JOURNAL OF NATURAL REMEDIES
DOI: 10.18311/jnr/2025/48515
REVIEW ARTICLE

Revolutionizing Natural Product Science with AI: Integrating Diverse Data and Knowledge Graphs

Pooja Dave[1*], Krunal Panchal[2], Aayush Gandhi[3] and Brahmdutta Raval[4]

[1]Department of Pharmaceutics, RK University - 3600320, Rajkot, Gujarat, India; poojadave.pharma@gmail.com

[2]Department of Computer Science, University of Massachusetts - 02125, Boston, Massachusetts, USA

[3]Department of Computer Science, Stevens Institute of Technology, Hoboken - 07030, New Jersey, USA

[4]Department of Regulatory Affairs, Northeastern University, Boston - 02115, Massachusetts, USA

Abstract

Drug discovery and development is overly complex and time-consuming process because traditional trial-and-error methods depend on inefficient data utilization approaches. Traditional methods have limited scopes and delays in identification of lead compound. Natural products are more difficult to identify, isolation, synthesis, and evaluation by traditional methods. However, recent development of scientific advances applied in field of data analysis, biologically, chemically, medicinally by using computational transformations. Artificial intelligence is the most revolutionary tool in drug discovery, development, and predicting data which includes machine learning and deep learning. Natural Products (NPs) offer invaluable value during drug discovery because they provide the basis for numerous effective compounds. The growth of NP-related databases together with AI-based analytics has enabled increased exploration of bioactive compounds. The application of AI technologies in NP research enables new drug discoveries by using predictive models together with generative AI and large-scale data synthesizing methods. This review demonstrates the latest developments in AI applications for NP drug discovery and genomics by explaining their effects on natural product biology research as well as chemical adaptations and therapeutic analysis. Through AI integration in NP research scientists will speed up drug development and detect novel bioactive molecules within an ethical regulatory framework for computational drug discovery in its current development phase.

Major Findings: The review discusses the transformation redefining natural product research that artificial intelligence and its integration with knowledge graphs are driving by providing unified access to multimodal data, enhancing the predictive power of bioactivity and structural elucidation, and speeding up drug discovery pipelines. AI-based models eliminate the data fragmentation and complexity to allow the quick discovery, optimization and sustainable production of new bioactive compounds.

Keywords: Biosynthetic, Deep Learning, Machine Learning, Natural Products Drug Discovery, Predicting Targets, Target Identification

1. Introduction

The drug development field includes separate elements of Artificial Intelligence (AI), Machine Learning (ML), and Computational drug delivery systems which cooperate to yield improved results. AI creates self-operating machines through human-like simulation of learning processes along with reasoning and decision-making functions. The wide applications of this technology support drug discovery while it helps maximize therapy methods and it generates predictions and detects new drug properties. The process of human problem-solving through AI systems depends on vast datasets with complicated algorithms for function replication. Machine learning functions as a subset of AI which develops learning algorithms that

*Author for correspondence

Article Received on: 07.02.2025 Revised on: 29.03.2025 Accepted on: 29.07.2025

enable machines to extract data patterns before making predictions and decisions. Drug development benefits from ML by using big datasets to determine drug properties and bioactivity attributes and toxicity effects and lead compound optimization. Computational drug delivery focuses on medical formulation techniques and delivery methods for drugs unlike AI and ML which concentrate on better decision support and predictive analysis. Computational Drug Delivery utilizes numerical models to build better drug delivery methods which control dosage release mechanisms and choose body targets and better kinetic drug effects. Experimental data and simulations within computational drug delivery predict the drug release profiles together with optimizations in bioavailability. The identification along with optimization of drug candidates is a core aspect of AI and ML operations but the focus of computational drug delivery is to enhance drug delivery approaches and body interaction to maximize drug effectiveness. AI transforms scientific fields through its evolving benefits for protein folding analysis using AlphaFold along with its capability to generate efficient summaries of large bodies of literature by applying large language models. Natural product research benefits minimally from artificial intelligence utilization at present. Natural product data faces main difficulties because it consists of diverse multi-modal data that lacks uniformity as well as being dispersed. The various data patterns hinder the ability of standard AI models that depend on structured unconstructed data to work efficiently with these data sets. Establishing complete knowledge graphs for natural product data reveals the chance to create AI models that duplicate the complex decision systems used by scientists who work in this field. This paper investigates the barriers that prevent AI from reaching its complete potential in natural product research and proposes different strategies to overcome them[1-4].

The combination of AI with natural drug discovery demonstrates powerful potential to transform the entire industry. Drug discovery has always found its wealth in natural products that generate several life-saving medications through bioactive compounds discovered in nature. Labor-intensive along with time-consuming approaches currently exist for discovering and studying natural compounds. Fast tracking of biological discoveries becomes possible through AI because it provides efficient tools to analyze large databases. The pursuit of natural drug discovery faces an essential problem because natural product data exists with considerable complexity and variability. Natural product data include multiple input formats that combine chemical structures as well as biological activities alongside ecological surroundings. Multidimensional natural product data faces two major problems: (1) These datasets are unstandardised along with being (2) Distributed through multiple databases and repositories. The processing and learning effectiveness of regular AI models becomes limited when it comes to working with such datasets because they demand standardized relational data[1,2].

Creating comprehensive knowledge graphs represents a promising solution against these problems because they combine several types of natural product data into a structured network format. A knowledge graph consolidates information across several sources by creating direct links between chemical structures and their biological activities and ecological functions and genomic information. AI-based models function better when they view data holistically so that they can detect previously hidden patterns which deliver precise prediction results. AI models train knowledge graphs to replicate the decision-making procedures which natural drug scientists use in their work. The implementation of AI decisions generates better bioactive compound selection while simultaneously forecasting their operational methods and proposing new drug molecule development. Through its capabilities AI helps scientists discover natural products which were undetected previously allowing them to extend their natural drug discovery activities[2,3].

This paper explores the integration boundaries and prospects between Artificial Intelligence (AI) technology and knowledge graphs when used for natural drug discovery techniques. The discussion reveals how technological advancements make natural product data more accessible and speed up the discovery process until new effective natural drugs can be produced. Addressing the existing hurdles will unlock AI's maximum potential for natural drug innovation thus making forward-thinking medicinal therapies achievable[4]. In the Figure 1, AI based natural product development by various methods shown.

2. The Future of Natural Product Science: An AI Perspective

Research involving natural products analyzes multiple data forms which include genomic information together with proteomic and metabolomic data as well as spectroscopic and biochemical results. The different data types analyse the same biochemical entities from exclusive viewpoints while providing mutual enhancement for each other's analysis. Genomics reveals genetic lines that establish natural product manufacturing in biological organisms through its analysis while metabolomics identifies generated metabolites. The structural features of molecules can be determined from spectroscopic analyses while enzymatic pathways become understandable through biochemical studies. Multiple research viewpoints help scientists obtain complete knowledge regarding natural product analytical structures and functional properties. The current scientific data available in natural product research exists as multiple separate domains. The collection of many datasets includes known biases whose extent remains unclear because proper descriptions of covered areas are missing for genetic and proteinogenic and chemical structures. Many data collection projects handle diverse categories of artifacts with different levels of annotation combined with inconsistent metadata and structural features. Standardizing data collection becomes complex because various sources present different types of information. The solution requires new methods that transform current data sources into one standardized data repository to achieve comprehensive cross-referencing functionality across all-natural product-related datasets. By uniting genomic information that explains natural product development genetics with metabolomic detection of natural products scientists can better understand natural product biosynthesis pathways. By merging spectroscopic data that provides elaborate information on natural product structural features with biochemical data for understanding enzymatic pathway systems researchers can discover correspondence between natural product structures and functions. A single consolidated data warehouse including genomic, proteomic, metabolomic, spectroscopic and biochemical records enable standardization and reference comparison of acquired information. The mentioned repository functions as a

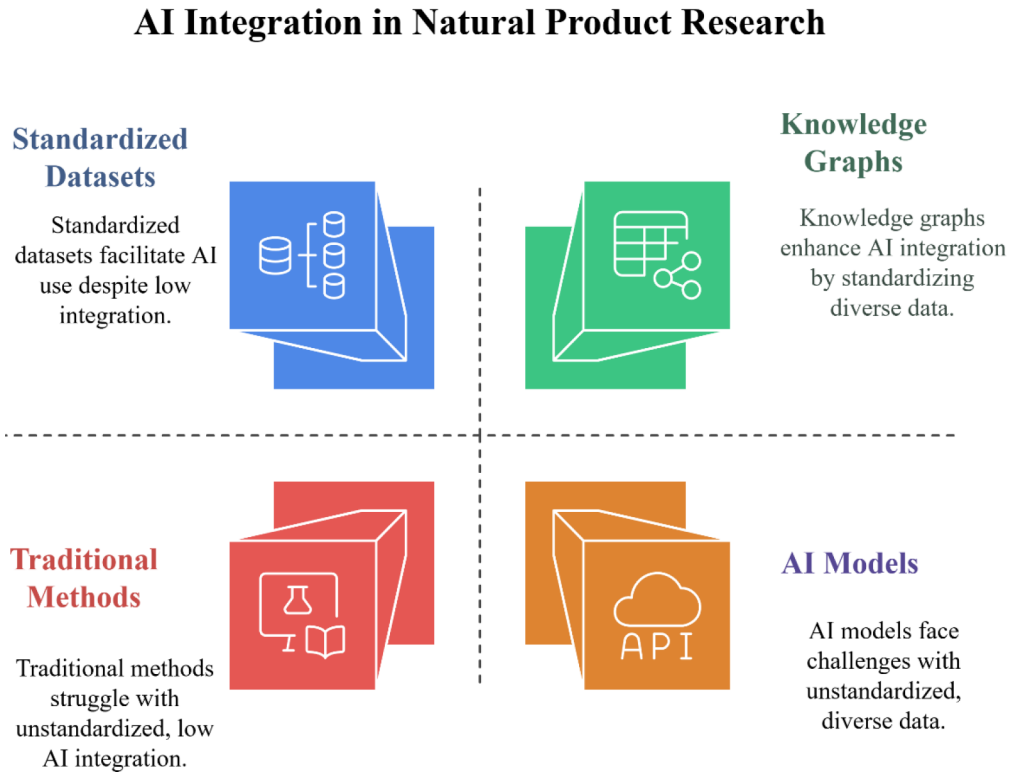


Figure 1. AI integration in natural product development.

resource for building AI models that determine natural product structures and functions alongside biosynthetic pathways. A merged data repository that connects genomic proteomic metabolomic spectroscopic and biochemical data allows researchers to discover better ways of finding and employing medically important natural products[5-8].

