Artificial Intelligence in the 21st century: The treasure hunt for systematic mining of natural products

Janani Manochkumar[a] and Siva Ramamoorthy[a, *]

[a] School of Bio Sciences and Technology, Vellore Institute of Technology, Vellore 632014, India

Corresponding author

Siva Ramamoorthy,

School of Bio Sciences and Technology,

Vellore Institute of Technology,

Vellore 632014, Tamil Nadu, India

Email: siva.ramamoorthy@gmail.com

ORCID ID: 0000 – 0001-7509- 8602

Abstract

Advancements in genome mining, high-throughput sequencing, and experimental techniques have generated an enormous amount of data on natural products. This led to the design and development of advanced machine learning and artificial intelligence algorithms which simplified the hunt for novel natural product discovery in the 21st century. These algorithms could effectively analyze the chemical structure of natural products and predict their biological function. These algorithms could also effectively analyze large sets of data in a sophisticated manner. In this context, this manuscript reviews the various AI/ML algorithms employed in natural product-based drug discovery. Particular attention is paid to case studies employing AI tools in plant and microbial research. Challenges associated with the use of AI tools for natural product research have also been discussed.

Keywords

Artificial Intelligence, Dereplication, Drug Discovery, Machine Learning, Natural Products.

Significance:

The recent progress in the AI field led to the efficient mining of natural products. The existing and emerging AI/ML-based tools for effective screening of bioactive metabolites from plants and microbes were discussed. This article highlights the importance of AI algorithms in sophisticating the identification of natural products.

Abbreviations

ADME, Absorption, Distribution, Metabolism, and Excretion; AI, Artificial Intelligence; ANN, Artificial Neural Network; antiSMASH, antibiotics and Secondary Metabolites Analysis SHell; ARTS, Automated Resource Tracking System; BGCs, Biosynthetic Gene Clusters; BIG-SCAPE, Biosynthetic Gene Similarity Clustering and Prospecting Engine; BMRB, Biological Magnetic Resonance Data Bank; CMNPD, Comprehensive Marine Natural Products Database; CNN, Convolutional Neural Network; DNN, Deep Neural Network; DeepDTA, Deep Drug-Target binding Affinity prediction; DNP, Dictionary of Natural Products; DL, Deep Learning; ELINA, Eliciting Nature's Activities; GNPS, Global Natural Product Social Molecular Networking; HMDB, Human Metabolome Database; HMM, hidden Markov model; HRMS, High Resolution Mass Spectrometry; IMG/ABC, Integrated Microbial

Genomes; IMS, Imaging Mass Spectrometry; KronRLS, Kronecker Regularized Least Squares; LBVS, Ligand Based Virtual Screening; MALDI-TOF, Matrix-Assisted Laser Desorption/ionization Time-of-Flight mass; MetaBGC, Metagenomic identifier of Biosynthetic Gene Clusters; MetEx, Metabolomics Explorer; MIBiG, Minimum Information about a Biosynthetic Gene cluster; ML, Machine Learning; MN, Molecular Networking; NaPLeS, Natural Product-Likeness Software Suite and Database; NMR, Nuclear Magnetic Resonance; NPASS, Natural Product Activity and Species Source Database; NPCARE, Natural Products for Cancer Regulation; NP-MRD, Natural Products Magnetic Resonance Database; NPs, Natural Products; NuBBE DB, Nuclei of bioassays, ecophysiology and biosynthesis of Natural Products Database; PADME, Protein and drug molecule interaction Prediction; PDA, Photodiode Array; pHMMs, profile hidden Markov models; QSAR, Quantitative Structure-Activity Relationships; RF, Random Forest; SBVS, Structure Based Virtual Screening; SMART, Small Molecule Accurate Recognition Technology; SIMILE, Significant Interrelation of MS/MS Ions via Laplacian Embedding; SPiDER, Self-Organizing Map-Based Prediction of Drug Equivalence Relationship; SVM, Support Vector Machine; TCM, Traditional Chinese Medicine; UNaProd, Universal Natural Product Database.

1. Introduction

Artificial intelligence (AI) utilizes computers for performing complicated tasks, analyzing huge data files, and evaluating them based on advanced algorithms. It is well known that AI has a plethora of applications in various fields of research for controlling and processing tasks as it analyses effectively as well as interprets rapidly with minimized human faults and reveals complex data structures[1]. Recently, AI is also used by researchers for the identification of molecular characteristics, automatic processing, genome mining, dereplication, and prediction of targets and bioactivity. The fruitful advancements in machine learning (ML) and AI algorithms and information overload in databases and repositories have enabled researchers to gain free access to diverse data and utilize AI/ML techniques in the mining of natural products efficiently[2].

Natural products (NPs) have garnered proliferating attention in drug discovery as it is bio-friendly, less toxic, and evolve collaboratively along with their active sites[3,4], The high variation in the molecular structure and physicochemical properties of NPs makes them a treasured source of novel bioactive compounds with various applications in the agricultural, biotechnological, food, cosmetics, and pharmaceutical industries[5,6].

There are over 465,000 plant species existing on the earth of which 391,000 species are vascular plants[7]. One of the enthralling facts about plants is their unique metabolic pathway which corresponds to the synthesis of highly complex bioactive metabolites[8]. The diversity of plant metabolites is estimated to exceed 1 million with each plant contributing to more than 4.7 structurally unique compounds[9]. The use of plant extracts as a commercial product in food and flavor, cosmetic, and pharma industries has been predicted to reach USD 59.4 billion by 2025[10]. Plants have been alternatively used for the treatment of several diseases worldwide[11]. Based on this evidence, researchers are now focussing their investigations on plants and microbes'

potential to render natural products with beneficial therapeutic effects[8]. Over the last few decades, AI has been utilized in the screening of plant extracts, chemical taxonomy, chemical fingerprinting, phylogenetic studies, predicting toxic properties and determining the structure of phytochemicals based on the spectroscopic data[12].

In spite of the incomparable role of NPs in drug design and discovery, conventional techniques have several challenges like extraction, screening, purification, and structure elucidation from plant and microbial sources[13]. The repeated identification of already identified NPs, high demand for resources, increasing manual efforts, and time-consuming tasks have restrained the interest of scientists and industries in natural product research[14]. However, with the recent advancement in omic technologies including proteomics, genomics and metabolomics, it is now easy to retrieve enormous data regarding the biosynthetic pathway of secondary metabolites[15]. At present, omics-related tools and AI-based algorithms aid in the characterization, screening, and selection of chemical structures with desired bioactivity and physicochemical characteristics[16].

When compared to experimental techniques that only involve *in vitro* and *in vivo* testing, computational bioprospecting methodologies have been reported as effective, low-cost, low-labor, and less-time approaches[17]. In addition, some structural scaffolds derived from various classes of natural products, such as alkaloids, phenylpropanoids, polyketides, and terpenoids, have served as an inspiration to design new drug candidates[18]. The concept of AI in mining the various classes of plants and microbial secondary metabolites is illustrated in (Figure 1).

2. Role of computational methods in virtual screening of bioactive metabolites

Virtual screening strategies transformed the identification of novel bioactive metabolites by evaluating the *in-silico* large compound library aiding the exploration of their pharmacodynamics, pharmacokinetics and chemical space thus leading to less time, cost and

infrastructure involved in the discovery of novel metabolites[16]. Virtual screening strategies have immensely contributed to the identification of novel bioactive compounds by assessing the *in-silico* structural public libraries against relevant receptors through knowledge of AI and utilization of molecular models, and statistical and probability tools[16]. This has the added advantages of lessening cost, time, manual efforts, and infrastructure[19]. These techniques employ a series of consecutive and hierarchical procedures with the goal of separating out molecules with desirable physicochemical, pharmacodynamic, and Absorption, distribution, metabolism, and excretion (ADME) properties and rejecting those that do not meet the profile. The success of discovering novel bioactive compounds is increased when these techniques are integrated with experimental methodologies[20]. The virtual screening strategies will utilize both the computational techniques that aim to discover novel bioactive metabolite against a specific target[25]. These methods should examine the chemical space of natural products in order to identify the bioactive class of compounds and structural scaffolds of known compound. Some of these methods applies less restraining structural similarity cutoff and modelling of putatively derived structures of natural products[21]. The 3D structure depicts the configuration of structure and binding site of ligands. Therefore, virtual screening strategies have emerged to be an essential part of discovery of novel bioactive metabolites[16]. The overflow of the virtual screening strategy for identifying bioactive metabolites along with conventional computer aided discovery of natural products was depicted in (**Figure 2)**.

2.1. Ligand-based virtual screening (LBVS)

The LBVS approach uses a set of compounds with experimentally demonstrated bioactivity as a starting point and solely relies on the analysis of the inherent features of the compound's structure including physicochemical, electronic, structural, and topological characteristics that are related to its bioactivity[22]. Quantitative structure-activity relationship (QSAR), ML algorithms, ligand-based pharmacophore modelling, cheminformatics filters, and similarity

searches based on structure, fingerprint, 3D shape were some of the computer-generated strategies utilized in LBVS[23].

2.2. Structure-based virtual screening (SBVS)

In contrast, the SBVS strategy uses data on ligand's recognition site in receptor's structure as a starting point which includes the binding affinity of ligands, conformation of the receptor, charge on the surface of the molecule and configuration of molecules present in binding site[24]. These techniques require the receptor's 3D structure to be fully understood and, ideally, to be in intricate complex with the bioactive substance. Molecular dynamics simulation, structure-based pharmacophore modeling, and molecular docking are a few of the computational techniques used in the SBVS methodology[25]. Virtual screening techniques are currently a crucial component in the design and invention of novel bioactive molecules. Therefore, the applications of SBVS strategies have been increased in academics as well as industries[16].

2.3. AI-assisted virtual screening

AI has made immense progress in accelerating the identification and screening of bioactive metabolites with commercial applications. AI along with molecular modeling and cheminformatics have improved the efficiency of virtual screening strategies, thus allowing the users to explore the extremely diverse chemo-structural topographies of natural products[16]. AI-assisted virtual screening strategies have successfully predicted pharmacokinetic properties, molecular targets, bioactivities, the permeability of compounds across the blood-brain barrier, toxicity, and side effects[26]. AI algorithms utilized in ligand-based strategies have shown a higher success rate in identifying novel metabolites with less time[16]. Nevertheless, the virtual screening should be concerned with the decision of human experts in order to evade false findings and misinterpretation and to choose metabolites based on its unique features[16]. Some of those AI tools used for virtual screening and various fields of drug discovery were enlisted in (**Table 1**).

3. Applications of AI in NP-based drug discovery

The distinct properties of NPs still astonish computational experts as well as research scientists. As expected, scientists have created many computational tools with the aid of AI algorithms and implemented them in NPs-based drug discovery[27]. Over the past few decades, infinite datasets on molecular structure have been created which give data on the biochemical and physiological functions of metabolites as well. The rapid advancement of AI/ML algorithms and increasing datasets of chemical structure could proffer an exceptional chance for understanding the association between the structure and function of metabolites[28]. Similarly, those algorithms could also predict the function of NPs from biosynthetic gene clusters (BGCs)[29]. For instance, the progression of NP-based drug discovery has been gradually improving with the advancement of algorithms like Biosynthetic Gene Similarity Clustering and Prospecting Engine (BiG-SCAPE) and antibiotics and Secondary Metabolites Analysis SHell (antiSMASH) for mining of genome[30]. On the other hand, Small Molecule Accurate Recognition Technology (SMART 2.0) could predict the function of NPs effectively[31]. The identification of biosynthetic gene clusters of secondary metabolites could encode diverse structures which could be effectively predicted by PRISM 4[32]. These developments increase the availability of chemical structures of NPs which proposes a prodigious opportunity for researchers to link those structures to relevant functions using AI/ML algorithms[28]. Therefore, ML and AI algorithms have gradually paved the way for prominent research in the field of NP-based drug discovery. The most challenging task is the effective and accurate prediction of biological functions as innumerable NPs have been discovered in day-to-day life[28]. Case studies on the use of diverse algorithms in the fields of plant and microbial research have been discussed below.

