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

Janani Manochkumar and Siva Ramamoorthy*

School of Bio Sciences and Technology, Vellore Institute of Technology, Vellore 632 014, India

Advancements in genome mining, high-throughput sequencing and experimental techniques have generated an enormous amount of data on natural products. This has led to the design and development of advanced machine learning (ML) and artificial intelligence (AI) algorithms which have simplified the search for novel natural products in the 21st century. These algorithms could effectively analyse the chemical structure of natural products and predict their biological function. They could also effectively analyse large sets of data in a sophisticated manner. In this context, this article reviews the various AI/ML algorithms employed in natural products-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 products research have also been discussed.

Keywords: Artificial intelligence, dereplication, drug discovery, genome mining, machine learning, natural products.

ARTIFICIAL INTELLIGENCE (AI) utilizes computers for performing complicated tasks, analysing 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 being 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 (NPs) efficiently[2].

NPs have garnered proliferating attention in drug discovery as they are bio-friendly, less toxic and evolve collaboratively along with their active sites[3,4], The high variation in the molecular structure and physico-chemical 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 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 flavour, cosmetics, and pharma industries has been predicted to reach USD 59.4 billion by 2025 (ref. 10). Plants have also been used for the treatment of several diseases worldwide[11]. Based on this evidence, researchers are now focusing their studies on the potential of plants and microbes to render NPs 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 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]. Repeated identification of the already identified NPs, high demand for resources, increasing manual efforts, and time-consuming tasks have restrained the interest of scientists and industries in NPs 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, with low cost, less labour and consuming less time[17]. In addition, some structural scaffolds derived from various

*For correspondence. (e-mail: siva.ramamoorthy@gmail.com)

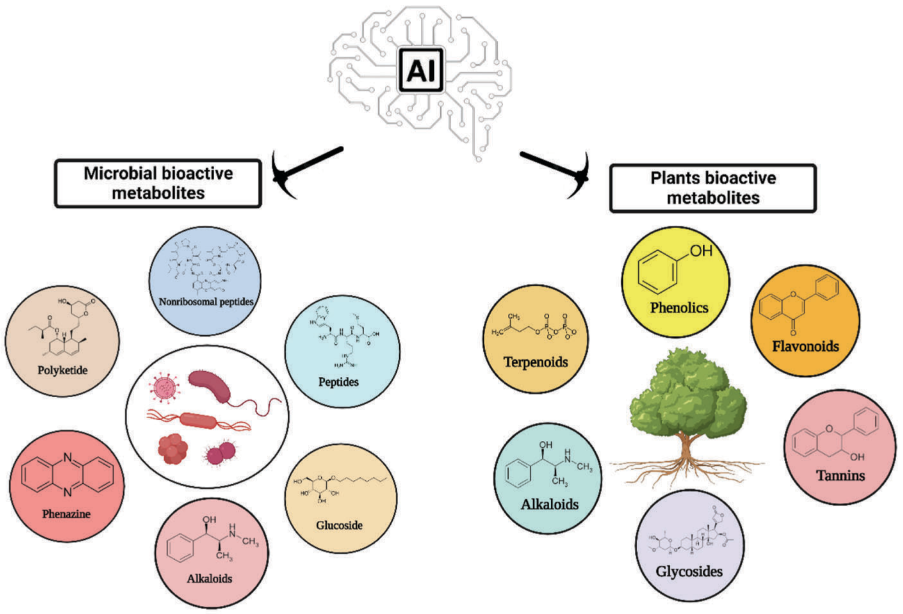

Figure 1. Artificial intelligence (AI) as a tool for mining plant and microbial secondary metabolites.

classes of NPs, such as alkaloids, phenylpropanoids, polyketides and terpenoids, have served as an inspiration to design new drug candidates[18]. Figure 1 illustrates the concept of AI in mining the various classes of plants and microbial secondary metabolites.

Role of computational methods in virtual screening of bioactive metabolites

Virtual screening strategies have 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 reducing 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 physico-chemical, pharmacodynamics 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 more 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 metabolites against a specific target[21]. These methods should examine the chemical space of NPs in order to identify the bioactive class of compounds and structural scaffolds of known compounds. Some of these methods apply less restraining structural similarity cut-off and modelling of putatively derived structures of NPs[22]. The 3D structure depicts the configuration of structure and binding sites of ligands. Therefore, virtual screening strategies have emerged as an essential part of the discovery of novel bioactive metabolites[16]. Figure 2 depicts the overflow of the virtual screening strategy for identifying bioactive metabolites along with conventional computer-aided discovery of NPs.

Ligand-based virtual screening

The ligand-based virtual screening (LBVS) approach uses a set of compounds with experimentally demonstrated bioactivity as the starting point and solely relies on the analysis of inherent features of the compound, including physico-chemical, electronic, structural and topological characteristics that are related to its bioactivity[23]. Quantitative structure-activity relationship (QSAR), ML algorithms, ligand-based pharmacophore modelling, cheminformatics filters, and similarity searches based on structure, fingerprint and 3D shape are some of the computer-generated strategies utilized in LBVS[24].

