

OPEN ACCESS

This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Parijaat College of Pharmacy,
Indore, Madhya Pradesh, India
ROR

Correspondence to: Harshit Shringi, shringiharshit45@gmail.com

Cite this as: Shringi H. Artificial Intelligence and Computational Modeling in Natural Product Drug Discovery. Premier Journal of Science 2026;22:100279

DOI: <https://doi.org/10.70389/PJS.100279>

Peer Review:

Received: 08 February 2026

Last revised: 03 March 2026

Accepted: 10 May 2026

Version accepted: 2

Published: 17 May 2026

Ethical approval: Yes

Consent: Yes

Funding: Yes

Conflicts of interest: N/a

Author contribution: Harshit Shringi: Conceptualization; Investigation; Supervision; Validation; Visualization; Writing – original draft; Writing – review; editing

Guarantor: Harshit Shringi

Provenance and peer-review: Unsolicited and externally peer-reviewed

Data availability statement: N/a

Artificial Intelligence and Computational Modeling in Natural Product Drug Discovery

Harshit Shringi

ABSTRACT

The continued investigation of natural products remains an integral part of drug discovery today and offers a tremendous resource of diverse and potentially bioactive molecules. The use of computational methodologies has transformed and improved the ability to discover, optimize, and validate leads with therapeutic potential from natural sources. Recent advances in *in silico* methods allow for the rapid screening of natural product databases, better characterization of pharmacokinetic properties, and assessment of molecular interactions with high accuracy. Quantitative structure–activity relationship (QSAR) modeling plays a central role in establishing the association of molecular descriptors with biological activity, which increases the predictability and efficiency of lead optimization. Structure- and ligand-based virtual screening methodologies also expedite the identification of promising natural scaffolds for a variety of disease processes. This chapter presents an overview of computationally guided methodologies in phytochemical research and how they contribute to a better drug discovery process, reduced costs, and better success in bioactive molecule development. By bridging traditional ethnopharmacology with data-driven modeling and predictive methods, a new paradigm towards sustainable and efficient drug discovery is presented.

Keywords: Bioactive lead discovery, Computational drug design, Ethnopharmacology, Lead optimization, Molecular docking, Natural products, Phytochemicals, QSAR modeling, Structure–activity relationship, Virtual screening

Introduction

Natural products (NPs) small molecules and their derivatives produced by plants, fungi, bacteria, and marine organisms constitute a historically and currently important reservoir of biologically active chemical matter for drug discovery.¹ NP-derived scaffolds have given rise to numerous approved drugs across therapeutic areas including anti-infectives, oncology, immunosuppression, metabolic disease, and central nervous system disorders.² The defining features of many natural products are high stereochemical complexity, rich three-dimensionality, and dense functionalization, properties that differentiate the NP chemical space from typical synthetic small-molecule libraries. Because natural selection has “pre-validated” many NP scaffolds against biological macromolecules, NPs often exhibit high affinity and specificity for protein targets that are difficult to drug with flat, planar synthetic compounds.³

The Significance of Natural Products in Contemporary Drug Discovery

Natural products have been a rich source of lead compounds for drug discovery because of their extensive diversity and biological significance. Many successful drugs over the years, including antibiotics, anticancer medications, immunosuppressants, and enzymes, have origins derived from nature or naturally occurring scaffolds, having a high degree of complexity in terms of stereochemistry and multiple-ring structures that are difficult if not impossible to create in the lab.⁴ The development of new technologies such as genomics and metabolomics has provided an immense opportunity to explore the biosynthetic potential of the vast array of microorganisms, plants, and other living organisms that produce compounds naturally; these opportunities are not accessible via traditional approaches. Artificial intelligence (AI) and machine learning (ML) are also allowing scientists to find out where they might find biosynthetic gene clusters, predict the chemical structure of metabolites, and determine if a metabolite has potential biological activity.⁵ This is providing many avenues for finding new drug candidates. Another aspect worth mentioning is that the way natural products exist is quite different from the way synthetic libraries exist; natural products are more complex and typically exist in three-dimensional arrangements due to the evolution of their function over time. Therefore, leads derived from natural products are desirable candidates for new drug discovery.⁶

Limitations of Conventional Screening Strategies

Traditionally, screening strategies that rely on bioassay-guided fractionation, chromatography, and structure elucidation as natural products are slow and resource-intensive. Normally, these strategies would require a large collection of samples, require an extended period of time (~3 months) of bench or wet laboratory work for each extract, and would typically take multiple passes through the process to purify each compound and dereplicate (to avoid redundancy with previously identified compounds) the known compounds.⁷ Additionally, the methods used to generate the ability to perform a high-throughput screen (HTS) on synthetic compound libraries have limitations when applied to the complex, heterogeneous crude extracts (often called “crude extracts”) from natural product sources. Crude extracts are made up of a multitude of molecules (usually over 100) that can interfere with the assays used to identify active compounds. Thus, while HTS methods are able to perform higher throughput, the process to screen complex samples often takes longer, resulting in fewer hits (active

compounds) and greater costs than if the samples had been screened using traditional laboratory screening methods. Finally, even when active compounds are identified, new structures will need to be isolated and characterized for structure using highly specialized instruments and resources, which leads to a significant bottleneck to obtaining meaningful translation of active compound structures to optimally active chemical compound structures.⁸ Table 1 compares the major differences between experimental conventional screening and the computational approaches to drug discovery in terms of efficiency, cost, and yield.

The Development of Computational Approaches Within the Field of Phytochemistry

To address these challenges, researchers have increasingly adopted computational approaches that integrate cheminformatics, bioinformatics, machine learning, and deep learning to accelerate natural product discovery. These AI-driven methodologies enable *in silico* prioritization of compounds, prediction of biological activity, target deconvolution, and efficient navigation of the vast chemical space defined by natural metabolites.¹⁰ For example, machine learning models trained on known actives and inactive can forecast the activity of untested natural compounds, while deep learning architectures can capture complex molecular features that traditional descriptors miss. AI has also been integrated into virtual screening pipelines to enrich compound libraries for likely binders before costly experimental validation, effectively converting the discovery workflow into a hybrid computational–experimental strategy. Moreover, beyond activity prediction, AI facilitates advanced tasks such as prediction of biosynthetic gene clusters (which reveal the genetic basis for NP production), metabolomic pattern recognition, and multi-omics data integration, collectively enhancing the efficiency and breadth of phytochemical research.¹¹

Purpose of the Review

This review will provide a detailed, current overview of how AI, machine learning, and advanced computing methods are being applied to drug discovery for natural products.

The review will cover the following:

1. The continuing importance of natural products as sources of new drugs in the present-day era of drug discovery.
2. How traditional methods of screening for natural product activity have limitations, thus creating a need for innovation through the application of computers.
3. How state-of-the-art applications of AI and ML techniques (e.g., activity prediction, virtual screening, generative modeling, and multi-omic integration) are being used in NP research.
4. The types of challenges present today, opportunities, and case studies showing how researchers have successfully used and been successful in using AI to find bioactive natural products.

This review will create a foundation for researchers to apply AI-based drug discovery processes within the area of natural product research by providing them with a combination of both theory and hands-on practical experience in terms of recent computationally and experimentally advanced research methodologies and develop a framework for researchers to use to create AI-enabled NP workflows.¹²

Methodology

Literature Search Strategy

The present review was conducted through a systematic and structured literature survey focusing on artificial intelligence (AI), computational modeling, and natural product-based drug discovery. Scientific literature was collected from internationally recognized databases including PubMed, Scopus, Web of Science, ScienceDirect, and Google Scholar. The search strategy incorporated combinations of keywords such as *natural products*, *phytochemicals*, *AI in drug discovery*, *machine learning*, *virtual screening*, *molecular docking*, *QSAR modelling*, *cheminformatics*, and *ADMET prediction*.

Studies published primarily between 2015 and 2025 were prioritized to ensure inclusion of recent technological developments and contemporary computational approaches. Classical foundational studies were also included where necessary to explain theoretical concepts and methodological evolution.

Table 1 | Comparison between conventional and computational approaches in natural product drug discovery⁹

Parameter	Conventional Drug Discovery Approaches	Computational (AI-Driven) Drug Discovery Approaches
Screening method	Wet-lab–based experimental screening	<i>In silico</i> virtual screening and predictive modeling
Time consumption	Time-consuming (months to years)	Rapid and time-efficient (days to weeks)
Coast	High cost due to extensive laboratory experiments	Cost-effective by reducing experimental burden
Compound handling	Limited number of compounds can be tested	Large chemical libraries can be screened simultaneously
Hit identification	Low hit rate and trial and error-based	Higher hit rate using rational and data-driven selection
Mechanistic insight	Limited understanding at early stages	Provides molecular-level interaction and mechanism insights
ADMET assessment	Performed at later experimental stages	Early prediction of ADMET and drug-likeness properties
Lead optimization	Sequential and resource-intensive	Efficient optimization using QSAR and molecular modeling
Reproducibility	Experimental variability may affect results	High reproducibility through standardized algorithms
Overall efficiency	Low efficiency with high failure rate	High efficiency with improved success probability

Inclusion and Exclusion Criteria

Articles were selected based on their scientific relevance to computational drug discovery involving natural products. Inclusion criteria consisted of the following:

- Peer-reviewed research articles and review papers related to AI-driven or computational drug discovery.
- Studies describing virtual screening, QSAR modeling, molecular docking, pharmacophore modeling, or AI-based predictive modeling.
- Publications discussing integration of computational and experimental validation approaches.

Exclusion criteria included the following:

- Non-peer-reviewed articles, conference abstracts without full methodology, and duplicate reports.
- Studies lacking computational relevance or insufficient methodological description.
- Articles focusing solely on synthetic compounds without relevance to natural product drug discovery.

Data Extraction and Analysis

Relevant information from selected studies was extracted and categorized according to the computational methodology, application area, and stage of drug discovery. Extracted parameters included computational tools used, modeling approaches, validation techniques, biological targets, and reported outcomes.

The collected data were organized into thematic categories including computer-aided drug design (CADD), virtual screening approaches, QSAR modeling, cheminformatics analysis, and AI-based predictive modeling. A comparative evaluation of methodologies was performed to identify the advantages, limitations, and emerging trends in computational natural product research.

Computational Workflow Considered in the Review

The methodological framework of this review follows a generalized AI-assisted natural product drug discovery workflow. The workflow includes the following:

1. Identification of medicinal plants and ethnopharmacological knowledge sources.
2. Compilation and curation of phytochemical databases.
3. Molecular descriptor calculation and chemical data preprocessing.
4. Virtual screening and molecular docking for target interaction analysis.
5. QSAR modeling and machine learning-based activity prediction.
6. ADMET and drug-likeness evaluation.
7. Lead prioritization, followed by experimental validation strategies.

This workflow reflects the integrated computational pipeline commonly adopted in modern natural product drug discovery studies.

Limitations of Methodological Approaches

As this study represents a narrative review, conclusions are dependent on the availability and quality

of published data. Variability in computational protocols, data set heterogeneity, and differences in validation strategies across studies may influence interpretation. Nevertheless, efforts were made to include diverse and high-quality sources to ensure balanced representation of current methodologies. Although elements of structured literature screening were applied to enhance transparency, this study represents a narrative and critical review rather than a fully systematic review. PRISMA guidelines were consulted to improve reporting clarity; however, no formal systematic review protocol registration (e.g., PROSPERO) or meta-analysis was conducted, as illustrated in Figure 1.

Nature of the Review

This article is designed as a narrative and critical review aimed at synthesizing contemporary developments in AI-assisted natural product drug discovery. While structured search strategies and transparent inclusion criteria were applied, the objective was conceptual synthesis and methodological evaluation rather than exhaustive systematic aggregation or quantitative meta-analysis. Therefore, the review emphasizes thematic integration, critical appraisal, and translational interpretation rather than comprehensive systematic evidence grading.

