and 128 neurons and a single output node. ReLU activation, batch normalization, and dropout of 0.3 were used. The model was trained with Adam using a learning rate of 10^{-3}, weight decay of 10^{-5}, binary cross-entropy loss, and a maximum of 150 epochs. Early stopping with a patience of 15 epochs was applied. For the combined representation, continuous descriptors and binary fingerprints were handled separately during preprocessing. Descriptor standardization was fitted using training data only to avoid information leakage.

2.9. Classification Performance Evaluation

Classification performance was assessed using accuracy, sensitivity, specificity, F1-score, Matthews correlation coefficient (MCC), ROC-AUC, and PR-AUC. ROC-AUC was used to measure ranking ability across classification thresholds, while PR-AUC provided additional information in the presence of class imbalance. MCC was included because it summarizes the four elements of the confusion matrix and is informative when the class distribution is not perfectly balanced.

2.10. Quantitative Structure–Activity Relationship Modelling

Quantitative modelling was conducted separately for pIC_{50} and pGI_{50}. Random forest, XGBoost, and DNN models were evaluated for continuous activity prediction. Performance was assessed using the coefficient of determination (R^2), root mean squared error (RMSE), mean absolute error (MAE), and Q^2_{F3}. The use of a held-out test set was maintained for final evaluation, while model and preprocessing decisions were determined without using test-set outcomes.

2.11. Reproducibility and Leakage Control

To reduce information leakage, molecular descriptors were standardized using training-set statistics. Fingerprint generation was performed directly from molecular structures, and test-set information was not used for hyperparameter optimization. Scaffold information was considered during data partitioning rather than after model development. The same final test set was retained for comparison of the principal models. Consequently, differences in reported test performance reflect the combination of molecular representation and model architecture under the same evaluation setting.

3. RESULTS AND DISCUSSION

3.1. Dataset Construction and Composition

The initial NPASS v3.0 [7] collection contained 135,240 bioactivity records. Restricting the database to plant-derived records reduced the dataset to 42,180 bioactivity observations representing 18,450 unique compounds. Subsequent cancer-related filtering yielded 14,280 records associated with 6,840 unique compounds. The quantitative endpoint distribution consisted of 9,250 IC_{50} records and 5,030 GI_{50} records. After structure standardization, duplicate aggregation, and activity-conflict filtering, 6,120 unique compounds remained for machine-learning analysis. The reduction in compound number during curation reflects the importance of structure-level quality control in large natural-product datasets. Natural products may be represented under multiple names, salt forms, stereochemical representations, or records originating from different experiments. Without standardization, the same chemical entity can be counted multiple times and can potentially appear in both training and testing subsets. The resulting dataset therefore provides a substantially cleaner molecular basis for supervised learning than the raw activity records.

3.2. Structure Standardization and Activity Distribution

The final dataset contained 2,693 active compounds and 3,427 inactive compounds, corresponding to an active proportion of approximately 44.0%. The IC_{50} and GI_{50} subsets showed similar activity proportions, at 42.99% and 41.99%, respectively. The similarity between the two endpoint-specific class distributions suggests that the overall classification task was not dominated by one particular quantitative assay type. Nevertheless, the measurements originate from heterogeneous biological experiments and should not be interpreted as directly interchangeable measurements of clinical efficacy. A subset of 1,105 compounds had activity at or below 1 μM. These highly potent observations represented approximately 18.06% of the final library, indicating that the database contained a substantial group of compounds potentially suitable for further prioritization.

3.3. Molecular Descriptor Profile

The 208 calculated descriptors covered multiple dimensions of molecular structure. Physicochemical descriptors captured molecular size, lipophilicity, polar surface characteristics, and related properties, whereas topological and E-state descriptors provided additional information about molecular connectivity and electronic environments. The descriptor distributions also illustrate the broad chemical range of the dataset. Molecular weight ranged from 118.15 to 1,248.35, while the reported physicochemical descriptor group had a mean molecular weight of 412.45 with a standard deviation of 142.18. Such diversity is characteristic of natural-product collections and presents both an opportunity and a challenge for machine learning. A broad chemical space can improve the potential usefulness of a model for screening, but it also increases the risk that a model will encounter structures substantially different from those used during training.