Structure-based virtual screening

In contrast, the structure-based virtual screening (SBVS) strategy uses data on the recognition site of the ligand in structure of the receptor as the starting point, which includes the binding affinity of ligands, conformation of the receptor, charge on the surface of the molecule and configuration of

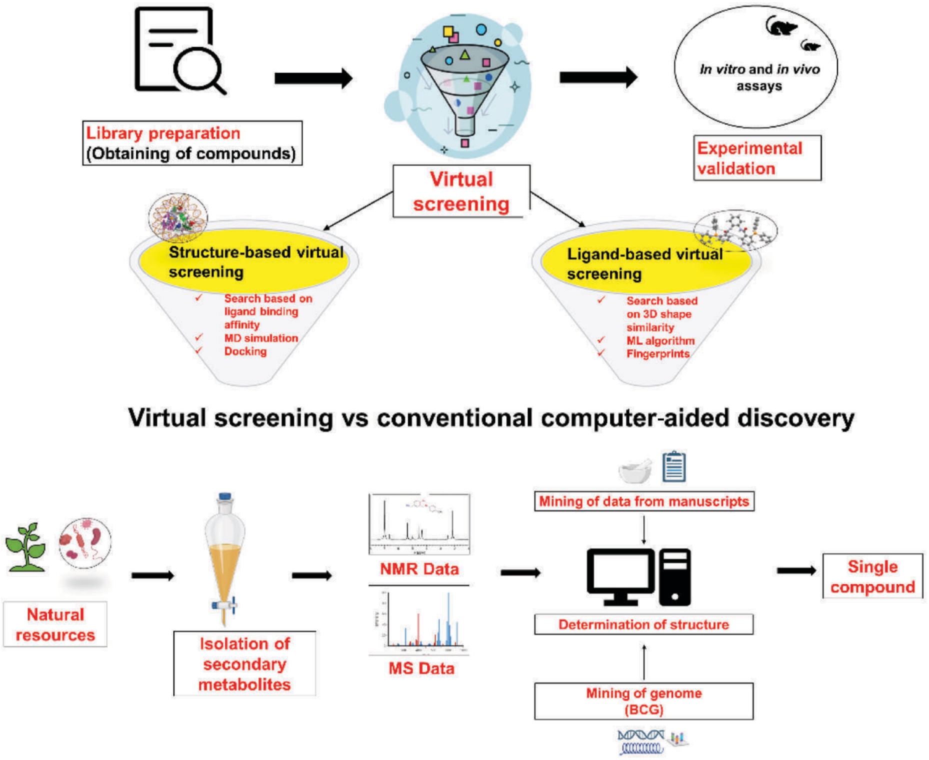
Library preparation
(Obtaining of compounds)
Virtual screening
In vitro and in vivo assays
Experimental validation
Structure-based virtual screening
Search based on ligand binding affinity
MD simulation
Docking
Ligand-based virtual screening
Search based on 3D shape similarity
ML algorithm
Fingerprints
Virtual screening vs conventional computer-aided discovery
Natural resources
Isolation of secondary metabolites
NMR Data
MS Data
Mining of data from manuscripts
Determination of structure
Mining of genome (BCG)
Single compound

Figure 2. Virtual screening versus 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 modelling, 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).

molecules present in the binding site[25]. These techniques require the 3D structure of the receptor to be fully understood and, ideally, to be in intricate complex with the bioactive substance. Molecular dynamics simulation, structure-based pharmacophore modelling, and molecular docking are a few of the computational techniques used in the SBVS methodology[21]. Virtual screening techniques are currently a crucial component in the design and discovery of novel bioactive molecules. Therefore, the applications of SBVS strategy been increased in academics as well as industries[16].

AI-assisted virtual screening

AI has made immense progress in speeding-up the identification and screening of bioactive metabolites with commercial applications. AI along with molecular modelling and cheminformatics have improved the efficiency of virtual screening strategies, thus allowing the users to explore the extremely diverse chemo-structural topographies of NPs[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 high success rate in identifying novel metabolites in 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 their unique features[16]. Table 1 lists some of those AI tools used for virtual screening and various fields of drug discovery.

Applications of AI in NPs-based drug discovery

The distinct properties of NPs still confound computational experts as well as research scientists. As expected, scientists have developed several 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 developed 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]. Those algorithms could also predict the function of NPs from biosynthetic gene clusters (BGCs)[29]. For instance,