Computational Approaches in Drug Discovery of Natural Products

Computational Methods are now a vital part of modern drug discovery because they allow scientists to identify, prioritize, and optimize new chemical entities before testing them. For natural products, this technology addresses some of the challenges associated with working with complex chemicals, the scarcity of samples, and the low number of experiments that can be performed to evaluate these compounds. When these technologies were combined in the field of natural products, known as “naturals,” they enabled researchers to develop a more structured and systematic (data-driven) approach to the design of new drugs, which has dramatically decreased the amount of time, money, and experimental failure rates for drug development of natural products.¹³

Introduction to Computer-Aided Drug Design (CADD)

Computer-aided drug design (CADD) is a collection of computer-assisted methods that help to assess how a molecule interacts with the biological system, insights that help to predict how much activity will be derived from a given lead compound, and therefore, aid the development of novel therapeutic agents. These computer-assisted methodologies are most useful for drug discovery from natural products because of their structural diversity, stereochemical complexity, and multi-target potential.¹⁴ Structure-based drug design (SBDD) and ligand-based drug design (LBDD) are two primary approaches used to create CADD methodologies. SBDD uses the availability of 3D structures of biological targets to directly investigate how ligands interact with their respective targets. Conversely, LBDD

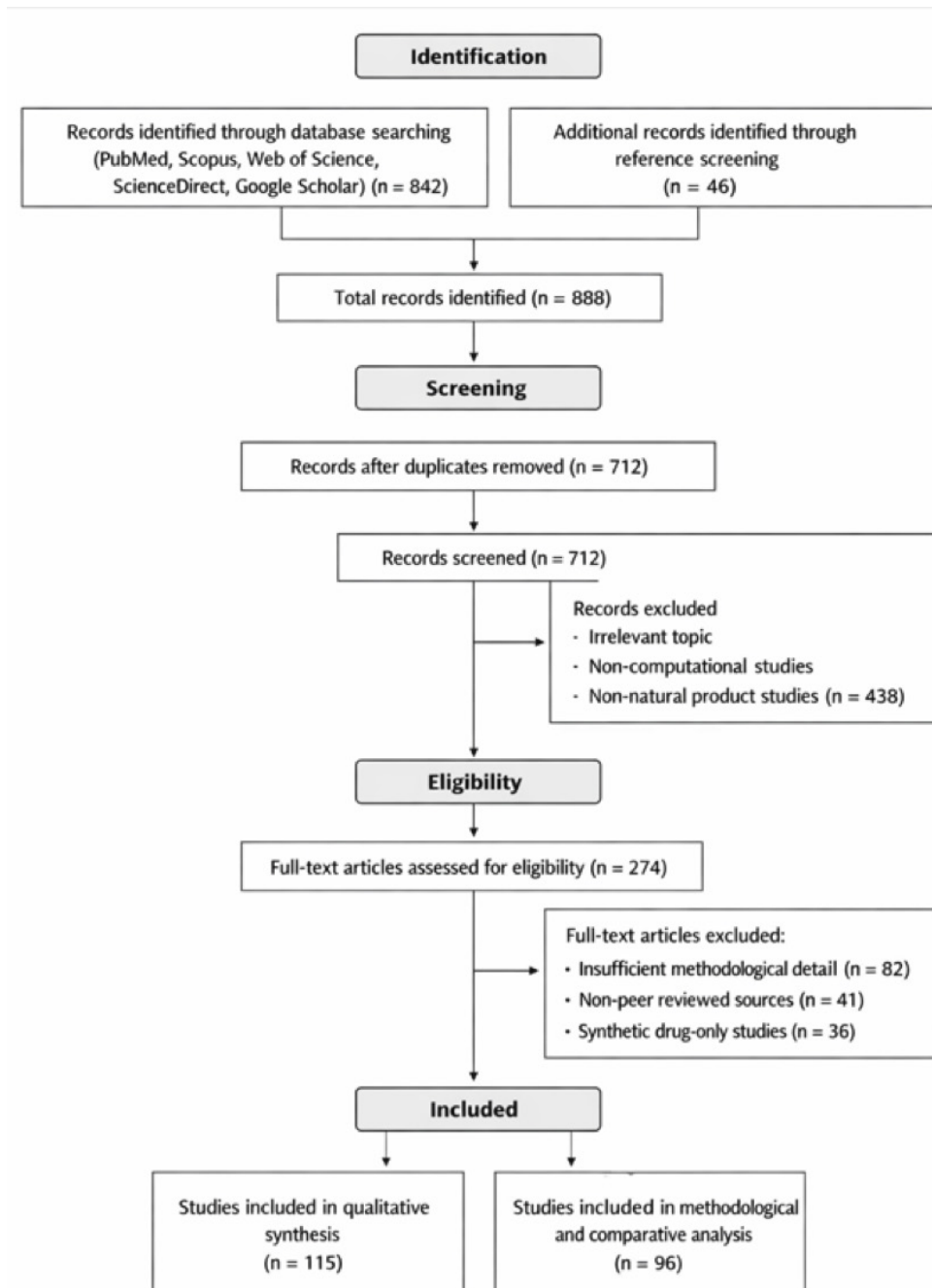


Fig 1 | PRISMA flow diagram illustrating the literature search and study selection process used in the present review. Records were identified through multiple scientific databases and additional sources, followed by duplicate removal, title and abstract screening, eligibility assessment through full-text evaluation, and final inclusion of studies for qualitative synthesis and methodological analysis according to PRISMA guidelines

is typically used when the target structure(s) of the compound(s) are not known, and instead, the goal of LBDD is to use structural knowledge about previously identified active compounds to evaluate whether active compounds should exhibit activity.¹⁵ In particular, many of the natural compounds can be found to violate standard drug-likeness rules; however, their activity remains robust. AI provides the opportunity to more systematically assess compounds that fall into this

categorization; by using these computer-enhanced CADD methodologies, we can identify the important aspects of molecular interactions and guide the design of lead compounds while maintaining biological activity.¹⁶ Additionally, through the use of recent advancements in machine learning, CADD methodologies will continue to evolve as more extensive databases containing information on natural products are created and made available to researchers interested in natural

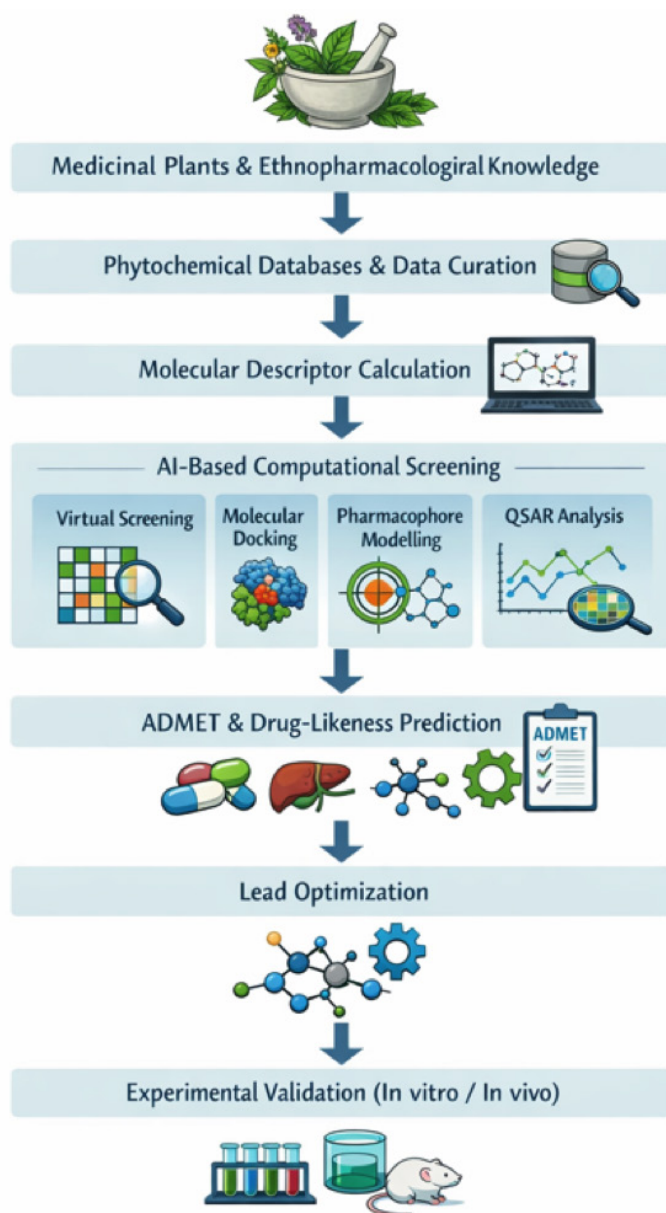


Fig 2 | Overall workflow of AI-driven natural product-based drug discovery. Schematic representation of the AI-driven workflow for natural product-based drug discovery. The process initiates with medicinal plants and ethnopharmacological knowledge, followed by phytochemical database generation and data curation. Molecular descriptor calculation enables AI-based computational screening approaches such as virtual screening, molecular docking, pharmacophore modeling, and QSAR analysis. The shortlisted compounds are further evaluated through ADMET and drug-likeness prediction, leading to lead optimization and subsequent experimental validation using in vitro and in vivo models

product-based drug discovery.¹⁷ An example of the AI-enabled CADD process from phytochemical identification to experimental validation is shown in Figure 2.

Virtual Screening and Molecular Docking Approaches

The virtual screening process is one of the main tools used to determine the most effective compounds from an extensive database for many different types

of biological targets.^{18,19} This technique is extremely useful for researchers attempting to identify a small number of phytochemicals from extensive molecular databases, before making tedious efforts to separate and evaluate their biological effectiveness through time-consuming methods. There are two categories for these strategies: ligand- and structure-based virtual screening.^{20,21} Ligand-based virtual screening uses the concept that “similar compounds will produce similar biological effects.” Structure-based virtual screening simulates the molecular interactions of ligands as they interact with the 3D structure of target proteins. The method of molecular docking is a structure-based method used to estimate the most favorable orientation of a ligand in the binding site of a target protein and to assess the degree of binding affinity between the ligand and its target protein through the application of scoring functions.²² Molecular docking studies are also commonly performed to help identify novel natural inhibitors/modulators for enzymes, receptors, and signaling proteins associated with cancer, neurodegenerative disorders, inflammation, and infectious diseases. Even though molecular docking is extensively used, it does have several limitations, including the inherent protein flexibility and the effects of solvents, as well as the fact that the scoring functions tend to have a greater degree of error for large flexible molecules that are of natural origin. Computational methods are now a vital part of modern drug discovery because they allow scientists to identify, prioritize, and optimize new chemical entities before testing them.^{23,24} For natural products, this technology addresses some of the challenges associated with working with complex chemicals, the scarcity of samples, and the low number of experiments that can be performed to evaluate these compounds. When these technologies were combined in the field of natural products, known as “naturals,” they enabled researchers to develop a more structured and systematic (data-driven) approach to the design of new drugs, which has dramatically decreased the amount of time, money, and experimental failure rates for drug development of natural products (Table 2).²⁵

Pharmacophore Modeling and Related Similarity Searches

Using a computational method known as “pharmacophore modeling,” it is possible to identify chemical characteristics that are essential to the pursuit of biologically active ligands and to produce a spatial arrangement of the pharmacophore. In the case of natural product drug discovery, the pharmacophore approach is especially useful when one may not have detailed target structural information, but several active compounds may be present.³⁰ The pharmacophore model captures the common pharmacophoric characteristics shared by many structurally dissimilar natural products, allowing researchers to identify new compounds that are likely to have similar biological properties as the active compounds they

Table 2 | Common computational tools and techniques used in AI-assisted natural product research^{26–29}

