



REVIEW

Open Access



Artificial intelligence-based screening of phytochemicals for targeted cancer therapy

Livia Ramos Santiago¹ , Estéfani Alves Asevedo¹ , Maria Eduarda Jeunon de Oliveira¹ , Karen Cota Pereira¹ , Maria Fernanda da Silva Trindade¹ , Ana Gabriela Silva Oliveira¹ , Marina Andrade Rocha¹ , Sojin Kang² , Amama Rani² , Moon Nyeo Park² , Michel William Tan³ , Rony Abdi Syahputra³ , Bonglee Kim² and Rosy lara Maciel de Azambuja Ribeiro^{1*}

Abstract

Cancer remains one of the leading causes of death worldwide and continues to pose a serious public health challenge. The limited success of many current treatments—often due to toxicity, poor selectivity, and the development of drug resistance—highlights the need for new and more effective therapeutic options. Phytochemicals have emerged as a valuable source of anticancer agents, offering rich structural diversity and a wide range of biological activities. However, identifying promising compounds from the vast chemical space of natural products remains difficult using conventional screening methods, which are typically slow, costly, and inefficient. In recent years, artificial intelligence (AI) has begun to transform phytochemical-based drug discovery. Machine learning and deep learning approaches are now used to support key steps in the discovery process, including metabolite identification, virtual screening, target prediction, and toxicity assessment. By integrating chemical, biological, and multi-omics data, AI enables a more systematic and data-driven exploration of natural product diversity. Despite these advances, challenges persist, particularly the scarcity of high-quality experimental data, the structural complexity of phytochemicals, and their limited representation in public databases. This review critically examines current AI applications in phytochemical-based anticancer drug discovery and discusses emerging strategies aimed at overcoming these limitations. Overall, AI-driven phytochemical screening represents a promising path toward accelerating the development of next-generation cancer therapies.

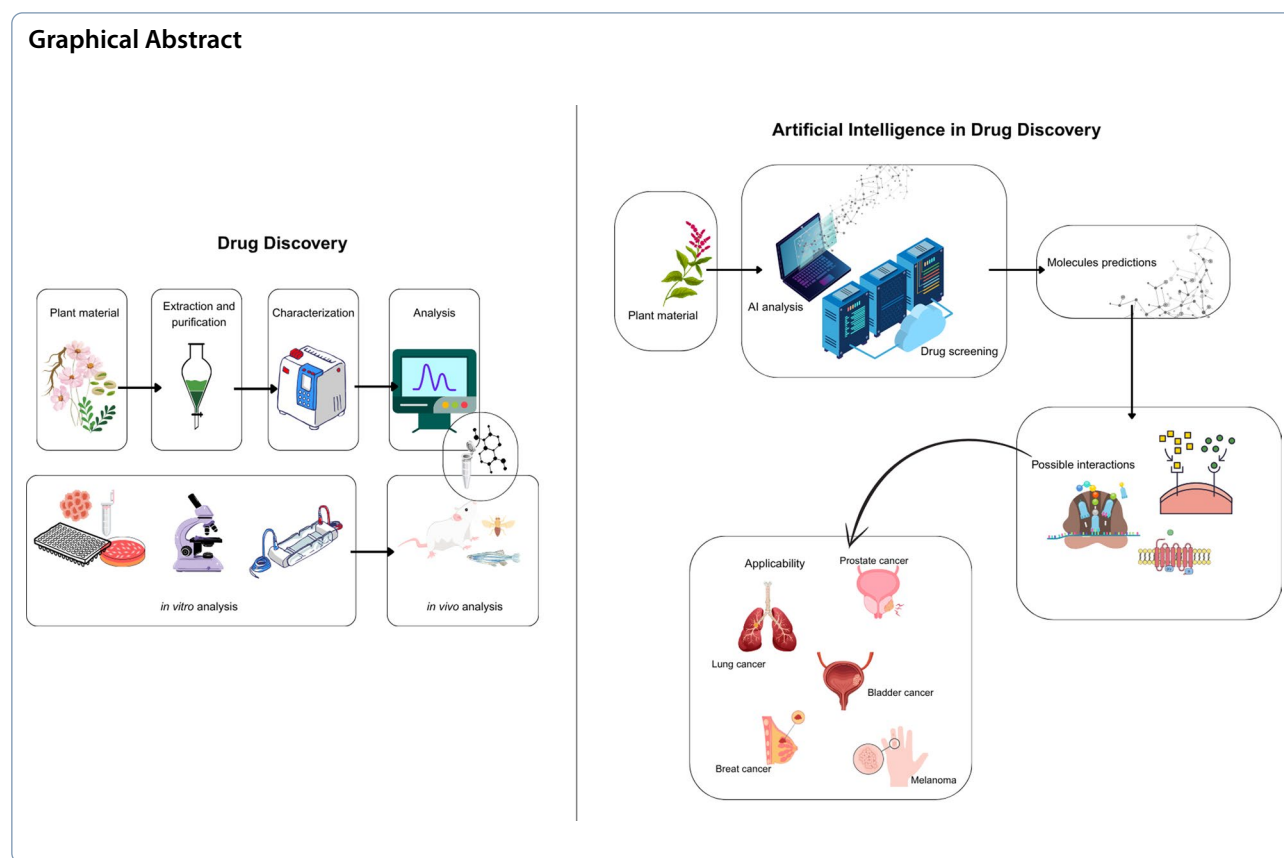
Keywords Phytochemicals, Artificial intelligence, Machine learning, Deep learning

*Correspondence:

Rosy lara Maciel de Azambuja Ribeiro
rosy@ufsj.edu.br

Full list of author information is available at the end of the article

Graphical Abstract



1 Introduction

Cancer remains one of the leading causes of death globally, accounting for approximately one in every six deaths. Despite advances in conventional therapies such as surgery, radiotherapy and chemotherapy, limitations including severe side effects, limited selectivity, and growing drug resistance continue to hamper treatment success [1, 2]. These limitations continue to drive the search for alternative and complementary therapeutic strategies that can improve efficacy while minimizing toxicity.

In this context, phytochemicals have emerged as an attractive source of novel anticancer agents. These naturally occurring bioactive compounds, derived mainly from medicinal plants and dietary sources, exhibit diverse mechanisms of action and have been shown to enhance the therapeutic efficacy of conventional chemotherapeutic agents [3]. Due to their structural diversity and broad pharmacological activities, many phytochemicals exhibit favorable attributes such as low toxicity, biocompatibility and dual therapeutic–nutritional benefit [4]. Compared with many synthetic drugs, phytochemicals are often associated with improved safety profiles and a reduced propensity for the development of multidrug resistance [5]. Notable

plant-derived anticancer drugs approved by the Food and Drug Administration (FDA) are Paclitaxel, isolated from the extract of *Taxus brevifolia*, Camptothecin, isolated from the extract of *Camptotheca acuminata* Decne, and the alkaloids vinblastine, vincristine, vindesine and vinorelbine, isolated from the extract of *Catharanthus roseus* [6–8]. Nevertheless, despite the vast chemical diversity of the plant kingdom, the systematic discovery of anticancer phytochemicals remains slow, largely due to the limitations of traditional experimental screening approaches, which are time-consuming, costly, and poorly suited for large-scale exploration [9].

In recent years, artificial intelligence (AI) has emerged as a powerful enabling technology capable of addressing these challenges in natural product–based drug discovery. AI-driven computational methods allow for the high-throughput analysis of complex chemical and biological datasets, supporting tasks such as virtual screening, target prediction, activity modeling, and lead optimization for phytochemicals with anticancer potential [9]. Importantly, phytochemical research poses unique computational challenges, including high structural complexity, limited annotated datasets, and chemical features that deviate from typical drug-like space, which often reduce the effectiveness of

conventional modeling approaches. AI and machine-learning techniques are particularly well suited to overcoming these barriers by learning non-linear patterns from heterogeneous data and enabling predictive modeling even in data-constrained settings. Beyond reducing costs and development time, AI-driven methods hold the potential to revolutionize precision oncology by enabling the design of therapies that target specific cancer pathways with greater accuracy [10].

Although numerous AI tools and platforms have been developed for drug discovery, existing literature often presents these methods in a fragmented manner, with limited critical comparison, performance evaluation, or guidance on tool selection for specific phytochemical screening scenarios. Moreover, many available reviews either focus broadly on AI in drug discovery or on natural products without systematically integrating both perspectives in the context of anticancer research. To address these gaps, this review aims to provide a structured and critical overview of AI applications in phytochemical-based anticancer drug discovery, with emphasis on methodological principles, comparative strengths and limitations of existing tools, representative case studies, and practical considerations for tool selection across different research objectives. By doing so, we seek to offer not only a summary of current advances but also a decision-oriented framework to guide future research in this rapidly evolving field.

2 Anticancer phytochemicals and current limitations to their discovery

Natural products are characterized by presenting scaffolding diversity and structural complexity. Compared with most synthetic small molecules, phytochemicals often exhibit higher molecular weights, increased rigidity, and a greater number of stereocenters and functional groups, features that contribute to their rich biological activity but also complicate conventional drug discovery workflows. These properties allow natural products to serve as valuable scaffolds for drug discovery, even for compounds that fall outside Lipinski's Rule of Five. Consistent with this trend, the average molecular weight of newly approved drugs has increased in recent years, highlighting a gradual shift away from strict adherence to traditional drug-likeness criteria [11]. Despite this, natural products continue to be the foundation of most approved anticancer drugs, supported by resources such as the National Cancer Institute's Natural Products Repository, comprising more than 180,000 extracts.

In general, after identifying bioactive extracts, the next steps involve isolating individual components and characterizing their chemical structures. However,

this process is often labor-intensive and technically challenging, particularly when dealing with complex mixtures and low-abundance active compounds. Practical challenges associated with handling natural products, including difficulties in sample reconstitution, dilution, liquid transfer, variability, viscosity and precipitation, all of which can limit reproducibility and scalability in high-throughput screening campaigns [12]. These limitations represent a major bottleneck in the systematic discovery and optimization of anticancer phytochemicals.

One promising approach to overcoming drug resistance in tumor cells involves the use of natural products as adjuvants. Numerous studies have demonstrated that natural compounds can resensitize cancer cells to conventional chemotherapeutics by modulating multiple resistance-associated pathways [12]. For example, treatment with the phytochemical curcumin combined with docetaxel resulted in a downregulation of drug resistance genes in MCF-7 human breast cancer cells [13]. Another example is that *Myrciaria tenella* extracts caused cell death in human cervical adenocarcinoma (HeLa) and human gastric adenocarcinoma (SFA) cell lines. In addition, the extracts did not show cytotoxicity in the 3T3 non-tumor cell line [14]. Such findings support the growing interest in phytochemicals as multitarget, potentially less toxic alternatives or complements to conventional single-target chemotherapeutic agents [15].

