Natural Resources for Human Health

visagaa publishing house

Review

View Article online

Check for updates

Received 18 March 2026
Revised 19 May 2026
Accepted 16 June 2026
Available Online 07 September 2026

Edited by Mohmed Isaqali Karobari

KEYWORDS:
Machine learning
Natural products
Bioactivity prediction
Validation practices

Natr Resour Human Health 2026; 6 (4): 659–675
https://doi.org/10.53365/nrfhh/224641
eISSN: 2583-1194

AI/ML Based Prediction of Experimentally Derived Bioactivity in Natural Products: A Systematic Review of Models, Validation Practices, and Performance Reporting

Shashikanth Kharat[1,A-F], Shubham Kumar[1,B-D], Toufiq Noor[2,C,E-F], Ankita Mathur[2,C,E-F], Vini Mehta[2,A,C,E-F]

[1]Global Research Cell, Dr. D. Y. Patil Dental College & Hospital, Dr. D. Y. Patil Vidyapeeth (Deemed to be University), India
[2]Dental Research Cell, Dr. D. Y. Patil Dental College & Hospital, Dr. D. Y. Patil Vidyapeeth (Deemed to be University), India

A – Research concept and design, B – Collection and/or assembly of data, C – Data analysis and interpretation, D – Writing the article, E - Critical revision of the article, F – Final approval of the article

ABSTRACT: Researchers increasingly use AI/ML to predict bioactivity of natural products from molecular structure, but datasets, validation designs, and reporting practices vary widely across studies, making it hard to judge how trustworthy reported performance is? We conducted a systematic review of scholarly databases from 2015 to 2026 to identify AI/ML studies that predict experimentally measured bioactivity of natural products. Two reviewers screened records, extracted data, and applied an AI focused risk of bias framework covering dataset size, data leakage, validation design, applicability domain assessment, and reproducibility. Model performance was summarised using median AUC and related metrics from held out, external, or prospective validation, grouped by validation tier, dataset size, and model family. We included 16 studies (52 predictive models) among 2,705 records, spanning diverse natural product sources and pharmacological targets most models used classical machine learning. while deep learning and graph based methods were mainly used for larger or more complex datasets. Feature selection performed before data splitting and the absence of external validation were judged high risk in 56.3% studies. Models trained on small datasets with fewer than 100 compounds showed the highest median best AUC (0.97). whereas studies in higher validation tiers (Tier 4-5) often reported very high median best AUC (0.98-0.999) despite limited robust external testing, suggesting performance inflation in some settings. Overall, our findings indicate that improving validation design, explicitly analysing applicability domains, and routinely sharing code and data are likely to be more important for reliable natural product bioactivity prediction than further escalation in model complexity.

* Corresponding author.
E-mail address: vmehta@statsense.in (Vini Mehta)

1. INTRODUCTION

Natural products have provided many clinically useful drugs and leads. Their chemical diversity and biological activity have been explored in areas such as oncology, infectious disease, and metabolic disorders. Their evolutionarily refined bioactivity profiles and chemical complexity continue to inspire lead discovery in oncology, infectious disease, and metabolic disorders [1,2]. Despite this enduring promise, the systematic identification of bioactive natural products faces substantial logistical constraints. Traditional experimental screening approaches whether high-throughput in vitro assays or resource-intensive in vivo models are limited by the sheer scale and structural heterogeneity of natural product space, compounded by challenges in compound isolation, purification, and assay standardization [2,3]. These bottlenecks restrict the efficient translation of natural product diversity into validated leads.

The emergence of artificial intelligence (AI) and machine learning (ML) has provided new ways to predict biological activity directly from molecular structure and to prioritise compounds for experimental testing [4,5]. A wide range of modelling approaches has been explored over the past decade. These include classical quantitative structure–activity relationship (QSAR) models, ensemble methods such as random forests and gradient boosting, deep neural networks, and graph-based architectures such as graph neural networks (GNNs) [6] and self-attentive message-passing networks [3,7–10]. More recently, transformer-based encoders have been adapted to chemical representation, further broadening the set of tools available for this type of work [3]. Together, these approaches aim to speed up the discovery of bioactive natural products by using larger datasets and by capturing structure–activity patterns that are difficult to recognise with traditional mechanistic methods.

At the same time, the growing number of AI and ML studies in natural product bioactivity has revealed a lot of methodological variation. Studies draw on very different data sources, ranging from large public repositories such as ChEMBL and PubChem to smaller, manually curated databases focused on natural products. They often use different rules for cleaning and combining these data [1,3]. The quality, standardisation, and representativeness of the underlying datasets strongly affect model performance, yet details of data curation, assay harmonisation, and handling of heterogeneous assays are not always reported in a consistent way [1,3]. Choices about how molecules are represented whether by fingerprints and descriptors, graph-based encodings, or learned embeddings also shape model behaviour, and authors do not always explain why a particular representation was chosen [7–9]. Validation design remains a major concern. Many studies rely only on internal cross-validation without a clearly independent test set, make limited use of scaffold-based splits to reduce data leakage [11], or omit a formal applicability-domain analysis, all of which can inflate reported performance and reduce reproducibility [1,9,12]. In addition, information on code availability, hyperparameter settings, and the exact performance metrics used is sometimes incomplete, which makes it harder to compare models and to reproduce published results.

Given the rapid growth of AI-driven work on natural product bioactivity and the clear variation we see in how models and validation schemes are implemented. There is a need to bring the existing evidence together in a structured way [13]. In this systematic review, we aim to identify and describe AI and ML models that predict experimentally measured bioactivity of natural products, with a particular focus on dataset sources, molecular representations, validation strategies, and performance reporting practices. By summarising these aspects across published studies, we aim to map the current landscape, point out good practices and common gaps, and provide practical guidance for future model development and for linking computational predictions to experimental work in natural product drug discovery. However, prior overviews of AI and ML in natural product research have rarely examined, in a systematic way, how validation strategies and performance metrics are implemented and reported in primary modelling studies. This review directly addresses that gap by pairing model-level data extraction with an AI-focused risk-of-bias framework to evaluate validation rigour, applicability-domain assessment, and performance reporting.

2. METHODS

2.1. Study design and reporting standards

This systematic review was conducted in accordance with the PRISMA 2020 guidelines (Supplementary Table S6) [14]. Because of substantial anticipated heterogeneity in model architectures, molecular representations, validation strategies, and outcome metrics, we prespecified a structured narrative synthesis with stratified quantitative summaries rather than a formal meta-analysis. During protocol development, we discussed alternative quantitative options (e.g., meta-regression, subgroup syntheses) but judged them inappropriate given the diversity and non-comparable reporting of performance metrics across studies.

2.2. Research question

The review addressed the following question: 1) What artificial intelligence (AI) and machine learning (ML) models

have been applied to predict the bioactivity of natural products. 2) what molecular representations and validation strategies do they employ. and 3) how does reported predictive performance vary according to validation rigor and dataset characteristics?

2.3. Eligibility criteria

Eligibility criteria were defined using a modified PICOS framework adapted for computational predictive modelling studies. Studies were included if they: (1) investigated natural products, including plant-, marine-, or microbial-derived compounds or their metabolites (2) applied AI/ML/deep learning-based predictive models (e.g. random forest, support vector machine, neural networks, graph neural networks) (3) predicted experimentally derived bioactivity endpoints, such as IC_{50}, EC_{50}, Ki, GI_{50}, or binary activity classifications (4) reported quantitative predictive performance metrics (e.g. AUC/ROC, accuracy, F1-score, MCC, RMSE, R^2) and (5) described a validation strategy (e.g. cross-validation, internal test split, independent external dataset, or prospective validation).

"Experimentally derived bioactivity" was defined as activity labels originating from validated in vitro or in vivo assays, either generated in-house or obtained from curated databases such as ChEMBL, BindingDB, or published literature. We included original research articles published in English between January 2014 and March 2026 and excluded reviews, editorials, commentaries, conference abstracts without full methodological information, studies relying solely on molecular docking without an ML component, work not focused on natural product chemical space, and studies whose predicted endpoints were unrelated to bioactivity (e.g. drug-likeness only).

2.4. Protocol and deviations

A detailed protocol prespecifying the review question, eligibility criteria, data extraction items, risk-of-bias domains, and analysis plan was drafted before screening commenced and is available in Supplementary Appendix A. Because this review does not involve primary data from human or animal participants and focuses on methodological features of artificial intelligence and machine learning models, it did not meet the eligibility criteria for registration in PROSPERO, which currently covers reviews of human health and wellbeing and animal pre-clinical studies with a direct link to human health. No substantive deviations from the original protocol were made. Minor clarifications included refining the wording of the AI-specific risk-of-bias domains and adding a descriptive stratification of performance by dataset size and model family, both of which were consistent with the original analytic objectives.