Modern artificial intelligence specifies the development of systems to fulfill responsibilities done by human intelligence. The field extends into multiple sub-domains that contain Machine Learning (ML) as one component which studies algorithms whose automatic improvement stems from gained experience. The field of AI incorporates algorithms that demonstrate parts of human cognitive abilities by representing and reasoning with knowledge along with deriving outcomes from their conceptualized information. AI functions as the broad field which contains ML as a specific part. Large datasets benefit from ML techniques because these systems learn examples to reach higher algorithm performance (Figure 2). It proves difficult for professionals to explain extensive information about their results[8-10].

The application of machine learning models during recent years includes three fundamental uses for molecular predictions and functional allocations and structural creation. The systems establish fundamental patterns in training data that enable them to forecast new cases from the same data area. Natural product chemical space analysis receives its foundation from AI techniques including unsupervised approaches and representation learning together with Natural Language Processing (NLP) and text mining methods for its encoding and representation functions. NLP techniques extract chemical information from huge database collections to support drug discovery activities. Through AI computations scientists now achieve faster results for drug discovery through predictions concerning molecular properties and novel molecule development. AI has established powerful tools to plan synthetic pathways which enable the production of complex natural products through specific routes. Natural product drug discovery processes benefit extensively from machine learning methods because they provide tools to detect bioactive compounds

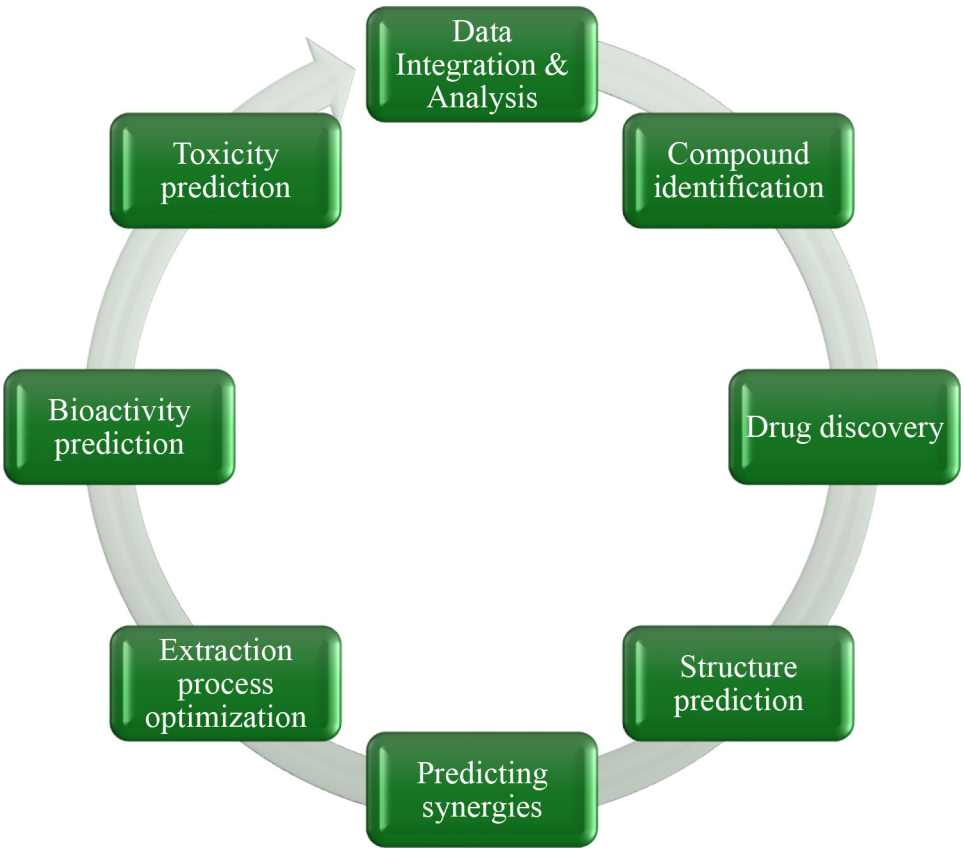


Figure 2. Illustration representing the uses of AI and ML in natural product research.

while recognizing their structural characteristics for designing drugs. Computer-Aided Synthesis Planning (CASP) serves complex natural products along with their mimetic compounds. The main obstacle today exists in creating mathematical models which both detect and use core patterns across multiple complex data types in natural products research[8-13].

The analysis of fundamental patterns requires an analytic technique known as causal inference because it stands separate from predictive analysis. Scientists must use causal inference to understand cause-effect dynamics while prediction operates as a method to forecast results from existing informational data. Scientific discovery through the fusion of different data relationships leads researchers to predict new natural product chemical findings. We identify chlorophyll presence in green plants as the simplest observation of natural product chemistry. Research would achieve natural product expectations through analysis of existing data connections which operate across multiple data categories. The practice of natural product anticipation exceeds the estimation of chemical structures since it encompasses all fundamental aspects of natural product science. Scientific personnel make predictions about natural products through genomics or collection data without necessitating the use of metabolomics information directly. The microbial populations inhabiting the producer organism make possible the prediction of natural product bioactivity. Scientists achieve better natural product research discoveries through data integration coupled with relationship understanding between different data sets[10-13].

To show the power of a natural product knowledge graph, one way would be to discover the relationships between tandem mass spectrometry fragmentation patterns and the metabolic building blocks of known or predicted natural products by way of retro biosynthesis prediction. It can be linked to predicted (meta) genome-identified biosynthetic gene clusters for which biosynthetic natural product pathways were predicted. These relationships can be visualized as a graph structure, and the graph structure can be used for estimating the natural product chemistry of a microbiome sample through accessing only one of those data types, Metabolomics or Metagenomics. Nodes (vertices) and edges are constituting a graph structure. Entities are assigned to nodes and the edges denote the relation between 2 nodes. A digraph is a directed graph, meaning that edges in a digraph indicate one way relationships (if one vertex is on the digraph, then on cannot 'visit' the other vertex 'without reaching it from the other way round') whereas in an undirected graph, they indicate bidirectional relationships (if one vertex is on the graph, then on can visit the other vertex without having to reach it through the other vertex first.). Edges in graphs can have costs, so they are weighted, or may not have costs, in which case the graph is unweighted. The graph model is used to represent entities and the relationship between entities. Heterogeneous graphs (*i.e.*, graphs with nodes and edges having different properties) naturally accommodate multimodal and interconnected datasets. This contrasts with the often-used non-relational tabular data, where each sample (*i.e.*, row) expects value for every feature (*i.e.*, column), making it difficult to combine non-relational tabular datasets containing non-overlapping sets of features. Furthermore, dependencies and other types of relationships can easily be encoded in a graph by linking nodes with (weighted) undirected and directed edges, and even hyperedges. Hyperedges are edges that connect multiple nodes simultaneously. This approach allows for a more flexible and comprehensive representation of complex relationships within the data. It is the ideal framework for a comprehensive graph structure that would integrate all available data in natural product science to facilitate large scale causal inference within the field. Knowledge graphs, which are these comprehensive graph structures, are known. Structured information about knowledge is represented in a knowledge graph that can be machine readable. It is constructed in terms of a network of linked data points in which each data point represents an element of information or idea and the relationships between them are explicitly defined. Knowledge graphs store and organize data in a format to facilitate information retrieval, data analysis or reasons.

Knowledge graphs have greater predictive abilities than traditional tabular data because unconstructed data exists primarily for new prediction tasks rather than predictive analysis for new node and edge creation. A professionally designed knowledge graph provides the needed base for artificial intelligence models to understand and imitate the logical processes of natural

product chemists and researchers. The AI system would both complete missing details in metabolic pathways and establish complete pathways from phenotypic data in the way a natural product scientist operates. Knowledge graphs originated from Tim Berners-Lee's work in the 1990s although they have not gained widespread use in natural product science research. Graph data presents data through tables in data organization to show different node entities along with their connections between them. These entities connect with one another through their relationships by using separate linking tables. The main separation between graph-based tables and single-table datasets emerges from their different priority between connecting entities in graph tables versus their independent data structure format in single tables. Knowledge graphs have proved their effectiveness in biomedical research through current practical applications. Text-mining solutions help conduct integrative biology through the extraction and integration of data from large datasets. The BioBLP (BERT for Link Prediction) framework demonstrates through its work that multimodal biomedical knowledge graphs hold value in predicting complex biological interactions to increase our comprehension of biological systems. The FORUM framework serves as a notable example because it links scientific literature to public databases to find meaningful associations and make hypothesis predictions in metabolomics studies. Research benefits enormously from well-constructed knowledge graphs because such systems deliver a full understanding of existing data[14-16].