Computational Technique	Commonly Used Tools/Software	Primary Application in Natural Product Research
Virtual screening	AutoDock Vina, Schrödinger Glide, GOLD, DOCK	Rapid screening of large phytochemical libraries to identify potential hits
Molecular docking	AutoDock, AutoDock Vina, MOE, Schrödinger	Prediction of binding modes and interaction energies with biological targets
Pharmacophore modeling	LigandScout, Discovery Studio, Phase	Identification of key structural features required for biological activity
QSAR modeling	PaDEL-Descriptor, Dragon, KNIME, WEKA	Correlation of molecular descriptors with biological activity for prediction
Molecular dynamics simulation	GROMACS, AMBER, NAMD	Evaluation of stability and conformational changes of ligand–target complexes
ADMET prediction	SwissADME, pkCSM, ADMETlab	Early assessment of pharmacokinetic and toxicity profiles
Cheminformatics analysis	RDKit, Open Babel, ChemAxon	Chemical data handling, curation, and molecular similarity analysis
AI/machine learning methods	Random Forest, Support Vector Machine, Neural Networks	Predictive modeling and lead optimization from complex data sets

have already identified, using comparison to the shared pharmacophoric characteristics of the active compounds.³¹ Similarity search technologies can be used in conjunction with pharmacophore modeling to identify compounds that have similar 2D or 3D signatures to a given compound.³² These technologies are very useful for finding and analyzing the chemical diversity within natural products. Additionally, when pharmacophore modeling and advanced scoring and machine learning techniques are combined, novel natural scaffolds can be identified that are very reliable and have greater predictive power. Pharmacophore modeling is a ligand-based computational approach that identifies the essential chemical features required for biological activity and defines their spatial arrangement.³³ These features typically include hydrogen-bond donors and acceptors, hydrophobic regions, aromatic rings, and charged moieties. In natural product drug discovery, pharmacophore modeling is particularly useful when multiple active compounds are known, but detailed target structural information is unavailable. By capturing shared pharmacophoric features among structurally diverse natural compounds, this approach enables the identification of novel candidates with similar biological potential. Similarity search techniques complement pharmacophore modeling by identifying compounds with comparable two-dimensional fingerprints or three-dimensional shapes. Such methods are highly effective in exploring the chemically diverse and biologically relevant space occupied by natural products. The integration of pharmacophore modeling with advanced scoring and machine learning tools further enhances the identification of novel natural scaffolds with improved predictive reliability.³⁴

Cheminformatics Fusion and Use of Molecular Databases

Cheminformatics is the computational backbone for managing, analyzing, and interpreting chemical information for natural products. Cheminformatics enables the calculation of molecular descriptors, physicochemical properties, fingerprints, and similarity metrics that are critical for screening, modeling, and mining data virtually. Currently, there are multiple databases dedicated to storage of the chemical structure, biological activity, and physicochemical properties of natural

products that can be incorporated easily into computational processes.^{35,36} There are several benefits to using large-scale databases as part of cheminformatics, such as enabling high-throughput screening, predictive modeling, and hypothesis generation during the drug discovery process for natural products. Additional concepts related to cheminformatics include the “fusion” of multiple chemical structure types, bioactivity, target annotation, and pharmacokinetic information into a single analysis. Fusion of multiple sources of data provides improved prediction ability and allows for multimodal optimization of chemical leads derived from natural sources.³⁷ The primary challenges encountered during the cheminformatics process include data diversity (the diversity of the source of data), lack of consistency in annotation of data sources, and experimental bias. To ensure that AI-based cheminformatics can effectively support the process of development of drugs from natural products, there is a need for standardized curation and transparency in reporting data.

Quantitative Structure–Activity Relationship (QSAR) Modeling

The quantitative structure–activity relationship (QSAR) model has become one of the best forms of computational drug discovery by providing a mathematical model to correlate the chemical structure of a compound to its activity in cells (biological activity) or within the human body. Natural products (NPRs) drug discovery utilizes QSAR as a basis to rationally explain activities at different levels in plant-based medicines (phytomedicines).³⁸ By correlating numerical descriptors with the characteristics of each phytochemical, prediction, prioritization, and optimization of the best leads through a reduction in the experimental workload become achievable. Natural products have their own individual challenges/benefits for QSAR modeling with respect to their varied chemical compositions, thicknesses, and layers of activity at a multitude of sites. Whereas synthetic (manmade) libraries are more uniform in nature, phytochemicals often consist of complex ring structures, multiple chiral centers, and multiple functional groups, all of which have varying degrees of influence on the actual interaction with the “target” of research (biological target) through different pathways/mechanisms.³⁹ QSAR modeling provides a systematic approach to identify these interactions

and develop an understanding of what constitutes structure–activity relationships, and therefore, significantly aids in the development of new drugs.

Introduction and Theoretical Background

The theoretical foundation of QSAR is based on the assumption that the biological activity of a compound is a function of its molecular structure and physicochemical properties. Compounds that share similar structural features are expected to exhibit comparable biological responses when interacting with the same target or biological system. QSAR formalizes this relationship by constructing mathematical models that quantitatively link molecular descriptors to measured biological activity. Early QSAR models were primarily linear, focusing on simple relationships between lipophilicity, electronic effects, and steric factors.⁴⁰ Over time, the QSAR methodology has evolved to accommodate non-linear and complex biological systems through the application of multivariate statistics and machine learning. This evolution is particularly relevant to phytochemicals, whose biological activity often arises from synergistic interactions between multiple structural features rather than a single dominant property. Modern QSAR frameworks integrate concepts from physical chemistry, statistics, and artificial intelligence, allowing models to capture subtle and non-intuitive relationships within highly diverse natural product data sets.⁴¹ These models not only predict activity but also provide mechanistic insights into the structural determinants of bioactivity, thereby supporting rational decision-making in lead discovery and optimization.

Molecular Descriptor Calculations and Data Preprocessing

Molecular descriptors are numerical representations of chemical information that encode various aspects of molecular structure and properties. In QSAR modeling, descriptors serve as independent variables that capture the size, shape, topology, electronic distribution, lipophilicity, hydrogen-bonding capacity, and conformational flexibility of molecules. For phytochemicals, descriptors are particularly important because they translate complex molecular architecture into interpretable and computable features.⁴² Descriptor calculation is followed by rigorous data preprocessing to ensure model reliability and reproducibility. This step includes chemical structure standardization, removal of salts and duplicates, correction of inconsistent representations, and harmonization of biological activity data. Since natural product data sets often originate from diverse experimental sources, careful preprocessing is essential to minimize noise and bias. Feature selection and dimensionality reduction play a critical role in QSAR modeling by eliminating redundant or irrelevant descriptors. Reducing descriptor space improves model interpretability, prevents overfitting, and enhances predictive performance. Data scaling and normalization are also applied to ensure that descriptors with different numerical ranges contribute appropriately during model training.⁴³

Model Building and Validation, and Assessment of Performance

QSAR model development involves selecting suitable algorithms to establish relationships between molecular descriptors and biological activity. Depending on the data set size and complexity, models may range from simple linear regression approaches to advanced machine learning techniques capable of capturing non-linear patterns. Algorithm selection is guided by the nature of the biological endpoint, the data set size, and the desired balance between interpretability and predictive power.⁴⁴ Model validation is a critical component of QSAR analysis and ensures that the developed model is both robust and generalizable. Internal validation techniques assess the stability of the model by testing its performance on subsets of training data, while external validation evaluates predictive accuracy using independent data sets. These validation strategies are essential for determining whether a model captures genuine structure–activity relationships or merely reflects statistical artifacts. Performance assessment relies on quantitative metrics that measure goodness of fit, predictive accuracy, and classification reliability. Interpreting these metrics collectively provides insight into model strengths and limitations.⁴⁵ A well-validated QSAR model not only predicts activity accurately but also demonstrates consistency, transparency, and reproducibility, which are critical for acceptance in drug discovery workflows. The application of the general workflow of QSAR modeling to phytochemicals, including descriptor calculation, model development, validation, and activity prediction, is depicted in Figure 3.

Examples of Successful Applications of QSAR-Driven Discovery

QSAR-driven discovery has been successfully applied across a wide range of therapeutic areas involving natural products. In antimicrobial research, QSAR models have enabled the identification of phytochemicals with enhanced activity against resistant pathogens. In oncology and neuropharmacology, QSAR has guided the selection and optimization of natural compounds targeting enzymes, receptors, and signaling pathways associated with disease progression. QSAR approaches have also been widely used to predict antioxidant, anti-inflammatory, enzyme-inhibitory, and neuroprotective activities of plant-derived compounds. By screening large phytochemical libraries computationally, QSAR models significantly reduce the experimental workload while improving hit rates. These successes highlight the versatility and effectiveness of QSAR as a core component of AI-driven natural product drug discovery.⁴⁸

QSAR Applicability in Lead Optimization of Phytochemicals

QSAR modeling plays a pivotal role in lead optimization by guiding rational modifications of bioactive phytochemicals. Once an initial lead is identified, QSAR models help predict how changes in the molecular

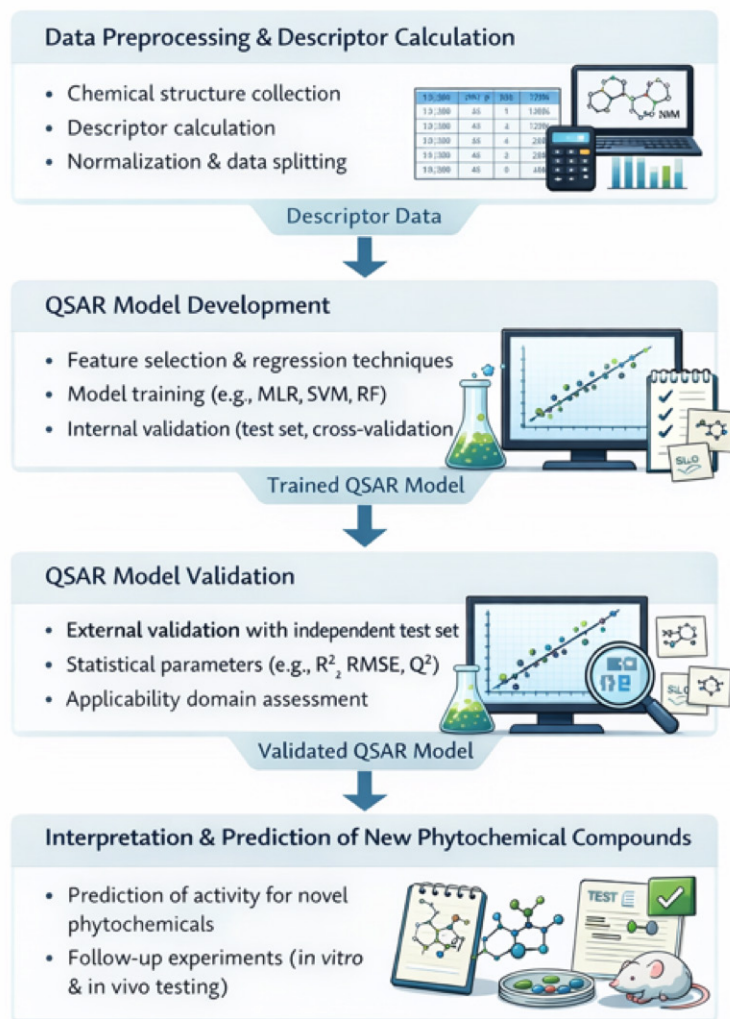


Fig 3 | QSAR modeling process for phytochemicals, including descriptor calculation, model development, and validation. Illustration of the quantitative structure–activity relationship (QSAR) modeling workflow applied to phytochemicals. The process involves data preprocessing and molecular descriptor calculation, followed by feature selection and model development using statistical and machine learning techniques. Model performance is evaluated through internal and external validation using statistical parameters such as R^2 , RMSE, and Q^2 , along with applicability domain assessment. The validated QSAR model is subsequently used for prediction of the biological activity of new phytochemical compounds, supporting lead identification and experimental validation

structure influence biological activity, selectivity, and physicochemical properties. This predictive capability allows researchers to prioritize structural modifications that enhance potency while minimizing undesirable properties. In natural product research, lead optimization often involves balancing bioactivity with drug-like characteristics such as solubility, permeability, and metabolic stability.⁴⁶ QSAR models facilitate this balance by identifying descriptors associated with both efficacy and pharmacokinetic behavior. This enables multi-objective optimization strategies that simultaneously improve therapeutic potential and developability. QSAR also supports the design of semi-synthetic analogs of natural products, preserving the core bioactive scaffold while fine-tuning functional groups. Such rational optimization reduces experimental trial- and

error, and accelerates the progression of phytochemicals from initial hits to viable drug candidates.⁴⁷

Virtual Screening of Compound Libraries Generated From Plants

The emergence of computer-based virtual screening of plant-derived compound libraries has helped revolutionize natural product-based drug discovery as scientists increasingly utilize time-sensitive herbaceous plant-based materials. While herbs contain a vast diversity of secondary metabolites that exhibit a wide range of pharmacological properties, experimentation with the enormous chemical diversity found in herbs through physical experiments would take far too much material, time, and funds to be practical.⁴⁹ Virtual screening enables scientists to virtually screen the enormous number of plant-based phytochemicals using *in silico* methods to evaluate phytochemicals on a mass scale, which allows them to identify potential candidates for biological activity more quickly than they could through traditional experimental methods and then screen for the best possible hits. By combining computational techniques, such as cheminformatics, and molecular and artificial intelligence (AI) models, scientists use virtual screening as a structured, hypothesis-driven approach to conduct plant-based drug discovery. This method of virtual screening improves hit rates for plant-derived candidates by finding and focusing on the most likely candidates, making it much more conceivable to use herbaceous plant materials for drug development and speeding up the drug development process.