3.1. Case studies on the use of AI/ML algorithms on plant

Plants have always been the center of attraction owing to their numerous beneficial effects on humans[33]. The tribute to an immense increase in plant research extends to the wide variety of secondary metabolites synthesized in a limited range[34]. Nevertheless, several biotic and abiotic factors affect the biosynthetic pathway of secondary metabolite production. Therefore, a lot of time, cost, and manual effort was needed to screen these novel bioactive metabolites. Considering this, one effective alternative includes using AI, an *in-silico* tool for plant research. It is surprising that AI was used to even predict the best suitable culture medium and phytohormones for the *in-vitro* growth of plants[35]. For predicting the role of phytohormones in plant growth, the data from *in-vitro* experimental studies are exposed to computational modeling which will imply the impact of various factors[33]. For instance, using computational techniques, an artificial neural network (ANN) was used to predict the growth requirements and bulk synthesis of biomass in *Centella asiatica*[36]. AI predicts the correlation between the influencing factors using ANN and provides the mineral inequity in plants. Hence, by this, the factors affecting the plant's growth could be optimized[37]. Recently, AI along with micro-fluidics was used to speed up the process of drug discovery[33]. On the other hand, ML was used to increase the bioactive metabolite synthesis in *Bryophyllum*[38]. This work paved way for the synthesis of plant secondary metabolites on a larger scale. AI could also predict the extinct and endangered medicinal plants and therefore could aid in the conservation of plants with high therapeutic value[39]. For instance, maximum entropy model, an ML algorithm was used for predicting the distribution of a critically endangered medicinal plant, *Lilium polyphyllum* in Indian Western Himalayan Region[40]. Similarly, seven machine learning models were used to model the habitat suitability for *Ferula gummosa* medicinal plant in mountainous region to avoid the extinction in the future[41]. It could also be used for the identification of different plant leaves using an image processor and prediction of the interaction of herbal targets[42]. Recently,

the application of ML techniques in various fields of photosynthetic research including studies on photosynthetic pigment studies have been reviewed and discussed diverse strategies on how to employ ML in enhancing crop yield[43]. ML was used to increase the bioactive metabolite synthesis in plants on large scale for commercialization purposes[44]. ANN organizes plants based on morphological characteristics like size, color, and the dimension of leaves. ML uses ANN and SVM for predicting the interconnection between photodissociation and its bioactivity[33]. The different AI algorithms used in various fields of plant research like enhancement of secondary metabolites, plant tissue culture, drug design and discovery, and disease treatment were tabulated in (**Table 2**).

3.2. Case studies on the use of AI/ML algorithm on microbes

3.2.1. Natural products from microbes: Selection and screening

The preliminary step in natural product discovery is the selection of the organism. Among various microbes, actinomycetes have been overmined as a significant source of therapeutic compounds which led to the repetitive discovery of known compounds. This led to a lack of identification of novel compounds[2]. Even though, the whole process of extraction of natural products is challenging and laborious, cautious exploration of unexplored sources enhances the chance of finding novel scaffolds[2]. The conventional way of isolation of natural products is a time-consuming process, hence with the advancement in AI/ML and omic techniques, it is possible to predict microbes proficiently[45]. For instance, the convolutional neural network (CNN) was now used to identify diverse shapes of gram-positive and gram-negative bacterial strains by high throughput imaging[46]. This technique could be expanded to identify and classify microbes using ML tools[2]. Scientists have developed, IDBac using ML for the classification of microbes based on their ability to synthesize secondary metabolites using matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS)[47]. Using this technique, the authors have categorized *Bacillus subtilis* depending on its capability to

synthesize cyclic peptide antibiotics. Similarly, ML models have been used to predict the antibacterial activity of fungal secondary metabolites from biosynthetic gene cluster data[48]. Recently, multi-omic techniques have been combined with ML algorithms for characterizing the marine metabolites datasets thus providing an unprecedented opportunity for discovering novel bioactive compounds from marine environment[49]. In the future, integration of AI/ML techniques with MALDI-TOF could be a possible technique to rapid the process of screening and extraction of NPs. MALDI has now emerged with imaging MS which could be utilized for mapping the spatial arrangement of secondary metabolites[2].

3.2.2. Genome mining

Recently, next-generation sequencing and bioinformatics have paved the way for the identification of secondary metabolites with the use of genome mining[50]. In spite of the huge diversity of NPs, their relevant BGCs are extremely conserved in micro-organisms. These BGCs belong to classes of non-ribosomally synthesized peptides, polyketide synthases, and ribosomally synthesized and post-translationally modified peptides, terpenes and alkaloids[51]. This approach starts with identifying known and unknown new BGCs from genome and characterizing them for analysis. ML algorithms aid in analyzing the big data for the prediction of these BGCs and reputed determined structures[52].

The AI algorithms employed in various fields of microbial research was enlisted in (**Table 3**). Using genome mining, gladiolin has been extracted from *Burkholderia galdioli* from a cystic fibrosis patient[53]. ML and Deep learning (DL) approach also contributed to the identification of mysterious BGCs, lanthipeptides[54]. With the help of genome mining and ML and DL approaches, it is possible to extract novel metabolites directly from uncultured microbes[55]. It is possible to identify novel compounds from human microbiota by using the hidden Markov model (HMM) algorithm. It identifies BGCs from metagenome samples[56]. Mostly some BGCs

exist silently which hinders the synthesis of secondary metabolites. However, it is possible to predict those genes using elicitors and ML/AI algorithms aid in expressing them[57]. The major disadvantage of the discovery of NPs is to identify secondary metabolites from unconventional environmental sources or biological niches without microbial cultivation. But now with the advancement of AI/ML and metagenome, NPs could be predicted directly from biotic and environmental sites[56].

3.2.3. Metabolite expression and synthesis:

Using bioinformatic tools and genome sequencing, it is predicted that *Myxococcus* and *Streptomyces* possess huge BGCs of secondary metabolites. But these BGCs remain silent without expression[58]. Recently, AI/ML algorithms have been applied to screen and monitor metabolite synthesis. For instance, deep reinforcement learning of AI was used to control the coculture of microbes in a fermentor[59]. Through this technique, the parameters of growth and the relevant output could be regulated. Hence for the synthesis of NPs, this technique could be used to control countless factors. Similarly, a high throughput strategy was used for the activation of these silent unexpressed BGCs in several organisms. Here imaging mass spectrometry (IMS) was used to screen the elicitors for inducing the secondary metabolite synthesis. The integration of this technique with laser ablation coupled electrospray ionization MS, led to the identification of a novel glycoprotein from *Amycolatopsis keratiniphila*[2].

3.2.4. AI/ML in the dereplication of NPs

Many drugs were discovered during the golden age of NPs progress, which were used even today as therapeutic agents. Yet, the repetitive discovery of already-known compounds gradually slowed down the discovery of NPs[2]. Hence for the reduction of time of analysis and resource availability, rapid recognition of identified bioactive metabolites is essential. One such process widely used to rapidly identify already known metabolites in microbial extracts is

dereplication[2]. As the extracts of microbes were enriched with several compounds, the dereplication approach could possibly reduce repetition and offers data on novel compounds. Therefore, engagement of highly accurate ML/AI tools could make this crucial task easier. Conventionally, dereplication was done by HPLC coupled with a UV/Photodiode array (PDA) detector which has integral library databases[60]. But this could not give data on structure and hence instruments with advanced multispectroscopic detectors is needed for capturing the compound's additional spectral characteristics[2].

3.2.5. AI/ML in Mass spectrometry-assisted dereplication

MS is extensively used for NPs dereplication as it is accurate, rapid, and highly sensitive. MS has the added advantage of retrieving huge amounts of structure-related data even from very less samples using a non-targeted strategy. The integration of mass-related data with UV/PDA could be used to recognize compounds with the aid of databases like MarinLit[61], NPEdia[62], Dictionary of Natural Products[63] and the Natural Product Atlas[64]. This technique was used to dereplicate the bioactive metabolites of many actinomycetes[65]. The efficient screening of bioactive metabolites could be achieved by LC-MS but the challenging part is the data analysis. But for this, scientists have to screen and search various UV spectra, mass spectra, and micro-organisms data in various databases[2]. Therefore, the use of ML techniques will be a possible way to analyze and identify natural products based on their spectral data without searching the databases manually.

The major disadvantage concerned with MS was that the molecular mass of several parent molecules of various metabolites overlaps depending on the MS spectra[66]. Hence, advanced techniques like tandem MS could detect the metabolites with high sensitivity depending on the MS/MS separation[67]. However, analysis of MS/MS data is a time-consuming and labor-intensive manual task. Hence, ML algorithms were used recently to evaluate these hugely

resolved MS spectrums with decreased noise[2]. THRASH, XCMS, MS-Dial, MZmine, Decon2LS, and MetaboAnalyst are some of the AI/ML tools used for the analysis and processing of MS data[2]. Nowadays commercialized suppliers like Thermo Fisher and Agilent are equipped with algorithms like MassHunter and XCalibur for manual prediction of metabolites with high confidence[68].

Recently, molecular networking (MN) was used to dereplicate novel bioactive metabolites from diverse sources. It evaluates the complicated data files of MS spectra and images them into network depiction. GNPS has a collection of reference spectra of a wide variety of compounds deposited from various sources which could be analyzed by MN[69]. This integrated approach is termed as Global Natural Products Social Molecular Networking. MN identifies compounds depending on the similarity of MS/MS spectra and it links the novel metabolites with known compounds by utilization of alike fragments. Dereplication could be accomplished using MN with high success probability. For instance, around 260 microbial strains from various sources have been screened using MN. Through this, the metabolome of *Pseudomonas* contributed to the identification of bananamide and poaeamide B[70]. Similarly using MN, conulothiazole C and isoconulothiazole B were identified from blue-green algae[71]. Recently, a conventional metabolomics strategy coupled with integrated untargeted liquid chromatography-tandem MS along with synchronized detection of protein affinity via native MS was created. A novel inhibitor of serine protease, rivulariapeptolides was discovered using this approach[72]. This could be a significant way for drug discovery from natural products in the future.

An advanced algorithm, DEREPLICATOR+ has been developed to aid the identification of various classes of NPs like terpenes, alkaloids, polyketides, benzenoids, and flavonoids[73]. The major issue involved in the identification of NPs is the extraction of bioactive metabolite during

the purification of the extract. As a result, integrated bioinformatics coupled with bioactivity-based MN was developed. This could be used for mapping the score of bioactivities[74].

It is easy to predict the structure of already known compounds with the available MS tools but it is difficult to predict the unknown compound's structure. But with ML it became possible. For instance, SIRIUS 4, a web-based tool uses SVM for the identification of structure[75]. An improved version, ZODIAC was developed which is 16.5 times more advanced than SIRIUS 4 and could even predict the molecular formula of compounds. Then, Deep Neural Network (DNN) was developed for the prediction of unidentified metabolites for which no structure or spectra-related data was available[75]. Another tool, MS2DeepScore predicts the unknown compounds based on the MS similarity and identifies them by grouping[69]. Hence, using MN for dereplication would be a successful hit and therefore could be utilized in the future in combination with ML for interpretation of structure for novel compounds[2].

3.2.6. Dereplication of NPs using NMR

Interpretation of metabolite's structure is another crucial task. Even though unambiguous and precise interpretation of structure was provided by X-ray crystallography, its application is very limited as it requires a single crystal[76]. On the other hand, Nuclear magnetic resonance (NMR) was widely used spectroscopic technique which infers structural data depending on the spectrum[77]. NMR-based databases like CHNMR-NP, NAPROC-13, BMRB, and Spektraris were available, they possess many disadvantages and hence could not quench the natural product discovery. As a result, NP-MRD, a database based on NMR was developed which has data on >41,000 NPs extracted from over 7400 sources[78]. This database is still in progress and in the future, this allows efficient elucidation of structure and also dereplicates in an automatic manner. Then, SMART 2.0 was developed which analyses and characterizes complex mixture of compounds leading to the characterization of novel NPs[31]. Using SMART 2.0, symplocolide

a novel macrolide was identified and annotated. Then from $^{1}H–^{13}C$ HSQC NMR spectra, SMART-miner was developed for identifying the complex metabolites using CNN. For training this tool, around 657 chemical compounds retrieved from Biological Magnetic Resonance Data Bank (BMRB) and Human Metabolome Database (HMDB) have been analyzed. This tool could identify these molecules from amalgamated mixture with 88% accuracy.