Recent research by Gaudry *et al.* presented the Experimental Natural Products Knowledge Graph (ENPKG) as an example which illustrates the process of transforming unstructured unpublished data into connected public information. The ENPKG represents a trailblazing implementation that demonstrates how semantic web technologies maintain and reorganize large metabolomics datasets toward the detection of novel bioactive compounds. The new method presents a path toward creating Artificial Intelligence that discovers comprehensive patterns in natural product science. Natural product science should establish a dynamic knowledge graph structure called natural product science knowledge graph to fulfill its vision. The natural product science knowledge graph should contain information from multiple fields including natural product chemical structures together with metabolomics data (mass spectra) genomic data (Biosynthetic Gene Clusters or BGCs) assay data (bioactivity assay read-outs) and expert description (experimental design and data quality comments). The federation of knowledge through a framework represents the optimal method. Natural product researchers must continue the example started by LOTUS by creating a free open knowledge base through which they consolidate referenced structure-organism pairs. The distributed knowledge management systems can make use of the centralized resource framework just like ENPKG operates. Data access and analysis requires no centralization of data to achieve comprehensive results through this method. Such knowledge graphs need to progress through additional inclusion of orphan data that includes compounds lacking BGCs or mass spectra as well as mined metabolic gene clusters with no associated compounds. The research will continue through studies on data with reduced connections provided by this approach. Wikidata development together with individual knowledge graph evolutions should proceed naturally through contributions from the scientific community and their requirements. For reproducible science, the practice demands both change tracking and version control of these resources. The method ensures experimental data remains easy to reproduce regardless of how the knowledge base continues to develop. Scientific studies stay intact while remaining easily reproducible through mechanisms that enable users to select specific previously versioned linked resources and prevent newly added entities from entry after predetermined time periods. The establishment of this vision depends on continuous collaboration between community members. Two successful examples of these projects exist through the Earth Metabolome Initiative and the Human Microbiome Project. Scientific research benefits profoundly from coordinated community projects which produce substantial advancement. Natural drug discovery research stands to undergo substantial transformation through the combined use of AI systems and knowledge graphs which improve data examination as well as pattern detection leading to fresh bioactive compounds[15-23]. In Table 1, AI based applications and examples were given which is used for natural products.

Table 1. Natural drug discovery and drug identification by AI

Application	Role of AI	Examples
Bioactive compound screening	AI analyzes large natural product databases to identify potential drug candidates.	Machine learning models identify alkaloids from medicinal plants as anti-cancer agents.
Genome mining for Biosynthetic Gene Clusters (BGCs)	AI predicts the presence of genes responsible for synthesizing bioactive compounds.	AI-guided genome mining identifies new antibiotics from *Streptomyces* bacteria.
Molecular docking and target prediction	AI predicts interactions between natural compounds and biological targets, improving hit selection.	Deep learning models predict curcumin's binding affinity to inflammatory proteins.

3. Unlocking the Value of Data Diversity in Natural Product Science

The operation of multimodal models depends on integrating text and images along with audio and video content and extra data types to achieve improved data comprehension. An effective model functions with matched datasets that need each sample to have its specific corresponding data points. The training process requires available values for certain images but samples without such values must be eliminated from the training set. The lack of properly matched datasets affects natural product science strongly because these paired datasets are both sparse and incomplete for different data categories. The available datasets typically do not follow standardized well-curated ontologies or controlled vocabularies to enable effortless study integration. The worldwide database of natural product diversity remains incomplete because it contains numerous empty spaces where some data types are missing, and annotations need improvement. Practical limitations prevent researchers from removing samples because metadata is incomplete particularly when costly data collection methods such as metagenomics or RNA sequencing are considered. Scientific research cannot tolerate disposing of essential examples solely because different annotation levels exist in these samples. Every point of information available from this scarce data set maintains its own independent worth. The application of knowledge graphs as a solution helps create mappings between data points even if some pieces of information remain incomplete while using unique identifiers for linking purposes. The connection among data points enables the prediction of incomplete properties as well as the implementation of weight distribution for samples according to their metadata density. The fundamental mechanism states that when different data collections include both high-quality and lower-quality datasets they can form a better understanding of the natural product data system. A knowledge graph which includes experimental observations such as cytological profiles or bioactivity assay results becomes useful for predicting protein targets or bioactivities when essential labels remain unavailable. The combination of structured data relationships from various diverse datasets enables us to supplement incomplete labels which expand the analytical processing capabilities of multimodal models[24-26].

Diverse data types have different quality levels combined with varying expenses which lead to assessment difficulties because of the necessary expertise within separate scientific areas. Different data types represent either inexpensive low-value information or expensive difficult-to-acquire valuable insights. The information value of long-read genome assemblies exceeds short-read genome assemblies despite their parallel assembly format output. The complexity and accuracy levels of molecular conformer calculations depend entirely on the calculation method employed. Working with limited high-value BGC bioactivity annotation information requires maximizing accuracy since mistakes have significant impacts on the results. Domain experts encounter difficulties when assessing multidisciplinary datasets and this complication leads many open-access datasets to become non-FAIR due to missing findability along with accessibility, integration, and reusability requirements. Quality assessments in metadata will help natural product science field researchers gain complete understanding of research materials. Datasets with various annotation stages must include clear explanations about their evaluation system rather than using unexplained methods to prefer newer information over established annotations. It is equally important to keep accessible the whole line

of modified annotations. The integration of different quality levels in a knowledge graph becomes possible through weighting relationships based on the density of metadata available. The methodology allows incomplete data entry while it uses quality levels as a parameter to optimize model predictions thus increasing analytical precision.

A knowledge graph built from natural product science data would offer deeper insights into nature's underlying patterns and preferences. A fundamental challenge in data collection is the tendency to overlook negative results. A comprehensive knowledge graph integrating all natural product science data would be a powerful asset for the field and beyond. It would not only enable the extraction of multimodal datasets for supervised learning tasks that rely on non-graph structured databut also facilitate direct learning on graph-structured data. This capability is particularly useful for predictive tasks such as link or node prediction, where the graph can help identify potential biomarkers and uncover previously unrecognized relationships within data[26-30].

3.1 Utilizing Natural Product Data for Natural Product Discovery with AI

The definition of an artificial intelligence breakthrough in natural product science involves a model achieving human-level inference while extrapolating information from various data types. Advanced AI systems possess the capability to analyze mass spectrometry information and determine chemical structures at the level of expertise seen in chemists and use microbiologist-level experience to predict microbial metabolites. Natural product science breakthroughs require the creation of unified knowledge representations from diverse data sources which serve as an educational basis for modeling complex relations in the field. Current AI applications focus on dataset learning but do not leverage AI's potential to generate hypotheses along with guiding experimental designs in research practice[31-34].

3.2 Natural Product Extraction, Targeting, and Isolation

The discovery process of natural products traditionally involved manual testing of different extraction methods and culture approaches to maximize output. The initial steps involve recognizing biological habitats to obtain living organisms before they may be cultured in laboratory environments using co-culture methods to study metabolite generation under simulated natural conditions. The trial of laboratory conditions which match regional environments proves difficult because natural chemical outputs change in correspondence with life stages and environmental elements. The actual chemical requirements for developing targeted compounds face challenges because of complex biological connections that prevent total replication. Extracting natural products faces a major challenge because different scientists utilize unstandardized extraction methods which create sizeable variations throughout the extraction process. This variability involves solvent choice and pH conditions as well as choice of grinding techniques plus sonication intensity along with extraction temperature and duration. The disparities between these methods negatively influence metabolite stability which may lead to partial extraction together with false experimental results. Following extraction natural substances require personnel identification before purification for establishing their biological activity *in vitro*. Precision is essential during research since scientists need to distinguish unreported bioactive compounds from compounds with similar properties. The bioactivity-based method has seen improvement through the integration of data-based methods utilizing metabolomics and genomics platforms which accelerate compound dereplication and selection of promising investigations.

A knowledge graph focused on natural product science that includes biosynthetic gene clusters and their host organisms as well as sampling locations and key environmental factors already enables community diversity analysis to discover unfamiliar ecological niches. Future AI models should use experimental designs from synthetic community research to predict phenotypes which will help scientists understand unobserved interactions and results. Knowledge graphs enable researchers to detect rare microbial communities with distinctive morphologies and activities that constitute optimal opportunities for discovering previously unknown biochemical compounds including bioactive molecular structures and new biosynthetic mechanisms. The current strength of newly developed knowledge graphs requires a wider

integration of natural product science data to achieve optimal artificial intelligence capabilities in natural product research and novel bioactive compound prediction. Natural product science benefits highly from multi-criteria decision aiding because this approach bases problem modeling on expert insights instead of following data-driven methods. The decision-making process relies on expert preferences together with experience alongside data inputs for resolving conflicts between criteria. These established mathematical models demonstrate interpretability and transparency and deliver explainable results which make them well-suited for complex decisions including natural product selection and structural elucidation of their functions. Expert-driven frameworks assist decision-making processes of multifactorial tasks by offering systematic procedures to handle selection conflicts between various factors. This decision theory-based strategy developed for natural product chemistry now receives improvements from recent machine learning models designed to model human decision behaviors. AI improves natural product targeting through analysis of decision success rates and historical data recorded in large objective datasets particularly knowledge graphs that record chemist's selection processes[35-40].