Natural compounds can interfere with essential cellular processes such as DNA replication, topoisomerase functioning, and microtubule dynamics, encouraging their broad investigation in preclinical and clinical studies. Among the drugs approved by the Food and Drug Administration (FDA), Camptosar, derived from the plant *Camptotheca acuminata*, used in the treatment of metastatic colorectal cancer; Etoposide, obtained from the podophyllotoxin present in *Podophyllum peltatum*, used for lung cancer and lymphomas; and Taxol, isolated from the bark of *Taxus brevifolia*, widely used in the treatment of non-small cell lung cancer, neuroblastoma, and Wilms' tumor. In addition to plant-derived agents, marine- and microorganism-derived drugs such as Halaven, Cytarabine, and Synribo further illustrate the broad diversity of natural sources contributing to modern oncology. Table 1 summarizes the major FDA-approved natural product-derived anticancer drugs, presenting their respective sources of natural products, the indications for cancers to be treated, mechanisms of action, and bibliographic references.

Despite their proven therapeutic value, the development of natural product-based anticancer drugs remains constrained by several fundamental limitations. One major challenge is the availability and sustainable supply of source organisms. Many microorganisms that

Table 1 List of FDA-approved natural product-derived anticancer drugs

Commercial name	Source (Natural products)	Specific cancer suppressed	Mechanism of action	Refs.
Camptosar	Extract of <i>Camptotheca acuminata</i>	Metastatic colorectal cancer	SN-38 acts as a topoisomerase I inhibitor. It binds to the DNA-topoisomerase I complex, preventing the religation of DNA single-strand breaks. The accumulation of DNA breaks during replication, interrupting the cell cycle and inducing apoptosis	[16]
Cytarabine	Spongouridine nucleoside from the sponge <i>Tethya crypta</i> (<i>Cryptotethya crypta</i>)	Hematologic neoplasm	It is a deoxycytidine analog that acts on cells during DNA synthesis, specifically killing them in the S phase of the cell cycle, and can also block the transition of cells from the G1 phase to the S phase	[17]
Etoposide	Derives from podophyllotoxin, a toxin found in the <i>Podophyllum peltatum</i>	Lung cancer, germ cell tumors, and refractory lymphomas	Etoposide acts by inhibiting topoisomerase II, causing double-strand breaks in DNA and forming a complex by binding to the DNA-topoisomerase II, causing an impediment to the religation of the double-strand breaks in DNA. As a result, there is cell cycle arrest and induction of apoptosis	[18]
Halaven	Isolated from the marine sponge <i>Halichondria okadai</i>	Metastatic breast cancer	Binds with high specificity to the ends of microtubules, inhibiting polymerization without affecting depolymerization, causing disorganization of the mitotic spindle, arrest in the G2/M phase of the cell cycle, prolonged mitotic blockade, and induction of apoptosis	[19]
NAVELBINE	Derived from <i>Vinca rosea</i>	Non-small cell lung cancer	Inhibition of microtubule formation in mitotic spindle, resulting in an arrest of dividing cells at the metaphase stage	[20]
Synribo	Vincristine Sulfate Injection; extract from leaves of <i>Cephalotaxus</i> ; alkaloid from <i>Vinca rosea</i>	Chronic myeloid leukemia; Leukemia, Hodgkin's disease, non-Hodgkin's lymphomas, rhabdomyosarcoma, neuroblastoma, Wilms' tumor	Inhibits protein synthesis and microtubule formation in mitotic spindle, arresting cells at metaphase	[21]
Taxol	Originally isolated from the bark of <i>Taxus brevifolia</i>	Non-small cell lung cancer	Promotes the assembly of microtubules from tubulin dimers and stabilizes microtubules by preventing depolymerization	[22]

produce bioactive metabolites are difficult to cultivate outside their natural habitats, which can compromise the continuous production of these compounds. To address this issue, alternative strategies such as *in situ* analysis, synthetic induction of biosynthetic pathways, and heterologous expression of biosynthetic gene clusters in model organisms have been explored [23]. In addition, crude extracts often contain reduced concentrations of the active ingredients, which makes their isolation and characterization difficult.

Another significant challenge concerns the distribution and bioavailability of natural compounds. Many natural compounds undergo extensive first-pass hepatic metabolism, leading to reduced plasma concentrations and diminished therapeutic efficacy [24]. Often, high doses are required to achieve clinically relevant effects. In this context, the use of lipid carrier systems has shown promise. These carriers, owing to their biocompatibility, biodegradability, and capacity for controlled release, offer promising strategies to enhance the stability, absorption, and bioavailability of phytochemicals [25].

Beyond pharmacokinetic constraints, the clinical development of herbal medicines faces additional challenges, including poor aqueous solubility, chemical instability, and specific regulatory requirements. For example, classification as a Traditional Herbal Product (THP) often requires evidence of long-term traditional use, which may not align with modern drug development paradigms. To overcome these barriers, various pharmaceutical strategies, such as nanoparticles, liposomes and hydrogels, have been investigated with the aim of overcoming these limitations, promoting greater stability, solubility and sustained release. However, despite encouraging preclinical findings, robust and well-designed preclinical and clinical investigations remain essential to fully establish the efficacy and safety and translational viability of these approaches [26].

Collectively, these challenges underscore the need for innovative, scalable, and integrative strategies to accelerate the discovery and development of anticancer phytochemicals—an area where computational and AI-driven approaches offer considerable promise, as discussed in subsequent sections.

3 Artificial intelligence in drug discovery and screening

The development of new drugs is increasingly benefiting from artificial intelligence (AI) and big data, which allow for faster and more efficient processes, reduced reliance on animal testing, and minimized human bias. These technologies also enable more accurate analysis and handling of complex chemical and biological data compared with traditional manual approaches. A notable

example of this is QSAR (quantitative structure–activity relationship), a computational method used to predict the biological activity of chemical compounds. AI-enhanced QSAR models improve predictive performance by capturing complex, non-linear relationships between molecular descriptors and biological activity. Such models are now applied across multiple stages of drug discovery, from target identification and validation to guiding clinical trial design [27].

Artificial Intelligence (AI) is rapidly reshaping the landscape of drug discovery and development, addressing solutions to long-standing challenges such as inefficiency, high costs, and ethical concerns, including the need to reduce animal testing and mitigate human bias.

Advanced AI techniques, including machine learning (ML), deep learning (DL), and natural language processing (NLP), are now widely adopted to predict molecular properties, identify druggable targets, and optimize lead compounds with improved precision and speed.

By leveraging vast and diverse datasets, ranging from genomics and chemical structures to clinical records, AI models can uncover hidden patterns and generate actionable hypotheses far more rapidly than traditional methods. A comprehensive review by Vamathevan et al. highlights how AI-driven platforms can significantly accelerate hit identification while reducing attrition during early-stage drug development [27, 28].

Furthermore, the tools allow us to identify whether natural compounds can inhibit specific targets. In cancer research, signaling pathways play a central role in therapeutic development, with the mitogen-activated protein kinase (MAPK) pathway being one of the most extensively studied. The MAPK pathway is thought to be involved in glioma development. This pathway interacts with several bioactive regulators. In carcinogenesis, dysregulation of this pathway has been associated with both therapeutic resistance and sensitivity. The MAPK signaling pathway is upregulated in glioma. To model pathway-associated prognostic risk, least absolute shrinkage and selection operator (LASSO) regression has been applied to select relevant genes and coefficients for machine-learning-based risk models. As a result, it was demonstrated that high-risk groups tend to have shorter survival times, while the low-risk group has a more dispersed representation in the analysis range, resulting in a wider survival time [29].

Zhang et al. (2019) demonstrated the use of a computational modeling strategy to evaluate the response of the HCT116 colorectal cancer cell line to cisplatin and TRAIL at the single-cell level. Their models predicted that combined cisplatin and TRAIL treatment increased the rate of cell death compared with cisplatin

alone and revealed a bimodal distribution in the timing of tumor cell death. Partial least squares regression, expected forest, logistic regression, and support vector machine analysis were also employed. The study identified key determinants of drug sensitivity, including BCL family proteins, p53 regulatory feedback loops, and the processing rate of Bcl-2 homology 3 (BH3) domain proteins [30].

One of AI's most transformative applications is in virtual screening and molecular design. AI-driven algorithms now enable rapid in silico screening of millions of compounds, predicting binding affinities and interactions with target proteins at unprecedented speed. Tools such as graph neural networks (GNNs), variational autoencoders (VAEs), and reinforcement learning models are being leveraged to generate novel, drug-like molecules optimized for specific biological targets. A notable example is Exscientia, which advanced AI-designed drug candidates to clinical trials in less than one year, highlighting the translational potential of these approaches. As reviewed by Lavecchia (2021), deep learning models have significantly improved the efficiency and accuracy of structure-based virtual screening [31, 32].

AI also plays a critical role in predicting ADMET properties, Absorption, Distribution, Metabolism, Excretion, and Toxicity. Early identification of compounds with unfavorable pharmacokinetic or toxicity profiles is essential to reducing late-stage failure rates and development costs. Machine-learning models trained on experimental ADMET data help flag potentially unsafe compounds and prioritize those with desirable properties. Recent studies, including work by Wu et al., demonstrate that AI-based toxicity prediction, particularly using ensemble and deep learning approaches—often outperforms traditional QSAR models across diverse chemical spaces [33, 34].

Despite the promising advances, the integration of AI into drug discovery pipelines still faces several challenges. Limitations related to data quality, dataset bias, biological relevance, and model interpretability can hinder the reliability and translational impact of AI-driven predictions. Many AI models operate as “black boxes,” lacking transparency, which complicates experimental validation and regulatory acceptance. As highlighted by Walters & Barzilay [35], the development of explainable AI (XAI) frameworks is critical for building trust and facilitating adoption in clinical and regulatory settings. Looking ahead, the successful application of AI in drug discovery will depend on continued collaboration between computational scientists, chemists, and biologists. Bridging these disciplines will be key to

harnessing AI's full potential, enabling smarter, faster, and more ethical drug development.

4 AI tools in phytochemical for drug discovery

AI can support various stages of the drug discovery process, for example, by analyzing spectrometry data, performing automated searches across chemical and biological databases, and predicting compound identities. Once candidate compounds are identified, AI-based tools can further assist in toxicity prediction, target interaction analysis, and mechanism-of-action elucidation, which are critical steps in both anticancer drug development and translational research (Fig. 1).

Machine learning (ML) models, including supervised, unsupervised, and semi-supervised approaches, have been widely applied and analyze data related to natural compounds with potential anticancer activity, providing insights and predicting properties of novel compounds [36, 37]. ML typically operates based on labeled or structured data, which determines the nature of the learning process as summarized in Table 2. A more specialized subset of ML is deep learning (DL), which employs multi-layered neural networks capable of automatically extracting informative features from raw molecular or biological data. This hierarchical representation allows DL models to capture complex, non-linear relationships that are often inaccessible to traditional ML approaches. As a result, DL methods have shown particular promise in drug discovery, especially when dealing with high-dimensional biological data and structurally complex natural products [38].

Table 3 provides an overview of several ML-tools used to predict chemical properties and molecular interactions. These ligand-based approaches rely primarily on chemical similarity, molecular descriptors, or learned structure–activity relationships derived from curated bioactivity databases. DEcRyPT (Drug–Target Relationship Predictor) is based on random forest regression that uses the ChEMBL database as a training set and from there predicts continuous affinity values from molecular descriptors [44].