2.5. Information sources and search strategy

A systematic search was conducted in PubMed/MEDLINE, Scopus, and Web of Science Core Collection. The final search was performed on 12 January 2026 and was limited to title, abstract, and keyword fields. Search strategies combined three concept blocks:

Concept block	Example search terms (truncated)
AI/ML terminology	"Artificial intelligence", "machine learning", "deep learning"
Natural product terms	"Natural product", "phytochemical", "marine", "microbial"
Bioactivity prediction	"bioactivity", "activity prediction", "pharmacological activity"

Briefly, our search strategy combined terms related to natural products (e.g., 'natural product*', 'phytochemical*', 'herbal', 'plant-derived'), bioactivity (e.g., 'bioactivity', 'biological activity', 'pharmacological activity'), and artificial intelligence or machine learning (e.g., 'machine learning', 'deep learning', 'neural network*', 'artificial intelligence'), using Boolean operators (AND/OR) and database-specific subject headings. Database-specific syntax adaptations were applied, and full search strings for each database are provided in Supplementary Table S2. No restrictions other than publication language (English) and period were applied at the search stage. In addition, backward and forward citation tracking of included studies was undertaken to identify any further eligible articles.

2.6. Study selection

All records retrieved from the databases were imported into a reference management system for duplicate removal and then into Rayyan for screening. After automatic and manual removal of duplicates, titles and abstracts of the remaining records were screened independently by two reviewers (SDK & SM) against the predefined eligibility criteria. Potentially eligible studies were retrieved in full text and assessed independently by the same reviewers, discrepancies were resolved by discussion, with involvement of a third reviewer when needed. The overall selection process is presented in a PRISMA 2020 flow diagram (Figure 1).

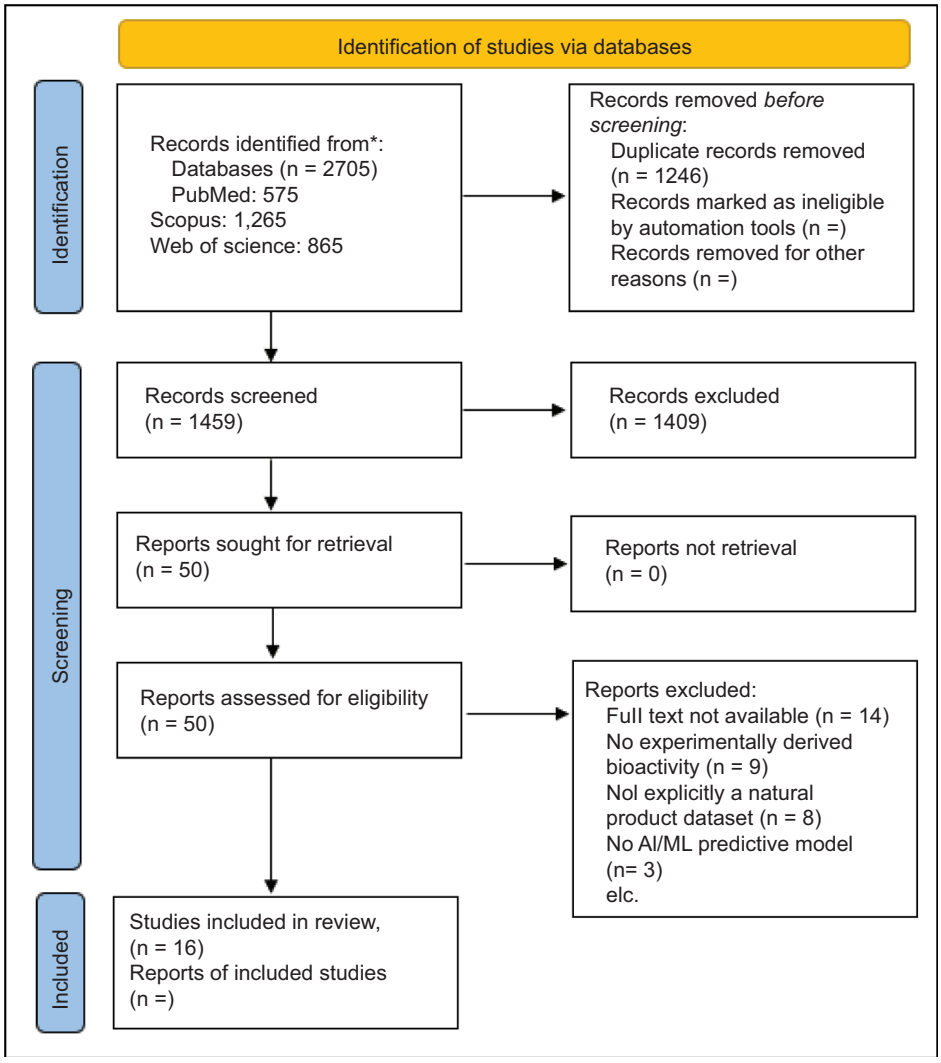

Figure 1. Prisma Flow Diagram. PRISMA 2020 flow diagram summarising the study selection process, including numbers of records identified through database searching, records screened, full-text articles assessed for eligibility, and studies included in the final systematic review.

2.7. Data extraction

A structured data extraction framework was developed prior to full extraction. At the study level, we extracted publication year and country, natural product source (plant, marine, microbial, mixed), dataset type (NP-only, mixed NP + synthetic, NP-derived), bioactivity class, endpoint type, dataset size, experimental origin of activity labels, and presence of prospective experimental validation.

At the model level, each predictive model within a study was extracted separately. Variables included model type (e.g. RF, SVM, CNN, GNN), learning paradigm, task type (classification or regression), molecular representation (descriptors, fingerprints, graph-based embeddings, metabolomics features, etc.), feature engineering steps and timing of feature selection, hyperparameter optimisation strategy, validation approach, presence of independent external validation, use of scaffold-based splitting, and applicability domain assessment.

Performance metrics were extracted in long format (one row per model × validation context × metric), capturing AUC/ROC, accuracy, F1-score, MCC, RMSE, R^2, and any additional performance indices reported. To minimise performance inflation, we prioritised metrics derived from held-out test sets, independent external datasets, or prospective experimental validation in the synthesis, and did not use training-only metrics as primary outcomes.

2.8. Validation tier classification

To enable structured comparison of validation rigour, models were categorised into five validation tiers: Tier 1,

internal random split only (e.g. a single train–test split within one dataset). Tier 2, internal cross-validation only. Tier 3 combined internal random split plus cross-validation. Tier 4, validation on an independent external dataset (retrospective). Tier 5, prospective experimental validation.

Tier 5 (prospective experimental validation) was defined as models for which predictions were prospectively tested in new wet-lab experiments designed after model development. Tier 4 (independent external dataset) captured retrospective evaluation on datasets or libraries not used for training or tuning. When multiple validation strategies were reported for a given model or study, we assigned the highest applicable tier based on the most rigorous validation actually used. For example, a model using both cross-validation and an independent external test set was classified as Tier 4, while one with prospective wet-lab confirmation of predictions was classified as Tier 5.

In practice, most studies that included laboratory follow-up evaluated only a small number of predicted hits. As a result, although Tier 5 (prospective experimental validation) is conceptually stronger than Tier 4 (retrospective external validation), the available Tier 5 data did not consistently represent a well-powered prospective evaluation in this corpus. For this reason, and to avoid overinterpreting very small strata, Tier 4 and Tier 5 models were grouped together as Tier 4–5 in the stratified analyses, and Tier 5 results are interpreted as indicating the presence of some prospective testing rather than a fully robust prospective validation tier.

2.9. Risk-of-bias assessment

Because the included studies were computational predictive modelling investigations rather than clinical trials or observational studies, conventional tools such as RoB 2 or ROBINS-I were not applicable. Instead, we adapted an AI-specific, domain-based risk-of-bias framework tailored to computational predictive modelling studies. For each included study, we assessed nine domains: (1) dataset adequacy, (2) class imbalance, (3) data leakage risk, (4) feature selection timing, (5) hyperparameter tuning leakage, (6) external validation, (7) applicability domain assessment, (8) code availability, and (9) data availability. Each domain was judged as low risk, high risk, or unclear based on prespecified criteria no aggregate numerical quality score was calculated to avoid artificial precision. In this context, applicability domain assessment encompassed common QSAR practices for delineating the reliable prediction space, including distance-based methods (e.g., similarity or k-nearest neighbour distances) and leverage-based approaches that identify compounds lying far outside the training-set descriptor space. The nine domains were chosen a priori, drawing on recurring sources of bias described in the AI/ML prediction-model literature and the core structure of tools such as PROBAST and its AI-focused updates. We did not apply PROBAST/PROBAST+AI directly because they are geared toward clinical prediction studies with patient-level data, so we instead translated the underlying principles (e.g. data leakage, validation design, applicability and reproducibility) into an AI-specific framework tailored to cheminformatics-style bioactivity models.

2.10. Data synthesis

Given the diversity of targets, endpoints, dataset sizes, molecular representations, and validation designs, we did not perform a meta-analysis. Instead, data were synthesised using descriptive statistics and stratified quantitative summaries. The focus was on median "best" performance metrics, taken from held-out, external, or prospective validation contexts rather than internal estimates.