3.3 Structure Elucidation and Biological Assessment of Natural Products

A compound isolation success requires the determination of its chemical structure followed by biological activity assessment through various assays. A complete structural characterization needs the joint evaluation of NMR in combination with HRMS and MS/MS and IR spectrometry and VCD or ECD for understanding the compound's complete framework. To build a unified structural hypothesis these spectroscopic and spectrometric data points need meticulous assessment as an integrated solution. When genetic sequence data exists structural predictions of natural products occur directly from sequencing information. This provides scientists with an additional strategy during structural analysis. Ultimately, using a knowledge graph-driven AI model could enable the development of data-agnostic frameworks capable of ingesting diverse analytical readouts from sample extracts and suggesting optimal analysis pathways for further research[41].

The activity testing phase requires two approaches - the 1st tests protein-binding abilities through assays and the second conducts both phenotype and function screens to measure biological outcomes. The essential quality of antibiotics in their testing phase is identifying whether they work against various microorganisms or only specific pathogens. The assessment of resistance development plays an essential role in determining how much time a compound will retain its therapeutic value. Human pharmacology favors compounds with targeted action modes since such treatments deliver better safety outcomes to patients through reduced side effects and microbiota disruption potential. In drug development, a common strategy involves designing or modifying derivatives of existing molecular scaffolds, often prioritizing compounds that incorporate known warheads or pharmacophores. Alternatively, compounds can be designed entirely de novo. In both cases, understanding Quantitative Structure-Activity Relationships (QSARs) is essential. Advanced machine learning models attempt to infer these relationships by analyzing known connections between compounds, phenotypes, targets, and functional bioactivities. However, a persistent challenge is the lack of data on missing links- specifically, conclusive evidence of which phenotypes, targets, and bioactivities are not associated with a given compound. This issue is further complicated when dealing with frequent hitters- compounds that show activity across multiple screening assays, making it difficult to discern their true specificity. These gaps in data pose significant challenges for developing generative models and multi-label classifiers, as not all relationships are explicitly known, requiring AI-driven approaches to account for uncertainty and incomplete annotations[42-45].

3.4 Biosynthesis, Optimization, Characterization, and production of Natural Products by Using AI

Research advances compound synthesis after scientists verify which substance causes specific effects and establish its structure and targeting properties and activity levels. The traditional organic synthesis remains one of multiple available synthesis approaches. Research-based biological approaches like heterologous expression along with cell-free biosynthesis now serve as efficient and sustainable alternatives to produce complex natural

products. Scientists possess multiple discreet approaches for optimizing molecule compounds throughout their development process. The iterative use of Computer-Assisted Structure Prediction (CASP) leads to improved chemical synthesis because it aids compound design and optimization processes. Researchers combine algorithms and computational tools to study an extensive array of chemical compositions for discovering new potent molecular structures with valuable specific characteristics. Most widely used AI tools in natural product development shown in Table 2, Table 3 and Figure 3. The synthesis of drugs becomes faster along with reduced resource use through cost-efficient proposals from CASP models. CASP technology integrates natural product science knowledge graphs to create complicated structural solutions including genetics which human experts find difficult to manage. The current high research costs for heterologous BGC expression become significantly more expensive when efforts focus on industrial production along with additional complexity. A knowledge graph improved by AI functions as an interactive system where researchers can obtain optimal growth settings by supplying questions through either structured algorithmic recommendations or AI-driven chat interfaces. Synthesis outcomes undergo detection and analysis against the knowledge graph to find discrepancies and abnormal findings in experimental results. The integration of AI models with knowledge graph systems should develop beyond basic forecasting abilities to perform reasoning-based predictions in situations without direct training data. A breakthrough in natural product discovery would become possible through this technology which enables more efficient scalable synthesis adapted to emerging scientific requirements[45-50]. The discovery of drugs from natural products receives significant expansion through AI-based databases including IMPART and CoCoNut and traditional Chinese medicine platforms. These databases evaluate both the drug-likeness characteristics alongside ADMET properties of NPs for identifying top drug candidate prospects. The predictive models of machine learning algorithms on IMPART platforms show researchers data about natural compound profiles regarding drug absorption and toxicity attributes including solubility and permeability measurements together with bioavailability levels and toxicological assessment. The CoCoNut system enables researchers to understand NP chemical properties and biological target interactions through computational analyses of molecular interactions. TCM-focused databases establish a different methodology to systematically organize plant-based compounds and their linked bioactivities while leveraging AI for predicting drug discovery potential with modern pharmaceutical applications. Artificial intelligence platforms serve essential capabilities for NPs complexity management since these compounds display layout intricacies and multiple target interactions. As a prediction tool IMPART supports scientists to select drug candidates from NPs by assessing their pharmacokinetic properties thus minimizing experimental testing requirements. CoCoNut helps scientists study complex NPs through its ability to apply molecular fingerprinting and graph-based methods which simplify the analysis of chemical diversity in NPs leading to discovery of hidden bioactive compounds. Traditional Chinese medicine databases merge traditional herbal wisdom with contemporary computational resources to perform rapid screening tests on compound mixtures that generate predictions about multi-target engagements. AI plays a crucial role in identifying synergistic relationships between TCM compounds since many compounds in TCM operate together as single units. The potential AI-driven platforms struggle to resolve all intricate aspects which characterize NP drug discovery research. The shortage of sufficient and high-quality data limits the use of AI models in research applications. Model accuracy on IMPART and CoCoNut relies on the amount of available and diverse non-protein data. The prediction capabilities of AI models remain strong for pharmacological properties but the data should incorporate experimental laboratory testing for complex NPs. The complex biological mechanisms of NPs prove difficult to simulate using AI methods because their multi-target interactions are not easily discernible. Consequently, this limits the precision of safety and effectiveness predictions for NPs[1,57].

The deployment of artificial intelligence in natural product assessment has led to major progress through the application of complex computational models together with machine learning methods to study these compounds. The AI platforms develop advanced understanding of natural compounds molecular structures as well as biological activities and therapeutic potential which minimizes the requirement of prolonged laboratory work. Large datasets help AI systems find connective patterns

that demonstrate how natural products combine with biological systems since this reveals essential information about their action mechanics. Scientific databases become accessible through AI which enables researchers to find new compounds that show previously unreachable bioactivities. Through its technology optimization capabilities, the synthesis of natural products becomes simpler which helps drug manufacturers accelerate their commercial activities. Multi-omics data integration through AI establishes relationships between natural substance chemical properties and biological outcomes to present holistic understandings of their therapeutic value. AI platforms deliver improved personalized medicine by using predictive technologies to determine how human bodies will respond to natural products thus leading to individualized effective treatments. AI applications for natural product examinations facilitate drug discovery by enabling more effective and secure natural-based therapeutic development[1,9,57,74].

4. Unlocking AI-driven Breakthroughs in Natural Product Science

As part of this vision for the future of natural product science, AI and linked open data is offered as a means for accelerating discoveries. A natural product tricorder- a device capable of instantaneously identifying an organism's taxonomic, genomic, proteomic, metabolomic, and functional data by scanning or sampling - could be developed from a fully integrated knowledge graph. This would enable real time insights for fast cross reference as well as the prediction of novel discoveries. A true artificial intelligence revolution in natural product research is beyond merely the improvement of the classifier for bioactivity, or of the genome mining for the biosynthetic gene cluster. To harness AI models that allow scientists to take more informed hypotheses into unknown natural products, suggest additional hypothesis forming research that will increase success with limited resources. This is done by initiatives like the ENPKG and LOTUS initiatives, along with Wikidata, as all data should be structured in the format of linked open data to make it accessible and enable a global collaboration[55]. There are two key steps of transitioning to open linked natural product science data. The first is entity recognition, requiring that compounds, biological sources, and bioactivities inside the datasets are identified and linked. These entities must be mapped with relationships between them such as connecting a compound to its source organism or its biological activity as well as attributes like molecular weight and its functional properties.

Table 2. Important AI tools useful in natural product development and evaluation

Tools	Use
Wikidata	Central repository for linked natural product data and metadata.
Drug Bank	Database of FDA-approved drugs, chemical structures, drug interactions, and pharmacological data.
Natural products atlas	Integrates data from various natural product databases, bioactivity data, and chemical structures.
ChemSpider	Identification of natural products, chemical property analysis, and predicting biological activity
Super Natural 2	Prediction of pharmacological potential and discovery of bioactive compounds from natural sources.
KNAPSACK	Bioactivity analysis and medicinal evaluation of natural products
Deepchem	Machine learning framework for drug discovery and material sciences
MetaCyc	Understanding the systemic biological effects and mechanisms of natural products within metabolic pathways
Tox21	Machine learning algorithms to predict toxicity, safety, and biological effects of compounds.
STITCH	Chemical-protein interaction database, interaction networks, data from chemical, biological, and computational research.
AutoDock	Predicts molecular docking of natural compounds with biological targets
MolGAN	AI tool for de novo molecular design of natural product analogs.
KNIME Analytics	Integrates cheminformatics and bioinformatics workflows.
RDKit	Toolkit for cheminformatics and molecular modeling.