3.4. Fingerprint Characteristics

The fingerprint comparison showed that ECFP4 representations were relatively sparse, particularly at 2,048 bits. The 2,048-bit ECFP4 representation had a mean of 52.6 nonzero bits and a density of 2.57%. MACCS keys were considerably denser because of their shorter fixed representation. The sparse ECFP4 representation is particularly useful for capturing local molecular environments. In contrast, conventional descriptors encode global physicochemical and topological characteristics. The complementary nature of these representations motivated the use of a combined fingerprint–descriptor representation.

3.5. Chemical Scaffold Diversity

The final dataset contained 2,145 unique Bemis–Murcko scaffolds, including 1,280 singleton scaffolds. The standardized compound collection had a scaffold diversity value of 0.350, while the active subset showed a value of 0.403. These values indicate that cancer-associated activity was distributed across a chemically diverse collection rather than being concentrated exclusively in a few dominant frameworks. The relatively low coverage of the most common scaffold in the standardized dataset (3.68%) further supports this observation. The scaffold diversity also provides an important context for interpreting the machine-learning results. A model that performs well across such a chemically heterogeneous collection must identify patterns that extend beyond a small set of highly represented structural classes.

3.6. Scaffold-Aware Evaluation and Generalization

The scaffold-aware partition produced 4,896 training compounds, 612 validation compounds, and 612 test compounds. The active fractions were highly similar across the three partitions, ranging from 43.9% to 44.1%. The central advantage of this evaluation strategy is that compounds were assigned with scaffold information taken into account. This reduces the risk that close molecular analogues will be distributed arbitrarily between the training and test subsets. The approach therefore provides a more demanding assessment than a simple random molecular split. However, it should not be interpreted as equivalent to external validation. All three subsets originated from the NPASS-derived dataset, and the test compounds may still share biological and experimental characteristics with the development data. Under this evaluation, the combined ECFP4-2048 and RDKit descriptor representation produced ROC-AUC values of 0.852, 0.858, 0.874, and 0.883 for RF, SVM, XGBoost, and DNN, respectively. The corresponding MCC values increased from 0.559 to 0.573, 0.610, and 0.627. The consistency across several evaluation metrics indicates that the improvement of the DNN over the other models was not restricted to a single performance measure. Nevertheless, the results are best described as **good predictive performance under scaffold-aware internal evaluation**, rather than definitive evidence of broad chemical-space generalization.

3.7. Random Forest Classification

The random forest results are presented in Table 8.

Table 8. Random forest performance for different molecular representations

Representation	ROC-AUC	PR-AUC	Accuracy	F1	MCC	Sensitivity	Specificity
RDKit 208	0.812	0.774	0.748	0.702	0.485	0.685	0.798
MACCS 166	0.785	0.741	0.721	0.668	0.428	0.648	0.778
ECFP4 1024	0.834	0.801	0.765	0.724	0.520	0.711	0.807
ECFP4 2048	0.841	0.810	0.773	0.732	0.536	0.719	0.816
RDKit FP 2048	0.828	0.792	0.758	0.715	0.506	0.700	0.804
Combined ECFP4 2048 + RDKit 208	**0.852**	**0.823**	**0.784**	**0.746**	**0.559**	**0.733**	**0.825**

The random forest model performed better with molecular fingerprints than with conventional descriptors alone. ECFP4-2048 improved ROC-AUC from 0.812 for the RDKit descriptor representation to 0.841. The strongest performance was obtained when ECFP4-2048 and RDKit descriptors were combined. ROC-AUC increased to 0.852 and MCC to 0.559. This improvement supports the view that local structural information and global molecular properties provide complementary information for activity prediction.

3.8. Support Vector Machine Classification

Table 9. Support vector machine performance

Representation	ROC-AUC	PR-AUC	Accuracy	F1	MCC	Sensitivity	Specificity
RDKit 208	0.825	0.788	0.758	0.714	0.505	0.696	0.807
MACCS 166	0.792	0.752	0.730	0.680	0.448	0.663	0.784
ECFP4 1024	0.842	0.811	0.776	0.735	0.542	0.722	0.819
ECFP4 2048	0.848	0.818	0.781	0.741	0.553	0.730	0.822
RDKit FP 2048	0.835	0.801	0.768	0.726	0.526	0.711	0.813
Combined ECFP4 2048 + RDKit 208	**0.858**	**0.830**	**0.791**	**0.753**	**0.573**	**0.741**	**0.830**

The SVM model showed a similar pattern. ECFP4 representations outperformed MACCS and the descriptor-only representation, while the combined representation achieved the best results. The combined model obtained ROC-AUC of 0.858, PR-AUC of 0.830, and MCC of 0.573. These results were slightly higher than the corresponding random forest values, suggesting that the nonlinear RBF kernel was able to exploit relationships in the combined molecular feature space effectively.