Table 1. Application of artificial intelligence/machine learning (AI/ML) tools in virtual screening and various fields of natural products (NP)-based drug 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	LSA employs a conventional similarity and substructure match algorithm to align the structure for virtual screening	A structure-based alignment tool for virtual screening of pharmaceutical compounds[98]
	LigGrep	ML	Filtration of docked models for enhancing the hit ranks of virtual screening[99]
	Trix X	ML	Structure-based molecular indexing tool enabled for the fastest and largest virtual screening[87]
	Drug finder	ML	*In silico* virtual screening tool intended for validation while screening the compounds[100]
	LS-align	ML	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 binding of the drug to the target[102]
Drug design and discovery	ChemDes	Chemopy, Pybel	An integrated on-line software used for the computation of molecular descriptors and fingerprints[103]
QSAR modelling	ChemGrapher	Deep learning	Recognizes chemical compounds using an optical graph[104]
	ChemSAR	ChemoPy	Generates molecular SAR model benefitting cheminformatics[105]
	ANFIS	Neuro-fuzzy modelling	A QSAR model used for the evaluation of physico-chemical characteristics of chemical molecules[106]
	OntoQSAR	ML	Interpretation and evaluation of biological and chemical data[107]
Drug repurposing	GIPAE	Gaussian interaction profile	A drug repositioning tool used to recognize novel signs in existing drugs[108]
	DrugNEt	ML	Integrates heterogenous information by prioritizing the interaction of drugs and target[109]
Drug repurposing	RCDR	Collaborative filtering model	Gives high preference for the candidate drugs against diseases[110]
	DrPOCS	ML	It predicts potential associations between drugs and diseases with matrix completion and projection onto convex[42]
	Pred-binding	Vector machine	Predicts the binding of proteins with ligands on a large scale[111]
Physico-chemical properties and bioactivity prediction	CSM-lig	ML	A web-based tool to compare and evaluate affinity of proteins to small molecules[112]
	mCSM-AB	ML	Quantifies mutational effects on the affinity of proteins to small molecules in genetic diseases[113]
	Chembranch	ML	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	ML classification model	Identifies and recognizes pharmacokinetic properties[117]
	RosENet	Convolutional neural network	Predicts the accurate binding efficiency of proteins with ligand[118]
	DeepPurpose	Deep learning	Open library available for predicting the interaction of drug for target[119]
Molecular target prediction	PASS	NB	Predicts the bioactivity, mechanism of action and pharmaceutical properties[120]
	TiGER	Multiple self organizing maps (SOMs)	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 to evaluate the probable side effects[121]
	SEA	Kruskal algorithm	Prediction of chemical similarity of proteins to ligands[122]

the progression of NPs-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 (ref. 32). These developments increase the availability of chemical structures of NPs and provide an opportunity for the researchers to link these structures to the relevant functions using AI/ML algorithms[28]. Therefore, ML and AI algorithms have gradually paved the way for prominent research in the field of NPs-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 are discussed below.

Case studies on the use of AI/ML algorithms on plants

Plants have always been the centre of attraction owing to their numerous beneficial effects to humans[33]. The enormous advancement in plant-based research provides a testament to the vast array of limited secondary metabolites synthesis[34]. Nevertheless, several biotic and abiotic factors affect the biosynthetic pathway of secondary metabolites production. Therefore, lot of time, cost and manual efforts are needed to screen these novel bioactive metabolites. Considering this, an effective alternative is 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]. Data from *in vitro* experimental research were utilized in computational modelling to study the impact of various factors in predicting the involvement of phytohormones in plant growth[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 nutritional imbalance in plants. Hence, the factors affecting plant growth could be optimized[37]. Recently, AI along with microfluidics has been used to enhance 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 the Indian Western Himalayan Region[40]. Similarly, seven ML models were used to model the habitat suitability for the medicinal plant *Ferula gummosa* in mountainous regions to avoid extinction in the future[41]. They can also be used for the identification of different 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 photosynthetic pigment studies have been reviewed and diverse strategies on how to employ ML in enhancing crop yield have been discussed[43]. ML was used to increase the bioactive metabolites synthesis in plants on a large scale for commercialization purposes[44]. ANN organizes plants based on morphological characteristics like size, colour and the dimension of leaves. ML uses ANN and square-support vector machine (SVM) for predicting the interconnection between photodissociation and its bioactivity[33]. Table 2 shows 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.

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

NPs from microbes – selection and screening: The preliminary step in NPs discovery is selection of the organism. Among various microbes, actinomycetes have been overmined as a significant source of therapeutic compounds, which has led to the repetitive discovery of known compounds and the lack of identification of novel compounds[2]. Even though the whole process of extraction of NPs is challenging and laborious, cautious exploration of unexplored sources enhances the chance of finding novel scaffolds[2]. The conventional method of isolation of NPs 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 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, *Bacillus subtilis* has been categorized 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 the marine environment[49]. In the future, integration of AI/ML techniques with MALDI-TOF could be a possible method to enhance 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].

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].