Table 3. AI and computational tools for natural product discovery

Name of Tool	Key Features	Primary Functions	Applications for Natural Product Research
MZmine	Open-source platform for MS (Mass Spectrometry) data processing.	Graphical interaction, peak detection, alignment, and feature extraction.	Used in untargeted metabolomics for natural product characterization.
AntDAS-DDA	Automates processing of UHPLC-HRMS data in DDA mode.	Detects and annotates molecular features from high-resolution MS data.	Enhances identification and quantification of natural products.
MetEx	Automated annotation of LC-MS-based untargeted data.	Identifies unknown compounds and extracts spectral features.	Speeds up metabolite profiling in complex mixtures of natural compounds.
MS-DIAL	Open-source software for DIA-based metabolomics.	Identifies and quantifies natural products in complex samples.	Used in secondary metabolite research and drug discovery.
XCMS	Feature detection algorithm for LC-MS data.	Detects close and overlapping features, retention time correction.	Helps in comparative metabolomics and natural product differentiation.
CAMERA	Post-processing tool for XCMS output.	Groups related features based on isotopic and adduct information.	Helps in compound annotation and metabolite classification.
MetaboScape	All-in-one software for non-targeted workflows in NP identification.	Data visualization, spectral library searching, and annotation.	Used in high-throughput natural product discovery.
MetAlign	Aligns up to one thousand datasets and corrects mass spectral artifacts.	Performs baseline corrections, peak-picking, and mass alignment.	Improves accuracy in comparative analysis of natural product samples.
OpenMS	Open-source platform for reproducible MS data analysis.	Allows flexible processing of high-throughput data.	Applied in metabolomics and proteomics studies of natural products.
W4M	Workflow for Metabolomics—enables reproducible computational analysis of raw MS data.	Processes and analyze large-scale metabolomics datasets.	Supports high-throughput screening in natural product discovery.
MN (Molecular Networking)	Graph-based analysis of molecular similarity.	Identifies related chemical structures in MS datasets.	Helps discover new natural products and their derivatives.
Lindenane Extractor	Extracts characteristic MS/MS ions of lindenane sesquiterpenes.	Builds molecular networks for targeted natural product discovery.	Used for sesquiterpene-based drug discovery.
Single-node-based MN	Filters and optimizes molecular network single nodes.	Connects promising molecules with similar scaffolds.	Helps in dereplication and identification of novel compounds.
IIMN	Integrates chromatographic peak shape correlation into MN.	Connects with different ion species of the same molecule.	Improves annotation accuracy of natural products.
Building Block Extractor	Targets NP discovery based on specified MS/MS features.	Extracts meaningful structural elements from MS spectra.	Enhances structural elucidation of natural compounds.
MetGem	Combines GNPS-style MN and t-SNE clustering for compound connectivity.	Addresses non-connected compounds in MN analysis.	Improves classification and identification of natural product families.
MN with GC-MS data	Integrates GC-MS with MN for volatile/derivatized molecules.	Allows identification of small volatile compounds.	Applied in essential oil and volatile secondary metabolite studies.
NAP	Improves in silico annotation using MN topology and structural similarity.	Enhances accuracy of molecular candidate ranking.	Helps in dereplication and classification of natural compounds.

Table 3. Continued...

Name of Tool	Key Features	Primary Functions	Applications for Natural Product Research
MS2LDA	Identifies co-occurring mass spectral fragment patterns (Mass2Motifs).	Recognizes common molecular substructures.	Used in fragment-based natural product discovery.
MolNetEnhancer	Combines multiple annotation tools for a comprehensive overview.	Integrates MS2LDA, NAP, and ClassyFire for better compound classification.	Enables multi-layered analysis of natural product libraries.
SNAP-MS	Annotates molecular subnetworks from in-house and reference spectra.	Predicts compound families and corrects annotation errors.	Supports large-scale annotation of complex natural products.
MetWork	Predicts NP structures based on molecular frameworks.	Identifies compound families from unknown spectra.	Helps in reconstructing biosynthetic pathways.
ConCISE	Establishes hypothetical classification combining MN, spectral matching, and AI-based predictions.	Enhances molecular classification and subnetwork organization.	Provides accurate compound annotation for natural products.
ISDB	Combines MN with an in-silico MS/MS fragmentation database.	Enhances dereplication of known and novel compounds.	Used for efficient annotation in metabolomics studies.
Dereplicator	Identifies peptide, polyketide, terpenoid, alkaloid, flavonoid, and benzenoid classes.	Matches unknown spectra to known chemical databases.	Helps rapidly identify known natural products.
DFMN-ISD	Enhances annotation of natural products with similar chemical scaffolds.	Analyzes structural similarities in molecular networks.	Increases accuracy in structural prediction.
VarQuest	Identifies variants in peptide-based natural product datasets.	Expands the known diversity of peptidic natural products.	Helps in discovering novel bioactive peptides.
SIRIUS	AI-powered molecular structure identification tool.	Predicts molecular formulas and fragmentation patterns.	Used for chemical structure elucidation of natural products.
CSI:FingerID	Searches molecular structure databases using tandem MS data.	Matches MS/MS spectra with known chemical structures.	Supports high accuracy dereplication in NP research.
MESSAR	AI-based substructure recommendation system.	Identifies meaningful molecular fragments.	Enhances structural prediction for novel natural products.
RiPPquest	Identifies ribosomally synthesized and post-translationally modified peptides (RiPPs).	Matches MS spectra to microbial RiPP databases.	Used in identifying RiPP-based antibiotics and bioactive peptides.
ADAPTIVE	Generates representations of compound-specific spectrum-structure pairs.	Enhances chemical annotation through adaptive learning.	Used for metabolomics and drug discovery research.
CFM-ID	Web-based AI tool for automated NP identification from MS/MS data.	Predicts fragmentation patterns and molecular structures.	Assists in high throughput dereplication of natural products.

However, for batch effects, metadata such as experimental methods and details of instrumentation, should also be included (for instance, follow the Paired Omics Data Platform model). The second step is to give these next entities Uniform Resource Identifiers (URIs) in accordance with Resource Description Framework (RDF) standards to get consistent reference to the datasets. To be readable for machines and interoperable, this data should be serialized into structured formats such as RDF or XML and published using HTTP URIs to be accessible to anyone. By integrating entities into related datasets, they are

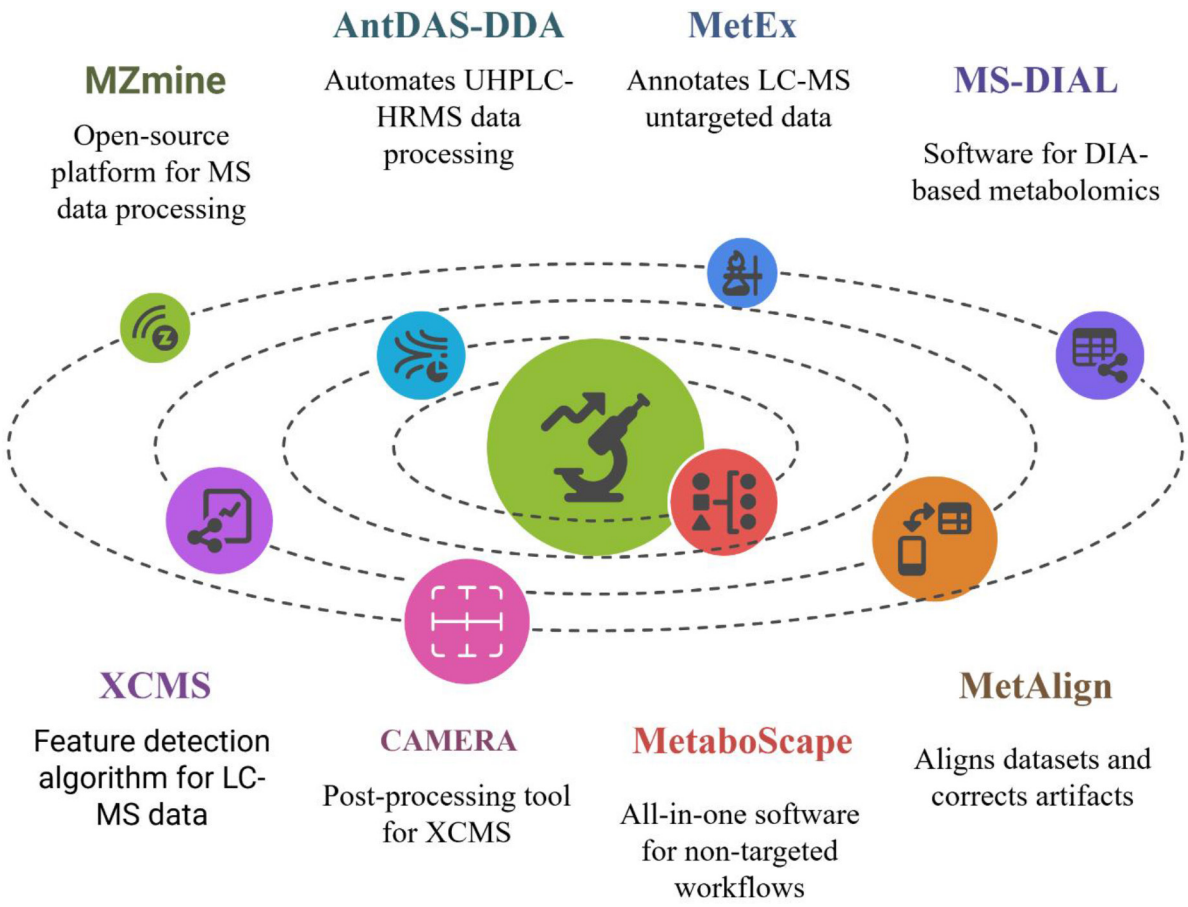


Figure 3. Various tools for natural research products.

linked with other datasets into the bigger linked ecosystem for AI based insights and smooth data sharing.This is a major challenge in natural product science, where datasets are fragmented and not standardized, for instance for bioactivity screening or for relationships between chemical structures, BGCs, mass spectrometry spectra, and uncharted organisms. Wikidata can be used as a centralized, open access repository of as much data as possible, where a proposed solution to this problem is to contribute as much data as possible to Wikidata. Along with that, knowledge graphs can be created that are specialized to the domain and that enable researchers to structure the domain specific knowledge, while integrating older knowledge graphs to preserve historic insights.The final desired outcome is documentation of the raw data as well as annotations, interpretations, and reviews, with full disclosure and machine readability. This will lead to novel discoveries, facilitate conceptual synthesis, and support the predictive power of AI in natural product science[50-55].