PASS is a target fishing model that predicts a spectrum of biological activity based on structural patterns using a Multilevel Neighborhoods of Atoms (MNA) substructure descriptor algorithm. For each biological activity, numerical values, Pa and Pi, are provided, corresponding to the probability of the compound exhibiting that activity or not, respectively. Pramey and Raj et al. used PASS to predict the biological activities of *Azadirachta indica* A. Juss, for each phytoconstituent, a series of activities were listed, with their respective Pa and Pi values [45]. Kadir et al. used PASS to predict the biological activities of *Vitex negundo*, the antioxidant and

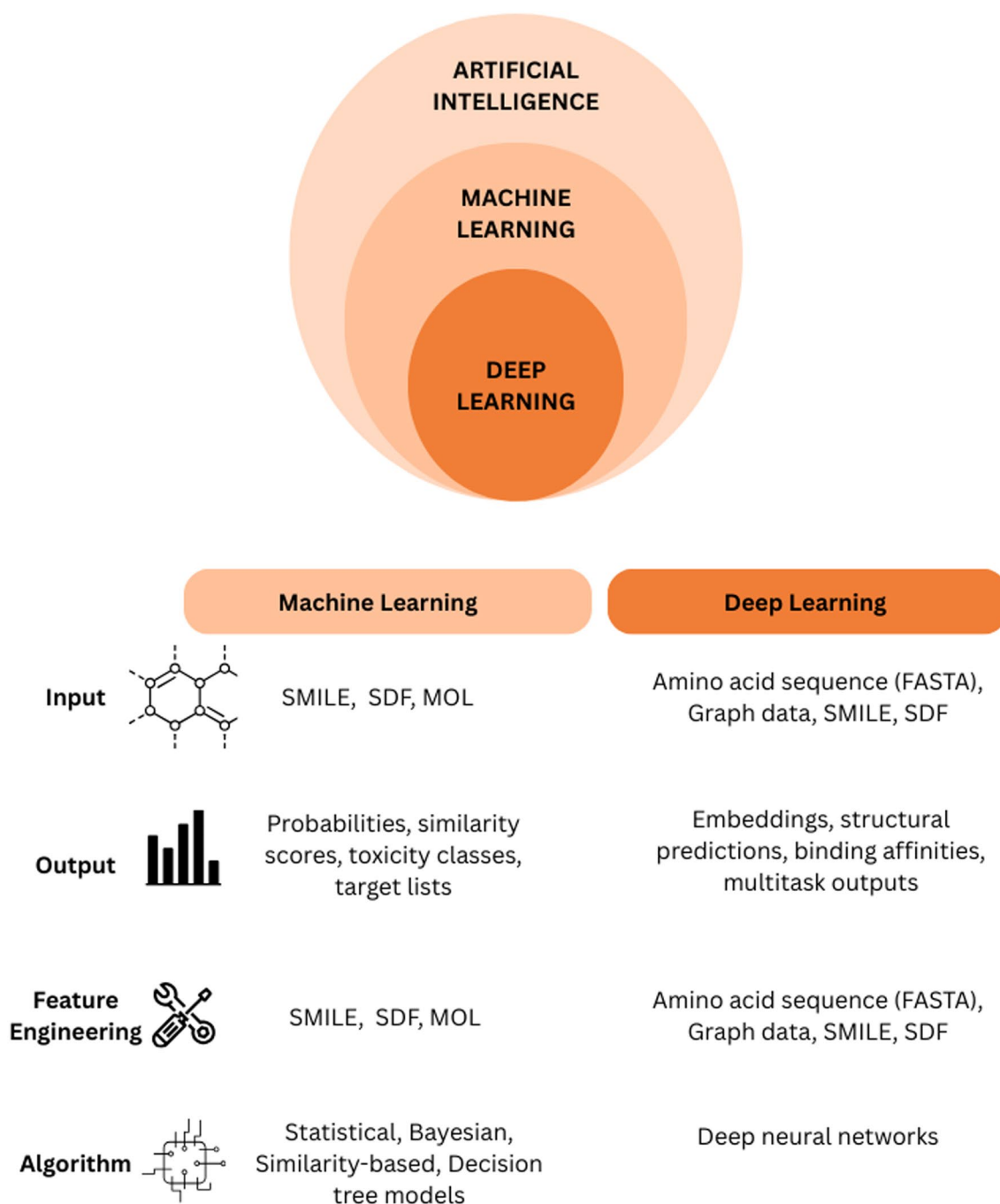


Fig. 1 Comparison between Machine Learning and Deep Learning approaches in artificial intelligence for molecular and biological data. The concentric diagram illustrates the hierarchical relationship between Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL). The table compares ML and DL based on four aspects: input data, output predictions, feature engineering, and algorithms.

antiproliferative activities were experimentally validated [46].

ProTox-II allows rapid toxicity prediction based on the structural similarity to experimentally characterized compounds combined with toxicophore-based modeling.

This approach allows early-stage safety assessment and prioritization of natural compounds before experimental validation. Asevedo et al. used the model to predict the LD50 of natural compounds in a Review Paper that highlighted the pharmacological properties of *Caesalpinia*

Table 2 Machine learning and deep learning

Machine learning		
Methods	Description	Refs.
Supervised learning	Predictions are made based on a set of labeled data by modeling the relationship between a set of variables (features or predictors) and the output variable of interest	[39]
Unsupervised learning	Identify underlying dimensions, components, clusters, or trajectories within a data structure	[40]
Semi-supervised learning	Uses sparsely labeled data and a large amount of auxiliary unlabeled data often drawn from the same underlying data distribution as the labeled data	[41]
Deep learning		
Methods	Description	Refs.
Deep neural networks	Artificial neural network that is incorporated with many layers amid input layers and output layers of the network	[42]
Graph-based approaches	Table structure reconstruction steps from detected building components	[43]
Natural language processing	Applied to extract information on DDIs, as well as on drug–drug interactions, from unstructured text sources, such as the scientific literature, electronic health records, and drug labels	[36]
Text mining	Extracting structured information from unstructured text	
Generative modeling-based approaches	Learning the data's underlying distribution with the intent of generating new samples	

Table 3 Machine learning–based target fishing tools

Software	Strategy	Algorithm	Description	Website	Refs.
DEcRyPT	Ligand-based	Random forest (Regression)	Drug–target interactions prediction	–	[44]
PASS Online	Ligand-based	Neural network (Feed-forward, Classical ANN)	Biological activity prediction	https://www.way2drug.com/passonline/	[52]
ProTox-II	Ligand-based virtual screening	Random forest (Ensemble Methods)	Toxicity prediction	https://tox.charite.de/protox3	[53]
SEA	Ligand-based	Similarity-based method (Fingerprint comparison)	2D molecular similarity	https://sea.bkslab.org/	[54]
SPiDER	Ligand-based	Similarity-based method (Pharmacophore descriptors)	Drug Equivalence Relationships	–	[50]
STP	Ligand-based	Machine learning (k-Nearest neighbors, logistic regression, 2D/3D similarity)	2D and 3D molecular similarity	http://www.swisstargetprediction.ch	[55]

sappan [47]. Boukhira et al. used ProTox to predict toxicological endpoints such as hepatotoxicity, carcinogenicity, immunotoxicity, mutagenicity, and other activities of *Thymus broussonetii* Boiss. essential oil [48].

SEA (Similarity Ensemble Approach) is based on chemical similarity between known ligands, linking 2D similarity with statistical significance for each target using bioactivity data derived from ChEMBL. Mayr et al. evaluated potential molecular targets of known natural products using three target prediction tools, including SEA. The authors incorporated statistical calculations across the three tools to summarize the evidence, and finally, the in silico findings were experimentally confirmed [49].

SPiDER (Self-organizing map-based Prediction of Drug Equivalence Relationships) predicts molecular targets based on pharmacophoric and physicochemical

descriptors, generating ranked lists of potential protein targets [50]. Rodrigues et al. used SPiDER to identify potential targets of β -lapachone and refined predictions using DEcRyPT. Experimental validation confirmed the inhibition of a tumor-associated enzyme, illustrating the complementary use of multiple AI tools in natural product research [44].

Swiss Target Prediction (STP) is a ligand-based approach combining 2D and 3D metrics with ML classifiers. Adnan et al. STP and SEA to identify molecular targets of *C. sappan* phytochemicals, linking predicted targets to genes associated with type 2 diabetes mellitus and performing protein–protein interaction network analysis [51].

Table 4 presents deep learning–based tools for predicting molecular properties, drug–target interactions, and biological activities. AlphaFold2 is a deep-learning model

Table 4 Deep learning–based target fishing tools

Software	Strategy	Algorithm	Description	Website	Refs.
AlphaFold	Structure-based	Deep neural network (Transformer + Convolutional Neural Network)	Protein 3D structure prediction	https://alphafold.ebi.ac.uk/	[67]
ChemBERTa	Ligand-based	Transformer (BERT adapted for SMILES)	Biological activity prediction	https://deepchem.io/tutorials/transfer-learning-with-chemberta-transformers/	[68]
DeepChem	Framework	Random Forest, Support Vector Machine, Convolutional Neural Network, Graph Neural Network, Transformer, etc	Molecular property prediction, Drug-target interaction prediction, Toxicity prediction, Bioactivity prediction, etc	https://deepchem.io/	[69]
DeepNPD	Ligand-based + network-based	Ensemble de Deep Neural Networks	Synergistic anticancer potential	–	[60]
PyTorch Geometric (PyG)	Graph-based molecular modeling	Graph Neural Network	Drug-target interaction prediction, Bioactivity prediction	https://graph-neural-networks.github.io/	[61]
MolBERT	Ligand-based	Transformer (BERT adapted for SMILES)	Biological activity prediction	https://github.com/BenevolentAI/MolBERT	[70]
MT-DTI	Ligand + target-based	Transformer (Molecule Transformer)	Drug-target interaction prediction	https://github.com/deargen/mt-dti	[71]
STarFish	Ligand-based	Deep Neural Network (Multi-task, Fully Connected)	Biological activity prediction	https://github.com/ntcokroft/STarFish	[66]

for protein structure prediction. Although not specific to natural products, AlphaFold2 provides near-experimental protein structures that substantially improve structure-based studies of phytochemical–target interactions [56, 57].

Jung et al. demonstrated the effectiveness of the ChemBERTa by predicting ADMET properties of DrugBank compounds, showing its utility for early-stage screening and optimization [58]. Zhou et al. compiled a library of natural compounds with therapeutic potential for Alzheimer's disease and applied DeepChem to predict protein–ligand affinity, ADMET properties, and computational ranking [59].

Cao developed DeepNPD to predict synergistic natural products acting on shared targets via distinct mechanisms. Notably, the model demonstrated robustness when trained on the HERB database, performing well despite limited data availability and high chemical complexity characteristic of natural products [60].

PyTorch Geometric (PyG) supports graph neural networks (GNNs) that represent molecules as graphs, enabling accurate modeling of atomic connectivity. When integrated with generative graph models such as GraphVAE, GraphGAN, and MolGAN, PyG facilitates the generation of optimized molecular derivatives beyond traditional fingerprint-based approaches [61]. PyTorch Geometric is a tool for Graph Neural Networks (GNN). Because it receives data structured as graphs, it is highly flexible, with considerable architectural freedom allowing

the creation of custom Graph Neural Networks. In the context of drug discovery, it can predict the biological activities of new molecules. When combined with other generative graph models (GraphVAE, GraphGAN, MolGAN), it can generate molecules with optimized derivatives. The advantage is that by using graphs, the GNN captures the true structure and connectivity of atoms, something that traditional methods based on fingerprints or fixed vectors cannot fully represent [61].