Recently, DEEP picker, an AI tool based on DNN was developed for the analysis of the 2D NMR spectrum[79]. [80] used the ML technique for the prediction of various classes of NPs from ^{13}C-NMR spectral data. As far as dereplication is concerned, High-resolution mass spectrometry (HRMS) is preferred rather than NMR owing to its high sensitivity. But NMR could predict the optical isomers accurately and identify organic molecules in the extract[81]. MixONat based on ^{13}C-NMR was developed for the identification of structurally similar NPs and optical isomers. This dereplication software was able to identify xanthones from *Calophyllum brasiliense*[82]. Another tool based on ^{1}H-NMR, Eliciting Nature's Activities (ELINA) was developed for the detection of the chemical characteristics correlating with the biological activity prior to extraction of compounds. Hence, this tool identified novel lanostane triterpenes from the fungal extract of *Fomitopsis pinicola*[83].

4. Other applications of AI/ML tools

4.1. Prediction of bioactivity and identification of target using AI/ML

Generally, the bioactivity of NPs was identified depending on the phenotypic characteristics or screening by high-throughput techniques owing to the diverse structures and extensive chemical gaps[84]. On the other hand, experimental identification of targets was conventionally performed using chemical proteomics and genomics. But validation of targets was difficult, time-consuming, and requires more effort[85]. Hence, computational strategies in turn could

reduce these constraints and limit the search for target screening[86]. The various applications of AI algorithms in various fields of NPs based drug discovery is depicted in (**Figure 3)**.

When compared to conventional ligand-based and structure-based computational identification of targets, AI/ML-based strategies have several pros and hence can be engaged as a successful approach for the identification of NP targets[2]. Recently, advanced features of AI algorithms improve the prediction of binding affinity by considering the similarity between the drug compound and its relevant target. The widely used AI/ML tools for target identification and bioactivity prediction were enlisted in (**Table 4**). From a research standpoint, the validity and accuracy of such algorithms remain a key limitation. In order to increase the accuracy and precision of AI-based algorithms through selected and substantial data input, a comprehensive study should be conducted[87].

4.2. Prediction of physicochemical properties

It is eminent that each compound possess diverse physicochemical properties like solubility, degree of ionization, partition, and permeability co-efficient that may interfere with the molecule's pharmacokinetic qualities and drug-target binding effectiveness[88]. To aid this, many AI-based techniques for predicting the chemical compound's physicochemical characteristics have been created. Molecular fingerprinting, SMILES format, Coulomb matrices, and potential energy measurements are among those AI-based tools[89]. A QSAR model was recently created by [90] to forecast the six different physiochemical characteristics of eco-friendly agents taken from environmental protection agency data. Later, six AI-based systems for the prediction of chemical absorption in the human digestive tract were developed. SVM, k-nearest neighbor, probabilistic neural network, ANN, Partial least square (PLS), and linear discriminate model are among the constructed approaches. SVM has a greater accuracy at 91.54% than the other models mentioned above[91]. An ML-based model was created in 2017

by Zang et al. to predict the physicochemical characteristics of foreign chemicals like bioconcentration factors, solubility in water, octanol-water partition co-efficient, melting and boiling point and vapor pressure[87].

Furthermore, several AI-based tools like ALOGPS 2.1 (http://www.vcclab.org/lab/alogps/), E-BABEL (http://www.vcclab.org/lab/babel/0), E-DRAGON (http://www.vcclab.org/lab/edragon/), PCLIENT (http://www.vcclab.org/lab/pclient/), ASNN (http://www.vcclab.org/lab/asnn/), ChemSpider (http://www.chemspider.com/), SPARC (http://sparc.chem.uga.edu/sparc/) and OSIRIS property explorer (https://www.organic-chemistry.org/prog/peo/) have been created. The quantitative structural toxicity of tyrosine derivatives intended for effective, safe inflammatory treatment was further predicted by [92] using ORISIS property explorer. Only 19 of the 55 bioactive compounds were found to be effective cyclooxygenase-2 inhibitors, according to the data generated by ORISIS. In a similar vein, models based on Random Forest (RF) and DNN were developed to forecast human intestinal absorption of various chemical substances. Therefore, it must be inferred from the instances that the AI-based strategy significantly contributes to drug discovery and development through the prediction of physicochemical features[87].

5. Challenges and limitations in NP-based drug discovery

5.1. Virtual screening-exclusion of compounds

In comparison with the application of conventional methods for the extraction of novel bioactive metabolites, computational strategies were known to be prognostic, low-cost, and beneficial. Nevertheless, regardless of these advantages, they also have challenges and limitations and mostly these techniques were susceptible to bias[93]. Analysis of diverse chemical structures and bioactivity of NPs by similarity-based computational tools mostly

provides biased data as it has a postulation that novel compounds might be similar to well-known bioactive compounds[93]. This hypothesis mostly leads to errors in the construction of models and hence can decrease the diversity of newly identified chemical structures. Hence, it is obvious that some compounds could be excluded from the screening process and could possibly lessen the exploration of novel chemical compounds with unique biological activity.

5.2. Generation of inaccurate data

The major challenge associated with NP-based drug targets was exploring and identifying the mechanism of action and their relevant side effects which is an expensive and time-consuming process[94]. In spite of several advantages, use of AI/ML tools could generate inaccurate data, and only already known targets can be predicted and validated[95]. On the other hand, the selection of a drug molecule depends on whether it has any side effects or toxicity. But this requires a prolonged time and it is an expensive process. This requires validation of the molecule by *in-vitro* and *in-vivo* experimental studies for assessing the toxicity[2]. Hence, computational toxicology could be used for screening several compounds simultaneously thus reducing the time of performing animal studies. But this could also generate inaccurate data[2].

5.3. Molecular featurization (Technical issue)

Over past few decades, infinite datasets on molecular structure have been created which give data on the biochemical and physiological functions of metabolites as well. The rapid advancement of AI/ML algorithms and increasing datasets of chemical structure could proffer an exceptional chance for understanding the association between the structure and function of metabolites[26]. Similarly, those algorithms could also predict the function of NPs from BGCs[29].

The most challenging task is the effective and accurate prediction of biological functions as innumerable NPs have been discovered in day-to-day life[28]. The next challenge for the

development of successful ML/AI models lies in the featurization of molecular structures of NPs. Molecular featurization is a process that converts the chemical structure of NPs to computer-readable formats[96]. NPs predominantly exist as high molecular weight compounds with diverse physicochemical properties and complex structures. On the other hand, these molecular featurization tools are designed and optimized for targeting smaller molecules. Hence, current featurization tools could not be used when the structural and physicochemical properties of NPs deviate from those of smaller molecules[28]. Firstly, the performance of existing featurization tools could be examined with different NPs having complex structures. Based on this data new featurization tools may be developed which will tailor structurally complex NPs in a better way.

5.4. Interpretation of predicted data

The next challenge lies in the interpretations of data predicted by AI/ML models. As NPs possess numerous biological functions, understanding the bioactivity and mechanism of the action itself is a complicated task as many factors were involved. Therefore, the predicted outcomes from ML/AI models should be explicable for a proper understanding of NPs biochemical properties[28]. ML coupled with biochemistry approaches could employ various computational tools for predicting the cellular, molecular and biological activities of NPs. Therefore bioactivity, targets, and toxicity predicted by AI/ML tools could provide hints on the mechanism of action of NPs.

6. Conclusion and future prospects

Natural products have instigated many successful drug discovery stories but challenges like limited yield, unfriendly extraction, unidentified functions, unpredicted targets, and intricate chemical synthesis contributed to the decline of NPs-based drug discovery. AI and ML algorithms gradually integrated various stages of NP drug discovery by assisting in finding and

elucidating the bioactive structures and capturing the molecular patterns of these structures for target prediction. In conclusion, we extensively review the latest AI/ML algorithms employed in various fields of NP-based drug discovery. These applications have been extensively growing in the last few decades, fuelled by the exceptional success of AI/ML-based approaches in diverse fields of science and technology.

The advancement of AI/ML techniques has unlocked innovative approaches to determine novel industry-oriented applications of NPs by just minimizing the economic and time constraints required for the exploration. Yet, AI algorithms could not be utilized completely for the successful exploration of NPs. The extensive diversity and structural complexity of NPs impose a great challenge for computational experts to develop a novel AI algorithm that could analyze different classes of metabolites efficiently. Therefore, the design and development of an AI tool that could analyze enormous data and different classes of secondary metabolites efficiently could contribute to fruitful outcomes in the future.

There exists a significant gap between wet lab (experimental) and computational research. Researchers from NPs research and computational experts could collaborate for successful characterization of the NPs function. Scientific researchers will understand the objective of the study and could elaborate the complicated NPs physicochemical properties whereas experts in computers could develop suitable AI tools and featurization methods for better predictions. Finally, NPs scientists could analyze and validate those predictions generated by AI. Therefore, collaboration between diverse fields of research may contribute to the efficient mining of NPs and better characterization of their functions.

CRediT authorship contribution statement

Janani Manochkumar: Conceptualization, Investigation, Literature research, Writing-Original draft preparation. **Siva Ramamoorthy:** Conceptualization, Supervision, Validation, Writing-Reviewing, and Editing.

Declaration of competing interest

The authors declare no conflict of interest.

Acknowledgment

We thank VIT, Vellore campus for providing the library and open-access facilities of the Institute for accessing journals during the preparation of the manuscript.

1. Jiménez-Luna, J., Grisoni, F. and Schneider, G., Drug discovery with explainable artificial intelligence. *Nat. Mach. Intell*., 2020, **2**(10), pp.573-584.
2. Sahayasheela, V. J., Yu, Z., Hirose, Y., Pandian, G. N., Bando, T. and Sugiyama, H., Inhibition of GLI-mediated transcription by cyclic pyrrole-imidazole polyamide in cancer stem cells. *Bull. Chem. Soc. Jpn*., 2022, ***95***(4), 693-699.
3. Siva, R., 2010. Plant dyes. In Industrial crops and uses (pp. 349-357). Wallingford UK: CABI.
4. Siva, R., Doss, F.P., Kundu, K., Satyanarayana, V.S.V. and Kumar, V., 2010. Molecular characterization of bixin—An important industrial product. Industrial Crops and products, **32**(1), pp.48-53.
5. Atanasov, A. G., Zotchev, S. B., Dirsch, V. M. and Supuran, C. T., Natural products in drug discovery: advances and opportunities. *Nat. Rev. Drug Discovery*., 2021, **20**(3), 200-216.
6. Siva, R., 2014. Food colourants and health issues: are we aware?. Current Science, **106**(2), p.143.
7. Cobb, A.H., Herbicides and plant physiology. John Wiley & Sons, 2022.
8. Bernardini, S., Tiezzi, A., Laghezza Masci, V. and Ovidi, E., Natural products for human health: an historical overview of the drug discovery approaches. *Nat. Prod. Res*., 2018, **32**(16), 1926-1950.
9. Afendi, F.M., Okada, T., Yamazaki, M., Hirai-Morita, A., Nakamura, Y., Nakamura, K., Ikeda, S., Takahashi, H., Altaf-Ul-Amin, M., Darusman, L.K. and Saito, K., KNApSAcK family databases: integrated metabolite–plant species databases for multifaceted plant research. Plant Cell Physiol., 2012, 53(2), pp.e1-e1.
10. Tsugawa, H., Rai, A., Saito, K. and Nakabayashi, R., Metabolomics and complementary techniques to investigate the plant phytochemical cosmos. *Nat. Prod. Rep.*, 2021, **38**(10), 1729-1759.
11. Tariq, A. *et al.*, Systematic review on ethnomedicines of anti-cancer plants. *Phytother. Res.*, 2017, **31**(2), 202-264.
12. Sarker, S. D. and Nahar, L., An introduction to computational phytochemistry. *In Computational phytochemistry*, Elsevier, 2018, pp. 1-41.
13. Silver, L. L., Challenges of antibacterial discovery. *Clin. Microbiol. Rev.*, 2011, **24**(1), 71-109.
14. Lyddiard, D., Jones, G. L. and Greatrex, B. W., Keeping it simple: lessons from the golden era of antibiotic discovery. *FEMS Microbiol. Lett.*, 2016, **363**(8).
15. Hautbergue, T., Jamin, E. L., Debrauwer, L., Puel, O. and Oswald, I. P., From genomics to metabolomics, moving toward an integrated strategy for the discovery of fungal secondary metabolites. *Nat. Prod. Rep.*, 2018, **35**(2), 147-173.
16. Santana, K., Do Nascimento, L. D., Lima e Lima, A., Damasceno, V., Nahum, C., Braga, R. C. and Lameira, J., Applications of virtual screening in bioprospecting: facts, shifts, and perspectives to explore the chemo-structural diversity of natural products. *Front. Chem.*, 2021, **9**, 662688.
17. Trujillo-Correa, A. I., Quintero-Gil, D. C., Diaz-Castillo, F., Quiñones, W., Robledo, S. M. and Martinez-Gutierrez, M., In vitro and in silico anti-dengue activity of compounds obtained from Psidium guajava through bioprospecting. *BMC Complement Altern. Med.*, 2019, **19**, 1-16.
18. Davison, E. K. and Brimble, M. A., Natural product derived privileged scaffolds in drug discovery. *Curr. Opin. Chem. Biol.*, 2019, **52**, 1-8.
19. Macalino, S. J. Y., Gosu, V., Hong, S. and Choi, S., Role of computer-aided drug design in modern drug discovery. *Arch. Pharmacal Res.*, 2015, **38**, 1686-1701.