3.9. XGBoost Classification

Table 10. XGBoost performance

Representation	ROC-AUC	PR-AUC	Accuracy	F1	MCC	Sensitivity	Specificity
RDKit 208	0.838	0.802	0.771	0.730	0.532	0.715	0.816
MACCS 166	0.805	0.765	0.742	0.694	0.472	0.678	0.792
ECFP4 1024	0.855	0.824	0.788	0.748	0.567	0.733	0.830

ECFP4 2048	0.862	0.832	0.796	0.758	0.583	0.744	0.836
RDKit FP 2048	0.849	0.815	0.781	0.740	0.552	0.726	0.825
Combined ECFP4 2048 + RDKit 208	**0.874**	**0.848**	**0.809**	**0.773**	**0.610**	**0.759**	**0.848**

XGBoost produced a further improvement over RF and SVM. The combined representation achieved ROC-AUC of 0.874 and PR-AUC of 0.848. The MCC of 0.610 indicates a stronger overall correspondence between predicted and observed classes than obtained with the preceding models. The improvement was accompanied by increases in both sensitivity and specificity, suggesting that the performance gain was not simply a consequence of favoring the majority class.

3.10. Deep Neural Network Classification

Table 11. Deep neural network performance

Representation	ROC-AUC	PR-AUC	Accuracy	F1	MCC	Sensitivity	Specificity
RDKit 208	0.842	0.808	0.776	0.736	0.543	0.722	0.819
MACCS 166	0.812	0.772	0.748	0.701	0.485	0.685	0.798
ECFP4 1024	0.861	0.830	0.794	0.755	0.579	0.741	0.836
ECFP4 2048	0.870	0.841	0.802	0.765	0.596	0.752	0.842
RDKit FP 2048	0.856	0.823	0.789	0.749	0.568	0.733	0.833
Combined ECFP4 2048 + RDKit 208	**0.883**	**0.857**	**0.817**	**0.782**	**0.627**	**0.770**	**0.854**

The DNN achieved the strongest overall classification performance. Using the combined molecular representation, ROC-AUC reached 0.883, while PR-AUC was 0.857. The accuracy of 0.817 and MCC of 0.627 also indicate that the model maintained useful performance when both classes were considered simultaneously. Sensitivity of 0.770 and specificity of 0.854 show that the model was capable of identifying active compounds while retaining relatively strong discrimination of inactive compounds. The performance improvement over ECFP4 alone suggests that conventional molecular descriptors contributed information not fully encoded by the fingerprint representation. This is consistent with the interpretation that local substructural patterns and global physicochemical characteristics provide complementary predictive signals.

3.11. Comparison of the Best Models

Table 12. Comparison of the best combined-representation models

Model	ROC-AUC	PR-AUC	Accuracy	F1	MCC	Sensitivity	Specificity
RF	0.852	0.823	0.784	0.746	0.559	0.733	0.825
SVM	0.858	0.830	0.791	0.753	0.573	0.741	0.830
XGBoost	0.874	0.848	0.809	0.773	0.610	0.759	0.848
DNN	**0.883**	**0.857**	**0.817**	**0.782**	**0.627**	**0.770**	**0.854**

The comparison demonstrates a consistent ranking across the principal metrics, with DNN achieving the highest values followed by XGBoost, SVM, and RF. The progression from RF to DNN suggests that the ability to learn nonlinear relationships becomes increasingly important as the feature representation combines sparse structural fingerprints with continuous physicochemical descriptors. However, the differences should not be interpreted as evidence that DNN is universally superior. Model selection remains dependent on dataset size, representation, reproducibility, interpretability, and external validation. The performance of XGBoost is particularly noteworthy because it achieved ROC-AUC of 0.874 and

MCC of 0.610 while using a considerably simpler learning architecture than the DNN. For practical deployment, this trade-off between predictive performance, computational complexity, and interpretability may be relevant.

3.12. Quantitative Activity Prediction

The classification analysis was complemented by quantitative modelling of pIC_{50} and pGI_{50}.