MolBERT is a SMILES-based language model trained to generate molecular embeddings for bioactivity prediction. Norinder demonstrated that MolBERT outperformed conventional methods for toxicity prediction, achieving higher sensitivity, specificity, and balanced accuracy [62]. Related models, including SMILES-BERT and MolCLR, have also shown strong performance in predicting physicochemical properties such as lipophilicity and solubility [63]. Li and Jiang further reported that MolBERT outperformed other models on standard benchmarks, including BBBP, Tox21, SIDER, and ClinTox datasets [64]. MolBERT is a SMILES-based language model that generates molecular representations (embeddings) and is trained to represent molecules and predict bioactive properties. Norinder compared MolBERT with traditional approaches to predict toxicity endpoints, and the tool performed better with high sensitivity, specificity, and balanced accuracy [62]. In addition to MolBERT, other models such as SMILES-BERT and MolCLR are also highlighted for showing high accuracy in predicting physicochemical

properties, such as lipophilicity, solubility and distribution coefficient [63]. Li and Jiang also demonstrated that MolBERT outperformed other tested models for predicting four standard datasets: Blood–Brain Barrier Penetration (BBBP), toxicity (Tox21), Side Effect Resource (SIDER), and clinical toxicity (ClinTox) [64].

Lee et al. applied MT-DTI to identify natural compounds targeting the TRPV ion channel, predicting the flavonoid troxerutin as a promising antagonist. These predictions were subsequently validated through molecular docking, in vitro assays, and a small-scale clinical trial [65]. Lee et al. used MT-DTI to identify potential natural product molecules with antagonistic activity on the TRPV ion channel. The flavonoid troxerutin was predicted to be a promising antagonist, which was corroborated by molecular docking, in vitro assays, and a small-scale clinical trial [65].

The application of ML and DL tools to natural product target fixing faces two major challenges: limited representation of phytochemicals in public databases and high structural complexity. Consequently, models trained predominantly on synthetic compounds may exhibit reduced accuracy when applied to natural products. To address this challenge, Cockcroft et al. developed a model-stacking approach, combining multiple algorithms with a meta-classifier for reverse prediction. Although trained

on synthetic compounds from ChEMBL, STarFish was evaluated using a natural product benchmark and demonstrated improved prediction performance for nuclear receptors, enzymes, and GPCRs [66].

Table 5 presents selection of bioinformatics tools designed in the analysis of mass spectrometry. These tools use AI or advanced computational algorithms to annotate fragmentation spectra, assign molecular formulas, and suggest candidate compound identities, thereby supporting large-scale metabolomic and phytochemical studies. In addition to compound identification, these tools are commonly used for data sharing and statistical analyses, allowing for the quantification and comparison of compounds across samples.

Such tools frequently interface with plant-specific and metabolomic databases that provide molecular, genomic, taxonomic, and phylogenetic context, including HMDB, MassBank, METLIN, GNPS, ReSpec, and MoNA [72]. For proteomics, UniProt, PRIDE, PeptideAtlas, and Mascot support peptide and protein identification, while lipidomics resources such as LIPID MAPS and LipidBlast facilitate lipid annotation. Databases including DrugBank, T3DB, and PubChem further support the annotation of xenobiotics and bioactive compounds.

Studies in the literature have shown that the use of these tools is promising. Studies by Mu et al. evaluated the primary targets that herbs present in the

Table 5 Bioinformatics tools specialized in the analysis of mass spectrometry data

Function	Software	Details	Website	Refs.
Annotation of molecular fragments	MS2LDA	Decompose sets of molecular fragmentation	ms2lda.org	[73]
	MAGMA	Automatic chemical annotation of mass spectrometry data	https://research-software-directory.org/software/magma	[74]
Compound identification	SIRIUS 4	Identify metabolites using single and tandem mass spectrometry	https://bio.informatik.uni-jena.de/software/sirius-4-0-1/	[75]
	CFM-ID	Predict, annotate and interpret tandem mass (MS/MS) spectra of small molecules	https://cfmid.wishartlab.com	[76]
	MetFrag	Identification of the structure of molecules by annotation of high-precision tandem mass spectra of metabolites	https://msbi.ipb-halle.de/MetFrag/	[77]
	MS-FINDER	Annotation program that supports EI-MS (GC/MS) and MS/MS spectral mining	https://systemsomicslab.github.io/comprms/msfinder/main.html	[78]
Spectrum sharing and fragmentation	GNPS	Share raw, processed, or annotated fragmentation mass spectrometry data (MS/MS)	https://gnps.ucsd.edu/ProteoSAFe/static/gnps-splash.jsp	[79]
Statistical analysis and interpretation	MetaboAnalyst	Metabolomic data analysis and interpretation of NMR and MS data	https://www.metaboanalyst.ca/	[80]
Visualization and quantification	MZmine 2	Visualization and basic statistics analyses	https://mzio.io/	[81]
	XCMS	Used to process LC/MS-based metabolomic data	https://xcmsonline.scripps.edu/landing_page.php?pgcontent=mainPage	[82]

Sargentodoxa cuneata formulation commonly used in traditional Chinese medicine, and the results obtained in silico corroborated the results obtained in vitro [83]. The same tools were developed to elucidate the mechanisms of action of Bazhen Decoction [84] and Quercetin [85], and both were validated in vitro using different tumor cell lines.

AI tools have also enabled the identification of compound–target affinities and guided molecular optimization. For example, indirubin was identified as a key inhibitor of RPS20 in colorectal cancer, and AI-guided optimization led to the design of 20 derivatives with enhanced activity [86]. Another study screened 18 varieties of date pits using AI methods, revealing substantial variability in bioactive composition and functional properties with prediction accuracies reaching 92.57% [87]. Collectively, these studies demonstrate the ability of AI to integrate bioactivity, bioavailability, and tissue-specific information into phytochemical research pipelines [88].

The tools and studies discussed illustrate that AI enables large-scale bioactivity prediction, toxicity assessment, structural elucidation, and data-driven discovery. Ligand-based platforms such as PASS Online, ProTox-II, SEA, SPiDER, STP, and DEcRyPT generally exhibit good predictive performance on benchmark datasets (AUC $\approx$ 0.80–0.90), but their accuracy declines for structurally novel phytochemicals due to limited chemical space coverage and increased false-positive rates [53, 54, 89]. Deep learning approaches, including ChemBERTa, MolBERT, DeepChem, DeepNPD, PyG, MT-DTI, and STarFish, consistently outperform traditional fingerprint-based models in QSAR and target prediction benchmarks, achieving up to 10% improvement in predictive metrics, albeit with higher computational cost and data requirements [68, 70, 90]. AlphaFold has further revolutionized protein structure prediction with near-experimental accuracy, indirectly supporting phytochemical research through improved protein–ligand interaction modeling [91]. For metabolomics, GNPS, SIRIUS 4, CFM-ID, MetFrag, and MS-FINDER demonstrate high annotation accuracy for high-resolution MS/MS data, while MS2LDA and MAGMa provide complementary substructure-level insights [75, 76, 79, 92]. Despite these advances, there remains a strong need for more robust, interpretable, and natural-product-specific AI tools, particularly for poorly characterized or novel phytochemicals [93].

5 Proposed AI-driven framework for phytochemical discovery

Artificial intelligence (AI) has increasingly been applied in drug discovery to integrate machine learning, molecular modeling, and large-scale biological data analysis, thereby enabling faster and more cost-effective identification of bioactive compounds [94, 95]. However, most AI-driven discovery pipelines have been created and verified using datasets that are primarily composed of synthetic small molecules, which restricts how well they work with natural products and phytochemicals that have more complex structures, extensive stereochemistry, and frequent multi-target pharmacological effects [11, 96]. Numerous studies have demonstrated that traditional molecular representations, drug-likeness filters, and single-target screening methods frequently fail to adequately include the distinctive chemical space and biological properties of plant-derived compounds [97]. As a result, AI models optimized for synthetic libraries often show reduced generalizability when applied to phytochemicals. There is an increasing agreement in the literature that phytochemical discovery necessitates AI frameworks particularly tailored to incorporate natural product–focused data, multi-target bioactivity prediction, and experimental feedback mechanisms [98, 99].

This section describes NP-ScreenR, an AI-driven framework that integrates natural product knowledge, metabolomics, and multi-level computational screening to systematically prioritize plant-derived anticancer phytochemicals (Fig. 2) [100]. NP-ScreenR is designed as a modular, decision-guided pipeline in which each stage addresses a specific limitation of conventional natural product discovery workflows identified in earlier sections.

5.1 Plant/NP source

Plant-derived natural products remain central to cancer therapy, with approximately 60% of approved anticancer drugs derived from or inspired by phytochemicals. Consequently, selecting promising plant sources is a crucial initial step. AI approaches now integrate ethnopharmacological knowledge with large chemical and bioactivity databases to prioritize plants with high anticancer potential. By learning patterns from historical usage, chemical composition, and genomic information, machine learning models can rank plant species or extracts according to predicted therapeutic relevance. This integration streamlines the traditionally slow discovery process and provides a focused, data-driven foundation for NP-ScreenR, directing subsequent analyses toward high-value natural sources [101].