20. Coimbra, J. R., Baptista, S. J., Dinis, T. C., Silva, M. M., Moreira, P. I., Santos, A. E. and Salvador, J. A., Combining virtual screening protocol and in vitro evaluation towards the discovery of BACE1 inhibitors. *Biomolecules*, 2020, **10**(4), 535.
21. Skinnider, M.A., Dejong, C.A., Franczak, B.C., McNicholas, P.D. and Magarvey, N.A., Comparative analysis of chemical similarity methods for modular natural products with a hypothetical structure enumeration algorithm. J. Cheminform., 2017, 9,1-15.
22. Garcia-Hernandez, C., Fernandez, A. and Serratosa, F., Ligand-based virtual screening using graph edit distance as molecular similarity measure. *J. Chem. Inf. Model.*, 2019, **59**(4), 1410-1421.
23. Yan, X., Liao, C., Liu, Z., T Hagler, A., Gu, Q. and Xu, J., Chemical structure similarity search for ligand-based virtual screening: methods and computational resources. *Curr. Drug Targets.*, 2016, **17**(14), 1580-1585.
24. Maia, E. H. B., Assis, L. C., De Oliveira, T. A., Da Silva, A. M. and Taranto, A. G., Structure-based virtual screening: from classical to artificial intelligence. *Front. Chem.*, 2020, **8**, 343
25. Wang, Z. *et al.*, Combined strategies in structure-based virtual screening. *Phys. Chem. Chem. Phys.*, 2020, **22**(6), 3149-3159.
26. Lin, Y., Zhang, Y., Wang, D., Yang, B. and Shen, Y. Q., Computer especially AI-assisted drug virtual screening and design in traditional Chinese medicine. *Phytomedicine*, 2022, 154481.
27. Nugroho, A. E. and Morita, H., Computationally-assisted discovery and structure elucidation of natural products. *J. Nat. Med.*, 2019, **73**, 687-695.
28. Jeon, J., Kang, S. and Kim, H. U., Predicting biochemical and physiological effects of natural products from molecular structures using machine learning. *Nat. Prod. Rep.*, 2021, **38**(11), 1954-1966.
29. Prihoda, D.*et al.*, The application potential of machine learning and genomics for understanding natural product diversity, chemistry, and therapeutic translatability. *Nat. Prod. Rep.*, 2021, **38**(6), 1100-1108.
30. Blin, K.*et al.*, AntiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res.*, 2019, **47**(W1), W81-W87.
31. Reher, R., Kim, H. W., Zhang, C., Mao, H. H., Wang, M., Nothias, L. F. and Gerwick, W. H., A convolutional neural network-based approach for the rapid annotation of molecularly diverse natural products. *J. Am. Chem. Soc.*, 2020, **142**(9), 4114-4120.
32. Skinnider, M. A.*et al.*, Comprehensive prediction of secondary metabolite structure and biological activity from microbial genome sequences. *Nat. Commun.*, 2020, **11**(1), 6058.
33. Singh, H. and Bharadvaja, N., Treasuring the computational approach in medicinal plant research. *Prog. Biophys. Mol. Biol.*, 2021, **164**, 19-32.
34. Seca, A. M. and Pinto, D. C., Biological potential and medical use of secondary metabolites. *Medicines*, 2019, **6**(2), 66.
35. Arab, M.M., Yadollahi, A., Eftekhari, M., Ahmadi, H., Akbari, M. and Khorami, S.S., Modeling and optimizing a new culture medium for in vitro rooting of G× N15 Prunus rootstock using artificial neural network-genetic algorithm. *Sci. Rep.*, 2018, **8**(1), 1-18.
36. Prasad, A., Prakash, O., Mehrotra, S., Khan, F., Mathur, A. K. and Mathur, A., Artificial neural network-based model for the prediction of optimal growth and culture conditions for maximum biomass accumulation in multiple shoot cultures of Centella asiatica. *Protoplasma*, 2017, **254**, 335-341.

37. García-Pérez, P., Lozano-Milo, E., Landin, M. and Gallego, P. P., Machine Learning unmasked nutritional imbalances on the medicinal plant Bryophyllum sp. cultured in vitro. *Front. Plant Sci.*, 2020a, **11**, 576177.
38. García-Pérez, P., Lozano-Milo, E., Landin, M. and Gallego, P. P., From ethnomedicine to plant biotechnology and machine learning: the valorization of the medicinal plant Bryophyllum sp. *Pharmaceuticals*, 2020b,**13**(12), 444.
39. Wearn, O. R., Freeman, R. and Jacoby, D. M., Responsible AI for conservation. *Nat. Mach. Intell.*, 2019, **1**(2), 72-73.
40. Dhyani, A., Kadaverugu, R., Nautiyal, B.P. and Nautiyal, M.C., Predicting the potential distribution of a critically endangered medicinal plant Lilium polyphyllum in Indian Western Himalayan Region. Reg. Environ. Change., 2021, 21, 1-11.
41. Mohammady, M., Pourghasemi, H.R., Yousefi, S., Dastres, E., Edalat, M., Pouyan, S. and Eskandari, S., Modeling and prediction of habitat suitability for Ferula gummosa medicinal plant in a mountainous area. Nat. Resour. Res., 2021, 30, 4861-4884.
42. Wang, Y., Jafari, M., Tang, Y. and Tang, J., Predicting Meridian in Chinese traditional medicine using machine learning approaches. *PLoS Comput. Biol.*, , 2019, **15**(11), e1007249.
43. Varghese, R., Cherukuri, A.K., Doddrell, N.H., Doss, C.G.P., Simkin, A.J., Ramamoorthy, R., Machine learning in photosynthesis: Prospects on sustainable crop development. Plant Sci., 2023, **335**, 111795.
44. García-Pérez, P., Lozano-Milo, E., Landín, M. and Gallego, P. P., Combining medicinal plant in vitro culture with machine learning technologies for maximizing the production of phenolic compounds. *Antioxidants*, 2020c, **9**(3), 210.
45. Hugenholtz, P., Chuvochina, M., Oren, A., Parks, D. H. and Soo, R. M., Prokaryotic taxonomy and nomenclature in the age of big sequence data. *the ISME Journal*, 2021, **15**(7), 1879-1892.
46. Smith, K. P., Kang, A. D. and Kirby, J. E., Automated interpretation of blood culture gram stains by use of a deep convolutional neural network. *J. Clin. Microbiol.*, 2018, **56**(3), e01521-17.
47. Clark, C. M., Costa, M. S., Sanchez, L. M. and Murphy, B. T., Coupling MALDI-TOF mass spectrometry protein and specialized metabolite analyses to rapidly discriminate bacterial function. *Proc. Natl. Acad. Sci.*, 2018, **115**(19), 4981-4986.
48. Riedling, O., Walker, A.S. and Rokas, A., Predicting fungal secondary metabolite activity from biosynthetic gene cluster data using machine learning. bioRxiv., 2023, 2023-09.
49. Manochkumar, J., Cherukuri, A.K., Kumar, R.S., Almansour, A.I., Ramamoorthy, S. and Efferth, T., A critical review of machine-learning for "multi-omics" marine metabolite datasets. Comput. Biol. Med., 2023, 165,107425.
50. Baltz, R. H., Genome mining for drug discovery: progress at the front end. *J. Ind. Microbiol. Biotechnol.*, 2021, **48**(9-10), kuab044.
51. Hai, Y., Huang, A. and Tang, Y., Biosynthesis of Amino Acid Derived α-Pyrones by an NRPS–NRPKS Hybrid Megasynthetase in Fungi. *J. Nat. Prod.*, 2020, **83**(3), 593-600.
52. Scherlach, K. and Hertweck, C., Mining and unearthing hidden biosynthetic potential. *Nat. Commun.*, 2021, **12**(1), 3864.
53. Song, L.*et al.*, Discovery and biosynthesis of gladiolin: a Burkholderia gladioli antibiotic with promising activity against Mycobacterium tuberculosis. *J. Am. Chem. Soc.*, 2017, **139**(23), 7974-7981.

54. Kloosterman, A. M.*et al.*, Expansion of RiPP biosynthetic space through integration of pan-genomics and machine learning uncovers a novel class of lanthipeptides. *PLoS biology*, 2020, **18**(12), e3001026.
55. Miller, S. J. and Clardy, J., Beyond grind and find. *Nat. Chem.*, 2009, **1**(4), 261-263.
56. Sugimoto, Y., Camacho, F. R., Wang, S., Chankhamjon, P., Odabas, A., Biswas, A. ... Donia, M. S., A metagenomic strategy for harnessing the chemical repertoire of the human microbiome. *Science*, 2019, **366**(6471), eaax9176.
57. Banf, M., Zhao, K. and Rhee, S. Y., METACLUSTER—an R package for context-specific expression analysis of metabolic gene clusters. *Bioinformatics*, 2019, **35**(17), 3178-3180.
58. Bader, C. D., Panter, F. and Müller, R., In depth natural product discovery-Myxobacterial strains that provided multiple secondary metabolites. *Biotechnol. Adv.*, 2020, **39**, 107480.
59. Treloar, N. J., Fedorec, A. J., Ingalls, B. and Barnes, C. P., Deep reinforcement learning for the control of microbial co-cultures in bioreactors. *PLoS Comput. Biol.*, 2020, **16**(4), e1007783.
60. Hook, D. J., More, C. F., Yacobucci, J. J., Dubay, G. and O'Connor, S., Integrated biological—physicochemical system for the identification of antitumor compounds in fermentation broths. *J. Chromatogr. A.*, 1987, **385**, 99-108.
61. Blunt, J. W., Carroll, A. R., Copp, B. R., Davis, R. A., Keyzers, R. A. and Prinsep, M. R., Marine natural products. *Nat. Prod. Rep.*, 2018, **35**(1), 8-53.
62. Tomiki, T.*et al.*, RIKEN natural products encyclopedia (RIKEN NPEdia), a chemical database of RIKEN natural products depository (RIKEN NPDepo). *J Comput Aid Chem*, 2006, **7**, 157-162.
63. Buckingham, J. (Ed.)., *Dictionary of natural products. supplement*, **4** (11). CRC press, 1997.
64. Van Santen, J. A.*et al.*, The natural products atlas: an open access knowledge base for microbial natural products discovery. *ACS Cent. Sci.*, 2019, *5*(11), 1824-1833.
65. Mehetre, G. T., Vinodh, J. S., Burkul, B. B., Desai, D., Santhakumari, B., Dharne, M. S. and Dastager, S. G., Bioactivities and molecular networking-based elucidation of metabolites of potent actinobacterial strains isolated from the Unkeshwar geothermal springs in India. *RSC Adv.*, 2019, **9**(17), 9850-9859.
66. Caesar, L. K., Kellogg, J. J., Kvalheim, O. M. and Cech, N. B., Opportunities and limitations for untargeted mass spectrometry metabolomics to identify biologically active constituents in complex natural product mixtures. *J. Nat. Prod.*, 2019, **82**(3), 469-484.
67. Hoffmann, T., Krug, D., Hüttel, S. and Müller, R., Improving natural products identification through targeted LC-MS/MS in an untargeted secondary metabolomics workflow. *Anal. Chem.*, 2014, **86**(21), 10780-10788.
68. Kumar, V., Kumar, A. A., Joseph, V., Dan, V. M., Jaleel, A., Kumar, T. S. and Kartha, C. C., Untargeted metabolomics reveals alterations in metabolites of lipid metabolism and immune pathways in the serum of rats after long-term oral administration of Amalaki rasayana. *Mol. Cell. Biochem.*, 2020, **463**, 147-160.
69. Wang, M.*et al.*, Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.*, 2016, **34**(8), 828-837.
70. Nguyen, D. D.*et al.*, Indexing the Pseudomonas specialized metabolome enabled the discovery of poaeamide B and the bananamides. *Nat. Microbiol.*, 2016, **2**(1), 1-10.