5. Collaboration between AI and Pharma Companies

Innovative collaboration between AI and pharma companies improves natural product research through faster drug discovery, improved formulation, and more efficient clinical trials[62]. In the last few decades, natural products have been a continuing source of therapeutic compounds, but their complex structures and biosynthetic pathways have made traditional research methods impractical. The problems it attempts to overcome with AI are fast analysis of large datasets, the prediction of molecular interactions and choosing promising candidates for drugs. Using a machine learning approach, AI can screen through vast natural product libraries, identify possible bioactives, and predict their modes of action, quickly and cheaply reducing the time and expense of the drug development process[56]. Molecular docking and target prediction is one of the most interesting AI applications in the field of

natural product research. Molecules hit by AI powered algorithms can also be modeled to interact with the biological targets more precisely while improving the precision of hit selection in the development stage of drug discovery. The capability of selecting novel drug candidates of high efficacy with minimal false positives enables pharmaceutical companies to identify possible novel drug candidates. AI also helps in optimizing drug formulations by making natural compounds have improved bioavailability, solubility, and stability. AI predicts the ideal excipients and delivery method to enhance drug formulation efficiency to develop better treatments. AI is also changing the way in which biosynthetic pathway engineering is carried out, by figuring out the Biosynthetic Gene Cluster (BGC) responsible for producing bioactive compounds. Advancement allows the use of these approaches to synthetic biology for sustainable and scalable drug production, and offloading reliance on rare natural sources. Another contribution of AI to drug safety is that by predicting toxicity of drugs before the implementation of clinical trials, the risk is minimized, and the safety of the patients is improved. Another area in which AI is making strides is personalized medicine, which uses natural product-based therapies targeted towards the genetic profiles of individuals. AI can predict how an individual would react to a certain natural compound and that in turn can make their treatments more effective and safer. The natural product research field is now being transformed by AI driven collaborations between pharmaceutical companies and biotech firms. Through these partnerships, we are tapping the potential of AI to facilitate drug discovery, formulation, safety assessment and clinical trials of innovative, effective, and sustainable natural product derived medicines. The future of drug development and the improvement of healthcare outcomes worldwide are coming from the development of pharmaceutical research as AI technology evolves[55-63]. Here, in Table 4 collaboration research between pharma companies and AI data was given with its applications.

Table 4. Collaborative research between AI and pharma companies

Collaboration Area	Role of AI	Pharmaceutical Applications	Examples of AI-Pharma Partnership
Drug discovery and development	AI analyzes natural product databases to identify new bioactive compounds.	Accelerates the identification of lead compounds from natural sources.	Benevolent AI and AstraZeneca: AI-driven drug discovery for new therapies.
Molecular docking and target prediction	AI predicts interactions between natural molecules and drug targets.	Enhance precision in selecting potential drug candidates.	Schrodinger and Sanofi: AI-driven molecular modeling for natural product drugs.
Formulation optimization	AI suggests optimal excipients and delivery methods for natural compounds.	Improves bioavailability, stability, and solubility of natural drugs.	*In-silico* medicine and Pfizer: AI-driven drug formulation research.
Biosynthetic pathway engineering	AI deciphers Biosynthetic Gene Clusters (BGCs) for producing bioactive compounds.	Enables synthetic biology approaches for sustainable drug production.	Zymergen and Ginkgo bioworks: AI-powered biosynthesis of pharmaceuticals.
Toxicity and safety prediction	AI predicts potential toxicity and side effects of natural compounds.	Reduces failure rates in clinical trials by early risk assessment.	Exscientia and Bayer: AI-driven toxicity screening for natural product drugs.
Personalized medicine	AI tailor's natural product-based therapies to individual genetic profiles.	Improves patient-specific treatment efficacy and safety.	IBM Watson and GSK: AI-driven personalized drug discovery.
Supply chain optimization	AI optimizes sourcing, extraction, and production of natural compounds.	Ensures cost-effective and sustainable pharmaceutical manufacturing.	Recursion Pharmaceuticals and Roche: AI-powered supply chain management.
Clinical trial design and analysis	AI streamlines patient selection and trial data analysis.	Accelerates drug approval processes and reduces costs.	DeepMind and Novartis: AI-driven clinical trial simulations.
Regulatory compliance and documentation	AI automates data structuring for regulatory submissions.	Ensures compliance with FDA or EMA guidelines for natural product-based drugs.	BioXcel and Eli Lilly: AI-assisted regulatory compliance for drug approvals.

6. Challenges of AI in Evaluation of Natural Products

AI contributes substantially to NP drug discovery even though multiple research obstacles remain because of NP-specific complexities. Data scarcity and complex NP structures alongside compound synthesis limitations and biological interaction complexities and the necessity of transparent models cause problems in the application of artificial intelligence for natural product drug discovery[64]. Lack of data availability represents a major difficulty in the process. The inadequate content in NP database information about chemical structures and biological activities as well as pharmacological properties creates problems for building precise models[5,65]. Advanced techniques including transfer learning along with active learning and data synthesis help reduce this problem yet a solution dedicated to addressing the data shortage in NP drug discovery remains important. The difficulty of analyzing NP structures increases due to their complex nature. The complex structures of NPs contain many stereocenters in addition to multiple functional groups along with diverse isomers[65]. The prediction of bioactivity together with toxicity levels and additional properties becomes challenging because of these conditions. The complexity of NP structures demands additional advanced methods including graph-based approaches to handle this complexity effectively. Compound optimization together with virtual screening techniques have not managed to overcome the major challenge of synthesizing intricate NP structures. The AI tool known as AiZynth Finder demonstrates effective synthetic pathway design capabilities for basic molecules although it encounters challenges with complex structures featuring multiple ring systems and chiral centers[66]. The development of AI-driven synthesis methods needs additional improvements, especially to address diverse chemical structures with complex structures. Predictable chemical synthesis tools are expected to undergo improvement through quantum computing as well as cheminformatics techniques. The high biological complexity of NPs makes analyzing them difficult[67]. The capability of NPs to engage multiple biological targets and mechanisms makes present prediction models ineffective at determining both their safety and efficacy accurately. The accuracy of predictions requires both reliable data and biological process understanding to enhance future prediction outcomes in this field. The main drawback of AI models is their inability to provide clear explanations because they lack interpretability. Researchers struggle to determine the process behind prediction methods which reduces their confidence in the outcomes. The process of making AI models more understandable includes developing methods based on saliency maps as well as feature attribution and attention mechanisms. The reliable and transparent operation of these models requires additional partnership between AI experts and domain specialists. Advancing the field of NP drug discovery through artificial intelligence requires the resolution of data accessibility issues and structure simplification as well as biological intricacies and synthesis challenges and model interpretation limitations[68-75].

7. Conclusion and Future Prospective

By integrating Artificial Intelligence and knowledge graphs, natural product science can now operate more efficiently, with the management of data being more effective, generating more hypotheses, and discovering more drugs. Rather, AI models create structures out of various datasets such as genomic, metabolomic, spectroscopic, and biochemical data to find patterns that would otherwise be masked. Initiatives like ENPKG, LOTUS, Wikidata enable open data implementation, making the data accessible, interoperable, and globally collaborating. For instance, a fully integrated graphical modeling could support the development of a natural product tricorder, a tool that could rapidly identify an organism's taxonomic, genomic, and metabolomic profiles to speed up and conjure discovery processes. This goes beyond conventional classifiers for bioactivity or genome mining and delivers predictive frameworks that potentially lead researchers to new discoveries in resource-limited environments. The architecting of this AI-driven revolution requires entity recognition, relationship mapping, and the adoption of URIs to standardize entities and entities for the process of accelerated unified data integration. However, AI can leverage predictive modeling cohesively to solve data fragmentation, lack of standardization and more to improve natural product research, maximize novel

drug formulation, and elucidate structure. AI and knowledge graphs have enormous potential in the research area that can radically transform the discovery into application. It also facilitates a meaningful constructive collaboration of various scientific tools that can advance drug development, but it cannot solve it alone. AI driven expansion of therapeutic innovation in natural product science will be enhanced by the novel direct access data structure.

7. Author Contribution

PD supervision wrote and edited final draft, KP and BR collected the data and wrote, AG prepared the manuscript, figures, and tables. All authors have read and agreed to the published version of the manuscript.