Table 13. Quantitative structure–activity relationship performance

Endpoint	Model	Train R^2	Test R^2	RMSE	MAE	Q^2_{F3}
pIC_{50}	RF	0.882	0.642	0.782	0.598	0.638
pIC_{50}	XGBoost	0.915	0.685	0.734	0.552	0.680
pIC_{50}	DNN	0.932	**0.712**	**0.701**	**0.524**	**0.708**
pGI_{50}	RF	0.865	0.618	0.812	0.621	0.612
pGI_{50}	XGBoost	0.898	0.661	0.765	0.581	0.655
pGI_{50}	DNN	0.918	—	—	—	—

For pIC_{50}, the DNN produced the highest test-set R^2 of 0.712, accompanied by RMSE of 0.701 and MAE of 0.524. Its Q^2_{F3} value of 0.708 was also higher than those obtained by RF and XGBoost. XGBoost achieved an intermediate test R^2 of 0.685, whereas RF produced 0.642. The gradual improvement across the three models is consistent with the classification results. The training and test R^2 values also reveal a reduction in performance when moving from model development to held-out prediction. This difference is expected in a chemically diverse dataset and emphasizes why training performance alone should not be used to judge model quality. For pGI_{50}, RF and XGBoost produced test R^2 values of 0.618 and 0.661, respectively. The DNN achieved a training R^2 of 0.918; however, the remaining DNN test statistics were not available in the established results and are therefore not reported here. They should be calculated from the held-out predictions before the final manuscript is submitted. Overall, the quantitative results support the use of machine learning for activity estimation but also indicate that predictive uncertainty and applicability domain should be considered when ranking compounds for experimental investigation.

3.13. Observations

The results demonstrate three consistent observations.

First, molecular representation strongly influenced model performance. MACCS keys and descriptor-only representations generally produced lower performance than ECFP4 fingerprints. This indicates that local molecular environments captured by circular fingerprints are particularly informative for distinguishing active and inactive natural products. **Second**, combining ECFP4-2048 with 208 RDKit descriptors consistently improved performance across all four classification algorithms. The improvement suggests that molecular fingerprints and physicochemical descriptors capture complementary aspects of chemical information. **Third,** the ranking of the four models remained consistent under the same held-out evaluation. DNN achieved the highest performance, followed by XGBoost, SVM, and RF. The DNN's ROC-AUC of 0.883 and MCC of 0.627 represent the strongest reported classification performance in the present analysis.

Importantly, the use of scaffold-aware partitioning makes these values more informative than results obtained from an ordinary random split, but they should not be interpreted as equivalent to independent prospective validation. The dataset remains derived from NPASS and therefore retains the biological and experimental biases of its source. The framework is consequently best viewed as a **compound prioritization system**. Its purpose is to reduce the size of a large natural-product search space and identify candidates worthy of additional investigation. Experimental confirmation remains necessary because computationally predicted activity does not establish pharmacological efficacy, selectivity, safety, bioavailability, or therapeutic potential. For future development, the framework could be extended with applicability-domain estimation, probability

calibration, repeated scaffold splits, Y-randomization tests, uncertainty quantification, and explainable machine-learning analyses. Integration with ADME and toxicity prediction, molecular docking, molecular dynamics, and experimental screening would further improve the translational value of the computational pipeline.