Performance metrics were extracted in long format, with one row per model × validation context × metric. For each model, we prioritised results from held-out test sets, independent external datasets, or prospective experimental evaluations. Internal cross-validation and training-only results were considered lower priority. When multiple values of the same metric were reported within these prioritised contexts (for example, several AUC values), we selected the highest reported value and treated it as that model's "best" performance.

Study-level summaries were then constructed by identifying the top-performing model within each study. Group-level estimates—stratified by validation tier, dataset size (<100, 100–1,000, >1,000 compounds), and model type (classical ML versus deep learning/graph-based)—were calculated as the median of these study-level best values. Associations between validation rigor, dataset size, and reported performance were examined to explore possible patterns of performance inflation.

Risk-of-bias findings were summarised using frequency tables and visual heatmaps. Because several strata included only a small number of studies, the reported medians should be read as descriptive patterns, not precise pooled effect estimates.

In multi-target or multi-task studies (e.g. multiple enzyme targets or multi-label cell-line responses), each distinct prediction task was treated as a separate model entry in the performance extraction sheet. For study-level summaries, we selected the task-specific model with the highest prioritised 'best' AUC, so each study contributed a single best-performing model to the stratified analyses.

3. RESULTS

3.1. Study selection

The database search identified 2,705 records (PubMed: 575 Scopus: 1,265 Web of Science: 865). After removal of 1,246 duplicates, 1,459 unique records remained for title and abstract screening. Of these, 1,409 were excluded as they clearly did not meet the predefined eligibility criteria. Fifty full-text articles were assessed for eligibility, and 16 studies fulfilled all inclusion criteria and were included in the systematic review. The study selection process is summarized in Figure 1 (PRISMA 2020 flow diagram), and reasons for exclusion at the full-text stage were recorded and are summarised in Supplementary Table S3 (full-text articles excluded, with reasons, in line with PRISMA 2020).

3.2. Study characteristics

In total, 16 studies published between 2018 and 2025 were included (Table 1). Publication numbers rose noticeably from 2022 onwards, indicating a recent increase in AI/ML work on natural product bioactivity (Figure 2). The studies were carried out in a range of countries, most often China, South Korea, Brazil, Portugal, Morocco, and Thailand, with a few multi-country teams (Table 1).

3.2.1. Natural product sources and dataset types

Natural product sources encompassed plant-derived compounds, marine-derived metabolites, microbial products, essential oils, and mixed natural product plus synthetic libraries (Table 1). Several studies focused on curated plant or herbal preparations, such as Thai medicinal plants, Cuban plant essential oils, xanthoceras extracts, and Ayurvedic anti-epileptic herbs, while others used broader mixed datasets that combined natural products with synthetic or semi-synthetic analogues. Dataset types varied accordingly, including NP-only collections, NP-derived subsets from larger repositories, EO-level compositions, and mixed NP + synthetic libraries drawn from public and in-house sources (Table 1).

3.2.2. Bioactivity classes and endpoints

The bioactivity endpoints modelled in the included studies were predominantly pharmacological and target-specific (Table 1). Reported bioactivity classes included enzyme inhibition (e.g. CYP3A4, COX-2, anti-hyperglycaemic enzymes), antiproliferative or anticancer activity, antiprotozoal and schistosomicidal effects, anti-neuroinflammatory responses, anti-Alzheimer's targets, anti-epileptic interactions, antioxidant activity, and anthelmintic endpoints. Endpoints were represented as either continuous measure (e.g. pIC_{50}, pMIC, other potency scales) or binary/multiclass activity labels (active vs inactive, multi-class cell-line response), with several studies reporting both regression and classification tasks.

3.2.3. Dataset size and experimental origin

Dataset sizes varied widely, ranging from very small QSAR-style sets of fewer than 30 compounds to large screening collections exceeding 40,000 molecules (Table 1). When grouped by size, 4 studies used small datasets with fewer than 100 compounds, 6 studies used medium-sized datasets

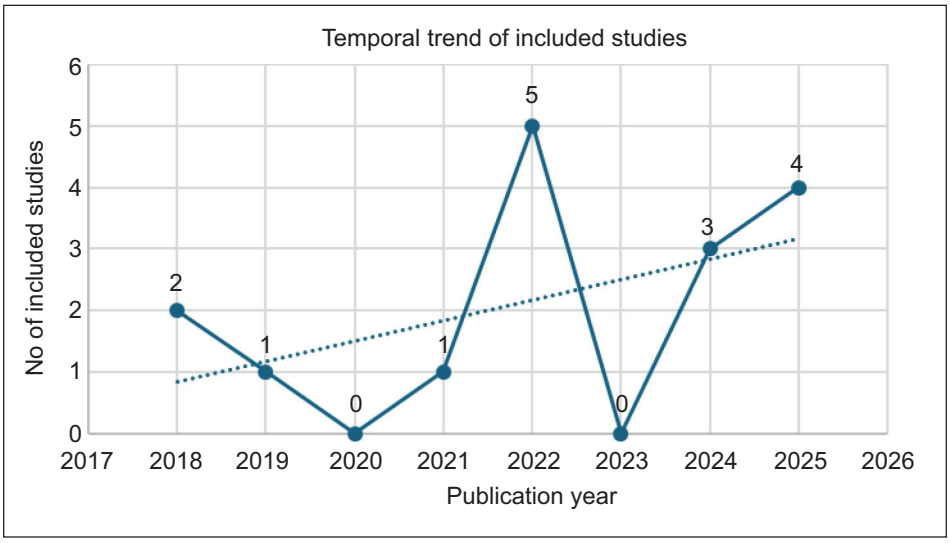

Figure 2. Temporal Trend of Publications. Annual number of included AI/ML studies on experimentally derived natural product bioactivity published between 2018 and 2025, illustrating the recent increase in publications from 2022 onwards.

Table 1
Study Characteristics.

Study_ID	Year	Country	Natural product source	Dataset type	Bioactivity class	Dataset size	Experimentalo rigin	Prospective experimental validation
Pang2018_RSCAdvances	2018	China	Mixed NP + synthetic	Mixed NP + synthetic	CYP3A4 inhibition / DDI risk	5,540 (2,830 actives)	BindingDB	Yes
Lahyaoui2025_ ChemPhysImpact	2025	Morocco	Plant (Wen Zhou et al. curation)	NP-derived	Anticancer (pIC50)	24	Literature curated	No
Qian2025_NewJChem	2025	China	Plant (HPL samples)	NP-only	Antioxidant (DPPH scavenging)	Not reported (31 extracts)	In-house assay	Partial (docking/MD + literature)
Lee2024_NO-Classifier	2024	South Korea	Mixed (literature + in-house)	Literature-mined + in-house	Anti-neuroinflammatory (NO inhibition)	554 (main) + 27 independent	Literature + in-house cell assay	No (retrospective only)
Gahl2024_PECAN	2024	USA / South Korea	NCI-60 (incl. natural products)	Public cytostatic screen	Antiproliferative (59 cancer lines)	49,126 compounds (~3M pairs)	NCI-60 DTP assay	No (retrospective + external NPL-720)
Yang2022_JNK1-NP-AI	2022	China	ChEMBL JNK1 + ZINC NP subset	ChEMBL + virtual screening	JNK1 kinase inhibition	1,423 (ML) + 4,112 (VS)	ChEMBL (in vitro)	Yes (3 hits wet-lab validated)
Barros2022_EOs-Protozoa	2022	Brazil / Cuba / USA	Cuban plant essential oils	EO-level curated dataset	Antiprotozoal	45 EO samples	Literature in vitro assays	No
Barros2022_Schisto-Alkaloids	2022	Brazil / USA	Alkaloids (Menispermaceae/ Apocynaceae)	ChEMBL + in-house NP database	Schistosomicidal	309 (QSAR) + 1,000 (VS)	ChEMBL + SistematX	No (in silico only)
Srisongkram2022_ ThaiHerbs	2022	Thailand	29 Thai medicinal plants	In-house experimental	Anti-hyperglycemic (α-amylase/glucosidase)	31 extracts	In-house enzyme inhibition assays	No
Ojo2021_ AutoQSAR-AD	2021	Nigeria (multi)	Plant flavonoids + ChEMBL inhibitors	QSAR datasets	Anti-Alzheimer's (AChE/ BChE/MAO)	74/47/71 per target	Literature (ChEMBL)	No
Ai2019_COX2-NP-QSAR	2019	China	Plants & microorganisms	QSAR + virtual screening	COX-2 inhibition	176 (QSAR) + 502 (VS)	Literature (Kauffman & Jurs)	Yes (5 hits cell-based validated)
Dias2019_MRSA-NP-QSAR	2019	Portugal	Mixed synthetic + marine actinobacteria	Regression + NMR QSAR	Anti-MRSA	6,645 (A) + 155 (B)	Literature + in-house MIC	Yes (Approach B: 4 external NPs)
Qian2025	2025	China	Xanthoceras sorbifolia	Target-specific QSAR	Anti-RA (7 targets)	Not reported (ChEMBL-derived)	ChEMBL	No
Gonzalez2022	2022	Ecuador (multi)	Agave brittoniana	QSAR classification	Anthelmintic	517 (train 352 / test 165)	Literature curated	Yes (in vitro + in vivo)
Riswanto2024	2024	Indonesia	LOTUS database	SBVS optimization	AChE inhibition	DUDE-Z 5,034 + LOTUS 59k screened	DUDE-Z benchmark	No
Shi2025	2025	Multi (China/ India/Oman)	Ayurvedic anti-epileptic herbs	Bipartite graph link prediction	Anti-epileptic	607 positive interactions	Compiled multi-database	No

Table 2
Model & Validation Characteristics.