8. References

1. Willighagen E. Artificial intelligence for natural product drug discovery. Nature Reviews Drug Discovery. 2023; 22(11):895-916. https://doi.org/10.59350/dtyms-yt012
2. Ozturk H, Ozgur A, Schwaller P, Laino T, Ozkirimli E. Exploring chemical space using natural language processing methodologies for drug discovery. Drug Discovery Today. 2020; 25(4):689-705. https://doi.org/10.1016/j.drudis.2020.01.020
3. Sadybekov AV, Katritch V. Computational approaches streamlining drug discovery. Nature. 2023; 616(7958):673-685.
4. Klucznik BM, Golebiowska P, Bayly AA, Popik O, Klucznik T, Szymkuc S, Gajewska EP, Dittwald P, Krajewska OS, Beker W, Badowski T. Computational planning of the synthesis of complex natural products. Nature. 2020; 588(7836):83-88.
5. Arora S, Chettri S, Percha V, Kumar D, Latwal M. Artifical intelligence: A virtual chemist for natural product drug discovery. J Biomol Struct Dyn. 2024; 42(7):3826-3835. PMID: 37232451. https://doi.org/10.1080/07391102.2023.2216295
6. Levin I, Liu M, Voigt CA, Coley CW. Merging enzymatic and synthetic chemistry with computational synthesis planning. Nat Commun. 2022; 13(1):7747.
7. Michoel T, Zhang JD. Causal inference in drug discovery and development. Drug Discov Today. 2023; 28(10):103737. https://doi.org/10.1016/j.drudis.2023.103737
8. Hogan A, Blomqvist E, Cochez M, D'Amato C, Melo GD, Gutierrez C, Kirrane S, Gayo JEL, Navigli R, Neumaier S. Knowledge graphs. ACM Computing Surveys. 2021; 54(4):1-37. https://doi.org/10.1145/3447772
9. 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
10. Yang JS, Tsai SC, Hsu YM, Bau DT, Tsai CW, Chang WS, Kuo SC, Yu CC, Chiu YJ, Tsai FJ. Integrating natural product research laboratory with artificial intelligence: Advancements and breakthroughs in traditional medicine. Biomedicine. 2024; 14(4):1-14. PMID: 39777110. PMCID: PMC11703400. https://doi.org/10.37796/2211-8039.1475
11. Wang Z, Yang L. Natural-product-based, carrier-free, noncovalent nanoparticles for tumor chemo-photodynamic combination therapy. Pharmacol Res. 2024; 203:107150. PMID: 38521285. https://doi.org/10.1016/j.phrs.2024.107150
12. Sun Y, Sun H, Zhang Z, Tan F, Qu Y, Lei X, Xu Q, Wang J, Shu L, Xiao H, Yang Z, Liu H. New insight into oxidative stress and inflammatory responses to kidney stones: Potential therapeutic strategies with natural active ingredients. Biomed Pharmacother. 2024; 179:117333. https://doi.org/10.1016/j.biopha.2024.117333
13. Wei D, Peslherbe GH, Selvaraj G, Wang Y. Advances in drug design and development for human therapeutics using artificial intelligence- I. Biomolecules. 2022; 12(12):1846. PMID: 36551273. PMCID: PMC9775020. https://doi.org/10.3390/biom12121846
14. Xu M, Chen Z, Zheng J, Zhao Q, Yuan Z. Artificial intelligence-aided optical imaging for cancer theranostics. Semin Cancer Biol. 2023; 94:62-80. PMID: 37302519. https://doi.org/10.1016/j.semcancer.2023.06.003
15. Rutz A, Sorokina M, Galgonek J, Mietchen D, Willighagen E, *et al.* The LOTUS initiative for open natural products research: Knowledge management through Wikidata. eLife. 2021. https://doi.org/10.7554/eLife.70780
16. Meijer D, Beniddir MA, Coley CW, Mejri YM, Ozturk M, Hooft JJV, Medema MH, Skiredj A. Empowering natural product science with AI: Leveraging multimodal data and knowledge graphs. Nat Prod Rep. 2025; 42(4):654-662. PMID: 39148455. PMCID: PMC11327853. https://doi.org/10.1039/d4np00008k
17. Daza D, Alivanistos D, Mitra P, Pijnenburg T, Cochez M, Groth P. BioBLP: A modular framework for learning on multimodal biomedical knowledge graphs. J Biomed Semant. 2023; 14(1):20.
18. Delmas M, Filangi O, Paulhe N, Vinson F, Duperier C, Garrier W, Saunier PE, Pitarch Y, Jourdan F, Giacomoni F, Frainay C. FORUM: Building a knowledge graph from public databases and scientific literature to extract associations between chemicals and diseases. Bioinformatics. 2021; 37(21):3896-3904. https://doi.org/10.1093/bioinformatics/btab627
19. Gaudry A, Pagni M, Mehl F, Moretti S, Guerrero LMQ, Cappelletti L, *et al.* A sample-centric and knowledge-driven computational framework for natural products drug

discovery. ACS Cent Sci. 2024; 10(3):494-510. https://doi.org/10.1021/acscentsci.3c00800

20. Rutz A, Sorokina M, Galgonek J, Mietchen D, Willighagen E, *et al*. The LOTUS initiative for open knowledge management in natural products research. Elife. 2022; 11:e70780. PMID: 35616633. PMCID: PMC9135406. https://doi.org/10.7554/eLife.70780
21. Meng X, Fang Y, Ding M, Zhang Y, Jia K, Li Z, Collemare J, Liu W. Developing fungal heterologous expression platforms to explore and improve the production of natural products from fungal biodiversity. Biotechnol Adv. 2022; 54:107866. PMID: 34780934. https://doi.org/10.1016/j.biotechadv.2021.107866
22. Takino J, Kotani A, Ozaki T, Peng W, Yu J, Guo Y, Mochizuki S, Akimitsu K, Hashimoto M, Ye T, Minami A. Biochemistry-guided prediction of the absolute configuration of fungal reduced polyketides. Angew Chem. 2021; 60(43):23403-23411. https://doi.org/10.1002/anie.202110658
23. Jacobsen A, Azevedo RM, Juty N, Batista D, Coles S, *et al*. FAIR Principles: Interpretations and Implementation Considerations. Data Intelligence. 2020; 2(1-2):10-29. https://doi.org/10.1162/dint_r_00024
24. Liu Y, Li H, Duran AG, Niepert M, Rubio DO, Rosenblum DS. MMKG: multi-modal knowledge graphs. 16th International conference. ESWC. 2019. https://doi.org/10.48550/arXiv.1903.05485
25. Wu Y, Wu X, Li J, Zhang Y, Wang H, Du W, He Z, Liu J, Ruan T. Mmpedia: A large-scale multi-modal knowledge graph. The Semantic Web - ISWC. LNCS. 2023; 14266:18-37.
26. Auer S, Kovtun V, Prinz M, Kasprzik A, Stocker M, Vidal ME. Towards a knowledge graph for science. Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics. 2018: pp. 1-6. https://doi.org/10.1145/3227609.3227689
27. Silva MC, Eugenio P, Faria D, Pesquita C. Ontologies and knowledge graphs in oncology research. Cancers. 2022; 14(8):1906. PMCID: PMC9029532. PMID: 35454813. https://doi.org/10.3390/cancers14081906
28. Taneja SB, Callahan TJ, Paine MF, Gill SLK, Kilicoglu H, Joachimiak MP, Boyce RD. Developing a knowledge graph for pharmacokinetic natural product-drug interactions. J Biomed Inform. 2023; 140:104341. PMID: 36933632. PMCID: PMC10150409. https://doi.org/10.1016/j.jbi.2023.104341
29. Peterson JC, Bourgin DD, Agrawal M, Reichman D, Griffiths TL. Using large-scale experiments and machine learning to discover theories of human decision-making. Science. 2021; 372(6547):1209-1214. https://doi.org/10.1126/science.abe2629
30. Tietz JI, Mitchell A. Using genomics for natural product structure elucidation. Curr Top Med Chem. 2016; 16(15):1645-1694. PMID: 26456468. https://doi.org/10.2174/1568026616666151012111439
31. Yuan Y, Shi C, Zhao H. Machine learning-enabled genome mining and bioactivity prediction of natural products. ACS synthetic biology. 2023 Aug 22;12(9):2650-62.
32. Yang JS, Chiang JH, Tsai SC, Hsu YM, Bau DT, Lee KH, Tsai FJ. *In silico* de novo curcuminoid derivatives from the compound library of natural products research laboratories inhibit COVID-19 $3CL^{pro}$ activity. Nat Prod Commun. 2020; 15(9). https://doi.org/10.1177/1934578X20953262
33. Wu SY, Pan SL, Xiao ZY, Hsu JL, Chen MC, Lee KH, Teng CM. NPRL-Z-1, as a new topoisomerase II poison, induces cell apoptosis and ROS generation in human renal carcinoma cells. Plos One. 2014; 9(11):e112220. https://doi.org/10.1371/journal.pone.0112220
34. Liu SQ, Shen BB, Li HY, Yao YX, Li B, Yu HH, Tian X, Zhou XD, *et al*. Integrating UPLC-Q-exactive orbitrap/MS, network pharmacology and experimental validation to reveal the potential mechanism of *Kadsura coccinea* roots in colon cancer. J Ethnopharmacol. 2025; 337(Part 3):118934. PMID: 39401665. https://doi.org/10.1016/j.jep.2024.118934
35. Dong SH, Duan ZK, Bai M, Huang XX, Song SJ. Advanced technologies targeting isolation and characterization of natural products. TrAC Trends Anal Chem. 2024; 175:117711. https://doi.org/10.1016/j.trac.2024.117711
36. Peterson JC, Bourgin DD, Agrawal M, Reichman D, Griffiths TL. Using large-scale experiments and machine learning to discover theories of human decision-making. Science. 2021; 372(6547):1209-1214. https://doi.org/10.1126/science.abe2629
37. Ai CB, Li LN. Stereostructure of salvianolic acid B and isolation of salvianolic acid C from *Salvia miltiorrhiza*. J Nat Prod. 1988; 51(1):145-149. https://doi.org/10.1021/np50055a023
38. Sahayasheela VJ, Lankadasari MB, Dan VM, Dastager SG, Pandian GN, Sugiyama H. Artificial intelligence in microbial natural product drug discovery: Current and emerging role. Nat Prod Rep. 2022; 39(12):2215-2230. https://doi.org/10.1039/d2np00035k
39. Qian H, Xiao Z, Su L, Yang Y, Tian X, Wang X. Prediction of anti-rheumatoid arthritis natural products of *Xanthocerais lignum* based on LC-MS and artificial intelligence. Comb Chem High Throughput Screen. 2025; 28(4):627-646. PMID: 38299408. PMCID: PMC12174899. https://doi.org/10.2174/0113862073282138240116112348
40. Said RP, Gawad JB, Kalaskar M, Gurav N, Beldar VG. Isolation, fractionation, and purification of natural products. Pharmacognosy and Phytochemistry: Principles, Techniques, and Clinical Applications. Wiley. 2025: pp. 207-234.
41. Hu G, Qiu M. Machine learning-assisted structure annotation of natural products based on MS and NMR