71. Teta, R.*et al.,* A joint molecular networking study of a Smenospongia sponge and a cyanobacterial bloom revealed new antiproliferative chlorinated polyketides. *Org. Chem. Front.*, 2019, **6**(11), 1762-1774.
72. Reher, R., Aron, A. T., Fajtová, P., Stincone, P., Wagner, B., Pérez-Lorente, A. I. and Petras, D., Native metabolomics identifies the rivulariapeptolide family of protease inhibitors. *Nat. Commun.,* 2022, **13**(1), 4619.
73. Mohimani, H.*et al.,* Dereplication of microbial metabolites through database search of mass spectra. *Nat. Commun.*, 2018, **9**(1), 4035.
74. Nothias, L. F.*et al.,* Bioactivity-based molecular networking for the discovery of drug leads in natural product bioassay-guided fractionation. *J. Nat. Prod.*, 2018, **81**(4), 758-767.
75. Dührkop, K., Fleischauer, M., Ludwig, M., Aksenov, A. A., Melnik, A. V., Meusel, M. and Böcker, S., SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat. Methods.*, 2019, **16**(4), 299-302.
76. Buevich, A. V. and Elyashberg, M. E., Synergistic combination of CASE algorithms and DFT chemical shift predictions: a powerful approach for structure elucidation, verification, and revision. *J. Nat. Prod.*, 2016, **79**(12), 3105-3116.
77. Reynolds, W. F, Natural product structure elucidation by NMR spectroscopy. In *Pharmacognosy,* Academic Press, 2017, pp. 567-596.
78. Wishart, D. S.*et al.,* NP-MRD: the natural products magnetic resonance database. *Nucleic Acids Res.*, 2022, **50**(D1), D665-D677.
79. Li, D. W., Hansen, A. L., Yuan, C., Bruschweiler-Li, L. and Brüschweiler, R., DEEP picker is a deep neural network for accurate deconvolution of complex two-dimensional NMR spectra. *Nat. Commun.*, 2021, **12**(1), 5229.
80. Martinez-Trevino, S. H., Uc-Cetina, V., Fernandez-Herrera, M. A. and Merino, G., Prediction of natural product classes using machine learning and 13C NMR spectroscopic data. *J. Chem. Inf. Model.*, 2020, **60**(7), 3376-3386.
81. Vignoli, A.*et al.,* High-throughput metabolomics by 1D NMR. *Angew. Chem. Int. Ed.*, 2019, **58**(4), 968-994.
82. Bruguière, A., Derbré, S., Dietsch, J., Leguy, J., Rahier, V., Pottier, Q. and Richomme, P., MixONat, a software for the dereplication of mixtures based on 13C NMR spectroscopy. *Anal. Chem.*, 2020, **92**(13), 8793-8801.
83. Grienke, U., Foster, P. A., Zwirchmayr, J., Tahir, A., Rollinger, J. M. and Mikros, E., 1H NMR-MS-based heterocovariance as a drug discovery tool for fishing bioactive compounds out of a complex mixture of structural analogues. *Sci. Rep.*, 2019, **9**(1), 1-10.
84. Moffat, J. G., Vincent, F., Lee, J. A., Eder, J. and Prunotto, M., Opportunities and challenges in phenotypic drug discovery: an industry perspective. *Nat. Rev. Drug Discovery.*, 2017, **16**(8), 531-543.
85. Zeng, X.*et al.,* Target identification among known drugs by deep learning from heterogeneous networks. Chem. Sci., 2020, **11**(7), 1775-1797.
86. Langley, G. R.*et al.,* Towards a 21st-century roadmap for biomedical research and drug discovery: consensus report and recommendations. *Drug discovery today,* 2017, **22**(2), 327-339.
87. Gupta, R., Srivastava, D., Sahu, M., Tiwari, S., Ambasta, R. K. and Kumar, P., Artificial intelligence to deep learning: machine intelligence approach for drug discovery. *Mol. Diversity.*, 2021, **25**, 1315-1360.
88. Lynch S.R., Bothwell T. and Campbell, L., A comparison of physical properties, screening procedures and a human efficacy trial for predicting the bioavailability of

commercial elemental iron powders used for food fortification. *Int J Vitam Nutr Res.*, 2007, **77** (2), 107-124.
89. Schneider, P.*et al.*, Rethinking drug design in the artificial intelligence era. *Nat. Rev. Drug Discovery.*, 2020, **19**(5), 353-364.
90. Zhang, W., Pei, J. and Lai, L., Computational multitarget drug design. *J. Chem. Inf. Model.*, 2017, **57**(3), 403-412.
91. Kumar, R., Sharma, A., Siddiqui, M. H. and Tiwari, R. K., Prediction of human intestinal absorption of compounds using artificial intelligence techniques. *Curr. Drug Discov. Technol.*, 2017, **14**(4), 244-254.
92. Puratchikody, A., Sriram, D., Umamaheswari, A. and Irfan, N., 3-D structural interactions and quantitative structural toxicity studies of tyrosine derivatives intended for safe potent inflammation treatment. *Chem. Cent. J.*, 2016, **10**(1), 1-19.
93. Sieg, J., Flachsenberg, F. and Rarey, M., In need of bias control: evaluating chemical data for machine learning in structure-based virtual screening. *J. Chem. Inf. Model.*, 2019, **59**(3), 947-961.
94. Chen, X.*et al.*, Target identification of natural medicine with chemical proteomics approach: probe synthesis, target fishing and protein identification. *Signal Transduction Targeted Ther.*, 2020, **5**(1), 72.
95. Rodrigues, T., Reker, D., Schneider, P. and Schneider, G., Counting on natural products for drug design. *Nat. Chem.*, 2016, **8**(6), 531-541.
96. Wu, Z. *et al.*, MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.*, 2018, **9**(2), 513-530.
97. Spiegel, J. O., and Durrant, J. D., AutoGrow4: an open-source genetic algorithm for de novo drug design and lead optimization. *J. Cheminf.*, 2020, **12**(1), 1-16.
98. Li, J.*et al.*, Identification of target genes at Juvenile idiopathic arthritis GWAS loci in human neutrophils. *Front. Genet.*, 2019, **10**, 181.
99. Ha, E. J., Lwin, C. T. and Durrant, J. D., LigGrep: A tool for filtering docked poses to improve virtual-screening hit rates. J. Cheminf., 2020, **12**(1), 1-12.
100. Lagarde, N. *et al.*, A free web-based protocol to assist structure-based virtual screening experiments. *Int. J. Mol. Sci.*, 2019, **20**(18), 4648.
101. Hu, J., Liu, Z., Yu, D. J. and Zhang, Y., LS-align: an atom-level, flexible ligand structural alignment algorithm for high-throughput virtual screening. *Bioinformatics*, 2018, **34**(13), 2209-2218.
102. Rifaioglu, A. S., Nalbat, E., Atalay, V., Martin, M. J., Cetin-Atalay, R. and Doğan, T., DEEPScreen: high performance drug–target interaction prediction with convolutional neural networks using 2-D structural compound representations. *Chem. Sci.*, 2020, **11**(9), 2531-2557.
103. Dong, J. *et al.*, ChemDes: an integrated web-based platform for molecular descriptor and fingerprint computation. *J. Cheminf.*, 2015, **7**(1), 1-10.
104. Oldenhof, M., Arany, A., Moreau, Y. and Simm, J., ChemGrapher: optical graph recognition of chemical compounds by deep learning. *J. Chem. Inf. Model.*, 2020, **60**(10), 4506-4517.
105. Dong, J. *et al.*, ChemSAR: an online pipelining platform for molecular SAR modeling. *J. Cheminf.*, 2017, **9**, 1-13.
106. Buyukbingol, E., Sisman, A., Akyildiz, M., Alparslan, F. N. and Adejare, A., Adaptive neuro-fuzzy inference system (ANFIS): a new approach to predictive modeling in QSAR applications: a study of neuro-fuzzy modeling of PCP-based NMDA receptor antagonists. *Bioorg. Med. Chem.*, 2007, **15**(12), 4265-4282.
107. Angelo, R.M., Io, A.K., Almeida, M.P., Silveira, R.G., Oliveira, P. R., Alcazar, J. J. and Bettanin, F., OntoQSAR: An ontology for interpreting chemical and biological

data in quantitative structure-activity relationship studies. In *2020 IEEE 14th International Conference on Semantic Computing* (ICSC), IEEE, 2020, pp. 203-206.
108. Jiang, H. J., Huang, Y. A. and You, Z. H., Predicting drug-disease associations via using gaussian interaction profile and kernel-based autoencoder. *Biomed Res. Int.*, 2019.
109. Martinez, V., Navarro, C., Cano, C., Fajardo, W. and Blanco, A., DrugNet: network-based drug–disease prioritization by integrating heterogeneous data. *Artif. Intell. Med.*, 2015, **63**(1), 41-49.
110. Sadeghi, S. S. and Keyvanpour, M., RCDR: a recommender-based method for computational drug repurposing. In *2019 5th conference on KBEI*, IEEE, 2019, pp. 467-471.
111. Shar, P. A. *et al.*, Pred-binding: large-scale protein–ligand binding affinity prediction. *J. Enzyme Inhib. Med. Chem.*, 2016, **31**(6), 1443-1450.
112. Pires, D. E. and Ascher, D. B., CSM-lig: a web server for assessing and comparing protein–small molecule affinities. *Nucleic Acids Res.*, 2016, **44**(W1), W557-W561.
113. Pires, D. E., Blundell, T. L. and Ascher, D. B., mCSM-lig: quantifying the effects of mutations on protein-small molecule affinity in genetic disease and emergence of drug resistance. *Sci. Rep.*, 2016, **6**(1), 29575.
114. Capuzzi, S. J., Kim, I. S. J., Lam, W. I., Thornton, T. E., Muratov, E. N., Pozefsky, D. and Tropsha, A., Chembench: a publicly accessible, integrated cheminformatics portal. *J. Chem. Inf. Model.*, 2017, **57**(2), 105-108.
115. Patel, R. D., Prasanth Kumar, S., Pandya, H. A. and Solanki, H. A., MDCKpred: a web-tool to calculate MDCK permeability coefficient of small molecule using membrane-interaction chemical features. *Toxicol. Mech. Methods.*, 2018, **28**(9), 685-698.
116. Hornig, M. and Klamt, A., COSMO f rag: A Novel Tool for High-Throughput ADME Property Prediction and Similarity Screening Based on Quantum Chemistry. *J. Chem. Inf. Model.*, 2005, **45**(5), 1169-1177.
117. Montanari, F., Knasmüller, B., Kohlbacher, S., Hillisch, C., Baierová, C., Grandits, M. and Ecker, G. F., Vienna LiverTox workspace—a set of machine learning models for prediction of interactions profiles of small molecules with transporters relevant for regulatory agencies. *Front. Chem.*, 2020, **7**, 899.
118. Hassan-Harrirou, H., Zhang, C. and Lemmin, T., RosENet: improving binding affinity prediction by leveraging molecular mechanics energies with an ensemble of 3D convolutional neural networks. *J. Chem. Inf. Model.*, 2020, **60**(6), 2791-2802.
119. Yang, J., He, S., Zhang, Z. and Bo, X., NegStacking: Drug− Target Interaction Prediction Based on Ensemble Learning and Logistic Regression. *IEEE/ACM Trans. Comput. Biol. Bioinf.*, 2020, **18**(6), 2624-2634.
120. Lagunin, A., Stepanchikova, A., Filimonov, D. and Poroikov, V., PASS: prediction of activity spectra for biologically active substances. *Bioinformatics*, 2000, **16**(8), 747-748.
121. Reker, D., Rodrigues, T., Schneider, P. and Schneider, G., Identifying the macromolecular targets of de novo-designed chemical entities through self-organizing map consensus. *Proc. Natl. Acad. Sci.*, 2014, **111**(11), 4067-4072.
122. Keiser, M. J., Roth, B. L., Armbruster, B. N., Ernsberger, P., Irwin, J. J. and Shoichet, B. K., Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.*, 2007, **25**(2), 197-206.
123. He, J., Chen, L., Chu, B. and Zhang, C., Determination of total polysaccharides and total flavonoids in Chrysanthemum morifolium using near-infrared hyperspectral imaging and multivariate analysis. *Molecules*, 2018, **23**(9), 2395.