Study ID	Main Model Type	Task Type	Molecular Representation	Feature Selection Timing	Validation Strategy	Validation Tier	External Validation	Scaffold Split
Pang2018	NB / SVM / RP	Binary classification	2D descriptors + ECFP-6	Before split	5-fold CV + external test	Tier 4	Yes	No
Lahyaoui2025	PLS / PCR / MLR	Regression	MOE 2D/3D descriptors + PCA	Before split	Random 19/5 split	Tier 4	Yes	No
Qian2025_NewJChem	Bagging-MLPR (best)	Regression	Metabolomics peak intensities	Unclear	Train/prediction hold-out	Tier 1	No	No
Lee2024_NO-Classifier	CNN / MLP / RNN / LASSO	Binary classification	3D conformers → E3FP + TF-IDF	During model (LASSO)	10-fold CV + 20% test + scaffold sets	Tier 4	Yes	Partial
Gahl2024_PECAN	MLP (weighted / resampled)	6-class multi-label	Morgan fingerprints (6144-bit)	None (fixed FP)	80/10/10 split + external NPL-720	Tier 4	Yes	No
Yang2022_JNK1	RF / SVM / ANN + ensembles	Binary classification	208 RDKit descriptors (pruned to 120)	Before split	10-fold CV + 20% test	Tier 3	Partial (wet-lab)	No
Barros2022_EOs	SOM + RF	Binary classification	EO composition vectors	None	5-fold external CV	Tier 2	No	No
Barros2022_Schisto	RF + MuDRA	Binary classification	VolSurf+ 3D + multiple fingerprints	Within CV	5-fold external CV	Tier 2	No	No
Srisongkram2022	RF (on PCA scores)	Binary classification	Phytochemical rank scores → PCA	Before split	5-fold CV	Tier 2	No	No
Ojo2021_AutoQSAR	AutoQSAR (kpls / pls)	Regression (pIC50)	AutoQSAR descriptors + fingerprints	Automatic (within training)	77/23 random split	Tier 1	No	No
Ai2019_COX2	FS-RF + SOM	Binary classification	Mold2 0D–2D descriptors	Before split	SOM-based single split	Tier 1	Yes (cell assay)	No
Dias2019_MRSA	RF + consensus (CM2 / CM_NMR)	Regression + binary	PaDEL FP + NMR binned descriptors	Within training	OOB + external test + 4 external NPs	Tier 4	Yes	No
Qian2025	Integrated GCN + ML ensemble	Binary classification	Descriptors + Morgan + MACCS + graphs	None	80/20 stratified + 10-fold CV	Tier 3	No	No
Gonzalez2022	LDA (QuBiLS-MAS quadratic indices)	Binary classification	Bond-based 2D quadratic indices	Before split	Train/test + LMO + Y-scrambling	Tier 4	Yes	No
Riswanto2024	Optimized RPART decision tree	Binary classification	ensPLIF from docking	Implicit in tree	Retrospective DUDE-Z + y-scrambling	Tier 1	No	No
Shi2025	GraphSAGE (global best)	Binary link prediction	Node2Vec + 5 centrality measures	Implicit in GNN	64/20/20 + 5-fold CV	Tier 1	No	No

of 100–1,000 compounds, and 6 studies used large datasets with more than 1,000 compounds. Experimental activity labels were derived from curated public repositories (e.g. BindingDB, ChEMBL, NCI-60), literature-derived potency data, and in-house assays, either singly or in combination (Table 1). Prospective experimental validation of model predictions was reported in a minority of studies, most commonly as targeted follow-up testing of a small number of predicted hits.

3.3. Model and validation characteristics

Across the 16 studies, 52 predictive models or modelling pipelines were extracted for detailed analysis (Table 2, Supplementary Table S1). These models differed in algorithmic family, task type, molecular representation, feature selection strategies, and validation design (Table 2).

3.3.1. Model types and task formulations

Classical machine learning approaches were the most frequently reported model family, with 11 studies using methods such as random forest, support vector machines, naïve Bayes, linear and logistic regression, partial least squares, principal component regression, and decision trees (Table 2). Deep learning and graph-based architectures, including convolutional neural networks, multilayer perceptrons, recurrent neural networks, graph convolutional networks, and GraphSAGE, were described in 5 studies, typically in the context of higher-dimensional or network-based data. Most models were framed as binary classification tasks (e.g. active vs inactive), although several studies implemented regression models for potency prediction and one study addressed a multi-label, multi-class anticancer response setting (Table 2).

3.3.2. Molecular representations

Molecular representations included traditional 2D and 3D descriptors, a variety of fingerprint schemes, graph-based embeddings, metabolomics-derived features, and docking-derived interaction fingerprints (Table 2). Descriptor-based QSAR approaches using software-derived 2D/3D descriptors and standard fingerprints (e.g. ECFP/Morgan, MACCS, VolSurf, PaDEL descriptors) were common, particularly in small to medium-sized datasets. Graph-based and deep learning models employed learned graph embeddings, node2vec features, and 3D conformer-based encodings, while some studies used composition vectors for essential oils or NMR-derived features for complex mixtures (Table 2, Supplementary Table S1).

3.3.3. Validation practices and tiers

Validation designs showed substantial heterogeneity across studies (Table 2, Figure 3 which summarises the distribution of validation tiers across included models). When categorized using the predefined validation tier framework, 8 studies were classified as Tier 1–2 (internal random split and/or cross-validation only), 2 studies as Tier 3 (combined internal split plus cross-validation), and 6 studies as Tier 4–5 (independent external dataset and/or prospective experimental validation). Scaffold-based splitting was rarely implemented one

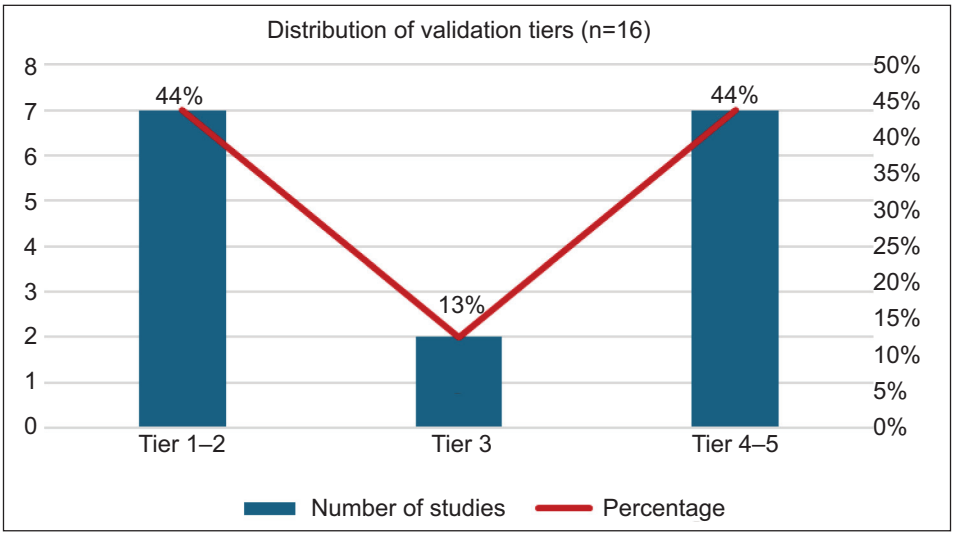

Figure 3. Validation Tier Distribution. Distribution of included studies across validation tiers (Tier 1–5) based on the most rigorous validation strategy reported for each study, highlighting the predominance of internal-only validation and the limited use of independent external or prospective evaluation.