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 (SVM)	*Chrysenthenum morifolium*	AI was used to estimate the total flavonoid and polysaccharide contents[123]
Artificial neural network (ANN)	*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 the explant[126]
Neuro-fuzzy logic	*Prunus armeniaca*	To predict shoot multiplication using hormones, nutrients and vitamins[127]
Intelligent image analysis using ANN	*Solanum tuberosum*	To predict the characteristic features of the 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 ANN	*Cuminum cyminum*	To predict the formation of callus, and determine its volume and fresh weight[130]
Backpropagation neural network	*Chlorophytum borivilianum*	To predict the development of shoots in a fermentor and fresh weight of plantlets[131]
Multivariate adaptive regression splines algorithm	*Fragaria ananassa*	To predict the nutrients required for the culture of strawberry and to predict the responses like shoot quality, multiplication and leaf colour responses[132]
Multilayer perception	*Pinus taeda*	To predict the impact of nitrogen source on organogenesis of the shoot[133]
Multilayer perception-based modelling	*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 chemical composition for the production of callus[136]
Neurofuzzy logic	*Prunus* sp.	To predict the best medium for rootstock micropropagation[137]
Regression analysis and ANN analysis	*Pyrus communis*	To predict the *in vitro* culture medium macronutrients for rootstock propagation, and analyse 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 conditions[139]
Algorithm	**Target**	**Applications**
Drug design and discovery		
ML algorithm	Drug-induced liver injury	To predict the upsurge/reduction in the efficacy of multiple drug interactions, and evaluate the inhibition rate of drugs[140]
ML algorithm – random forest (RF) and SVM	Drug–ADR association	To identify different adverse drug reactions, and predict the intensity of outcome and the developed ML model could predict the death due to adverse drug reactions with 91% accuracy[141]
SVM	*Schizophrenia* and depression/anxiety	Drug repositioning – to predict indications for a disease based on drug expression profiles[142]
Supervised learning (SVM)-neural network	Drug–ADR association	To predict adverse drug interactions[143]
ML 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, RF, and SVM algorithms	Drug repurposing	To explore the unknown medicinal properties of herbal bioactive compounds; has identified novel indications for 20 known drugs and 31 herbal compounds[144]
Regularized least square (semi-supervised based new modelling)	Drug repurposing	To identify the novel pharmacological significance of existing drugs for viral infections[145]
ML approach	Drug discovery	To elucidate the medicinal value of Xiaoxuming decoction to be utilized as a neuroprotective agent[146]

(*Contd*)

Table 2. (*Contd*)

Algorithm	Target	Applications
Ontology-based AI model	AI-based traditional Chinese medicine (TCM) screening	To predict the side effects of prescription[147]
AI in disease treatment		
Neuro-fuzzy	Disease treatment	To evaluate the pharmacological aspects 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]
ML 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. There is 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 medicine[154]
ML algorithm	Biomass production	To predict the accumulation of biomass in microalgal suspension[155]

In spite of the huge diversity of NPs, their relevant BGCs are extremely conserved in microorganisms. 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 the genome and characterizing them for analysis. ML algorithms aid in analysing big data for the prediction of these BGCs and reputed determined structures[52].

Table 3 lists the AI algorithms employed in various fields of microbial research. Using genome mining, gladiolin was extracted from *Burkholderia galdioli* in a cystic fibrosis patient[53]. ML and deep learning (DL) approaches have also contributed to the identification of mysterious BGCs, viz. 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 also possible to identify novel compounds from human microbiota using the hidden Markov model (HMM) algorithm. It identifies BGCs from metagenome samples[56]. 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. Now with the advancement of AI/ML and metagenome, NPs can be predicted directly from biotic and environmental sites[56].

Metabolite expression and synthesis: Using bioinformatic tools and genome sequencing, it has been predicted that *Myxococcus* and *Streptomyces* possess huge BGCs of secondary metabolites. However, 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]. Using 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 employed for the activation of these silent, unexpressed BGCs in several organisms. Here imaging mass spectrometry (IMS) was used to screen the elicitors for inducing secondary metabolite synthesis. The integration of this technique with laser ablation coupled electrospray ionization mass spectroscopy, led to the identification of a novel glycoprotein from *Amycolatopsis keratiniphila*[2].

AI/ML in the dereplication of NPs: Many drugs were discovered during the golden age of the progress of NPs, which are 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 are enriched with several compounds, the dereplication approach could possibly reduce repetition and offer data on novel compounds. Therefore, engagement of highly accurate ML/AI tools could make this crucial task easier. Conventionally, dereplication was done

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

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 bacterial antimicrobial peptides[158]
	MIBiG	Large curated database on biosynthetic gene clusters[159]
	IMG-ABC	Database on biosynthetic laboratory 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 metagenomes	MetaBGC	Algorithm used to detect BGC in the 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 products dereplication and structure elucidation		
Databases	DNP	Contains the physical and chemical properties of more than 226,000 natural products[63]
	NPEdia	Exclusive database on natural products[62]
	StreptomeDB	Contains chemical and biological data on natural products isolated from streptomyces[64]
	MarinLit	Exclusive database on marine natural products[166]
	NuBBE DB	Contains over 2200 chemical structures of diverse natural molecules acquired from various Brazilian habitats[167]
	CMNPD	Inclusive and organized data on natural products derived from marine sources contains over 32,000 structures of marine compounds along with its physical, chemical and ADME properties[168]
	NaPLeS	Free access MySQL database of natural compounds that process NP-likeness score of huge compound libraries[169]
	UNaProd	On-line 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[170]
MS-based dereplication	DEREPLICATOR	Integration of molecular network with dereplication[73]
	SIRIUS-4	To identify molecular structures from MS[171]
	GNPS	On-line 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]

using HPLC coupled with a UV/photodiode array (PDA) detector which has integral library databases[60]. However, this could not give data on the structure, and hence instruments with advanced multispectroscopic detectors are needed for capturing the additional spectral characteristics of the compounds[2].