124. Deep, K. and Katiyar, V. K., Multi objective extraction optimization of bioactive compounds from gardenia using real coded genetic algorithm. In *6th World Congress of Biomechanics (WCB 2010). August 1-6, 2010 Singapore: In Conjunction with 14th International Conference on Biomedical Engineering (ICBME) and 5th Asia Pacific Conference on Biomechanics* Springer Berlin Heidelberg, *(APBiomech)* 2010, pp. 1463-1466.
125. Farhadi, S., Salehi, M., Moieni, A., Safaie, N. and Sabet, M. S., Modeling of paclitaxel biosynthesis elicitation in Corylus avellana cell culture using adaptive neuro-fuzzy inference system-genetic algorithm (ANFIS-GA) and multiple regression methods. *PloS one*, 2020, **15**(8), e0237478.
126. Begue, A., Kowlessur, V., Singh, U., Mahomoodally, F. and Pudaruth, S., Automatic recognition of medicinal plants using machine learning techniques. *Int J Adv Comput Sci Appl.*, 2017, **8**(4), 166-175.
127. Gago, J., Pérez-Tornero, O., Landín, M., Burgos, L. and Gallego, P. P., Improving knowledge of plant tissue culture and media formulation by neurofuzzy logic: a practical case of data mining using apricot databases. *J. Plant Physiol.*, 2011, **168**(15), 1858-1865.
128. Dutta Gupta, S. and Pattanayak, A. K., Intelligent image analysis (IIA) using artificial neural network (ANN) for non-invasive estimation of chlorophyll content in micropropagated plants of potato. *In Vitro Cell. Dev. Biol.*, 2017, **53**, 520-526.
129. Mridula, M. R., Nair, A. S. and Kumar, K. S., Genetic programming based models in plant tissue culture: an addendum to traditional statistical approach. *PLoS Comput. Biol.*, 2018, **14**(2), e1005976.
130. Mansouri, A., Fadavi, A. and Mortazavian, S. M. M., An artificial intelligence approach for modeling volume and fresh weight of callus–A case study of cumin (Cuminum cyminum L.). *J. Theor. Biol.*, 2016, **397**, 199-205.
131. Mohd, Z. R., Arun, K. K. and Narendra, S. B., "Retraction: Plant Regeneration In Chlorophytum Borivilianum Sant. Et Fernand. From Embryogenic Callus and Cell Suspension Culture and Assessment of Genetic Fidelity of Plants Derived Through Somatic Embryogenesis." *Physiol. Mol. Biol. Plants*., 2012, **18** (3), pp. 253-263.
132. Akin, M., Eyduran, S.P., Eyduran, E. and Reed, B.M., Analysis of macro nutrient related growth responses using multivariate adaptive regression splines. *PCTOC*, 2020, **140**, 661-670.
133. Barone, J. O., Use of multiple regression analysis and artificial neural networks to model the effect of nitrogen in the organogenesis of Pinus taeda L. *PCTOC*, 2019, **137**(3), 455-464.
134. Gago, J., Landín, M. and Gallego, P. P., A neurofuzzy logic approach for modeling plant processes: A practical case of in vitro direct rooting and acclimatization of Vitis vinifera L. *Plant Sci.*, 2010, **179**(3), 241-249.
135. Hameg, R., Arteta, T. A., Landin, M., Gallego, P. P. and Barreal, M. E., Modeling and optimizing culture medium mineral composition for in vitro propagation of Actinidia arguta. *Front. Plant Sci.*, 2020, **11**, 554905.
136. Munasinghe, S. P., Somaratne, S., Weerakoon, S. R. and Ranasinghe, C., Prediction of chemical composition for callus production in Gyrinops walla Gaetner through machine learning. *Inf. Process. Agric.*, 2020, **7**(4), 511-522.
137. Alanagh, E. N., Garoosi, G. A., Haddad, R., Maleki, S., Landín, M. and Gallego, P. P., Design of tissue culture media for efficient Prunus rootstock micropropagation using artificial intelligence models. *PCTOC*, 2014, **117**, 349-359.

138. Jamshidi, S., Yadollahi, A., Ahmadi, H., Arab, M. M. and Eftekhari, M. Predicting in vitro culture medium macro-nutrients composition for pear rootstocks using regression analysis and neural network models. *Front. Plant Sci.*, 2016, **7**, 274.
139. Zhang, Q., Deng, D., Dai, W., Li, J. and Jin, X., Optimization of culture conditions for differentiation of melon based on artificial neural network and genetic algorithm. *Sci. Rep.*, 2020, **10**(1), 1-8.
140. Ancuceanu, R., Hovanet, M. V., Anghel, A. I., Furtunescu, F., Neagu, M., Constantin, C. and Dinu, M.,Computational models using multiple machine learning algorithms for predicting drug hepatotoxicity with the DILIrank dataset. *Int. J. Mol. Sci.*, 2020, **21**(6), 2114.
141. Islam, T., Hussain, N., Islam, S. and Chakrabarty, A., Detecting adverse drug reaction with data mining and predicting its severity with machine learning. In *2018 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, IEEE, 2018, pp. 1-5.
142. Zhao, K. and So, H. C., Drug repositioning for schizophrenia and depression/anxiety disorders: a machine learning approach leveraging expression data. *IEEE J. Biomed. Health. Inf.*, 2018, **23**(3), 1304-1315.
143. Ning, A., Lau, H. C., Zhao, Y. and Wong, T. T., Fulfillment of retailer demand by using the MDL-optimal neural network prediction and decision policy. *IEEE Trans. Ind. Inf.*, 2009, **5**(4), 495-506.
144. Kim, E., Choi, A. S. and Nam, H., Drug repositioning of herbal compounds via a machine-learning approach. *BMC Bioinf.*, 2019, **20**(10), 33-43.
145. Mercorelli, B., Palù, G. and Loregian, A., Drug repurposing for viral infectious diseases: how far are we?. *Trends Microbiol.*, 2018, **26**(10), 865-876.
146. Yang, X., Wang, Y., Byrne, R., Schneider, G. and Yang, S., Concepts of artificial intelligence for computer-assisted drug discovery. *Chem. Rev.*, 2019, **119**(18), 10520-10594.
147. Yao, Y., Wang, Z., Li, L., Lu, K., Liu, R., Liu, Z. and Yan, J., An ontology-based artificial intelligence model for medicine side-effect prediction: taking traditional Chinese medicine as an example. *Comput. Math. Methods Med.*, 2019.
148. Kazemipoor, M., Hajifaraji, M., Shamshirband, S., Petković, D. and Kiah, M. L. M., Appraisal of adaptive neuro-fuzzy computing technique for estimating anti-obesity properties of a medicinal plant. *Comput. Methods Programs Biomed.*, 2015, **118**(1), 69-76.
149. Dudek, G., Grzywna, Z. J. and Willcox, M. L., Classification of antituberculosis herbs for remedial purposes by using fuzzy sets. *Biosystems*, 2008, **94**(3), 285-289.
150. Zha, Q. L. *et al.*, Predictive role of diagnostic information in treatment efficacy of rheumatoid arthritis based on neural network model analysis. *Zhong xi yi jie he xue bao= J Chin Integr Medicine.*, 2007, **5**(1), 32-38.
151. Tao, W., Xu, X., Wang, X., Li, B., Wang, Y., Li, Y. and Yang, L., Network pharmacology-based prediction of the active ingredients and potential targets of Chinese herbal Radix Curcumae formula for application to cardiovascular disease. *J. Ethnopharmacol.*, 2013, **145**(1), 1-10.
152. Xu, X. *et al.*, Identification of herbal categories active in pain disorder subtypes by machine learning help reveal novel molecular mechanisms of algesia. *Pharmacol. Res.*, 2020, **156**, 104797.
153. Keum, J., Yoo, S., Lee, D. and Nam, H., Prediction of compound-target interactions of natural products using large-scale drug and protein information. *BMC Bioinf.*, 2016, **17**(6), 417-425.

154. Tan, C., Wu, C., Huang, Y., Wu, C. and Chen, H., Identification of different species of Zanthoxyli Pericarpium based on convolution neural network. *PloS one*, 2020, **15**(4), e0230287.
155. Expósito, N., Kumar, V., Sierra, J., Schuhmacher, M. and Papiol, G. G., Performance of Raphidocelis subcapitata exposed to heavy metal mixtures. *Sci. Total Environ.*, 2017, **601**, 865-873.
156. Dumolin, C. *et al.*, Introducing SPeDE: High-throughput dereplication and accurate determination of microbial diversity from matrix-assisted laser desorption–ionization time of flight mass spectrometry data. *Msystems*, 2019, **4**(5), e00437-19.
157. Clark, C. M., Costa, M. S., Sanchez, L. M. and Murphy, B. T., Coupling MALDI-TOF mass spectrometry protein and specialized metabolite analyses to rapidly discriminate bacterial function. *Proc. Natl. Acad. Sci.*, 2018, **115**(19), 4981-4986.
158. Hammami, R., Zouhir, A., Le Lay, C., Ben Hamida, J. and Fliss, I., BACTIBASE second release: a database and tool platform for bacteriocin characterization. *BMC Microbiol.*, 2010, **10**(1), 1-5.
159. Kautsar, S. A. *et al.*, MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Res.*, 2020, **48**(D1), D454-D458.
160. Palaniappan, K. *et al.*, IMG-ABC v. 5.0: an update to the IMG/Atlas of Biosynthetic Gene Clusters Knowledgebase. *Nucleic Acids Res.*, 2020, **48**(D1), D422-D430.
161. Skinnider, M. A. *et al.*, Genomes to natural products prediction informatics for secondary metabolomes (PRISM). *Nucleic Acids Res.*, 2015, **43**(20), 9645-9662.
162. Mungan, M. D., Alanjary, M., Blin, K., Weber, T., Medema, M. H. and Ziemert, N., ARTS 2.0: feature updates and expansion of the Antibiotic Resistant Target Seeker for comparative genome mining. *Nucleic Acids Res.*, 2020, **48**(W1), W546-W552.
163. Sugimoto, Y. *et al.*, A metagenomic strategy for harnessing the chemical repertoire of the human microbiome. *Science*, 2019, **366**(6471), eaax9176.
164. Reddy, B. V. B., Milshteyn, A., Charlop-Powers, Z. and Brady, S. F., eSNaPD: a versatile, web-based bioinformatics platform for surveying and mining natural product biosynthetic diversity from metagenomes. *Chem. Biol.*, 2014, **21**(8), 1023-1033.
165. Covington, B. C. and Seyedsayamdost, M. R., MetEx, a metabolomics explorer application for natural product discovery. *ACS Chem. Biol.*, 2021, **16**(12), 2825-2833.
166. Tomiki, T. *et al.*, RIKEN natural products encyclopedia (RIKEN NPEdia), a chemical database of RIKEN natural products depository (RIKEN NPDepo). *J Comput Aid Chem*, 2006, **7**, 157-162.
167. Moumbock, A. F. *et al.*, StreptomeDB 3.0: an updated compendium of streptomycetes natural products. *Nucleic Acids Res.*, 2021, **49**(D1), D600-D604.
168. Pilon, A. C. *et al.*, NuBBEDB: an updated database to uncover chemical and biological information from Brazilian biodiversity. *Sci. Rep.*, 2017, **7**(1), 7215.
169. Lyu, C., Chen, T., Qiang, B., Liu, N., Wang, H., Zhang, L. and Liu, Z., CMNPD: a comprehensive marine natural products database towards facilitating drug discovery from the ocean. *Nucleic Acids Res.*, 2021, **49**(D1), D509-D515.
170. Sorokina, M. and Steinbeck, C., NaPLeS: a natural products likeness scorer—web application and database. *J. Cheminf.*, 2019, **11**(1), 55.
171. Naghizadeh, A. *et al.*, UNaProd: a universal natural product database for Materia Medica of Iranian traditional medicine. *eCAM.*, 2020.
172. Dührkop, K., Fleischauer, M., Ludwig, M., Aksenov, A. A., Melnik, A. V., Meusel, M. and Böcker, S., SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat. Methods.*, 2019, **16**(4), 299-302.
173. Madhukar, N. S. *et al.*, A Bayesian machine learning approach for drug target identification using diverse data types. *Nat. Commun.*, 2019, **10**(1), 5221.