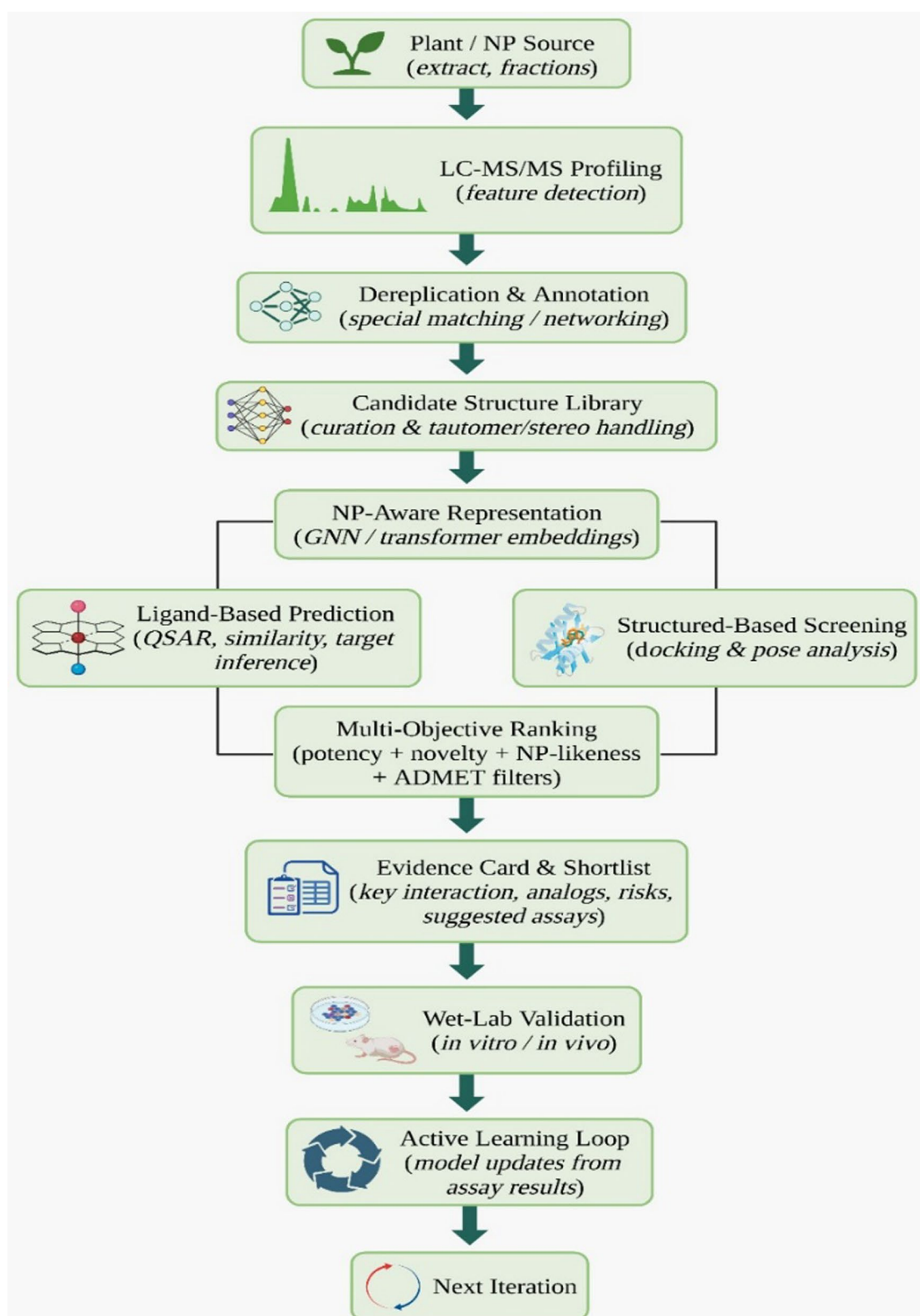


Fig. 2 Artificial intelligence-based workflow for natural product drug discovery. The pipeline integrates LC MS MS metabolite profiling, compound dereplication and annotation, AI based molecular representation, ligand-based prediction, structure based virtual screening, and multi objective ranking to prioritize candidates for experimental validation and iterative model refinement

5.2 LC–MS/MS profiling

After a plant source is chosen, NP-ScreenR uses LC–MS/MS to rapidly profile its chemical constituents. High-resolution MS generates complex metabolite spectra, which are interpreted using AI-assisted metabolomic workflows. Molecular networking platforms such as GNPS cluster MS/MS spectra by similarity, revealing compound families and enabling early dereplication by linking unknown features to known natural products. In parallel, machine learning models trained on large spectral libraries infer structural features from fragmentation patterns and rank candidate structures. Together, these AI-integrated approaches convert raw MS/MS data into annotated, prioritized molecular features efficiently [72, 102].

5.3 Dereplication and annotation

At this stage, NP-ScreenR focuses on rapid dereplication and annotation to prevent rediscovery of known compounds. Dereplication determines whether detected molecules correspond to previously reported structures. AI-based tools such as SIRIUS combined with CSI:FingerID predict molecular formulas and substructure fingerprints from MS/MS data using machine learning, then match these predictions against natural product databases to generate ranked identifications [72, 103]. In parallel, spectral library matching and molecular networking approaches annotate compounds based on similarity to known metabolites. Additional AI classifiers can assign biosynthetic or scaffold classes, providing contextual information about compound origin and potential bioactivity. As a result, this step yields a curated list of candidates, separating known compounds from potentially novel natural products prioritized for further study [102, 104].

5.4 Candidate structure library

All distinct structures obtained after dereplication, both known compounds and predicted novel ones, are assembled into a candidate structure library representing the chemical space for in silico screening. NP-ScreenR expands this library by integrating public natural product databases, such as large curated repositories of plant-derived compounds, to improve coverage and leverage existing bioactivity annotations. To further explore underrepresented regions of natural product chemical space, AI-based generative models can propose NP-like virtual molecules with realistic structural features. By combining experimentally observed compounds with database entries and AI-generated analogs, NP-ScreenR produces a curated and comprehensive library ready for downstream computational bioactivity screening [105, 106].

5.5 NP-aware representation

Before virtual screening, NP-ScreenR converts each molecule into an NP-aware representation suitable for AI analysis. Because natural products have complex scaffolds, stereochemistry, and biosynthetic features, standard descriptors alone are insufficient. Deep learning models such as transformer-based SMILES encoders generate dense embeddings that capture subtle structural patterns, while NP-specific classifiers encode biosynthetic class information directly into the representation. In parallel, graph-based embeddings and classical fingerprints are used to preserve atomic connectivity and topological information. By integrating general chemical embeddings with NP-informed features, NP-ScreenR provides a comprehensive molecular representation that supports accurate downstream bioactivity prediction [90, 104].

5.6 Ligand-based prediction

Using the derived molecular representations, NP-ScreenR applies ligand-based prediction to prioritize candidates with potential anticancer activity. QSAR and machine learning models trained on large bioactivity datasets estimate the likelihood of cytotoxic or anticancer effects based on NP-aware descriptors. In parallel, target-specific ligand-based screening is performed using similarity-based target fishing tools, which compare candidate structures to known ligands to infer probable protein targets. This stage is particularly effective for identifying phytochemicals that resemble known inhibitors of cancer-relevant targets such as kinases, topoisomerases, or signaling proteins. This ligand-based stage efficiently narrows the candidate pool to phytochemicals with promising predicted activities, guiding subsequent structure-based evaluation [90, 107].

5.7 Structure-based screening

As a complement to ligand-based approaches, NP-ScreenR applies structure-based methods to assess how well each candidate interacts with specific cancer targets. Top-ranked phytochemicals are docked in silico into three-dimensional protein structures relevant to cancer biology. When experimental structures are unavailable, high-quality AI-predicted models enable inclusion of otherwise inaccessible targets. Docking algorithms position each compound within the binding site and estimate binding strength, while accounting for the size and complexity typical of natural products. Promising complexes may be further refined using AI-assisted scoring or rapid simulation approaches. This step connects phytochemicals to plausible molecular targets and prioritizes those with strong predicted binding for further study [10, 91].

5.8 Multi-objective ranking

At this stage, NP-ScreenR typically yields several promising phytochemicals with different strengths, such as predicted potency, novelty, or drug-likeness. To prioritize the most viable leads, the framework applies multi-objective ranking that balances key drug discovery criteria. Rather than focusing on activity alone, NP-ScreenR integrates predicted efficacy, target selectivity, ADMET properties, and novelty into a unified evaluation using multi-criteria decision analysis. Algorithms such as Pareto ranking identify candidates that achieve optimal trade-offs. Pareto-based ranking allows flexible prioritization strategies, enabling researchers to emphasize safety, novelty, or potency depending on project goals. This AI-driven prioritization ensures that top hits combine biological promise with practical development potential [108, 109].

5.9 Evidence cards and shortlist

Following ranking, NP-ScreenR produces a shortlist of top-priority phytochemical candidates. To support transparent decision-making, each candidate is summarized in an Evidence Card integrating key predictions, including chemical structure, source information, predicted targets, docking results, and ADMET flags. Where available, supporting literature, analog bioactivity, and omics evidence are also incorporated. These cards provide a concise yet comprehensive rationale for selection, resulting in a well-supported shortlist recommended for experimental validation [90].

5.10 Wet-lab validation and active learning loop

In the final stage, NP-ScreenR transitions from computational prediction to experimental validation. Shortlisted phytochemicals are extracted or synthesized and tested using *in vitro* and *in vivo* assays to confirm anticancer activity and target engagement. Crucially, NP-ScreenR incorporates an active learning loop in which experimental outcomes are fed back into the AI models. Both successful hits and false positives are used to retrain prediction and ranking algorithms, improving performance over time. This iterative integration of computation and experimentation enables continuous refinement of the framework and accelerates the discovery of effective plant-derived anticancer agents [101, 110].

6 Limitations and future perspectives

6.1 Quality, scarcity and inadequate experimental data integration

AI has become integral to modern drug discovery, supporting key stages such as target identification,

compound screening, lead optimization, ADMET prediction and even clinical trial design. Despite these advances, several persistent barriers continue to limit the full potential of AI-driven approaches. Chief among these is the challenge of obtaining high-quality, consistent, and representative experimental data. The complex nature of drug–target interactions within biological systems often defies simple binary classification, complicating the training of supervised learning models [111, 112].

Furthermore, the volume of available high-quality data is still limited, especially when compared to the large amount of potentially usable but currently inaccessible or unstructured biological and chemical information. This imbalance restricts the ability of AI models to learn generalizable patterns across diverse chemical spaces. Rodrigues also reported that current datasets inadequately present full diversity of natural products. Numerous plant-derived compounds exhibit unique structural characteristics that are infrequently found in conventional drug-like molecule libraries [98]. Therefore, AI models trained on such datasets often exhibit reduced predictive reliability when applied to phytochemicals.

AI systems also face limitations arising from incomplete, inconsistent, and insufficiently labeled datasets, compounded by the inherent biological complexity of cancer systems and challenges related to model interpretability. Deep learning models are highly-data dependent and without sufficient, high-quality data, their predictive reliability and long-term utility remain constrained [113].

Another major challenge lies in data heterogeneity and inconsistency across studies. As Bender and Cortés-Ciriano noted, the quality of data within natural product databases is often compromised by inconsistencies, missing annotations, and variability in experimental conditions. These issues complicate data integration and significantly hinder the development of robust, generalizable models [114].

Addressing these limitations will require rigorous data standardization, improved curation practices, and enhanced experimental reproducibility. As emphasized by Gaulton et al., the implementation of standardized data protocols is essential to ensure the reliability of AI-driven predictions and to facilitate meaningful comparisons [115, 116]. Only through improved data quality and integration strategies can the full potential of AI in natural product-based drug discovery be realized.

6.2 Data biases of AI models and transparency

Bias in AI can significantly undermine the reliability of its outcomes, often emerging at multiple stages of model development. These biases typically stem from flawed algorithm or issues within the training data and can lead

to inequities in application [117]. In the context of drug discovery, AI models are particularly vulnerable to bias due to the limitations of publicly available biological and pharmacological datasets, which tend to overrepresent well-studied compounds. This overemphasis restricts the exploration of novel therapeutic agents and may hinder innovation [118].

Kumar et al. [119] also highlighted that dataset imbalances, biased feature selection, and a general lack of transparency in model design can further compromise both the fairness and accuracy of AI systems. Moreover, inconsistencies between data sources across experimental platforms such as variations in *in vitro* and *in vivo* studies results can weaken the reproducibility and trustworthiness of AI-generated predictions [120].

Mitigating these issues requires deliberate efforts to diversify training datasets, standardize data collection protocols, and improve transparency in model development. Such measures are essential for enhancing the robustness, fairness, and credibility of AI applications in phytochemical drug discovery [121].