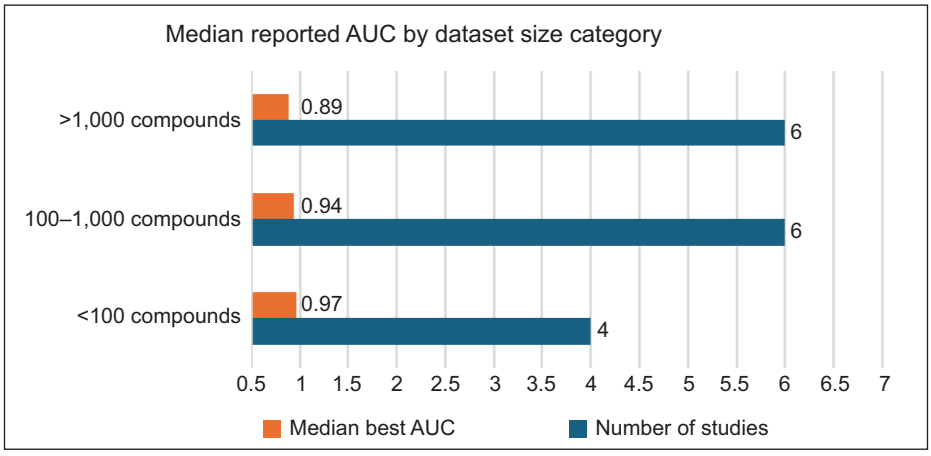

Figure 4. Dataset Size vs Median AUC. Relationship between dataset size categories (<100, 100–1,000, >1,000 compounds) and median best AUC, showing higher reported AUC values in smaller datasets and more moderate AUC values in larger datasets.

study reported partial use of scaffold-based sets, while most relied on random or stratified splits without explicit scaffold constraints (Table 2). External validation using independent compound sets, benchmark collections, or prospective assays was reported in fewer than half of the studies, and truly prospective wet-lab confirmation of predictions was limited to a small subset of investigations.

3.4. Risk-of-bias assessment

The AI-specific risk-of-bias assessment summarized the methodological characteristics of the 16 studies across nine domains (Table 3, Figure 5, which presents a heatmap of risk-of-bias judgments across domains and studies).

Table 3
AI-specific risk-of-bias summary (n=16).

Domain	Low Risk n (%)	High Risk n (%)	Unclear n (%)
Dataset Size Adequate (>100?)	13 (81.3)	2 (12.5)	1 (6.3)
Severe Class Imbalance	9 (56.3)	3 (18.8)	4 (25.0)
Data Leakage Risk	7 (43.8)	5 (31.3)	4 (25.0)
Feature Selection Before Split	4 (25.0)	**9 (56.3)**	3 (18.8)
Hyperparameter Tuning Leakage	5 (31.3)	6 (37.5)	5 (31.3)
External Validation Absent	5 (31.3)	**9 (56.3)**	2 (12.5)
Applicability Domain Missing	6 (37.5)	8 (50.0)	2 (12.5)
Code Availability	5 (31.3)	7 (43.8)	4 (25.0)
Data Availability	10 (62.5)	2 (12.5)	4 (25.0)

3.4.1. Dataset adequacy and class imbalance

Most studies (13/16, 81.3%) were judged to have adequate dataset size (more than 100 compounds) for their modelling aims, with only 2 studies (12.5%) categorized as high risk due to very small sample sizes (Table 3). Severe class imbalance was less common, with 9 studies (56.3%) rated low risk, 3 (18.8%) high risk, and 4 (25.0%) unclear, often reflecting limited reporting of class distributions (Table 3, Supplementary Table S4).

3.4.2. Data leakage and feature selection

Potential data leakage was identified as a frequent concern, with 5 studies (31.3%) classified as high risk for data leakage and a further 4 (25.0%) rated unclear (Table 3). Feature selection conducted before dataset splitting was the single most common methodological issue, recorded as high risk in 9 studies (56.3%) and low risk in only 4 (25.0%), while the remaining 3 (18.8%) did not clearly describe the timing of feature selection (Table 3, Supplementary Table S4).

3.4.3. Hyperparameter tuning and validation rigor

Hyperparameter tuning procedures were sometimes applied on full datasets or on test sets, leading to 6 studies (37.5%) being rated high risk and 5 (31.3%) unclear for hyperparameter tuning leakage (Table 3). External validation was absent in 9 studies (56.3%), which were therefore rated high risk in this domain only 5 studies (31.3%) were judged low risk for external validation, with 2 (12.5%) unclear (Table 3).

3.4.4. Applicability domain, code, and data availability

Applicability domain assessment was missing in 8 studies (50.0%), with 6 (37.5%) rated low risk (i.e. explicit AD analysis provided) and 2 (12.5%) unclear (Table 3). Code

Study_ID	Dataset Size Adequate (>100?)	Severe Class Imbalance	Data Leakage Risk	Feature Selection Before Split	Hyperparameter Tuning Leakage	External Validation Absent	Applicability Domain Missing	Code Availability	Data Availability
Pang2018_RSCAdvances	Low	Low	Unclear	High	Unclear	Low	High	Unclear	Unclear
Lahyaoui2025_ChemPhysImpact	High	Unclear	Unclear	High	Unclear	Low	Low	Unclear	Low
Qian2025_NewJChem	Unclear	unclear	Unclear	Unclear	Unclear	High	High	Not reported	Not reported
Lee2024_NO-Classifier	Low	Low	Unclear	Unclear	High	Low	Low	Unclear	Low
Gahl2024_PECAN	Low	Low	Unclear	Unclear	High	Low	Low	Unclear	Low
Yang2022_JNK1-NP-AI	Low	Low	Unclear	High	Unclear	Low	Low	Unclear	Low
Barros2022_EOs-Protozoa	High	Low	Unclear	Unclear	Unclear	High	High	Not reported	Low
Barros2022_Schisto-Alkaloids	Low	Low	Unclear	Unclear	Unclear	High	High	Not reported	Low
Srisongkram2022_ThaiHerbs	High	Low	Unclear	High	Unclear	High	High	Not reported	Low
Ojo2021_AutoQSAR-AD	High	Low	Unclear	High	Unclear	High	High	Not reported	Low
Ai2019_COX2-NP-QSAR	Low	Low	Unclear	High	Unclear	Low	Low	Unclear	Low
Dias2019_MRSA-NP-QSAR	Low	Low	Unclear	Unclear	High	Low	Low	Unclear	Low
Qian2025	Unclear	Unclear	Low	No	Low	High	High	Low	High
Gonzalez2022	Low	Low	Low	No	unclear	No	No	High	High
Riswanto2024	Low	High	Low	No	Low	Yes	Yes	High	High
Shi2025	Low	Low	Low	No	Low	Yes	Yes	Not reported	High

Low Risk | High Risk | Unclear

Figure 5. Risk-of-Bias Heatmap (Matrix). Heatmap summarising domain-level AI-specific risk-of-bias judgements (low, high, unclear) across the nine domains for all included studies, illustrating recurrent concerns such as feature selection before splitting, absence of external validation, and missing applicability-domain assessment.

availability was limited, with 7 studies (43.8%) categorized as high risk for this domain and only 5 (31.3%) explicitly providing or referencing code repositories data availability was comparatively better, with 10 studies (62.5%) rated low risk due to use of public datasets or shared data files (Table 3, Supplementary Table S4).

3.5. Performance patterns

Performance outcomes were synthesized using held-out test, independent external, or prospective validation metrics, as available, and summarized in Table 4 and Supplementary Table S1. To minimize inflation, training-only metrics and cross-validation results without an independent test context were not prioritized in the main synthesis.

3.5.1. Performance by validation tier

Across validation tiers, median best AUC values appeared higher in models with some form of external or prospective evaluation (Tier 4–5, 0.98–0.999) than in models using only internal validation (Tier 1–2, 0.89; Tier 3, 0.92), although these estimates are based on small numbers of studies within each tier (Table 4). Within the combined Tier 4–5 group, only a minority of models had any prospective experimental testing, and these usually involved small numbers of follow-up assays; therefore, Tier 4–5 should be interpreted as indicating at least one form of external validation, rather than

Table 4
Performance stratified analysis.

Category	N studies	Median best AUC	Median best MCC/F1 score	% with external validation
By Validation Tier				
Tier 1–2	8	0.89	0.78–0.85	86%
Tier 3	2	0.92	0.77–0.81	50%
Tier 4–5	6	0.98–0.999	0.92–0.97	14%
By Dataset Size				
< 100 compounds	4	0.97	0.89–0.94	25%
100 – 1,000 compounds	6	0.94	0.82–0.88	33%
> 1,000 compounds	6	0.89	0.77–0.84	67%
By Model Type				
Classical ML	11	0.94	0.82–0.90	36%
Deep Learning / GNN	5	0.96	0.85–0.95	40%

as a homogeneous group of fully prospective studies. Median MCC/F1 score values followed a similar pattern, with Tier 1–2 studies ranging from 0.78 to 0.85, Tier 3 from 0.77 to 0.81, and Tier 4–5 from 0.92 to 0.97. The proportion of models with any external validation was 86% in Tier 1–2, 50% in Tier 3, and 14% in Tier 4–5. This pattern reflects differences in how validation was defined and applied, and in many cases, reliance on narrow or closely related external datasets rather than consistently rigorous external benchmarking (Table 4).

3.5.2. Performance by dataset size

When stratified by dataset size, studies with fewer than 100 compounds reported a median best AUC of 0.97 and median best MCC/F1 score of 0.89–0.94, whereas medium-sized datasets (100–1,000 compounds) showed somewhat lower median best AUC (0.94) and MCC/F1 score (0.82–0.88), and large datasets (>1,000 compounds) showed the lowest median best AUC (0.89) and MCC/F1 score (0.77–0.84) (Table 4). External validation was reported for 25% of small, 33% of medium, and 67% of large datasets, suggesting that more modest performance estimates are generally associated with larger samples and more rigorous validation.