data. Nat Prod Rep. 2023; 40(11):1735-1753. https://doi.org/10.1039/D3NP00025G

42. Yang R, Zhao G, Yan B. Discovery of novel c-jun N-terminal kinase 1 inhibitors from natural products: Integrating artificial intelligence with structure-based virtual screening and biological evaluation. Molecules. 2022; 27(19):6249. PMID: 36234788. PMCID: PMC9572546. https://doi.org/10.3390/molecules27196249
43. Ai S, Lin G, Bai Y, Liu X, Piao L. QSAR classification-based virtual screening followed by molecular docking identification of potential COX-2 inhibitors in a natural product library. J Comput Biol. 2019; 26(11):1296-1315. PMID: 31233340. https://doi.org/10.1089/cmb.2019.0142
44. Rajan K, Brinkhaus HO, Agea MI, Zielesny A, Steinbeck C. DECIMER. AI: An open platform for automated optical chemical structure identification, segmentation and recognition in scientific publications. Nat Commun. 2023; 14(1):5045.
45. Ni Y, Bai J, Zhang Y, Qiao H, Liang L, Wan J, Zhu Y, Cao H, Li H, Zhao Q. Exploring and designing potential inhibitors of SIRT2 in natural products by Artificial Intelligence (AI) and molecular dynamics methods. Letters Drug Des Discov. 2024; 21(16):3542-3554. https://doi.org/10.2174/0115701808288696240308052948
46. Aware CB, Patil DN, Suryawanshi SS, Mali PR, Rane MR, Gurav RG, Jadhav JP. Natural bioactive products as promising therapeutics: A review of natural product-based drug development. S Afr J Bot. 2022; 151(Part B):512-528. https://doi.org/10.1016/j.sajb.2022.05.028
47. Merk D, Grisoni F, Friedrich L, Schneider G. Tuning artificial intelligence on the de novo design of natural-product-inspired retinoid X receptor modulators. Commun Chem. 2018; 1(1):68.
48. El-Naggar NE, Bashir SI, Rabei NH, Saber WI. Innovative biosynthesis, artificial intelligence-based optimization, and characterization of chitosan nanoparticles by *Streptomyces microflavus* and their inhibitory potential against *Pectobacterium carotovorum*. Scientific Reports. 2022; 12(1):21851. PMID: 36528632. PMCID: PMC9759534. https://doi.org/10.1038/s41598-022-25726-w
49. Seshadri K, Abad AN, Nagasawa KK, Yost KM, Johnson CW, Dror MJ, Tang Y. Synthetic biology in natural product biosynthesis. Chem Rev. 2025; 125(7):3814-3931. https://doi.org/10.1021/acs.chemrev.4c00567
50. Li G, Lin P, Wang K, Gu CC, Kusari S. Artificial intelligence-guided discovery of anticancer lead compounds from plants and associated microorganisms. TiC. 2022; 8(1):65-80. https://doi.org/10.1016/j.trecan.2021.10.002
51. Medema MH, Kottmann R, Yilmaz P, Cummings M, Biggins JB, Blin K, Bruijn I, *et al*. Minimum information about a biosynthetic gene cluster. Nat Chem Biol. 2015; 11(9):625-631.
52. Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, *et al*. Sharing and community curation of mass spectrometry data with global natural products social molecular networking. Nat Biotechnol. 2016; 34(8):828-837.
53. Wolfender JL, Litaudon M, Touboul D, Queiroz EF. Innovative omics-based approaches for prioritisation and targeted isolation of natural products-new strategies for drug discovery. Nat Prod Rep. 2019; 36(6):855-868. PMID: 31073562. https://doi.org/10.1039/c9np00004f
54. Victoria IP. Natural products dereplication: Databases and analytical methods. Prog Chem Org Nat Prod. 2024; 124:1-56. PMID: 39101983. https://doi.org/10.1007/978-3-031-59567-7_1
55. Meijer D, Beniddir MA, Coley CW, Mejri YM, Ozturk M, Hooft JJVD, Medema MH, Skiredj A. Empowering natural product science with AI: Leveraging multimodal data and knowledge graphs. Nat Prod Rep. 2025; 42(4):654-662. PMID: 39148455. PMCID: PMC11327853. https://doi.org/10.1039/d4np00008k
56. Mullowney MW, Duncan KR, Elsayed SS, Garg N, Hooft JJVD, Martin NI, *et al*. Artificial intelligence for natural product drug discovery. Nat Review Drug Discov. 2023; 22(11):895-916.
57. Sampene AK, Nyirenda F. Evaluating the effect of artificial intelligence on pharmaceutical product and drug discovery in China. Future J Pharm Sci. 2024; 10(1):58.
58. Mirakhori F, Niazi SK. Harnessing the AI/ML in drug and biological products discovery and development: The regulatory perspective. Pharmaceuticals. 2025; 18(1):47. PMID: 39861110. PMCID: PMC11769376. https://doi.org/10.3390/ph18010047
59. Kapustina O, Burmakina P, Gubina N, Serov N, Vinogradov V. User-friendly and industry-integrated AI for medicinal chemists and pharmaceuticals. Artif Intell Chem. 2024; 2(2):100072. https://doi.org/10.1016/j.aichem.2024.100072
60. Gude S, Gude YS. The synergy of artificial intelligence and machine learning in revolutionizing pharmaceutical regulatory affairs. Transl Regul Sci. 2024; 6(2):37-45. https://doi.org/10.33611/trs.2024-005
61. Terranova N, Renard D, Shahin MH, Menon S, Cao Y, Hop CE, Hayes S, *et al*. Artificial intelligence for quantitative modeling in drug discovery and development: an innovation and quality consortium perspective on use cases and best practices. Clin Pharmacol Ther. 2024; 115(4):658-672. PMID: 37716910. https://doi.org/10.1002/cpt.3053
62. Najmi A, Javed SA, Al Bratty M, Alhazmi HA. Modern approaches in the discovery and development of plant-based natural products and their analogues as potential therapeutic agents. Molecules. 2022; 27(2):349. https://doi.org/10.3390/molecules27020349

63. Sheela PJ, Chelliah DA, David SJ, Jebaraj JW, *et al.* Legal landscape and a few case studies of natural products sales in India: challenges and resolutions. Nat Prod Res. 2025: pp. 1-7. PMID: 40063064. https://doi.org/10.1080/14786419.2025.2475363
64. Gangwal A, Ansari A, Ahmad I, Azad AK, Sulaiman WM. Current strategies to address data scarcity in artificial intelligence-based drug discovery: A comprehensive review. Comput Biol Med. 2024; 179:108734. https://doi.org/10.1016/j.compbiomed.2024.108734
65. Boldini D, Ballabio D, Consonni V, Todeschini R, Grisoni F, Sieber SA. Effectiveness of molecular fingerprints for exploring the chemical space of natural products. J Cheminf. 2024; 16(1):35.
66. Genheden S, Thakkar A, Chadimova V, Reymond JL, Engkvist O, Bjerrum E. AiZynthFinder: A fast, robust and flexible open-source software for retrosynthetic planning. J Cheminf. 2020; 12(1):70.
67. Meijer D, Beniddir MA, Coley CW, Mejri YM, Ozturk M, Hooft JJJV, Medema MH, Skiredj A. Empowering natural product science with AI: Leveraging multimodal data and knowledge graphs. Nat Prod Rep. 2025; 42(4):654-662. PMID: 39148455. PMCID: PMC11327853. https://doi.org/10.1039/d4np00008k
68. Mazuz E, Shtar G, Shapira B, Rokach L. Molecule generation using transformers and policy gradient reinforcement learning. Sci Rep. 2023; 13(1):8799.
69. Kadurin A, Nikolenko S, Khrabrov K, Aliper A, Zhavoronkov A. druGAN: An advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties *in silico*. Mol Pharmaceutics. 2017; 14(9):3098-3104. https://doi.org/10.1021/acs.molpharmaceut.7b00346
70. Gupta A, Muller AT, Huisman BJ, Fuchs JA, Schneider P, Schneider G. Generative recurrent networks for de novo drug design. Mol Inform. 2018; 37(1-2):1700111. PMID: 29095571. PMCID: PMC5836943. https://doi.org/10.1002/minf.201700111
71. Lavecchia A. Advancing drug discovery with deep attention neural networks. Drug Discov Today. 2024; 28(8):104067. https://doi.org/10.1016/j.drudis.2024.104067
72. Karpov P, Godin G, Tetko IV. Transformer-CNN: Fast and reliable tool for QSAR. Quantitative Methods. 2019. https://doi.org/10.48550/arXiv.1911.06603
73. Proietti M, Ragno A, Rosa BL, Ragno R, Capobianco R. Explainable AI in drug discovery: self-interpretable graph neural network for molecular property prediction using concept whitening. Machine Learning. 2024; 113(4):2013-2044.
74. Luna JJ, Grisoni F, Schneider G. Drug discovery with explainable artificial intelligence. Nat Mach Intell. 2020; 2(10):573-584.
75. Madival SD, Mishra DC, Chaturvedi KK, Sharma A, Budhlakoti N, Angadi UB, Farooqi MS, Srivastava S, Jha GK. NaturePred: A tool for revolutionizing natural product classification with artificial intelligence. Curr Prot. 2025; 21(5):429-436. https://doi.org/10.2174/0115701646322417241101055512