6.3 Limited model interpretability and Mechanistic insights

The lack of interpretability in AI models continues to be a major barrier to their widespread adoption in drug discovery, as it diminishes confidence in their predictions [122]. Many advanced AI systems are frequently described as “black boxes”, meaning their internal decision-making processes are opaque and difficult to understand. Although methods such as visualization tools and attention mechanisms have been introduced to improve model interpretability, achieving a clear and comprehensive understanding remains a significant hurdle [123].

For example, Zhang et al. developed an AI-based framework to classify meridians by analyzing the topological structures of anti-cancer phytochemicals. While the model demonstrated potential, the authors acknowledged a key limitation including the absence of interpretability analysis related to meridian tropism. They emphasized the importance of integrating chemical profiling with advanced target prediction methods to improve both predictive accuracy and mechanistic insight, highlighting a recurring challenge in AI-based phytochemical research [124].

6.4 Computational resources and expertise constraint

AI and ML hold significant promise for accelerating the discovery of plant-derived anticancer agents by rapidly predicting bioactivity and compound optimization through advanced computational approaches [125]. Computational methods integrated with experimental

validation and advanced AI have the potential to revolutionize drug development by optimizing design and improving clinical success rate [126].

However, the development and deployment of AI models are hindered by several technical and resource-related barriers. These include the need for advanced programming expertise, large volumes of well-annotated training data, and access to high-performance computing infrastructure [127]. Ghosh et al. reviewed a range of computational techniques commonly used in phytochemical research including MDS, MM PBSA, ADMET profiling and molecular docking, highlighting both their value and their limitations. They emphasized ongoing challenges such as low accuracy, insufficient datasets, and the critical need for experimental validation to support computational findings [128].

Moreover, Varghese et al. documented that a significant barrier to AI adoption in the field of phytochemicals is the limited availability of user-friendly, accessible software platforms [72]. The lack of accessible tools and computational training among experimental researchers further restricts the widespread implementation of AI-driven discovery pipelines.

6.5 Future prospects

Artificial intelligence (AI) and machine learning have significantly advanced data interpretation and prediction modeling in biomedical research. As these technologies continue to evolve, their integration with multi-omics datasets—including genomics, transcriptomics, proteomics, and metabolomics—through multimodal learning frameworks is expected to further advance disease modeling, diagnosis, and personalized therapy development [129].

In particular cross attention transformer models [130], variational autoencoders [36] and Multiomics factors analysis (MOFA) [131] can be employed to fuse heterogeneous omics layers and identify robust biomarkers and therapeutic targets. Such holistic approaches offer a powerful strategy for discovering plant-derived compounds that selectively target cancer-specific pathways [132].

AI has also substantially advanced phytochemical research by enabling detailed analysis into biosynthetic pathways and metabolite profiling. One of its most notable contributions lies in the identification of novel therapeutic compounds for cancer treatment [100]. Vision transformers (ViTs) and CNN-transformers hybrids can be applied to spectral and imaging data for identification [133] while graph convolutional Networks (GCN), recurrent neural network (RNN), Message passing Neural Networks (MPNN) and graph attention network (GAT) can improve structure activity

relationship model of phytochemicals [124, 134]. In addition, SMILES-based transformers and diffusion models are increasingly used for de novo compound generation and optimization of anticancer activity [135–137].

For predictive screening multitasking deep learning architecture and gradient boosting models (XG Boost) can be implemented for simultaneously activity and ADMET prediction [36, 138]. Docking and binding pose refinement can be enhanced using AI assisted docking framework such as EquiBind and DeepDock thereby improving hit prioritization beyond classical scoring functions [139, 140]. Together, these advances are accelerating AI-driven identification of plant-based compounds for precision oncology.

To fully realize this potential, standardization and sharing of phytochemical-related multi-omics data should be prioritized through international collaborations. Secure and privacy-preserving data sharing may be facilitated through federated learning architectures and blockchain-enabled repositories, enabling cross-institutional collaboration without compromising confidentiality.

Despite its promise, AI-driven phytochemical discovery also raises ethical and regulatory considerations. Bias in data, algorithms, and decision-making processes must be addressed through inclusive dataset design, transparent model development, and rigorous validation strategies [141]. Explainable AI (XAI) techniques, such as SHAP, integrated gradients, and Grad-CAM, are increasingly important for improving transparency and reproducibility [142]. At the same time, the intersection of AI with intellectual property rights and regulatory frameworks is reshaping the governance of drug discovery [143].

Ultimately, multidisciplinary collaboration among computational scientists, biologists, clinicians, legal experts, and regulatory authorities will be essential to overcome both technical and ethical challenges. Through such partnerships, AI enabled phytochemical screening can evolve into a scientifically robust and ethically responsible platform for discovering targeted cancer therapies, ultimately shaping the future of precision oncology [144].

7 Conclusion

Artificial intelligence is reshaping the discovery of natural product-derived antitumor agents by enabling faster compound screening, more reliable pharmacological predictions, and more efficient optimization throughout the drug discovery process. By combining machine learning and deep learning approaches with multi-omics data and advanced bioinformatics tools, AI has greatly

expanded our ability to explore the chemical diversity of phytochemicals and to identify candidates with genuine anticancer potential.

When integrated with metabolomic, proteomic, and cheminformatic analyses, AI-driven strategies provide a more systematic and informed approach to prioritizing bioactive compounds and understanding their possible mechanisms of action. This represents a meaningful shift away from traditional trial-and-error screening toward more targeted and data-driven phytochemical discovery.

When integrated with metabolomic, proteomic, and cheminformatic analyses, AI-driven strategies provide a more systematic and informed approach to prioritizing bioactive compounds and understanding their possible mechanisms of action. This represents a meaningful shift away from traditional trial-and-error screening toward more targeted and data-driven phytochemical discovery.

Despite these clear advantages, important challenges remain. Issues related to data quality and standardization, limited model interpretability, computational demands, and ethical and regulatory considerations continue to limit the widespread adoption of AI in phytochemical drug discovery. Addressing these challenges will require improved data curation practices, the development of transparent and explainable AI models, and closer collaboration between computational scientists, experimental researchers, and clinicians.

Overall, AI-driven phytochemical screening stands out as both a necessary and promising strategy for the development of new chemotherapeutic agents. As AI technologies mature and become more accessible, their thoughtful integration with experimental validation and clinical insight is likely to play an increasingly important role in advancing precision oncology and natural product-based cancer therapies.

Acknowledgements

Authors would like to thank Federal University of São João del-Rei (Brazil), Universitas Sumatera Utara, (Indonesia) and Kyung Hee University (Republic of Korea).

Author's contribution

Livia Ramos Santiago: Writing – review and editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. Estéfani Alves Asevedo: Writing – review and editing, Writing – original draft, Visualization, Formal analysis, Data curation. Maria Eduarda Jeunon de Oliveira: Writing – original draft, Formal analysis, Data curation. Karen Cota Pereira: Writing – original draft, Formal analysis, Data curation. Maria Fernanda da Silva Trindade: Writing – original draft, Formal analysis, Data curation. Ana Gabriela Silva Oliveira: Writing – original draft, Formal analysis, Data curation. Marina Andrade Rocha: Writing – original draft, Formal analysis, Data curation. Sojin Kang: Writing – original draft, Formal analysis, Data curation. Amama Rani: Writing – original draft, Formal analysis, Data curation. Moon Nyeo Park: Writing – review and editing, Writing – original draft, Data curation. Michel William Tan: Writing – original draft, Formal analysis, Data curation. Rony Abdi Syahputra: Writing – review and editing, Writing – original draft, Data curation. Bonglee Kim: Writing – review and editing, Writing – original draft, Data curation. Rosy lara Maciel de

Azambuja Ribeiro: Writing – review and editing, Writing – original draft, Data curation, Conceptualization, Supervision.

Funding

This research was supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) (code 001), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Amparo à Pesquisa do Estado de Minas Gerais (Fapemig), Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2020R111A2066868), the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2020-NR049559), a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: RS-2020-KH087790), the Starting Growth Technological R&D Program (TIPS Program, No. RS-2024-00507224) funded by the Ministry of SMEs and Startups (MSS, Korea) in 2024, and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2024-00350362).

Data availability

Not applicable.

Declarations

Competing interests

The authors declare that they have no competing of interest.

Author details

¹Department of Experimental Pathology, Federal University of São João del-Rei, Sebastião Gonçalves Coelho Street, 400-Chanadour, Divinópolis, MG 35501-296, Brazil. ²Department of Pathology, Kyung Hee University, Seoul, Republic of Korea. ³Department of Pharmacology, Universitas Sumatera Utara, Sumatera Utara, Medan, Indonesia.