Databases for Phytochemicals and Ethnobotanical Compound Sources

Virtual screening provides researchers studying plant-based pharmaceuticals a means to conduct screening of phytochemicals and ethnobotanical compounds. This database houses chemical structures related to plant materials, along with information on the source plants; traditional medicinal uses of these compounds; and in some instances, biological activity. Ethnobotanical resources are useful for identifying appropriate leads to determine the molecular modes of action of phytochemical leads; this resource provides the best biological relevance (starting point) for discovery.⁵⁰ Typically, phytochemical data include a variety of phytochemical compounds and classes, including alkaloids, flavonoids, terpenoids, phenolics, glycosides, saponins, etc. Phytochemical resources provide the necessary molecular format in a consistent molecular structure for performing virtual screening of phytochemicals derived from plants. Increasingly extensive, precisely maintained resources related to virtual screens of phytochemicals support the continued enhancement and building of enhanced capability and reliability of virtual screening research for the development of a new class of antineoplastic drugs.⁵¹

Structure-Based Virtual Screening Approaches

SBVS utilizes three-dimensional structural knowledge to evaluate the ability of phytochemicals to bind to a

target biological system. The primary concept used in SBVS is molecular docking, which uses scoring algorithms to predict the interaction between the ligand and the target, and to calculate the interaction. Therefore, the SBVS technique is a good way to identify natural inhibitors or modulators of specific enzymes, receptors, and signaling proteins associated with disease development.⁵² Simulation of phytochemical binding through molecular docking is a means for SBVS to assess conformation flexibility and functional groups present on phytochemicals binding to their target, that is, how they interact with the key amino acids of the binding site. Advanced SBVS processes usually incorporate preparation of the protein, identification of the binding site, flexible molecular docking, and refinement of virtual binding as a way to improve the accuracy of predictions concerning ligand–target interactions. In addition to the above, consensus scoring of molecular docking results and re-scoring techniques address the weaknesses of individual scoring functions. SBVS can also identify potential binding conformations for phytochemicals with large and flexible structures that would be otherwise time-consuming, difficult to obtain, or impossible through laboratory methods.⁵³

Ligand-Based Screening and Predictive Filtering

When there is no or only limited access to structural data associated with a biological target, ligand-based virtual screening (LBVS) becomes the methodology of choice. The method depends on the principle that compounds with similar chemical structures will typically exhibit similar biological activities. LBVS utilizes methods for establishing molecular similarity, pharmacophore models, and machine learning to find potential phytochemical leads. Additionally, predictive filtering allows the determination of which leads will have higher drug-like quality and developability based upon physicochemical properties, as well as liver metabolic stability; those compounds that do not meet these criteria will be eliminated from the library of phytochemicals being screened.⁵⁴ For example, LBVS uses machine learning algorithms that have been trained to predict the likelihood of phytochemicals with specific characteristics to have the same or similar mechanism of action as drugs already on the market. These types of models are able to identify sophisticated (e.g., complex, non-linear) relationships between different physicochemical properties and biological activity, thus enabling LBVS to provide faster, more targeted pharmacological predictions of phytochemicals than would be available through traditional means. Overall, the development and use of predictive models support greater enrichment of active compounds while at the same time reducing the number of false positives in large-scale phytochemical screens.⁵⁵

High-Throughput Screening Pipeline, Interpretation, and Elaboration of Results

To identify and prioritize the best lead candidates, a high-throughput virtual screening pipeline uses an integrated workflow consisting of multiple computing

steps. A screening proposal typically includes a library preparation step, followed by a library filtering step, and final inspection and ranking. Docking scores, binding poses, interaction profiles, and predicted activity values (VAAs) should be assessed in detail to determine which results are genuine characterizations of physical interactions between the compound and its intended biological target(s). A full evaluation of the results in terms of visualization of ligand–target interactomes will enable a deeper understanding of the binding mechanisms and thereby inform rational candidate selection for subsequent experimental validation. In addition, an evaluation of all finishing elite candidates via refined secondary screening consensus methods and integration with QSAR and ADMET predictions allows for iterative consideration of the elite candidates to ultimately arrive at a select group of high-confidence lead candidates for experimental testing. Furthermore, through the use of ethnobotanical knowledge and biological relevance, high-throughput virtual screening pipelines substantially improve the efficiency and success of plant-based drug discovery programs.

Lead Bioactivity Identification and Optimization

A major milestone in the development of drug discovery based on natural products is the discovery and subsequent optimization of bioactive compounds that can serve as drug candidates. While natural products are believed to be a source of varied types of compounds with biological activity, only a small number of compounds found in nature are suitable for drug development due to their limited pharmacokinetic and pharmacodynamic properties.⁵⁶ The successful discovery of lead compounds requires the use of a multi-faceted approach, which includes both experimental and computational methods, to identify potential lead compounds, assess their drug-like characteristics, and improve bioactivity and safety for potential drug development candidates.¹⁶

Identifying Potential Leads From Natural Sources

Historically, a variety of bioactive natural product sources have been identified, since they contain many types of bioactive compounds that may have potential for use as therapeutics. The identification of potential leads often begins with ethnopharmacological data in which plants used historically to treat specific diseases are prioritized for exploration of therapeutic use. This strategy increases the likelihood of bioactivity when screening for active pharmaceuticals. Currently, the strategies for lead identification include bioassay-guided fractionation, high-throughput screening (HTS), and virtual bioinformatics screening. Crude extracts/fractions will then be screened *in vitro* for activity with defined endpoints including, but not limited to, target-based and disease model screening. The rapid advancement of separative techniques, such as HPLC and LC–MS, enables rapid access to bioactive phytochemical compounds. Bioinformatics are also becoming increasingly important for the

identification of leads through virtual screening, or ligand-based versus target-based. In addition, molecular docking and pharmacophore modeling are useful tools that aid in the estimation of binding affinities and specific interaction patterns of natural compounds with therapeutic targets. Through all these advanced developments, compounds that display high binding potential, selectivity, and beneficial interaction traits can then be selected for subsequent evaluation as leads.⁵⁷

ADMET Prediction and Drug-Like Assessment

Even with their potential to be bioactive, many natural compounds are eliminated during later phases of drug development because of poor ADMET properties. Therefore, it is imperative that ADMET parameters be identified early in order to lower attrition rates and save on costs of development. In silico tools are commercially available to estimate the oral bioavailability, intestinal absorption, permeability of the BBB, metabolic stability, and toxicity of a lead compound at an early stage. Bioavailability, absorption, permeability, stability, and toxicity parameters are predicted using acceptably defined likelihood rules, including Lipinski's rule of five, as well as various other potential filters, and compounds other than these parameters have a much lower chance of being useful or approved for clinical use.⁵⁸ Toxicity prediction models can provide information on potential toxicity including, but not limited to, hepatotoxicity, cardiotoxicity, mutagenicity, and inhibition of CYP450 enzymes. Therefore, on eliminating these compounds based on ADMET at an early stage, there is a greater likelihood of focusing on chemical messengers that will successfully make it to clinical use.⁵⁹ The compilation of ADMET parameters routinely assessed for pharmacological activity during the lead optimization process of phytochemicals is summarized in Table 3.

Multi-Target Drug Strategies and Network Pharmacology Approaches

Natural compounds frequently have multiple drug effects, as their naturally occurring active compounds

act on several different targets within the human body. Recent advancements in network pharmacology provide researchers with a means to leverage this site target action of natural compound active components by integrating systems biology with bioinformatics and computer-based modeling to reveal the interactions between various compounds with their respective targets/pathways, as well as how the combined actions of each of these compound/target/pathway interactions create a synergistic effect.⁶⁵ When constructing the compound–target pathway networks of phytochemicals, researchers may gain an understanding of which compounds and/or pathways may interact to produce maximum therapeutic effectiveness, which will aid in the development of rational polypharmacological drugs. The multi-target strategies provided by network pharmacology allow for greater therapeutic effectiveness and reduced risk of drug resistance or adverse effects typically associated with the use of single-targeted agents.⁶⁶

Structural Optimization and Analog Design Strategies

When researchers discover a lead that has the potential to be made into a new drug, the next step is to conduct structural optimization to increase potency/selectivity/pharmacological properties. Chemical modifications to NATURAL products that have complicated scaffolding(s) may increase “drug-like” characteristics but at the same time may also help maintain biological activity. Structure–activity relationship (SAR) studies are key to this process.⁶⁷ In SAR studies, researchers study how systematic modifications of functional groups will alter bioactivity. One example of how computer-aided drug design is carried out is through molecular dynamics simulations and quantitative structure–activity relationship (QSAR) models, where researchers simulate how a structural change will affect binding to the target and/or the stability of the modification. Analog design includes simplification of the complex newly designed structures, creation of bioisosters (or substituting atom for atom with other atoms), and optimal configuration of the

Table 3 | Key ADMET parameters considered during lead optimization of phytochemicals^{60–64}

ADMET Category	Parameter	Significance in Lead Optimization
Absorption	Aqueous solubility	Determines oral bioavailability and formulation feasibility
	Intestinal permeability (Caco-2)	Predicts absorption through intestinal epithelium
	P-glycoprotein interaction	Assesses efflux liability and bioavailability limitations
Distribution	Plasma protein binding (PPB)	Influences free drug concentration and tissue distribution
	Blood–brain barrier (BBB) penetration	Determines central nervous system exposure
Metabolism	Cytochrome P450 inhibition	Evaluates metabolic stability and drug–drug interaction risk
	Metabolic clearance	Predicts in vivo half-life and dosing frequency
Excretion	Renal clearance	Indicates elimination efficiency and accumulation risk
Toxicity	Hepatotoxicity	Assesses potential liver toxicity
	Cardiotoxicity (hERG inhibition)	Predicts the risk of cardiac arrhythmias
	Mutagenicity/carcinogenicity	Evaluates the long-term safety profile

stereochemistry of these analogs so that they possess good solubility/metabolic stability/target specificity. Through these iterative cycles, researchers will eventually create optimized lead candidates from their NATURAL product's "backbone" that have greatly increased therapeutic potential compared to the original.⁴²

Integration of Computational and Experimental Approaches

The use of both computational and experimental approaches together represents a new approach in drug discovery focused on using natural products, representing significant changes in how drug discovery can occur. Computational methods allow us to quickly predict and test large numbers of natural products for activity at low cost. However, we must test each natural product to determine whether it is biologically active and safe for human use. Computational methods used in combination with experimental approaches allow us to more accurately identify potential leads, eliminate false positives, and accelerate the discovery of new natural products that may be developed into drug products.⁶⁸ Because this combined approach provides more accurate information about the molecular basis, therapeutic potential, and safety of natural products, it improves the likelihood of development of natural products into drugs. Furthermore, this combined approach allows researchers to iteratively test and refine their hypotheses and methodologies, and to improve selection and optimization of the best natural products through cycle iterations. This iterative process leads to a more complete characterization of the molecular basis, therapeutic potential, and safety profile of natural products.⁶⁹

Validation of Computational Predictions In Vitro and In Vivo

Computational predictions, such as molecular docking scores, binding free energies, QSAR models, and ADMET assessments, provide valuable preliminary insights into the potential efficacy and safety of natural compounds. However, these predictions must be validated experimentally to establish their biological significance.