174. Walker, A. S. and Clardy, J., A machine learning bioinformatics method to predict biological activity from biosynthetic gene clusters. *J. Chem. Inf. Model.*, 2021, **61**(6), 2560-2571.
175. Nickel, J. *et al.*, SuperPred: update on drug classification and target prediction. *Nucleic Acids Res.*, 2014, **42**(W1), W26-W31.
176. Nascimento, A. C., Prudêncio, R. B. and Costa, I. G., A drug-target network-based supervised machine learning repurposing method allowing the use of multiple heterogeneous information sources. *Computational Methods for Drug Repurposing*, 2019, 281-289.
177. Beck, B. R., Shin, B., Choi, Y., Park, S. and Kang, K., Predicting commercially available antiviral drugs that may act on the novel coronavirus (SARS-CoV-2) through a drug-target interaction deep learning model. *Comput. Struct. Biotechnol. J.*, 2020, **18**, 784-790.
178. Lee, H. and Kim, W., Comparison of target features for predicting drug-target interactions by deep neural network based on large-scale drug-induced transcriptome data. *Pharmaceutics*, 2019, **11**(8), 377.

1084 **Table 1.** Application of AI/ML tools in virtual screening and various fields of NP-based drug
1085 discovery

Application	Tool and software	Method	Features
Structure and Ligand-based Virtual Screening	AutoGrow 4	Genetic algorithms	Optimization of Lead compound and de novo drug design[97]
	LSA	Conventional Similarity and a substructure match algorithms (GMA)	A structure-based alignment tool for virtual screening of pharmaceutical compounds[98]
	LigGrep	Machine learning	Filtration of docked models for enhancing the hit ranks of virtual screening[99]
	Trix X	Machine learning	Structure-based molecular indexing tool that is enabled for the fastest and largest virtual screening[87]
	Drug Finder	Machine learning	*In-silico* virtual screening tool intended for validation while screening the compounds[100]
	LS-align	Machine learning	A high-throughput screening method used to generate fast, reliable, and accurate atom-level structural alignment of ligands[101]
	DEEPScreen	Convolutional neural networks	A high-performance tool used for the prediction of the binding of the drug to target[102]
Drug design and Discovery	ChemDes	Chemopy, Pybel	An integrated online software used for the computation of molecular descriptors and fingerprints[103]
	ChemGrapher	Deep Learning	Recognizes chemical compounds using optical graph[104]
QSAR modeling	ChemSAR	ChemoPy	Generates Molecular SAR model benefiting cheminformatics[105]
	ANFIS	Neuro-fuzzy modeling	A QSAR model used for the evaluation of physicochemical characteristics of chemical molecules[106]
	OntoQSAR	Machine learning	Interpretation and evaluation of biological and chemical data[107]
Drug repurposing	GIPAE	Gaussian interaction profile	A drug repositioning tool used to recognize novel signs for existing drugs[108]
	DrugNEt	Machine learning	Integrates heterogenous information by prioritizing the interaction of drugs against target[109]
Drug repurposing	RCDR	Collaborative filtering model	Gives high preference for the candidate drugs against diseases[110]

	DrPOCS	Machine learning	Predicts the interaction of drugs and diseases based on projection onto convex[42]
	Pred-binding	Vector machine	Predicts the binding of proteins to ligand on a large scale[111]
Physico-chemical properties and bioactivity prediction	CSM-lig	Machine learning	A web-based tool to compare and evaluate affinity of proteins to small molecules[112]
	mCSM-AB	Machine learning	Quantifies the mutational effects on affinity of proteins to small molecules in genetic diseases[113]
	Chembranch	Machine learning	Publicly available, integrated Cheminformatics tool[114]
	MDCK pred	Regression model	Prioritizes small molecules by calculating MDCK permeability[115]
	COSMOfrag	Quantum Chemistry	A high-throughput technique used for predicting ADME properties and similarity screening[116]
	Vienna LiverTox	Machine learning classification model	Identifies and recognizes pharmacokinetic properties[117]
	RosENet	Convolutional neural network	Predicts the accurate binding efficiency of proteins with ligands[118]
	DeepPurpose	Deep Learning	Open library available for predicting the interaction of drug to target[119]
Molecular Target prediction	PASS	NB	Predicts the bioactivity, mechanism of action and pharmaceutical properties[120]
	TiGER	Multiple SOMs	It qualitatively predicts targets on a larger scale[121]
	STarFish	MLP, kNN	Predicts the prediction of small molecule binding to target[95]
	SPiDER	SOMs	Identification of novel compounds in chemical biology and evaluates the probable side effects[121]
	SEA	Kruskal algorithm	Prediction of chemical similarity of proteins to ligands[122]

1095 **Table 2.** Case studies on the utilization of AI algorithms in various fields of plant research

Algorithm	Plant	Applications
Enhancement of secondary metabolites in plants		
Least square-Support vector machine	*Chrysenthenum morifolium*	AI was used to estimate the total flavonoid and polysaccharide content[123]
Artificial neural network	*Bryophyllum sp.*	To maximize the production of chemical synthesis[38]
Real coded genetic algorithm (MI-LXPM)	*Gardenia*	To predict the optimal ideal condition for extraction of total phenolic compounds[124]
Neurofuzzy inference system genetic algorithm	*Corylus avellane*	To optimize the secondary metabolite concentration[125]
Plant Tissue Culture		
Multilayer perception	-	To optimize the surface sterilization protocol without causing damage to explant[126]
Neuro-fuzzy logic	*Prunus armeniaca*	To predict the number of shoot multiplication using hormones, nutrients and vitamins[127]
Intelligent image analysis by ANN	*Solanum tuberosum*	To predict the characteristic features of shoot[128]
Genetic algorithm (AI-based modelling)	*Wrightia tinctoria*	To optimize the environmental conditions to utilize charcoal for rhizogenesis and to lower caulogenesis[129]
Backpropagation algorithms in artificial neural network	*Cuminum cyminum*	To predict the formation of callus and to determine its volume and fresh weight[130]
Backpropagation Neural network	*Chlorophytum borivilianum*	To predict the development of shoots in fermentor and fresh weight of plantlets[131]
Multivariate Adaptive Regression Splines Algorithm	*Fragaria ananassa*	To predict the nutrients required for culture of strawberry and to predict the responses like shoot quality, multiplication and leaf color responses[132]
Multilayer perception	*Pinus taeda*	To predict the impact of nitrogen source on organogenesis of shoot[133]
Multilayer perception-based modeling	*Vitis vinifera*	To optimize the factors affecting *in-vitro* root formation[134]
ANN, fuzzy logic and genetic algorithms	*Actinidia arguta.*	To reduce mineral and salt content for enhancing the micropropagation[135]

ML algorithms and artificial neural network	*Gyrinops walla* Gaetner	To predict the chemical composition for production of callus[136]
Neurofuzzy logic	*Prunus* sp.	To predict the best medium for rootstock micropropagation[137]
Regression analysis and artificial neural network analysis	*Pyrus communis*	To predict the *in-vitro* culture medium macronutrients for rootstock propagation and to analyze the growth parameters like shoot tip necrosis, shoot tip length, explant growth rate, vitrification and chlorosis[138]
Neural networks and genetic algorithm	*Cucumis melo*	To optimize the in-vitro culture condition[139]

Drug design and discovery

Algorithm	Target	Application
ML algorithm	Drug-induced liver injury	To predict the upsurge/reduction in the efficacy of multiple drug interactions and to evaluate the inhibition rate of drugs[140]
ML algorithm-Random Forest and support vector machine	Drug-ADR association	To identify different adverse drug reactions and to predict the intensity of outcome and achieved a 91% accuracy rate in predicting the death causing adverse drug reactions[141]
Support vector machine	Schizophrenia and depression/anxiety	Drug repositioning-To predict the indications for disease based on the drug expression profiles[142]
Supervised learning (SVM)-neural network	Drug-ADR association	To predict adverse drug interactions[143]
Machine learning algorithm	Classification of Chinese herbs	To determine the molecular features of 646 Chinese herbs and their active constituents by structure-based fingerprints and ADME properties[42]
Logistic regression, random forest, and support vector machine algorithms	Drug repurposing	To explore the unknown medicinal properties of herbal bioactive compounds and has identified novel indications for 20 known drugs and 31 herbal compounds[144]
Regularised least square (semi-	Drug repurposing	To identify the novel pharmacological significance of

supervised based new modelling)		existing drugs for viral infections[145]
Machine learning approach	Drug discovery	To elucidate the medicinal value of *Xiaoxuming* decoction to be utilized as a neuroprotective agent[146]
Ontology-based AI model	AI-based TCM screening	To predict the side effects of prescription[147]
AI in disease treatment		
Neuro-fuzzy	Treatment of disease	To evaluate the pharmacological aspect of medicinal plants for the treatment of obesity[148]
Fuzzy logic	Disease treatment	To group plants with anti-tuberculosis properties based on botanical data[149]
Convolutional neural network	Rheumatoid arthritis	To predict the significance of traditional Chinese medicines against inflammatory rheumatoid disease[150]
Network pharmacology-based prediction	Cardiovascular disease	To predict the mechanism of phytocompounds of *Radix Curcumae* against cardiovascular diseases[151]
Machine learning algorithm	Pain disorders	To predict the mechanism of action of herbal phytocompounds at the atomic level against algesia[152]
Other fields of medicinal plant research		
Convolutional neural network	Compound-target interaction of natural products	To generate scoring energy functions of proteins and their ligands. Has an image processor to assist protein-ligand binding. To optimize the scoring for stable conformations[153]
Image-based convolutional neural network	TCM	To demarcate diverse species of *Zanthoxyli pericarpium* for aiding traditional Chinese medicines[154]
ML algorithm	Biomass production	To predict the accumulation of biomass in microalgal suspension[155]

1102 **Table 3.** Case studies on AI algorithms used for microbial research tasks

Task	AI/ML Tool	Features
Identification of microbes		
MALDI/TOF	SpeDE	Identifies microbes based on unique characteristics rather than universal similarity[156]
	IDBac	A bioinformatic tool that amalgamates integral protein and its metabolite for detection[157]
Genome mining		
Databases on Biosynthetic gene clusters	antiSMASH database	Most common and inclusive source on secondary metabolites[30]
	Bactibase	An open-access database exclusive for of bacterial antimicrobial peptides[158]
	MIBiG	Large curated database on biosynthetic gene clusters[159]
	IMG-ABC	Database on biosynthetic lab clusters retrieved from metagenomes and microbial genomes[160]
BGC identification from genomes	antiSMASH database	Detects biosynthetic gene clusters based on profile Hidden Markov Models[30]
	PRISM	Identifies biosynthetic gene clusters, biological activity and cheminformatic dereplication[161]
	ARTS	To prioritize the most capable gene cluster that encodes antibiotics with novel mode of action[162]
BGC identification from metagenome	MetaBGC	Algorithm used to detect BGC in data of metagenomic sequencing directly[163]
	DeepBGC	A deep learning approach based on genome mining to predict BGC clusters[164]
Metabolite production and expression		
Elicitor screening	MetEx	UPLC-MS based high throughput screening of elicitors[165]
Natural product dereplication and structure elucidation		
Databases	DNP	It contains the physical and chemical properties of more than 226,000 natural products[63]
	NPEdia	Exclusive database for natural products[166]
	StreptomeDB	Contains chemical and biological data on natural products isolated from streptomyces[64]
	MarinLit	Exclusive database on marine natural products[167]

	NuBBE DB	Contains over 2200 chemical structures of diverse natural molecules acquired from various Brazilian habitats[168]
	CMNPD	Inclusive and organized data on natural products derived from marine sources Contains over 32000 structures of marine compounds along with its physical, chemical and ADME properties[169]
	NaPLeS	Free access MySQL database of natural compounds that process NP-likeness score of huge compound libraries[170]
	UNaProd	Online database of natural compounds that was traditionally used as medicine by Iranians. Contains data on more than 2696 natural compounds derived from plants, animal and minerals[171]
MS-based dereplication	DEREPLICATOR	Integration of molecular network with dereplication[73]
	SIRIUS-4	To identify molecular structures from MS[172]
	GNPS	Online database that contains sample information for untargeted MS[69]
NMR-based structure elucidation	NP-MRD	Large NMR database containing more than 41,000 natural products[78]
	DEEP picker	Deconvulutes the complicated 2D NMR spectra based deep neural network[79]