Received: 11 November 2025 Accepted: 27 January 2026

Published online: 21 April 2026

References

- Kaur R, Bhardwaj A, Gupta S. Cancer treatment therapies: traditional to modern approaches to combat cancers. *Mol Biol Rep*. 2023;50(11):9663–76.
- Nasir A, et al. Nanotechnology, a tool for diagnostics and treatment of cancer. *Curr Top Med Chem*. 2021;21(15):1360–76.
- Park MN, et al. Phytochemical synergies in BK002: advanced molecular docking insights for targeted prostate cancer therapy. *Front Pharmacol*. 2025;16:1504618.
- Islam MR, et al. Colon cancer and colorectal cancer: prevention and treatment by potential natural products. *Chem Biol Interact*. 2022;368:110170.
- Jang SY, et al. Immunomodulatory effects of a standardized botanical mixture comprising *Angelica gigas* roots and *Pueraria lobata* flowers through the TLR2/6 pathway in RAW 264.7 macrophages and cyclophosphamide-induced immunosuppression mice. *Pharmaceuticals* (Basel). 2025. <https://doi.org/10.3390/ph18030336>.
- Tewari D, et al. Natural products targeting the PI3K-Akt-mTOR signaling pathway in cancer: a novel therapeutic strategy. *Semin Cancer Biol*. 2022;80:1–17.
- Kumar S, Singh B, Singh R. *Catharanthus roseus* (L.) G. Don: a review of its ethnobotany, phytochemistry, ethnopharmacology and toxicities. *J Ethnopharmacol*. 2022;284:114647.
- Banerjee S, et al. Anticancer properties and mechanisms of botanical derivatives. *Phytomedicine Plus*. 2023;3(1):100396.
- Li G, et al. Artificial intelligence-guided discovery of anticancer lead compounds from plants and associated microorganisms. *Trends Cancer*. 2022;8(1):65–80.
- Chunarkar-Patil P, et al. Anticancer drug discovery based on natural products: from computational approaches to clinical studies. *Biomed*. 2024. <https://doi.org/10.3390/biomedicines12010201>.
- Atanasov AG, et al. Natural products in drug discovery: advances and opportunities. *Nat Rev Drug Discov*. 2021;20(3):200–16.
- Simoben CV, et al. Challenges in natural product-based drug discovery assisted with *in silico*-based methods. *RSC Adv*. 2023;13(45):31578–94.
- Iqbal J, et al. Plant-derived anticancer agents: a green anticancer approach. *Asian Pac J Trop Biomed*. 2017;7(12):1129–50.
- Ribeiro ARC, et al. *Myrciaria tenella* (DC.) O. Berg (Myrtaceae) leaves as a source of antioxidant compounds. *Antioxidants*. 2019. <https://doi.org/10.3390/antiox8080310>.
- Hashim GM, et al. Plant-Derived Anti-Cancer Therapeutics and Biopharmaceuticals. *Bioengineering*. 2025;12(1):7.
- Fujita K, et al. Irinotecan, a key chemotherapeutic drug for metastatic colorectal cancer. *World J Gastroenterol*. 2015;21(43):12234–48.
- Niu Z-X, et al. Recent advance of clinically approved small-molecule drugs for the treatment of myeloid leukemia. *Eur J Med Chem*. 2023;261:115827.
- Montgomery B, Lin DW. Chapter 10 - TOXICITIES OF CHEMOTHERAPY FOR GENITOURINARY MALIGNANCIES. In: Taneja SS, editor. *Complications of Urologic Surgery* (Fourth Edition). Philadelphia: W.B. Saunders; 2010. p. 117–23.
- Menis J, Twelves C. Eribulin (Halaven): a new, effective treatment for women with heavily pretreated metastatic breast cancer. *Breast Cancer* (Dove Med Press). 2011;3:101–11.
- Gralla RJ, et al. Oral vinorelbine in the treatment of non-small cell lung cancer: rationale and implications for patient management. *Drugs*. 2007;67(10):1403–10.
- Nazha A, et al. Omacetaxine mepesuccinate (synribo) - newly launched in chronic myeloid leukemia. *Expert Opin Pharmacother*. 2013;14(14):1977–86.
- Gallego-Jara J, et al. A compressive review about Taxol®: history and future challenges. *Mol*. 2020. <https://doi.org/10.3390/molecules25245986>.
- Chihomvu P, et al. Phytochemicals in drug discovery—a confluence of tradition and innovation. *Int J Mol Sci*. 2024;25(16):8792.
- Naeem A, et al. Natural products as anticancer agents: current status and future perspectives. *Mol*. 2022. <https://doi.org/10.3390/molecules27238367>.
- Jacob S, et al. Solid lipid nanoparticles and nanostructured lipid carriers for anticancer phytochemical delivery: advances, challenges, and future prospects. *Pharmaceutics*. 2025;17(8):1079.
- Gupta PK, et al. Phytomedicines targeting cancer stem cells: therapeutic opportunities and prospects for pharmaceutical development. *Pharmaceutics*. 2021;14(7):676.
- Sahrawat TR. Role of artificial intelligence and machine learning in sustainable drug discovery. *Braz Arch Biol Technol*. 2024;67:e24240538.
- Kim H, et al. Artificial intelligence in drug discovery: a comprehensive review of data-driven and machine learning approaches. *Biotechnol Bioprocess Eng*. 2020;25(6):895–930.
- Liu H, Tang T. MAPK signaling pathway-based glioma subtypes, machine-learning risk model, and key hub proteins identification. *Sci Rep*. 2023;13(1):19055.
- Zhang Y, et al. Designing combination therapies with modeling chaperoned machine learning. *PLoS Comput Biol*. 2019;15(9):e1007158.
- Jiménez-Luna J, et al. Artificial intelligence in drug discovery: recent advances and future perspectives. *Expert Opin Drug Discov*. 2021;16(9):949–59.
- Lavecchia A. Machine-learning approaches in drug discovery: methods and applications. *Drug Discov Today*. 2015;20(3):318–31.
- Serghini A, Portelli S, Ascher DB. AI-driven enhancements in drug screening and optimization. *Methods Mol Biol*. 2024;2714:269–94.
- Wu Y, et al. The role of artificial intelligence in drug screening, drug design, and clinical trials. *Front Pharmacol*. 2024. <https://doi.org/10.3389/fphar.2024.1459954>.
- Walters WP, Barzilay R. Critical assessment of AI in drug discovery. *Expert Opin Drug Discov*. 2021;16(9):937–947. <https://doi.org/10.1080/17460441.2021.1915982>.
- Spanakis M, et al. Artificial intelligence models and tools for the assessment of drug-herb interactions. *Pharmaceutics*. 2025;18(3):282.
- Kwak MS, et al. Development of a machine learning model for the prediction of nodal metastasis in early T classification oral

- squamous cell carcinoma: SEER-based population study. *Head Neck*. 2021;43(8):2316–24.
38. Gupta R, et al. Artificial intelligence to deep learning: machine intelligence approach for drug discovery. *Mol Divers*. 2021;25(3):1315–60.
 39. Shaer I, Shami A. Data-driven methods for the reduction of energy consumption in warehouses: use-case driven analysis. *Internet of Things*. 2023;23:100882.
 40. Jiang T, Gradus JL, Rosellini AJ. Supervised machine learning: a brief primer. *Behav Ther*. 2020;51(5):675–87.
 41. Chen Y, et al. Semi-supervised and unsupervised deep visual learning: a survey. *IEEE Trans Pattern Anal Mach Intell*. 2024;46(3):1327–47.
 42. M G, Sethuraman SC. A comprehensive survey on deep learning based malware detection techniques. *Comput Sci Rev*. 2023;47:100529.
 43. Lee E, et al. Deep-learning and graph-based approach to table structure recognition. *Multimedia Tools Appl*. 2022;81(4):5827–48.
 44. Rodrigues T, et al. Machine intelligence decrypts β -lapachone as an allosteric 5-lipoxygenase inhibitor. *Chem Sci*. 2018;9(34):6899–903.
 45. Pramely R, Raj LS. Prediction of biological activity spectra of a few phytoconstituents of *Azadirachta indica* A. Juss. *J Biochem Technol*. 2012;3(4):375–9.
 46. Kadir FA, et al. PASS-predicted Vitex negundo activity: antioxidant and antiproliferative properties on human hepatoma cells—an in vitro study. *BMC Complement Altern Med*. 2013;13:343.
 47. Asevedo EA, et al. Unlocking the therapeutic mechanism of *Caesalpinia sappan*: a comprehensive review of its antioxidant and anti-cancer properties, ethnopharmacology, and phytochemistry. *Front Pharmacol*. 2025;15:1514573.
 48. Boukhira, S., et al. 2024 *The chemical composition and the preservative, antimicrobial, and antioxidant effects of Thymus broussonetii Boiss. essential oil: an in vitro and in silico approach*. *Frontiers in Chemistry* 12 1402310.
 49. Mayr F, et al. Finding new molecular targets of familiar natural products using in silico target prediction. *Int J Mol Sci*. 2020. <https://doi.org/10.3390/ijms21197102>.
 50. Reker D, et al. Identifying the macromolecular targets of de novo-designed chemical entities through self-organizing map consensus. *Proc Natl Acad Sci U S A*. 2014;111(11):4067–72.
 51. Adnan M, et al. Network pharmacology study to reveal the potentiality of a methanol extract of *Caesalpinia sappan* L. wood against type-2 diabetes mellitus. *Life*. 2022;12(2):277.
 52. Filimonov D, et al. Prediction of the biological activity spectra of organic compounds using the PASS online web resource. *Chem Heterocycl Comp*. 2014;50(3):444–57.
 53. Banerjee P, et al. ProTox-II: a webserver for the prediction of toxicity of chemicals. *Nucleic Acids Res*. 2018;46(W1):W257–W263.
 54. Keiser MJ, et al. Relating protein pharmacology by ligand chemistry. *Nat Biotechnol*. 2007;25(2):197–206.
 55. Gfeller D, et al. SwissTargetPrediction: a web server for target prediction of bioactive small molecules. *Nucleic Acids Res*. 2014;42:W32–8.
 56. Zhang H, et al. AlphaFold2 in biomedical research: facilitating the development of diagnostic strategies for disease. *Front Mol Biosci*. 2024;11:1414916.
 57. Yang Z, et al. AlphaFold2 and its applications in the fields of biology and medicine. *Signal Transduct Target Ther*. 2023;8(1):115.
 58. Jung W, et al. Absorption distribution metabolism excretion and toxicity property prediction utilizing a pre-trained natural language processing model and its applications in early-stage drug development. *Pharmaceuticals (Basel)*. 2024. <https://doi.org/10.3390/ph17030382>.
 59. Zhou J, et al. Bioinformatics and deep learning approach to discover food-derived active ingredients for Alzheimer's disease therapy. *Foods*. 2025;14(1):127.
 60. Cao EL. Natural product based anticancer drug combination discovery assisted by deep learning and network analysis. *Front Nat Prod*. 2024;2:1309994.
 61. Fey, M. and J.E. Lenssen, *Fast graph representation learning with PyTorch Geometric*. arXiv preprint [arXiv:1903.02428](https://arxiv.org/abs/1903.02428), 2019.
 62. Norinder U. Traditional machine and deep learning for predicting toxicity endpoints. *Molecules*. 2022. <https://doi.org/10.3390/molecules28010217>.
 63. Shilpa S, Kashyap G, Sunoj RB. Recent applications of machine learning in molecular property and chemical reaction outcome predictions. *J Phys Chem A*. 2023;127(40):8253–71.
 64. Li J, Jiang X. Mol-BERT: an effective molecular representation with BERT for molecular property prediction. *Wirel Commun Mob Comput*. 2021;2021(1):7181815.
 65. Lee J, et al. Drug-Target Interaction Deep Learning-Based Model Identifies the Flavonoid Troxerutin as a Candidate TRPV1 Antagonist. *Appl Sci*. 2023;13(9):5617.
 66. Cockroft NT, Cheng X, Fuchs JR. STarFish: a stacked ensemble target fishing approach and its application to natural products. *J Chem Inf Model*. 2019;59(11):4906–20.
 67. Jumper J, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596(7873):583–9.
 68. Chithrananda, S., G. Grand, and B. Ramsundar, *ChemBERTa: large-scale self-supervised pretraining for molecular property prediction*. arXiv preprint [arXiv:2010.09885](https://arxiv.org/abs/2010.09885), 2020.
 69. Ramsundar, B., et al., *Deep learning for the life sciences: applying deep learning to genomics, microscopy, drug discovery, and more*. 2019: O'Reilly Media.
 70. Fabian, B., et al., *Molecular representation learning with language models and domain-relevant auxiliary tasks*. arXiv preprint [arXiv:2011.13230](https://arxiv.org/abs/2011.13230), 2020.
 71. Shin, B., et al. 2019 *Self-attention based molecule representation for predicting drug-target interaction*. in *Machine learning for healthcare conference PMLR*.
 72. Varghese R, et al. Artificial intelligence driven approaches in phytochemical research: trends and prospects. *Phytochem Rev*. 2025. <https://doi.org/10.1007/s11101-025-10096-8>.
 73. Wandy J, et al. Ms2lda.org: web-based topic modelling for substructure discovery in mass spectrometry. *Bioinformatics*. 2018;34(2):317–8.
 74. Ridder L, van der Hooft JJJ, Verhoeven S. Automatic compound annotation from mass spectrometry data using MAGMa. *Mass Spectrom*. 2014;3(Special_Issue_2):S0033–S0033.
 75. Dührkop K, et al. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat Methods*. 2019;16(4):299–302.
 76. Wang F, et al. CFM-ID 4.0 – a web server for accurate MS-based metabolite identification. *Nucleic Acids Res*. 2022;50(W1):W165–74.
 77. Ruttkies C, et al. MetFrag relaunched: incorporating strategies beyond in silico fragmentation. *Journal of Cheminformatics*. 2016;8(1):3.
 78. Lai Z, et al. Identifying metabolites by integrating metabolome databases with mass spectrometry cheminformatics. *Nat Methods*. 2018;15(1):53–6.
 79. Wang M, et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat Biotechnol*. 2016;34(8):828–37.
 80. Xia J, Wishart DS. Using MetaboAnalyst 3.0 for comprehensive metabolomics data analysis. *Curr Protoc Bioinformatics*. 2016;55(1):14.10. 1-14.10. 91.
 81. Pluskal T, et al. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics*. 2010;11(1):395.
 82. Mahieu NG, Genenbacher JL, Patti GJ. A roadmap for the XCMS family of software solutions in metabolomics. *Curr Opin Chem Biol*. 2016;30:87–93.
 83. Mu BX, et al. Understanding apoptotic induction by *Sargentodoxa cuneata*-*Patrinia villosa* herb pair via PI3K/AKT/mTOR signalling in colorectal cancer cells using network pharmacology and cellular studies. *J Ethnopharmacol*. 2024;319(Pt 3):117342.
 84. Lu S, et al. Mechanism of Bazhen decoction in the treatment of colorectal cancer based on network pharmacology, molecular docking, and experimental validation. *Front Immunol*. 2023;14:1235575.
 85. Ji H, et al. Prediction of the mechanisms by which Quercetin enhances Cisplatin action in cervical cancer: a network pharmacology study and experimental validation. *Front Oncol*. 2021;11:780387.
 86. Ali N, et al. AI based natural inhibitor targeting RPS20 for colorectal cancer treatment using integrated computational approaches. *Sci Rep*. 2025;15(1):24906.