In Vitro Validation

In vitro assays are typically the first experimental step used to confirm computational findings. Cell- and enzyme-based assays are used to evaluate target inhibition, receptor binding, cytotoxicity, antioxidant activity, anti-inflammatory potential, or neuroprotective effects, depending on the disease model. For example, compounds predicted to interact strongly with a specific enzyme or receptor are tested for their inhibitory potency, selectivity, and dose-response relationships under controlled laboratory conditions. These assays also help verify predicted mechanisms of action by assessing downstream biological effects, such as changes in gene expression, protein levels, oxidative stress markers, or signaling pathway modulation. Importantly,

in vitro studies provide rapid feedback on compound efficacy while reducing the number of candidates progressing to more resource-intensive in vivo studies.⁷⁰

In Vivo Validation

In vivo investigations are crucial for verifying the significance of computational forecasting regarding pharmacology functioning within multilayered living systems. Animal investigations have been performed to test the pharmacodynamics, pharmacokinetics, biodistribution, toxicities, and overall safety of the investigational agents. The various parameters predicted using computational models (e.g., bioavailability, penetration through the blood-brain barrier, and metabolic stability) have been measured through laboratory testing to confirm that in silico models accurately simulate in vitro results. By directly comparing the data generated through computational predictions to those generated by in vivo models, discrepancies between theoretical performance and actual performance can be recognized and aid in the enhancement of predictive models, thereby improving their accuracy for future application. The validation of computational and in vivo models is particularly important in the context of the study of natural product-based investigational compounds, which frequently demonstrate complex metabolic pathways and a multiplicity of target sites.⁷¹

Linking Computational Data to Biological Assays

The successful integration of drug development through computer-based modeling and experimental design/interpretation is a significant barrier to developing new therapeutics. Computer-based models provide critical insight into how to choose the right experimental biological assays through the identification of key molecular targets, signaling pathways, and mechanisms associated with disease. For example, computer-generated molecular docking and pharmacophore modeling results provide the basis for ranking targets based on their likelihood of success in enzyme- or receptor-focused assays, while computer-aided network pharmacology analysis provides guidance on how to perform multi-target screening and pathway-focused studies.^{12,32} Quantitative structure-activity relationship (QSAR) models predict which chemical structures may produce specific biological activity based on chemical structural changes, allowing researchers to select the best dose and make chemical changes prior to performing experimental tests. Additionally, the use of computer-generated predictions aids in providing mechanism-based rationales for observed biological effects in experimental assays. The difference between computer prediction results and actual experimental results may highlight the limitations of the computer models or may provide evidence of a previously unrecognized biological interaction. Continued feedback between computer modeling and experimental validation or development enhances the development of more efficient computer prediction algorithms and experimen-

tal procedures, thereby increasing the certainty and translational relevance of drug discovery.⁶⁸ This schematic overview, showing the integration of computer-generated predictions into the validation process of drug discovery from natural products, is presented in Figure 4.

Success Stories in Drug Discovery Based on Natural Products

Many notable breakthroughs in drug discovery from natural products have come from effectively integrating both computational and experimental methods. Often, researchers are able to rapidly identify promising new candidates for drug development using

computer-based methods to screen through large collections of natural products. After identification of these leads on the computer, they are confirmed through laboratory testing. The ability to predict new targets and mechanisms of action has enabled researchers to develop new indications for already-existing natural products through the process of “repurposing” them for therapeutic use, utilizing computational tools. Structure-based modeling and quantitative structure–activity relationship (QSAR)-based optimization have allowed researchers to produce more potent, selective, and better pharmacokinetic profiles of analogs of the original products than was previously possible.¹⁶ All of the examples provided in this article illustrate the potential of combining computational and experimental approaches to overcome some of the historical barriers associated with studying natural product research, which were due to issues like complex structure, limited availability, and uncertain bioavailability. By reducing the amount of “trial and error” involved in a typical experiment and by concentrating efforts on only the most promising candidates, the combination of computer- and laboratory-based methodologies has accelerated the discovery and development of natural product-derived therapeutics.

Case Studies: AI-Guided Natural Product Discovery With Prospective Validation

The integration of artificial intelligence (AI), metabolomics, and genome mining has enabled several well-documented discoveries in natural product research where computational predictions were prospectively validated through experimental assays. The following case studies illustrate reproducible AI-guided discovery pipelines, detailing data sets, algorithms, validation protocols, and measurable outcomes.

Machine Learning-Guided Discovery of Abaucin: A Narrow-Spectrum Antibiotic

A landmark study demonstrated the use of a graph neural network (GNN) model to identify a narrow-spectrum antibiotic effective against *Acinetobacter baumannii*.

Data Sets

The model was trained using bacterial growth inhibition data derived from high-throughput screening data sets containing compounds with experimentally measured antibacterial activity. The trained model was subsequently applied to screen approximately 6,680 structurally diverse compounds from the Drug Repurposing Hub and additional compound libraries.

Algorithms and Computational Tools

A message-passing neural network (MPNN) architecture was used to encode molecular graphs and predict antibacterial activity. Compounds were represented as graph-based molecular structures derived from SMILES strings. The model was optimized using supervised learning and validated using internal cross-validation metrics prior to prospective screening.



Fig 4 | Integration of computational predictions with experimental validation in natural product drug discovery. Computational approaches, which include virtual screening, molecular docking, quantitative structure–activity relationship (QSAR) modeling, and ADMET prediction, facilitate the identification of potential lead compounds sourced from nature and help to prioritize the selection of those that are most readily available for experimental testing. Verified candidate leads are evaluated through in vitro and in vivo experiments for their biological activity and mechanism of action, safety, and effectiveness. Feedback from these evaluations is then considered as improvements to existing models and allows for further optimization of identified lead compounds. The combination of these two methods of discovery creates a robust and repeatable process for the efficient and consistent discovery of bioactive compounds

Validation Protocol

Top-ranked compounds predicted by the model were subjected to in vitro antibacterial assays against *A. baumannii*. Active candidates were further evaluated for species selectivity, cytotoxicity, and mechanism of action. Genetic and biochemical analyses identified disruption of lipoprotein trafficking as the primary mechanism. In vivo efficacy was tested using a murine wound infection model.

Measurable Outcomes

Nine compounds exhibited measurable antibacterial activity, with one compound, named abaucin, demonstrating potent and selective bactericidal effects against *A. baumannii*. In vivo studies confirmed therapeutic efficacy with reduced bacterial load in infected animals. This study represents a complete computational-to-preclinical validation pipeline.

Molecular Networking-Guided Discovery of Kyonggic Acids

Molecular networking using the Global Natural Products Social Molecular Networking (GNPS) platform has facilitated the identification of novel microbial natural products.

Data Sets

Non-targeted LC-MS/MS metabolomic data sets were generated from multiple *Massilia* bacterial strains. Spectral data were processed and uploaded to GNPS for feature-based molecular networking analysis.

Algorithms and Computational Tools

Feature-based molecular networking (FBMN) was applied to cluster MS/MS spectra based on fragmentation similarity. Spectral matching and database annotation enabled prioritization of clusters lacking reference matches, indicating potential chemical novelty.

Validation Protocol

Prioritized molecular clusters were subjected to large-scale fermentation and chromatographic isolation. Structural elucidation was performed using high-resolution mass spectrometry and nuclear magnetic resonance spectroscopy. Biological assays evaluated enzyme-inhibitory activity.

Measurable Outcomes

Several previously unreported kyonggic acid derivatives were isolated and structurally characterized. The compounds demonstrated measurable tyrosinase inhibitory activity with reported IC_{50} values, confirming the effectiveness of molecular networking-guided prioritization.

Integrated Molecular Networking and Bioassay Prioritization: Discovery of Methyl-Kalafunginate

An advanced workflow combining metabolomic molecular networking with orthogonal functional assays enabled the discovery of a novel pyranonaphthoquinone derivative.

Data Sets

LC-MS/MS metabolomic profiles were obtained from *Streptomyces tanashiensis* extracts. Additional biophysical screening data were generated through single-molecule interaction assays to prioritize bioactive fractions.

Algorithms and Computational Tools

Feature-based molecular networking was used to identify unique metabolite clusters. Integration of metabolomic signatures with functional assay outputs enhanced candidate prioritization prior to isolation.

Validation Protocol

Selected fractions were subjected to scale-up fermentation, chromatographic purification, and structural elucidation using NMR, HRMS, and stereochemical analysis. Cytotoxicity was evaluated across a panel of human cancer cell lines.

Measurable Outcomes

A novel compound, methyl-kalafunginate, was structurally characterized and demonstrated potent cytotoxic activity with sub-micromolar IC_{50} values in multiple cancer cell lines. This study highlights the effectiveness of combining AI-assisted metabolomic prioritization with experimental validation.

Genome Mining and Deep Learning for Biosynthetic Gene Cluster Discovery

Genome mining approaches using deep learning have enabled the identification of cryptic biosynthetic gene clusters (BGCs) leading to validated natural products.

Data Sets

Whole-genome sequencing data sets from microbial isolates were analyzed alongside curated biosynthetic gene cluster reference data sets. Complementary LC-MS/MS metabolomic data were integrated for compound detection.

Algorithms and Computational Tools

DeepBGC and related deep learning frameworks were used to identify and classify putative biosynthetic gene clusters based on sequence features and domain architectures. AntiSMASH was used for comparative annotation and cluster boundary prediction.

Validation Protocol

Predicted BGCs were prioritized based on novelty scores and biosynthetic potential. Selected clusters were activated through heterologous expression or optimized cultivation conditions. The resulting metabolites were isolated and structurally characterized. Gene knockout experiments confirmed biosynthetic linkage between predicted clusters and isolated compounds.

Measurable Outcomes

Multiple previously uncharacterized ribosomally synthesized and post-translationally modified peptides (RiPPs) and polyketide-derived compounds

were identified and experimentally validated. This genome-to-metabolome workflow demonstrates the translational power of AI-assisted genome mining.

Lessons for Reproducibility and Translational Impact

Across these case studies, several reproducible principles emerge as follows:

- Explicit reporting of data set origin and size.
- Clear specification of AI architecture and training strategy.
- Prospective experimental validation rather than retrospective fitting.
- Mechanistic investigation beyond activity screening.
- Applicability domain awareness to prevent overgeneralization.
- These examples collectively demonstrate that AI-guided natural product discovery can achieve translational relevance when computational predictions are rigorously integrated with systematic experimental validation.

Critical Synthesis: Quantitative Benchmarks, Predictive Performance, and Failure Modes

While computational methodologies have substantially accelerated natural product (NP) drug discovery, a critical evaluation of quantitative performance metrics and methodological limitations is essential to ensure translational reliability. Moving beyond descriptive accounts, this section synthesizes benchmark data and highlights common failure modes observed in AI-driven NP research.

Quantitative Benchmarks in AI-Guided Natural Product Discovery

Virtual Screening Performance Metrics

Structure-based virtual screening applied to NP-like libraries typically reports the following:

- ROC-AUC values ranging from 0.70 to 0.85 for well-curated data sets.
- Precision-recall AUC (PR-AUC) values between 0.40 and 0.75 depending on class imbalance.
- Enrichment factors (EF1% and EF5%) commonly between 5 and 25 when benchmarking against decoy sets.
- Docking-based enrichment often lower for macrocyclic or highly flexible phytochemicals.

However, performance variability is strongly dependent on the following:

- Data set quality,
- Target class,
- Decoy selection strategy, and
- Stereochemical specification.

QSAR and Machine Learning Models

For NP data sets, well-validated QSAR models typically demonstrate the following:

- Internal R^2 values between 0.70 and 0.90.
- Cross-validated Q^2 values above 0.60.
- External validation R^2 generally between 0.50 and 0.75.

- RMSE values dependent on endpoint variability.

Graph neural networks and deep learning models often show improved ROC-AUC (0.80+) compared to classical descriptor-based models; however, improvements may diminish under external validation when chemical space shifts occur.

Applicability domain (AD) assessment is critical, as extrapolation beyond NP chemical space frequently leads to inflated predictive claims.

Docking Versus Free-Energy Perturbation (FEP)

Comparative studies indicate the following:

- Standard docking scoring functions often exhibit moderate correlation with experimental binding affinity ($R \approx 0.30$ – 0.60).
- Free-energy perturbation (FEP) methods may achieve mean unsigned errors of ~ 1 – 2 kcal/mol for congeneric series.
- For NP-like ligands, docking performance often degrades due to the following:
 - high conformational flexibility,
 - multiple hydrogen-bonding networks, and
 - solvent-mediated interactions.

FEP, while more accurate, is computationally expensive, and sensitive to protonation and stereochemical states.

Failure Modes in AI-Guided Natural Product Discovery

Despite promising benchmarks, several systematic limitations persist.