3.5.3. Performance by model type

Classical ML models (n = 11 studies) achieved a median best AUC of 0.94, with median best MCC/F1 score in the range of 0.82–0.90 (Table 4). Deep learning and graph-based models (n = 5 studies) reported a slightly higher median best AUC of 0.96, with median best MCC/F1 score between 0.85 and 0.95 (Table 4). The proportion of models evaluated with external validation was similar between classical ML and deep learning/GNN studies (36% and 40%, respectively), suggesting that differences in reported performance were not solely attributable to the use of more complex model architectures (Table 4, Supplementary Table S1).

3.6. Summary of key findings

Across the 16 included studies, AI/ML approaches were increasingly applied to experimentally derived bioactivity prediction in natural products, with a clear rise in publications after 2022 and a predominance of classical machine learning methods, particularly in small to medium-sized QSAR-style datasets (Table 1, Figure 2). Model development relied on a diverse range of molecular representations, from traditional descriptors and fingerprints to graph-based embeddings and metabolomics-derived features, but validation practices and reporting standards were highly variable (Table 2, Figure 3). The AI-specific risk-of-bias assessment highlighted recurrent methodological issues, most notably feature selection performed before dataset splitting, absence of external validation, and limited applicability domain analysis (Table 3, Figure 5, Supplementary Table S4).

Stratified summaries of predictive performance showed that higher median AUC and classification metrics were typically reported in studies with weaker validation designs and smaller datasets, whereas larger datasets and more rigorous validation were associated with more moderate but likely more realistic performance estimates (Table 4, Figures 3–4 which depicts performance by validation tier; Figure 4, which shows performance stratified by dataset size, Analyses 2–3). Taken together, these results indicate a rapidly expanding but methodologically heterogeneous evidence base, providing the context for a more detailed examination of methodological quality, performance interpretation, and implications for AI-assisted natural product discovery in the subsequent Discussion section.

4. DISCUSSION

4.1. Principal findings

This systematic review shows that applications of AI and ML to experimentally derived bioactivity prediction in natural products have grown steadily, with most of the 16 included studies published from 2018 onwards and a visible increase after 2022 [9,15–23]. Classical machine learning methods such as random forest, support vector machines, naïve Bayes, linear and logistic regression remain the most frequently used approaches, particularly in smaller QSAR-style datasets

[16–19,24–27]. In contrast, deep learning and graph-based models are mainly adopted in larger or more complex data structures, including multi-label cell-line response [9] and phytochemical–protein interaction networks [23], or in high-dimensional 3D conformer representations [15].

Molecular representations span traditional 2D/3D descriptors and fingerprints [19,21,24,25,27], Vol-Surf and multiple fingerprints [28], metabolomics signals [22], graph embeddings [8,23], and docking-derived interaction fingerprints [20]. Descriptor-based QSAR remains particularly dominant in smaller datasets, while graph and deep models are selectively deployed for larger or more structured problems [9,15,23].

4.2. Methodological quality and risk of bias

The AI-specific risk-of-bias framework applied in this review revealed that, although most studies used datasets of at least moderate size and did not exhibit extreme class imbalance, several recurrent sources of bias were evident. Feature selection before data splitting [11], a known cause of optimistic bias in predictive modelling, occurred in a majority of studies [16,18,21,24], and hyperparameter tuning leakage was frequently suspected where tuning procedures were described on "test" datasets [9,15,29]. These practices contrast with best-practice recommendations in the broader ML literature, which emphasise nested cross-validation, strict separation of training, validation, and test sets, and avoidance of any model selection on held-out test data. Although our AI-specific, domain-based risk-of-bias tool is not identical to established instruments such as PROBAST/PROBAST+AI or reporting guidance like TRIPOD+AI, it is conceptually aligned with these frameworks in its focus on data quality, analysis choices, and applicability. In particular, domains addressing data leakage, hyperparameter tuning leakage, external validation, and applicability-domain analysis parallel key signalling questions and reporting items in PROBAST+AI and TRIPOD+AI, but are adapted here to the structure of cheminformatics and natural product bioactivity modelling studies rather than patient-level clinical prediction models [30,31].

External validation was absent in over half of the included studies, with many relying solely on random train–test splits or k-fold cross-validation within a single dataset [16,18,22,23,27]. Where external evaluation was conducted, it often involved retrospective application to related datasets or compound libraries [9,15,19,25,26], and truly prospective wet-lab validation was reported only in a minority of cases [16,19,24,26]. Formal applicability domain analysis an important safeguard for QSAR and ML models was provided in a subset of QSAR-focused studies [17,21,28], but was largely absent in deep learning and graph-based work [9,15,23]. These analyses typically relied on classical QSAR-style applicability domains, for example using similarity or distance-based criteria in descriptor or fingerprint space, or leverage-based diagnostics to flag compounds that lie outside the multivariate space spanned by the training set, where prediction error is expected to increase.

Reproducibility was also uneven. While many studies leveraged public bioactivity repositories (e.g. ChEMBL, BindingDB, NCI-60) and thus indirectly enable data reuse [8,9,16,19,20,24–27], only a minority shared code or fully documented modelling pipelines [9,23]. This aligns with wider concerns in AI for drug discovery about incomplete methodological reporting and limited availability of source code and scripts.

4.3. Interpretation of performance patterns

Our stratified summaries indicate that headline performance metrics need to be read in the context of dataset size and how validation was done. The highest median AUC values tended to come from small datasets (≤100 compounds), for example in some QSAR or extract-level models [18,21,27], whereas larger datasets (≥1,000 compounds), such as the NCI-60-based PECAN [9] models or large mixed libraries [24,26] usually produced more moderate but probably more realistic performance. This inverse relationship between dataset size and median reported performance, together with the lower use of external validation in small datasets, points towards a substantial risk of overfitting and performance inflation in highly curated, limited-size settings. For instance, QSAR-style models built on fewer than 100 compounds often reported best AUC values around 0.97–0.99 in the absence of rigorous external validation, whereas models trained on larger datasets with independent test sets more often showed AUCs in the 0.89–0.94 range, which is more compatible with realistic out-of-sample performance.

Differences between validation tiers tell a similar story. Models in Tier 1–2 often showed strong internal performance but were not tested on truly independent external datasets [16,18,22,23,27], whereas Tier 4–5 models did use external or prospective evaluation [9,15,19,24–26] yet sometimes still reported AUC values very close to 1.0. In those cases, extremely high AUC may reflect narrow chemical or endpoint diversity, particular ways of handling class imbalance, or residual data leakage, rather than genuinely superior general performance. This underlines the need to look carefully at validation design, label structures, and preprocessing choices. It is also important to note that the combined Tier 4–5 category mainly reflects retrospective external validation, with

only a small subset of models undergoing limited prospective testing, so Tier 5 should be viewed as indicating the presence of prospective assays rather than a large, standalone prospective evidence base.

When different model families were compared, deep learning and graph-based approaches [8,9,15,23] showed slightly higher median AUC than classical machine learning, but the share of models with rigorous external validation was similar in both groups [32]. This pattern suggests that careful data curation and robust validation design may matter at least as much as, and sometimes more than, increasing architectural complexity a view that is consistent with broader benchmarking work in cheminformatics, where simpler models with well-designed validation often perform on par with, or even better than, more complex models in realistic applications.

4.4. Implications for AI-assisted natural product discovery

Collectively, the included studies demonstrate that AIML methods can successfully prioritize bioactive natural products and guide hypothesis generation across diverse targets, including CYP3A4 inhibition [24], COX-2 inhibition [25], antiproliferative activity [9], anti-Alzheimer's enzymes [27], antiprotozoal and anti-schistosomal effects [17,28], anti-hyperglycemic enzymes [18], and anti-neuroinflammatory endpoints [15]. Several studies linked computational predictions to follow-up experimental testing, supporting the role of ML models as enrichment tools rather than stand-alone decision-makers in natural product discovery workflows [16,19,24–26].

At the same time, this review underscores that many published models may not yet be ready for uncritical deployment on new chemical space or different assay conditions. Persistent issues around leakage, lack of external validation, incomplete applicability domain assessment, and limited code transparency can undermine trust in reported performance and hinder reproducibility. To realise the full potential of AI in natural product research, future studies should adopt more stringent validation protocols such as scaffold-based splits, time-based splits, and independent external test sets alongside explicit applicability domain analyses and comprehensive reporting of preprocessing, feature selection, and hyperparameter optimisation [33].

The development of domain-specific reporting standards for AI in drug discovery analogous to PRISMA 2020 for systematic reviews and CONSORT-type extensions [34] for prediction models could also help standardize minimum reporting requirements and facilitate cross-study comparisons. Integration of FAIR (Findable, Accessible, Interoperable, Reusable) data principles [35], open code repositories, and pre-registered analysis plans would further strengthen methodological transparency and reproducibility in this rapidly evolving field [36,37].