AI/ML in mass spectrometry-assisted dereplication: Mass spectrometry (MS) is extensively used for dereplication of the NPs as it is accurate, rapid and highly sensitive. MS has the added advantage of retrieving huge amounts of structure-related data even from small amounts of 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 can be achieved by liquid chromatography-mass spectrometry (LC-MS), but the challenging part is data analysis. For this, scientists have to screen and search UV spectra, mass spectra and micro-organisms data in various databases[2]. Therefore, the use of ML techniques will be a possible way to analyse and identify natural products based on their spectral data without searching the databases manually.

The major disadvantage concerned with MS is 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 labour-intensive manual task. Hence, ML algorithms have been used recently to evaluate these hugely resolved MS spectra 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) has been used to dereplicate novel bioactive metabolites from diverse sources. It evaluates 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 analysed by MN[69]. This integrated approach is known 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 the 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 (ref. 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 has been formulated. A novel inhibitor of serine protease, rivulariapeptolides was discovered using this approach[72]. It could be a significant method 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 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 structure of unknown compound. However, this became possible with ML. For instance, SIRIUS 4, a web-based tool uses SVM for identification of the structure of compounds[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. Later, deep neural network (DNN) was developed for the prediction of unidentified metabolites for which no structure or spectra-related data were available[75]. Another tool, MS2DeepScore predicts the unknown compounds based on MS similarity and identifies them by grouping[69]. Hence, using MN for dereplication would prove successful and therefore could be utilized in the future in combination with ML for interpretation of the structure of novel compounds[2].

Dereplication of NPs using NMR: Interpretation of metabolite structure is another crucial task. Even though unambiguous and precise interpretation of structures was provided by X-ray crystallography, its application is limited as it requires a single crystal[76]. On the other hand, nuclear magnetic resonance (NMR) is a widely used spectroscopic technique which infers structural data depending on the spectrum[77]. NMR-based databases like CHNMR-NP, NAPROC-13, BMRB and Spektraris have many disadvantages and hence do not aid in the NPs 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]. The development of this database is ongoing and in the future, it will allow efficient elucidation of structure and also dereplicate in an automatic manner. SMART 2.0 analyses and characterizes a 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 the Biological Magnetic Resonance Data Bank (BMRB) and Human Metabolome Database (HMDB) were analysed. This tool could identify these molecules from an amalgamated mixture with 88% accuracy.

Recently, DEEP picker, an AI tool based on DNN has been developed for the analysis of 2D NMR spectra[79]. The ML technique has been used for the prediction of various