Macrocycles and Conformational Flexibility

Natural products frequently contain the following:

- Macrocyclic rings,
- Rotatable bonds, and
- Intramolecular hydrogen bonding networks.

Failure Mode:

- Docking algorithms inadequately sample conformational space.
- Scoring functions misestimate entropic penalties.
- ML descriptors fail to capture dynamic conformations.

Stereochemical Complexity

Natural products often possess the following:

- Multiple chiral centers,
- Epimers, and
- Atropisomerism.

Failure Mode:

- 2D SMILES without stereochemical encoding distort predictions.
- Docking may assign incorrect binding orientation.
- QSAR models may treat enantiomers as identical.

Recommendation:

Explicit stereochemical representation and 3D conformer ensemble generation.

Protein Flexibility

Most docking workflows treat proteins as rigid.

Failure Mode:

- Induced-fit effects are not captured.
- Allosteric site misidentification.
- Binding pocket rearrangement is overlooked.

Recommendation:

Use ensemble docking, molecular dynamics refinement, or flexible docking protocols.

Data Set Shift: Synthetic to Natural Product Space

Many AI models are trained predominantly on synthetic compound libraries.

Failure Mode:

- Distributional shift when applied to NP chemical space.
- Reduced external R^2 .
- Overestimation of predictive confidence.

Natural products occupy distinct chemical space characterized by the following:

- Higher sp^3 content,
- Greater stereochemical richness, and
- Increased scaffold complexity.

Applicability domain analysis and chemical space mapping are therefore mandatory.

Data Imbalance and Limited NP Annotations

NP data sets are often

- Small in size,
- Skewed toward certain bioactivities, and
- Derived from heterogeneous experimental conditions.

Failure Mode:

- Overfitting,
- Inflated internal validation metrics, and
- Poor real-world predictivity.

Toward Evidence-Based AI in Natural Product Research

To improve robustness and translational value, future studies should

- Report ROC-AUC and PR-AUC for classification tasks.
- Provide external R^2 and Q^2 for regression models.
- Include enrichment factor benchmarking.
- Explicitly define applicability domains.
- Compare docking results with higher-level methods when feasible.
- Quantify uncertainty in predictions.

Quantitative benchmarking and transparent failure analysis are essential to transition AI-driven NP discovery from exploratory modeling toward reliable therapeutic development.

Future Perspectives and Challenges

Despite tremendous advancement in computational and experimental methods, the identification of medicines developed from natural products still faces a variety of obstacles that are scientific, technical, and ethical in nature. Once these barriers are overcome, the use of new and developing technologies will enable transition from the current utilization of natural products to

a more predictive, sustainable, and translatable field. Future research should continue to focus on increasing the quality of data, increasing the interpretability of the modeling output, increasing the integration of multi-dimensional biological data, and ensuring that drug discovery processes function in an ecologically sustainable manner.⁶⁸

Data Quality, Reproducibility, and Interpretability Concerns

Input data quality and reliability represent the most significant challenges to the effective utilization and reproducibility of computationally based approaches in natural product research. There are numerous publicly available phytochemical databases that lack sufficient detail, or provide incomplete or inaccurate chemical and biological data. As a result of variations among experimental protocols, plant growth/harvesting conditions, phytochemical extraction, and bioassay design and implementation, standardization of data and reproducibility are hampered. Moreover, data reproducibility is hampered by the inability to accurately report the necessary experimental parameters and validate the computational models against multiple data sets collected independently. Predictive models based on either small numbers of data sets or data sets that exhibit sampling bias are likely to produce overfit results (which are not applicable to actual biological systems). Problems with these overfitted results adversely affect confidence in the accuracy of in silico predictive models and greatly decrease the potential for utilizing these models in developing new products (i.e., applying the predictive results for developing new natural products).⁷² Interpretability is an additional major concern with the increased use of advanced machine learning and deep learning algorithms (e.g., neural networks). Although these algorithms often achieve very high levels of predictive accuracy, their inherent “black-box” nature raises questions about the biological explanation of their predictions. Thus, methods to increase transparency (explainable artificial intelligence [XAI]) are important for creating trust in these models, allowing researchers to generate hypotheses, and for regulators to approve the use of computational predictive models.

Integration of Omics, Cheminformatics, and Predictive Modeling

Integrating multiple omics data sets with cheminformatics and predictive modeling will drive the future of drug discovery from natural sources through the creation of holistic views of how natural product compounds affect biological systems. To better understand the disease targets and pathways affected by natural product compounds, researchers can connect multiple “omics” sources (genomic, transcriptomic, proteomics, and metabolomics) together to gain insights at the level of a system response to a natural product. By combining omics data with cheminformatics, researchers can discover active phytochemical components of a natural product, make predictions regarding the potential for compounds to bind to a target, and elucidate the

biological mechanism of action at a higher resolution than possible using either approach alone. Additional tools for predictive modeling also allow researchers to further integrate data from multiple data sets to build networks specific to diseases, determine the key regulatory nodes within a disease network, and prioritize the natural product compound as a multi-target therapeutic. Integration of heterogeneous types of “omic” data sets remains complicated due to the differences in the formats, scales, and level of noise associated with each data set type. Improvement in data harmonization, development of standard ontologies, and the creation of open-access interoperable platforms are essential to achieve the full potential of an integrated omic-based drug discovery pipeline.⁷³

Sustainability and Biodiversity Concerns in Natural Product Research

Biodiversity conservation and environmental sustainability have become progressive agenda items in the field of natural product research. In addition, overharvesting of medicinal plants, degradation of their environments, and climate change are having a significant detrimental impact on continued access to global biological diversity and biological resources into the future. Thus, it is critical that all products and/or raw materials derived from biological sources are sourced in an environmentally and ethically sustainable manner, that all plants used for medicinal purposes are cultivated by environmentally and ethically sustainable practices, and that, when possible, non-exploitable and/or non-renewable resources are replaced with renewable resources in the drug discovery process. Developments in the areas of synthetic biology, metabolic engineering, and recombinant technology may provide us with some effective alternatives to utilizing endangered species for sourcing compounds needed in drug development and discovery. In fact, the use of predictive algorithms combined with virtual screening approaches will allow greater efficiency in terms of sample collection, experimental resource usage, and closing of the discovery gap between innovation and ecological conservation through intelligent pipeline design.^{74,75}

The New Era of Intelligent, Predictive Discovery Pipelines

Artificial intelligence, big data, and systems biology are creating a new wave of intelligent and predictive drug discovery. The next-generation drug discovery pipelines will be more comprehensive and data-driven than the traditional trial-and-error model. In addition, the drug discovery pipeline of the future will include automating of compound generation, automatic learning from experiment, and adaptive optimization methods. The creation of “closed-loop learning systems” that refine their predictions on multiple experimental outputs will improve the efficiency, accuracy, and translational relevance of drug discovery. These intelligent drug discovery pipelines have the potential to dramatically reduce the amount of time, cost, and success rates of natural product derivatives. However, in order

to achieve this goal, inter-disciplinary collaboration, good data governance practices, and continuous validation must exist to establish trustworthiness, ethical compliance, and clinical impact.

Ethical Governance, FAIR Data Implementation, and Responsible AI in Natural Product Research

The ethical and sustainable advancement of artificial intelligence (AI) in natural product (NP) drug discovery requires structured governance frameworks that prioritize transparency, reproducibility, ecological responsibility, and translational reliability. As AI-driven methodologies increasingly influence early-stage therapeutic discovery, it is imperative to establish standardized reporting and sustainability-oriented practices.

Operationalizing FAIR Data Principles in Natural Product Informatics

The FAIR (findable, accessible, interoperable, reusable) guiding principles should be systematically integrated into AI-assisted natural product research workflows.

1. Findability

- Deposition of curated phytochemical data sets in publicly accessible repositories.
- Use of persistent identifiers (DOIs, InChIKeys, PubChem CIDs).
- Standardized metadata, including
 - plant species (with taxonomic authority),
 - geographic origin,
 - extraction method,
 - assay conditions,
 - biological endpoints, and
 - target protein identifiers (UniProt ID).

2. Accessibility

- Open-access deposition of
 - descriptor matrices,
 - training/testing data sets,
 - docking parameter files, and
 - QSAR scripts.
- Clear licensing terms (e.g., CC-BY).

3. Interoperability

- Adoption of standardized formats:
 - SMILES/SDF for chemical structures,
 - FASTA/PDB for protein targets, and
 - CSV/JSON for descriptor data.
- Use of controlled vocabularies and ontologies (e.g., ChEBI, Gene Ontology).

4. Reusability

- Detailed methodological documentation.
- Transparent reporting of
 - preprocessing steps,
 - data cleaning procedures,
 - feature selection criteria, and
 - hyperparameter tuning strategy.

FAIR implementation ensures that AI models trained on NP data sets remain verifiable, extendable, and reproducible across independent laboratories.

Explainable Artificial Intelligence (XAI) for Trust and Regulatory Acceptance

While deep learning models often demonstrate high predictive accuracy, their “black-box” nature limits interpretability and regulatory confidence. The integration of Explainable AI (XAI) techniques is therefore essential.

Recommended practices include the following:

- Use of SHAP (Shapley additive explanations) or LIME for feature importance interpretation.
- Visualization of descriptor contribution in QSAR models.
- Reporting molecular substructures influencing predictions.
- Providing decision-boundary analysis for classification models.
- Reporting uncertainty estimation and confidence intervals.

XAI improves

- Biological plausibility assessment.
- Hypothesis generation.
- Regulatory acceptability.
- Clinical translation potential.

Trustworthy AI in NP discovery must prioritize interpretability alongside performance metrics.

Biodiversity Conservation and Ethical Sourcing Practices

Sustainable natural product research must align with biodiversity conservation and ethical bioprospecting standards.

Concrete recommendations include the following:

1. Compliance with the Nagoya Protocol on Access and Benefit Sharing.
2. Transparent documentation of
 - plant collection permits,
 - institutional ethics approvals, and
 - local community engagement.
3. Preference for
 - cultivated over wild-harvested species,
 - renewable biomass sources, and
 - microbial or synthetic biology alternatives for rare compounds.
4. Use of AI-based prioritization to
 - minimize ecological sampling burden,
 - reduce overharvesting, and
 - optimize compound selection before large-scale extraction

Sustainable AI pipelines can significantly reduce environmental footprint by limiting unnecessary field collection and experimental redundancy.

Minimal Reporting Standards for AI-in-Natural Product Studies

To enhance reproducibility, translational value, and scientific integrity, we propose a minimal reporting checklist for AI-assisted NP research.

A. Data set Transparency

- Total number of compounds.
- Source database(s).
- Data split strategy (training/validation/test).

- Class balance information.
- Inclusion/exclusion criteria.

B. Computational Methodology

- Software/tools used (with version numbers).
- Descriptor calculation method.
- Docking protocol (grid size, scoring function).
- QSAR algorithm type.
- Hyperparameter tuning approach.

C. Model Validation

- Internal validation method.
- External validation data set.
- Performance metrics (R^2 , Q^2 , RMSE, ROC-AUC, MCC).
- Applicability domain assessment.
- Y-randomization testing (if applicable).

D. Reproducibility Resources

- Code repository link.
- Data set availability statement.
- Random seed reporting.
- Hardware specifications (if relevant).

E. Biological Validation

- In vitro assay details.
- Replicate number.
- Statistical analysis method.
- Dose–response modeling.

F. Sustainability Statement

- Source authentication.
- Collection permits.
- Ecological risk assessment (if applicable).

Translational Impact and Future Policy Directions

Standardization of ethical AI practices, FAIR data stewardship, biodiversity-conscious sourcing, and transparent reporting frameworks will significantly enhance

- Cross-study reproducibility.
- Regulatory alignment.
- Industrial adoption.
- Global collaboration.
- Sustainable innovation.

Future policy frameworks should encourage

- AI model auditing.
- Open benchmarking data sets for phytochemicals.
- International biodiversity-AI integration guidelines.
- Journals mandating AI transparency checklists.

A responsible, transparent, and sustainability-oriented AI framework will be critical in transforming natural product drug discovery from exploratory research into a predictable, clinically translatable, and environmentally responsible discipline.

Resources, Data Infrastructure, and Reproducibility Considerations

Table 4 displays the major databases and computational resources in AI-driven natural product research.