87. Alqahtani NK, et al. Machine learning insights into the antioxidant and biomolecular shielding effects of polyphenol-rich 18 date palm pit extracts. *Food Chemistry: X*. 2025;27:102480.
88. Yoo S, et al. A deep learning-based approach for identifying the medicinal uses of plant-derived natural compounds. *Front Pharmacol*. 2020;11:584875.
89. Lagunin A, et al. PASS: prediction of activity spectra for biologically active substances. *Bioinformatics*. 2000;16(8):747–8.
90. Othman ZK, et al. Artificial intelligence for natural product drug discovery and development: current landscape, applications, and future directions. *Intelligence-Based Medicine*. 2025;12:100316.
91. Ren F, et al. AlphaFold accelerates artificial intelligence powered drug discovery: efficient discovery of a novel CDK20 small molecule inhibitor. *Chem Sci*. 2023;14(6):1443–52.
92. Tsugawa H, et al. MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nat Methods*. 2015;12(6):523–6.
93. Aigensberger M, et al. Modular comparison of untargeted metabolomics processing steps. *Anal Chim Acta*. 2025;1336:343491.
94. Vamathevan J, et al. Applications of machine learning in drug discovery and development. *Nat Rev Drug Discov*. 2019;18(6):463–77.
95. Chen H, et al. The rise of deep learning in drug discovery. *Drug Discov Today*. 2018;23(6):1241–50.
96. Harvey AL, Edrada-Ebel R, Quinn RJ. The re-emergence of natural products for drug discovery in the genomics era. *Nat Rev Drug Discov*. 2015;14(2):111–29.
97. Feher M, Schmidt JM. Property distributions: differences between drugs, natural products, and molecules from combinatorial chemistry. *J Chem Inf Comput Sci*. 2003;43(1):218–27.
98. Rodrigues T, et al. Counting on natural products for drug design. *Nat Chem*. 2016;8(6):531–41.
99. Zhang R, et al. Network pharmacology databases for traditional Chinese medicine: review and assessment. *Front Pharmacol*. 2019;10:123.
100. Mullooney MW, et al. Artificial intelligence for natural product drug discovery. *Nat Rev Drug Discov*. 2023;22(11):895–916.
101. Disha NS. From nature to neural networks: the role of artificial intelligence in selecting plants for cancer drug discovery. *Journal of Information Systems Engineering and Management*. 2025;10(49):35–50.
102. Nothias L-F, et al. Feature-based molecular networking in the GNPS analysis environment. *Nat Methods*. 2020;17(9):905–8.
103. Tu M, et al. Machine learning driven decoding of impurity fingerprint in imidacloprid material. *Microchem J*. 2025;212:113399.
104. Kim HW, et al. NPClassifier: A Deep Neural Network-Based Structural Classification Tool for Natural Products. *J Nat Prod*. 2021;84(11):2795–807.
105. Sorokina M, Steinbeck C. Review on natural products databases: where to find data in 2020. *J Cheminform*. 2020;12(1):20.
106. Tay DWP, et al. 67 million natural product-like compound database generated via molecular language processing. *Sci Data*. 2023;10(1):296.
107. Abdulhakeem Mansour Alhasbary A, Hashimah Hassain Malim N, Zuraidah Mohamad Zobir S. Exploring natural products potential: a similarity-based target prediction tool for natural products. *Comput Biol Med*. 2025;184:109351.
108. Bachorz RA, et al. Multi-criteria decision analysis in drug discovery. *Appl Biosci*. 2025;4(1):2.
109. Xiong G, et al. ADMETlab 2.0: an integrated online platform for accurate and comprehensive predictions of ADMET properties. *Nucleic Acids Res*. 2021;49(W1):W5–w14.
110. Wang L, et al. The present state and challenges of active learning in drug discovery. *Drug Discov Today*. 2024;29(6):103985.
111. Yang GR, et al. Task representations in neural networks trained to perform many cognitive tasks. *Nat Neurosci*. 2019;22(2):297–306.
112. Yang X, et al. Concepts of artificial intelligence for computer-assisted drug discovery. *Chem Rev*. 2019;119(18):10520–94.
113. Schneider P, et al. Rethinking drug design in the artificial intelligence era. *Nat Rev Drug Discov*. 2020;19(5):353–64.
114. Bender A, Cortes-Ciriano I. Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 2: a discussion of chemical and biological data. *Drug Discov Today*. 2021;26(4):1040–52.
115. Gaulton A, et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res*. 2011;40(D1):D1100–7.
116. Gaulton A, et al. The ChEMBL database in 2017. *Nucleic Acids Res*. 2017;45(D1):D945–54.
117. Nelson GS. Bias in artificial intelligence. *N C Med J*. 2019;80(4):220–2.
118. Mumtaz H, et al. Exploring alternative approaches to precision medicine through genomics and artificial intelligence—a systematic review. *Front Med (Lausanne)*. 2023;10:1227168.
119. Kumar R, et al. An Integration of blockchain and AI for secure data sharing and detection of CT images for the hospitals. *Comput Med Imaging Graph*. 2021;87:101812.
120. Rasool, S., et al 2024 *Innovations in AI-powered healthcare: Transforming cancer treatment with innovative methods*. BULLET: Jurnal Multidisiplin Ilmu 3 1
121. Han R, et al. Revolutionizing medicinal chemistry: the application of artificial intelligence (AI) in early drug discovery. *Pharmaceuticals*. 2023;16(9):1259.
122. Xu Y, et al. Deep learning for molecular generation. *Future Med Chem*. 2019;11(6):567–97.
123. Lundberg, S.M. and S.-I. Lee 2017 *A unified approach to interpreting model predictions*. Advances in neural information processing systems 30.
124. Zhang S, et al. Multi-target meridians classification based on the topological structure of anti-cancer phytochemicals using deep learning. *J Ethnopharmacol*. 2024;319:117244.
125. Orobator E, et al. Applications of artificial intelligence in plant-based anticancer drug discovery and development. *J Pharma Insights Res*. 2025;3(2):203–10.
126. Lin K, et al. Cyclic peptide therapeutic agents discovery: computational and artificial intelligence-driven strategies. *J Med Chem*. 2025. <https://doi.org/10.1021/acs.jmedchem.5c00712>.
127. Singla N, et al. A Pilot Study of Breast Cancer Histopathological Image Classification Using Google Teachable Machine: A No-Code Artificial Intelligence Approach. *Cureus*. 2025;17:7.
128. Ghosh I, Ng HKT. Hidden truncation models: theory and applications. *WIREs Comput Stat*. 2025;17(1):e70019.
129. Nguyen, P.-T. and H.-T. Nguyen, *Emerging trends in pharmacological research of herbal-based traditional medicine*. Advances in Traditional Medicine, 2025.
130. Baek B, Lee H. Crossfeat: a transformer-based cross-feature learning model for predicting drug side effect frequency. *BMC Bioinformatics*. 2024;25(1):324.
131. Brown, S., *Benchmarking multi-omics latent factor methods to predict anticancer drug response using baseline cancer cell line data*. 2022, University of Nottingham (United Kingdom).
132. Ali, S., et al 2025 *Comprehensive Insights into Natural Bioactive Compounds: From Chemical Diversity and Mechanisms to Biotechnological Innovations and Applications*. ChemistryOpen e202500469.
133. Ramesh, R., et al. *Applying Vision Transformers for Herbal Medicine Classification: A Novel Approach to Plant Identification*. in 2025 *International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI)*. 2025. IEEE.
134. Mahanta, H.J., et al 2025 *Exploring graph-based models for predicting active compounds against triple-negative breast cancer*. Molecular Diversity 1–19.
135. Honda, S., S. Shi, and H.R. Ueda, *Smiles transformer: Pre-trained molecular fingerprint for low data drug discovery*. arXiv preprint [arXiv:1911.04738](https://arxiv.org/abs/1911.04738), 2019.
136. Mswahili ME, Jeong Y-S. Transformer-based models for chemical SMILES representation: a comprehensive literature review. *Heliyon*. 2024. <https://doi.org/10.1016/j.heliyon.2024.e39038>.
137. Kim H, et al. A genotype-to-drug diffusion model for generation of tailored anti-cancer small molecules. *Nat Commun*. 2025;16(1):5628.
138. Jabeen, F., *Application of Machine Learning and Deep Learning Approaches Cheminformatic for Drug Discovery*. 2024, Carleton University.
139. Durojaye OA, et al. Harnessing AI-driven reverse docking in drug discovery: a comprehensive review of opportunities, challenges, and emerging trends. *J Mol Model*. 2025;31(9):256.
140. Zhang X, et al. Advancing ligand docking through deep learning: challenges and prospects in virtual screening. *Acc Chem Res*. 2024;57(10):1500–9.

141. Venkatasubbu S, Krishnamoorthy G. Ethical considerations in AI addressing bias and fairness in machine learning models. *J of Knowl Learn and Sci Tech*. 2022;1(1):130–8.
142. Bifarin OO, Fernández FM. Automated machine learning and explainable AI (AutoML-XAI) for metabolomics: improving cancer diagnostics. *J Am Soc Mass Spectrom*. 2024;35(6):1089–100.
143. Lal S, Singh B, Kaunert C. Role of Artificial intelligence (AI) and Intellectual property rights (IPR) in transforming drug discovery and development in the life sciences: legal and ethical concerns library of progress-library science. *Information Technology & Computer*. 2024;44:3.
144. Shirzad M, et al. Artificial intelligence-assisted design of nanomedicines for breast cancer diagnosis and therapy: advances, challenges, and future directions. *BioNanoScience*. 2025;15(3):354.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.