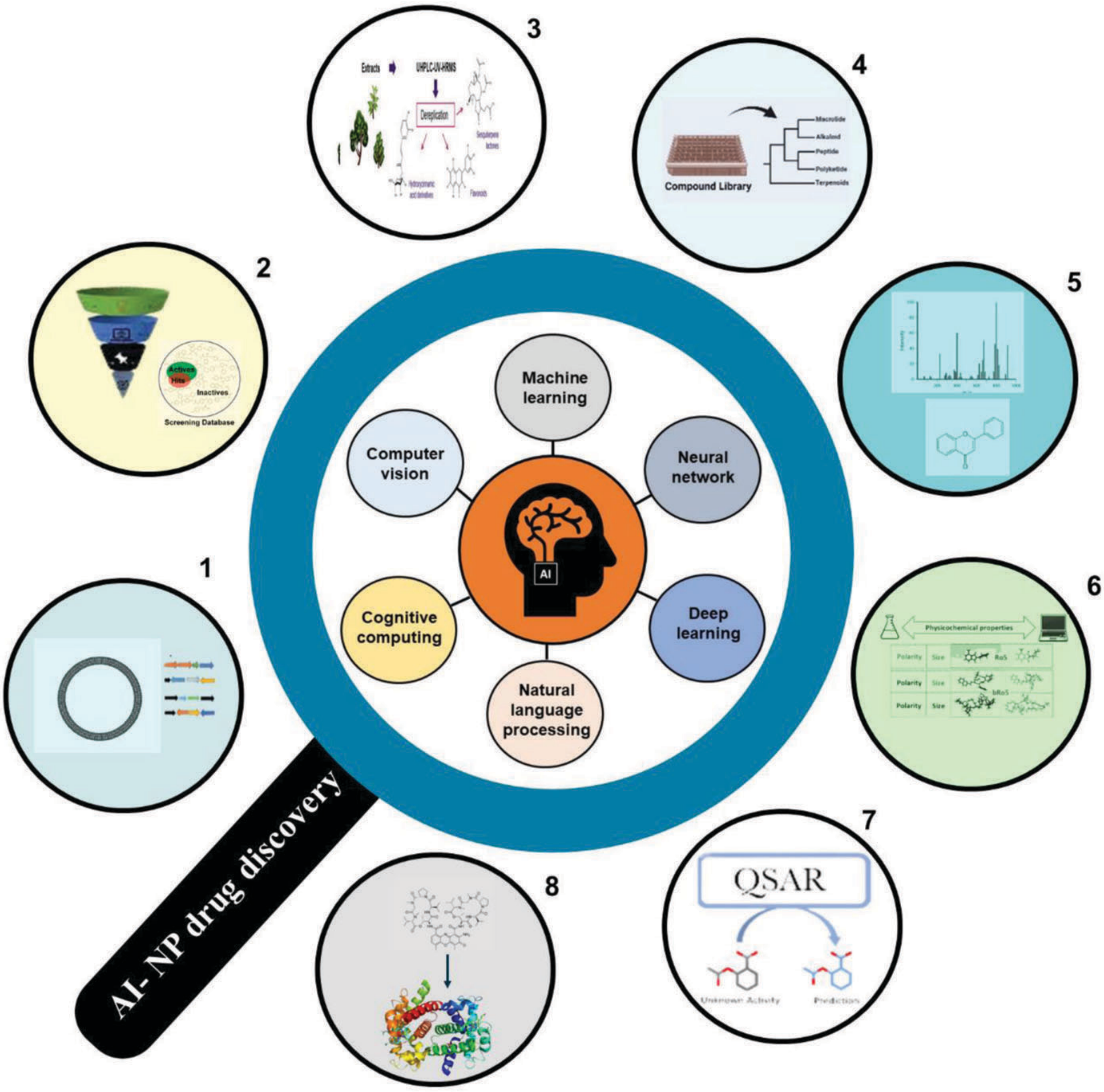

Figure 3. Applications of AI in natural products (NPs) drug discovery. 1, Genome mining (PRISM, BAGEL, antiSMASH, ARTS). 2, Selection and screening of natural products (IDBac, SPeDE, MALDI-TOF). 3, Dereplication of natural products (DEREPLICATOR, GNPS, SIRIUS-4). 4, Classification of metabolites. 5, Interpretation of structure (DEEP picker, DP4-AI, NAPROC-13). 6, Prediction of physico-chemical 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).

classes of NPs from ^{13}C-NMR spectral data[80]. As far as dereplication is concerned, high-resolution mass spectrometry (HRMS) is preferred over NMR owing to its high sensitivity. However, 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 detection of the chemical characteristics correlating with biological activity prior to the extraction of compounds. Hence, this tool identified novel lanostane triterpenes from the fungal extract of *Fomitopsis pinicola*[83].

Other applications of AI/ML tools

Prediction of bioactivity and identification of target using AI/ML

Generally, the bioactivity of NPs is identified depending on the phenotypic or screening by high-throughput techniques owing to the diverse structures and extensive chemical gaps[84]. On the other hand, experimental identification of targets has been conventionally performed using chemical proteomics and genomics. However, validation of the targets is difficult, time-consuming and requires more effort[85]. Computational strategies, in turn, could reduce these constraints and limit the search for target screening[86]. Figure 3

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

Tool	Features	Applications
BANDIT	Bayesian-based ML approach	Prediction of drug binding targets. Predicted more than 4000 molecules with 90% accuracy. Validation of 14 new microtubule inhibitors[172].
deepDTnet	Deep learning (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[173].
SPiDER	ML-based tool	Target identification for drugs and computer-generated scaffolds. Identification of novel fenofibrate-related compounds[121].
SuperPred	Prediction webserver	Classification of drugs and prediction of targets by considering 2D, 3D and fragment similarity. Alternative to chemoproteomics[174].
KronRLS	ML algorithm	Prediction of drug–target interaction[175] based on features and similarity.
DeepDTA	DL algorithm	Prediction of drug–target interaction based on 3D structure of the protein. Used to identify therapeutic efficacy of antiviral medicines against SARS-CoV-2 (ref. 176).
PADME	DL algorithm	Analyses drug-induced transcriptome data for prediction of drug–target interaction[177].
DeepAffinity	DL algorithm	Uses both convolutional neural network and recurrent neural network (RNN) to predict the binding affinity of drug to target[84].
DeepTox	DL algorithm	A DL tool that predicts toxicity[174].

depicts the applications of AI algorithms in various fields of NPs based drug discovery.

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 successfully for the identification of NP targets[2]. At present, the advanced features of AI algorithms help improve the prediction of binding affinity by considering the similarity between the drug compound and its relevant target. Table 4 lists the widely used AI/ML tools for target identification and bioactivity prediction. 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 needs to be conducted[87].