Data Curation, Stereochemistry, and Applicability Domain Considerations

While these databases and computational tools have accelerated AI-assisted natural product discovery,

Table 4 | Major databases and computational resources in AI-driven natural product research

Category	Resources	Description	Strengths	Key Limitations
NP databases	COCONUT	Collection of Open Natural Products database	Large open-access data set (>400k compounds)	Variable curation depth, stereochemistry inconsistencies
	NPAAtlas	Microbial natural product database	High-quality curated microbial NPs	Limited plant coverage
	SuperNatural II	Natural-like and natural compounds	Drug-likeness annotated	Redundant entries
	ZINC NP	Subset of purchasable natural products	Ready-to-dock formats	Commercial bias
	IMPPAT	Indian medicinal plants database	Ethnopharmacology linkage	Region-specific
	TCMID	Traditional Chinese Medicine database	Herb–compound–target networks	Variable validation
	AfroDB	African medicinal plant compounds	Underrepresented biodiversity	Smaller data set size
Metabolomics/Omics tools	GNPS	Molecular networking for MS data	Community-based annotation	Spectral misannotation risk
	antiSMASH	Biosynthetic gene cluster detection	Genome mining capability	Predictive uncertainty
	DeepBGC	ML-based BGC prediction	AI-enhanced accuracy	Training-data bias
	BiG-SLiCE	BGC clustering	Diversity assessment	Annotation incompleteness
Cheminformatics & AI tools	RDKit	Open-source cheminformatics toolkit	Flexible descriptor calculation	Requires scripting expertise
	DeepChem	ML for drug discovery	Integrated DL pipelines	Data preprocessing sensitivity
	Chemprop	Graph neural network model	Strong molecular prediction	Black-box interpretability limits

careful consideration of data integrity and chemical representation is essential.

Data Curation Challenges

Natural product data sets often contain

- Duplicate structures.
- Incomplete stereochemical annotation.
- Inconsistent protonation states.
- Variable assay conditions.
- Unstandardized plant taxonomy.

Improper curation may lead to

- Inflated performance metrics.
- Biased QSAR models.
- Poor external predictivity.

Best practice includes

- Structure standardization.
- Salt removal.
- Tautomer normalization.
- Stereochemical validation.
- Cross-database deduplication.

Stereochemistry Complexity in Natural Products

Natural products are highly stereochemically rich molecules with

- Multiple chiral centers.
- Conformational flexibility.
- Atropisomerism.
- Epimeric mixtures.

Many AI pipelines rely on simplified 2D representations (SMILES without stereochemical specification), which may

- Reduce biological relevance.
- Affect docking orientation.
- Distort descriptor calculations.
- Decrease translational accuracy.

Recommendation:

- Explicit stereochemical encoding.
- 3D conformer generation.
- Ensemble docking.
- Chirality-aware descriptors.

Applicability Domain (AD) in AI Models

AI and QSAR models trained on limited NP data sets may fail when applied to

- Structurally distant scaffolds.
- Underrepresented compound classes.
- Rare macrocycles.
- Highly flexible molecules.

Therefore, studies must report

- Applicability domain assessment method.
- Chemical space coverage.
- Distance-based AD metrics.
- Domain extrapolation limits.

Without AD evaluation, predictive performance claims may lack generalizability.

Conclusion

Computational methods and approaches for identifying new candidate drugs from natural plant sources have created a new and revolutionary paradigm that links traditional natural product-related research with the potential of modern computer-driven technologies. Novel in silico approaches include molecular modeling, virtual screening, QSAR analysis, network pharmacology, and ADMET prediction, all of which greatly enhance the ability to efficiently identify, validate, and develop therapeutics found in natural products. Significant advances have been made toward systematic discovery and identification of bioactive phytochemicals, establishing a rational basis for lead selection, and optimizing their pharmacologic and pharmacokinetic properties. The use of modern computational tools has reduced both the amount of chemical space available for searching for potential leads (thus eliminating the

need for trial-and-error approaches) and allowed researchers to discover new, molecular-based insights into the complex relationships between many different targets and paths as they relate to difficult-to-treat chronic diseases. These developments have greatly accelerated the pace of lead discovery and also led to an improved understanding of the molecular basis and function of the bioactivity of natural products derived from plants. While the field has made significant progress, many challenges lie ahead before we can unlock the full value of computational tools for natural product research. The key challenges that must be overcome include quality of data issues, reproducibility and interpretability of models, and integration of diverse biological data. To unlock the potential of computer-aided discovery in developing new therapeutic agents from plant materials, we will need to leverage the combined power of AI/multi-omic technology, cheminformatics, and systems biology to create smart, adaptable, and predictive processes for identifying drug candidates. In the future, computer-guided discovery is positioned to drive the evolution of the next generation of plant-based drugs. This collaborative effort, along with sustainable and ethical research practices, and improvement of predictive models with experimental data, provides the best possible path to faster, more innovative, and environmentally friendly drug development. The combination of computing power with the expansive chemical range available from plant sources can help fulfill many of today's unmet health care needs.

List of Abbreviations

AI – Artificial intelligence
 ADMET – Absorption, distribution, metabolism, excretion, and toxicity
 BBB – Blood–brain barrier
 CADD – Computer-aided drug design
 CNS – Central nervous system
 DL – Deep learning
 HTS – High-throughput screening
 LBVS – Ligand-based virtual screening
 ML – Machine learning
 MOE – Molecular operating environment
 MD – Molecular dynamics
 NP – Natural product
 NPR – Natural product research
 PPB – Plasma protein binding
 QSAR – Quantitative structure–activity relationship
 RMSE – Root mean square error
 SBDD – Structure-based drug design
 SAR – Structure–activity relationship
 SVM – Support vector machine
 VAA – Virtual activity assessment
 VS – Virtual screening
 XAI – Explainable artificial intelligence

Transparency Statement

The author confirms that this review article was developed based on information obtained from peer-reviewed scientific literature and publicly accessible academic databases. All sources were appropriately cited, and efforts were made to ensure accuracy,

transparency, and scientific integrity throughout the manuscript. The methodology used for literature selection and data interpretation was conducted in accordance with accepted academic standards. No data fabrication, manipulation, or intentional bias was involved in the preparation of this review.

Limitations

This review has certain limitations associated with literature-based analyses. The conclusions presented are dependent on the availability and quality of published studies, which may vary in computational methodologies, data sets, and validation strategies. Differences in experimental design and reporting standards across studies may influence interpretation. Furthermore, the rapidly evolving nature of artificial intelligence and computational drug discovery may lead to the emergence of new methodologies beyond the scope of the present review. Nevertheless, efforts were made to include recent and relevant literature to provide a comprehensive overview of the topic.

References

- Newman DJ, Cragg GM. Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *J Nat Prod*. 2020;83(3):770–803. <https://doi.org/10.1021/acs.jnatprod.9b01285>
- Atanasov AG, Zotchev SB, Dirsch VM, Supuran CT, International Natural Product Sciences Taskforce. Natural products in drug discovery: advances and opportunities. *Nat Rev Drug Discov*. 2021;20(3):200–216. <https://doi.org/10.1038/s41573-020-00114-z>
- Hong J. Role of natural product diversity in chemical biology. *Curr Opin Chem Biol*. 2011;15(3):350–354. <https://doi.org/10.1016/j.cbpa.2011.03.004>
- Singh K, Gupta JK, Chanchal DK, et al. Natural products as drug leads: exploring their potential in drug discovery and development. *Naunyn Schmiedeberg's Arch Pharmacol*. 2025;398(5):4673–4687. <https://doi.org/10.1007/s00210-024-03622-6>
- Chigozie VU, Ugochukwu CG, Igboji KO, Okoye FB. Application of artificial intelligence in bioprospecting for natural products for biopharmaceutical purposes. *BMC Artif Intell*. 2025;1(1):4. <https://doi.org/10.1186/s44398-025-00004-7>
- Nelson A, Karageorgis G. Natural product-informed exploration of chemical space to enable bioactive molecular discovery. In: *RSC Med Chem*. 12, 2020. p. 353–362. In: p. 3.
- Gaudêncio SP, Bayram E, Lukic Bilela L, et al. Advanced methods for natural products discovery: bioactivity screening, dereplication, metabolomics profiling, genomic sequencing, databases and informatic tools, and structure elucidation. *Mar Drugs*. 2023;21(5):308. <https://doi.org/10.3390/md21050308>
- Bugni TS, Richards B, Bhoite L, Cimbora D, Harper MK, Ireland CM. Marine natural product libraries for high-throughput screening and rapid drug discovery. *J Nat Prod*. 2008;71(6):1095–1098. <https://doi.org/10.1021/np800184g>
- Gangwal A, Lavecchia A. Artificial intelligence in natural product drug discovery: current applications and future perspectives. *J Med Chem*. 2025;68(4):3948–3969. <https://doi.org/10.1021/acs.jmedchem.4c01257>
- Caesar LK, Montaser R, Keller NP, Kelleher NL. Metabolomics and genomics in natural products research: complementary tools for targeting new chemical entities. *Nat Prod Rep*. 2021;38(11):2041–2065. <https://doi.org/10.1039/D1NP00036E>
- Mullowney MW, Duncan KR, Elsayed SS, et al. Artificial intelligence for natural product drug discovery. *Nat Rev Drug Discov*. 2023;22(11):895–916. <https://doi.org/10.1038/s41573-023-00774-7>
- Othman ZK, Ahmed MM, Kasimieh O, et al. Artificial intelligence for natural product drug discovery and development: current landscape, applications, and future directions. *Intell*