4.5. Strengths and limitations

This review has several strengths that, together, provide a structured view of how AI and ML are currently applied to natural product bioactivity prediction. We drew on three major bibliographic databases, complemented by citation tracking, and applied predefined eligibility criteria with dual independent screening and extraction, which reduces selection and extraction bias. The study- and model-level framework allowed us to examine not only which algorithms were used but also how datasets, molecular representations, and validation schemes were configured in practice. A further strength is the development of an AI-specific, domain-based risk-of-bias tool that captures issues such as data leakage, hyperparameter tuning leakage, and applicability domain assessment that are not addressed by traditional clinical tools.

Important limitations should also be acknowledged. We conducted a descriptive synthesis without meta-analysis because of extensive heterogeneity in targets, endpoints, and metrics, so pooled performance estimates are not provided. Restriction to English-language, peer-reviewed articles may have introduced language and publication bias and excludes potentially informative negative or inconclusive studies [38]. The risk-of-bias framework, although tailored to common ML pitfalls, has not yet been externally validated and relies on reporting quality, which was sometimes incomplete. Finally, the rapid pace of AI development means that additional methods and studies are likely to have emerged after the final search date. In addition, in order to enable a standardised comparison across heterogeneous studies, we focused our stratified analyses on the best-performing model from each article and on the highest prioritised performance metrics per model. While this simplifies cross-study comparison, it may overemphasise optimally tuned or selectively reported models and therefore contribute to an optimistic impression of achievable performance, relative to the full range of models and experiments described in the primary studies.

4.6. Future research directions

The patterns observed in this review suggest several priorities for strengthening AI/ML applications in natural product bioactivity prediction. First, there is a clear need for community-agreed reporting guidance tailored to cheminformatics and natural product modelling, analogous to existing AI

extensions of CONSORT and SPIRIT [39], to standardise description of datasets, preprocessing, validation design, and performance metrics. Such guidance should explicitly address data leakage, feature selection timing, and applicability domain analysis, which were frequent sources of concern in the included studies. Second, future work should emphasise rigorous validation strategies. These include scaffold-based and temporal splitting, independent external test sets, and, where feasible, prospective experimental confirmation of model predictions. Benchmarking efforts on shared, openly available natural product datasets would help disentangle the contributions of model choice versus data and validation design. Third, wider adoption of FAIR principles, with deposition of curated datasets, feature-generation pipelines, and trained models in public repositories, would substantially enhance transparency and reusability [40]. Finally, closer integration of ML pipelines with iterative experimental workflows such as active learning-guided screening of natural product libraries offers a pragmatic route to demonstrating real-world impact while continuously stress-testing models on genuinely novel chemistry.

5. CONCLUSION

This systematic review shows that AI and ML are now being applied across a diverse range of natural product datasets and bioactivity endpoints, but that current practice remains methodologically heterogeneous and often falls short of best-practice standards for predictive modelling. In particular, frequent use of small, highly curated datasets, feature selection before data splitting, limited applicability-domain analysis, and the absence of independent external validation contribute to optimistic performance estimates and complicate meaningful cross-study comparison.

Going forward, progress in this field will depend less on ever more complex architectures and more on careful attention to data curation, transparent reporting, and rigorous validation protocols, including scaffold-based or temporal splits, external test sets, and, where feasible, prospective experimental confirmation. Adoption of shared reporting standards, FAIR data and code practices, and benchmark datasets for natural product bioactivity prediction would provide a common foundation on which genuinely generalizable and trustworthy models can be built. By aligning methodological innovation with these principles, AI- and ML-driven approaches can move beyond proof-of-concept studies to become reliable partners in prioritizing natural products, de-risking early discovery, and ultimately expanding the space of bioactive chemotypes available for therapeutic development.

DECLARATIONS

Originality statement

The authors confirm that this manuscript is original, has not been published previously, and is not under consideration for publication elsewhere, in whole or in part.

Ethics statement

This systematic review is based exclusively on previously published studies and did not involve the collection of new data from human participants or animals. Consequently, ethical approval and informed consent were not required.

Conflict of interest statement

The authors declare that they have no financial or personal relationships that could inappropriately influence (bias) the work reported in this paper.

Funding statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Authors' contributions

S.K. conceived and designed the review, developed the protocol, performed the literature searches, screened studies, extracted and curated the data, conducted the analyses, prepared the tables and figures, and drafted the manuscript. S. contributed to study screening, cross-checked data extraction, and assisted with literature management. T.N. and A.M. critically reviewed the analyses and contributed to revising the manuscript for important intellectual content. V.M. provided overall supervision, contributed to the interpretation of findings. All authors read and approved the final version of the manuscript and agree to be accountable for all aspects of the work.

Data availability statement

The data underlying this systematic review (standardised extraction tables and performance matrices) are provided in the supplementary materials. Any additional data or analysis

files can be obtained from the corresponding author upon reasonable request.

USE OF AI-ASSISTED TOOLS

Artificial intelligence tools were used to assist with grammar checking, language refinement, and rephrasing during the manuscript preparation process. No content, ideas, data interpretation, or conclusions were generated by AI. All intellectual content and critical analysis remain the sole responsibility of the authors.

ACKNOWLEDGEMENT

The authors gratefully acknowledge Dr. Sanjana Mali, Patient Counsellor, Qi Spine Clinic, Qi Lifecare Pvt. Ltd., Pune, for her encouragement and practical support during the conduct of this work. The authors also thank Dr. Gayatri Korade, Independent Researchers, Pune, for the meaning full inputs and supports.

ORCID

Shashikanth Kharat 0009-0007-0205-4001
Shubham Kumar 0009-0000-7690-950X
Toufiq Noor 0009-0006-7296-8952
Ankita Mathur 0000-0002-9004-9072
Vini Mehta 0000-0003-4174-907X

REFERENCES

1. Master of Science in Chemistry University of New Haven, West Haven, Connecticut, USA, Tasnim S. Quantitative structure-activity relationship (qsar) modeling of bioactive compounds from mangifera indica for anti-diabetic drug development. Am J Adv Technol Eng Solut. 2022 Jun 1;02(02):01–32. https://doi.org/10.63125/ffkez356
2. Periwal V, Bassler S, Andrejev S, Gabrielli N, Patil KR, Typas A, et al. Bioactivity assessment of natural compounds using machine learning models trained on target similarity between drugs. Maranas CD, editor. PLOS Comput Biol. 2022 Apr 25;18(4):e1010029. https://doi.org/10.1371/journal.pcbi.1010029
3. Zhang R, Ren S, Dai Q, Shen T, Li X, Li J, et al. InflamNat: web-based database and predictor of anti-inflammatory natural products. J Cheminformatics. 2022 Dec;14(1):30. https://doi.org/10.1186/s13321-022-00608-5
4. AI's emerging role in natural product drug discovery | CAS [Internet]. [cited 2026 Mar 2]. Available from: https://www.cas.org/resources/cas-insights/ais-emerging-role-in-natural-product-drug-discovery
5. Dara S, Dhamercherla S, Jadav SS, Babu CM, Ahsan MJ. Machine Learning in Drug Discovery: A Review. Artif Intell Rev. 2022;55(3):1947–99. https://doi.org/10.1007/s10462-021-10058-4 PubMed PMID: 34393317; PubMed Central PMCID: PMC8356896.
6. Kojima R, Ishida S, Ohta M, Iwata H, Honma T, Okuno Y. kGCN: a graph-based deep learning framework for chemical structures. J Cheminformatics. 2020 May 12;12(1):32. https://doi.org/10.1186/s13321-020-00435-6
7. Liu Z, Huang D, Zheng S, Song Y, Liu B, Sun J, et al. Deep learning enables discovery of highly potent anti-osteoporosis natural products. Eur J Med Chem. 2021 Jan;210:112982. https://doi.org/10.1016/j.ejmech.2020.112982
8. Qian H, Xiao Z, Su L, Yang Y, Tian X, Wang X. Prediction of Anti-rheumatoid Arthritis Natural Products of Xanthocerais Lignum Based on LC-MS and Artificial Intelligence. Comb Chem High Throughput Screen. 2025 Feb;28(4):627–46. https://doi.org/10.2174/0113862073282138240116112348
9. Gahl M, Kim HW, Glukhov E, Gerwick WH, Cottrell GW. PECAN Predicts Patterns of Cancer Cell Cytostatic Activity of Natural Products Using Deep Learning. J Nat Prod. 2024 Mar 22;87(3):567–75. https://doi.org/10.1021/acs.jnatprod.3c00879
10. Ferreira FJN, Carneiro AS. AI-Driven Drug Discovery: A Comprehensive Review. ACS Omega. 2025 Jun 17;10(23):23889–903. https://doi.org/10.1021/acsomega.5c00549 PubMed PMID: 40547666; PubMed Central PMCID: PMC12177741.
11. Masand VH, Mahajan DT, Nazeruddin GM, Hadda TB, Rastija V, Alfeefy AM. Effect of information leakage and method of splitting (rational and random) on external predictive ability and behavior of different statistical parameters of QSAR model. Med Chem Res. 2015 Mar;24(3):1241–64. https://doi.org/10.1007/s00044-014-1193-8
12. M. De Angelo R, G. Maltarollo V, G. Lago JH, Honorio KM. Machine Learning Models for the Identification of Natural Products with Anti-Schistosomiasis Activity. J Braz Chem Soc. 2025. https://doi.org/10.21577/0103-5053.20250070
13. AI-driven drug discovery from natural products. Adv Agrochem. 2024 Sep 1;3(3):185–7. https://doi.org/10.1016/j.aac.2024.06.003
14. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. Int J Surg. 2021 Apr;88:105906. https://doi.org/10.1016/j.ijsu.2021.105906 PubMed PMID: 33789826.
15. Lee SE, Ahn S, Kumar S, Kim M hyun. NO classifier prediction of anti neuroinflammatory agents using text mining of 3D molecular fingerprints. Sci Rep. 2024 Nov 16;14(1):28338. https://doi.org/10.1038/s41598-024-78823-3
16. Yang R, Zhao G, Yan B. Discovery of Novel c-Jun N-Terminal Kinase 1 Inhibitors from Natural Products: Integrating Artificial Intelligence with Structure-Based Virtual Screening and Biological Evaluation. Molecules. 2022 Sep 22;27(19):6249. https://doi.org/10.3390/molecules27196249
17. Barros De Menezes RP, Scotti L, Scotti MT, García J, González R, Monzote L, et al. Machine Learning Analysis of Essential Oils from Cuban Plants: Potential Activity against Protozoa Parasites. Molecules. 2022 Feb 17;27(4):1366. https://doi.org/10.3390/molecules27041366
18. Srisongkram T, Waithong S, Thitimetharoch T, Weerapreeyakul N. Machine Learning and In Vitro Chemical Screening of Potential