Prediction of physico-chemical properties

It is clear that each compound possesses diverse physico-chemical properties like solubility, degree of ionization, partition and permeability coefficient that may interfere with the pharmacokinetic qualities of a molecule and drug–target binding effectiveness[88]. To assist with this, many AI-based techniques for predicting the physico-chemical characteristics of chemical compounds have been developed. Molecular fingerprinting, SMILES format, Coulomb matrices and potential energy measurements are among the AI-based tools[89]. A QSAR model was developed[90] to forecast six different physio-chemical characteristics of eco-friendly agents taken from the US Environmental Protection Agency data. Later, six AI-based systems for the prediction of chemical absorption in the human digestive tract were developed. These include SVM, *k*-nearest neighbour, probabilistic neural network, ANN, partial least square (PLS) and linear discriminate model. SVM has a greater accuracy at 91.54% than the other models mentioned above[91]. An ML-based model was developed to predict the physico-chemical characteristics of foreign chemicals like bioconcentration factors, solubility in water, octanol–water partition coefficient, melting and boiling point, and vapour 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 developed. The quantitative structural toxicity of tyrosine derivatives intended for effective and safe inflammatory treatment was further predicted using ORISIS Property Explorer[92]. 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 physico-chemical features[87].

Challenges and limitations in NPs-based drug discovery

Virtual screening–exclusion of compounds

In comparison with the application of conventional methods for the extraction of novel bioactive metabolites, computational strategies are known to be prognostic, low-cost and beneficial. Nevertheless, regardless of these advantages, they also have challenges and limitations, and most of them are susceptible to bias[93]. Analysis of diverse chemical structures and bioactivity of NPs by similarity-based computational tools provides biased data as it has a postulation that novel compounds might be similar to well-known bioactive compounds[93]. This hypothesis leads to errors in the development 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 minimize the investigation of novel chemical compounds with unique biological activity.

Generation of inaccurate data

The major challenge associated with NPs-based drug targets is identifying the mechanism of action and their relevant side effects, which is an expensive and time-consuming process[94]. In spite of several advantages, the 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. However, this requires a prolonged time-period and it is an expensive process. It also requires validation of the molecule by *in vitro* and *in vivo* experimental studies for assessing toxicity[2]. Hence, computational toxicology could be used for screening several compounds simultaneously, thus reducing the time of performing animal studies. However, this could also generate inaccurate data[2].

Molecular featurization (technical issue)

Over past few decades, infinite datasets on molecular structure have been developed which provide 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, these algorithms can 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 structure 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 physico-chemical properties and complex structures. On the other hand, these molecular featurization tools are designed and optimized for targeting smaller molecules. Hence, current featurization tools cannot be used when the structural and physico-chemical properties of NPs deviate from those of smaller molecules[28]. First, the performance of existing featurization tools could be examined with different NPs having complex structures. Based on these data, new featurization tools may be developed which will tailor structurally complex NPs in a better way.

Interpretation of predicted data

The next challenge lies in the interpretation 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 are involved. Therefore, the predicted outcomes from ML/AI models should be explicable for a proper understanding of the biochemical properties of NPs[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 clues regarding the mechanism of action of NPs.

Conclusion and future prospects

NPs have encouraged several 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 NPs drug discovery by assisting in finding and elucidating the bioactive structures, and capturing their molecular patterns for target prediction. In this study, we have extensively reviewed the latest AI/ML algorithms employed in various fields of NPs-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 their exploration. Yet, AI algorithms cannot 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 analyse different classes of metabolites efficiently. Therefore, the design and development of an AI tool that could analyse enormous amount of data and different classes of secondary metabolites efficiently could contribute to fruitful outcomes in the future.

There exists a significant gap between wet laboratory (experimental) and computational research. Researchers working on NPs and computational experts could collaborate for successful characterization of the functions of NPs. Researchers could elaborate upon the complicated physico-chemical properties of NPs, whereas experts in computers could develop suitable AI tools and featurization methods for better prediction. Finally, researchers could analyse and validate the 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.

Conflict of interest: The authors declare that there is no conflict of interest.

1. Jiménez-Luna, J., Grisoni, F. and Schneider, G., Drug discovery with explainable artificial intelligence. *Nature Mach. Intell.*, 2020, **2**(10), 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., Plant dyes. In *Industrial Crops and Uses*, CABI, Wallingford UK, CABI, 2010, pp. 349–357.
4. Siva, R., Doss, F. P., Kundu, K., Satyanarayana, V. S. V. and Kumar, V., Molecular characterization of bixin – an important industrial product. *Ind. Crops Prod.*, 2010, **32**(1), 48–53.
5. Atanasov, A. G., Zotchev, S. B., Dirsch, V. M. and Supuran, C. T., Natural products in drug discovery: advances and opportunities. *Nature Rev. Drug Discov.*, 2021, **20**(3), 200–216.
6. Siva, R., Food colourants and health issues: are we aware? *Curr. Sci.*, 2014, **106**(2), 143.
7. Cobb, A. H., *Herbicides and Plant Physiology (Third Edn)*, John Wiley, West Sussex, UK, 2022, pp. 1–400.
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. *et al.*, KNApSAcK family databases: integrated metabolite–plant species databases for multifaceted plant research. *Plant Cell Physiol.*, 2012, **53**(2), 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 (First Edn)*, Elsevier, Amsterdam, The Netherlands, 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), 84.
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. Wang, Z. *et al.*, Combined strategies in structure-based virtual screening. *Phys. Chem. Chem. Phys.*, 2020, **22**(6), 3149–3159.
22. 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. Cheminformat.*, 2017, **9**, 1–15.
23. 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.
24. Yan, X., Liao, C., Liu, Z., 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.
25. 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.
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, **107**, 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. *Nature 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.*, 2020, **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*, 2020, **13**(12), 444.