Table 4. Identification of targets and prediction of bioactivity of natural products using AI/ML

Tool	Features	Application
BANDIT	Bayesian based ML approach	Prediction of drug binding targets. Predicted more than 4000 molecules with 90 % accuracy Validation of 14 new microtubule inhibitors[173]
deepDTnet	DL tool	Identifies target from heterogenous networks[2]
ML-classifier	ML based tool	Utilizes genome mining for prediction of biological activity Predicts the antifungal and antibacterial activity of natural products based on BGS with 80% accuracy[174]
SPiDER	ML based tool	Target identification for drugs and computer-generated scaffolds. Identification of novel fenofibrate related compounds[121]
SuperPred	Prediction webserver	Classification of drug and prediction of target by considering 2D, 3D and fragment similarity. Alternative to chemoproteomics[175]
KronRLS	ML algorithm	Prediction of drug-target interaction[176] based on features and similarity
DeepDTA	DL algorithm	Prediction of drug target based on 3D structure of protein Used to identify therapeutic efficacy of antiviral medicines against SARS-CoV-2[177]
PADME	DL algorithm	Analyzes drug-induced transcriptome data for prediction of drug target interaction[178]
DeepAffinity	DL algorithm	Uses both CNN and RNN to predict the binding affinity of drug to target[84]
DeepTox	DL algorithm	A deep learning tool that predicts toxicity[175]

Figure 1. AI as a tool for mining plant and microbial secondary metabolites

Figure 2. Virtual screening vs conventional computer-aided discovery of natural products

Virtual screening (Selection of bioactive NPs by virtual screening includes three major sequential steps: **Library preparation** -The bioactive metabolites are obtained from the compound library and then checked for correction of structures, generation of conformers, and file format conversion. **Virtual screening** -Structure-based and ligand-based pharmacophore modeling, Similarity search-based 3D shape and fingerprints, docking, molecular filters, and molecular simulation. **Experimental validation** of selected compounds by *in-vitro* and *in-vivo* assays).

Figure 3. Applications of AI in Natural product drug discovery:

1- Genome mining (PRISM, BAGEL, antiSMASH, ARTS); 2-Selection and screening of natural products (IDBac, SPeDE, MALDI-TOF); 3-Dereplication of natural products (DEREPLICATOR, GNPS, SIRIUS-4); 4-Classification of metabolites; 5-Interpretation of structure (DEEP picker, DP4-AI, NAPROC-13); 6-Prediction of physicochemical properties (OpenChem, ChemSpider, PCLIENT, E- BABEL); 7-Prediction of bioactivity (ML-classifier, Deep affinity, DeepTox, PADME, KronRLS) ; 8-Identification of Target (BANDIT, SPIDER, SuperPred, DEcRyPT).

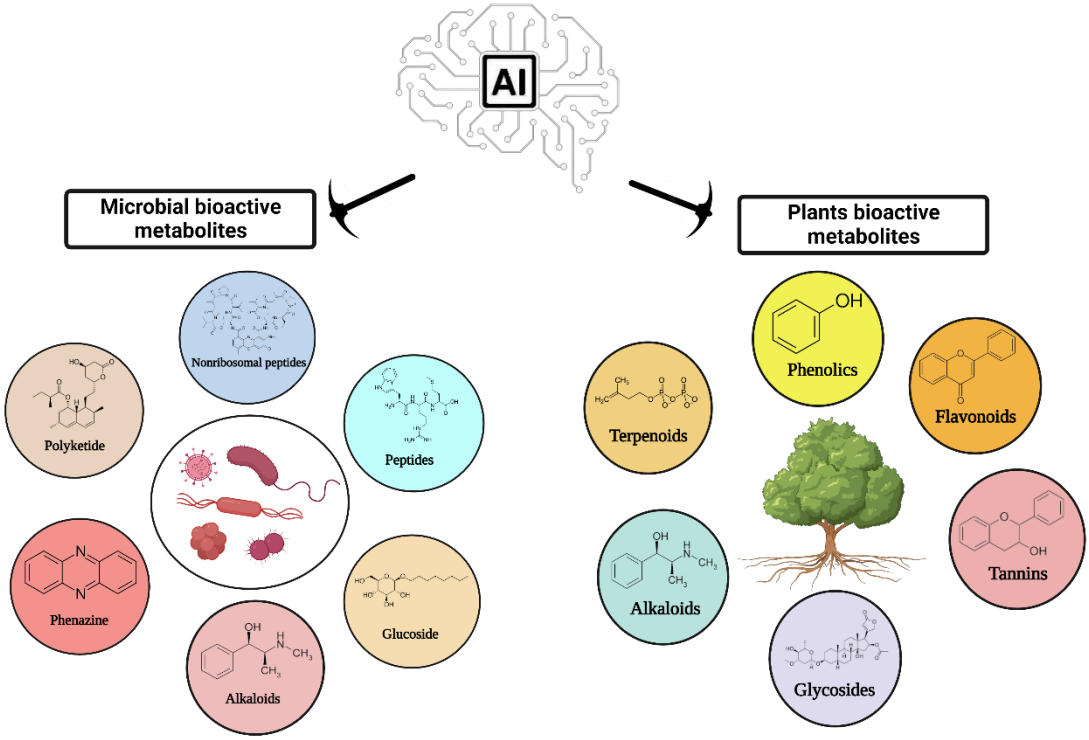

Figure 1. AI as a tool for mining plant and microbial secondary metabolites

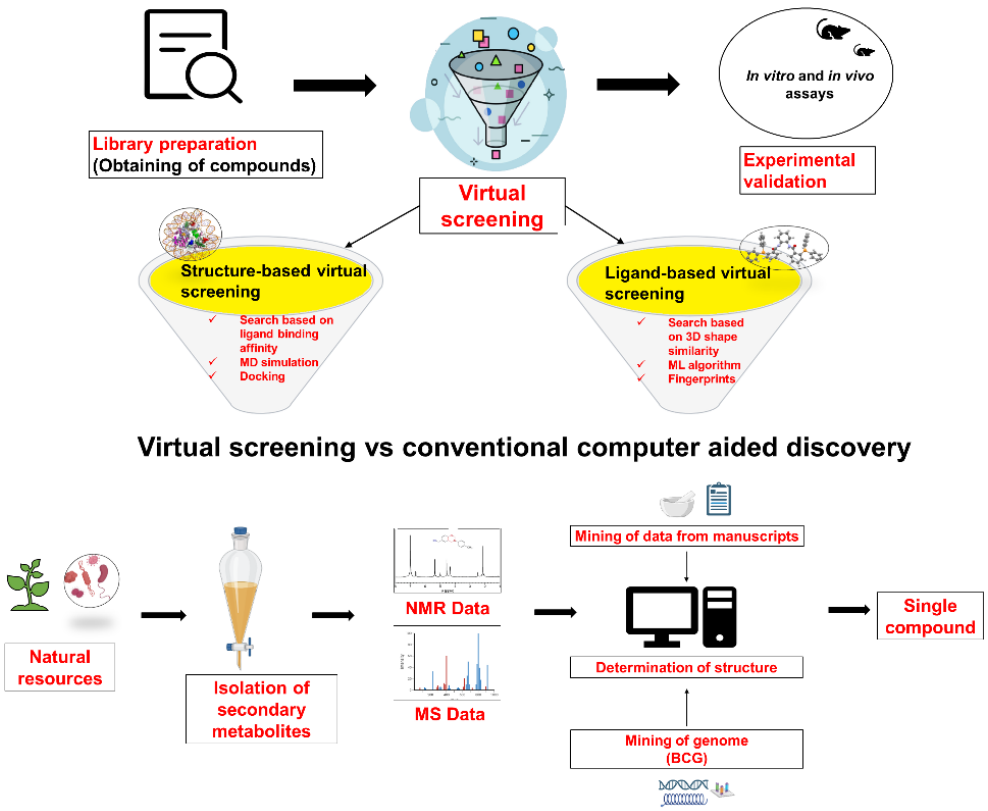

Figure 2. Virtual screening vs conventional computer-aided discovery of natural products

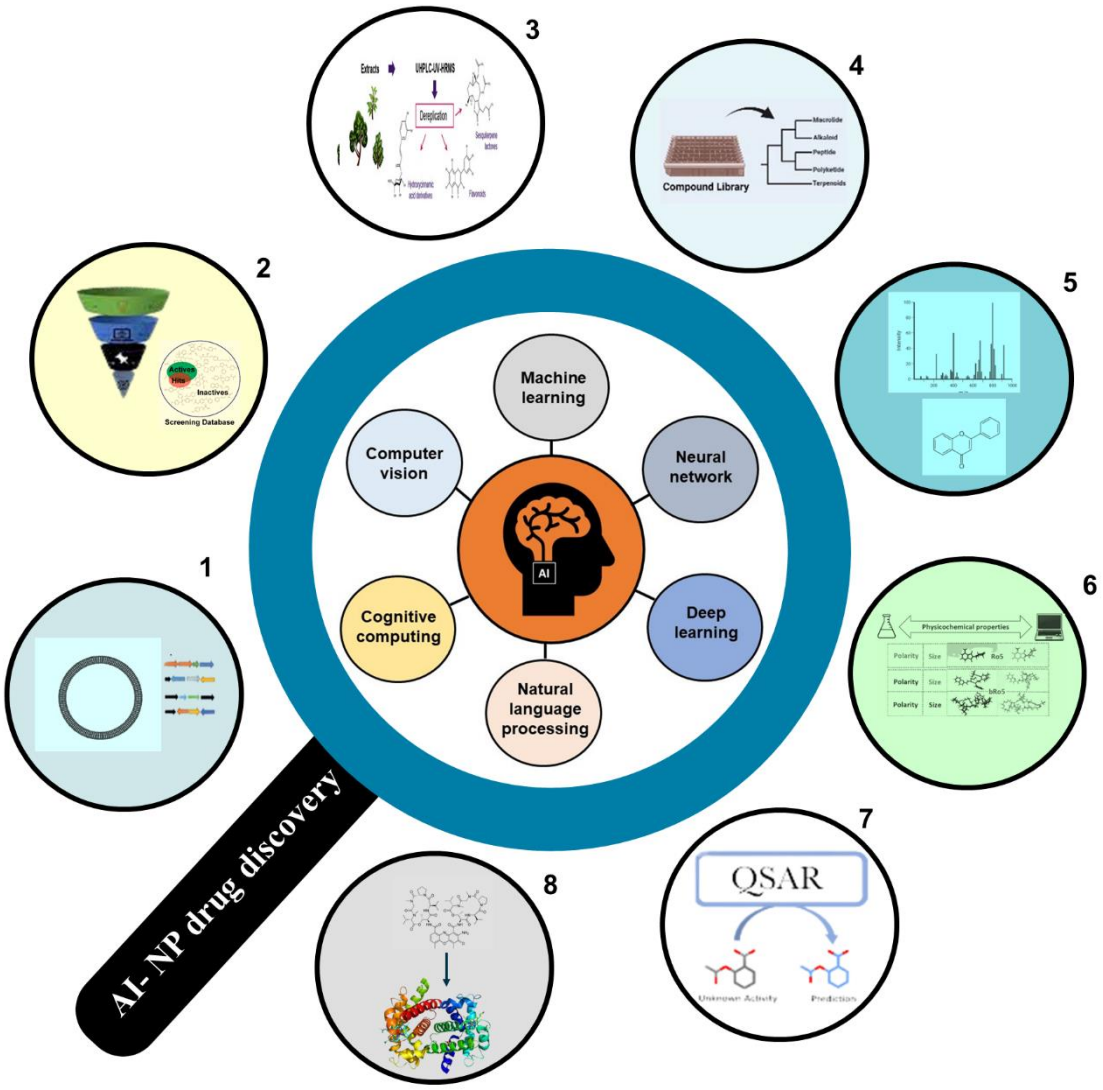

Figure 3. Applications of AI in Natural product drug discovery