α-Amylase and α-Glucosidase Inhibitors from Thai Indigenous Plants. Nutrients. 2022 Jan 9;14(2):267. https://doi.org/10.3390/nu14020267
19. González-Castañeda Y, Marrero-Ponce Y, Guerra JO, Echevarría-Díaz Y, Pérez N, Pérez-Giménez F, et al. Computational discovery of novel anthelmintic natural compounds from Agave Brittoniana trel. Spp. Brachypus. Bionatura. 2022 Dec 15;7(4):1–15. https://doi.org/10.21931/RB/2022.07.04.53
20. Dika Octa Riswanto F, Waskitha SSW, Yanuar MRS, Istyastono EP. Structure-based virtual screening on a new open-source natural products database LOTUS to discover acetylcholinesterase ınhibitors. J Res Pharm. 2024;28(4)(28(4)):1099–106. https://doi.org/10.29228/jrp.792
21. Lahyaoui M, El Yaqoubi M, Idrissi HE, Moumni B, Saffaj T, Ihssane B, et al. Design, quantitative structure–activity relationships and computational studies of acylshikonin derivatives: Insights into antitumor activity. Chem Phys Impact. 2025 Dec;11:100959. https://doi.org/10.1016/j.chphi.2025.100959
22. Qian J, Nie W, Zhang Z, Fang G, Li C, Li W. A machine learning-based strategy for screening bioactive compounds in natural products: a case study on *Hypericum perforatum* L. New J Chem. 2025;49(36):15709–22. https://doi.org/10.1039/D5NJ02230D
23. Shi X, Sundareswaran R, Shanmugapriya M, Jayaguptha N, Khan A. Graph based link prediction for epilepsy drug discovery. Sci Rep. 2025 Sep 30;15(1):34087. https://doi.org/10.1038/s41598-025-14550-7
24. Pang X, Zhang B, Mu G, Xia J, Xiang Q, Zhao X, et al. Screening of cytochrome P450 3A4 inhibitors *via in silico* and *in vitro* approaches. RSC Adv. 2018;8(61):34783–92. https://doi.org/10.1039/C8RA06311G
25. Ai S, Lin G, Bai Y, Liu X, Piao L. QSAR Classification-Based Virtual Screening Followed by Molecular Docking Identification of Potential COX-2 Inhibitors in a Natural Product Library. J Comput Biol. 2019 Nov 1;26(11):1296–315. https://doi.org/10.1089/cmb.2019.0142
26. Dias T, Gaudêncio SP, Pereira F. A Computer-Driven Approach to Discover Natural Product Leads for Methicillin-Resistant Staphylococcus aureus Infection Therapy. Mar Drugs. 2018 Dec 28;17(1):16. https://doi.org/10.3390/md17010016
27. Ojo OA, Ojo AB, Okolie C, Nwakama MAC, Iyobhebhe M, Evbuomwan IO, et al. Deciphering the Interactions of Bioactive Compounds in Selected Traditional Medicinal Plants against Alzheimer's Diseases via Pharmacophore Modeling, Auto-QSAR, and Molecular Docking Approaches. Molecules. 2021 Apr 1;26(7):1996. https://doi.org/10.3390/molecules26071996
28. Menezes RPBD, Viana JDO, Muratov E, Scotti L, Scotti MT. Computer-Assisted Discovery of Alkaloids with Schistosomicidal Activity. Curr Issues Mol Biol. 2022 Jan 15;44(1):383–408. https://doi.org/10.3390/cimb44010028
29. Logan M. SYSBIOSPHERE [Internet]. Global Optimization for Model Tuning: Enhancing Accuracy and Efficiency in Drug Discovery. Available from: https://www.sysbiosci.com/posts/global-optimization-for-model-tuning-enhancing-accuracy-and-efficiency-in-drug-discovery
30. Collins GS, Dhiman P, Navarro CLA, Ma J, Hooft L, Reitsma JB, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open. 2021 Jul 1;11(7):e048008. https://doi.org/10.1136/bmjopen-2020-048008 PubMed PMID: 34244270.
31. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Calster BV, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024 Apr 16;385:e078378. https://doi.org/10.1136/bmj-2023-078378 PubMed PMID: 38626948.
32. Koirala M, Yan L, Mohamed Z, DiPaola M. AI-Integrated QSAR Modeling for Enhanced Drug Discovery: From Classical Approaches to Deep Learning and Structural Insight. Int J Mol Sci. 2025 Sep 25;26(19):9384. https://doi.org/10.3390/ijms26199384 PubMed PMID: 41096653; PubMed Central PMCID: PMC12525248.
33. Guo Q, Hernandez-Hernandez S, Ballester PJ. UMAP-based clustering split for rigorous evaluation of AI models for virtual screening on cancer cell lines. J Cheminformatics. 2025 Jun 10;17(1):94. https://doi.org/10.1186/s13321-025-01039-8 PubMed PMID: 40495205; PubMed Central PMCID: PMC12153141.
34. Martindale APL, Llewellyn CD, de Visser RO, Ng B, Ngai V, Kale AU, et al. Concordance of randomised controlled trials for artificial intelligence interventions with the CONSORT-AI reporting guidelines. Nat Commun. 2024 Feb 22;15(1):1619. https://doi.org/10.1038/s41467-024-45355-3 PubMed PMID: 38388497; PubMed Central PMCID: PMC10883966.
35. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016 Mar 15;3:160018. https://doi.org/10.1038/sdata.2016.18 PubMed PMID: 26978244; PubMed Central PMCID: PMC4792175.
36. Chandrasekhar V, Rajan K, Kanakam SRS, Sharma N, Weißenborn V, Schaub J, et al. COCONUT 2.0: a comprehensive overhaul and curation of the collection of open natural products database. Nucleic Acids Res. 2025 Jan 6;53(D1):D634–43. https://doi.org/10.1093/nar/gkae1063 PubMed PMID: 39588778; PubMed Central PMCID: PMC11701633.
37. AI approaches for the discovery and validation of drug targets - PMC [Internet]. [cited 2026 Mar 2]. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC11383977/
38. Mckenzie T. The Hidden Value of Failure: Why Negative Data is Critical for AI-Driven Drug Discovery. Digital Chemistry [Internet]. 2025 Oct 2 [cited 2026 Mar 2]. Available from: https://www.digitalchemistry.ai/resources/white-papers/the-hidden-value-of-failure-why-negative-data-is-critical-for-ai-driven-drug-discovery/
39. Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ, Grupo de Trabajo SPIRIT-AI y CONSORT-AI, et al. [Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extensionDiretrizes para protocolos de ensaios clínicos com intervenções que utilizam inteligência artificial: a extensão SPIRIT-AI]. Rev Panam Salud Publica Pan Am J Public Health. 2024;48:e12. https://doi.org/10.26633/RPSP.2024.12 PubMed PMID: 38304411; PubMed Central PMCID: PMC10832304.
40. AI in drug discovery: predictions for 2026. Drug Target Review [Internet]. [cited 2026 Mar 2]. Available from: https://www.drugtargetreview.com/article/192962/ai-in-drug-discovery-predictions-for-2026/