4. CONCLUSION

This study presented a data-driven computational framework for the discovery and prioritization of anticancer-associated phytochemicals from medicinal plants by integrating curated natural-product bioactivity data with molecular representations and machine-learning models. The analysis was based on NPASS v3.0 and focused specifically on plant-derived compounds associated with cancer-related bioactivity. Following structure standardization, duplicate handling, activity-label assignment, and scaffold analysis, a final library of 6,120 unique compounds representing 2,145 molecular scaffolds was obtained. The resulting dataset contained 2,693 active and 3,427 inactive compounds, providing a sufficiently diverse chemical basis for supervised learning. The comparison of molecular representations demonstrated that combining circular fingerprints with conventional molecular descriptors provided a more informative representation of the structural and physicochemical characteristics associated with cancer-related activity. Among the evaluated models, the deep neural network using the combined ECFP4 and RDKit descriptor representation achieved the highest overall classification performance, with a ROC-AUC of 0.883, PR-AUC of 0.857, accuracy of 0.817, F1-score of 0.782, and MCC of 0.627 on the held-out test set. XGBoost provided the second strongest performance, while SVM and random forest showed comparatively lower but still meaningful predictive ability. These results indicate that nonlinear learning models can capture complementary relationships between molecular substructures and physicochemical properties that are not fully represented by either descriptor family alone. The quantitative analysis further supported the usefulness of the computational framework for activity estimation. The best-performing DNN model achieved a test-set R^2 of 0.712 for pIC_{50} prediction, with RMSE and MAE values of 0.701 and 0.524, respectively. XGBoost also produced competitive quantitative predictions, whereas random forest showed comparatively weaker extrapolation to structurally distinct compounds. The scaffold-aware evaluation is particularly important because conventional random splitting can produce overly optimistic estimates when closely related chemical structures occur simultaneously in training and testing subsets. The present results therefore provide a more conservative assessment of model performance. From an application perspective, the framework can be used as a prioritization tool rather than as a substitute for experimental validation. Compounds predicted to have high anticancer-associated activity can be ranked for subsequent biochemical, cellular, pharmacological, and toxicity evaluation. Such prioritization is especially relevant for natural-product libraries, where structural diversity and heterogeneous biological measurements make experimental screening expensive and time-consuming. Integration of plant taxonomy, molecular structure, bioactivity, and machine-learning predictions may consequently provide a practical route for narrowing large natural-product collections to experimentally tractable candidate sets. Several limitations should nevertheless be acknowledged. NPASS contains measurements obtained from different experimental systems, biological targets, laboratories, and assay conditions, and these sources of heterogeneity cannot be completely removed through computational standardization. In addition, activity against cancer cell lines does not necessarily establish a compound's therapeutic anticancer potential. The present models should therefore be interpreted as predictors of the curated cancer-associated activity represented in the dataset. External validation using independent experimental datasets, prospective testing, applicability-domain analysis, model calibration, and experimental confirmation of prioritized compounds would further strengthen the evidence for practical deployment. Overall, the study demonstrates that combining carefully curated natural-product data, scaffold-aware evaluation, molecular fingerprints, physicochemical descriptors, and machine-learning algorithms can provide a reproducible computational strategy for anticancer phytochemical prioritization. The framework establishes a foundation for future studies incorporating explainable machine learning, molecular docking, molecular dynamics, toxicity prediction, ADME profiling, and prospective experimental validation to move promising plant-derived compounds from computational prioritization toward evidence-based drug discovery.

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(11), 895–916. https://doi.org/10.1038/s41573-023-00774-7

[2] Das, A. P., & Agarwal, S. M. (2024). Recent advances in the area of plant-based anti-cancer drug discovery using computational approaches. *Molecular Diversity*, 28(2), 901–925. https://doi.org/10.1007/s11030-022-10590-7

[3] Chunarkar-Patil, P., Kaleem, M., Mishra, R., et al. (2024). Anticancer drug discovery based on natural products: From computational approaches to clinical studies. *Biomedicines*, 12(1), 201. https://doi.org/10.3390/biomedicines12010201

[4] Lin, H., Hou, D., Jiang, X., Yin, R., Yu, S., Lu, S., Wang, S., Ju, D., Zhao, H., Chen, Y., & Zeng, X. (2026). NPASS database update 2026: Comprehensive quantitative composition, bioactivity, and ADME-Tox data of natural products for biomedical research. *Nucleic Acids Research*, 54(D1), D1519–D1527. https://doi.org/10.1093/nar/gkaf1196

[5] Zeng, J., Zhang, X., Han, X., et al. (2018). NPASS: Natural product activity and species source database for natural product research, discovery and tool development. *Nucleic Acids Research*, 46(D1), D1213–D1219.

[6] Bemis, G. W., & Murcko, M. A. (1996). The properties of known drugs. 1. Molecular frameworks. *Journal of Medicinal Chemistry*, 39(15), 2887–2893.

[7] Lin, H., Hou, D., Jiang, X., Yin, R., Yu, S., Lu, S., Wang, S., Ju, D., Zhao, H., Chen, Y., & Zeng, X. (2026). NPASS database update 2026: Comprehensive quantitative composition, bioactivity, and ADME-Tox data of natural products for biomedical research. *Nucleic Acids Research, 54*(D1), D1519–D1527. https://doi.org/10.1093/nar/gkaf1196