- Based Med.* 2025;12:100316. <https://doi.org/10.1016/j.ibmed.2025.100316>
- 13 Saldivar-González FI, Aldas-Bulos VD, Medina-Franco JL, Plisson F. Natural product drug discovery in the artificial intelligence era. *Chem Sci (Camb)*. 2021;13(6):1526–1546. <https://doi.org/10.1039/D1SC04471K>
 - 14 Wu Z, Chen S, Wang Y, et al. Current perspectives and trend of computer-aided drug design: a review and bibliometric analysis. *Int J Surg*. 2024;110(6):3848–3878. <https://doi.org/10.1097/J9S.0000000000001289>
 - 15 Yadav V, Tonk RK. Ligand-based drug design (LBDD). In: *Computer Aided Drug Design (CADD): From Ligand-Based Methods to Structure-Based Approaches*. Elsevier; 2022: 57–99.
 - 16 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):111–129. <https://doi.org/10.1038/nrd4510>
 - 17 Meijer D, Benidrir MA, Coley CW, et al. Empowering natural product science with AI: leveraging multimodal data and knowledge graphs. *Nat Prod Rep*. 2025;42(4):654–662. <https://doi.org/10.1039/D4NP00008K>
 - 18 Mandujano-Lázaro G, Torres-Rojas MF, Ramírez-Moreno E, Marchat LA. Virtual screening combined with molecular docking for the identification of new anti-adipogenic compounds. *Sci Prog*. 2025;108(1):368504251320313. <https://doi.org/10.1177/00368504251320313>
 - 19 Cheng T, Li Q, Zhou Z, Wang Y, Bryant SH. Structure-based virtual screening for drug discovery: a problem-centric review. *AAPS J*. 2012;14(1):133–141. <https://doi.org/10.1208/s12248-012-9322-0>
 - 20 Vázquez J, López M, Gibert E, Herrero E, Luque FJ. Merging ligand-based and structure-based methods in drug discovery: an overview of combined virtual screening approaches. *Molecules*. 2020;25(20):4723. <https://doi.org/10.3390/molecules25204723>
 - 21 Dagur P, Rakshit G, Ghosh M. Virtual screening of phytochemicals for drug discovery. In: *Phytochemistry, Computational Tools and Databases in Drug Discovery*. Elsevier; 2023: 149–179.
 - 22 Meng XY, Zhang HX, Mezei M, Cui M. Molecular docking: a powerful approach for structure-based drug discovery. *Curr Comput Aided Drug Des*. 2011;7(2):146–157. <https://doi.org/10.2174/157340911795677602>
 - 23 Agu PC, Añikwa CA, Orji OU, et al. Molecular docking as a tool for the discovery of molecular targets of nutraceuticals in diseases management. *Sci Rep*. 2023;13(1):13398. <https://doi.org/10.1038/s41598-023-40160-2>
 - 24 Paggi JM, Pandit A, Dror RO. The art and science of molecular docking. *Annu Rev Biochem*. 2024;93(1):389–410. <https://doi.org/10.1146/annurev-biochem-030222-120000>
 - 25 Fu C, Chen Q. The future of pharmaceuticals: artificial intelligence in drug discovery and development. *J Pharm Anal*. 2025;15(8):101248. <https://doi.org/10.1016/j.jpha.2025.101248>
 - 26 Cosconati S, Forli S, Perryman AL, Harris R, Goodsell DS, Olson AJ. Virtual screening with AutoDock: theory and practice. *Expert Opin Drug Discov*. 2010;5(6):597–607. <https://doi.org/10.1517/17460441.2010.484460>
 - 27 Bartuzi D, Kaczor AA, Targowska-Duda KM, Matosiuk D. Recent advances and applications of molecular docking to G protein-coupled receptors. *Molecules*. 2017;22(2):340. <https://doi.org/10.3390/molecules22020340>
 - 28 Schaller D, Šribar D, Noonan T, et al. Next generation 3D pharmacophore modeling. *Wiley Interdiscip Rev Comput Mol Sci*. 2020;10(4):e1468. <https://doi.org/10.1002/wcms.1468>
 - 29 Ponzoni I, Sebastián-Pérez V, Requena-Triguero C, et al. Hybridizing feature selection and feature learning approaches in QSAR modeling for drug discovery. *Sci Rep*. 2017;7(1):2403. <https://doi.org/10.1038/s41598-017-02114-3>
 - 30 Seidel T, Wieder O, Garon A, Langer T. Applications of the pharmacophore concept in natural product inspired drug design. *Mol Inform*. 2020;39(11):e2000059. <https://doi.org/10.1002/minf.202000059>
 - 31 Kandakatta N, Ramakrishnan G. Ligand based pharmacophore modeling and virtual screening studies to design novel HDAC2 inhibitors. *Adv Bioinform*. 2014;2014(1):812148. <https://doi.org/10.1155/2014/812148>
 - 32 Muhammed MT, Aki-yalcin E. Pharmacophore modeling in drug discovery: methodology and current status. *J Turk Chem Soc A Chem*. 2021;8(3):749–762. <https://doi.org/10.18596/jotcsa.927426>
 - 33 Giordano D, Biancaniello C, Argenio MA, Facchiano A. Drug design by pharmacophore and virtual screening approach. *Pharmaceuticals (Basel)*. 2022;15(5):646. <https://doi.org/10.3390/ph15050646>
 - 34 MacKenzie NM. New therapeutics that treat rheumatoid arthritis by blocking T-cell activation. *Drug Discov Today*. 2006;11(19–20):952–956. <https://doi.org/10.1016/j.drudis.2006.08.007>
 - 35 Chen Y, Kirchmair J. Cheminformatics in natural product-based drug discovery. *Mol Inform*. 2020;39(12):e2000171. <https://doi.org/10.1002/minf.202000171>
 - 36 Jónsdóttir SÓ, Jørgensen FS, Brunak S. Prediction methods and databases within chemoinformatics: emphasis on drugs and drug candidates. *Bioinformatics*. 2005;21(10):2145–2160. <https://doi.org/10.1093/bioinformatics/bti314>
 - 37 Raslan MA, Raslan SA, Shehata EM, Mahmoud AS, Sabri NA. Advances in the applications of bioinformatics and chemoinformatics. *Pharmaceuticals (Basel)*. 2023;16(7):1050. <https://doi.org/10.3390/ph16071050>
 - 38 Ferreira LT, Borba JV, Moreira-Filho JT, Rimoldi A, Andrade CH, Costa FT. QSAR-based virtual screening of natural products database for identification of potent antimalarial hits. *Biomolecules*. 2021;11(3):459. <https://doi.org/10.3390/biom11030459>
 - 39 Ganguly A. QSAR modelling for analysis of different medicinal and toxicological properties of different phytochemicals extracted from rare and endangered plants in India. *Int J Adv Sci Res*. 2025;16(6):31–40.
 - 40 Li J, Zhao T, Yang Q, Du S, Xu L. A review of quantitative structure-activity relationship: the development and current status of data sets, molecular descriptors and mathematical models. *Chemom Intell Lab Syst*. 2025;256:105278. <https://doi.org/10.1016/j.chemolab.2024.105278>
 - 41 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;26(19):9384. <https://doi.org/10.3390/ijms26199384>
 - 42 Vasilev B, Atanasova MA. (comprehensive) review of the application of quantitative structure–activity relationship (QSAR) in the prediction of new compounds with anti-breast cancer activity. *Appl Sci (Basel)*. 2025;15(3):1206. <https://doi.org/10.3390/app15031206>
 - 43 Kumar D, Wal P, Wal A, et al. Novel Nanoformulations to Overcome Obstacles in Herbal drug delivery for Alzheimer's disease. *Curr Top Med Chem*. 2025;25(28):3234–3250. <https://doi.org/10.2174/0115680266362594250527112018>
 - 44 De P, Kar S, Ambure P, Roy K. Prediction reliability of QSAR models: an overview of various validation tools. *Arch Toxicol*. 2022;96(5):1279–1295. <https://doi.org/10.1007/s00204-022-03252-y>
 - 45 Ojha PK, Mitra I, Das RN, Roy K. Further exploring rm2 metrics for validation of QSPR models. *Chemom Intell Lab Syst*. 2011;107(1):194–205. <https://doi.org/10.1016/j.chemolab.2011.03.011>
 - 46 Lahyaoui M, El Idrissi H, Moumni B, et al. Design, quantitative structure–activity relationships and computational studies of Acylshikonin derivatives: insights into antitumor activity. *Chem Phys Impact*. 2025;11:100959. <https://doi.org/10.1016/j.chphi.2025.100959>
 - 47 Gedeck P, Lewis RA. Exploiting QSAR models in lead optimization. *Curr Opin Drug Discov Devel*. 2008;11(4):569–575.
 - 48 Jukić M, Bren U. Machine learning in antibacterial drug design. *Front Pharmacol*. 2022;13:864412. <https://doi.org/10.3389/fphar.2022.864412>
 - 49 Governa P, Manetti F, eds. Virtual screening of natural product databases for drug discovery. *Pharmaceuticals*. 2023. [cited January 4, 2026].
 - 50 Ingle SG, Gade AK, Hedawoo GB. Systematic review on phytochemicals structure and activity databases. *Phytomed Plus*. 2024;4(4):100644. <https://doi.org/10.1016/j.phyplu.2024.100644>
 - 51 Karthikeyan Mohanraj KM, Karthikeyan BS, Vivek-Ananth RP, et al. IMPPAT: a curated database of Indian Medicinal Plants, phytochemistry and therapeutics. *Sci Rep*. 2018;8:4329.
 - 52 Alamri MA, Altharawi A, Alabbas AB, Alossaimi MA, Alqahtani SM. Structure-based virtual screening and molecular dynamics of phytochemicals derived from Saudi medicinal plants

- to identify potential COVID-19 therapeutics. *Arab J Chem*. 2020;13(9):7224–7234.
- 53 Chandrasekhar G, Srinivasan E, Sekar PC, Venkataramanan S, Rajasekaran R. Molecular simulation probes the potency of resveratrol in regulating the toxic aggregation of mutant V30M TTR fibrils in Transthyretin mediated amyloidosis. *J Mol Graph Model*. 2022;110:108055. <https://doi.org/10.1016/j.jmglm.2021.108055>
 - 54 Li Q, Ma Z, Qin S, Zhao WJ. Virtual screening-based drug development for the treatment of nervous system diseases. *Curr Neuropharmacol*. 2023;21(12):2447–2464. <https://doi.org/10.2174/1570159X20666220830105350>
 - 55 Gupta S, Shankar R. miWords: transformer-based composite deep learning for highly accurate discovery of pre-miRNA regions across plant genomes. *Brief Bioinform*. 2023;24(2):bbad088. <https://doi.org/10.1093/bib/bbad088>
 - 56 Newman DJ, Cragg GM. Natural products as sources of new drugs over the last 25 years. *J Nat Prod*. 2007;70(3):461–477. <https://doi.org/10.1021/np068054v>
 - 57 Fox Ramos AE, Evanno L, Poupon E, Champy P, Beniddir MA. Natural products targeting strategies involving molecular networking: different manners, one goal. *Nat Prod Rep*. 2019;36(7):960–980. <https://doi.org/10.1039/C9NP00006B>
 - 58 Ancuceanu R, Popovici PC, Drăgănescu D, Busnatu Ș, Lascu BE, Dinu M. QSAR regression models for predicting HMG-CoA reductase inhibition. *Pharmaceuticals (Basel)*. 2024;17(11):1448. <https://doi.org/10.3390/ph17111448>
 - 59 Lee H, Kim J, Kim JW, Lee Y. Recent advances in AI-based toxicity prediction for drug discovery. *Front Chem*. 2025;13:1632046. <https://doi.org/10.3389/fchem.2025.1632046>
 - 60 Dulsat J, López-Nieto B, Estrada-Tejedor R, Borrell JI. Evaluation of free online ADMET tools for academic or small biotech environments. *Molecules*. 2023;28(2):776. <https://doi.org/10.3390/molecules28020776>
 - 61 Ferreira LL, Andricopulo AD. ADMET modeling approaches in drug discovery. *Drug Discov Today*. 2019;24(5):1157–1165. <https://doi.org/10.1016/j.drudis.2019.03.015>
 - 62 Jung W, Goo S, Hwang T, 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;17(3):382. <https://doi.org/10.3390/ph17030382>
 - 63 Guan L, Yang H, Cai Y, et al. ADMET-score—a comprehensive scoring function for evaluation of chemical drug-likeness. *MedChemComm*. 2018;10(1):148–157. <https://doi.org/10.1039/C8MD000472B>
 - 64 Lin J, Sahakian DC, de Morais SM, Xu JJ, Polzer RJ, Winter SM. The role of absorption, distribution, metabolism, excretion and toxicity in drug discovery. *Curr Top Med Chem*. 2003;3(10):1125–1154. <https://doi.org/10.2174/1568026033452096>
 - 65 Noor F, Tahir UI, Qamar M, Ashfaq UA, et al. Network pharmacology approach for medicinal plants: review and assessment. *Pharmaceuticals (Basel)*. 2022;15(5):572. <https://doi.org/10.3390/ph15050572>
 - 66 Li L, Kar S. Leveraging network pharmacology for drug discovery: integrative approaches and emerging insights. *Med Drug Discov*. 2025;27:100220. <https://doi.org/10.1016/j.medidd.2025.100220>
 - 67 Dulak J. New anti-angiogenic drugs. In: *Book of Abstracts*. IEEE; 2010:42.
 - 68 Medina-Franco JL. Computational approaches for the discovery and development of pharmacologically active natural products. *Biomolecules*. 2021;11(5):630. <https://doi.org/10.3390/biom11050630>
 - 69 Simoben CV. Challenges in natural product-based drug discovery assisted within silico-based methods. *RSC Adv*. 2023;13(45):31578–31594. <https://doi.org/10.1039/D3RA06831E>
 - 70 Thomford NE, Senthebane DA, Rowe A, et al. Natural products for drug discovery in the 21st century: innovations for novel drug discovery. *Int J Mol Sci*. 2018;19(6):1578. <https://doi.org/10.3390/ijms19061578>
 - 71 Purba G. A systematic review of *in vitro* approaches for evaluating bioactive natural compound–target interactions in early drug discovery. *Sci Clin Pharm Res J*. 2025;2(4):11.
 - 72 Rodrigues T, Reker D, Schneider P, Schneider G. Counting on natural products for drug design. *Nat Chem*. 2016;8(6):531–541. <https://doi.org/10.1038/nchem.2479>
 - 73 Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. *Genome Biol*. 2017;18(1):83. <https://doi.org/10.1186/s13059-017-1215-1>
 - 74 Wilkinson MD, Dumontier M, Aalbersberg JJ, et al. The FAIR guiding principles for scientific data management and stewardship. *Sci Data*. 2016;3(1):160018. <https://doi.org/10.1038/sdata.2016.18>
 - 75 Romano JD, Tatonetti NP. Informatics and computational methods in natural product drug discovery: a review and perspectives. *Front Genet*. 2019;10:368. <https://doi.org/10.3389/fgene.2019.00368>