39. Wearn, O. R., Freeman, R. and Jacoby, D. M., Responsible AI for conservation. *Nature 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. and 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*, 2020, **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. *ISME J.*, 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. USA*, 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. *Nature 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 Biol.*, 2020, **18**(12), e3001026.
55. Miller, S. J. and Clardy, J., Beyond grind and find. *Nature Chem.*, 2009, **1**(4), 261–263.
56. Sugimoto, Y. *et al.*, 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. Aided Chem.*, 2006, **7**, 157–162.
63. Buckingham, J. (ed.), *Dictionary of Natural Products, Supplement*, CRC Press, New York, 1997, vol. 4(11), pp. 1–598.
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. *Nature 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. *Nature 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. *Nature 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. *Nature 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. Nature Prod.*, 2016, **79**(12), 3105–3116.
77. Reynolds, W. F, Natural product structure elucidation by NMR spectroscopy. In *Pharmacognosy*, Academic Press, Boston, MA, USA, 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. *Nature 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 ^{13}C NMR spectroscopic data. *J. Chem. Inf. Model.*, 2020, **60**(7), 3376–3386.
81. Vignoli, A. *et al.*, High-throughput metabolomics by ^{1}D NMR. *Angew. Chem. 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 ^{13}C 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., ^{1}H 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. *Nature Rev. Drug Discov.*, 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 Discov. 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. Divers.*, 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. *Nature Rev. Drug Discov.*, 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 Transduct. Target. 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. Cheminformat.*, 2020, **12**(1), 1–16.
98. Li, X. *et al.*, LSA: a local-weighted structural alignment tool for pharmaceutical virtual screening. *RSC Adv.*, 2019, **9**(7), 3912–3917.
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. Cheminformat.*, 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. Cheminformat.*, 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 IEEE 14th International Conference on Semantic Computing, San Diego, CA, USA, 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 IEEE Fifth Conference on Knowledge Based Engineering and Innovation (KBEI), 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 interaction 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. Bioinformat.*, 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. USA*, 2014, **111**(11), 4067–4072.

122. Keiser, M. J., Roth, B. L., Armbruster, B. N., Ernsberger, P., Irwin, J. J. and Shoichet, K., Relating protein pharmacology by ligand chemistry. *Nature 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 Proceedings of the 6th World Congress of Biomechanics in Conjunction with 14th International Conference on Biomedical Engineering and 5th Asia-Pacific Conference on Biomechanics, Springer, Berlin Germany, 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), 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, 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 IEEE Region 10 Humanitarian Technology Conference, Piscataway Township, NJ, USA, 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. Informat.*, 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 Bioinformat.*, 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, **2019**, 1–7.

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. *J. Chin. Integr Med.*, 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 Bioinformat.*, 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. USA*, 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. Moumbock, A. F. *et al.*, StreptomeDB 3.0: an updated compendium of streptomycetes natural products. *Nucleic Acids Res.*, 2021, **49**(D1), D600–D604.
167. Pilon, A. C. *et al.*, NuBBEDB: an updated database to uncover chemical and biological information from Brazilian biodiversity. *Sci. Rep.*, 2017, **7**(1), 7215.
168. 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.
169. Sorokina, M. and Steinbeck, C., NaPLeS: a natural products likeness scorer – web application and database. *J. Cheminformat.*, 2019, **11**(1), 55.
170. Naghizadeh, A. *et al.*, UNaProd: a universal natural product database for *Materia medica* of Iranian traditional medicine. *Evid. Based Complement. Altern. Med.* (*eCAM)*, 2020, **2020**, 1–14.
171. 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. *Nature Methods*, 2019, **16**(4), 299–302.
172. Madhukar, N. S. *et al.*, A Bayesian machine learning approach for drug target identification using diverse data types. *Nature Commun.*, 2019, **10**(1), 5221.
173. 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.
174. Nickel, J. *et al.*, SuperPred: update on drug classification and target prediction. *Nucleic Acids Res.*, 2014, **42**(W1), W26–W31.
175. 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. *Comput. Meth. Drug Repurpos.*, 2019, **1903**, 281–289.
176. 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.
177. 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.

ACKNOWLEDGEMENT. We thank the management of VIT for their constant support.

Received 5 June 2023; revised accepted 22 November 2023

doi: 10.18520/cs/v126/i1/19-35